Metric Grammars

Whitney Tabor (whitney.tabor@uconn.edu)

Hyosun Lee (hyosun.lee@uconn.edu)

Department of Psychological Sciences-U1020
University of Connecticut
Storrs, CT 06269-1020 USA

Abstract

Many of the most interesting and vexing problems in linguistic analysis concern structures that have a hybrid character—they show evidence of belonging to two independently motivated types. Proposals often assign them to one or the other class, requiring complication of the theory to handle their exceptionality. We suggest that there is no satisfactory answer to such conundrums under standard, type-based representational theories, for those theories are founded on discrete topologies. These are ill-suited for blending. As an alternative, we propose *Metric Grammars*–grammatical systems founded on connected topologies (specifically on manifolds on which we can define continuous-valued metrics—e.g., Euclidean distance). Such topologies are the natural residence of dynamical systems. Indeed, dynamical systems theory has identified a formal phenomenon–the bifurcation—that naturally models phenomena of intermediacy. Bifurcations are also often used to model structural transitions. Therefore, to explore the possible relevance of our novel representational approach, we focus on transitional linguistic phenomena, namely diachronic grammaticalization episodes. These are cases where, over the course of the history of a language, a morpheme (or combination of morphemes) gradually changes its grammatical status. A *metric grammar*, a recurrent map with a neural network at its core, changes its grammatical system slightly with each instance of language use. Focusing on a particular episode from the history of English—the development of "sort of" and "kind of" from Noun-Preposition structures into adverbs—we provide evidence that metric grammars exhibit statistical anticipation of categorical change, a phenomenon that has been documented for several grammaticalization episodes and is difficult to account for with discrete-topology models.

Keywords: metric grammars; grammaticalization; morphosyntax; language change; self-organization; dynamical systems; bifurcation; emergent structure; context free grammars; neural networks; phrase embedding

Introduction

Large Language Models (LLMs) have vigorously evolved in the past decade. Recently, a number of researchers have offered evidence that including phrase embeddings (vector space encodings of phrasal units), in addition to word embeddings, allows LLMs to better capture semantic subtleties (Borensztajn et al., 2009; Le & Zuidema, 2014; Li et al., 2021; Nikzad-Khasmakhi et al., 2021; Park & Kim, 2022; Wang et al., 2021; Wu et al., 2020, 2024). This suggests potentially beneficial interaction between formal linguistic theory and LLM engineering, but at present such bridging is hampered by the analytic opaqueness of LLMs. Here we propose a model called a "Metric Grammar", which adopts similar formal assumptions to LLMs, but can be closely related to linguistic hierarchical analyses as they manifest in Probablistic Context Free Grammars (PCFGs).

Linguistic intermediacy

Having approached our topic from the engineering side, we now approach it from the theoretical side. Many core challenges for linguistic analysis center around elements that have a mixed character with respect to well-motivated grammatical categories: English gerunds (Jackendoff, 1991) and Korean *hada* constructions (Chae, 1997), for example, have mixed nominal and verbal properties; clitics across languages have the distributional properties of independent words, suggesting they are atoms of syntax, but the phonological properties of bound morphemes, making them more like lexicon-internal elements (Inkelas & Zec, 1990); certain prepositions that have evolved from verbs (especially in serial verb languages) retain some properties associated with their verbal origins (Lord, 1973).

Mental topology

We suggest that these cases are challenging because the theories are built on an inappropriate topology. To formalize things, we can think of mental states as points in a space. *Topology of the space* refers to fundamental assumptions we make about relationships between the states. In a *discrete topology*, the states are all separated from each other, there are no partial relationships between states. Standard model-theoretic treatments of language syntax and semantics operate on discrete topologies—this is sensible if one is mainly interested in properly classifying mental entities into distinct categories. In a *connected topology*, by contrast, there is no way to cleanly divide the space into separated subsets of points—in a sense, all points are in partial relationships to one another. One type of connected topology can be generated by positing real-valued distances between points that range from 0 to some (or all) positive values. This is called a *metric space*. Euclidean space has a connected topology with its standard metric, Euclidean distance. Artificial neural network models assign, typically via training, model states to a finite set of observed data points (their training set), and in doing so induce predictions about a continuum of other possible behavioral circumstances. They accomplish this by embedding that finite set of hypothesized mental states into a connected metric topology, thereby positing specific degrees of proxim-

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

ity between pairs of mental states.

Here, we use connected metric topologies as the foundation for mental representations of sentences. An advantage of working in a connected topology is that the framework not only handles intermediacy, but there is a way to model categories: they are collections of experience-prompted encodings that cluster together in the metric space. In particular, the theory posits an important role for *usage* in the sense of usage-based theories of grammar (J. Bybee, 2006; J. L. Bybee et al., 1994; Schmid, 2020): a pattern that gets used frequently creates a dense region of encoding which, if it is sufficiently dense and separated from other clusters, will warrant characterization as a category.

This paper is primarily a conceptual paper, but the conceptual account is backed up by a computer implementation. Near the end, we will report on an initial simulation result which provides evidence that, in this system, usage systematically affects form.

Bifurcations

Dynamical systems theory, which studies state-change functions on (connected) metric spaces, has useful tools for understanding intermediacy and transformations of form. We focus on discrete dynamical systems which have the form (1)

$$\mathbf{x}_{t+1} = f(\mathbf{x}_t, t) \text{ for } \mathbf{x} \in \mathbf{S} \tag{1}$$

Here, $\mathbf{S}$ is the connected state space, $\mathbf{x}$ is a vector specifying the mental state, x_t is the mental state at time t, f is a function from $\mathbf{S}$ to $\mathbf{S}$. The inclusion of the time index, t, as a parameter of f amounts to letting f characterize the environment of the system, which stimulates the system in various ways at various times, in addition to inherent system biases which put constraints on the evolution of $\mathbf{x}$. Here, we will take the environment to be a corpus of sentences, and the inherent biases to be knowledge of how to parse the sentences.

In a metric grammar, f has tunable parameters—these are instantiated as real-valued weights in an artificial neural network. In dynamical systems theory, such parameters are called *control parameters*. A particular setting of the control parameters typically generates a specific system whose behavior is organized around stable trajectories called *attractors*—these are sets in the state space such that if the system is started near one such set, it will converge on it under the dynamics (f) over time (this will generally depend on the environment behaving in a particular, reliable way). A *bifurcation point*, $\mathbf{z}$, is a point in control parameter space such that if the system follows a continuous path in this parameter space which passes through $\mathbf{z}$, then the attractor configuration just before it reached $\mathbf{z}$ is categorically different from the configuration just after it has passed $\mathbf{z}$. In the present case, the attractor configuration not only determines which parses (if any) of a perceived word sequence are perfectly grammatical, but it also assigns graded grammaticality values (called *harmonies*) to all possible sequences.

We use metric grammars to model change over historical time of the inventory of parses that speakers find grammatical. The model we describe is a model of the systematic language knowledge of an individual speaker (Chomsky, 1986). However, unlike classical I-language models (e.g., Gibson & Wexler, 1994) which are not able to track arbitrarily fine-grained variations in statistical language behavior because they have only finitely many parameter settings, metric grammars are sensitive to such variation and model it as part of their language system knowledge. Moreover, ensembles of metric grammars can show swarm behavior, and are thus relevant for modeling social contagion phenomena which arguably play a central role in grammar change (Enfield, 2008).

Grammaticalization

The morphemes that form a language can largely be divided into two types: content morphemes and function morphemes. *Content morphemes* are nouns, verbs, and adjectives—they directly specify elements of a world that speakers wish to talk about (English examples: *house, plum, algebra, drone, titter, blunt, delicious*). *Function morphemes* are grammatical morphemes—e.g., prepositions (*on*, *near*), complementizers (*that*, *whether*), affixes (*un-*, *-ly*), etc. Over historical time, new function morphemes often evolve from instances of content morphemes used in a particular context. Examples include the development of a Romance adverb-forming suffix from the Latin noun *mente* "mind" (e.g., Latin *pura mente* "(of) a pure mind" [Adjective Noun] > Italian *puramente* "purely" [Adverb]; Detges (2015)); Ewe *bé* "say" [Verb] > "that" [Complementizer] (Lord, 1976); English *sort/kind of* "'type of" [Noun Preposition] > "to a mild degree" [Adverb] (Tabor, 1995).

Such structural changes are generally accompanied by changes in the statistical behavior of the language. For example, the frequency of the grammatical morpheme-to-be substantially rises. This makes sense because grammatical morphemes are generally much more frequent than content morphemes. De Smet (2012) and Tabor (1995) provide evidence that not only do frequencies change during the process of grammaticalization, but they change in an anticipatory way—certain frequency changes foreshadow the appearance of the form in new structures. For example, Tabor (1995) offers evidence that, prior to the first clear uses of English *sort of* and *kind of* as adverbs (1-c), there was a rise in the frequency of *sort/kind of* in environments ambiguous between Noun-Preposition and Adverb senses (1-b). We will show via computational simulation that an evolving metric grammar system, driven by usage that simply heightens the occurrence rate of *sort/kind of* before adjective becomes more likely (albeit only very slightly in simulations run so far) to generate a novel adverbial form.

(1) a. We found a sort of crab. [Original: Noun + Prep]

b. We sought a sort of large pebble. [Ambiguous]

c. We sort of changed our plan. [Modern: Adverb]

Metric Grammars

It is helpful to introduce metric grammars in the context of formal languages. Let Σ be a finite alphabet of symbols. A *formal language* is a subset of Σ^*, where Σ^* is the set of finite strings of symbols drawn from Σ. Context Free Grammars (CFGs) are generation and recognition devices that specify one particular class of formal languages (Context Free Languages) that is arguably relevant to the characterization of natural languages. A *Context Free Grammar* is a finite set of rules of the form $M \rightarrow D_1 D_2 \ldots D_n$ (n finite; at least one of the rules is designated a "Starting Rule"). The symbol on the left of each rule is called the "Mother" symbol. The symbols on the right are called "Daughters". A *Probabilistic Context Free Grammar* (PCFG) is a CFG in which the rules are associated with probabilities (probabilities on rules with the same mother symbol sum to 1). (Figure 1) PCFGs support a corpus generation process in which the system repeatedly starts with a starting rule, recursively connecting mother nodes below to daughters above, under the constraints that the labels must match if a mother is to attach to a daughter, and multiple matching options are sampled according to their probabilities. The resulting hierarchical analysis is called a (P)CFG tree. Figure 2 shows a CFG tree of (1-a) under Grammar 1.

Metric Grammar Trees. Metric grammars employ *Metric Grammar Trees*. These are similar to CFG trees, but lie in a connected space. The space is high dimensional and is occupied by points which lie in clusters which correspond closely (but not exactly) to the classes of mother nodes in a Context Free Grammar. The dimensions of the high-dimensional space encode the statistics of observed word and phrase forms—this makes them analogous to Large Large Language Model (LLM) vector encodings, but we do this in a way, explained below, that makes the vector dimensions precisely interpretable. The elements of a metric grammar are not points, however, but *metric grammar treelets*—these are either triples of points consisting of a mother and two ordered daughter nodes (these correspond to branching rules in a standard CFG) or doubles of points (these correspond to lexical rules). Each lexical treelet specifies a word form. Figure 3 illustrates. This figure can be thought of as a 2-dimensional projection of a higher-dimensional space (14-dimensional in the simulations described below), where we have taken the liberty of positioning the clusters of points in a way that makes the geometry of the structure on the page roughly resemble the geometry of a standard tree diagram. The black lines indicate the treelets that are involved in the parse of sentence (1-a). However, all the points have treelet lines connecting them to other points—these are generally oriented similarly to the ones shown in the diagram, but we have omitted them to make the figure clearer.

Besides being specified by points in a connected metric space, Metric Grammar Trees are different from CFG trees in that that the featural specification of the mother of a treelet

1.00	S	→	NP VP
0.33	NP	→	Det N'
0.67	VP	→	V NP
0.33	VP	→	Adv VP
0.27	N'	→	AdjP N'
0.18	N'	→	N' PP
0.67	AdjP	→	Adj
0.33	AdjP	→	Adv AdjP
1.00	PP	→	P N'
1.00	Adv	→	Adj ly
	NP	→	0.37 *we, 0.18 they, 0.12 I*
	N'	→	0.30 *bear, 0.15 crab, 0.10 sort*
	Adj	→	0.67 *real, 0.33 mild*
	P	→	0.67 *of, 0.33 near*
	V	→	0.55 *found, 0.27 loved, 0.18 deplored*
	Det	→	0.67 *a, 0.33 this*

Figure 1: Grammar 1: A PCFG description of a fragment of English that includes relevant Early Modern English distributional characteristics of *sort of* and *kind of*. In the simulation reported below, for efficiency of training, we simplified the grammar by omitting determiners and pronouns and eliminating all but one lexical item of each type.

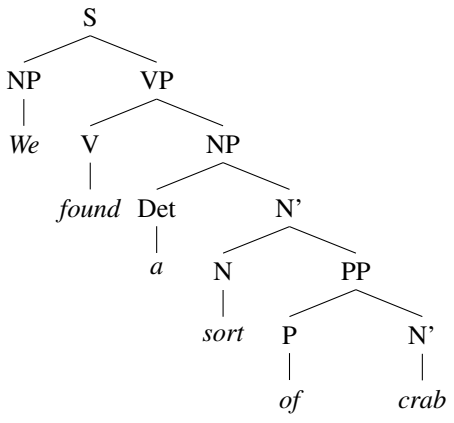

Figure 2: A plausible Early Modern English (as well as Modern English) parse of sentence (1-a) generated by Grammar 1.

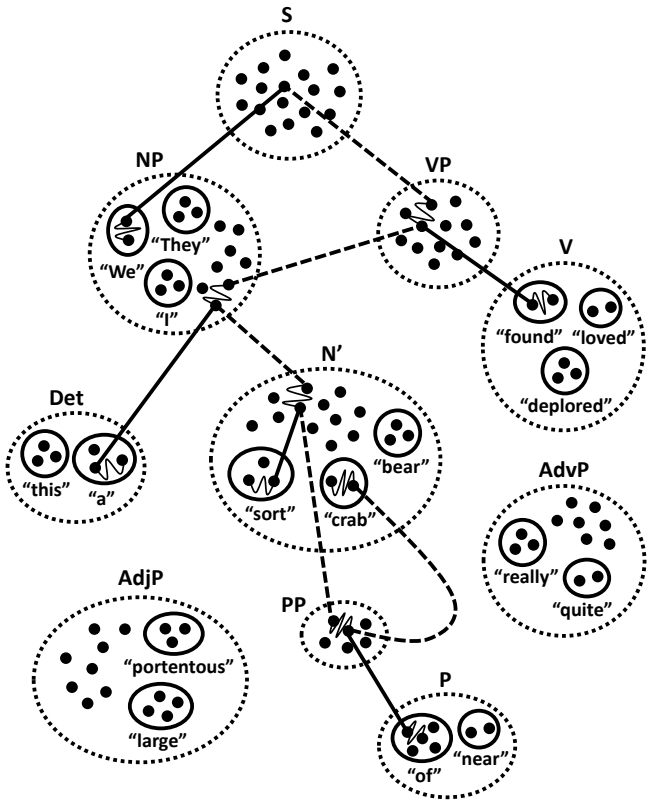

Figure 3: Schematic diagram of a metric grammar parse corresponding to the PCFG parse in Figure 2. Branching treelets are shown with two types of lines: solid lines for left daughters and dashed lines for right daughters. Squiggly lines highlight the distance from 'mother below' to 'daughter above' (the longer the squiggly lines, the lower the harmony of the parse). Two types of clusters are highlighted by circles in the diagram: solid circles surround lexical node clusters and dotted circles surround phrasal node clusters.

below that attaches to the daughter of a treelet above need not exactly match. This property is emphasized in Figure 3 by the wavy lines that connect daughters-above to mothers-below. Based on these separations, we define a measure of the grammaticality of the structure called its *harmony*, h as shown in (2).

$$h = \sum_i log(1 - d_i/r) \tag{2}$$

Here, i indexes the mother nodes in the parse, d_i is the distance between the i'th mother-below and the corresponding daughter-above, and r is a constant that is slightly bigger than the diameter of the entire set of points in the metric grammar space. Note that harmony is always non-positive but highly grammatical sentences approach a harmony value of 0.

Generation. Generation is accomplished similarly to the way it occurs in a PCFG, with clusters of points playing the role of rules with the same mother label (cf. Eisner, 1996). There is a labeled starting cluster (the S cluster in Figure 3). A point is picked at random from this cluster. Then the daughters of this point are considered in order. For each daughter, if the daughter is not a lexical daughter, a point is picked near the daughter by sampling among all the points in the metric space with nearby points more likely to be sampled (see 3). If the daughter is a lexical daughter, then the corresponding word is generated. This process is recursively applied until it ceases. The result is an ordered string of words.

$$p_i = \frac{e^{-\alpha \cdot d_i}}{\sum_j e^{-\alpha \cdot d_j}} \tag{3}$$

p_i is the probability of choosing point i,

d_i is the distance from the current locus to point i

α is a free parameter

Treelet coding. Now we turn to the question of how the points that form the treelets are located in the metric space. The encodings are initialized with a PCFG that models sentence distribution in a language of interest. The PCFG is used to generate a large corpus of parsed sentences. To align the account with the description of metric grammars, we refer to each rule as a Context Free Grammar (CFG) "treelet". When such a CFG treelet occurs in the corpus, it spans a sequence of lexical items, called its *lexical span*—e.g., the treelet $VP \rightarrow V\ NP$ in Figure 2 spans the sequence, *found-a-sort-of-crab*. For each treelet-plus-lexical-span (TpLS), we catalog the sequences of preceding and following words in every place where it occurs. This info is compiled into a branching probability tree, specifying the probability of each next word given the sequence of preceding words. In the general model, the branching probability trees are compressed into a fixed-width vector which functions as the encoding of the TpLS. Here, to keep the model simple, we only consider one word of preceding and following context for each TpLS. The encoding of a TpLS is the concatenation of the vectors of preceding and following word probabilities. For the grammar at hand, this method creates a unique encoding locus (meaning a small volume of nearby points) for each mother node in the PCFG. As noted above, one can think of this encoding as a simple-minded form of LLM encoding, where the dimensions are directly interpretable as next-word and preceding-word probabilities. Note that different TpLS's that belong to the same syntactic type under the CFG (e.g., $\langle$VP $\rightarrow$ V NP + found-a-sort-of-crab$\rangle$ and $\langle$VP $\rightarrow$ V NP + deplored-this-bear$\rangle$), tend to be nearby but not at the same point on account of the random sampling of the PCFG generation process. It is this variation that gives rise to the variance in each cluster illustrated in Figure 3.

Parsing. The model parses by using a trained neural network to invert the generation process (Borensztajn et al., 2009; Le & Zuidema, 2014). The network has an input layer with $2N$ units and N output units, where N is the dimensionality of the metric grammar space. The network is trained

to invert the map from mothers to daughters that was used in the generation process. Each branching treelet specifies a training pattern: the two daughter points are concatenated together to make an input vector of length $2N$; the mother point is the target. We used simple backpropagation to train the network. We achieved good results in the simulation below using a network with one hidden layer.

With the trained network at hand, parsing of any word-sequence proceeds as follows. All unlabeled binary branching trees over the word-sequence are considered. Each such tree is traversed in a bottom up fashion. First, each word is considered in the context of the surrounding words–this specifies a left + right probability vector with 100% probability on the observed surrounding words. A lexical point close to this vector is chosen for each word. Then, the first binary branching tree is considered. If two successive lexical items are the daughters of a single mother in this tree, then, to be parsed as a unit, they need a mother in the metric grammar space to which they can be joined. The two daughter point locations are fed into the neural network, and its output is taken to be their mother location. Then, distance-based-sampling as described above is used to find a daughter-above in the metric grammar that is proximal to this mother location. The process is repeated until mothers-below and daughters-above are found for each connection locus in the binary branching tree. In this fashion a metric grammar tree with the structure of the current binary branching tree is built. Once this has been done for all binary branching trees, the harmony of each parse is evaluated and the parses are sampled with probability proportional to $e^{harmony}$. The selected tree becomes the chosen metric grammar parse of the sentence.

Language evolution. With both generation and parsing defined, the system can be used to model language evolution. Language evolution is modeled as a discrete dynamical system, albeit one that makes very small changes in its state at each time step. The core assumption that permits the modeling of evolution is that every instance of language comprehension slightly modifies the metric grammar. How does this work? A metric grammar consists of a finite sample of treelets (in the simulations reported below, we used 1000 treelets). If the metric grammar is presented with a sentence, it parses it in the manner just defined. If the sentence has k words, then its parse has 2k-1 treelets (k lexical treelets, and k-1 branching treelets). Presented with such a sentence, the system randomly selects 2k - 1 of its treelets and deletes them. It then installs k-1 branching treelets of the new parse and it adjusts each lexical treelet of the new parse by slightly increasing the probability mass on the dimensions corresponding to the observed preceding and following words, decreasing the others to compensate. In this fashion, the metric grammar always maintains the same number of treelets but the structure of the system subtly shifts.

We noted in the introduction that dynamical systems can either evolve autonomously or be driven. We first consider the metric grammar system evolving autonomously. In this case, the metric grammar is used to generate a sentence using the random generation procedure just described, removing 2k - 1 treelets and adding 2k-1 treelets. This process is iterated many times. At each iteration, the neural network does a few trials of training on the new treelets. In this case, since the network is generating the very sentences that it is parsing, the network tends to find mothers for daughter sequences that are close to the locations of the original generating mothers (recall that it was originally trained on these very daughter-mother pairings). Therefore, the structure of the point distribution in the metric space does not change much; nor do the neural net's weights change much. Now consider the case where this is going on all the time, but additionally, forces originating in the extra-linguistic world cause some shift in the distribution of what is being said. In the case of content morpheme distribution changes (e.g., people used to talk a lot about "population growth"; now they talk a lot about "climate change"), the shift in distribution will not have a big effect on the higher grammatical structure of the metric grammar. This is because the grammatical co-occurrence privileges of different content morphemes belonging to the same class are largely similar. However, as noted, there are changes relevant to grammatical behavior that can trend statistically over time. This can be carried by social valuation—when there are alternative forms for expressing different ideas, social relationships often mediate the form choice (Labov, 1973). Additionally, such shifts can be driven by processing factors which favor one form over another one (Kroch, 1989; Hawkins, 2014). A novel prediction that metric grammars make is that such mere statistical adjustments can, in certain cases, result in alteration of the grammatical system, causing new grammatical behaviors to emerge. To explain how this works, we turn to our case study of English *sort/kind of*.

Case study: English *sort/kind of*

As we mentioned, Tabor (1995) presented evidence that, prior to the first appearance of the novel adverbial uses of *sort of* and *kind of*, there were statistical shifts in the use of these phrases that seemed to foreshadow the structural shift. In particular, the rate of use of these expressions before adjectives increased substantially in comparison to the rate of use of other Noun-*of* sequences over the period 1550-1850. Tabor (1995) identify the first clear adverb use in 1804. We initiated a metric grammar with a PCFG similar to Grammar 1, which models the Early Modern English distribution relevant to *sort/kind of*. Although we have not yet fully implemented an evolution simulation that brings about the innovative change, we have done a test of the model that indicates that it has the right causal properties to produce this result. We noted that, if the frequency of sentences of the form (1-b) shifts upward relative to their expected rate of occurrence under Grammar 1 in the input to the metric grammar, then some instances of lexical treelet *sort* will shift their encodings in the direction of more following *of*. Relatedly, subsequent instances of the word *of*, shift their distribution toward

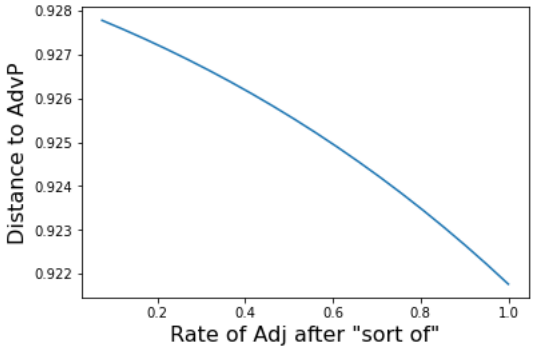

Figure 4: Distance of *sort of* mother from the Adverb locus in the metric grammar as a function of the degree to which the word sequence is correlated with a following Adjective.

more preceding *sort* and more following Adjective. This last change has the consequence that the distribution after *of* becomes more similar to the distribution after Adverbs, which are often followed by Adjectives. This suggests performing an experiment on the neural network to see how it responds to a potential combination of *sort* and *of* as a constituent (prior to the point in history, there had been no grammatical parses with *sort* and *of* joined into a single constituent). We tested the neural network of Metric Grammar 1 by giving it, as input, a series of points along a 1-dimensional manifold in the 2N dimensional encoding space of its input layer. This series points specified a sequence of inputs in which the frequency of "sort" + "of" before Adjective gradually increased as described above. This adjustment approximates a control parameter change, as described in *Bifurcations* above–we are interested in how this continuous change in frequency values may affect the structure of the grammar system. Figure 4 shows, on the x-axis, the degree of adjustment along the 1-dimensional manifold in frequency space, and it shows, on the y-axis, the distance between the output vector (Mother node location) and the average of the Adverb class.

What does this imply about interaction of the metric grammar with mere statistical changes in the distribution of existing forms? The model chooses parses in proportion to their harmony values. The Figure 4 result suggests that the control parameter manipulation increases the relative likelihood of picking an adverbial parse for *sort of* when presented with sentences structured like example (1-b) because it makes the harmony of the adverbial parse slightly higher than it was before. Producing this parse would establish a foothold in the treelet population for an adverb-like *sort of*. The treelet with the novel mother value discussed above would then become a member of the population of treelets and would thus be a part, albeit a rare part initially, of the grammatical system. This, in turn, implies that this treelet would have a chance of being selected not only with a following Adjective, but also, because it is near the Adverb Phrase locus, with a following verb. (In PCFG 1 and therefore also in the derived metric grammar, Adverbs tend to modify both Adjectives and Verb Phrases). If that becomes a significant tendency, we can say that *sort of* has become a member of the Adverb class. This does not mean it has ceased being a Noun-Preposition sequence because there were originally multiple "sort" and "of" lexical items, and only a proper subset of these have made the transition away from the Noun and Preposition loci. In this way, the model captures the A $\rightarrow$ A/B pattern that is typical of grammaticalization episodes (Hopper & Traugott, 1993).

Conclusions

We have described metric grammars, language generation and parsing systems that reside in connected metric spaces. We have suggested that these systems offer a way of modeling phenomena that point to a close relationship between formal language properties and the statistics of language use. In particular, we focused on cases in which grammatical category change is anticipated by frequency changes which appear to prepare the ground for a structural shift. Classical symbolic models (like PCFGs) fail to predict such correlations because their statistical properties are independent of their structural properties. Neural network models predict such correlations (Tabor, 1994) but their encoding systems are black boxes; metric grammars predict them in an analytically interpretable way: grammaticalization innovations come about through cluster disturbances prompted by external forces that modulate usage frequencies, and these snowball into category bifurcations.

The model constitutes an integration of linguistic theory, neural networks, and dynamical systems theory. It is well-known that recurrent neural networks, as well as gradient descent learning mechanisms are dynamical systems and are thus plausibly organized around complexly structured attractors. This observation seems promising because attractors are category-like, and thus seem naturally suited to treating cognitive phenomena which show a lot of evidence of complex categorical patterning. So far, however, this promise has not borne great fruit, possibly because we do not yet adequately understand the kinds of attractor structures that cognition relies. The current framework offers a way to explore the dynamics of attractors that capture complex syntactic patterns in natural languages, potentially enriching both dynamical systems theory and cognitive science.

Our model is still very nascent. One challenge is that we have only implemented a 1-word window radius for characterizing linguistic context. An important question is how to handle more distant correlations that are important for characterizing word and phrase classes. We have also not addressed pragmatics, widely acknowledged to be at play in grammaticalization developments (Traugott, 1988).

Recognizing the validity of these questions, we nevertheless suggest that metric grammars offer a potentially helpful fresh perspective on the challenging problem of relating language form to language use.

References

Borensztajn, G., Zuidema, W., & Bod, R. (2009). The hierarchical prediction network: towards a neural theory of grammar acquisition. In *Proc. cogsci.*

Bybee, J. (2006). *Frequency of use and the organization of language*. New York: Oxford University Press.

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

Chae, H.-R. (1997). Verbal nouns and light verbs in Korean. *Language Research*, *33*(4), 581–600. Retrieved 2024-05-11, from

Chomsky, N. (1986). *Knowledge of language : its nature, origin, and use*. New York: Praeger.

De Smet, H. (2012). The course of actualization. *Language (Baltimore)*, *88*(3), 601-633.

Detges, U. (2015). The Romance adverbs in *-mente*: a case study in grammaticalization. In *An International Handbook of the Languages of Europe, Berlin: Mouton De Gruyter.* Berlin: Mouton De Gruyter.

Eisner, J. M. (1996). Three new probabilistic models for dependency parsing: An exploration. In *COLING 1996 volume 1: The 16th international conference on computational linguistics.*

Enfield, N. (2008). Transmission biases in linguistic epidemiology. *Journal of Language Contact*, *2*(1), 299–310.

Gibson, E., & Wexler, K. (1994). Triggers. *Linguistic inquiry*, *25*(3), 407-454.

Hawkins, J. A. (2014). *Cross-linguistic variation and efficiency*. Oxford: Oxford University Press.

Hopper, P. J., & Traugott, E. C. (1993). *Grammaticalization, 2nd ed*. Cambridge, England: CUP.

Inkelas, S., & Zec, D. (Eds.). (1990). *The Phonology-Syntax Connection*. Chicago, IL: University of Chicago Press.

Jackendoff, R. (1991, December). Parts and boundaries. *Cognition*, *41*(1), 9–45. doi: 10.1016/0010-0277(91)90031-X

Kroch, A. S. (1989). Function and grammar in the history of English: Periphrastic *do*. In R. W. Fasold & D. Schiffrin (Eds.), *Language change and variation* (p. 134-169). Philadelphia: John Benjamins.

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

Le, P., & Zuidema, W. (2014, October). The Inside-Outside Recursive Neural Network model for Dependency Parsing. In A. Moschitti, B. Pang, & W. Daelemans (Eds.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 729–739). Doha, Qatar: Association for Computational Linguistics. doi: 10.3115/v1/D14-1081

Li, R., Yu, Q., Huang, S., Shen, L., Wei, C., & Sun, X. (2021). Phrase embedding learning from internal and external information based on autoencoder. *Information processing management*, *58*(1), 102422-.

Lord, C. (1973). Serial verbs in transition. *Studies in African Linguistics*, *4*(3), 269-296.

Lord, C. (1976). Evidence for syntactic reanalysis: from verb to complementizer in Kwa. In *Papers from the parasession on diachronic syntax, chicago linguistic society* (pp. 179–91). University of Chicago Press.

Nikzad-Khasmakhi, N., Feizi-Derakhshi, M.-R., Asgari-Chenaghlu, M., Balafar, M.-A., Feizi-Derakhshi, A.-R., Rahkar-Farshi, T., ... Ranjbar-Khadivi, M. (2021). *Phraseformer: Multimodal key-phrase extraction using transformer and graph embedding.*

Park, S., & Kim, H. M. (2022). Phrase embedding and clustering for sub-feature extraction from online data. *Journal of mechanical design (1990)*, *144*(5).

Schmid, H.-J. (2020). *The dynamics of the linguistic system: Usage, conventionalization, and entrenchment* (First ed.). Oxford: Oxford University Press.

Tabor, W. (1994). *Syntactic innovation: A connectionist model.* (Ph.D. dissertation, Stanford University. Available at: http://www.sp.uconn.edu/p̃s300vc/Papers/)

Tabor, W. (1995). Lexical change as nonlinear interpolation. In J. D. Moore & J. F. Lehman (Eds.), *Proceedings of the 17th annual cognitive science conference.* Lawrence Erlbaum Associates.

Traugott, E. C. (1988, October). Pragmatic strengthening and grammaticalization. *Annual Meeting of the Berkeley Linguistics Society*, 406–416. doi: 10.3765/bls.v14i0.1784

Wang, S., Thompson, L., & Iyyer, M. (2021). Phrase-bert: Improved phrase embeddings from BERT with an application to corpus exploration. *CoRR*, *abs/2109.06304*.

Wu, Y., Pan, X., Li, J., Dou, S., Dong, J., & Wei, D. (2024). Knowledge graph-based hierarchical text semantic representation. *International journal of intelligent systems*, *2024*, 1-14.

Wu, Y., Zhao, S., & Li, W. (2020). Phrase2vec: Phrase embedding based on parsing. *Information sciences*, *517*, 100-127.