COGNITIVE SCIENCE Vol 23 (4) 1999, pp. 491–515 ISSN 0364-0213

Dynamical Models of Sentence Processing

WHITNEY TABOR
University of Connecticut

MICHAEL K. TANENHAUS
University of Rochester

We suggest that the theory of dynamical systems provides a revealing general framework for modeling the representations and mechanism underlying syntactic processing. We show how a particular dynamical model, the Visitation Set Gravitation model of Tabor, Juliano, and Tanenhaus (1997), develops syntactic representations and models a set of contingent frequency effects in parsing that are problematic for other models. We also present new simulations showing how the model accounts for semantic effects in parsing, and propose a new account of the distinction between syntactic and semantic incongruity. The results show how symbolic structures useful in parsing arise as emergent properties of connectionist dynamical systems.

I. INTRODUCTION

The Dynamics of Sentence Processing

Linguistic input is typically consistent with multiple syntactic possibilities as it unfolds over time. Because syntax strongly constrains interpretation, the processing system must determine the set of possible syntactic hypotheses, maintain some or all in memory, and update them as new input arrives.

Behavioral evidence from sentences with temporary syntactic ambiguities has clearly established that readers and listeners have strong preferences for some structures over others. When subsequent input becomes inconsistent with the preferred structure, processing difficulty ensues. Patterns of processing difficulty for these "garden-path" sentences taken primarily from reading experiments provide an empirical benchmark for evaluating theories of syntactic processing (see Tanenhaus & Trueswell, 1995).

Direct all correspondence to: Whitney Tabor, Department of Psychology, University of Connecticut, Storrs, CT 06269; E-Mail: tabor@uconnvm.uconn.edu.

Within traditional symbolic systems, syntactic hypotheses are computed by a parser—a set of procedures that maps the input onto partial syntactic structures using categories such as Noun and Noun Phrase utilizing a knowledge base defined by a grammar. Explanations for structural preferences are typically couched in terms of the complexity of structure building operations, and/or memory demands in an initial stage of structure building (see Frazier & Clifton, 1996; Gibson, 1998, for recent reviews). A second set of procedures guides recovery from misanalysis when the initial structure is disconfirmed.

However, recent evidence indicates that syntactic processing is simultaneously affected by multiple sources of constraints, including semantic and discourse-based information (for reviews see MacDonald, Pearlmutter, & Seidenberg, 1994; Tanenhaus & Trueswell, 1995). Moreover, reading time is correlated with graded properties of the linguistic input not easily reduced to purely structural factors, such as the relative frequency with which lexical forms occur in different environments (MacDonald et al., 1994; Trueswell, 1996). Models in which multiple sources of constraint provide weighted evidence for competing syntactic analyses provide a natural account of these phenomena (Cottrell, 1985; Cottrell & Small, 1984; MacDonald et al., 1994; Spivey-Knowlton, 1996; St. John & McClelland, 1990; Waltz & Pollack, 1985). Processing difficulty occurs when input that is inconsistent with the previously biased alternative is encountered.

Connectionist models are a variety of constraint-based models in which learning plays a central explanatory role. In connectionist models of parsing, some of the systematic properties of language which motivate the positing of specialized structures in a symbolic paradigm are hypothesized to arise as "emergent properties" under connectionist learning. Crucially, the emergent counterparts of symbolic structures may differ from them in important ways.

However, much previous connectionist modeling of syntactic structures has been inexplicit about what these emergent structures are and how, exactly, they differ from their symbolic counterparts. Following Tabor, Juliano, and Tanenhaus (1997), we argue that a connectionist, learning-based system can be explicit about "emergent properties" by using the constructs of *dynamical systems theory*, a theory commonly applied to dynamically changing natural systems such as swinging pendulums, orbiting planets and circulating fluids. Constructs useful in analyzing such systems include trajectories, fixed points (or stable states), attractors, basins, and saddlepoints (see Abraham & Shaw, 1984, and Strogatz, 1994 for introductions). The model we describe here has two components: a network similar to a Simple Recurrent Network or "SRN", (Elman, 1990, 1991) and a dynamical gravitation module.

The input to the network component is a sequence of words generated by a probabilistic finite state or context free grammar. As it learns to process the input, the model forms representations of parse states in its hidden unit space, placing words that are likely to be followed by similar constructions nearby one another (Christiansen, 1994; Elman, 1990, 1991; Tabor, 1994). Thus, the learning process strongly influences the final performance of the system (Christiansen, 1994; Christiansen & Chater, 1999a; MacDonald & Christiansen, 1998).

The gravitation module is a dynamical processor which transforms the representations produced by the SRN into unique parse hypotheses, requiring varying amounts of time to

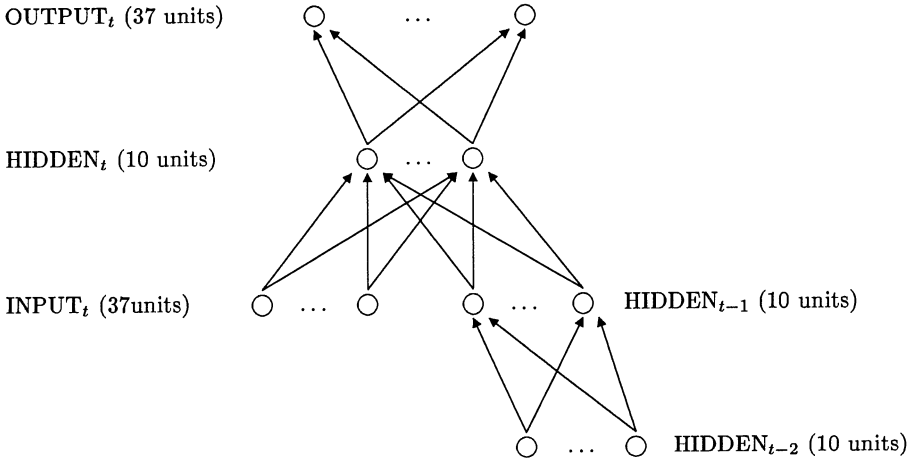

Figure 1. Three layer network with recurrent connections in the hidden layer (implemented as partial unfolding across time).

do so. The model's processing time is taken as an analog of human reading time. The dynamical component operates on the set of states visited by the network when it is processing a large random sample of text. It uses a gravitational mechanism to group these into distinct classes, thus providing useful structural information about the SRN's representation. Thus, we refer to the model as the *Visitation Set Gravitation* (or VSG) model.

II. THE VSG MODEL AND RELATED DYNAMICAL MODELS

The VSG Model

The network and training procedure were closely modeled on Elman (1991). Each word in a corpus was assigned a unique localist vector (one element was equal to 1 and all the others were equal to 0). Vectors were presented on the input layer of the network in the order in which the corresponding words occurred in the training corpus. The network was trained using the backpropagation algorithm (e.g., Rumelhart, Hinton, & Williams, 1986) to predict on the output layer which word was coming next for each input. We used a three-layer feed-forward network with a "context layer" feeding into the hidden layer at each timestep (Elman, 1991). This network has the same relaxation dynamics as a three layer network with complete interconnection among its hidden units (on the assumption that each unit is updated exactly once each time a word is presented, with the input and context units updated first, then the hidden units, and then the output units). Thus, the training procedure is an approximation of the Backpropagation Through Time (BPTT) algorithm (Rumelhart et al., 1986).) Error propagation was carried through two hidden-layer time steps while adjusting input-to-hidden weights only on the basis of the current time step (see Figure 1). The network had 37 input units, 10 hidden units, and 37 output units. The hidden units at all time steps had fixed sigmoid activation functions ($y_i =$

$1/(1 + e^{-net_i})$ where net_i is the net input to unit i). The output units as a group had normalized exponential (or softmax) activation functions ($y_i = e^{net_i}/\sum_{j \in Outputs} e^{net_j}$). The output error for input p was thus defined for "1-of-n" classification by Equation (1).

$$E_p = log \prod_{j \in Outputs} y_j^{t_j} \tag{1}$$

where y_j is the activation of unit j for input p, and t_j is the target for unit j on that input, (Rumelhart, Durbin, Golden, & Chauvin, 1995) and backpropagated through the unfolded network.[1] Weights were adjusted after every input presentation. The network was trained on the output of a simple grammar approximating relevant features of the phenomenon to be modeled. The details of training are described in Section 3.

The gravitation module of the VSG operates on the hidden layer representations of the trained SRN. The SRN places words in contexts with similar distributional characteristics near one another in the hidden unit space. Thus, if we sample the hidden unit activations of the trained network over a wide range of constructions from the training language, we may find a set of clusters of points, where points in the same cluster correspond to grammatically equivalent states of the generating language. The gravitation module is a clustering mechanism which finds such equivalence classes of states. The trained network was presented with a large random sample of sentences generated by the grammar. All the hidden unit states visited during processing were recorded. Each point was treated as a massive body in the 10-dimensional hidden unit space. Each such massive body had a mass of one unit and was fixed in its position. The processing of a particular word-in-context was tested by treating its hidden unit location as a test mass (also of unit magnitude) which was free to move under the gravitational influence of all the fixed masses. Typically, the test mass was near the center of mass of some dense cluster and would gravitate into that cluster. We modeled processing time as the time required to gravitate into the cluster. The fixed masses can be thought of as representing typical previous experiences with the language. Thus, the gravitational mechanism implements the idea that, in responding to a new instance of a word-in-context, the processor analogizes that word-in-context to its previous experiences and gravitates to a cluster corresponding to the most-similar previous experience. The points or small regions in the centers of the clusters where the system is stable are called *attractors*. The set of all starting points from which the system gravitates into a particular attractor is called its *basin*. Under an appropriate parameterization of the gravitational system, the system's basin structure defines a partition of the set of words-in-context into equivalence classes. In the cases we analyzed, these classes corresponded to states of the grammar (Hopcroft & Ullman, 1979) that generated the training data.

The change in position of the test mass was defined by Equation (2).

$$\frac{\Delta x}{\Delta t} = \nu \sum_{i=1}^{N} \frac{\vec{x}_i - \vec{x}}{r_i^p} \tag{2}$$

where $\vec{x}$ is the position of the test mass, N indexes the fixed masses, $\vec{x}_i$ is the position of the i'th fixed mass, r_i is the Euclidean distance between $\vec{x}_i$ and $\vec{x}$ at time t, and p is a gravitational strength parameter which determines the pulling power of each test mass. This equation is an approximation of Newton's Law of Universal Gravitation when (i) the test mass starts with zero velocity at infinite distance from the starting point, (ii) $p = 2$, and (iii) ν is the Universal Gravitation Constant.

Equation 2 implies that every point in the visitation set is a singular point (i.e., a point where the velocity goes to infinity). To avoid infinite velocities, which make the structure of the system hard to detect, we introduced a threshold r_{min} and set $r_i = r_{min}$ whenever r_i became smaller than r_{min}. This made the trajectories less prone to wild jumps. The parameters, N, ν, r_{min}, Δt, and p are all free parameters of the model. The first four of these are primarily relevant to making the performance of the model easy to interpret.[2] The last one, (p, or gravitational strength), is undesirably unconstrained—we set it to a value that made the attractor basins correspond to distinct parse states as defined by the training grammar. Fortunately, this parameter narrows the range of possible basin structures to a relatively small set. Because the parameter is closely tied to the constraints on learning there may be a way to bind it less stipulatively. Under these assumptions, the test mass typically sped up as it approached a fixed point near the center of mass of a cluster, overshot the fixed point (because it is unlikely that the mass will land exactly on a fixed point for positive Δt), and then headed back toward the center of mass for another "fly-by". Our algorithm for determining gravitation times thus computed the number of steps it took the test mass to reverse direction for the first time (where a direction reversal is a turn of more than 90° in one step). Note that the gravitation module operates completely independently of the recurrent network: the outcome of the relaxation dynamics does not affect the network's processing of the subsequent word.

In sum, the VSG model is trained like an SRN. It generates predictions of reading times as follows:

1. Feed a sentence one word at a time to the trained network using SRN relaxation dynamics.
2. For each word of the sentence, use the gravitation module to determine a gravitation time.
3. Compare gravitation time profiles (e.g., across words in a sentence) to reading time profiles.

Previous Related Models

In order to motivate the VSG model, we will briefly review previous related models. Most connectionist models are standard dynamical systems: their operation can be described by a differential equation for which the state change is a continuous function of the parameters (or weights) of the network. One can distinguish two important dynamical regimes within the connectionist framework: learning dynamics, and relaxation dynamics. Learning dynamics involve slow adjustment of connection weights to find the minimum

of a cost function. Relaxation dynamics involve rapid adjustment of activation values in order to compute an output. In the VSG model, relaxation dynamics (albeit in a non-connectionist system) reveal structural properties of the learning dynamical system for one type of network (the SRN).

The first connectionist processing models (e.g., Cottrell, 1985; Cottrell & Small, 1984; Waltz & Pollack, 1985) focused on the relaxation dynamics of hand-designed models using localist representations. These models had many of the properties we make use of here. For example: 1) Competition between simultaneously valid parses increased processing time. 2) Magnitudes of real-valued weights reflected contrasts in frequency and thus gave rise to biases in favor of more frequently-encountered interpretations (e.g., Cottrell & Small, 1984). 3) Sometimes, "spurious" attractive states arose which corresponded to no interpretation (e.g., Cottrell & Small, 1984). In the Emergence of a Syntax/Semantics Distinction Section we show that certain "spurious states" may provide a plausible model of parsing of an ungrammatical string (cf. Plaut, McClelland, Seidenberg, & Patterson, 1996). 4) Syntactic and semantic information simultaneously constrained parsing (e.g., Cottrell & Small, 1983).

The development of the backpropagation algorithm (Rumelhart et al., 1986) led to a new class of learning-based connectionist parsing models. Currently, the most successful models are Elman's SRN (Elman, 1990, 1991) and its variants (e.g., St. John & McClelland, 1990—see Christiansen & Chater, this issue, for a review). Elman's model can approximate the word-to-word transition likelihoods associated with a simple text corpus, thus embodying information relevant to the syntax and semantics of the language of the corpus to the degree that these are reflected in distributional properties.

While the learning dynamics of Elman's model are complex and interesting, the relaxation dynamics are uniform and uninformative. Since each node is updated exactly once after a word is presented, the network's processing time is identical from word to word and cannot plausibly be interpreted as a model of human processing time. Several researchers (Christiansen & Chater, 1999a; MacDonald & Christiansen, in press) have shown that a well-chosen definition of SRN output error can be mapped onto processing times. A desirable next step is to model word-to-word processing explicitly in the relaxation dynamics. Such explicitness is one goal of the VSG approach.

Moreover, as in many connectionist simulations, the principles governing the Elman model's specific predictions are not usually easy to surmise: the trained network's model of its environment is a complexly shaped manifold in a high-dimensional space. Although 1-dimensional quantities such as error measures and cost functions can give insight into local properties of this manifold, they do not tell us much about its structure. A useful addition would be some summarizing category information, indicating which pieces of the manifold are important, and what role they play in organizing the linguistic task. Thus, a second aim of the VSG approach is to use dynamical systems theory to reveal this summarizing category information by approximating certain basins, attractors, saddle-points etc. which are implicit in the SRN's learning dynamics. For example, as we noted above, the attractors of the VSG model map onto distinct parse states of the language learned by the network (see The Gravitation Mechanism Section).

Although, the VSG model is inelegant in that it is a hybrid of two distinct dynamical systems, we view it is a useful stepping stone to a more mathematically streamlined and more neurally plausible model. In particular, the dynamics of the gravitation module are roughly paralleled by the dynamics of recurrent connectionist networks which settle to fixed points after each word presentation. In current work, we are exploring the use of the recurrent backpropagation (RBP) algorithm (Almeida, 1987; Pineda, 1995) to train such networks on sentence processing tasks. In these models, the learning process drives the formation of attractor basins, so the free parameter p is eliminated and the categorization system stems from independently motivated constraints such as the number of hidden units and the nature of the activation function. However, the task of learning complex syntax in an RBP network is harder. Thus, an advantage of the VSG model is that it permits us to use the currently more syntactically capable SRN to explore the effectiveness of dynamical constructs. If the predictions are borne out, then the motivation for solving the learning challenges facing RBP becomes greater.

Although the attractor basins defined by the VSG model are primarily valuable for the insight they provide into the representations learned by an SRN, they also have an independent functional motivation. Interpreting language probably requires making some discrete choices. For example, Waltz and Pollack (1985) note that although we can comprehend the multiple meanings of wholly ambiguous sentences (e.g., *Trust shrinks*; *Respect remains*; *Exercise smarts*—p. 52) we seem to flip-flop between them rather than comprehend them as a single composite. Moreover, it is clearly important to be able to conclude that in a sentence like *Jack believed Josh was lying*, *Josh* is not an object of the matrix clause but a subject of the embedded clause, even though processing evidence suggests that we temporarily entertain the former hypothesis. It is not obvious how to map the real-valued states of an SRN onto representations that could support such discrete choices. The VSG model provides a principled solution to this problem.[3]

Previous VSG Results

Focusing on the interaction between lexical and syntactic category ambiguity, Tabor, Juliano, and Tanenhaus (1997) showed that the VSG model predicts word-by-word reading times for phenomena that are challenging for other models, For example, lexical category ambiguities involving the word “that”, exhibit an interesting mix of contingent frequency effects (Juliano & Tanenhaus, 1993; Tabor et al., 1997). The sentences in (1) illustrate that “that” can be either a determiner (a and c) or a complementizer (b and d). The number of the noun disambiguates “that” as either a determiner (singular) or a complementizer (plural).

(1) a. That marmot whistles.
    b. That marmots whistle is surprising.
    c. A girl thinks that marmot whistles.
    d. A girl thinks that marmots whistle.

Reading times for these sentences are predicted by the hypothesis that readers slow down when they encounter words that violate their expectations about typical usage, as

determined from a corpus analysis. In particular, "that" is more frequent as a determiner than as a complementizer sentence-initially, but it is more frequent as a complementizer than as a determiner post-verbally. Thus, (1a) is easier than (1b), while (1c) is harder than (1d) at the words following "that" (see Tabor et al., 1997, for details). The VSG model predicts these effects because the denser visitation clusters associated with more-frequent continuations give rise to stronger gravitational pull and hence more rapid gravitation. However, the correlation between expectancy and reading time is also skewed by category structure. For example, after strictly transitive verbs like *visited*, *that* and a following adjective (2a) was read more slowly than *those* and a following adjective (2b) even though these two determiners occur equally frequently after transitive verbs.

(2) a. The writer visited that old cemetery.
b. The writer visited those old cemeteries.

Attractor competition also predicts these results. *That* following a transitive verb bears a distributional resemblance to *that* following a sentence-complement verb. Therefore, the position assigned by the recurrent net to *that* following a transitive verb is intermediate in the gravitation field between the attractor for *that* following a sentence-complement verb and the attractor for unambiguous determiners following transitive verbs. By contrast, since *those* is not ambiguous, *those* following a transitive verb starts very close to the appropriate attractor. Since *that* starts farther away from the attractor and its gravitation is slowed by the presence of a nearby attractor, it is processed more slowly than *those* (Tabor et al., 1997).

These cases illustrate two advantageous properties of the VSG model: 1) it is consistent with the pervasive evidence showing that reading time is inversely correlated with class frequency; and 2) it diverges appropriately from the frequency-based predictions in cases where class similarity effects distort these. The VSG model predicts the latter, *smoothing* effects by letting similarities between categories distort the internal structure of the attractor basins associated with the categories. While, it is possible that a similar prediction can be made by a model that computes expectations based on a probabilistic grammar (e.g., Jurafsky, 1996), some kind of as-yet-unspecified statistical smoothing (Charniak, 1993) across grammatical classes is required. It is also possible that a model which treats reading time as a kind of output error in an SRN (e.g., MacDonald & Christiansen, in press) would predict divergences from frequency-based predictions due to class similarity since position contrasts in the hidden unit space tend to map to position contrasts in the output space. But, because one-dimensional measures do not encode information about direction of displacement, it would be difficult to tell whether similarity is indeed the source of the error.

III. CASE STUDY: THEMATIC EXPECTATION

The Tabor et al. (1997) simulations focused primarily on syntactic contrasts in that the complement requirements of the verbs and the agreement requirements of the determiners were categorical. The simulations presented here investigated the less categorical cases

that arise in association with "semantic" distinctions. Thematic role assignment is a clear example. Almost any noun can fill any role, but if a noun that is unsuitable for a given role is forced to play that role, the result is a "semantically strange" or "incongruous" sentence (3).

(3) a. # The waitress was served by the jukebox.
b. # The car accused the pedestrian of cheating.

Linguistic theories generally posit distinct mechanisms for explaining semantic and syntactic incongruity. Semantic incongruity is detected using world knowledge, whereas syntactic incongruity results from violating rules of grammar. Psychophysiological data from studies using event-related potentials suggesting that these two kinds of violations may result in qualitatively different patterns of brain responses (Ainsworth-Darnell, Shulman, & Boland, 1998; Garnsey, 1993; Hagoort, Brown, & Groothusen, 1993; Osterhout & Holcomb, 1993).

The distinction between syntactic and semantic incongruity is especially interesting from the perspective of connectionist models. Since both semantic and syntactic constraints affect the distributional structure of words, a connectionist device trained on distributional information might be able to model both classes of constraints. Weckerley and Elman (1992) showed plausible "semantic" influences on the processing of center-embedded constructions in such a model, although they did not analyze the mechanism. Here we analyze the VSG model's representation, showing that it exhibits what might best be called a *graded qualitative distinction* between semantic and syntactic incongruity.

The claim that "semantic" information can be learned by a model which only interacts with corpus data needs to be qualified. Clearly, a model without an extra-linguistic world cannot simulate the relationship between language and the extra-linguistic world, and thus cannot be a full *semantic model*. However, corpora contain a good deal of information beyond what the syntax of a language provides. Indeed, Burgess & Lund (1997) and Landauer & Dumais (1997) among others have shown that information which is standardly termed "semantic" can be extracted from a corpus by evaluating co-occurrence statistics. Much of this information is about which words *tend* to be used in combination with which other words. The usual strategy in linguistic modeling is to assume that knowledge about language does not incorporate knowledge of the world that can be learned independently of language. But this may be misguided: since information about tendencies of usage is available in the input, the "language mechanism" may be shaped by this usage as well as by abstract grammatical constraints.

We examined the role of thematic fit in syntactic ambiguity resolution, focusing on the results of McRae, Spivey-Knowlton, & Tanenhaus (1998).

The Phenomenon

McRae et al. used sentences such as: (4a) and (4b).

(4) a. The cop / arrested by / the detective / was guilty / of taking / bribes.
b. The crook / arrested by / the detective / was guilty / of taking / bribes.

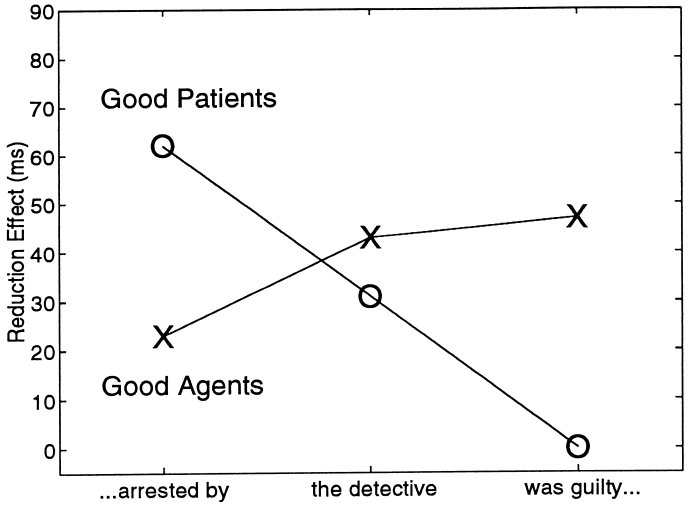

Figure 2. Crossed and smoothed latencies in the main clause/reduced relative ambiguity (after McRae et al. (in press)). The "X" sentences began with "Good Agents"; the "O" sentences began with "Good Patients".

Similarity ratings, (e.g., How common is it for a cop to arrest someone?) were used to group "cop" and similarly rated nouns together as "Good Agents" and "crook," and similarly rated nouns together as "Good Patients". Good Agents provide initial support for the main clause hypothesis at the first verb (e.g., "arrested"), whereas Good Patients provide support for a reduced relative hypothesis. Figure 2 summarizes the results from a self paced reading study using sentences like those in the example. The "reduction effect" is the difference between the reading time of sentences like (4) and the corresponding unreduced (unambiguous) cases in which "who was" was inserted before the first verb. The regions were the pairs of words between slashes.

Three properties of the data are worth highlighting. 1) There is an immediate effect of thematic bias: in the verb+"by" region, the Good Patients give rise to higher reading times than the Good Agents. 2) Reading times are longer where there is a conflict between the biases of the preceding context and the biases of the current word, e.g., at the (agentive) verb after a Good Patient subject, and at the NP following a Good Agent subject and verb. 3) Reading times show an "inertia" effect. Even when the linguistic input provides information that could, in principle be used to strongly reject a previously entertained parse, (e.g., the word "by" after the verb), the processor seems to shift only gradually over to the new hypothesis.

McRae et al. showed that the reading time profiles can be plausibly interpreted as stemming from competition between two alternative syntactic hypotheses: 1) the first verb (e.g., "arrested") is the main verb of the sentence, or 2) it is a verb in a reduced relative clause. For Good Patients, there is competition between these two hypotheses beginning at the first verb, which resolves quickly when supporting evidence for the reduced relatives comes from the "by"-phrase. For the Good Agents there is a strong initial bias

for the main clause, with competition beginning when disconfirming information is encountered in the "by-phrase".

McRae et al. formalized the competition account by using Spivey-Knowlton's (1996) Normalized Recurrence algorithm, in which multiple constraints provided support for two competing structures: main clause and reduced relative. The strength of the constraints was determined by norms and corpus analysis. The weights were free parameters set to model fragment completion data using the same materials. The same weights were then used to successfully predict on-line reading times. In the simulation described next we show that competition, lexical sensitivity, and inertia emerge in a VSG model trained on corpus data resembling the significant distributional properties of the McRae et al. materials.

Simulation

The Training Grammar. The simulation grammar shown in Table 1 generates a simple, symmetrical set of strings which share a number of properties with the English sentences of interest. The labels in the grammar and in the following discussion (e.g., "Good Agt", "Good Pat") make this analogy explicit for the convenience of the reader. However, the analogy is rough, and the model is not intended to map precisely onto human behavior. Instead, we view the study of formally simple grammars like this as a useful tool for gaining insight into the behavior of a connectionist network that shares a number of interesting properties with humans.

The first two words of each sentence in the grammar can be classified as belonging to one of two classes, X and Y, which give rise to different expectations about which words are likely to occur next. X and Y correspond to "Good Agent" and "Good Patient" respectively. Sentences starting with X's outnumber sentences starting with Y's by a ratio of 2:1, reflecting the greater frequency of agentive constructions in English. Also, as in English, there are initial X's and initial Y's of a range of different frequencies. The second word is of the type labeled V. It corresponds conceptually to the English verbs in McRae et al.'s study in the following way: both X's and Y's are followed by the same set of Vs, but, depending on which first word and V occurred, there is a bias as to how the sentence will end. Sentences that begin with X words and are followed by V words with letter labels alphabetically close to "a" tend to end with the most common members of the X2 and X3 categories (ignoring, for a moment, the words with "1" in their labels). Sentences that begin with Y and are followed by V's with letter labels alphabetically close to "f" tend to end with the most common members of the Y2 and Y3 categories. In fact, the members of the categories X2 and Y2 are the same, as are the members of the categories X3 and Y3, but if the generating category is X2 or X3, then there is a bias toward words with labels alphabetically close to "a", and if the generating category is Y2 or Y3, there is a bias toward words with labels alphabetically close to "f". The word "p" is an end-of-sentence marker, or "period".

The non-absolute biases of many of the words in this grammar mirror the fact that in natural language, many words can be constituents of many constructions and thus do not

TABLE 1
Training Grammar for the Thematic Bias Simulation

0.67 S → X VX VPX p ("MC")	
0.33 S → Y VY VPY p ("RR")	
0.67 X → xa ("Good Agt")	0.02 Y → ya ("Good Pat")
0.17 X → xb ("Good Agt")	0.03 Y → yb ("Good Pat")
0.07 X → xc ("Good Agt")	0.04 Y → yc ("Good Pat")
0.04 X → xd ("Good Agt")	0.07 Y → yd ("Good Pat")
0.03 X → xe ("Good Agt")	0.17 Y → ye ("Good Pat")
0.02 X → xf ("Good Agt")	0.67 Y → yf ("Good Pat")
0.67 VX → xa ("MC Bias Verb")	0.02 VY → va ("RR Bias Verb")
0.17 VX → vb ("MC Bias Verb")	0.03 VY → vb ("RR Bias Verb")
0.07 VX → vc ("MC Bias Verb")	0.04 VY → vc ("RR Bias Verb")
0.04 VX → vd ("MC Bias Verb")	0.07 VY → vd ("RR Bias Verb")
0.03 VX → ve ("MC Bias Verb")	0.17 VY → ve ("RR Bias Verb")
0.02 VX → vf ("MC Bias Verb")	0.67 VY → vf ("RR Bias Verb")
0.67 VPX → 1a X2 X3 ("MC")	0.02 VPY → 1a X2 X3 ("MC")
0.17 VPX → 1b X2 X3 ("MC")	0.03 VPY → 1b X2 X3 ("MC")
0.07 VPX → 1c X2 X3 ("MC")	0.04 VPY → 1c X2 X3 ("MC")
0.04 VPX → 1d Y2 Y3 ("RR")	0.07 VPY → 1d Y2 Y3 ("RR")
0.03 VPX → 1e Y2 Y3 ("RR")	0.17 VPY → 1e Y2 Y3 ("RR")
0.02 VPX → 1f Y2 Y3 ("RR")	0.67 VPY → 1f Y2 Y3 ("RR")
0.67 X2 → 2a ("MC")	0.02 Y2 → 2a ("RR")
0.17 X2 → 2b ("MC")	0.03 Y2 → 2b ("RR")
0.07 X2 → 2c ("MC")	0.04 Y2 → 2c ("RR")
0.04 X2 → 2d ("MC")	0.07 Y2 → 2d ("RR")
0.03 X2 → 2e ("MC")	0.17 Y2 → 2e ("RR")
0.02 X2 → 2f ("MC")	0.67 Y2 → 2f ("RR")
0.67 X3 → 3a ("MC")	0.02 Y3 → 3a ("RR")
0.17 X3 → 3b ("MC")	0.03 Y3 → 3b ("RR")
0.07 X3 → 3c ("MC")	0.04 Y3 → 3c ("RR")
0.04 X3 → 3d ("MC")	0.07 Y3 → 3d ("RR")
0.03 X3 → 3e ("MC")	0.17 Y3 → 3e ("RR")
0.02 X3 → 3f ("MC")	0.67 Y3 → 3f ("RR")

Note. MC, Main Clause; RR, Reduced Relative. The quoted labels specify the analogy with English.

provide a categorical signal, independent of their context, as to which parse hypothesis is correct; but many of these same words have statistical tendencies which can be used to compute a bias toward one construction or another in the absence of a fully constraining context. In the model, the local ambiguity of the words is essential for predicting inertia effects: it forces the network to use its context representation to compute expectations. As a result, the network tends to retain the parse bias it had at earlier stages, only relinquishing it gradually.

There are, however, some words in natural languages, the "closed class" or "function" words that provide fairly unambiguous cues as to which parse hypothesis is correct. The word "by" is one such word in the McRae et al. materials. Here, the members of the 1 category provide this kind of categorical constraining information. "1a" through "1c" are only compatible with an X2 X3 ending, while "1d" through "1f" are only compatible with a Y2 Y3 ending. Note that both X and Y initial words can be followed by both kinds of

endings, but there is a bias for X initial words to be followed by X2 X3 endings and for Y initial words to be followed by Y2 Y3 endings.

Training the Network. The grammar was used to generate data for training the network described in the VSG Model Section. Before training began, the weights and biases of the network were assigned uniformly distributed random values in the interval $[-0.5, 0.5]$. The network's learning rate was set at 0.05 and momentum was not used. The grammar defines ten states (states are distinct if they induce different distributions over the set of all possible future sequences—Crutchfield, 1994; cf. Hopcroft & Ullman, 1979). The network was trained until it was distinguishing and reasonably approximating the transition likelihoods of all ten states. The grammar sanctions $12 \times 6^4 = 15552$ grammatical strings. Each juncture-between-words in a string is associated with a probability distribution over next-words which can be computed from the grammar. We compared the network's output for each juncture to the grammar-generated distribution for that juncture and asked if the distance between these two distributions was less than one half the minimum distance between any two grammar-determined distributions.[4] We stopped training when a hand-picked sample of such comparisons yielded positive outcomes, and then evaluated this comparison for the whole language to find that the comparison yielded a positive outcome for 94% of the $15552 \times 6 = 93312$ junctures between words. At this point, the network had been trained on 50,000 word presentations. We re-initialized the weights and retrained the network five times for the same number of word presentations. We determined by inspection that the visitation set had nearly identical (10-cluster) structure in three out of the six cases, and similar structure in all cases. The results reported below are based on the first case.

The Gravitation Mechanism. We set the gravitation module parameters to $n = 2000$, $r_{min} = 0.01$, $\mu = 0.0002$, and $p = 2.7$. With these settings, the dynamical processor had an attractor corresponding to each state associated with the training grammar. There were two attractors associated with initial words, V words, 1 words, and 2 words. The attractors corresponded to the X ("Main Clause") reading and the Y ("Reduced Relative") readings, respectively, in that sentences with a high likelihood of finishing with letter labels alphabetically near "a" were consistently drawn into the X attractor and those with a high likelihood of finishing with labels near "f" were consistently drawn into the Y attractor. There was one attractor for the 3 position and one for the end-of-sentence marker, "p".

Reading Time Results

We compared the reading times on a Y ("Reduced Relative") continuation for sentences beginning respectively, with X ("Main Clause bias") words and Y ("Reduced Relative bias") words. Because our grammar did not include the option of disambiguating the V ("Verb") word syntactically, prior to its occurrence (as in English *The cop who was arrested . . .*) we were not able to use such disambiguated cases as a baseline. A sample result is shown in Figure 3. The "Good Patients" curve shows word-by-word mean gravitation times for eight sentences like "yd vc 1d 2d 3d p" ("Crook arrested by detective

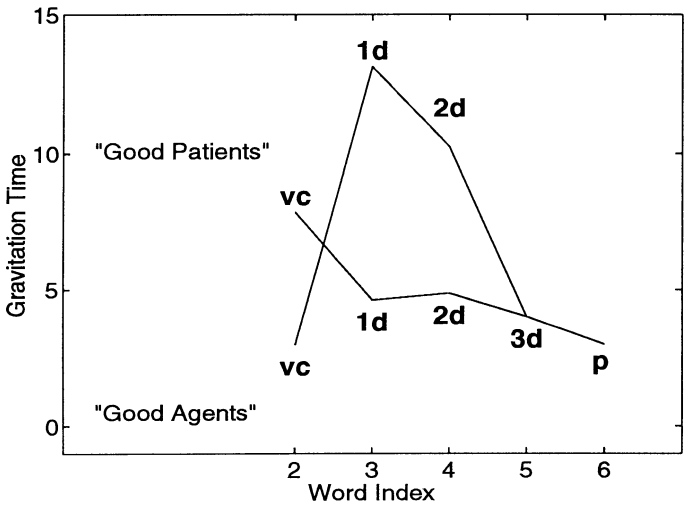

Figure 3. Gravitation times for the thematic bias simulation.

escaped."), which start with a Y word (yd or ye), continue with an X-biased verb (vb or vc), and finish with a type Y-ending (1d . . . or 1e . . .); the "Good Agents" curve shows word-by-word mean gravitation times for eight sentences like "xc vc 1d 2d 3d p" ("Cop arrested by detective escaped."), which start with an X-word (xb or xc) and continue with the same possibilities as the "Good Patients" examples. The means are significantly different across the two sentence types in word positions 2 ($t(14) = 46.09$, $p < .001$), 3 ($t(14) = 241.55$, $p < .001$), and 4 ($t(14) = 36.46$, $p < .001$), but not in positions 5 and 6 ($t(14) = 0$). Successive means within each sentence type are also significantly different in all cases except for the contrast between Word 3 and Word 4 in the "Good Patient" sentences ($t(14) = 0.209$, $p > .6$). The pattern thus shows the central properties of the human reading time data: 1) immediate effects of new information, even though the information is merely semantically biasing (at the V word, for example, there is an effect of the bias of the immediately preceding N word) 2) cross-over of the magnitudes of the reading times during the course of the sentence (first the Y or RR-bias sentence shows a spike in reading time; then the X or MC-bias sentence shows one), and 3) inertia in the parse choice (each spike has a tail which dwindles over the course of several following words).

Induction of Competition Effects. Examining the representations and the processing dynamics of the VSG model, reveals that the model is predicting the human data by implementing a competition mechanism very much like that used by McRae et al. (in press).

Figure 4 provides a global view of the visitation set for the simulation. This image was obtained by performing Principal Component Analysis (PCA) on the set of 2000 hidden unit locations used in the gravitational model. PCA (Jolliffe, 1986) is a way of choosing coordinate axes that are maximally aligned with the variance across a set of points in a

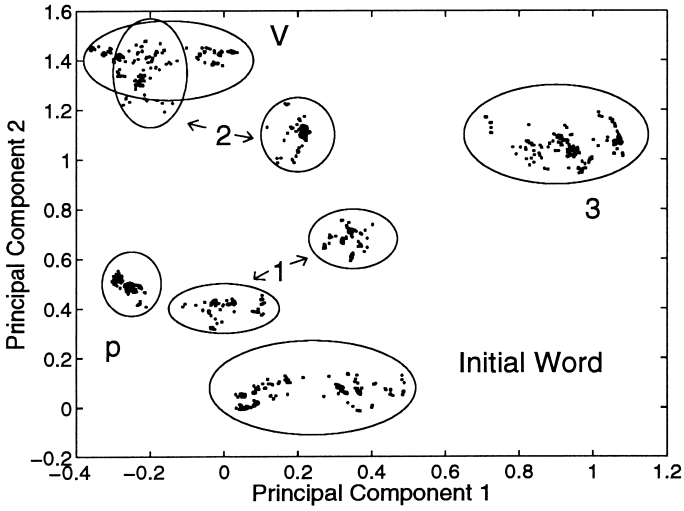

Figure 4. Global view of the visitation set for the thematic bias simulation.

space.[5] It is used here simply as a way of viewing the visitation set, and plays no role in the predictions made by the model.

The visitation set is grouped into major regions corresponding to the six major categories of the grammar (initial word, V word, 1 word, 2 word, 3 word, and final word, "p").[6] Two of these categories overlap in the two-dimensional reduced image ("V" and "2"), but they do not overlap in the 10-dimensional space. Several of the major regions seem to have two distinct clusters within them in Figure 4. These correspond to the two parse hypotheses, X ("Main Clause") and Y ("Reduced Relative").

To see this more clearly, it is helpful to zero in on one of the major clusters. Figure 5 shows a new PCA view of the points where the connectionist network places the system when its input layer is receiving a V word (the new PCA is based on all and only the V word points). Here, we can clearly see the two clusters corresponding to the X and Y readings. These clusters give rise to two attractors which are at the centers of the circles in the diagram.[7]

Three trajectories are shown corresponding to sentences which start "xc vc . . . ", "yd vc . . . " and "yf vf . . . ". The "xc vc . . . " and "yf vf . . . " cases, roughly analogous to "cop arrested" and "evidence examined", are the beginnings of normal sentences which typically give rise to X ("Main Clause") and Y ("Reduced Relative") interpretations respectively. The processor lands close to the appropriate attractor when the V word is presented and gravitation takes only two time steps.[8] By contrast, the sentence that starts with "yd vc" (analogous to "crook arrested") has conflicting information in it. The first word ("yd") indicates that the processor should favor the Y attractor, but the second word ("vc") is predominantly associated with an X continuation. As a result, the processor lands in an intermediate position when the second word is presented. It gravitates to the X attractor, but gravitation takes eight time steps.

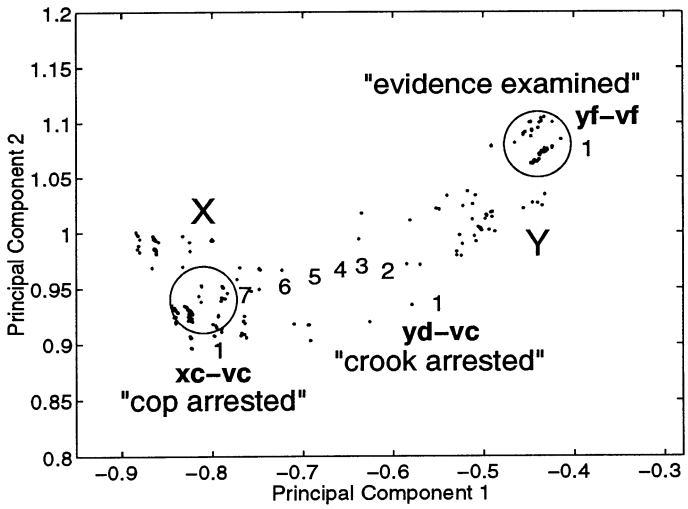

Figure 5. Three trajectories in the "V" region. (The label "yc-vc" identifies the starting point of the trajectory that ensued when "vc" had been presented on the input layer after "yc". The numbers '1', '2', '3', etc. proceeding from this label indicate the trajectory itself. The other labels have similar interpretations.)

Figure 6 (yet a different PCA) presents a close-up of the "1" region of the visitation set. Here we can observe one-word continuations of the sentences shown in Figure 4. The case of central interest is "xc vc 1d". This case is analogous to a fragment like "Cop arrested by . . . ", which begins with a Good Agent followed by an agentive verb, but continues with a "by"-phrase which strongly signals the unexpected Reduced Relative interpretation. Recall that human reading times at the agentive noun phrase in the "by"-phrase were long compared to when the sentence began with a Good Patient subject (e.g., "Crook arrested by detective . . . "). In the simulation, the first word cuing the switch, namely "1d", lands the processor in an intermediate position and has a correspondingly long gravitation time (13 time steps). The corresponding case, "yd vc 1d", ("Crook arrested by . . . ") produces a nonminimal trajectory at "1d" as well, but the starting point is closer to the "Y" attractor and the gravitation time is shorter (5 time steps). Thus, at the "V" words and "1" words, the crossing latency pattern in McRae et al.'s data is reproduced. However, the simulation shows the crossing pattern more immediately in response to the disambiguating information than do the human subjects. This may be due to "by" being read parafoveally; it may also reflect the more complex ambiguity of natural language "by".

For comparison, Figure 6 also shows a case of gravitation to the "X" attractor in the "1" region: the partial sentence, "yf-vf-1a" (presumably comparable to something like, "The evidence examined him . . ."). The first two words strongly favor a "Y" ("Reduced Relative") continuation, whereas the third word requires an "X" ("Main Clause") continuation. Reading times are elevated at "1a" but not as elevated as they were for, "xc-vc-1d". This less elevated reading time occurs because the most recent biasing information ("1a") is such a strong piece of evidence for the "X" parse (high in frequency and strongly

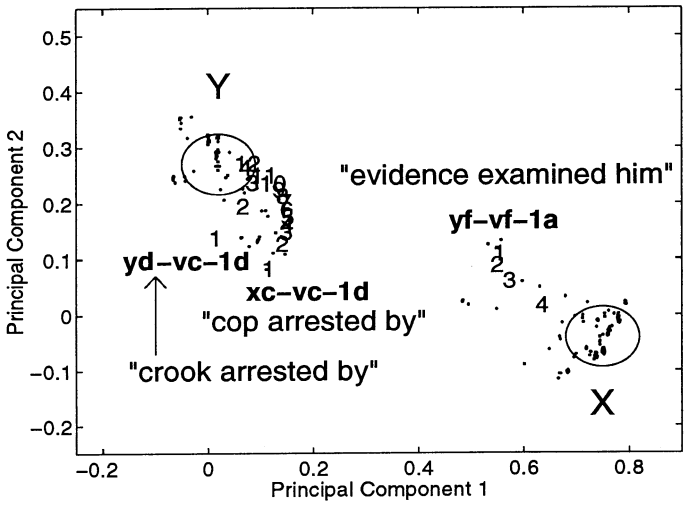

Figure 6. Three trajectories in the "1" region. (See previous figure for explanation of labels.)

correlated) that it overcomes the prior opposing bias more easily than the low-frequency, weakly correlated "1d" in "xc-vc-1d".

The competition effects we have illustrated occur whenever the same sentences are presented with different preceding contexts, and when we choose appropriately biased cases which are distributionally similar. The pattern becomes distorted if we make one or another bias especially strong, or change the directions of some of the biases. However, the cases we have focused on here seem most closely analogous to the relevant cases that psycholinguists have studied in natural language. Thus when given the constraint that the gravitation mechanism needs to form a distinct attractor basin for every syntactically distinct context, the model derives the competition mechanism hypothesized by McRae et al. from the distributional properties of its training corpus.

Emergence of a Syntax/Semantics Distinction. The VSG model also induces a distinction between "syntactic" and "semantic" types of violations. Processing syntactically well-formed strings (including semantically strange sentences) involves gravitation directly into an attractor whereas processing syntactic anomalies involves gravitation first into a saddlepoint (a fixed point which attracts trajectories from one region of the state space and repels them into another), and only later into an attractor. To illustrate this point, we extend the analogy between natural language and the Thematic Bias Grammar. In the training grammar, none of the five sequential categories is ever omitted and the elements always follow one another in the same order. Thus skipping or repeating categories is analogous to a natural language grammaticality violation.

Figure 7 shows two sample trajectories, one corresponding to a semantic violation and one corresponding to a syntactic violation. The semantic violation is one of the cases depicted in Figure 6. It occurs at the word "1d" in the sentence, "xc vc 1d 2d 3d p" (analogous to "Cop arrested by detective left."). The bias of the first two words toward X

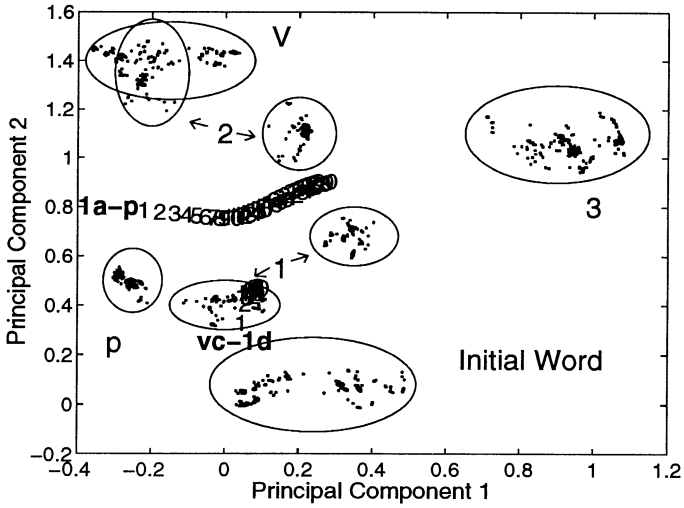

Figure 7. Trajectories for a semantic anomaly (labeled 'vc-1d' and a syntactic anomaly (labeled '1a-p'). The semantic anomaly occurs at the word '1d' in the sentence 'xc-vc-1d-2d-3d-p'. The syntactic anomaly occurs at the word 'p' in the string 'xb-va-1a-p'.

continuations is contradicted by the bias of the third word toward Y continuations so the processor slows down substantially at this word (13 time steps). Nevertheless, the string is grammatical in the sense that its category sequence is sanctioned by the grammar.

The syntactic violation in Figure 7 occurs at the word 'p' in the string, "xb va 1a p" (analogous to "Cop arrested the.") This string is ungrammatical because it ends after the third word, skipping the 2 and 3 categories. The VSG model's response in this case is substantially different from its response in the previous case. The starting point of the trajectory (labeled "1a-p") is remote from all of the clusters that are associated with normal sentence processing. Moreover, the trajectory stretches for a long way across empty space and gets pulled into what looks like an attractor midway between the "X 1" region and the "X 2" region. After 30 time steps it still has not gravitated into one of the clusters associated with normal processing. The apparent attractor is a saddlepoint. If gravitation proceeds for a much longer time, the trajectory will eventually reach an attractor. But it is clearly waylayed in a significant way compared to the trajectory of the semantic violation.

To explore the hypothesis that semantic violations involve direct gravitation into an attractor and syntactic violations involve delay by a saddlepoint, we studied the model's response to a sample of 20 semantic violations and 20 syntactic violations.

When the wrong category word was encountered, (a syntactic anomaly) the starting point of the trajectory tended to be a compromise between the contextually appropriate attractor and the attractor associated with the anomalous word. Thus syntactic anomalies nearly always placed the processor far from any of the attractors. In every case the trajectory was in the basin of the contextually appropriate attractor. In some cases, the trajectory was drawn into a saddlepoint close to the contextually appropriate attractor. The

'1a-p' trajectory in Figure 7 is a case like this: the contextually appropriate attractor is the 'X 2' attractor (east of the label "2" in Figure 7). In other cases, the trajectory went quickly into the contextually appropriate attractor despite the anomaly. An example is the word 'ye' after 'yf' in the sentence, "yf ye vf 1f 2e 3f p", which resulted in gravitation into the 'Y V' region in 3 timesteps. Often, then, the next word produced a trajectory that was still trapped behind a saddlepoint at the 30th time step. Thus, the model sometimes showed delayed sensitivity to an anomaly. We do not at this point know why the long reaction times were sometimes coincident with the anomalous word and sometimes delayed by a word or two, but we note that this behavior may be consistent with human behavior and bears further looking into. We assessed the outcome of the trials by examining the trajectory for the anomalous word and the word following it and tabulating results for whichever of these trajectories was longer. Figure 8 plots the velocity profiles of these maximally anomalous trajectories for the sets of syntactic and semantic anomalies. The velocity between two successive points $\vec{x}_i$ and $\vec{x}_j$ on a trajectory is taken to be the distance between $\vec{x}_i$ and $\vec{x}_j$ (since each step takes unit time). The difference between the mean maximal gravitation times across the two classes of anomaly is significant ($t(38) = 7.97$, $p < .001$).

Figure 8 suggests that the difference between grammatical and ungrammatical strings is a *graded qualitative difference*. At one extreme are the parses with short, direct trajectories into an attractor and thus short processing times. At the other extreme are trajectories which land on what is called the *stable manifold* of a saddlepoint. The stable manifold contains those points which happen to be at the balance point between the competing attractors and from which the system gravitates into the saddlepoint itself, never reaching an attractor. These two kinds of behaviors are qualitatively distinct: in the first case the processor arrives at a representation which is associated with an interpretation; in the second case it never arrives at such an interpretation. However, almost all real examples are a mixture of these two types: even the clearest grammatical examples show very slight influence of deflection by saddlepoints; even the worst grammatical anomalies are unlikely to land on a stable manifold of a saddlepoint, and thus will eventually gravitate into an attractor. Nonetheless, there is a clear clustering of strings into two classes: grammatical and ungrammatical.

This framework provides a useful new conceptualization of the notion of grammaticality judgments. It also makes making several clear predictions that differentiate the VSG model from models that make an absolute distinction between grammatical and ungrammatical sentences and models (e.g., the SRN) that treat all contrasts as graded: 1) people should show gradations of reading times on "grammatical" and "ungrammatical" sentences even when they make clear binary grammaticality judgments, 2) lexical and other contextual biasing can lead to a semantic anomaly behaving like a syntactic anomaly and 3) there should be variation in whether a syntactic violation leads to an immediate or delayed increase in reading time.

IV. CONCLUSIONS

Drawing upon dynamical systems theory, the VSG model provides a useful set of constructs for understanding the representational properties of high-dimensional learning

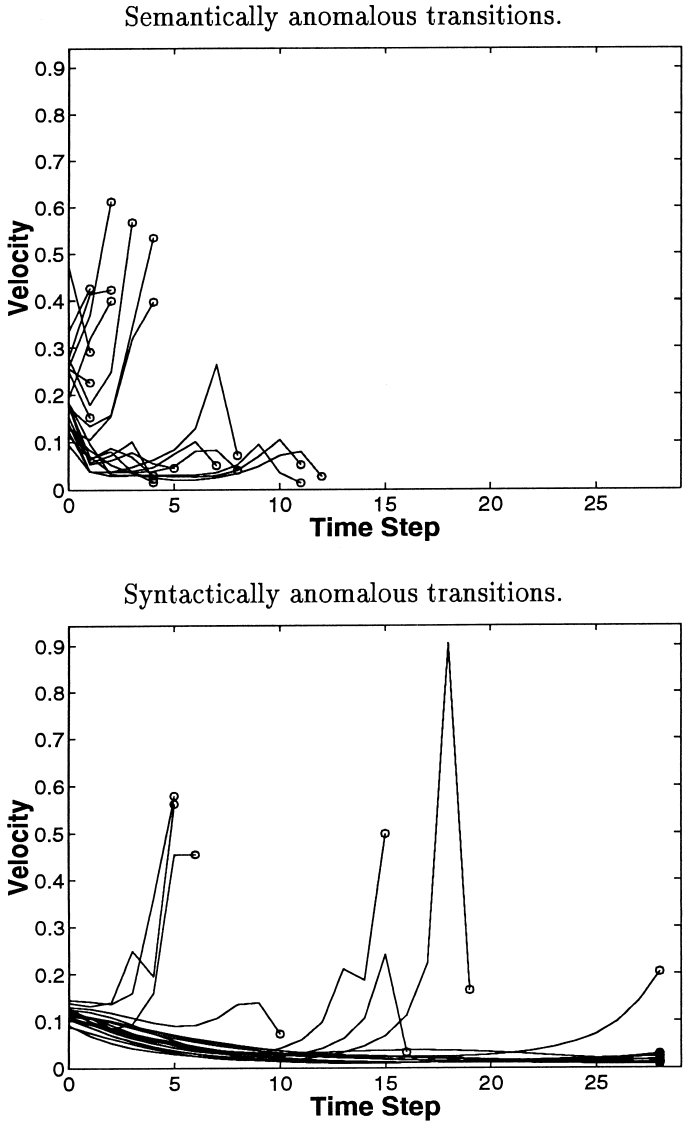

Figure 8. Velocity profiles for 20 semantically and 20 syntactically anomalous transitions. The profile is pictured for either the word at which the anomaly occurred or the word following this word, whichever had a longer gravitation time.

models like Elman's (SRN). In particular, the VSG model can be tuned so its attractor basins identify clusters in the SRN's representation space which correspond to states of the generating process. Clustering seems to be an important step in mapping from the continuous representations of learning models to the discrete representations of linguistic models. The current results suggest that the VSG model provides an improvement over hierarchical clustering methods of discretizing connectionist representations (e.g., Elman,

1990; Pollack, 1990; Servan-Schreiber, Cleeremans, & McClelland, 1991), for these provide no obvious way of picking out a linguistically or statistically relevant subset of a cluster hierarchy.

More specifically, we extended the results of Tabor et al. (1997) by modeling reading times for thematic effects on processing sentences with reduced relatives and we showed how a competition mechanism which has been used to model ambiguity effects by Michael Spivey and his colleagues arises as an emergent property of the model. We also found that the processing of grammatical strings tended to involve gravitation directly into an attractor, whereas the processing of ungrammatical strings usually led to gravitation into a saddle point which greatly delayed arrival at an attractor. This result provides a way of mapping the graded representation of an SRN (it rules out no string) onto the intuitively observable contrast between semantic and syntactic violation.

Future Challenges for the VSG Model

First, the link between the SRN and the gravitation mechanism depends upon an external constraint (the requirement that attractor basins line up with parse states) to set the parameter p. If varying the parameter p over all possible values could produce arbitrary attractor basin configurations, then the dynamical component would contributing little insight. But the model is, in fact, fairly tightly constrained: experimentation suggests that varying p leads to a relatively small range of basin configurations, with a simple case in which there is only one basin ($p = 0$) and a limiting case in which every point in the visitation set has its own basin. Nonetheless, it would be desirable if the value of p could be determined independently of a grammatical oracle and we are currently investigating this possibility.

Second, we have only analyzed VSG behavior on a simple formal language. While this provides a necessary foundation for future work, it will be important to study more realistic cases—e.g., one could incorporate a number of specific correlations between subjects and verbs, like the fact that "cop" is a good subject for "arrest" and "employee" is a good object for "hired" rather than a binary contrast between two biases (Good Agent vs. Good Patient). To this end, it is also important to address the question of how to represent phrase structural relationships as well as simple contrasts between states in a finite-state language. Rodriguez, Wiles, & Elman (1999), Tabor (1998), and Wiles and Elman (1995) provide some insight into this problem by looking at how SRNs and related devices can represent context free grammars. A central question is, How should the learning mechanism generalize from its finite training experience to an infinite-state language?

Third, the faster processing of semantic violations compared to grammatical violations in the current simulation is not surprising, given that the model is likely to have seen most semantic violations in training. It will be important in future research to demonstrate that the model exhibits generalization ability by removing a random sample of grammar-generated strings from the training data and using these as test cases. However, real semantic anomaly is not randomly distributed across grammatically legitimate combina-

tions: it is associated with the juxtaposition of particular word classes. Thus, a more interesting test requires using a grammar in which certain classes of words never directly co-occur, although they have a strong higher-order correlation, to see whether the gravitation mechanism will be able to appropriately group clusters of clusters into the same attractor basin (e.g., "dogs" don't "meow" or "purr" but they "eat", "run", "play", "sleep", etc.—things that "meow"-ing and "purr"-ing individuals commonly do).

These challenges are nontrivial, but they arise from asking the challenging question that motivates the VSG model: How can the relativistic perspective of a learning model be mapped in a principled way to the more absolutist perspective which supports categorical decisions? It is not obvious that there is any universally right way of taking this step. However, simplicity is desirable and dynamical systems theory gives insight into how rich structure can emerge from fairly simple assumptions. The dynamical perspective thus provides a promising way for connectionist natural language modeling to handle linguistic complexity without losing its useful relativism.

Acknowledgments: We thank Nick Chater, Morten Christiansen, Garrison Cottrell, William Turkel, Michael Spivey-Knowlton and Gary Dell for helpful comments. We would also like to give special thanks to Cornell Juliano whose involvement with the predecessor of this paper helped to steer us in the direction of our recent results. W. Tabor was supported by NIH Grant 5 T32 MH19389. M. K. Tanenhaus was supported by NIH Grant HD 27206.

NOTES

1. Thus, the hidden-to-output weights were adjusted according to

$$\Delta w_{ji} \propto y_i \delta_j = y_i (t_j - y_j)$$

while the input-to-hidden and hidden-to-hidden weights were adjusted according to

$$\Delta w_{ji} \propto y_i \delta_j = y_i f'(net_j) \sum_k w_{kj} \delta_k$$

where w_{ji} is the weight from unit i to unit j and $f'(net_j) = y_j(1 - y_j)$ is the derivative of the sigmoid activation function.

2. N must be large enough to make the cluster structure of the visitation set discernible; ν controls the rate of gravitation but does not affect relative rates of gravitation, so it can be scaled for implementational convenience. Without loss of generality, then, we assume $\Delta t = 1$.
3. We noted earlier that constraint-satisfaction models have been proposed as an alternative to "two-stage" models of sentence processing (Frazier & Clifton, 1996). The VSG model also performs computations in two distinct stages—the recurrent network computation, and the gravitation computation. But there are important differences between the VSG model and traditional two-stage models. In the VSG model, there is no early stage during which some information is systematically ignored. Rather, all information is present from the beginning of each word's settling process. Moreover, the second stage does not involve deconstructing and rebuilding parse trees, but rather migration in a continuous space. Finally, systematic biases in favor of one structure over another stem mainly from greater experience with the preferred structure, not from an avoid-complexity strategy (see MacDonald & Christiansen, 1998).

4. For this grammar, the minimum distance between grammar-determined distributions is 0.9410—this is, for example, the distance between the distribution associated with the partial string "xa . . . " and the distribution associated with the partial string, "ya . . . ".
5. In the case at hand, the original hidden unit space had 10 dimensions. The first two principal components captured 56 percent of the variance.
6. To make Figure 4 interpretable, we have circled and labeled the regions corresponding to distinct classes based on our knowledge of which words correspond to which points.
7. The circles were drawn as follows: an estimation of the location of the attractor was computed by averaging the second- and third-to-last positions of the trajectory for several trajectories and a circle of fixed radius was drawn with this point as its center. Recall that the trajectory is considered at an end when it makes a turn of more than 90 degrees on one step. This happens immediately after it has passed by the attractor. Therefore the attractor is usually located somewhere between the second- and third-to-last positions, so their average provides a reasonable estimate of its location. The circle radii have no explanatory significance—they are just a method of identifying the attractor location without obscuring the view by putting a label right on it.
8. The first step of each trajectory is marked by "1"; the second step brings the trajectory into the attractor and is not shown in order to make the diagram easier to read.

REFERENCES

Abraham, R. H., & Shaw, C. D. (1984). *Dynamics—the geometry of behavior, Books 0–4.* Aerial Press, Inc., P.O. Box 1360, Santa Cruz, CA.

Ainsworth-Darnell, K., Shulman, H., & Boland, J. E. (1998). Dissociating brain responses to syntactic and semantic anomalies: Evidence from event-related potentials. *Journal of Memory and Language, 38,* 112–130.

Almeida, L. B. (1987). A learning rule for asynchronous perceptions with feedback in a combinatorial environment. In M. Caudil & C. Butler (Eds.), *Proceedings of the IEEE First Annual International Conference on Neural Networks* (pp. 609–618). San Diego, CA: IEEE.

Burgess, C., & Lund, K. (1997). Modeling parsing constraints with a high-dimensional context space. *Language and Cognitive Processes, 12,* 177–210.

Charniak, E. (1993). *Statistical language learning.* Cambridge, MA: MIT Press.

Christiansen, M. H. (1994). *Infinite languages, finite minds: Connectionism, learning, and linguistic structure.* Unpublished doctoral dissertation, University of Edinburgh.

Christiansen, M. H., & Chater, N. (1999). Toward a connectionist model of recursion in human linguistic performance. *Cognitive Science, 23,* 157–205.

Christiansen, M. H., & Chater, N. (1999). Connectionist natural language processing: The state of the art. *Cognitive Science, 23,* 000–000.

Cottrell, G. W. (1985). Connectionist parsing. In *Proceedings of the Seventh Annual Meeting of the Cognitive Science Society* (pp. 201–11). Irvine, CA: Cognitive Science Society.

Cottrell, G. W., & Small, S. (1983). A connectionist scheme for modeling word sense disambiguation. *Cognition and Brain Theory, 6,* 89–120.

Cottrell, G. W., & Small, S. (1984). Viewing parsing as word sense discrimination: A connectionist approach. In B. Bara, & G. Guida (Eds.), *Computational models of natural language processing* (pp. 91–119). Amsterdam: North Holland.

Crutchfield, J. P. (1994). The calculi of emergence: Computation, dynamics, and induction. *Physica D, 75,* 11–54.

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.

Frazier, L., & Charles Clifton, J. (1996). *Construal.* Cambridge, MA: MIT Press.

Garnsey, S. M. (1993). Event-related brain potentials in the study of language: An introduction. *Language and Cognitive Processes, 8,* 337–356.

Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition, 68(1),* 1–76.

Hagoort, P., Brown, C., & Groothusen, J. (1993). The syntactic positive shift (SPS) as an ERP measure of syntactic processing. *Language and Cognitive Processes, 8,* 439–483.

Hopcroft, J. E., & Ullman, J. D. (1979). *Introduction to automata theory, languages, and computation.* Menlo Park, CA: Addison-Wesley.

Jolliffe, I. T. (1986). *Principal component analysis.* New York: Springer-Verlag.

Juliano, C., & Tanenhaus, M. (1993). Contingent frequency effects in syntactic ambiguity resolution. In *Proceedings of the Fifteenth Annual Meeting of the Cognitive Science Society* (pp. 593–598). Mahwah, NJ: Lawrence Erlbaum Associates.

Jurafsky, D. (1996). A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive Science, 20,* 137–194.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104,* 211–240.

MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological Review, 101,* 676–703.

MacDonald, M. C., & Christiansen, M. H. (in press). Reassessing working memory: A reply to Just & Carpenter and Waters & Caplan. *Psychological Review.*

McRae, K., Spivey-Knowlton, M. J., & Tanenhaus, M. K. (1998). Modeling the influence of thematic fit (and other constraints) in online sentence comprehension, *Journal of Memory and Language, 38,* 283–312.

McClelland, J., & Rumelhart, D. (1981). An interactive activation model of context effects in letter perception, Part i. *Psychological Review, 88,* 375–402.

McElree, B., & Griffith, T. (1995). Syntactic and thematic processing in sentence comprehension: Evidence for a temporal dissociation. *Journal of Experimental Psychology: Language, Memory, and Cognition, 21,* 134–157.

Osterhout, L., & Holcomb, P. J. (1993). Event-related potentials and syntactic anomaly: Evidence for anomaly detection during the perception of continuous speech. *Language and Cognitive Processes, 8,* 413–437.

Pineda, F. J. (1995). Recurrent backpropagation networks. In Y. Chauvin & D. E. Rumelhart (Eds.), *Back-propagation: Theory, architectures, and applications* (pp. 99–136). Hillsdale, NJ: Erlbaum.

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. E. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review, 103,* 56–115.

Pollack, J. B. (1990). Recursive distributed representations. *Artificial Intelligence, 46,* 77–106.

Rodriguez, P., Wiles, J., & Elman, J. (1999). A recurrent neural network that learns to count. *Connection Science, 11*(1), 5–40.

Rumelhart, D., Durbin, R., Golden, R., & Chauvin, Y. (1995). Backpropagation: The basic theory. In *Back-propagation: Theory, architectures, and applications.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986a). Learning internal representations by error propagation. In Rumelhart, D. E., McClelland, J. L., & the PDP Research Group, (Eds.), *Parallel distributed processing* (Vol. 1, pp. 318–362). Cambridge, MA: MIT Press.

Rumelhart, D. E., McClelland, J. L., & the PDP Research Group (1986b). *Parallel distributed processing. Explorations in the microstructure of cognition.* (Vol. 1). Cambridge, MA: MIT Press.

Selman, B., & Hirst, G. (1985). A rule-based connectionist parsing system. In *Proceedings of the Seventh Annual Meeting of the Cognitive Science Society* (pp. 212–221). Irvine, CA: Cognitive Science Society.

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

Spivey-Knowlton, M. (1996). *Integration of visual and linguistic information: Human data and model simulations.* Unpublished doctoral dissertation, University of Rochester.

St. John, M. F., & McClelland, J. L. (1990). Learning and applying contextual constraints in sentence comprehension. *Artificial Intelligence, 46,* 217–257.

Strogatz, S. (1994). *Nonlinear dynamics and chaos.* Reading, MA: Addison-Wesley.

Tabor, W. (1994). *Syntactic innovation: A connectionist model.* Unpublished doctoral dissertation, Stanford University.

Tabor, W. (1998). *Dynamical automata.* Technical Report No. TR98–1694, Cornell Computer Science Department. Available at http://cs-tr.cs.cornell.edu/.

Tabor, W., Juliano, C., & Tanenhaus, M. (1996). A dynamical system for language processing. In Cottrell, G. W. (Ed.), *Proceedings of the Eighteenth Annual Meeting of the Cognitive Science Society* (pp. 690–695). Hillsdale, NJ: Lawrence Erlbaum Associates.

Tabor, W., Juliano, C., & Tanenhaus, M. (1997). Parsing in a dynamical system: An attractor-based account of the interaction of lexical and structural constraints in sentence processing. *Language and Cognitive Processes, 12,* 211–271.

Tanenhaus, M. K., & Trueswell, J. C. (1995). Sentence comprehension. In Miller, J., & Eimas, P., (Eds.), *Handbook of perception and cognition*. (Vol. 1, pp. 217–262). San Diego: Academic Press.

Trueswell, J. C. (1996). The role of lexical frequency in syntactic ambiguity resolution. *Journal of Memory and Language, 35,* 566–585.

Waltz, D. L., & Pollack, J. B. (1985). Massively parallel parsing: A strongly interactive model of natural language interpretation. *Cognitive Science, 9,* 51–74.

Weckerley, J., & Elman, J. (1992). A PDP approach to processing center-embedded sentences. In *Proceedings of the Fourteenth Annual Meeting of the Cognitive Science Society* (pp. 414–419). Mahwah, NJ: Lawrence Erlbaum Associates.

Wiles, J., & Elman, J. (1995). Landscapes in recurrent networks. In Moore, J. D., & Lehman, J. F., (Eds.), *Proceedings of the Seventeenth Annual Meeting of the Cognitive Science Society* (pp. 482–487). Hillsdale, NJ: Erlbaum.

Williams, R. J., & Peng, J. (1990). An efficient gradient-based algorithm for on-line training of recurrent network trajectories. *Neural Computation, 2,* 490–501.