Lexical Change as Nonlinear Interpolation

Whitney Tabor
Center for the Sciences of Language and Department of Brain and Cognitive Sciences
University of Rochester
Rochester, NY 14627
whitney@psych.rochester.edu

[In Proceedings of the 17th Cognitive Science Conference, Moore and Lehman, eds. Lawrence Erlbaum.]

Abstract

Current, rule-based theories of grammar do not provide much insight into how languages can develop new behaviors over time. Yet, textual data indicate that languages usually evolve new grammatical patterns by gradually extending existing ones. I show how a grammar model that is sensitive to prototype structure can model innovation as a process of extrapolation along salient dimensions of the category clusters. A Connectionist network provides a usefully interpretable implementation. Confirming evidence comes from a study of the development of English *be going to* as a marker of future tense.

1. Introduction

How can a highly structured system evolve new behaviors without undergoing dissolution in the process? Natural language is an especially good domain in which to investigate this question: linguistic theory reveals the highly categorical, rule-governed character of its behavior at any point in time; but historical texts and studies of usage in communities show that change takes place by gradual metamorphosis of the existing system. Here I concentrate on grammatical class membership change, taking as a case study the development of the future auxiliary usage of English *be going to*. A key step in developing a model of this process is to take into account information about the relative frequencies of words in grammatically-defined contexts.

I model language change by positing a grammar corresponding to every point in time. Following common linguistic practice, I refer to a single grammar at one point in time as a "synchronic" model, and to a sequence of grammar-states as a "diachronic" model. It is useful to examine a diachronic model by drawing a diagram of (part of) its behavior-space. Consider *be going to*. Originally, *go* was strictly a verb of motion and hence could only be used in the kinds of environments that accomodated verbs like *walk*, *run*, *ride*. However, during the past five or six centuries, the particular use of *go* in the expression *be going to* has expanded its capabilities and become a marker of future tense as well as a motion verb. Thus one can now say

(1) It is going to rain tomorrow. [Non-agentive VP Compl]

where the complement of *be going to* is a Non-Agentive verb phrase (VP) and a motion interpretation is not plausible. In this usage it is reasonable to say that *be going to* is a type of auxiliary verb. Clearly, English is currently in the middle phase of this transition for we can still say:

(2) She is going to Sarajevo. [Place-denoting NP Compl]

where the complement is a Place-denoting Noun Phrase (NP) and only a motion interpretation is possible. A type that played an important role in the transition from old to new, and which I will say more about below, is the now-ambiguous case,

(3) I am going to deliver this letter. [Agentive VP Compl]

with an Agentive VP complement. If we consider the relative likelihood with which a given grammar predicts each of these types, then the behavior at any given time can be characterized as a 3-dimensional vector. Moreover, since the three values form a probability distribution, they are restricted to appearing in the triangle with vertices (1,0,0), (0,1,0), and (0,0,1) in 3-space.[1] Figure 1, showing just the triangle, shows points corresponding to motion verbs and canonical auxiliary verbs like *will* based on instance counts in a corpus.[2] These verbs' behaviors have not changed much in the relevant regards during the course of the 16th through the 20th centuries (see Warner (1990) for information on *will*), so Figure 1 is a reasonable approximation of the behavior diagram for these verbs at every point during this period.

I model the change in *be going to*'s status as a process of lexical reclassification, treating *be going to* as a single word. This is a nontrivial simplification but it permits the formulation of a model that is both reasonably accurate and easily interpretable.

Standard linguistic grammars do not normally make predictions about relative frequencies, but there is a natural extension of any generative model which turns it into a statistical model: probabilities can be assigned to the generative rules or parameters.[3] Under such models, with their categorical treatment of lexical representation,

[1] This representation is called a *ternary diagram*.

[2] For the auxiliary verbs, I sampled *will*, *may*, and *seem to*. For the motion verbs, I sampled *walk*, *run*, and *move*.

[3] Examples include the probabilistic context free phrase structure grammars used by computational linguists (see Charniak, 1993), the Competing Grammars model of Kroch (1989), the probabilistic Principles and Parameters model of Clark and Roberts (1991).

Table 1: Quantitative data from the history of *be going to*—3 dimensions.

Year	Source	1. Place	2. Agt	3. Nonagt	Tokens
1590	Shakespeare	64%	35%	0%	31
1695	Defoe	47%	45%	8%	62
1796	Austen	48%	43%	10%	150
1841	Dickens	22%	64%	15%	151
1884	Hardy	18%	60%	22%	149
1907	Lawrence	24%	47%	31%	142
1911	Joyce	19%	58%	23%	124
1970	London Lund	12%	48%	38%	139
MOTION	(Start state)	88%	12%	0%	164
AUXILIARY	(End state)	0%	34%	66%	374

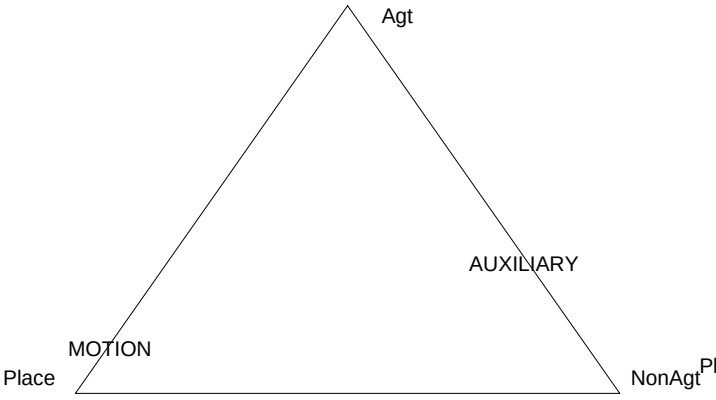


Figure 1: Relative positions of Motion Verbs and Canonical Auxiliary Verbs.

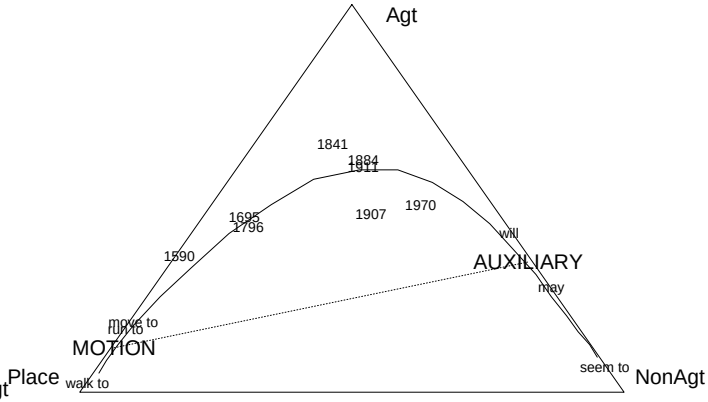


Figure 2: The succession of states of *be going to*—3 dimensions. The dotted line is the Linear Model. The solid curve is the Nonlinear Model.

it is expected that all behaviors which arise during the process of lexical reclassification can be modelled by assigning words, with varying probabilities to one or more of the existing grammatical classes. As a consequence, lexical reclassification is expected to involve purely linear interpolation between behavior states in all but a few cases.[4]

If this Linear Interpolation Model is correct, then the study of lexical reclassification is not of much interest from the standpoint of learning about innovation, because linear interpolation involves only mixture of existing types. Here, however, I propose an alternative nonlinear interpolation model with a smoothness constraint, which suggests a more interesting scenario. The notion of prototype structure plays a central role: if words that would be assigned to a single monolithic class in a standard linguistic grammar show statistical variation that is asymmetrically distributed around a prototype, then the model forms a representation that is locally alligned with the prototype structure. Because of the smoothness constraint, this allignment affects interpolated states as well, so the trajectories of words that change classes are expected to move along channels dictated by the prototype structure. In some cases, this influence produces a trajectory far-removed from the prediction of the linear model—one that can more reasonably be said to involve the occurrence of novel behavior.

The current work complements its closest relative, Hare and Elman (1992 and 1993), which shows how prototype-structured categories can appropriately model the tendency of certain linguistic classes to attract new elements in historical change. I show how such categories properly constrain the trajectories of changing elements even when they don't capture them as permanent members. Independent support for the notion that languages have prototype-structured categories has been provided by researchers in Cognitive Grammar (Langacker, 1987; Lakoff, 1987). Evidence that prototype-based categories play an important constraining role in historical change has been provided by Warner (1990) and Kemmer & Israel (to appear).

A Connectionist network provides a useful implementation of the Nonlinear Interpolation model. Consider a 3-layer network with a hidden layer that is smaller than its output layer. Suppose we train this network as follows: lexical items are given distinct indexical bit representations[5] on the input layer and behaviors (e.g. the behavior of occurring in a particular position in particular sentence frame) are given distinct indexical bit representations on the output layer; each training example consists of a single input paired with a single output; the relative probability of each input-output pair is the observed relative frequency of the item-with-behavior in a sample corpus. For each input unit, we want the activation of each output unit to converge on the likelihood that the item corresponding to the input unit will exhibit the behavior corresponding to the output unit. Thus, the output activations form a probability distribution. Given these specifications, a network with multinomial error function, fixed-sigmoid hidden units, and softmax output units, trained with backpropagation is appropriate (Rumelhart et al., 1995).

I model lexical reclassification in this framework by training a network on a set of elements belonging to different classes and considering a straight-line trajectory in the hidden unit space from a location associated with one class (Motion verb) to a location associated with a different class (Auxiliary verb). Although the trajectory is linear in the hidden unit space, it may not be linear in the output space, so the model's predictions differ from those of the standard grammar models.

2. Case Study: English *be going to*

Quantitative data on the change of *be going to* in the part of behavior space outlined in the previous section are shown in Table 1.

Figure 2 shows the corresponding ternary diagram along with the predictions of the Linear and Nonlinear Interpolation Models. The Nonlinear model is a Connectionist network with 10 input units, 1 hidden unit, and 3 output units. The inputs fall into two classes of 5 members each whose behaviors are clustered around the points labelled "MOTION" and "AUXILIARY". Crucially, the MOTION inputs vary only along dimensions 1 and 2 (Place and VP Agentive), while the AUXILIARY inputs vary only along dimensions 2 and 3 (VP Agentive and VP Non-Agentive). This kind of restricted prototype scatter reflects a situation common in language use: there is high variation along dimensions that are simultaneously allowed by a categorical grammar but essentially no variation along dimensions that are disallowed.

Clearly the historical trajectory is skewed in the direction predicted by the Nonlinear Model. This is an interesting finding not only because it shows how quantitatively unusual behavior can arise in the course of a simple reclassification episode, but also because it indicates that by monitoring subtle quantitative changes, it may be possible to make predictions about subsequent categorical changes: in Shakespeare (1590), *be going to* shows no instances of novel behaviors (Danchev and Kytö, 1991), but it's distribution is highly skewed in the direction that indicates imminent appearance of such behaviors.

It turns out that the history of *be going to* is more complex than this simple portrayal indicates. Linguistic theories generally distinguish two types of auxiliary verbs, called *Equi* and *Raising*. Criteria often considered diagnostic of Raising status include ability to take "dummy" subjects (*It seemed/appeared/tended to rain.*, *There seemed to be a thundercloud on the horizon.*) and ability to intervene in idioms (*The cat seems to be out of the bag.*). Equi verbs contrast in both regards. (e.g., McCawley, 1988). We can add the fact that Raising verbs

[4] The exceptions are situations where the changing word appears more than once in a grammatical structure defining a behavior. Such cases are rare and are not usually pertinent to assessing the main properties of a word's behavior so I will not consider them further here.

[5] By an *indexical bit representation* I mean a vector with a value of 1 on one dimension and 0 elsewhere.

Table 2: Quantitative data from the history of *be going to*—4 dimensions.

Year	Source	1. Place	2. Agt	3a. Sent/Nonagt	Nonsent/Nonagt	Tokens
1590	Shakespeare	64%	35%	0%	0%	31
1695	Defoe	47%	45%	8%	0%	62
1796	Austen	48%	43%	9%	1%	150
1841	Dickens	22%	64%	15%	0%	151
1884	Hardy	18%	60%	15%	7%	149
1907	Lawrence	24%	47%	19%	12%	142
1911	Joyce	19%	58%	13%	10%	124
1970	London Lund	12%	48%	20%	18%	139
MOTION	(Start state)	88%	12%	0%	0%	164
EQUI		0%	91%	8%	1%	415
RAISING	(End state)	0%	34%	37%	29%	374

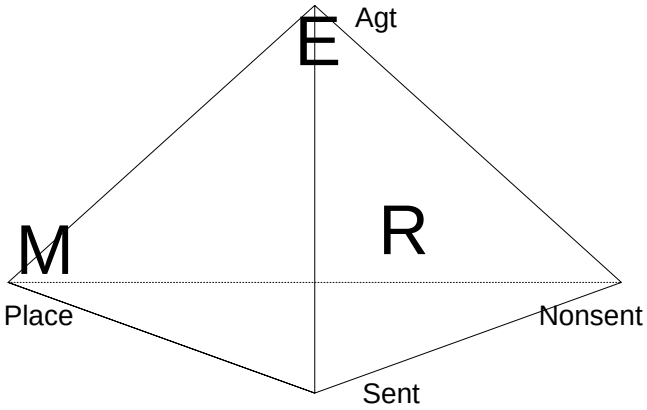


Figure 3: Motion, Equi, and Raising Verb locations.

permit inanimate subjects while Equi verbs do not, except in an anthropormorphic 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 (Nunberg, Sag, and Wasow 1994). By these diagnostics, the auxiliary verb *will* is a raising verb in its use as a future marker, for one can say things like *It will rain* and *The table will fit in the corner.*

Perez (1990) notes that *be going to* went through a stage in the 17th-19th centuries in which it seemed to have a meaning like *intend* before it developed the very general distribution it has today. Thus we might hypothesize that *be going to*'s transition was actually a three stage process: Motion → Equi → Raising. This provides an interesting challenge for the diachronic model: can it nonstipulatively generate the intermediate stage?

To clearly distinguish Equi and Raising verbs, it will be useful to make the behavior space 4-dimensional by breaking Behavior 3 down into two parts: nonagentives with sentient subjects and nonagentives with nonsentient subjects.

(3a) The man was going to faint. [Sentient Subject, Non-Agentive VP]

(3b) There is likely to be an eclipse. [Non-Sentient Subject, Non-Agentive VP]

This makes Raising verbs categorically different from Equi verbs since Equi verbs do not normally occur with nonsentient subjects (** There expected to be an eclipse*), although they sometimes occur with nonagentive complements (*Gregor expected to faint*). Table 2 gives quantitative data under this portrayal.

Under the constraint that they form a probability distribution in 4-space, the outputs are now restricted to a three-dimensional subset in the form of a tetrahedron. Figure 3 shows the mean locations of Motion, Equi, and Raising verbs on the tetrahedron. Figure 3 is unambiguous given the knowledge that all the depicted points lie on exposed surfaces (and edges) of the tetrahedron. However, since *be going to*'s trajectory need not lie entirely on the surface, I now switch to a different display scheme. Figure 4a shows a three-dimensional "parallel coordinate" image of the trajectory predicted by the Linear Interpolation Model. The X-axis marks time-points, the Y-axis marks behavioral category, and the Z-axis measures relative probability. Figure 4b shows the trajectory predicted by a nonlinear, connectionist model (time has been scaled for a good fit). This model has 15 input units, 2 hidden units, and 4 output units. As in the last example, it was trained on data with prototype scatter along dimensions that are categorically positive for each word-type. Note the significant skewing of the trajectory in the direction of Equi status in the middle of the transition (indicated by the peak in behavior 2 around 1800). Figure 4c shows the historical data.[6] Indeed, this diagram reveals a significant skewing in the direction of Equi behavior, confirming Perez's impressionistic observations and confirming the predictions of the Nonlinear Model. It is interesting to note, however, that *be going to* does not actually ever inhabit a canonical Equi state but only passes near such a state. This seems in keeping with the observation that while there are many historical examples during the 17th–19th centuries which can reasonably be assigned an *intend* interpretation, there seem to be none that definitively require it. Example (4) is a typical case.

(4) c. 1695 He was going to reply...but he heard his sister coming, Defoe, *Moll Flanders.*

3. Conclusion

I started by touting natural language as a worthy domain for studying change in highly structured systems. Focusing on lexical reclassification episodes, I noted that standard models predict nothing that could reasonably called "innovation", for they only perform linear interpolation, which is equivalent to mixture of existing types. By contrast, a simple Connectionist model performs nonlinear interpolation and hence predicts interestingly novel states. I noted that data distributed around reduced-dimension prototypes, as is typical in natural language, interacts with this Nonlinear Interpolation model in a strong way: transitions are expected to be locally constrained by the prototype structure. This prediction is born out by data from the history of English *be going to*. One conclusion of interest is that subtle quantitative shifts may anticipate significant categorical developments and hence have predictive value. Also, quantitative variation is intimately connected with optionality: linguistic items show significant quantitative variation only when grammar is indeterminate as to how something should be said. While it is tempting to think of optional decisions as inherently unbinding (if one chooses A one day, and A is optional, then one may legitimately choose $\neg A$ the next), the present study indicates that a sequence of correlated optional decisions can, in the domain of grammar, bring about revision of the rules, so that what was once optional becomes mandatory or vice versa. The conditions under which such creeping systemic revision can occur must be of interest not only to historical linguists, but also to biologists and sociologists—even politicians!

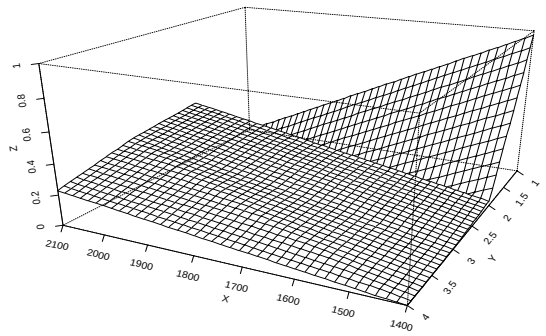
Acknowledgements

Thanks to Robert Jacobs, Michael Spivey-Knowlton, and anonymous reviewers for helpful comments. The text corpora where made available by the Oxford Text Archive, BookStacks, Project Gutenberg, and the Trent University Archive. The research has been supported in part by postdoctoral fellowship funding to the Center for the Sciences of Language (NIH-NIDCD 5T32DC0035-04).

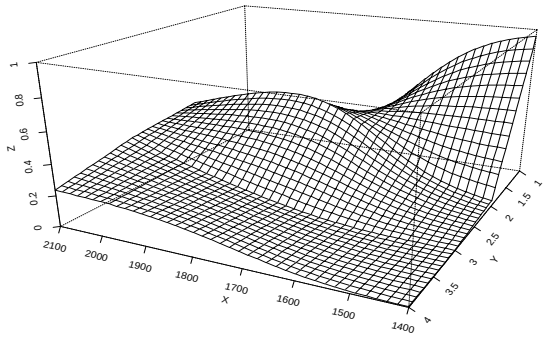
[6] The pure motion verb starting state and the (hypothetical) pure future auxiliary ending state have been included at the dates 1400 and and 2100, respectively, to facilitate comparison with the two interpolation models.

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

a. Linear interpolation.



b. Nonlinear interpolation.



c. Historical Data.

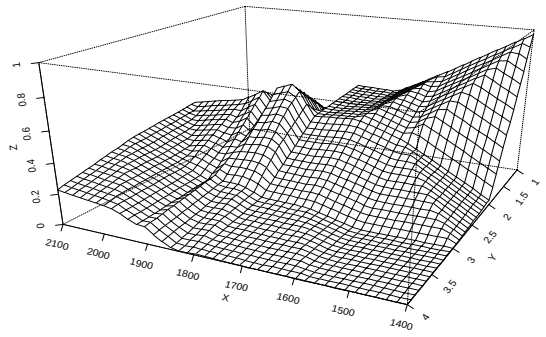


Figure 4: Comparison of Models—4 dimensions.

References

Charniak, E. (1993). *Statistical Language Learning.* MIT Press, Cambridge, Massachusetts.

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.

Hare, M. and Elman, J. L. (1992). A connectionist account of English inflectional morphology: evidence from language change. In *14th Cognitive Science Proceedings*, pages 265–270. Lawrence Erlbaum Associates.

Hare, M. and Elman, J. L. (1993). From *weared* to *wore*: a connectionist account of language change. In *14th Cognitive Science Proceedings*, pages 265–270. Lawrence Erlbaum Associates.

Kemmer, S. and Israel, M. ((to appear)). Variation and the usage-based model. In *CLS 30 Parasession on Variation and Linguistic Theory.* University of Chicago Press.

Kroch, A. S. (1989). Reflexes of grammar in patterns of language change. *Journal of Language Variation and Change*, 1(3):199–244.

Lakoff, G. (1987). *Women, Fire, and Dangerous Things.* University of Chicago Press, Chicago.

Langacker, R. W. (1987). *Foundations of Cognitive Grammar, v. 1.* Stanford University Press, Stanford, California.

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

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

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.

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

Warner, A. (1990). Reworking the history of English auxiliaries. In Adamson, S., Law, V., Vincent, N., and Wright, S., editors, *Papers from the Fifth International Conference on English Historical Linguistics.* John Benjamins.