

Continuity in Language Change: Implications for the Theory of Grammar

Whitney Tabor
University of Rochester

Running Head: Continuity in Language Change

Thanks to Elizabeth Traugott, Paul Kiparsky, David Rumelhart, Tom Wasow, Robbie Jacobs, Roberto Zamparelli, Mike Tanenhaus, and Phillip LeSourd for helpful comments. This work is a distillation of my thesis which I did at Stanford University. The distillation occurred mostly at the University of Rochester. The work was supported by a university fellowship from Stanford University, by U.S. Federal Work-Study funding, and by postdoctoral fellowship funding from the U.S. National Institutes of Health (NIH-NIDCD 5T32DC0035-04).

Please address correspondence to

Whitney Tabor
Department of Brain and Cognitive Sciences
University of Rochester
Meliora Hall—River Campus
Rochester, NY 14627

Email: whitney@bcs.rochester.edu

Abstract

Connectionist and other language models based on learning from examples, graded categories, optimization, and nonlinear dynamics offer some advantages over standard linguistic treatments because they make explicit hypotheses about how representations are derived from data. However, it is not yet clear what representational revision (if any) these models imply. This paper examines a little-studied phenomenon in language change, “Q-divergence”, which sheds light on this question: prior to many categorical revisions of the grammar there are anticipatory quantitative changes which make the language more similar to the language of the revised grammar. This phenomenon requires nonlinear, similarity-based generalization—something standard grammar models have little to say about. The results indicate the need for a revised theory of grammar which defines distances between representations, employs prototype-centered categories, and uses a smoothness constraint to navigate the realm of nonlinear interpolations. Moreover, the results suggest a solution to the notorious *actuation problem* of historical linguistics: how do grammar changes get started?

1. INTRODUCTION

The last decade has seen a surge of interest in computational models that learn by exposure to examples, that employ categories of a graded nature, that use optimization to find good solutions to problems, and that exhibit nonlinear dynamics. I am thinking, in particular, of Connectionist models, Genetic algorithms, Reinforcement learning, and related systems. Ballard (to appear) adapting a term from Richards (1988) refers to the general approach as “Natural Computation”.

A number of Natural Computation models, primarily Connectionist models, have been applied to problems in natural language, sometimes with some empirical success (e.g., Smolensky et al., 1992; Dell and O’Seaghdha, 1992; Goldsmith and Larson, 1992; Seidenberg and McClelland, 1989; Clark and Roberts, 1991). Nevertheless, a question that remains largely unanswered is what revision of traditional representational hypotheses such models urge. In particular, it is desirable to be able to compare the representational claims of such models to standard linguistic models which are currently much more clearly articulated at the level of abstract categories and rules (see Smolensky et al., 1992). Here I show how investigation of a hitherto little-studied phenomenon in historical language change sheds new light on this representation question.

Language change is a particularly revealing domain in which to examine representation in Natural Computation devices. Any language at a particular point in time seems to have a rigid set of rules governing grammatical usage, but these rules tend to change across time, and the processes of change, observed at the level of the statistical properties of corpora, are often gradual. While rigid rule systems are the forté of traditional linguistic models, gradual evolution is not. Natural Computation devices are appealing because they are able to exhibit rule-like behavior but can also evolve in a continuous fashion.

In fact, statistical gradualness in language change could easily be modeled using standard rule and category structures if it were possible to characterize it simply by associating probabilities with the rules and categories. Indeed such an approach is the basis of Variationist accounts of historical change (e.g., Weinreich, Labov and Herzog, 1968; Kroch 1989). Here, however, I present evidence for the existence of a phenomenon that runs counter to this assumption. Prior to the point at which an evolving linguistic element first takes on categorically new behavior, it may undergo subtle quantitative shifts¹ that make it less like other members of its original type and more like members of the new type it eventually becomes. I refer to this phenomenon as *Q-divergence* for “quantitative divergence”.

Q-divergence is not expected under models of historical change based on standard linguistic theory because such models treat information about sub-categorical (i.e. merely statistical) similarity as irrelevant to the determination of grammatical representation. By contrast, the phenomenon can be modeled well by Natural Computation devices because such models are sensitive to similarity at both the categorical and sub-categorical levels and they permit interaction between the levels. An examination of how these models predict Q-divergence effects provides a clear glimpse of the similarities and differences between them and the standard models. In a nutshell, the analysis suggests the following revision of grammatical theory:

¹By “quantitative shifts” I mean changes in the relative frequencies with which the word occurs in different syntactic environments.

1. Discrete categories are replaced by clusters of points in a continuous representation space.
2. The clusters are organized around prototypes (or central members) in such a way that the distance of a point from the prototype provides a measure of how different its behavior is expected to be from that of a prototypical element. Thus elements that are far from any prototypes are expected to fit poorly into traditional categorization schemes.
3. The space between points in the clusters is filled-in according to an interpolation hypothesis. The interpolated surface or *manifold* is continuous and smooth so many intermediate behaviors are predicted to be possible and nearby elements are expected to show a high degree of similarity. In many cases, the interpolation is curved, which means that the predicted intermediate behaviors are not simple mixtures of existing types but have distinctive characteristics.
4. Syntactic constituents are modeled as sequences of regions corresponding to categories.

In this paper, I concentrate mainly on the interpolation claims (point 3). The key to making plausible nonlinear interpolations is the smoothness constraint. To emphasize the importance of this constraint, I refer to the grammar model I study here as the *Smooth Manifold Model*. I examine a Connectionist implementation of this Smooth Manifold Model.

In short, Q-divergence is quantitative change that anticipates categorical change. The phenomenon amounts to a kind of quantitative hedging of the abrupt, categorical shifts one expects under non-probabilistic discrete-category models of historical change (e.g. Lightfoot, 1979). Q-divergence is often associated with processes of *grammaticalization*, a widespread phenomenon in language change in which meaning-laden elements (e.g., nouns and verbs) evolve into grammatical markers (prepositions, complementizers, auxiliary verbs, affixes, etc.) or grammatical markers evolve into different grammatical markers (see Hopper and Traugott, 1993, for review) so it pertains to general questions about the nature of “grammar” and the difference between “grammatical” and “lexical” elements, a cornerstone of most linguistic theories.

1.1 Examples of Q-divergence

Before examining the Smooth Manifold model in detail, I give some examples of Q-divergence.

Andersen (1987) describes a case of grammaticalization in Polish in which a clitic turns into an inflectional affix, or at least becomes more like an inflectional affix than it was many centuries ago.² The case is that of a person/number/tense marker derived from a verb of being. In Old Polish this marker often occurred in the Wackernagel Position typical of

²A *clitic* is an element that has the distribution of a word but the phonological behavior of a part of a word. An example from English is the reduced form of the auxiliary verb *will* that is written *'ll*. Affixes are elements whose distribution *and* phonology are typical of part-of-word elements. An example is the plural *-s* marker in English (Zwicky and Pullum, 1983). Affixes, but not clitics, are subject to phonological rules which take single words as their domains of application (Inkelas, 1989). It is often said that a clitic is “syntactically independent” but “phonologically dependent”.

clitics,³ and showed all the signs of being syntactically independent. Between the 1300s and the present, the marker gradually came to occur less frequently in the Wackernagel position and more frequently adjacent to the verb (which itself can occur in a variety of positions). Around the 1500s, those instances of the marker that occurred immediately to the right of a verb, began counting as a syllable with respect to a phonological rule which places stress on the penultimate syllables of words (1).

- (1) 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.'

This rule generally applied (and still generally applies) only to word-units, so the change suggests that the instances of the markers that occur next to verbs have switched status, and should now be classified as inflectional endings rather than clitics. Thus this is a case in which two quantitative trends (increasing adjacency to the verb and increasing deviation from Wackernagel placement) are correlated with and partly precede a categorical development (onset of participation in the penultimate stress rule). One of the distinguishing features of many words is that their parts usually have to occur adjacent to one another in a certain fixed order; thus, the trend toward right-hand adjacency to the verb made the marker more affix-like. Likewise, the increasing deviation from the Wackernagel position made the marker less clitic-like. Hence there is a clear sense in which the quantitative changes led to the qualitative one.

A second example of Q-divergence involves a major shift in lexical category. Modern English has two uses for the expressions *sort of* and *kind of*. They can be Noun-Preposition sequences (2) or they can be Degree Modifiers (3).

- (2) What sort/kind of jacket did you bring?
- (3) It is sort/kind of windy today.

In Tabor (1994a) I provide evidence that prior to the first attestations of unambiguous Degree Modifier usage, the rate of use of *sort/kind of* before adjectives in constructions like (4) rose dramatically.

- (4) c. 1675 The way [is] pretty good except 4 or 5 miles they call the Severalls, a sort of deep moore ground and woody. *Great Journey*, p. 144 [HELS]

Before the Degree Modifier usage arose, the increasing cooccurrence of *sort/kind of* with adjectives made the expressions much more like other Degree Modifiers (e.g., *rather*, *somewhat*, *quite*) since such degree modifiers often modify adjectives. Again, the quantitative trend preceded and very plausibly facilitated the categorical development.

³The *Wackernagel Position* is the position after the first constituent of the clause. What counts as a constituent is somewhat complex—see Halpern (1993) for further discussion.

A third example of Q-divergence is provided by a set of synchronic data. Craig (1991) argues that several cognate relational markers in Rama, a Chibchan language of Nicaragua, appear in three distinct syntactic environments: as independent postpositional elements with an adjacent overt NP (5); as “clitic preverbs” without an overt NP present (6); and as “lexical preverbs” with an overt NP present (7).

- (5) Nsu-suluk **u** angka nsu-uung-i
 our-finger PostP/with can't 1Pl-make-SUB
 'With our fingers we can't do it.' [C 465]
- (6) 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]
- (7) ngulkang banku **yu**-an-siik-u kaing
 wild pig now LexPV/with-3Pl-come-Tns Disc.
 'They brought the wild pig now.'

Craig distinguishes the second and third types from the first on two counts: (a) they are marked with a different form (*yu* versus *u* in the case 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 idiosyncractic semantics.

Given these criteria for distinguishing the types, Craig uses a set of spoken narratives to study the distribution of the relational markers quantitatively. Although these data are based on 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 common in grammaticalization (Givon, 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 1. The different markers are placed along the horizontal axis in the order that is most plausible given the surface gradualness of linguistic change.⁴ Interpreting the five ordered markers as successive stages of a grammaticalization process, we find evidence that prior to the development of Lexical Preverb (L) behavior, the distribution over Postposition (P) and Clitic (C) behaviors shifts in the direction of greater Clitic behavior. Since the Clitics are more similar to the Lexical Preverbs than the Postpositions in their positioning with respect to the verb and in their phonological form it is plausible that the increase in the Clitic usage facilitates the appearance of the Lexical Preverb behavior. These data involve small numbers of tokens, so

⁴The five 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.

they cannot be taken as substantial evidence for Q-divergence. Nevertheless they illustrate the hypothesized phenomenon clearly.

Insert Figure 1 about here

1.2 Historical Texts and the Theory of Grammar

The evidence I present below for Q-divergence comes from a sequence of historical English texts separated by fairly small time intervals (roughly 50 years on average). The texts are listed in Table 2. I chose the texts to span a period of time during which the expression *be going to* underwent some significant distributional changes. The texts are literary—mostly novels, although Shakespeare’s plays have been included since they were the only substantial, available corpus covering the late 16th century. I included only texts by authors from England on the assumption that this would reduce distortion of the data due to dialectal contrast, although there is certainly still substantial variation among them in this regard. My arguments here are based on the assumption that these historical texts tell us something about the changes in English speakers’ cognitive representations over a period of several centuries.

Insert Table 2 about here

There are, of course, some difficulties with using written texts to gain insight into cognitive representations.

On the basis of the gradual statistical developments observed in the texts, I claim that the cognitive representations of successive individuals correspond to points on a smooth function. But one must be careful, in making such an assertion, to rule out the possibility that the texts representing each period are a mixture of the outputs of multiple speakers whose grammars are underlyingly categorical (not continuous) and that what is changing gradually over time is the relative percentages of speakers with different grammars (see Bailey, 1973). For this reason, I have used works written by one author to represent each point in time. Of course, one might still be concerned about the fact that authors often draw from a variety of dialects, even in the course of a single work. However, if such a tendency to mix dialects were responsible for the persistent change in the distribution of *be going to*, then we would expect there to be a systematic change in the rate of use of different dialects across the authors that correlates with the grammatical change. This is not likely. Nearly all the authors employ a mixture of literary language and colloquial speech. Most of them employ essentially one form of colloquial speech and one literary voice. Defoe’s *Moll Flanders* (dated at approximately 1695) is exceptional in being presented as written entirely in the vernacular of one female commoner. However, even this extreme case fits in well with the overall quantitative trend (see Figure 4).

One may also question the use of written texts to study cognitive representations on the basis of the belief, widespread in generative linguistics, that spoken language produced

in an unselfconscious manner in a fairly casual setting is the best source of information on linguistic cognitive representations. Unfortunately, such data are not available for historical periods. It would have been preferable, from this standpoint, to use a written genre that reflected the most casual style—perhaps diaries, or plays written exclusively in the vernacular. But, because I needed to evaluate a range of statistical claims, it was important to have electronic access to a genre in a reasonably large quantity. Fictional prose is currently the only substantially available on-line historical genre.

Although it is quite different from spoken language at any point in time, written language, including fictional prose, clearly provides a systematic reflection of speech behavior. It is generally the case that persistent changes in the spoken language spread to the written language after a time lag. Thus, the written language may not reflect the casual output of speakers at the time of writing, but it may still show the same sequence of changes over a period of time. It may be, however, that written language is nonuniformly censored with respect to the spoken language. In a study of variation in currently spoken Brazilian Portuguese, Naro and Lemle (1976) and Naro (1981) found evidence that the degree of saliency of a divergence from received usage (measured in terms of phonological contrast and stress) was a predictor of the order in which changes occurred: less salient changes were invoked before more salient ones. Perhaps something similar is true of the relationship between written and spoken languages. It is certainly plausible that mere changes in the frequencies of existing grammatical constructions are less salient than changes in the natures of the constructions themselves (Kiparsky, 1978). Such a situation would seem to be especially damning for the Q-divergence hypothesis. Perhaps quantitative changes occur before categorical ones in the written language simply because writers are initially censoring the output of any part of their spoken language grammar that reflects a categorical divergence from the received way of speaking. Only when the frequency of the new forms becomes sufficiently high in the spoken language does it begin to impinge on written usage. But this possibility does not actually mitigate the strength of the claim about cognitive representations: if writers can tell which aspects of their spoken language are divergent, then their representations must be coding the differences. Moreover, if the degree of censorship undergoes a relatively smooth quantitative gradation in individual representations over time (as I find in the data I present below), then the censorship mechanism must be employing some kind of continuous representation. The Smooth Manifold model treats the stratification of language into writable and speakable registers as analogous to the stratification of language into currently legitimized language and “soon-to-be-invented language”. Thus by studying the flux across the writable versus speakable divide, we can also learn something about the process of innovation in spoken language.

1.3 Overview

Insert Figure 2 about here

Insert Figure 3 about here

The gist of the Smooth Manifold account of Q-divergence is as follows. Instead of modeling categories as boxes into which elements are placed according to type, we model them as clusters of points in a continuous space (Figure 2). Such a continuous space is called a *connected manifold* in differential geometry. Following colloquial usage, I'll refer to it simply as a *manifold*. The large-scale relationships between the clusters on the manifold recapitulate the information in a standard, discrete-category model. The small-scale relationships within the clusters reflect quantitative differences between items belonging to the same category. It is thus evident how such a model predicts that persistent gradual quantitative change can lead to categorical change: such change can move an element out of one cluster and into another one. Moreover, a manifold of a given finite dimension may be embedded in a space of higher dimension and be curved in various ways (Figure 3). When category clusters are on different sides of a curve in a manifold, and a linguistic item makes a gradual transition between the clusters along the manifold, then certain features of the item's behavior undergo significant quantitative change before others do (the different dimensions of the containing space correspond to different behavioral features). Q-divergence effects stem from this property: currently grammatical features change in frequency before new features become grammatical.

The Smooth Manifold Model has the potential to shed light on one of the most challenging questions in historical linguistics: How do grammar changes get started? (Weinreich, Labov, and Herzog, 1968). By positing that all events of language usage are miniscule events of language change, the model makes it comprehensible how substantial categorical changes can result from a series of completely ordinary linguistic events. A fundamental difference between the standard, discrete-category models and the Smooth Manifold Model is that the standard models can interpolate only linear manifolds while the Smooth Manifold Model permits curved ones. There are many ways to interpolate nonlinearly between the members of a set of points. Thus, in adopting a specific nonlinear interpolation claim in a principled way, the Smooth Manifold Model is making a strong, falsifiable claim about the nature of language representation. In this paper, I offer empirical evidence that this claim is close to the mark for one transitional episode in the history of English.

The remainder of this paper is organized as follows. Section 2 provides a detailed look at one historical episode involving Q-divergence: the development of the future auxiliary usage of English *be going to* (e.g., *This madness is going to be the death of me.*) from its original motion usage (e.g., *The abbot is going to Manasarovar.*). Section 3 develops the Smooth Manifold Model, and shows how it makes appropriate predictions about the case study. Section 4 summarizes the paper, identifies potential sources of further empirical evidence in both diachronic and synchronic linguistics, and shows how the model sheds new light on fundamental theoretical questions.

2. CASE STUDY: ENGLISH *be going to*

Pérez (1990) and Bybee, Pagliuca, and Perkins (1991) note that motion verbs have developed, via grammaticalization, into markers of future tense in many diverse languages. Example (8) shows typical, earlier uses of *be going to* as a motion verb. Example (9) shows typical later uses of *be going to* as a future auxiliary. In accord with all previous research on the topic, I find evidence that the motion use with a Noun Phrase (NP) complement

(8a) predates the motion use with a Verb Phrase (VP) complement (8b). And, in accord with Pérez (1990), I find evidence that the earliest auxiliary instances had something like *Equi* status, where *be going to* plausibly ascribed intention to its subject (9a), while most later auxiliary instances are clearly *Raising* constructions, where intention is not part of the meaning (9b).

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

The systematic difference between “Equi” and “Raising” verbs was recognized early in the history of transformational linguistics (e.g. Rosenbaum, 1967). Criteria often considered diagnostic of Raising status for English verbs include ability to take “dummy” subjects (*It seemed/appeared/tended to rain.*, *There seemed/appeared/tended 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 Nunberg, Sag, and Wasow, 1994). If *be going to* is interpreted as an Equi verb, it seems to have a meaning similar to that of the verb *intend* (e.g. *He was going to speak, but then he thought better of it.*) Thus, in evaluating historical examples below, I consider whether substitution of *intend* significantly changes the meaning. Agentivity of the embedded predicate is also a pertinent feature: Equi verbs tend to occur with agentive predicates like *reply* and *seek*;⁵ Raising verbs can occur with such predicates but they often occur with nonagentive predicates like *be able* and *rain* as well.

2.1 Historical Emergence by First Examples

The syntactic and semantic distinctions help identify a list of featural contrasts which are useful in tracing the history of *be going to* systematically: NP (Place) complement vs. VP complement, motion sense versus future sense, agentive vs. nonagentive embedded predicate, sentient vs. nonsentient subject, and, within the nonsentient subject cases, dummy subject vs. nonsentient “thing” subject.⁶

To give a sense of the process of emergence, this section reviews the times of first attestation of various combinations of values of these features. From this picture, it will be possible

⁵By *agentive predicates* I mean predicates where conscious will or intention is part of the meaning.

⁶It is probable that a more refined treatment of the data could be achieved by the inclusion of a feature for immediate future (10) versus general future (11).

to identify three successive intervals during which *be going to* was, respectively, only a Motion Verb, both a Motion Verb and an Equi Verb, and both a Motion Verb and a Raising Verb. In keeping with the hypothesis that linguistic categories have prototype structure, the data indicate somewhat blurry boundaries for these intervals.

As noted above, the earliest attested *be going to* type is the Place-complement type. When this is the only type, *be going to* is clearly a pure motion verb. (12) and (13) give Old English and Middle English examples respectively.⁷

(12) 855 thu ... 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)]

(13) 1450 To abide in prison...without goyng to bayle, abston, or mainprise ... *RParl.*
 5.201a [MED, Pérez (1990): 56]

The earliest known VP-complement cases (14) occur in Middle English. All have agentive VP complements and sentient subjects.

- (14) a. (early 1300s) Phillip (...) was going too the ouer Greece, *King Alisaunder*, 1.901
 [Danchev and Kytö (1991): 3]⁸
 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 going to see the Queen’s picture. Shakespeare, *A Winters Tale*, V ii. [OXF]
 d. c. 1590 I am going to visit the prisoner, Fare you well. Shakespeare, *Measure for Measure*, III i.
 e. c. 1675 When we were going to fight the Dutch, I had such a paine in my right arme that could not use but very litle. *Private Letters*, 3.15.

Note that the later cases of this type (c–e) permit *intend* to be substituted for *go* without too severe a distortion of the meaning. Thus, they are arguably early Equi instances, although the context in each case involves people in the act of travelling somewhere so the motion interpretation seems most likely.

(10) c. 1730 Adams was going to answer, when a most hideous Uproar began in the Inn. Fielding, *Joseph Andrews* [Match 4].

(11) [W]e all felt that we were going to be only half a mile apart, and were sure of meeting every day. Austen, *Emma* [Match 3].

It looks impressionistically as though immediate future senses of *be going to* were especially common in the early stages of the emergence of the future usage. I hesitated when I began this project to try to code this feature because I thought that the judgements would be too subjective. In retrospect I think enough case are probably clearcut that some useful statistics could be gathered.

⁷Middle English is taken to be the period from around 1150 to 1500.

⁸This example has a VP reading if Mossé is right in claiming that *the* is a version of OE *theon* ‘thrive’, but Danchev and Kytö (1991) are skeptical (p. 3).

In the 16th and 17th centuries, the first instances in which a motion interpretation is not plausible occur. Note that all these early cases have sentient subjects and agentive embedded predicates. *Intend*-substitution is again felicitous in these clear-cut Equi instances.

- (15) a. 1567 when you are going to lay a tax upon the people, Burton, *Parl. Diary* [Danchev and Kytö (1991): 7]
 b. 1695 He was going to reply... but he heard his sister coming, Defoe, *Moll Flanders* [OTA]
 c. 1699 Gad, I have forgot what I was going to say to you. Congreve, *The Way of the World*
 d. c.1703 The council sat upon it, and were going to order a search of all the houses about the town. (Burnet, *Burnet's history of my own time* p. 1, II, 163-4) [D&C, p. 13]

Persuasive evidence that 17th century *be going to* could be used as a future auxiliary verb occurs in a direct comment about grammar:

- (16) 1646 About to, or going to, is the signe of the Participle of the future. . . : as, my father when he was about [to] die, gave me this counsell: I am [about]; or going [to] read (Poole 1646: 26; square brackets in source text, emphasis added)
 [D&C, p. 12]

There is one very early instance of *be going to* with a sentient subject and a nonagentive embedded predicate:⁹

- (17) 1482 ...[W]hile thys onhappy sowle by the vycторыse pompys of her enmyes was goyng to 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 minystrys of the deuyll. ware dullyd and made onmyghty... *The Revelation to the Monk of Evesham* [p. 43]

Since *intend*-substitution is quite implausible here, this may be an early Raising instance. Such an early Raising example might be taken as evidence that *be going to* was reanalyzed as a Raising verb long ago, soon after the motion verb *go* began appearing with *to*-marked infinitive complements, and that it never went through any intermediate Equi-like stage. The context of the example does not completely corroborate this interpretation for it is also compatible with a motion interpretation: it is describing a scene in which a crowd of wicked spirits is driving the soul of a dead woman through an earthly town on the way to Hell. But even if this is a Raising instance, given that it seems to be the only such instance among many others, then it is not of great consequence for a theory like the Smooth Manifold model in which grammatical representations are formed on the basis of statistical evidence.

The next sentient subjects with nonagentive predicates occur in the 17th century:

- (18) a. 1628 He is fumbling with his purse-strings, as a School-boy with his points, when hee is going to be Whipt (Earle, *Micro-cosmographie*, p. 71) [D&C, p. 12]

⁹Before Danchev and Kytö found instance (14b) above, (17) was cited by most researchers as the first known case of *be going to* with a VP complement.

- b. c. 1675 my Unckle is going to 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]
- c. c. 1695 How little does he think that having Divorc'd a Whore, . . . he is going to Marry one that has lain with two Brothers, Defoe *Moll Flanders*.
- d. c. 1695 When a malefactor, who has the halter about his neck, is tied up, and just going to be turned off, and has a reprieve brought to him - I say, I do not wonder that they bring a surgeon with it. . . Defoe, *Robinson Crusoe*.

Example (18a) may have a motion sense. Example (18c) is perhaps best described as a borderline case since *marry* in the active voice is typically used agentively, but the embedded predicate as a whole clearly does not describe the intention of the marrier. Since *intend*-substitution is implausible in these cases, these are arguably early Raising instances. However, it is questionable to claim Raising status for a verb on the basis of cooccurrence with nonagentive predicates alone, because many Equi verbs can also occur in such contexts (e.g., *Jeremiah expected to fall. Samantha yearned to be chosen.*).

The first indisputable Raising examples, those with nonsentient subjects, occur in the 18th century (19). Nonsentient subjects inevitably take nonagentive complements.

- (19) a. c. 1730 She was well contented that other violent methods were now going to be used in favour of another man. Fielding, *Tom Jones*
- b. c. 1796 standing side by side, exactly as if the ceremony were going to be performed. Austen, *Mansfield Park*, p. 88
- c. c.1796 Something is going to happen. . . . Austen, *Mansfield Park*, p.25
- d. c. 1865 She went down to look about her, and to wonder what was going to happen next. Carroll, *Alice in Wonderland*

The first dummy subjects occur in the mid 19th century (20). Here also, nonagentive predicates are inevitable and Raising status is certain.

- (20) a. c. 1865 Do you think it's going to rain? Carroll, *Alice Through the Looking Glass*
- b. 1890 It seems as if it were going to rain. *Chamb. Jrnl* 14 June 370/2 [OED]
- c. c. 1894 There is going to be a shooting and somebody is going to get hurt. Doyle, *Sherlock Holmes* v. 1, p. 568.
- d. c. 1911 Mr. Bloom looked back towards the choir. Not going to be any music. Pity. Joyce, *Ulysses*

Except for the unusual nonagentive case in 1482 these data argue for the chronology shown in Table 2.

Insert Table 2 about here

This review provides two kinds of evidence for categories structured around prototypes: (i) uncertain classifications at the Motion/Equi and Equi/Raising boundaries; (ii) the fourth, fifth, and sixth types constitute progressively more distinctively Raising structures.

However, it is hard to be very confident in first-instance data alone: since the frequencies of new forms are generally low, the variance across samples of the observation-times of first forms is high. Moreover, first-instance data do not allow us to distinguish two significantly distinct claims about grammar change: (1) all new forms actually become grammatical at the same time, but we observe the new forms associated with different syntactic environments in series because some environments are more frequent than others, and (2) the new forms appear in series because they become grammatical in series. I'll refer to the first hypothesis as the *simultaneous emergence* claim and the second hypothesis as the *sequential emergence* claim. The distinction between these two hypotheses is conceptually very similar to the distinction between what Kroch (1989b) calls *simultaneous actuation* versus *sequential actuation*. The difference is that *emergence* refers to the period of time when the most rapid influx of a new form occurs (the middle of the S-shaped curve that is typical of relative frequency changes), while *actuation* refers to a hypothesized point early on when a discrete change in the grammar occurs and the S-shaped curve jumps off the zero line. Under the Smooth Manifold model, there is no actuation, and the interest of the model is the sequential emergence predictions it makes. In Section 4.2.1.1 I explain why it makes sense to emphasize emergence rather than actuation. It turns out that Q-divergence depends on the possibility of sequential emergence. Therefore, it is important to be able to distinguish the sequential from the simultaneous emergence claims. In this regard, a more revealing picture can be achieved by examining the *be going to* development quantitatively.

2.2 Quantitative Characterization

Insert Table 3 about here

Table 3 provides a quantitative characterization of the history of *be going to* based on the corpus sample described in Section 1.2. To facilitate coding, I used the featural decomposition developed in the previous section. Every example of *be going to* that occurred in the corpora I sampled fell into one of the 12 feature bundles listed across the top of the table. The uncertain classification marked by the 6 question marks consisted of items for which even the surrounding context did not make it clear what all the relevant feature values were. Fortunately, these items comprised only a small portion of the sample (a maximum of $2/64 = 3.1\%$ per source). For columns 1 through 11, the values in each row express the percentage of the total number of definitively classified instances for the time period corresponding to the row. Thus columns 1 to 11 sum to 100% in each row.¹⁰

Insert Figure 4 about here

¹⁰Columns 4 and 5 in this table correspond to patterns that *be going to* never exhibited but which other motion verbs, shown in Table 5 and discussed in Section 3 below, do exhibit. I include them here to facilitate comparison between the two tables.

Figure 4 provides a graphical summary of the data by collapsing the 11 classes into four significant ones: Place complement (<P>), Sentient Subject with Agentive VP complement (<A>), Sentient Subject with Non-agentive VP complement (<S>), Non-Sentient Subject with Non-agentive VP complement (<N>). The figure provides confirmation of the chronology developed in the previous section. A good way of assigning a single date to the “heyday” of a particular construction (assuming it has only one heyday) is to use the center of mass.¹¹ The centers of mass of the four constructions graphed in Figure 4 are shown in Table 4. Although better estimates of the true heydays would be obtained by examining a wider interval, the short interval is sufficient for getting an estimate of the relative chronology. Note that Table 4 is consistent with the data based on first attestations.

Insert Table 4 about here

Figure 4 vaguely suggests Q-divergence: (i) prior to the first significant appearance of sentient subjects with nonagentive complements (<S>) (1695), there is a rise in the frequency of agentive complement (<A>) constructions; (ii) prior to the first significant appearance of nonsentient subjects with nonagentive complements (<N>) (1730), there is a rise in the frequency of sentient subjects with nonagentive complements (<S>). The quantitative divergences plausibly “facilitated” the categorical changes since (i) increasing agentive complement (<A>) use, and hence VP-complement use in the 17th century makes *be going to* look more and more like an Equi auxiliary, and Equi auxiliaries typically show some sentient subject, nonagentive complement (<S>) behavior and (ii) increasing sentient subject, nonagentive complement (<S>) use makes *be going to* look more and more like a Raising Verb and Raising verbs typically show some nonsentient, nonagentive (<N>) behavior. Other evidence provides additional support for case (i): OE and early ME *be going to* seems to have had a more typical motion verb meaning so its distribution was probably closer to that of canonical motion verbs like *walk to*, *run to*, *move to*, etc. A tabulation over all the corpuses of *walk to*, *run to*, and *move to* places their average rate of agentive complement (<A>) behavior at 13.3% and their average rate of place complement (<P>) behavior at 86.7% (s.d. = 9.1). These verb+*to* collocations show essentially no nonsentient subject (<S> or <N>) behavior. Thus, it is reasonable to suppose that the <A> curve in Figure 4 rose quite a bit and the <P> curve came down quite a bit prior to the leftmost point on the graph.

Although the time-course analysis of Figure 4 is suggestive of Q-divergence, the graph looks very noisy. The noise might be due to measurement error, but it might also be due to variance in the grammatical progressiveness of the writers I have sampled. If this latter hypothesis is correct, then despite the large fluctuations in the frequency versus time plots, we should see consistency within individual writers—a high rate of Place-complement (<P>) usage should imply low rates of Nonagentive (<S> and <N>) usage; a high rate of Agentive (<A>) usage should imply intermediate rates of Place and Nonagentive usage; a high rate of Nonagentive usage should imply relatively low rates of Agentive usage and Place usage. It’s hard to discern to what extent these claims are valid from Figure 4. The next section will develop a method of graphical analysis which will allow us to sort out the effects of writer

¹¹By *center of mass* I mean $\sum f_i t_i / \sum f_i$ where f_i is the relative frequency at time t_i .

progressiveness from other noise. The same method will make it possible to distinguish the sequential and serial emergence hypotheses.

3. IMPLICATIONS FOR THE THEORY OF GRAMMAR

3.1 The Standard, Discrete Category Model

The methods of this section will allow us to make some precise claims about the implications of Q-divergence for the theory of grammatical representation. To address this issue it is useful to look first at standard grammatical models and see why they fail to predict the effects. In fact, standard generative syntactic models (e.g., Categorical Grammar, GB, HPSG, LFG, Relational Grammar, etc.) do not normally make any claims about quantitative properties of language use, so they trivially fail to make predictions about Q-divergence. But there is a natural augmentation of any generative model that turns it into a model of quantitative distributional data: one simply assigns a real-valued probability to each production of each generative rule or parameter. Examples are the Probabilistic Context Free Grammars (PCFGs) often used in computational linguistics (see Charniak (1993)), the Competing Subsystems models of Kroch and colleagues (Kroch, 1989a, 1989b; Santorini, 1989; 1992; Pintzuk, 1991; Taylor, 1992; Fontana, 1993) and the Variable Rule models studied by Variation theorists generally (Weinreich, Labov, and Herzog, 1968; Labov, 1969; Cedergren and Sankoff, 1974). Since standard generative grammars treat category membership as an all-or-nothing thing, whether or not there are probabilities attached to the generative rules, I refer to a sequence or continuum of generative grammars as a “discrete category diachronic model”.

Consider, in this light, the case of *be going to*. The changes that occur in this construction are syntactically non-trivial in the sense that an adjunct (the purpose clause of the motion verb construction) seems to have given rise to a subcategorized constituent (the infinitive clause of the Raising construction), and a multi-morpheme sequence (*go ing to*) may be becoming monomorphemic (as evidenced by its special phonological reduction behavior: *gonna*). Unfortunately, there are two impediments to examining Smooth Manifold representations of these complex developments: (i) the theory of constituent structure representation in Smooth Manifold models is very much in its infancy (see Smolensky 1990; Elman, 1991; Servan-Schreiber, Cleermans, and McClelland, 1991; Plate 1994) and (ii) if one uses a reasonably plausible Smooth Manifold representation of syntax (see Section 3.3.3 below), interpreting its Q-divergence behavior involves a fairly complex analysis of the model. Therefore, to provide a clear exposition, I start by examining a highly simplified model in which the historical episode is treated as a case of mere *lexical reclassification*: *be going to* is treated as a single word which used to be only a motion verb but is now a motion verb and an auxiliary verb. In Section 3.3.3, I show how the results extend to a more complex case with a more realistic treatment of syntax.

Insert Figure 5 about here

There is an obvious way of modeling lexical reclassification in a discrete category framework. We assume that prior to some time t_1 , word w is only a member of lexical class C_1 .

After some time t_2 , word w is only a member of class C_2 . Since there is usually an intermediate phase during which the element shows mixed behavior, we can posit that $t_2 > t_1$ and that during the intervening period, w is a member of both classes. Figure 5 illustrates for the case of *be-going-to* with a hypothetical probabilistic discrete-category grammar. Note that we must assume that present-day *be going to* is still in the intermediate phase since it occurs as both a motion verb and an auxiliary verb. The central hypothesis of the discrete category model is that lexical reclassification episodes can be modelled by specifying two functions of time: the probability, $P(w|C_1)$, that C_1 is realized as w and the probability, $P(w|C_2)$, that C_2 is realized as w .

3.2 Discrete Category Models do Linear Interpolation

It turns out that no matter what the functions $P(w|C_1)$ and $P(w|C_2)$ are, the discrete category model treats lexical reclassification as *linear interpolation* in the space of assignments of probabilities to syntactic structures. To see this, it is useful to consider a low-dimensional slice of this space. Consider, for example, the place complement behavior ($\langle P \rangle$), the agentive complement behavior ($\langle A \rangle$), and the union of the two nonagentive complement behaviors discussed in section 2.2 ($\langle S \rangle \cup \langle N \rangle$). I'll refer to this 3-dimensional portrayal as the "simple *be going to* data".

Insert Figure 6 about here

Insert Table 5 about here

Recall that one of the hypothesized cases of Q-divergence described above involved an increase in the relative rate of use of agentive VP complements ($\langle A \rangle$) prior to the appearance of the first nonagentive VP complements ($\langle S \rangle \cup \langle N \rangle$). It is thus useful to plot the relative frequencies of the three types, $\langle P \rangle$ (Place), $\langle A \rangle$ (Agt), and $\langle S \rangle \cup \langle N \rangle$ (NonAgt), at each point in time. Since, the conversion to relative frequencies eliminates one degree of freedom, the points all lie in the two-dimensional triangular region with vertices (1,0,0), (0,1,0), and (0,0,1). Figure 6, showing just the triangle, identifies the positions corresponding to typical Motion Verbs ("MOTION") and typical Raising Auxiliaries ("RAISING"). I ignore Equi Auxiliary verbs for the moment in order to make the illustration simple, returning to them in Section 3.3.2. The figure is based on the numerical data in Table 5.¹² Note that the average locations can be thought of as prototypes of the Motion and Raising classes.

¹²The data are based on a sample of approximately 25 verbs per corpus for each Raising verb, approximately 20 verbs per corpus for each Equi Verb, and approximately 8 verbs per corpus for each Motion Verb. The Motion and Equi samples are smaller because it was hard to find very many of them in the texts. The Motion and Equi samples also draw less heavily on the first two corpora (Shakespeare and Defoe) than on the five more recent corpora. Despite these shortcomings, there seems to be relatively little variation in the distributions of these types of verbs in the relevant categories over the 300 year period of interest, so I have assumed that Figure 6 is a reasonable approximation of the states of these two types throughout this period.

Following the lexical reclassification hypothesis, I assume that these two points represent the starting point and the (hypothetical, future) ending point of *be-going-to*'s transition. My claim is that the discrete category model implies a linear trajectory between these two states.

To see why this is so, note that under the discrete category model, *be-going-to*'s rate of appearance in the three clause types depends solely on the values of $P(\textit{be-going-to}|\textit{V}[\textit{Motion}])$ and $P(\textit{be-going-to}|\textit{V}[\textit{Raising}])$, which are specified in the grammar. Thus the location of *be going to* in the state space can always be written as a linear combination of the positions of MOTION and RAISING where the coefficients are positive and sum to 1 (see Appendix 1 for a formalization). Therefore, the only possible transitional states of *be going to* lie on the line segment connecting the start state and the (hypothetical) end state (Figure 7). This result depends, of course, on the assumption that *be going to*'s behavior is always a probabilistic mixture of just the two categories, $\textit{V}[\textit{Motion}]$ and $\textit{V}[\textit{Raising}]$. If more categories are involved the standard model can accommodate a wider range of behaviors but it becomes undesirably permissive (see Section 3.3.2).

Nonlinear interpolation implies sequential emergence. If *be going to* follows a nonlinear trajectory between two states, then there must be exchanges in the sizes of the relative rates of change of different behaviors: first one change will make a lot of progress and another only a little—later, the second change will make a lot of progress, and the first only a little—this follows from the assumption that trajectories are continuous in relative probability space (see Appendix 2 for a proof).

Insert Figure 7 about here

What does *be going to*'s trajectory in the state space look like? If we plot the corpus data described in Section 2.2 in the state space of Figure 6, the result is Figure 7. This diagram makes it evident that *be going to* has undergone a significant skewing in the direction of using Agentive VP complements on its way toward the most recent position at which I measured it (1907). Note that the first Q-divergence phenomenon discussed in Section 2.2 is a feature of this general skewing: up through the time of Shakespeare (c. 1590), *be going to* shows no evidence of being able to combine with a nonagentive complement, and yet its position at Shakespeare's time is displaced significantly away from the canonical motion verb position along the Place \leftrightarrow Agt axis. The linear discrete category model (dotted line) fails to predict this skewing.

Insert Figure 8 about here

We might ask, then, what kind of model could do a better job. Clearly we want a non-linear model. But, unfortunately, while there is only one linear interpolation between two points, there are infinitely many nonlinear interpolations. What kind of theory will pick a reasonable one? Here the notion of prototype-structure is helpful. Note that the representation of all motion verbs by one point and all auxiliary verbs by one point in Figures 6 and

7 is a simplification. If we plot the motion verbs and auxiliary verbs individually, they form clusters centered around these points (Figure 8). Moreover, for each type, all the variation in the clusters is along dimensions corresponding to behaviors allowed by the grammar. There is no variation along dimensions disallowed by the grammar.¹³ Conveniently, these reduced-dimension prototype-centered clusters are approximately alligned with the change trajectory we want to predict. Thus, if we restrict the interpolation space to one dimension and adopt some kind of optimizing, smooth-interpolation model, we will get an interpolation that looks much more like the historical data (Figure 8—solid curve). We want smooth interpolation so we don't predict abrupt changes in the direction of a trajectory. We want optimization (e.g. maximal smoothness) to ensure that skewing is in the direction of the ambiguous (Agt) behavior rather than, say, toward the Place↔NonAgt axis.

One kind of optimizing smoother is a Connectionist network. In the next section, I describe a network that is suitable for modeling the phenomenon at hand. Admittedly, Connectionist networks are very powerful curve-fitters, absurdly so if the goal is merely to fit the simple *be-going-to* data just described. But their additional power is desirable in the syntactically more realistic case which I consider in Section 3.3.3.

3.3 Smooth Manifold Models do Nonlinear Interpolation

A *connectionist model* or *neural network*¹⁴ is a particular type of Natural Computation model that is useful for modeling cognitive processes. Inspired by information gleaned from neurology about the structure of the brain, connectionist models consist of groups of nodes (like neurons) with directed connections (like synapses) between them. Each node is associated with a number called its *activation* and each connection is associated with a number called its *weight*. When no sequence of successive, directed connections loops back on itself, the network is called a *feedforward* network. Otherwise, it is called a *recurrent network*. Over time, in a process called *settling*, each unit computes the weighted sum of the activations of the units on lines feeding into it and adopts a new activation value that is a function of this sum. An appealing property of certain types of networks is that they can be made to approximate a function by adjusting their weights incrementally on the basis of a sample of input-output pairs from the function. This process is called *training* or *learning*. A function-approximating network needs units to represent inputs and outputs. It is often useful to give it additional units as well, called *hidden units*, which can be used to encode higher-order relationships between the inputs and the outputs. These hidden units are of particular interest for the theory of language representation because the representations they adopt play a role similar to that of the abstract category and rule structures of linguistic theory.

Consider the feedforward network shown in Figure 9. I trained this network in the following way: words were given localist representations on the input layer¹⁵ and behaviors

¹³Even in a corpus of natural speech like that a language learner is exposed to during the process of linguistic maturation (as opposed to the written texts I am using here), it is quite likely that variation along grammatically allowed dimensions (what we might call *competence-internal* variation, following Chomsky, 1957) is much greater than variation along grammatically disallowed dimensions (what we might call *competence-external* or *performance* variation.)

¹⁴See Rumelhart and McClelland, 1986; Hertz et al., 1991; Haykin 1994.

¹⁵By a *localist representation* on a layer I mean a layer with one unit “on” and all the others “off”.

(e.g., the behavior of occurring in a particular position in a particular sentence frame) were given localist representations on the output layer. The relative probability of each input-output pair was the observed relative frequency of the item-with-behavior in the world. For each input unit, I wanted the activation of each output unit to converge on the probability that the item corresponding to the input would exhibit the behavior corresponding to the output. Thus, the desired output activations formed a probability distribution. Under these conditions, a network with softmax output units, fixed-sigmoid hidden units, trained using the delta rule (and backpropagation of error) is appropriate (Rumelhart et al., 1995). See Appendix 3 for a formal description of the network architecture.

3.3.1 The Simple *be going to* Data

Insert Figure 9 about here

To model the simple *be going to* data, the network needs just three output units, one corresponding to each of the three construction types, <P> (Place), <A> (Agt), and <S> \cup <N> (NonAgt). It needs one input unit for each of the verbs that are relevant to the problem, i.e., one for each motion verb and raising auxiliary verb (Figure 9). Section 3.2 showed that a smooth one-dimensional curve can fit the data reasonably well. This indicates that one hidden unit will suffice. The mapping involved here is very simple and the network has no trouble learning it. In what follows, I study the representation adopted by the trained network.

Insert Figure 10 about here

In the standard grammar models, it is clear what it means to say *be going to* underwent “lexical reclassification” because there are categories in the grammar corresponding to the two classifications involved, V[motion] and V[aux]. In the Connectionist model, it is not, at first, apparent what lexical reclassification could mean since the categories themselves are not built into the architecture. But the categorical structure is recapitulated as an emergent property of the trained network: after learning, the hidden locations associated with each verb are clustered into two groups corresponding to the motion verbs and the auxiliary verbs (Figure 10). Moreover, there is a natural architectural analog of lexical representation in the network: the weights from each input unit to the hidden layer encode the information that is specific to each verb. Thus, a sensible way of modeling lexical reclassification is to consider a case in which a large number of verbs exhibit distinctly clustered and relatively static behavior over a period of time, thus inducing a stable category structure. Meanwhile, one verb changes against the backdrop of this category structure. Lexical reclassification is lexical representation change (i.e. INPUT \rightarrow HIDDEN weight change) that takes an item out of one cluster and brings it into another. The history of *be going to* is consistent with this *Stable Backdrop Assumption* for the probability distributions of most Motion, Equi, and Raising verbs have not changed much over the period in question.

Insert Figure 11 about here

Here, we are interested in what happens when the changing item is in transit between clusters. To characterize the scope of the possibilities it is useful to consider the image of the hidden space in the output space. Under the Stable Backdrop Assumption, all trajectories of change must lie in this image. If the output space is of sufficiently low dimension, we can examine this image by sampling the hidden unit space uniformly at small intervals and plotting the corresponding output points. The solid line in Figure 11 shows such a plot for a network trained on the simple *be going to* data. As hoped, the network interpolates a curve that approximates the historical data much better than the linear model (dotted line). To compare the two models quantitatively, it is useful to define the *Model Error* as the sum of the squares of the distances between each historical data point and its closest neighbor on the interpolated curve:

(21)

$$\text{Model Error} = \sum_i \| \vec{m}_i - \vec{d}_i \|^2$$

where m_i is the closest model point to data point d_i . The Model Error for the network model of the simple *be going to* data is approximately 0.063. The Model Error for the linear model is approximately 0.959.

Section 2 above motivated the foregoing model of the quantitative properties of the *be going to* episode by noting evidence for the Q-divergence phenomenon: prior to the first appearance of categorical evidence that an element has changed category, there are quantitative skewings that lead in the direction of the change. Now, it is clear why the Discrete Category Model fails to predict this effect and the Smooth Manifold Model predicts it. The discrete model claims that when an item changes category, it immediately becomes a canonical member of the new category, albeit one with a very low frequency of occurrence. Consequently, all behaviors associated with the new category are expected to start occurring immediately, and if we restrict our attention to just those instances when the item is behaving in its new capacity, the relative frequencies of the different behaviors that the new category licenses are expected to be precisely those of canonical members of the new category. Thus the discrete model implies a simultaneous emergence scenario. The Smooth Manifold model, by contrast, predicts that gradual category change should be constrained by the curvature of the interpolated manifold, whose shape is subject to a smoothness constraint. Consequently, when the endpoints of a transition are categories with some behaviors in common, there will be a quantitative skewing in the direction of those common behaviors during the transition. As noted above, the existence of nonlinear interpolations implies sequential emergence. The manifold analysis thus provides a reliable way of choosing between the sequential and simultaneous emergence claims.

A subtle but important difference between the two models is that the discrete model assigns ungrammatical behaviors zero probability while the Smooth Manifold Model assigns

them very small, positive probabilities. In fact, every possible behavior is assigned a positive probability under the Smooth Manifold Model, so the difference between grammatical and ungrammatical is best modelled in terms of a threshold: for some positive number θ , all behaviors assigned a probability less than θ are expected to be ungrammatical while all behaviors assigned a probability greater than θ are expected to be grammatical. For simple simulations like the *be going to* case at hand, the existence of an empirically accurate value of θ is guaranteed as long as the network can be trained to approximate the target function sufficiently closely.

Insert Figure 12 about here

Given this notion of grammaticality, we can see how the Smooth Manifold Model model predicts Q-divergence effects. In the simulation at hand, the least probable grammatical behavior has probability 0.064 while the most probable ungrammatical behavior has probability 0.017. So we can choose θ to be any value in the interval (0.064, 0.017), say $(0.064 + 0.017)/2$. The threshold partitions the output space into regions of categorically distinct behavior. We are particularly interested in the emergence of the novel nonagentive complement behavior so it is useful to draw the partition line that separates grammaticality of nonagentive complements from ungrammaticality of nonagentive complements (Figure 12). Because of the curvature of the representation manifold, a verb transiting from Motion behavior to Raising behavior will necessarily show an increase in the frequency of agentive complements before it crosses the threshold, i.e., before nonagentive complements start being grammatical. This is how the model predicts Q-divergence effects.

3.3.2 Equi vs. Raising Verbs

There is, in fact, a way of getting the discrete category model to do a better job of approximating the nonlinearity observed in the data. One may posit, following a suggestion by Pérez (1990), that *be going to* made a two-step transition, starting from its Motion verb state, it first became an Equi verb, and then became a Raising verb.

To clearly distinguish Equi from Raising verbs, it is useful to use the 4-fold division outlined in Section 2 (<P> (Place) vs. <A> (Agt.) vs. <S> (Sent. Subj., Non-Agt.) vs. <N> (Non-Sent. Subj., Non-Agt.)).

Insert Figure 13 about here

Insert Figure 14 about here

Insert Figure 15 about here

Insert Figure 16 about here

Four-dimensional probability distributions lie in a three-dimensional subset of 4-space that has the form of a tetrahedron. Figure 13, based on Table 5, shows the positions of the prototypes (average locations) of Motion, Equi, and Raising verbs on this tetrahedron. The diagram is unambiguous given the knowledge that the three types all lie on the surface of the tetrahedron. Because the trajectory of *be going to* is not limited to the surface of the tetrahedron in Figure 13 graphical depiction of it is difficult, so I switch to a different illustration scheme. Figure 14 shows a *parallel coordinate* image of the historical trajectory of *be going to*. The X-axis marks time points, the Y-axis marks behavior category, and the Z-axis measures relative probability. Figure 15 shows the (to be rejected) linear model (Model Error = 0.808). Figure 16 shows a parallel coordinate representation of the (to be rejected) two-step model (Model Error = 0.061).

The two-step trajectory gives a better fit than the linear model, but there are reasons to be dissatisfied with it. It predicts that *be going to* should have lost pure Motion behavior completely before it showed any clear signs of Raising status; but it is clear that *be going to* has exhibited Motion behavior throughout its history as it still does today. The discrete category model can, of course, generate mixtures of all three types by positing simultaneous membership in all three categories. In this case, the trajectory of *be going to* is predicted to lie in the triangular region whose vertices are at the Motion, Equi, and Raising loci. But under this weaker constraint, there is no reason to expect nonlinear skewing in the direction of Equi verbs. It is just as possible to have a direct transition between Motion and Raising. And yet there is reason to believe that the lean toward Equi is not coincidental: Equi is quantitatively more similar to Motion than Raising is. The discrete category model has trouble predicting the appropriate skewing because it assumes that mere quantitative similarity has no bearing on grammatical representation. Thus, one essential problem with the multiple category-membership approach to improving the empirical performance of discrete-category models (first identified in Section 3.2) is that the resulting account is undesirably stipulative. Moreover, this is the best case scenario. The multiple category-membership approach can only succeed in representing Q-divergence effects if there happen to be existing categories in the language which span a region containing the Q-divergence trajectory. Although this condition is roughly satisfied in the case of *be going to*, it is sometimes not satisfied. For example, when *sort of* and *kind of* evolved into degree modifiers (e.g. *We sort/kind of laughed*—see Section 1.1 above and Tabor, 1994a, 1994b) they went through a period of exhibiting an inordinately strong tendency to occur in constructions of the form <Det sort/kind of Adj N>. It is unlikely that any other elements in the language showed such a predilection for this construction. This means that assigning *sort of* and *kind of* multiple category membership would probably provide no improvement over a two-category linear interpolation model.

Insert Figure 17 about here

Returning to the *be going to* case, a Connectionist network for modeling these data needs four outputs and three different types of verb inputs. Moreover, since the variation in

Raising behavior extends in two dimensions, it is necessary to use a two-dimensional hidden unit space to get a reasonable fit to the training data. In this two-dimensional space there are a large variety of continuous trajectories that *be going to* might follow in transiting from Motion status to Raising status. But all of them involve Q-divergence in the neighborhood of Motion behavior (as well as in the neighborhood of Equi behavior) and all of them pass fairly near, though not exactly through, the Equi region. To make a specific prediction about *be going to*'s trajectory I make the assumption that the change trajectory is linear in the hidden unit space. Figure 17 gives a parallel coordinate depiction of this trajectory. The model fits the historical data much better than the linear model and almost as well as the two-step model (Model Error = 0.079).

The Smooth Manifold Model makes a good prediction by positioning the hidden manifold in such a way that Equi verbs are approximately intermediate between Raising verbs and Motion verbs. It adopts this solution because Equi verbs are behaviorally intermediate between the two types—this fact can be readily observed by comparing the distances between prototype loci in Figure 13. The discrete category model fails to provide an explanation for the Equi intermediacy because similarity considerations play no role in its interpolation mechanism.

3.3.3 Constituent Structure

Insert Figure 18 about here

The previous two sections have shown how the Connectionist implementation of the Smooth Manifold Model makes appropriate predictions when the problem is framed as a (verb) classification problem. This framework, however, permits only rudimentary encoding of syntactic structural relationships. Ideally, we should employ a Smooth Manifold Model which contains analogs of the complex constituent structures revealed by linguistic analyses (e.g., McCawley, 1988; Radford, 1988). It is not unreasonable to suppose that a Connectionist model trained on word-distribution data could develop representations that resemble those of linguistic parse-trees, for much of the evidence for the parse-trees is distributional in nature (see Finch, 1993). In this section, then, as a step in the desired direction, I examine a diachronic Connectionist model which encodes cross-phrasal dependencies to make sure that the analysis of Q-divergence given above extends to a case involving constituent-structure revision.

The model is based on Elman (1990) and (1991)'s paradigm for constituent-structure learning. Elman generated a corpus of sentences with an artificial grammar, presented localist representations of the words from the corpus in order on the input layer of a network, and trained the network on the task of predicting, at each juncture, a localist representation of the next word on the output layer. His model was similar to the networks described in the previous sections in that it had a three-layer (input, hidden, output) Connectionist network at its core. But, to allow the model to store information about temporal dependencies between inputs, Elman let the hidden layer at each time step receive activation from its own state at the previous time step as well as from the input layer. Thus the network was partially *recurrent* (Figure 18). As Rumelhart, Hinton, and Williams (1986) point out, a

recurrent network can be translated into an equivalent feedforward network by “unfolding” the time dimension in space. Thus the same backpropagation learning algorithm can be used and its convergence properties are identical provided the training examples (which may be infinite in number) determine a well-defined gradient. Elman trained on an approximation of the full gradient (backpropagating error only one time-step into the past) and found that the outputs converged on values approximating the in-context transition probabilities defined by the grammar, even in relative clause structures where long distance dependencies were involved. Thus he developed a reasonably sophisticated representation of the syntactic information contained in the grammar.

Insert Figure 19 about here

To incorporate this training arrangement into a diachronic model, I wrote a simple probabilistic phrase-structure grammar (Figure 19) to approximate the distribution of *be going to* prior to the 16th century (i.e., prior to the first changes of interest) and used the output of the grammar to train a network. For simplicity, I omitted the word *be* from the *be going to* collocation—this is essentially an artifice that makes training easier, but it is defensible on the grounds that the *be* of the *be going to* collocation does not participate in the word fusion, *go ing to* → *gonna*, and continues to behave exactly like other types of *be* with regard to inflection and adverb placement, so it is reasonable to view its behavior as independent of the constituent-boundary changes. This new network had 47 input units, 8 hidden units, and 47 output units. A larger number of hidden units were needed in this simulation than in the previous ones because the task is much more complicated. I trained the network until the average per-pattern error was stable.

In this case, the output and hidden spaces are too high-dimensional to visualize, but we are interested in just a small subset of the behaviors, so we can again focus on a normed subspace. Note that the net assigns a likelihood of occurrence to every sentence¹⁶ and thus to every type of sentence.¹⁷ Thus we can examine the relative likelihood assigned to each of the four sentence types of the previous section. The previous models generated a trajectory for *be going to* by identifying the hidden unit starting point and the hidden unit ending point and assuming linear interpolation in the hidden unit space. In this model, since *(be) go ing to* is not being treated as a single lexical item, there are several hidden unit states corresponding to any given state of *(be) going to*. Moreover, for the ending point of the trajectory, it is not possible to identify these hidden states with the states of existing elements (as in the previous simulations) because no member of the stable backdrop has

¹⁶To determine what likelihood it assigns to a given sentence, one sets the activations of the hidden layer to the state it is in at the end of every sentence and then feeds the words of the sentence to it in order, noting at each juncture, the output probability assigned to the next word. These probabilities are independent, so one can multiply them to obtain a likelihood for the sentence.

¹⁷By a type of sentence I mean a set of sentences. For example, one might be interested in the type <N V> where *N* can be any noun from the vocabulary and *V* can be any verb. The probability of the type is the probability of observing some member of the type—i.e., the sum of the probabilities of the members.

exactly the desired properties.¹⁸ However, there is a plausible way to generate a trajectory which closely parallels the interpolation models and permits us to investigate the curvature of the representation manifold. We can characterize *(be) going to*'s ending state in terms of the corpus-generating grammar: the morpheme sequence *(be) go ing to* has the distribution of a Raising verb. Thus, a sensible strategy is to take a network that has been trained to asymptote on the output of the pre-16th century grammar (Figure 19) and then switch the training corpus to one generated by a grammar that is identical in all regards except that *(be) going to* is distributed as a Raising verb. I'll refer to this second stage of training as *post-training*. During post-training, the network conveniently implements the Stable Backdrop condition because the preponderance of unchanging behaviors keep the weights near the minimum arrived at during initial training. Only those weights associated with the lexical item *go* on the input and output layers are able to change significantly. Thus, whatever trajectory *go* follows in the process of changing from being a motion verb to being a part of a raising collocation, it is constrained to move along a manifold defined by the general syntax.¹⁹

Insert Figure 20 about here

The results are shown in Figure 20 which should be compared to Figure 14. The recurrent network fits the data only slightly better (Model Error = 0.076) than the corresponding feedforward network and almost as well as the two-step model. However, it is performing a much more complex task—modeling the phrase structure of the changing grammar—in addition to getting the diachronic facts right. I take this result as indicative of the viability of the Smooth Manifold paradigm as a general syntactic model. In Section 4.2.2.1, I describe a way in which the grammaticalization phenomenon investigated here may help shed light on how to do constituent structure representation in Smooth Manifold models.

4. CONCLUSION

4.1 Summary

At the outset, I identified the phenomenon of Q-divergence: categorical grammar changes are often preceded by facilitating quantitative skewings. I suggested that the phenomenon urged a revision of standard discrete category representations in favor of continuous categories structured around prototypes. A case study of the changes in English *be going to* provided support for this thesis. Initially a pure motion verb, *be going to* acquired something like Equi status during the 18th and 19th centuries, before switching over to predominantly Raising behavior during the 20th century. Prior to the first appearance of each novel behavior, the relative frequencies of the existing behaviors shifted so that *be going to* became more

¹⁸Each of the other Raising Verbs has slightly different syntax from *be going to*: *seem* and *tend* can appear in any aspect or tense. Modal auxiliaries like *will*, *must*, *may*, are always the first verb in their clause and never take tense or aspect markers. By contrast, *go* can only be a Raising verb if it is in the present tense and has progressive (*-ing*) aspect. Moreover, Raising *be going to* needn't occur as the first verb in its clause.

¹⁹The manifold of interest is the space of lexical behaviors defined by the (effectively) fixed weights associated with lexical items other than *go*.

like the novel type. I showed that a Smooth Manifold Model implemented in a Connectionist network predicted these Q-divergence effects. The essential mechanism was that of fitting the quantitative distributional data with a low-dimensional, curved manifold. Traversing the manifold in time, *be going to* had to change its distribution substantially along one dimension before it could change significantly away from zero along a different dimension, and thus exhibit a novel behavior.

Probabilistic discrete-category models fail to predict Q-divergence because they perform only linear interpolation in the behavior space—that is to say, they lack curved manifolds. It is true that multiple linear manifolds can approximate curved manifolds—the two-step model of *be going to*'s development as a Motion-Equi mixture followed by an Equi-Raising mixture provided a crude example, and one might speculate that by introducing a richer feature base, one could discern a range of intermediate stages and thus approximate quite closely the predictions of the nonlinear model. Does this mean that there is really no fundamental difference between the Smooth Manifold and Discrete Category models? No. The discrete model must stipulate the ordering of the intermediate states—for it defines no notion of intermediacy. It also depends on the chance existence of appropriately situated intermediate categories (like the Equi-verbs), but these don't always exist where the model needs them (Tabor, 1994a, 1994b). The Smooth Manifold Model, on the other hand, derives the intermediate stages and their ordering from the similarity structure of the data.

Thus the main new feature of the grammatical theory outlined here is the appeal to subcategorical (i.e. purely quantitative) similarity in the definition of grammatical structure. Quantitative contrasts are cashed out as distances in the representation space and distance from a prototype reflects degree of divergence from prototypical behavior. The orientation of the dimensionally-restricted prototype-centered clusters determines the nature of the interpolation, that is to say, the generalization properties of the theory. The interpolated manifold is smooth and optimal in the sense that similar structures are assigned nearby representations. Thus there is an important role for a similarity metric in grammatical theory. Subcategorical (i.e. merely quantitative) similarity is crucial so it would not be effective simply to define a metric on standard categorical representations. The smoothness of the representation manifold coupled with the hypothesis that lexical change proceeds by incremental traversal of the manifold gives rise to the prediction that such change is never abrupt.

4.2 Sources of Further Evidence

To my knowledge, the phenomenon of Q-divergence has not yet been systematically investigated by anyone other than myself. Consequently, more empirical work is needed to confirm the results.

4.2.1 Pertinent Studies

There are many potential case studies which could provide confirming or disconfirming evidence. It is worth emphasizing that both synchronic and diachronic predictions stem from the Smooth Manifold Model. The synchronic predictions are the counterpart to the set of predictions about unobserved usages made by the categories and rules of standard linguistic theories. In both cases, it is the interpolated manifold that constitutes the set of

predicted unobserved behaviors.²⁰ In this section, I describe existing case studies in both the diachronic and synchronic literatures which point to good ways of testing the Smooth Manifold hypothesis.²¹

4.2.1.1 Diachronic Studies

Lichtenberk (1991) studies a comitative preposition,²² *bia*, in the Oceanic language, To'aba'ita, that is developing a use as a coordinating conjunction meaning 'and', and is consequently encroaching on the domain of the previously-existing coordinating conjunction, *ma*, which also means 'and'. He finds that *bia* has encroached the most on *ma* in "an environment that was less different from the typical environment of its prepositional function than the other potential environments were, that is with human NP's to its right (p. 64)". He gives quantitative data showing *bia* being used in 48% of human/human NP coordination cases, in 3% of inanimate/inanimate NP coordination cases, and in 0% of VP coordination cases. This looks like a case of Q-divergence: the rate of human/human coordination with *bia* is substantially elevated compared to the rate of inanimate/inanimate *bia*-coordination. Partial confirmation could be provided by converting Lichtenberk's semantic-domain based relative frequencies to the form-based relative frequencies considered here and see if they lie on the manifold induced by other prepositions and conjunctions in the language. It would also be revealing to collect quantitative data on To'aba'ita's close relative, Sa'a, in which the cognate of *bia* does participate in VP-coordination (p. 64) and thus seems to be at a more advanced stage in the grammaticalization process.

The Smooth Manifold Model presents a challenge to the theory of diachronic syntax advocated by Kroch (1989a, 1989b), Pintzuk (1991), Santorini (1992), Fontana (1993), and Taylor (1994). These authors propose the *Competing Grammars* model of syntactic change. This model is concerned with unified grammatical changes that affect many syntactic environments. It assumes that within each environment, two competing options can be identified, one corresponding to each of two competing grammars. The relative frequencies of one option versus the other are computed in each environment across time. Typically, the resulting curves are S-shaped so they are modeled with logistic functions— $f(t) = 1/(1 + e^{-st+b})$, where $f(t)$ is relative frequency, t is time, s is called the "slope" and b is called the "intercept". There needs to be one logistic function for each environment, since some environments show

²⁰It may seem surprising that no qualitative distinction is made, under the current theory, between interpolation that predicts possible changes the language can undergo and interpolation that predicts the general unobserved "grammatical" behaviors, which are the bread and butter of synchronic linguistic theory. In fact, under the current theory, *every* act of using the language effects a change in the language's state. It just so happens that most changes are distributed around a mean which is close to the current state of the language. Thus they cancel each other out and the language appears to be largely static. The theory distinguishes synchronic and diachronic "change" by treating the former as the set of minute extensions in random directions from the current state and the latter as the set of protracted divergences from the current state along particular paths. Crucially, the two types of change blend into one another.

²¹It is, of course, highly desirable to obtain evidence from spoken usage. This is obviously a possibility for synchronic studies. It may even be a possibility for diachronic studies. The Smooth Manifold Model suggests that Q-divergence effects should be detectable in a large sample of usage over even a very short period of time (perhaps a decade or less). With the increasing textual transcription of spoken corpora, such a study may be feasible before long.

²²A *comitative preposition* is a preposition expressing accompaniment, typically of humans by humans, and is usually translated by English *with*.

greater advancement than others (and hence have different intercepts), but the data largely support the hypothesis that all grammatically related environments have the same slope parameter (Kroch 1989b, Pintzuk, 1991, Taylor 1994). It is claimed that the different degrees of advancement in the different environments reflect the different degrees to which different syntactic environments are favored by constant, grammar-independent contextual factors. Formal similarity between environments does not play a role. By contrast, the Smooth Manifold account maintains that environments rise in an order that reflects their formal similarity. For example, it holds that *be going to* moved earlier into Equi environments than Raising environments because Equi verbs resemble the initial motion verb more closely. While this similarity hypothesis seems plausible in the case of *be going to*, it does not seem particularly plausible in many of the cases studied by Kroch and colleagues: English periphrastic *do* (Kroch), English word-order (Pintzuk), Yiddish word order (Santorini), Greek clitics (Taylor). Only the case of Spanish clitic lexicalization (Fontana) seems to involve a succession of stages ordered by similarity (see Tabor 1994, Chapter 2). The critical difference may be that Fontana's case and the *be going to* case involve grammatical reclassification. It will be useful, as a way of testing the Smooth Manifold account, to analyze the Competing Grammars statistical data in the manner studied here: the frequencies of a form in different environments should be compared to each other, rather than with competitor forms. Then the Q-divergence hypothesis can be directly evaluated on the same data, and the role of reclassification can be evaluated.

The Competing Grammars model implies that the actuations of the different behaviors involved in a complex grammatical change are simultaneous. It puts the emphasis on actuation and claims that the order of emergence of the different behaviors is determined by extralinguistic factors that have nothing to do with grammar. There are three reasons to be dissatisfied with this perspective. First, it is hard to distinguish simultaneous from sequential actuation empirically because, at the point in time when actuation is claimed to occur, all the frequencies of the relevant forms are very low. Second, the employment of discrete cut-offs for the S-shaped curves makes the mathematical model difficult to work with. Third, by claiming that grammar plays no role in the order of emergence, the Competing Grammars account forgoes the possibility of making any predictions about when new constructions will become grammatical—the so-called *actuation problem*—see Section 4.3. Of course, it is not a forgone conclusion that grammar change should be constrained by the current grammar state, but given the evidence presented in Section 3 that emergence is strongly constrained by the quantitative properties of grammatical distributions, it seems important not to rule out the possibility.

4.2.1.2 Synchronic Studies.

Quantitative studies in synchronic linguistics can also serve as tests of the Q-divergence hypothesis.

Zec (1986) examines a person/number/tense marker, *će*, in Serbo-Croatian that seems to be lexicalizing in much the same way as the Polish person/number/tense marker mentioned in Section 1.1. The marker is generally restricted to second position, and when the first position is occupied by a word that is not a verb, it behaves like an ordinary clitic. However, when the verb is in first position, the marker triggers a lexical rule of palatalization and two strictly lexical tone rules (see Inkelas and Zec, 1988). Inkelas (1989) thus argues that *će*

is an affix in this particular context. Given its semantic and phonological similarity to its clitic counterpart, and the parallel between this situation and the Polish one, it seems highly likely that *će* has only recently acquired affixal status. Therefore, it is reasonable to expect that the signs of Q-divergence still linger: it is likely that the element has developed a strong quantitative tendency to occur on the verb. Since Serbo-Croatian has a wide variety of clitic elements, it should be possible to make a synchronic quantitative comparison between *će* and other, related elements which have not undergone lexicalization, and thus test the Q-divergence hypothesis.

Using the syntactically-parsed Penn Treebank corpus, Juliano and Tabor (1995) compute probability distributions over eight major complement-types for 156 English verbs.²³ One interesting result of their study is that although the verbs tend to occur in clusters in the behavior space, the clusters are quite diffuse, and many verbs lie in the regions between clusters. Although Q-divergence was not the focus of Juliano and Tabor's study, such a diffuse distribution permits a test of the Q-divergence hypothesis in a purely synchronic setting: one can view the verbs that are far away from the prototypes as if they are items in historical transition from one cluster to another. They are thus expected to lie on a smooth manifold interpolated between the clusters. This manifold should be inducible from a relatively small sample of the central members of each cluster alone (just as the Q-divergence trajectory was inducible from the stable backdrop in the above simulations). We can test this hypothesis by using Connectionist learning to induce a manifold from a sample of central members and measuring how well this manifold fits the intermediate data-points that were absent from the training set. If the model does better than linear interpolation, we have evidence for Connectionist nonlinear optimization. If it does about the same or worse, then we have evidence for the standard model.²⁴

In sum, a variety of existing synchronic and diachronic studies point to places in languages where we should expect Q-divergence effects to be detectable if the Smooth Manifold Model is accurate.

4.2.2 Theoretical Questions

There are a number of theoretical questions about Smooth Manifold Models which must be addressed if the models are to become full-fledged grammatical theories. Prominent among these are: (i) How can such models embody constituent structure generalizations? (ii) What theory underlies the notion of grammaticality-contrast in such models? The current study sheds new light on both of these issues.

4.2.2.1 Constituent Structure

The third simulation discussed above pertains directly to the constituent structure question. In that simulation, *go ing to* is treated as a sequence of three distinct input representations. Under the initial grammar, these three morphemes belong to three distinct classes

²³The complement types considered are Adjective Phrase, Adverb Phrase, Noun Phrase, Prepositional Phrase, Infinitive Sentence, Finite Sentence, Progressive (-*ing*) Verb Phrase, and Zero. Almost every clause in the corpus involves one of these complement types.

²⁴If the two models perform similarly, then the interpolated space is linear. Since the contrasts involved in this case are contrasts of category-type rather than word-order, the arguments given in the previous section do not particularly lead us to expect a linear outcome here.

whose members are largely interchangeable in the corresponding positions: *go* is one of several motion verbs; *ing* is one of several tense/aspect markers; *to* is either the infinitive marker or one of several prepositions. By contrast, in the distribution defined by the post-training grammar, these three morphemes, when used to form a future auxiliary, have the status of a frozen phrase. As a consequence, the three morphemes do not share distributional characteristics with any other morphemes. A recurrent network predicting morpheme-to-morpheme transitions performs best if it uses its hidden unit resources to code transitions that are common to many morphemes. Thus, if a network is designed in which there is an option to treat sequences of phonemes either as sequences of morphemes (and hence use valuable hidden unit resources to code their relationships) or as mono-morphemic units (and hence avoid using valuable hidden unit resources), it should tend to treat *go ing to* as a sequence of morphemes under the initial grammar and as a mono-morphemic unit in precisely those instances in the post-training grammar where it is behaving as an auxiliary verb. This line of reasoning suggests that the global cost-assessment mechanism of the Connectionist network might be used to predict the oft-noted phonological idiosyncrasy of auxiliary *go ing to* in the modern language: only auxiliary *go ing to* tolerates phonological fusion of the three morphemes (22).

(22) It's gonna snow.

* We're gonna Texas.

Thus we might view the cost-assessment mechanism as embodying a theory of the distinction between branching and monolithic constituents. This simulation thus suggests an interesting new way of addressing the question, "How is constituent structure represented in smooth manifold models?"

4.2.2.2 Grammaticality Contrast

Defining grammaticality contrast in smooth manifold models is of particular interest because it relates to a central question of information theory: how to distinguish signal from noise. In information theory, one is often concerned with the problem of interpreting a signal that has been transmitted via a noisy channel. If the noise itself contains no information, then the problem is easy—one can assume that all the information in the data is due to the signal. But if the noise has some structure, as is typical in real data, one needs a criterion for deciding which structure is worth paying attention to and which can be ignored. Otherwise, one will tend to "overfit" the data and choose a model which has poor generalization ability. The usual information theoretic approach is to invoke "Occam's Razor": prefer simpler models over more complex ones. The question of how to balance the desirableness of simplicity with the desirableness of fitting the data well is often made intuitively on a case-by-case basis and sometimes theoretically on the basis of a principle like Bayes' Rule (e.g., MacKay, 1992). To date, there is no firm agreement about what the right solution is in any domain, including language. From a psychological standpoint, it is of interest to ask if human learners embody a particular principle about how to make the simplicity vs. accuracy trade-off and, if they do, to ask what that principle is. In this regard, the grammaticalization data are relevant. We can find many cases, like the *be going to* example discussed here, where a certain systematicity in the input data is treated as noise by learners at a given point in time, but then grows in magnitude or becomes more pronounced in such a way that learners at a

later point treat it as part of the signal. We know when the learners have switched models because they start generalizing on the basis of a pattern which they previously ignored. “Q-divergence” refers to the period when the pattern is changing systematically but does not yet inspire novel generalization. The notion of the *grammaticality threshold*, introduced in Section 3.3.1, is a way of defining this important cross-over point in smooth manifold models: when the behavior of a form crosses a grammaticality threshold, we say that generalization has occurred. Thus by studying the properties of the grammaticality threshold, we may gain insight into how the human language mechanism handles the simplicity vs. accuracy trade-off.

4.3 Final Comment: The Actuation Problem

The Smooth Manifold model has the potential to shed light on what is generally viewed as the most challenging problem in language change: what Weinreich, Labov, and Herzog (1968) call the *actuation problem*. Weinreich *et al.* note that while we can make some headway in cataloging the possible changes that a given language can undergo using current tools of grammar, we are almost completely powerless to say when a change is going to *be actuated*, i.e., start happening. They suggest that the only way we can make headway on this problem is to take into account the myriad social and functional forces that exert an influence on language. Indeed it is hard to imagine how we could achieve understanding without such knowledge. But even if we had a detailed knowledge of social and functional forces, that knowledge would do us no good if we had no understanding of how those forces can impinge on the grammatical structures whose revision concerns us. It is completely unelucidating to claim that there is a social or functional force which turns main verbs into auxiliary verbs, adpositions into clitics, nouns into adverbs, etc. The present theory offers us a bridge between these two levels of understanding. It shows how small changes in the frequencies of items, which are known to be subject to social and functional influences (e.g., Labov 1973), can lead to the grammatical changes which we tend to think of as events of “language change”. Thus the theory provides a line of approach to the actuation problem.²⁵

Q-divergence thus suggests a way of formalizing the notion of structural innovation and explaining how it can come about via forces that are known to act upon language. The problem of innovation is central not only in language change but in many other areas of cognitive change and in biological evolution. The manifold analysis of neural network representations suggested by the study of Q-divergence points to a useful new geometric approach to interpreting neural networks in structural terms and also shows promise of shedding light on the simplicity versus accuracy question. Given the centrality of these issues in cognitive science, further investigation of the Q-divergence hypothesis seems especially worthwhile.

²⁵Kroch (1989a, 1989b) interprets the term “actuation” as referring to a point in time when a new grammar is hypothesized to come instantaneously into being. Weinreich *et al.* are not so explicit about the character of the mechanism so I take their question to be a general one about when grammar changes will happen. By advocating the Smooth Manifold model, I am proposing that the appropriate goal is to predict when the substantial changes in distribution will occur, i.e., to predict events of quantitative *emergence*, rather than events of instantaneous appearance.

Appendix 1. Probabilistic context-free grammars interpolate linearly.

Proof: Under a probabilistic context-free grammar, if we make the linguistically reasonable assumption that syntactic environments can be described in terms of tree-structure configurations of lexical classes, without reference to specific words, then the probability of finding word x in a given syntactic environment is, for each lexical class, C , a linear function of $P(x|C)$, the probability of choosing lexical item x as the instantiation of class C . Thus, if x is simultaneously a member of class O and class N , its position \vec{p}_x in the relative probability space in which dimensions correspond to syntactic environments can be written,

(23)

$$\vec{p}_x = \frac{P(x|O)\vec{q}_O + P(x|N)\vec{q}_N}{(P(x|O)\vec{q}_O + P(x|N)\vec{q}_N) \cdot \vec{1}}$$

where \vec{q}_O and \vec{q}_N are the probability distributions over syntactic environments of the classes O and N , respectively, $\vec{1}$ denotes the vector with all elements equal to 1, and $(P(x|O), P(x|N))$ can be any value in $[0, 1] \times [0, 1]$. We want to show that \vec{p}_x lies on the line segment between the locations of canonical original-class (O) words and canonical new-class (N) words.

Note that

(24)

$$\begin{aligned} \vec{p}_x &= \frac{P(x|O)\vec{q}_O \cdot \vec{1}}{(p_O\vec{q}_O + P(x|N)\vec{q}_N) \cdot \vec{1}} \frac{\vec{q}_O}{\vec{q}_O \cdot \vec{1}} + \frac{P(x|N)\vec{q}_N \cdot \vec{1}}{(p_O\vec{q}_O + P(x|N)\vec{q}_N) \cdot \vec{1}} \frac{\vec{q}_N}{\vec{q}_N \cdot \vec{1}} \\ &= c \frac{\vec{q}_O}{\vec{q}_O \cdot \vec{1}} + (1 - c) \frac{\vec{q}_N}{\vec{q}_N \cdot \vec{1}} \end{aligned}$$

where c is in $[0, 1]$. In the initial condition in which x is only a member of class O , suppose $P(x|O)_0$ is the the likelihood that class O is instantiated by x . Then the starting relative probability vector, \vec{p}_{start} is given by

(25)

$$\vec{p}_{start} = \frac{P(x|O)_0\vec{q}_O}{P(x|O)_0\vec{q}_O \cdot \vec{1}} = \frac{\vec{q}_O}{\vec{q}_O \cdot \vec{1}}$$

which does not depend on the within-class relative probability, $P(x|O)_0$ so it is the same for all items of class O . The corresponding result holds for \vec{p}_{end} . Thus we can write the general position as

(26)

$$\vec{p}_x = c\vec{p}_{start} + (1 - c)\vec{p}_{end} = \vec{p}_{start} + (1 - c)(\vec{p}_{end} - \vec{p}_{start}) \quad c \in [0, 1]$$

In other words, \vec{p}_x lies on the line segment connecting \vec{p}_{start} and \vec{p}_{end} .

Appendix 2. Nonlinear interpolation implies relative-velocity exchanges.

Proof: Let ρ be a continuous trajectory in a metric space of dimension ≥ 2 . Suppose ρ starts at point A and ends at point B. Consider the projection of ρ onto any two dimensional subspace defined by two of the behaviors being measured. Call the coordinates of the trajectory in these dimensions x and y , respectively. Without loss of generality, assume that point A projects to the origin in xy space. Let x_N be the net change in the x -coordinate over the path ρ . Let y_N be the net change in the y -coordinate over the path ρ . We can integrate along each dimension to get the net change on each dimension:

(27)

$$\int_{\rho} \frac{dx}{dt} dt = x_N \quad \int_{\rho} \frac{dy}{dt} dt = y_N$$

Therefore,

(28)

$$\int_{\rho} \frac{1}{x_N} \frac{dx}{dt} - \frac{1}{y_N} \frac{dy}{dt} dt = 0$$

If ρ does not lie on the straight line between A and B, then there will be some pair of behaviors such that x/x_N and y/y_N are not equal for some t . Consequently, the corresponding derivatives will be unequal. Therefore, by (28), there will be points along ρ where $(1/x_N) \frac{dx}{dt} > (1/y_N) \frac{dy}{dt}$ and there will be other points along ρ where $(1/y_N) \frac{dy}{dt} > (1/x_N) \frac{dx}{dt}$. In other words, the relative rates of change will exchange sizes.

Appendix 3. Network descriptions.

Input units: Bit vectors with only 1 bit on, each vector represents a word.

Target units: Probability distributions (determined by historical data and simple assumptions about the forms of grammars as described in the text).

Hidden units: The (fixed sigmoid) activation of hidden unit h_i is given by

(29)

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

where $net_i = \sum_j a_j w_{ij} + \beta_i$, a_j is the activation of input unit j , w_{ij} is the weight from unit j to unit i , and β_i is the bias on unit i (the bias is equivalent to a weight from a unit that is always on).

Output units: The (softmax) activation of output unit o_i is given by

(30)

$$o_i = \frac{e^{net_i}}{\sum_j e^{net_j}}$$

where j indexes output units.

Learning: Delta rule learning, backpropagated through the layers and through time (see Rumelhart et al., 1986; Pearlmutter, 1995).

(31)

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

p indexes patterns (words), a_{pj} is the activation of unit j when pattern p is presented on the input layer,

(32)

$$\delta_{pi} = t_{pi} - o_{pi}$$

for output units, where t_{pi} is the target value and o_{pi} is the output value of unit i when pattern p is presented on the input layer, and

(33)

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

for hidden units, where δ_{pk} is the δ associated with the k th unit on the next higher layer.

References

Andersen, H. (1987). From auxiliary to desinence. In Harris, M. and Ramat, P., editors, *Historical Development of Auxiliaries*. Mouton de Gruyter, Berlin.

Bailey, C.-J. (1973). *Variation and Linguistic Theory*. Center for Applied Linguistics, Washington.

Ballard, D. H. (To appear in 1996). *An Introduction to Natural Computation*. MIT Press, Cambridge, Massachusetts.

Bybee, J., Pagliuca, W., and Perkins, R. (1991). Back to the future. In Traugott, E. and Heine, B., editors, *Approaches to Grammaticalization, v. 2*, pages 17–58. John Benjamins.

Cedergren, H. and Sankoff, D. (1974). Variable rules: Performance as a statistical reflection of competence. *Language*, 50(2):333–355.

- Charniak, E. (1993). *Statistical Language Learning*. MIT Press, Cambridge, Massachusetts.
- Chomsky, N. (1957). *Syntactic Structures*. Mouton and Co., The Hague.
- Clark, R. and Roberts, I. (1991). A computational model of language learnability and language change. *Linguistic Inquiry*, 24(2):299–345.
- Danchev, A. and Kytö, M. (1991). The construction *be going to + infinitive* in Early Modern English. In Kastovsky, D., editor, *Papers from the Early Modern English Conference (EMEC), Tulln, 1991*. Mouton de Gruyter.
- Dell, G. S. and O’Seaghdha, P. G. (1992). Stages of lexical access in language production. *Cognition*, 42:287–314.
- 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.
- Fontana, J. M. (1993). Phrase structure and the syntax of clitics in the history of Spanish. Ph.D. Dissertation, Linguistics, University of Pennsylvania.
- Givón, T. (1984). *Syntax I*. John Benjamins, Amsterdam.
- Goldsmith, J. and Larson, G. (1992). Using networks in a harmonic phonology. In Canakis, C., Chan, G., and Denton, J., editors, *Proceedings of the 29th Regional Meeting of the Chicago Linguistic Society*. University of Chicago Press.
- Halpern, A. (1993). ?? Ph.D. Dissertation, Stanford University.
- Haykin, S. S. (1994). *Neural networks: a comprehensive foundation*. MacMillan, New York.
- Hertz, J., Krogh, A., and Palmer, R. G. (1991). *Introduction to the Theory of Neural Computation*. Addison-Wesley, Redwood City, California.
- Hopper, P. (1991). On some principles of grammaticization. In Traugott, E. C. and Heine, B., editors, *Approaches to Grammaticalization, v. 1*, pages 17–36. John Benjamins.
- Hopper, P. J. and Traugott, E. C. (1993). *Grammaticalization*. Cambridge University Press, Cambridge, England.
- Inkelas, S. (1989). Prosodic constituency in the lexicon. Ph.D. Dissertation, Stanford University.

Inkelas, S. and Zec, D. (1988). Serbo-croatian pitch accent: The interaction of tone, stress and intonation. *Language*, 64:227–248.

Juliano, C. and Tabor, W. (1995). Frequency contrast and grammatical representation: the case of frequency by regularity interaction. Poster presented at the CUNY Language Processing Conference in Tucson, Arizona.

Kiparsky, P. (1982). Analogical change as a problem for linguistic theory. In *Explanation in Phonology*, chapter 11. Foris.

Klein, E. and Sag, I. (1985). Type-driven translation. *Linguistics and Philosophy*, 8:163–202.

Kroch, A. S. (1989a). Function and grammar in the history of English: Periphrastic *do*. In Fasold, R. W. and Schiffrin, D., editors, *Language Change and Variation*, pages 134–169. John Benjamins, Philadelphia. 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.

Labov, W. (1969). Contraction, deletion, and inherent variability of the English copula. *Language*, 45:715–62.

Labov, W. (1973). *Sociolinguistic Patterns*. University of Pennsylvania Press, Philadelphia.

Lichtenberk, F. (1991). On the gradualness of grammaticalization. In Traugott, E. C. and Heine, B., editors, *Approaches to Grammaticalization, v. 1*, pages 37–80. John Benjamins.

Lightfoot, D. (1979). *Principles of Diachronic Syntax*. Cambridge University Press, London.

MacKay, D. J. C. (1992). Bayesian methods for adaptive models. Ph.D. Dissertation, California Institute of Technology.

McCawley, J. D. (1988). *The Syntactic Phenomena of English, v. 1–2*. The University of Chicago Press, Chicago.

Naro, A. J. (1981). The social and structural dimensions of a syntactic change. *Language*, 57(1):63–98.

Naro, A. J. and Lemle, M. (1976). Syntactic diffusion. In *Papers from the Parasession on Diachronic Syntax*, pages 221–39. Chicago Linguistics Society. Reprinted in *Ciência e Cultura* 29.259–68.

Nunberg, G., Sag, I. A., and Wasow, T. (1994). Idioms. *Language*, 70(3):491–538.

Pearlmutter, B. A. (1995). Gradient calculations for dynamic recurrent networks: A survey. *IEEE Transactions on Neural Networks*, 6(5):1212–1228.

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.

Plate, T. A. (1994). Distributed representations and nested compositional structure. Ph.D. Thesis, Computer Science, University of Toronto.

Radford, A. (1988). *Transformational Grammar, A First Course*. Cambridge University Press, Cambridge, England.

Richards, W. (1988). *Natural Computation*. MIT Press, Cambridge, Massachusetts.

Rosenbaum, P. S. (1967). *The Grammar of English Predicate Complement Constructions*. MIT Press, Cambridge, Massachusetts.

Rumelhart, D., Durbin, R., Golden, R., and Chauvin, Y. (1995). Backpropagation: The basic theory. In *Backpropagation: Theory, Architectures, and Applications*. Lawrence Erlbaum Associates.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986a). Learning internal representations by error propagation. In Rumelhart, D. E., McClelland, J. L., and the PDP Research Group, editors, *Parallel Distributed Processing, Volume I*, pages 318–362. MIT Press.

Rumelhart, D. E., McClelland, J. L., and the PDP Research Group (1986b). *Parallel Distributed Processing. Explorations in the Microstructure of Cognition. Volume 1*. MIT Press, Cambridge, Massachusetts.

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.

Seidenberg, M. S. and McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review*, 96:447–452.

Servan-Schreiber, D., Cleeremans, A., and McClelland, J. L. (1991). Graded state machines: The representation of temporal contingencies in simple recurrent networks. *Machine Learning*, 7:161–193.

Smolensky, P. (1990). Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence*, 46. Special issue on Connectionist symbol processing edited by G. E. Hinton.

Smolensky, P., Legendre, G., and Miyata, Y. (1992). Principles for an integrated connectionist/symbolic theory of higher cognition. Ms., Department of Computer Science, University of Colorado at Boulder.

Tabor, W. (1994a). The gradual development of degree modifier *sort of*: A corpus proximity model. In Beals, K., Cooke, G., Kathman, D., McCullough, K.-E., Kita, S., and Testen, D., editors, *Proceedings of the 29th Regional Meeting of the Chicago Linguistic Society*. University of Chicago.

Tabor, W. (1994b). Syntactic innovation: A connectionist model. Ph.D. dissertation, Stanford University.

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.

Taylor, A. (1994). The change from SOV to SVO in Ancient Greek. *Language Variation and Change*, 6:1–37.

Weinreich, U., Labov, W., and Herzog, M. (1968). Empirical foundations for a theory of language change. In Lehmann, W. P. and Malkiel, Y., editors, *Directions for Historical Linguistics*, pages 95–188. University of Texas Press.

Zec, D. (1986). Neki problemi vezani za razlikovanje klitika i afiksa [Some problems related to distinguishing clitics and affixes]. Paper presented at *Slavisticki susreti u Vokove dane*, Belgrade, Yugoslavia.

Zwicky, A. M. and Pullum, G. K. (1983). Cliticization v. inflection: English *n't*. *Language*, 59(3):502–513.

Figure 1: Graph of Rama relational marker relative frequencies [P = Postposition, C = Clitic Preverb, L = Lexical Preverb; 1 = ba(ng)- (33 tokens), 2 = u/yu- (112 tokens), 3 = a(ak)ya- (43 tokens), 4 = su/su- (38 tokens), 5 = ki/ki-] (92 tokens).

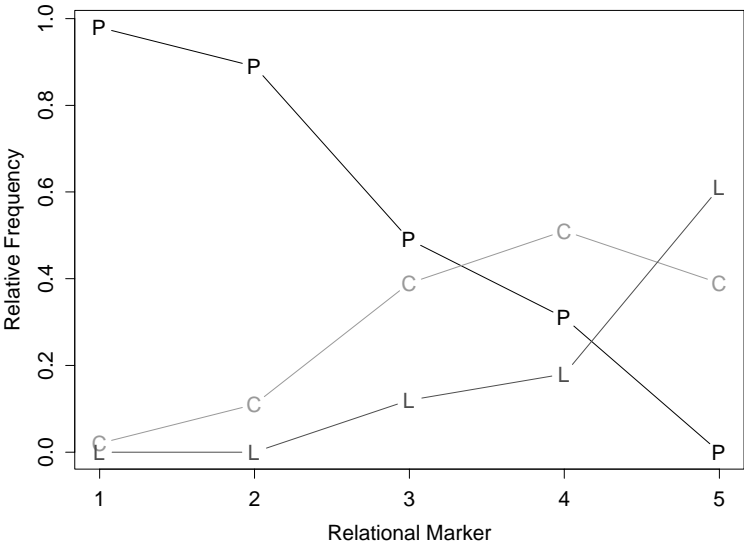


Figure 2: Lexical categories modeled as clusters of points in a continuous space.

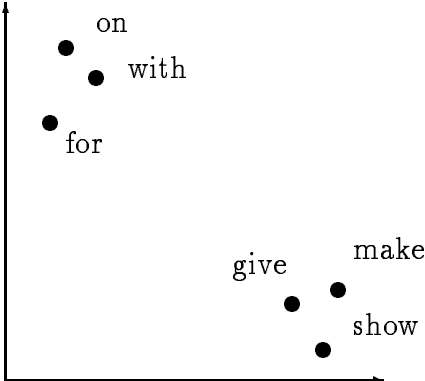


Figure 3: A Curved Manifold.

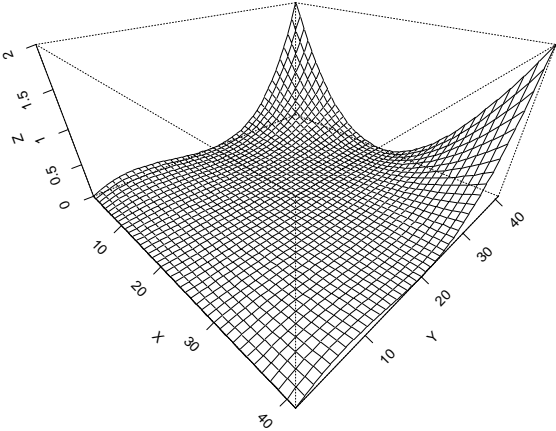


Figure 4: Graphical summary of the *be going to* data.

Legend

P = Place Complement
A = Agentive Complement
S = Sentient Subject, Nonagentive Complement
N = Nonsentient Subject, Nonagentive Complement

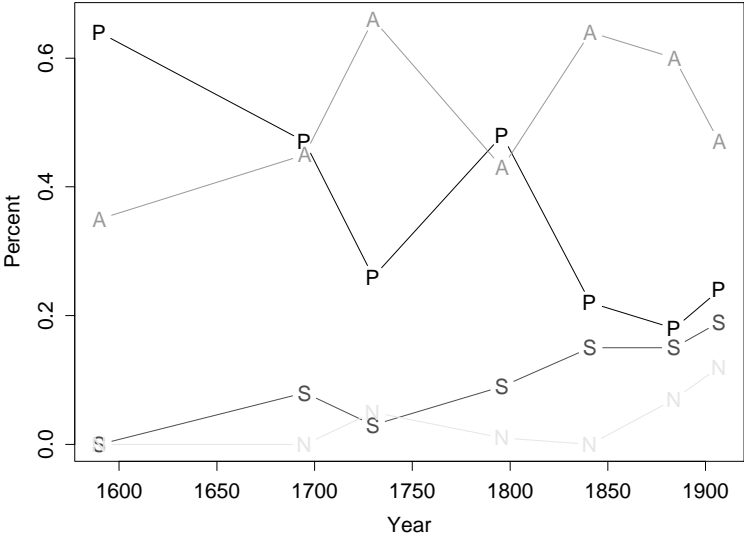


Figure 5: Lexical Reclassification in a Probabilistic Discrete-Category System. The numbers on the right give hypothetical relative probabilities of the rules during three successive intervals.

		$t < t_1$	$t_1 < t < t_2$	$t_2 < t$
S	→ NP VP	1.00	1.00	1.00
VP	→ Aux VP	0.80	0.80	0.80
VP	→ VP[Mot]	0.05	0.05	0.05
VP	→ VP[Percep]	0.15	0.15	0.15
VP[Mot]	→ V[Mot]	0.40	0.40	0.40
VP[Mot]	→ V[Mot] PP[Loc]	0.60	0.60	0.60
V[Mot]	→ walk[INFL]	0.30	0.60	1.00
V[Mot]	→ go[INFL]	<u>0.70</u>	<u>0.40</u>	<u>0.00</u>
X[INFL]	→ X	0.30	0.30	0.30
X[INFL]	→ X ed	0.50	0.50	0.50
X[INFL]	→ be X ing	0.20	0.20	0.20
Aux	→ will	1.00	0.95	0.55
Aux	→ be going to	<u>0.00</u>	<u>0.05</u>	<u>0.45</u>

Figure 6: Average Locations of Motion and Raising Auxiliary Verbs throughout the Modern English period.

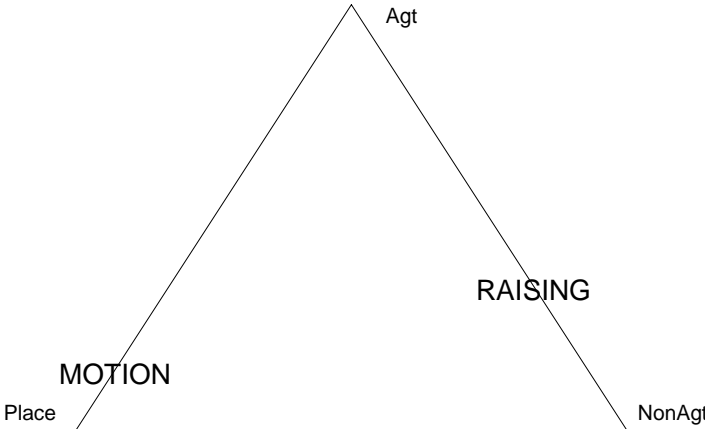


Figure 7: Historical *be going to* data and linear interpolation model.

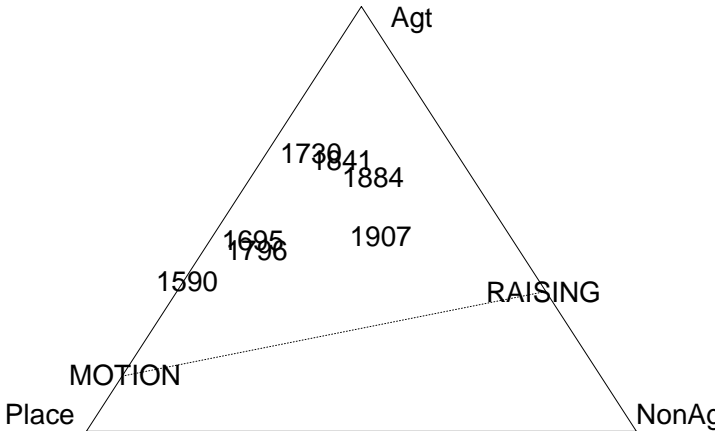


Figure 8: Verbs clustered around prototypes and an illustrative nonlinear interpolation.

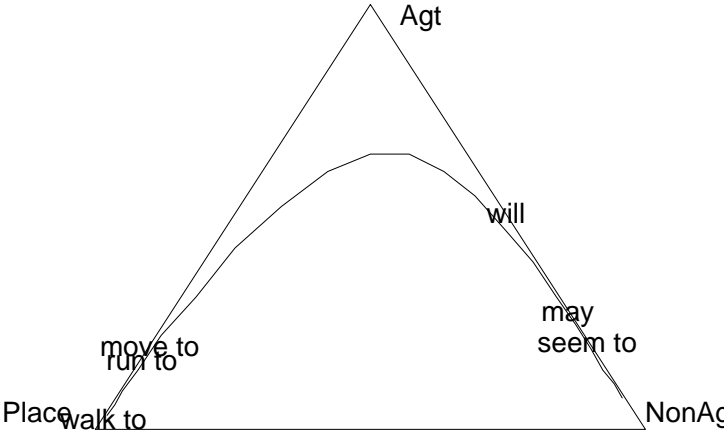


Figure 9: Network model—3 output dimensions.

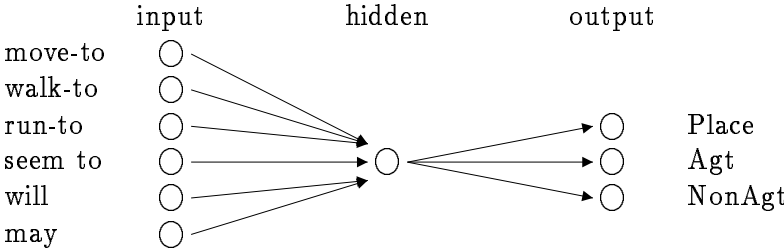


Figure 10: Two hidden unit clusters: motion and auxiliary verbs

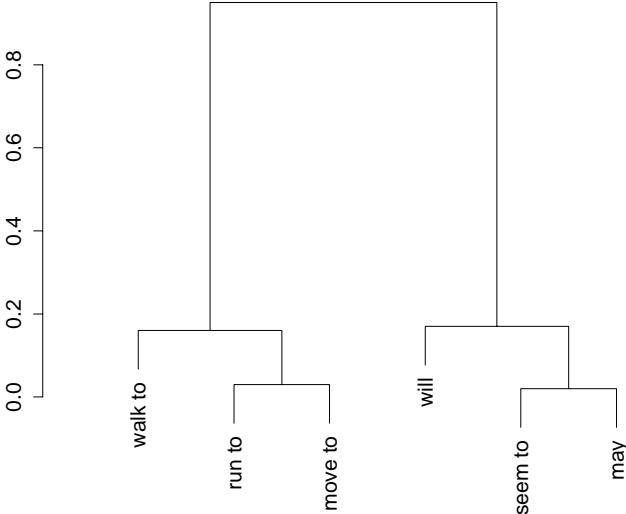


Figure 11: Comparison of linear model (dotted line), network model (solid line), and *be going to* data (numbers) in 3 output dimensions.

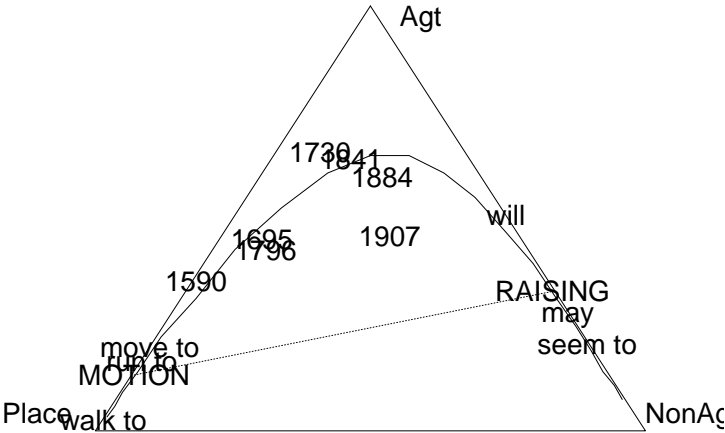


Figure 12: The threshold separating ungrammaticality from grammaticality of Non-Agentive VP Complements.

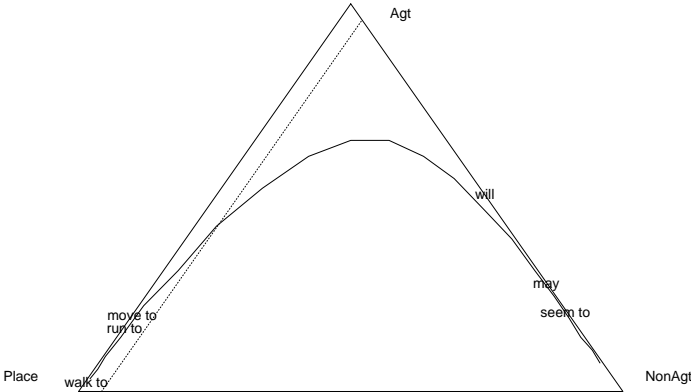


Figure 13: Locations of Motion, Equi, and Raising Clusters (4 Dimensions).

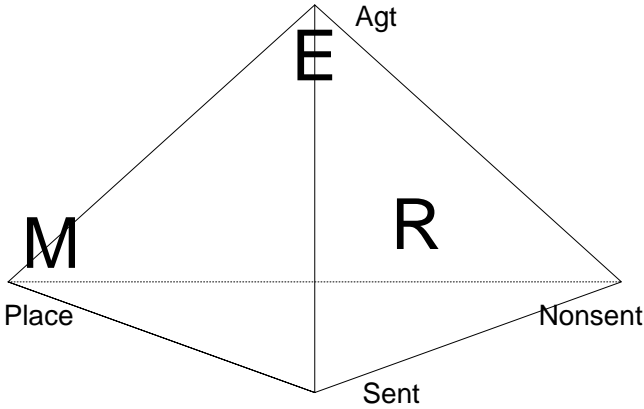


Figure 14: Historical Data (4 Dimensions).

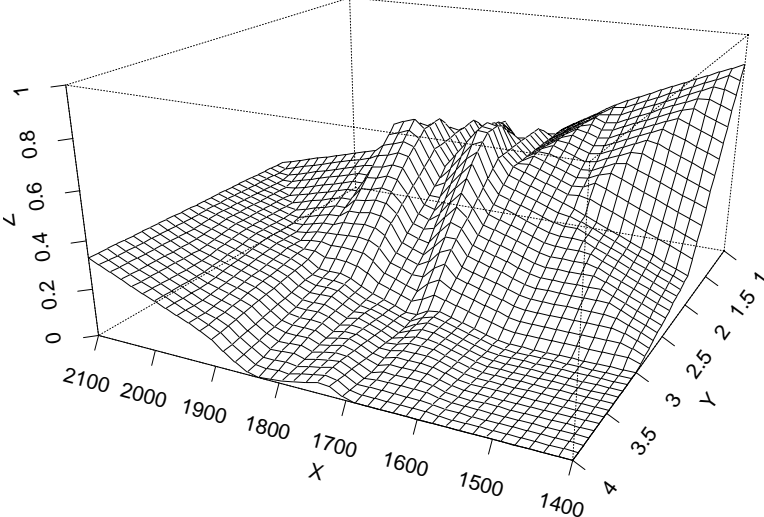


Figure 15: Linear Interpolation (4 Dimensions).

Legend	
X:	Year
Y:	1 = Place NP
	2 = Agentive VP Compl.
	3 = Sentient Subject + Non-Agentive Compl.
	4 = Non-Sentient Subject + Nonagentive Compl.
Z:	Relative Frequency

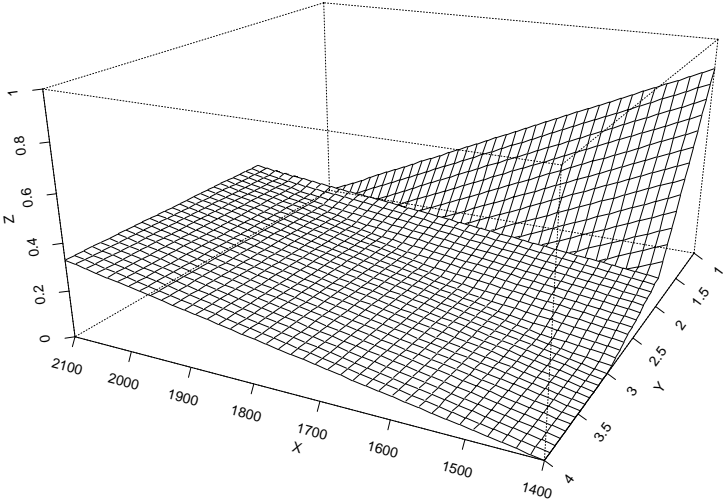


Figure 16: The Two-Step Model (4 Dimensions).

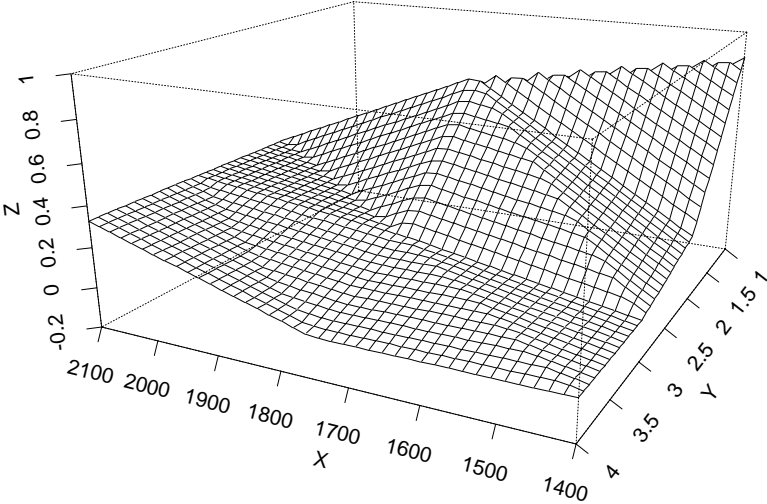


Figure 17: Smooth Manifold Model (4 Dimensions).

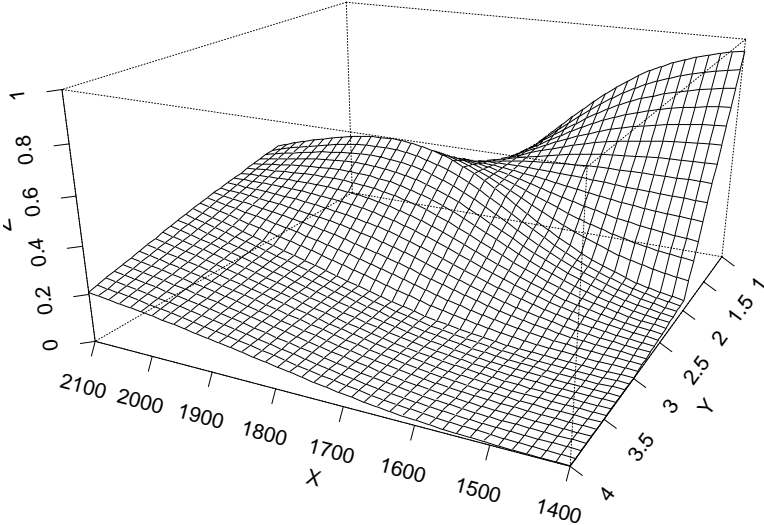


Figure 18: A 3 layer network for word-prediction: feedforward connections between the layers and complete innerconnectivity in the hidden layer (cf. Elman, 1990, 1991.)

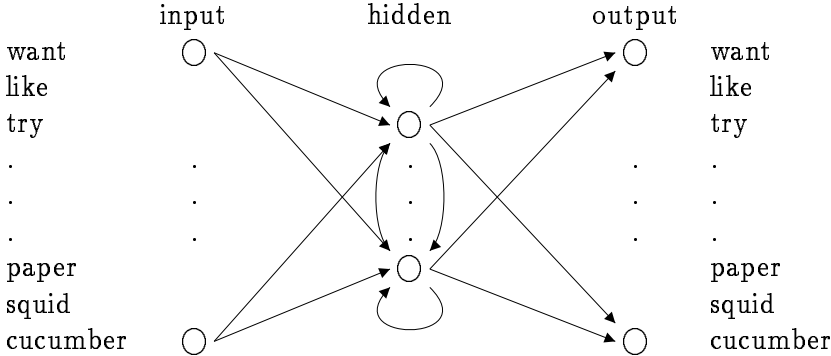


Figure 19: An approximation of the pre 16th century distribution of *be going to*. ((X / Y)) means “expands to X or expands to Y but not both”. “p” is the end-of-sentence marker.

0.25	S : Smain p
0.25	S : Smot p
0.25	S : Sequi p
0.25	S : Sraising p
0.25	Smain : NP[Sent] VP[Agent]
0.41	Smain : NP[Sent] VP[Nonagent]
0.24	Smain : NP[Thing] VP[Nonagent]
0.10	Smain : NP[Dummy] Vambient
0.88	Smot : NP[Sent] Vmot (ing) Prep NP[Place]
0.12	Smot : NP[Sent] Vmot (ing) to VP[Agent]
0.90	Sequi : NP[Sent] Vequi (ing) to VP[Agent]
0.10	Sequi : NP[Sent] Vequi (ing) to VP[Nonagent]
0.25	Sraising : NP[Sent] ((Vlightraising (ing) to / Vauxraising)) VP[Agent]
0.41	Sraising : NP[Sent] ((Vlightraising (ing) to / Vauxraising)) VP[Nonagent]
0.24	Sraising : NP[Thing] ((Vlightraising (ing) to / Vauxraising)) VP[Nonagent]
0.10	Sraising : NP[Dummy] ((Vlightraising (ing) to / Vauxraising)) Vambient
.	.
.	.
.	.
0.33	NP[Sent] : Lorrie, Bill, theGrocer
0.25	NP[Place] : Sutro, PointReyes, LA, theWharf
0.20	NP[Thing] : squid, cucumber, paper, TV, it
1.00	NP[Dummy] : it
.	.
.	.
.	.
0.25	Vmot : walk, run, move, go
0.33	Vequi : want, like, try
0.50	Vlightraising : seem, tend
0.25	Vauxraising : may, will, could, must

Figure 20: Recurrent Network *be going to* Trajectory (4 Dimensions).

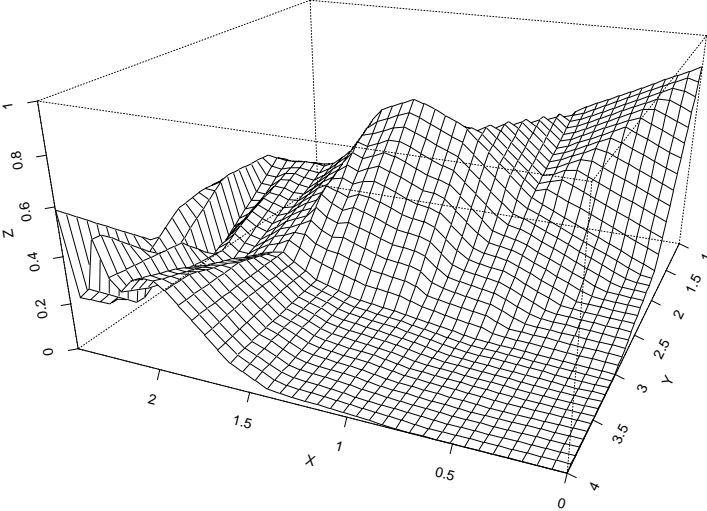


Table 1: Texts used in the quantitative study. The date assigned to each group of texts is the midpoint of the author's life.

Year	Author	Works	Source
1590	Shakespeare	All plays	OTA
1695	Defoe	Moll Flanders, Robinson Crusoe	OTA BOL
1730	Fielding	Tom Jones,	BS
		Joseph Andrews, Preface to Shamela, Shamela	OTA
1796	Austen	Emma, Mansfield Park, Northanger Abbey, Persuasion, Pride & Prejudice (part)	PG PG
1841	Dickens	The Chimes, Great Expectation, The Cricket on the Hearth, A Tale of Two Cities	OTA BS
1884	Hardy	Far From the Madding Crowd, Jude the Obscure, Return of the Native (part)	OTA ES
1907	Lawrence	Lady Chatterly's Lover, Sons and Lovers (part)	BOL

Key to Sources

BS	Book Stacks Library	http://www.books.com/lib1.htm
BOL	Books On-Line	http://www.cs.cmu.edu/afs/cs/misc/mosaic/common/omega/Web/books.html
ES	CMU English Server	http://english-www.hss.cmu.edu/
PG	Project Gutenberg	http://promo.net/pg/
OTA	Oxford Text Archive	email: archive@uk.ac.oxford.vax

Table 2: Emergence Chronology as indicated by First Attestations.

Advent	Form	Type
OE or before	Place Compl. (<P>)	Motion
ME	Sent. Subject, Agt. Compl. (<A>)	Motion/Equi
late 16th century	Sent. Subject, Agt. Compl. (<A>)	Equi
early 17th century	Sent. Subject, NonAgt. Compl. (<S>)	Equi/Raising
mid 18th century	NonSent. Subject, NonAgt. Compl. (<N>)	Raising
mid 19th century	Dummy Subject, NonAgt. Complement (<N>)	Raising

Table 3: Corpus data on the development of *be going to* from 1590 to 1907.

	1	2	3	4	5	6	7	8	9	10	11	12	Total
V	+	+	+	+	+	+	?	
P	+	+	+	+	+	?	
M	+	+	+	.	.	+	?	?	
S	+	+	.	+	.	+	+	+	+	.	.	?	
A	+	+	+	+	.	.	.	?	
D	+	.	?	
1590	61%	0%	3%	0%	0%	29%	6%	0%	0%	0%	0%	0	31
1695	34%	10%	3%	0%	0%	8%	0%	37%	8%	0%	0%	2	64
1730	25%	2%	0%	0%	0%	16%	0%	52%	3%	0%	3%	1	65
1796	47%	1%	0%	0%	0%	3%	3%	37%	9%	0%	1%	0	150
1841	18%	1%	3%	0%	0%	3%	1%	60%	15%	0%	0%	4	153
1884	17%	0%	1%	0%	0%	5%	3%	52%	15%	4%	3%	1	150
1907	23%	1%	0%	0%	0%	4%	0%	43%	19%	1%	11%	1	150
	<P>	<P>	<P>	<P>	<P>	<A>	<A>	<A>	<S>	<N>	<N>		

Features		Sources	
V	= VP complement	1590	Shakespeare
P	= Place complement	1695	Defoe
M	= Motion interpretation	1730	Fielding
S	= Sent. subject	1796	Austen
A	= Agt. complement (<i>intend</i> -sub.)	1841	Dickens
		1884	Hardy
D	= Dummy subject	1907	Lawrence

Table 4: Centers of mass of the relative frequency curves in Figure 1.

Construction	Center of Mass
Place <P>	1738
Agentive Compl. <A>	1788
Sentient Subj., Nonagt. Compl. <S>	1841
Nonsent. Subj., Nonagt. Compl. <N>	1861

Table 5: Corpus data on the distributions of Motion Verbs, Equi Verbs and Raising verbs in Modern English.

		1	2	3	4	5	6	7	8	9	10	11	12	Total
V		-	-	-	-	-	+	+	+	+	+	+	?	
P		+	+	+	+	+	-	-	-	-	-	-	-	
M		+	+	+	-	-	+	?	-	-	-	-	-	
S		+	+	-	+	-	+	+	+	+	-	-	-	
A		+	-	-	-	-	+	+	+	-	-	-	-	
D		-	-	-	-	-	-	-	-	-	+	-	-	
walk to	M	97%	0%	0%	0%	0%	3%	0%	0%	0%	0%	0%	0	65
run to	M	60%	0%	12%	4%	7%	16%	0%	0%	0%	0%	0%	2	69
move to	M	62%	0%	16%	0%	3%	12%	6%	0%	0%	0%	0%	0	32
like to	E	0%	0%	0%	0%	0%	0%	0%	73%	27%	0%	0%	2	130
want to	E	0%	0%	0%	0%	0%	0%	0%	81%	18%	0%	1%	2	143
try to	E	0%	0%	0%	0%	0%	0%	0%	89%	11%	0%	0%	1	144
seem to	R	0%	0%	0%	0%	0%	0%	0%	20%	43%	3%	34%	2	160
may	R	0%	0%	0%	0%	0%	0%	0%	28%	33%	11%	27%	0	193
will	R	0%	0%	0%	0%	0%	0%	0%	51%	22%	4%	22%	2	189
		<P>	<P>	<P>	<P>	<P>	<A>	<A>	<A>	<S>	<N>	<N>		

Features		Classes	
V	= VP complement	<P>	= Place compl.
P	= Place complement	<A>	= Sent. subj., Agt. compl.
M	= Motion interpretation	<S>	= Sent. subj., non-Agt compl.
S	= Sent. subject	<N>	= Non-Sent. subj., non-Agt. compl.
A	= Agt. complement		
D	= Dummy subject		