Manuscript Information Sheet

Authors: Whitney Tabor and Michael K. Tanenhaus

Title: Dynamical Models of Sentence Processing

Number of manuscript pages: 61

Number of equations: 2

Number of tables: 1

Number of figures: 8

Contact address:

Whitney Tabor

Department of Psychology

University of Connecticut

Storrs, CT 06269

Contact phone: (860) 486-4910

Contact fax: (860) 486-2760

Contact email: tabor@uconn.edu

Suggested running head: Dynamical Sentence Processing

Dynamical Systems for Sentence Processing

Whitney Tabor

University of Connecticut

Michael K. Tanenhaus

University of Rochester

Please address correspondence to:

Whitney Tabor

Department of Psychology

University of Connecticut

Storrs, CT 06269

tabor@uconn.edu

RUNNING HEAD: Dynamical Sentence Processing

July 3, 1999

1.1 The Dynamics of Sentence Processing

The syntactic constraints of a language strongly determine the interpretation that a reader or listener arrives at for a sentence. Thus, the human language comprehension system must develop and evaluate hypotheses as to how to map the linguistic input into appropriate syntactic units. Temporary ambiguity is the central problem faced by the system. Because the linguistic input is typically consistent with multiple syntactic possibilities, the processing system must determine the set of possible syntactic hypotheses, maintain some or all of these in memory, and update them as new input arrives.

Behavioral evidence from a rapidly expanding literature on how people read temporarily ambiguous sentences provides an empirical benchmark for evaluating theories of syntactic processing (see Tanenhaus & Trueswell, 1995). Evidence from intuitions and from empirical studies of processing difficulty—typically studies using reading time measures with temporarily ambiguous sentences—have clearly established that readers have strong preferences for some structures over others. When subsequent input becomes inconsistent with the preferred structure, the result is processing difficulty for the reader.

The sentence in example (1), taken from Bever (1970), is a classic example of such a "garden-path" sentence.

(1) The horse raced past the barn fell.

The reader assumes that the first six words of the sentence form a main clause in the active voice with "raced" being a past tense main verb, and "the horse" playing the role of "agent" of the racing event. This hypothesis is disconfirmed by the word

"fell", however, resulting in long reading times and confusion. Many readers are unable to arrive at the "grammatical" analysis in which "fell" is the main verb, and "the horse (which was) raced past the barn" as its subject.

Within traditional symbolic systems, syntactic hypotheses are typically computed by a parser—a set of procedures that maps the input onto partial syntactic structures that are consistent with the syntactic constraints of the language. These constraints are described by a grammar consisting of a set of rules/and or constraints defined over syntactic categories, such as Noun and Noun Phrase. The procedures that comprise the parser build structures using the knowledge base defined by the grammar.

A variety of structural hypotheses have been proposed to account for why some structures are preferred over others in an initial stage of structure building. These hypotheses are typically couched in terms of the complexity of structure building operations, and/or memory demands (see Frazier & Clifton, 1996; Gibson (to appear), for recent reviews). In such *two-stage* models, a second set of procedures is involved in recovering from misanalysis when the preferred structure selected in the initial parse is rejected.

However, recent research has highlighted a number of phenomena that are problematic for this view. A growing body of evidence indicates that syntactic processing is simultaneously affected by semantic, syntactic and discourse-based information (for reviews see MacDonald, et al., 1994; Tanenhaus & Trueswell, 1995). Moreover, processing is sensitive to differences among individual lexical items within syntactic categories. Thus sentences like (2) are much easier to process than sentences like (1) even though the sequence of standard lexical categories ("Det Noun Verb[ed form] Preposition etc.") is identical across the two examples.

(2) The salmon released in the ocean died.

In addition, the speed with which readers process the words of sentences like (1) and (2) is correlated with graded properties of the linguistic input that are not easily reduced to purely structural factors. For example, the degree to which readers have difficulty at the second verb in these sentences is negatively correlated with the degree to which the first verb tends to occur, in large natural language corpora, as a past participle (MacDonald et al, 1994; Trueswell, 1996).

Phenomena like these have led a number of researchers to propose constraint-based frameworks in which multiple sources of constraint provide probabilistic evidence in support of the most likely syntactic alternatives. Ambiguity resolution is viewed as a constraint-satisfaction process, involving competition among incompatible alternatives. As a sentence unfolds, the alternatives are evaluated using evidence provided by the current input as well as the preceding context (Cottrell and Small, 1983, 1984; Cottrell, 1985; Waltz and Pollack, 1985; St. John and McClelland, 1990; MacDonald et al., l994; Spivey-Knowlton, 1996). Processing difficulty occurs when input is encountered that is inconsistent with the previously biased alternative.

Connectionist (or neural network) models are one variety of constraint-based models. Typically, their properties as learning devices play a central role in their use as models. In particular, it is often suggested that the systematic properties of language which motivate the positing of specialized structures in a symbolic paradigm will arise as "emergent properties" under connectionist learning. The interest of this claim is not merely that it provides an explicit proposal about the grammar-induction part of the theory. It is also suggested that the emergent counterparts of symbolic mechanisms will

be different from them in important ways.

We endorse all of these claims here, but note that 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. This paper and Tabor et al. 1997 argue that a connectionist, learning-based system can be explicit about "emergent properties" by using the constructs of *dynamical systems theory*.

Dynamical systems theory is the theory of systems that are described in terms of how they change. Formally, this description has the form of a differential equation or an iterated map. Commonly studied examples of real dynamical systems are swinging pendulums, orbiting planets, circulating fluids, etc. Certain constructs are useful in analyzing such systems: trajectories, fixed points (or stable states), attractors, basins, saddlepoints (see Abraham and Shaw, 1984 and Strogatz, 1994 for introductions). Along with Cornell Juliano, we first explored the idea that such constructs might be useful in clarifying the principles underlying connectionist sentence processing (Tabor et al., 1997). There, we focussed on the way the dynamical approach allows us to handle class frequency effects and the interaction of formal similarity with structural constraints. The present paper continues this line of investigation by focusing on how the dynamical approach handles early effects of thematic role biases, often put forth as evidence for constraint-based modeling, without losing track of the structural constraints.

Our work is similar to other recent connectionist approaches to parsing in that we train a variant of a Simple Recurrent Network (SRN) using the prediction task developed by Elman (1991). The input to the model is a sequence of words generated by a finite state or context free grammar. The model forms representations of parse states in its hidden unit space. It places words that are likely to be followed by similar constructions

nearby one another in the hidden unit space (Elman, 1990, 1991; Christiansen, 1994, Tabor, 1994). In symbolic parameter-setting models of grammar learning (e.g., Lightfoot, 1991) syntactic structures are listed fully-formed in a mental warehouse of possibilities prior to learning. By contrast, in network models like ours, syntactic structure is built up by the network as it learns to process the input. The result is a correspondingly greater influence of the learning process on the final performance of the system (Christiansen, 1994; Christiansen and Chater, in press; MacDonald and Christiansen, 1998).

While prior learning connectionist models have revealed the important role that experience (or training) plays in adult sentence processing, they lack an explicit analog of reading times. Instead, they typically map output activations onto reading times with a formula (Christiansen & Chater, in press; MacDonald & Christiansen, 1998) Moreover, the representations developed by these and many connectionist models are hard to interpret in terms that reveal the structural principles underlying the models' empirical successes.

To address these shortcomings, we add a dynamical processor to the SRN. This processor transforms the sometimes ambivalent representations produced by the network into unique parse hypotheses, requiring varying amounts of time to 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 and it uses a gravitational mechanism to group these into distinct classes, thus providing useful structural information about the SRN's representation. We refer to the resulting model, which we first explored in Tabor et al. (1997) as the *Visitation Set Gravitation* (or VSG) model.

1.2 Overview of the paper

The remainder of this article is organized into three sections. In Section 2, we define the VSG model and motivate it by describing related modeling work. In Section 3 we show how the model can incorporate a kind of "semantic constraint" (the thematic role biases of nouns) into the resolution process. An appealing characteristic of the model is that it appears to make the intuitively appropriate distinction between syntactically and semantically incongruous sentences. Section 4 concludes.

2. The VSG model and related dynamical models

2.1 The VSG model

The VSG model has two components: a network similar to a Simple Recurrent Network or "SRN" (Elman 1990, 1991) and a gravitation module.

Elman (1991) describes a procedure for training a particular connectionist network, the SRN, to predict distributional information about a corpus of words. Each word in the corpus is assigned a unique *indexical bit vector* (a vector with one element equal to 1 and all others equal to 0). These vectors are presented on the input layer of the network in the order in which the corresponding words occur in the corpus. The network is trained on the task of predicting on the output layer, which word is coming next for each input.

Elman uses a 3-layer feedforward network with a "context layer" feeding into the hidden layer at each timestep (Elman, 1991). The context layer contains a copy of the

hidden unit activations on the previous time step. The network is trained using the backpropagation algorithm (e.g., Rumelhart, et al., 1986). This feedforward network with a context layer 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). Elman's training procedure is an approximation of the Backpropagation Through Time (BPTT) algorithm (Rumelhart, et al., 1986) in which the error propagation through the recurrent connections in the hidden units is cut off after it has been propagated back through just one hidden-layer time step. (Thus input-to-hidden weights only receive an error signal from the current-time hidden units.)[1]

In the simulations discussed below, we used the same recurrent architecture as Elman did for the relaxation dynamics, but carried error propagation through two hidden-layer time steps while still adjusting input-to-hidden weights only on the basis of the current time step (see Figure 1). The extra time step makes learning the longer-distance dependencies that occur in language tasks a little bit easier. Our 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,

———————— insert equation 1 about here ————————

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, et al., 1995) and backpropagated through the unfolded network.[2] Weights

were adjusted after every input presentation. The network was trained on the output of a simple grammar approximating those features of English syntax which appear to be relevant to the phenomenon we model. The phenomenon, the grammar, and the training specifics are described in detail in Section 3.

———— insert figure 1 about here ————

The second component of the VSG model, the gravitation module, operates on the hidden layer representations of the trained recurrent network. Elman's work and our earlier work (Tabor et al., 1996) suggests that words in context with similar distributional characteristics are placed near one another in the hidden unit space by an SRN. A consequence is that 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 VSG's gravitation module is a clustering mechanism which finds such equivalence classes of states. It operates as follows. Once the network is trained, we present it with a large random sample of sentences generated by the grammar and record all the hidden unit states visited (that is to say, the *visitation set*) during the processing of this sample. We treat each of these points as a fixed mass of unit magnitude in the 10-dimensional hidden unit space. We then test the processing of a particular word-in-context by treating the hidden unit location of that word-in-context as a test mass (also of unit magnitude) which is free to move under the gravitational influence of all the fixed masses. Typically, the test mass will be near the center of mass of some dense cluster and will gravitate into that cluster. We model processing time as the time required to gravitate into the cluster. One can think of the fixed masses 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 in the centers of the clusters where the system is stable are called *attractors*.[3] The set of all starting points from which the system gravitates into a particular attractor is called its *basin*. In the case we study, 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. These classes correspond to states of the grammar (Crutchfield, 1994, Hopcroft & Ullman, 1979) that generated the training data.

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

———— insert equation 2 about here ————

where x is the position of the test mass, N indexes the fixed masses, x_i is the position of the i'th fixed mass, r_i is the Euclidean distance between x_i and 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 has zero velocity at infinite distance from the fixed masses, (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 introduce a threshold r_{min} and set $r = r_{min}$ whenever r becomes smaller than r_{min}. This makes 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.[4] The last one, (p, or gravitational strength), is undesirably unconstrained—we

set it to a value that makes the attractor basins correspond to distinct parse states as defined by the training grammar. The fact that such a value for p has existed in nearly all the cases we have tried so far indicates that the model is quite restrictive, for varying p over all possible values defines a relatively small set of basin structures. Moreover, the choice of p is closely tied to the constraints on learning and so there may be a way to bind it less stipulatively (See the next subsection for discussion). Under these assumptions, the test mass will typically speed up as it approaches a fixed point (or chaotic attractor) near the center of mass of a cluster, overshoot the attractor (because it is unlikely that the mass will land exactly on a fixed point for positive Δt), and then head back toward the center of mass for another "fly-by". Our algorithm for determining gravitation times thus computes the number of steps it takes the test mass to reverse direction for the first time (where a direction reversal is a turn of more than 90° in one step).

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

(i) Feed a sentence one word at a time to the trained network using SRN relaxation dynamics.

(ii) For each word of the sentence, use the gravitation module to determine a gravitation time.

(iii) Compare gravitation time profiles (e.g., across words in a sentence) to reading time profiles.

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.

2.2 Previous related models

In order to motivate the model we have just described, we review previous, related models. Several of the models we discuss are connectionist models. It is worth noting that most connectionist models are dynamical systems of the standard sort: 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. In fact, one can distinguish two important dynamical regimes within the connectionist framework: learning dynamics and relaxation dynamics. Learning dynamics involve slow adjustment of connection weights in an attempt to find a minimum of a cost function. Relaxation dynamics involve rapid adjustment of activation values in order to compute an output. The VSG model clarifies the relationship between these two types of dynamical regimes by showing how there are relaxation dynamics (albeit in a non-connectionist system) that reveal structural properties of the learning dynamical system for one type of network (the SRN).

Earlier connectionist processing models (e.g., McClelland and Rumelhart, 1981; Cottrell and Small, 1983, 1984; Cottrell, 1985, Waltz and Pollack, 1985) usually examined the relaxation dynamics of hand-designed models. Nodes represented concepts that are naturally interpretable by human beings (e.g., [the word "throw"], [the concept, toss]), and all node properties were explicitly designed by the researchers—no learning was involved. These models had many of the properties that we make use of here. For example: (i) Competition between simultaneously valid parses increased processing time. (ii) The magnitudes of real-valued weights were adjusted to reflect contrasts in frequency and thus gave rise to biases in favor of more frequently-encountered interpre-

tations (e.g., Cottrell and Small, 1984). (iii) Sometimes, "spurious" attractive states arose which corresponded to no interpretation (e.g., Cottrell and Small, 1984). In some earlier models, these spurious states were considered a liability because some parsable sentences got stuck in them. Below, in Section 3.3.2, we show that certain "spurious states" are an asset in that they provide a plausible model of what happens when one attempts to parse an ungrammatical string (cf. Plaut et al., 1996);. (iv) Syntactic and semantic information were used simultaneously to constrain the parse (e.g., Cottrell & Small, 1983; Cottrell, 1985).

The development of the backpropagation algorithm for learning (Rumelhart et al., 1986) and its promotion as a useful tool in psychological modeling (Rumelhart and McClelland, 1986) led to a new class of connectionist parsing models. This algorithm made it possible to set weights and hidden node interpretations in a systematic way, without requiring as many subjective guesses as were needed in hand-designed models. Currently, the most successful learning connectionist models of sentence processing are Elman's Simple Recurrent Network or SRN (Elman 1990, 1991) and its variants (e.g., St. John & McClelland, 1990).[5] 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, in press; MacDonald & Christiansen, 1998)

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 like 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, saddlepoints 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 Section 3.2.3).

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.

The attractor basins defined by the VSG model are primarily valuable for the insight they provide into the representations learned by an SRN. One may reasonably wonder, though, if they have any motivation independent of the problem of predicting reading times. In fact, there is an independent functional motivation for having attractor basins: when we interpret language, we make, and probably need to make, discrete choices. 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 simultaneously understand both. 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 has not previously been evident how to map the real-valued states of an SRN onto such discrete interpretations. The VSG model provides a principled method of mapping from the SRN state vectors to discrete parse states which may be useful in distinguishing meanings.

We noted earlier that constraint-satisfaction models have been proposed as an alternative to "two-stage" models of sentence processing (Frazier and 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 migrating 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 and Christiansen, 1998).

2.3 Previous VSG results

In our earlier work (Tabor, Juliano, and Tanenhaus, 1997), we showed that the VSG model predicts word-by-word reading times in a set of cases that are challenging for other theories. We summarize the results here in order to situate our further exploration of the model.

In one simulation, we considered lexical category ambiguities involving the word "that", which exhibit an interesting mix of contingent frequency effects (Juliano & Tanenhaus, 1993; Tabor et al, 1997). The sentences in (3) 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).

(3) a. That marmot whistles.

b. That marmots whistle is surprising.

c. A girl thinks that marmot whistles.

d. A girl thinks that marmots whistle.

The results of Juliano and Tanenhaus (1993) indicate that processing times in these four 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, (3a) is easier than (3b), while (3c) is harder than (3d) at the words following "that". These results are consistent with a host of experimental results which suggest that reading times are correlated with the unexpectedness of continuations (see Jurafsky, 1996, for review). In Tabor et al. (1997), we showed that such effects fall out of the VSG model because of the denser visitation clusters associated with more-frequent continuations. Denser clusters give rise to stronger gravitational pull and hence more rapid gravitation.

On the other hand, the correlation between unexpectedness and reading times is not perfect. It seems to be skewed by the category structure of the grammar. For example, Juliano and Tanenhaus (1993) found that, after strictly transitive verbs like *visited*, the word *that* and a following adjective (4a) was read more slowly than the word *those* and a following adjective (4b).

(4) a. The writer visited <u>that old</u> cemetery.

b. The writer visited <u>those old</u> cemeteries.

Such a result cannot be attributed to the frequency of *that* vs. *those* after transitive verbs because the frequencies are essentially the same (at least in the Penn Treebank). The VSG model predicts the effect at the determiner as a case of attractor competition. The word *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* in the relevant examples.[6] Tabor et al. (1997) showed how a similar effect predicts the observed higher reading times at *the* after a pure sentence complement verb like *insisted* than at *the* after a transitive verb like *visited* (Juliano & Tanenhaus, 1993).

These two cases illustrate two advantageous properties of the VSG model: (i) it is consistent with the pervasive evidence showing that reading time is inversely correlated with class frequency; (ii) 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 similarities between categories distort the internal structure of the attractor basins associated with the categories. 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). However, it appears that some kind of as-yet-unspecified statistical smoothing (Charniak, 1993) across grammatical classes is required (Tabor et al., 1997). It is also possible that a model which treats reading time as a kind of output error in an SRN (e.g., Christiansen & Chater, in press; MacDonald & Christiansen, 1998) would also 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, as we have noted, such one-dimensional

measures do not encode information about direction of displacement so it is hard to tell, in such models, if similarity is indeed the source of the error.

3. Case Study: Thematic Expectation

The simulations just described only examined pure syntactic contrasts in the sense that the complement requirements of the verbs and the agreement requirements of the determiners were categorical. It is of some interest, then, to investigate how the VSG model performs in a case where the contrast is not categorical in this way. Such cases arise in association with what are generally thought of as "semantic" distinctions. A good example is thematic role assignment. 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 sentence that sounds "semantically strange" or "incongruous" (5).

(5) a. # The customer was served by the jukebox.

b. # The car accused the pedestrian of cheating.

Semantically strange sentences seem to violate our expectations about what is likely to happen in the world but they do not violate our expectations about what can happen in the language in the same way that ungrammatical sentences do. Linguistic theories generally posit two distinct mechanisms for handling semantic and syntactic incongruity. Semantic violation is thought to be detected on the basis of world knowledge, whereas syntactic incongruity results from violating rules of grammar. This set of assumptions is very useful in that it has allowed us to recognize the effectiveness of abstract syntactic mechanisms at organizing linguistic information. Moreover there is psychophysiologi-

cal data from studies using event-related potentials suggesting that the two kinds of violations result in qualitatively different patterns of brain responses (Garnsey, 1993; Osterhout & Holcomb, 1993; Hagoort et al., 1993; Ainsworth-Darnell et al., 1998).

The distinction between syntactic and semantic incongruity is especially interesting from the perspective of connectionist models. Both semantic and syntactic constraints affect the distributional structure of words in the language. This raises the possibility that a connectionist device trained on distributional information could model both classes of constraints. We show that this hypothesis is supported by the VSG model in the following sense: the gravitational mechanism, defined by the representation developed by a connectionist network, exhibits what might best be called a *graded qualitative distinction* (Section 3.3.2) between semantic and syntactic incongruity.

The claim that "semantic" information can be learned by a model which only interacts with corpus data is surprising. 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*, in one common sense of the term. However, we may ask if such a model provides the right "hooks" for interfacing with the world. In this case, corpus-based models may have something useful to contribute. Corpuses contain a good deal of information beyond what the syntax of a language provides. Indeed, Burgess & Lund (1997), Landauer & Dumais (1997) among others have shown that much information which is standardly termed "semantic" can be extracted from a corpus by evaluating co-occurrence statistics. Much of this information is information about which words *tend* to be used in combination with which other words, given that they can be so used. The usual strategy in linguistic modeling is to note that such information reflects properties of the world that can be learned independently of language—indeed some of

it is the kind of knowledge that animals and prelinguistic children seem to have—and to try to simplify the job of the language theory by assuming that it does not incorporate such knowledge. But this may be misguided: since the information about tendencies of usage is available in the speech we hear, it is possible that the "language mechanism" is actually shaped by this usage as well as by abstract grammatical constraints. In fact, a theory which posits such a "world-molded" language mechanism may be better suited to providing a full model of the language-world relationship than one which assumes strict independence, because it has the structures needed for interfacing. On the other hand, there is a potential problem with trying to let the language mechanism encode too much detail about the world: the theory may become too complex or unduly unrestrictive.[7] The dynamical systems framework is a way around the latter pitfall: the details of subtle differences in the "semantic" biases of words are encoded in small differences in position in the metric representation space; but the basin configuration of the system as a whole provides an organizing category structure which is computationally simple.

To explore these issues in a specific case, we examined the role of thematic fit in syntactic ambiguity resolution, focusing on the results of McRae et al. (1998).

3.1 The Phenomenon

McRae et al. (1998) examined the way the thematic properties of a subject and verb influenced readers' biases in favor of a reduced relative versus a main clause reading in sentences like (6a) and (6b).

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

b. The crook / arrested by / the detective / was guilty / of taking / bribes.

McRae et al. performed an offline rating task in which subjects were asked to answer questions such as "How common is it for a cop to arrest someone?" by providing a number on a scale of 1-7 (where 1 corresponds to very uncommon and 7 to very common). On this basis, they grouped "cop" and similarly rated nouns together as "Good Agents" and they grouped "crook" and similarly rated nouns together as "Good Patients". They then studied self-paced reading times in regions like those shown in (6). A summary of their results is graphed in Figure 2. The graph plots a "reduction effect", measured in milliseconds versus sentence region. The reduction effect is the difference between the reading time of sentences like (6) and the corresponding unreduced cases in which "who was" was inserted before the first verb.

———————— insert figure 2 about here ————————

Three properties of the data are worth highlighting. (i) 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. (ii) 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. (iii) 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: (Hypothesis X) the first verb (e.g., "arrested") is the main verb of the sentence, or (Hypothesis Y) 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 set to model fragment completion data using the same materials. The same weights were then used successfully to predict on-line reading times. In the simulation described next we extend this result by showing how the weights can be set via connectionist learning on corpus data resembling the significant distributional properties of the McRae et al. materials. The resulting model then predicts the three phenomena highlighted above: immediate semantic influences, competition-induced slow-downs, and inertia.

3.2 Thematic expectation simulation

3.2.1 The training grammar

The simulation grammar is shown in Table 1. This grammar generates a relatively simple, symmetrical set of strings which share a number of properties with the English sentences of interest. The quoted labels in the grammar and in the following discussion (e.g., "Good Agt", "Good Pat") make this analogy explicit for the sake of giving the reader some familiar labels to use as placeholders. Although the analogy is rough, and we do not intend that the model map precisely onto human behavior, it is designed to

make the central conceptual issues transparent. Such transparency is critical, we believe, for getting past the typical opaqueness of connectionist models.

The grammar in Table 1 is designed so that the first word of each sentence can be classified as belonging to one of two classes, X or 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. The dominance of agentive constructions in English is reflected in the fact that sentences starting with X's outnumber sentences starting with Y's by a ratio of 3:2. Also, as in English, there are initial X's and initial Y's of a range of different frequencies. The second word of each sentence 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 nonabsolute biases of the V, 1, and, 2 words mirror the fact that in natural language, many words can be constituents of many constructions and thus do not 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 a given context (Rohde and Plaut, in press). In the model, the local ambiguity of the words turns out to be essential to the prediction of 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 which 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. Following McRae et al.'s investigation, we will examine a case in which these tendencies are violated.

———— insert table 1 about here ————

3.2.2 Training the network

The grammar was used to generate data for training the network described in Section 2.1. 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. Momentum was not used. The grammar defines ten states (states are dis-

tinct if they induce different distributions over the set of all possible future sequences—Crutchfield, 1994; cf. Hopcroft and 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.[8] 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 determinded 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.

3.2.3 The gravitation mechanism

After some experimentation, 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 two attractors correspond to the X ("Main Clause") reading and the Y ("Reduced

Relative") readings, respectively, in the sense that sentences which had 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".

3.3 Reading Time Results

In analogy with McRae et al.'s study of reduced relatives after Good Agents and Good Patients, 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. However, in the simulation we only had a few relevant cases to measure and there was not much noise so the effect of contrasting initial word biases was evident without such baselining.

A sample result is shown in Figure 3. The dotted line shows gravitation times for the string "yd vc 1d 2d 3d p" ("Crook arrested by detective escaped.") and the solid line shows times for "xc vc 1d 2d 3d p" ("Cop arrested by detective escaped.") The pattern shows the central properties of the human reading time data: (i) 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) (ii) 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), (iii) inertia in the parse choice. (Each spike has a tail which dwindles over the course of several following words).

———————— insert figure 3 about here ————————

These results suggest that the VSG model can reproduce a number of significant features of human reading time patterns when trained on distributional information reflecting certain thematic role induced biases. The next subsection analyzes the model's predictions.

3.3.1 Induction of competition effects

The fact that the gravitation times of the VSG model show a similar pattern to the human data is encouraging. Examining the representations and the processing dynamics of the VSG model reveals that the VSG model is predicting the human data by implementing a competition mechanism very much like Spivey-Knowlton's Normalized Recurrence Algorithm.

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 space.[9] 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")).[10] 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").

———— insert figure 4 about here ————

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 two clusters give rise to two attractors which are at the centers of the circles in the diagram.[11]

Three trajectories are shown. These trajectories correspond to three sentences which start "xc vc...", "yc vc..." and "yf vf..." respectively. The "xc vc..." and "yf vf..." cases, roughly analogous to "cop arrested" and "evidence examined", are the the beginnings of normal sentences which typically give rise to X ("Main Clause") and Y ("Reduced Relative") interpretations respectively. Since their classification status is quite clearcut, the processor lands close to the appropriate attractor when the V word is presented and gravitation takes only two time steps.[12] By contrast, the sentence that starts with "yc vc" (analogous to "crook arrested") has conflicting information in it. The first word 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 the gravitation takes a long while (8 time steps).

———— insert figure 5 about here ————

In Figure 6, we show a close-up of the "1" region of the visitation set. Here we can observe one-word continuations of the sentences shown in Figure 5. The case of central interest is "xc vc 1d". We can think of this case as analogous to a partial sentence like "Cop arrested by...", which starts off with a Good Agent and is followed by an agentive verb, but continues with a "by"-phrase which strongly signals the unexpected Reduced Relative interpretation. In such cases, people showed latencies at the agentive noun phrase in the "by"-phrase which were quite high compared to a control case with a Good Patient subject (e.g., "Crook arrested by detective..."). While the model processes "xc vc" with ease, the subsequent "1d" lands it in an intermediate position and thus gives rise to a very long gravitation time (13 time steps). The corresponding control case, "yc vc 1d", (" Crook arrested by...") produces a nonminimal trajectory at "1d" as well, but the starting point is initially much closer to the "Y" attractor and the gravitation time is correspondingly shorter (5 time steps). Thus, these two strings, when compared at their "V" words and "1" words, produce the crossing latency-values pattern that distinguishes McRae et al.'s data. It is true that the simulation shows the crossing pattern more immediately in response to the disambiguating information than the human subjects appear to (i.e. at the first disambiguating word), but as we noted above, this may be due to the weakness of the parafoveal "by" signal; it may also reflect the more complex ambiguity of natural language "by" which we have noted.

———— insert figure 6 about here ————

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...."). In this case, the first two words strongly favor a "Y" ("Reduced Relative") continuation, while the third word requires an "X"

("Main Clause") continuation. The result is a nonminimal reading time at "1a" but not the superelevated reading time of the focus case, "xc-vc-1d". This intermediate reading time makes sense because high frequency bias-revising information ("1a") is more effective at overcoming the contextual bias than the relatively low frequency bias-revising information ("1d") of the focus case.

These examples indicate that the VSG model mirrors the two-attractor account of McRae et al. in the details of its dynamics. The behavior patterns illustrated are robust in the sense that they recur whenever the same sentences are presented with different preceding contexts, and they persist if 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. Nevertheless, the cases we have focused on here seem most closely analogous to the human subject cases which we are using as a standard. Thus, it appears that given the constraint that the gravitation mechanism needs to form a distinct attractor basin for every syntactically distinct context, the VSG model succeeds in deriving the hypothesized competition mechanism from the distributional properties of its training corpus.

3.3.2 Emergence of a syntax/semantics distinction

We now show that the gravitational component of the VSG model induces a distinction between types of violations which lines up in a plausible way with the distinction between syntactic and semantic violations as judged by human beings. The essence of the induced contrast is that grammatical processing (including the processing of semantically strange sentences) involves gravitation directly into an attractor while ungrammatical processing 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. Often, the saddlepoint delays convergence so long that we can say the processor fails to find an interpretation in "reasonable time".

In order to illustrate this point, we extend the analogy between natural language and our artificial Thematic Bias Grammar. The category sequence of the training grammar is very strict in the sense that 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 produces something analogous to a natural language grammaticality violation. We now describe a simulation in which we compared the VSG model's response to analogs of semantic violation with its response to analogs of syntactic violations.

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."). As we noted, the bias of the first two words toward X 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.

———— insert figure 7 about here ————

Figure 7 provides a suggestion that the VSG model is treating grammatical violations differently from semantic ones: semantic violations involve direct gravitation into an attractor; syntactic violations involve parse-blocking delay by a saddlepoint. To probe the legitimacy of this idea more thoroughly, we studied the model's response to a sample of 20 semantic violations and 20 syntactic violations. These examples were constructed by hand in an effort to test a range of types of conditions.

The model's response to syntactic anomaly turned out to have a fairly characteristic pattern, although the results were not as simple as the single test case described above suggests. Whenever a word of the wrong category occurred at a particular point in a string, the starting point of the trajectory tended to be a compromise between the contextually appropriate attractor and the attractor associated with the anomalous word. The result was that the syntactic anomalies nearly always placed the processor in a "no man's land", 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 which was 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, in this latter kind of case, the word following the anomalous word produced a trajectory that was still trapped behind a saddlepoint at the 30th time step. In this sense, the model sometimes showed delayed sensitivity to an anomaly.[13] 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. Because of the sometimes delayed response to the anomaly, we assessed the outcome of the 20-sentence trials by examining the trajectory for the anomalous word and the word following it and tabulating results for whichever of these trajectories was longer. We applied this method to both the semantically anomalous sentences and the syntactically anomalous strings. Figure 8 plots the velocity profiles of these maximally anomalous trajectories for the two sets of examples. 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 figure makes the contrast between the two sets apparent and a t-test shows a clear difference in the mean maximal gravitation times ($p < 0.0001$, 19 d.f.).

———————— insert figure 8 about here ————————

Figure 8 suggests thinking of the difference between grammatical and ungrammatical strings as a *graded qualitative difference*. At one extreme are the parses which involve short, direct trajectories into an attractor and thus result in very 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 most stoutly grammatical examples show very slight influence of deflection by saddlepoints; the most atrocious grammatical anomalies are very unlikely to land on a stable manifold of a saddlepoint, and thus will eventually gravitate into an attractor. But despite the gradedness of the difference, 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. The framework makes several kinds of testable predictions: (i) people should show gradations of reading times on grammatical and ungrammatical sentences even when they are happy to make binary judgments about them (ii) lexical or other contextual biasing can downgrade a semantic anomaly into an ungrammaticality and vice versa (iii) there can be variation in the location of anomalous latencies with respect to syntactic violations. These predictions differentiate the VSG model from all models that make an absolute distinction between grammatical and ungrammatical sentences as well as form models like the SRN, which treat all contrasts on a grey-scale.

4. Conclusions

This paper has described an application of the Visitation Set Gravitation (VSG) model, which was first described in Tabor et al. (1997), to sentence comprehension phenomena

involving graded differences in lexical information. We focused on the results of McRae et al. (1998), showing, in particular, that the VSG model correctly predicts (i) immediate sensitivity to graded lexical biases, (ii) a general association between elevated reading times and conflict between the parse biases of the previous context and the current word, (iii) inertia effects: that is, the tendency for the processor to resolve a conflict between parse biases gradually, over the course of reading several words, even if the words provide strongly constraining information. The inertia effect is important because it provides new evidence distinguishing the VSG model from closely related models which posit a systematic correlation between the unexpectedness of a word class and its reading time (Jurafsky, 1996). Tabor et al., 1997, described a related phenomenon, smoothing, in which class similarity effects cause reading times to diverge from expectation-based predictions.

The main theoretical insight of the paper, building on Tabor et al., 1997, is that dynamical systems theory provides a useful set of tools for understanding the representational properties of high-dimensional learning models like Elman's Simple Recurrent Network (SRN). We noted, in particular, that 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, which are good for research insight and good for discrete assignment of interpretations. The current results suggest that the VSG model provides an improvement over hierarchical clustering methods of discretizing connectionst representations (e.g., Elman, 1990; Pollack, 1990; Servan-Schreiber, et al., 1991), for these provide no obvious way of picking out a linguistically or statistically relevant subset of a cluster

hierarchy.

Finally, we identified a new case in which a construct of dynamical systems theory is useful in modeling a phenomenon in sentence processing: the contrast between semantic violation and syntactic violation. We found that the processing of grammatical strings (meaning those which could be generated by the training grammar) tended to involve gravitation directly into an attractor, while 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 entirely relativistic representation of an SRN (it rules out no string) onto the intuitively observable contrast between semantic and syntactic violation. It also makes contact with empirical work showing contrasts in brain activity for the two kinds of processing (e.g., Ainsworth-Darnell et al., 1998).

The VSG model has several shortcomings.

First, the link between the SRN and the gravitation mechanism is weak in that we invoke 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 no structural insight at all in virtue of the machine state correspondences. 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. This constrainedness suggests that the architectural assumptions of the model are doing some explanatory work. As we noted in Section 2.2, however, it would be desirable if the value of p could be determined independently of a grammatical oracle. Our current work is investigating this possibility.

Second, we have only analyzed VSG behavior on a very simple formal language. We feel it is useful to do this at first in order to build a foundation. It is desirable, however, 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. Wiles and Elman, (1995), Rodriguez et al. (to appear), and Tabor (1998) provide some insight into this problem by looking at how SRNs and related devices can represent context free grammars. Here, a central question is, How should the learning mechanism generalize from its finite training experience to an infinite-state language?

Third, as an anonymous reviewer emphasized, the quicker processing of semantic violations than grammatical violations in the current simulation is not surprising, given that the model is likely to have seen most semantic violations in training. A first step toward demonstrating that the model exhibits some generalization ability would be to filter small random samples of the 15552 possible sentences from the training data, and then test these examples to see if they behave like grammatical strings. Clearly, however, real semantic anomaly is not randomly distributed across grammatically legitimate combinations: it is associated with the juxtaposition of particular word classes. Thus, to make a more interesting test, we need to design a grammar in which certain classes of words never directly co-occur, although they have a strong higher-order correlation (e.g., "dogs" may never be said to "meow" or "purr" but they "eat", "run", "play", "sleep", etc.—things that "meow"-ing and "purr"-ing individuals commonly do). The question

is whether the gravitation mechanism will be able to appropriately group clusters of clusters into the same attractor basin in such a case.

These challenges are nontrivial, but it is encouraging to note that they are expected consequences of asking the challenging question that motivates the VSG model: How can one get, in a principled way, from the relativistic perspective of a learning model (where we prefer not to assume that anything is impossible) to an absolutist perspective which supports categorical choice-making. It is not obvious that there is any universally right way of taking this step. Nevertheless, simpler architectural assumptions seem desirable. Dynamical systems theory typically starts with a very simple assumption in the form of a class of equations. Many interesting structures emerge. Perhaps, this paper suggests, these structures are a kind of scaffolding via which emergentist cognitivists can hoist themselves out of the sea of relativism.

References

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

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 perceptrons 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.

Bever, T. (1970). The cognitive basis for linguistic structures. In J.R. Hayes (Ed.), *Cognition and the development of language* (pp. 279–362). New York: Wiley.

Burgess, C. & Lund, K. (1997). Modeling parsing constraints with a high-dimensional context space. *Language and Cognitive Processes*, **12(2/3)**, 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. (in press). Toward a connectionist model of recursion in human linguistic performance. *Cognitive Science*.

Christiansen, M.H. & Chater, N. (this issue). Connectionist natural language processing: The state of the art. *Cognitive Science.*

Cottrell, G. W. (1985). *A connectionist approach to word sense disambiguation (Tech Rep. No. 154).* University of Rochester, Dept. of Computer Science, Rochester, NY. Revised version published 1989 in the Pitman Publishers Research Notes in Artificial Intelligence Series.

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.

Elman, J. (1995). Language as a dynamical system. In Port, R. and van Gelder, T., (Eds.) *Mind as Motion: Explorations in the Dynamics of Cognition.* Cambridge, MA.: MIT Press.

Frazier, L. (1988). Sentence Processing: A Tutorial Review. In M. Coltheart (Ed.), *Attention and Performance* (pp. 559–586). Hillsdale, NJ: Lawrence Erlbaum Associates.

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.

von Gompel, R., Pickering, M., and Traxler, M. (1999). Making and revising syntactic analyses: evidence against current contraint-based and two-stage models. Submitted manuscript. Department of Psychology, University of Glasgow. Contact roger@psy.gla.ac.uk.

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.

Haykin, S. S. (1994). *Neural networks: a comprehensive foundation*. MacMillan, New York.

Hopcroft, J.E. & Ullman, J.D. (1979). *Introduction to Automata Theory, Languages, and Computation*. Menlo Park, CA: Addison-Wesley.

Jagota, A., Plate, T., Shastri, L., and Sun, R. (1999). Connectionist symbol processing: Dead or alive? *Neural Computing Surveys*, 1–40. Available at http://www.icsi.berkeley.edu/ jagota/NCS.

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.

Kawamoto, A. H. (1993). Nonlinear dynamics in the resolution of lexical ambiguity: A parallel distributed processing account. *Journal of Memory and Language*, **32**, 474–516.

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.

Lightfoot, D. (1991). *How to Set Paramters: Arguments from Language Change*. Cambridge, MA: MIT Press.

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

MacDonald, M. C. and Christiansen, M. H. (1998). Individual differences without working memory: A reply to Just & Carpenter and Waters & Caplan. Submitted mansucript. Draft version available at http://www-rcf.usc.edu/ mortenc/index.html#publications.

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(5)**, 375–402.

McClelland, J. L., Rumelhart, D. E., and the PDP Research Group (1986). *Parallel Distriubted Processing. Explorations in the Microstructure of Cognition. Volume 2.* Cambridge, MA: MIT Press.

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(1)**, 134–157.

Newmeyer, F. (1986). *Linguistic Theory in America.* Orlando: Academic Press.

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.

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

Perko, L. (1991). *Differential Equations and Dynamical Systems.* New York: Springer-Verlag.

Pineda, F.J. (1995). Recurrent backpropagation networks. In Y. Chauvin & D.E. Rumelhart (Eds.), *Backpropagation: Theory, Architectures, and Applications* (pp. 99–136). Hillsdale, NJ: Lawrence Erlbaum Associates.

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.

Port, R. and van Gelder, T. (Eds.). (1995). *Mind as Motion: Explorations in the Dynamics of Cognition*. Cambridge, MA: MIT Press.

Rodriguez, P., Wiles, J., & Elman, J. (in press). How a recurrent neural network learns to count. *Connection Science*.

Rohde, D. and Plaut, D. (in press). Language acquisition in the absence of explicit negative evidence: How important is starting small? To appear in *Cognition*.

Rumelhart, D., Durbin, R., Golden, R., & Chauvin, Y. (1995). Backpropagation: The basic theory. In *Backpropagation: 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, Volume 1* (pp. 318–362). Cambridge, MA: MIT Press.

Rumelhart, D.E., McClelland, J.L., & the PDP Research Group (1986b). *Parallel Distriubted Processing. Explorations in the Microstructure of Cognition. Volume 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. (1995). Lexical change as nonlinear interpolation. In Moore, J. D. and Lehman, J. F., (Eds.), *Proceedings of the 17th Annual Cognitive Science Conference.* Hillsdale, NJ: Lawrence Erlbaum Associates.

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(2/3)**, 211–271.

Tanenhaus, M.K. & Trueswell, J.C. (1995). Sentence comprehension. In Miller, J. & Eimas, P., (Eds.), *Handbook of Perception and Cognition: Volume 11* (pp. 217–262). San Diego: Academic Press.

Tiňo, P. and Dorffner, G. (1998). Constructing finite-context sources from fractal representations of symbolic sequences. Technical report No. TR-98-18, Austrian Research Institute for Articial Intelligence, Austria, 1998.

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. and 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: Lawrence Erlbaum Associates.

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

Williams, R.J. & Zipser, D. (1995). Gradient-Based Learning algorithms for Recurrent Networks In Y. Chauvin & D.E. Rumelhart (Eds.), *Backpropagation: Theory, Architectures, and Applications* (pp. 99–136). Hillsdale, NJ: Lawrence Erlbaum Associates.

Further Readings

The PDP volumes (McClelland & Rumelhart, 1986, and Rumelhart & McClelland, 1986b) provide an excellent introduction to the use of **connectionist networks** as cognitive models. For a nice visual introduction to **dynamical systems theory**, see the five volumes of Abraham and Shaw (1984). Strogatz (1994) is an enjoyable textbook: it blends theory with colorful examples. Perko (1991) is more rigorous and trenchant. On **dynamical systems in cognition**, see Port and van Gelder (1995) and van Gelder (1999). For good overviews of **recurrent connectionist networks**, see Williams and Zipser (1995) and Pearlmutter (1995). For a helpful explication of recurrent backpropagation networks (RBP) in particular see Haykin (1994). There has been a steady stream of work on **connectionist symbol processing**: Jagota et al. (1999) give a rundown. Tiňo and Dorffner (1998) and Tabor (1998) are recent developments on the question of representation. Elman (1995) motivates the dynamical systems approach to modeling natural language, summarizing the main insights of the two influential papers (Elman 1990, 1991) on the Simple Recurrent Network (SRN). Frazier (1988) is a helpful tutorial on two-stage (or garden path) models of **sentence processing**. Von Gompel, Pickering, and Traxler (1999) provide evidence that many current models of sentence processing, including current dynamical systems models, are missing an important probabilistic element. Gibson (1998) provides substantial evidence for effects of memory load, also an important challenge for dynamical and connectionist models (though see Christiansen and Chater, in press, for helpful insights). For a *tour de force* on the use of dynamical connectionist models for word recognition, see Kawamoto (1993).

Author Note

Thanks to Nick Chater, Morten Christiansen, Garrison Cottrell, William Turkel, Michael Spivey-Knowlton and Gary Dell for helpful comments. The first author was supported by NIH Grant 5 T32 MH19389. The second author was supported by NIH Grant HD 27206. 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.

Footnotes:

[1]See Williams and Peng for a discussion of other approximations to BPTT.

[2]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, k ranges over units that units that unit j sends activation to, and $f'(net_j) = y_j(1 - y_j)$ is the derivative of the fixed sigmoid activation function (and input-to-hidden weights were adjusted only on the basis of the current inputs).

[3]In fact, although some gravitational systems thus defined have fixed points near the centers of clusters, many appear to have chaotic attractors (see Strogatz, 1994). These chaotic attractors behave approximately like fixed points. Since we are only interested in approximations in the models anyway, we treat such attractors as if they are fixed points.

[4]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 $\Delta t = 1$.

[5]See also Selman and Hirst (1985) for a systematic method of setting weights in a Boltzmann Machine parser without learning.

[6]Gibson and Tunstall (1999 and personal communication) argue that all frequency effects in processing are either semantic (and hence outside of grammar) or lexical. They, in fact, provide evidence that the contrast between *that* and *those* after transitive verbs is due to the lexical preference of *that* for being a complementizer independent of syntactic context. They invoke a third, memory-based constraint system (Gibson, 1998) to handle the sentence-initial contrast between determiner *that* and complementizer *that*. The VSG model may well treat the effects after transitive verbs as essentially lexical (further simulation studies are needed). But even if the effect is lexical and expectations are semantic, the point still stands that some principled mechanism for mediating between local biases and contextually derived expectations is needed. The VSG account is appealing in this regard because it handles all of these phenomena with one formalism instead of three and specifies a mediating mechanism: the semi-pliable hidden unit manifold of the neural network (see Tabor, 1995).

[7]See Newmeyer (1986) for an articulation of the viewpoint that the theory of Generative Semantics foundered on such a shoal.

[8]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...".

[9]In the case at hand, the original hidden unit space had 10 dimensions. The first two principal components captured 56 percent of the variance.

[10]To make Figure 4 interpretable, we have circled and labeled the regions corresponding to distinct classes based on our knowledge of the grammar and of which words

correspond to which points.

[11]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.

[12]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.

[13]Note that all the anomalous sentences ended with sequences of words consistent with the anomalous word. This fact, in combination with the strong contextual dominance exhibited by the model makes it unsurprising that the word after an anomalous word was often associated with a long gravitation time: the model was still sticking to its original parse bias at this point. By the time three words had passed, however, the model typically shifted its hypothesis to the new perspective. These effects are thus a more extreme version of the inertia effects that we saw in conjunction with semantic anomalies.

Equation 1:

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

Equation 2:

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

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 → va ("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.

Figure Captions

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

Figure 2. Crossed and smoothed latencies in the main clause/reduced relative ambiguity (after McRae et al., 1998). The "X" sentences began with "Good Agents"; the "O" sentences began with "Good Patients".

Figure 3. Gravitation times for the thematic bias simulation.

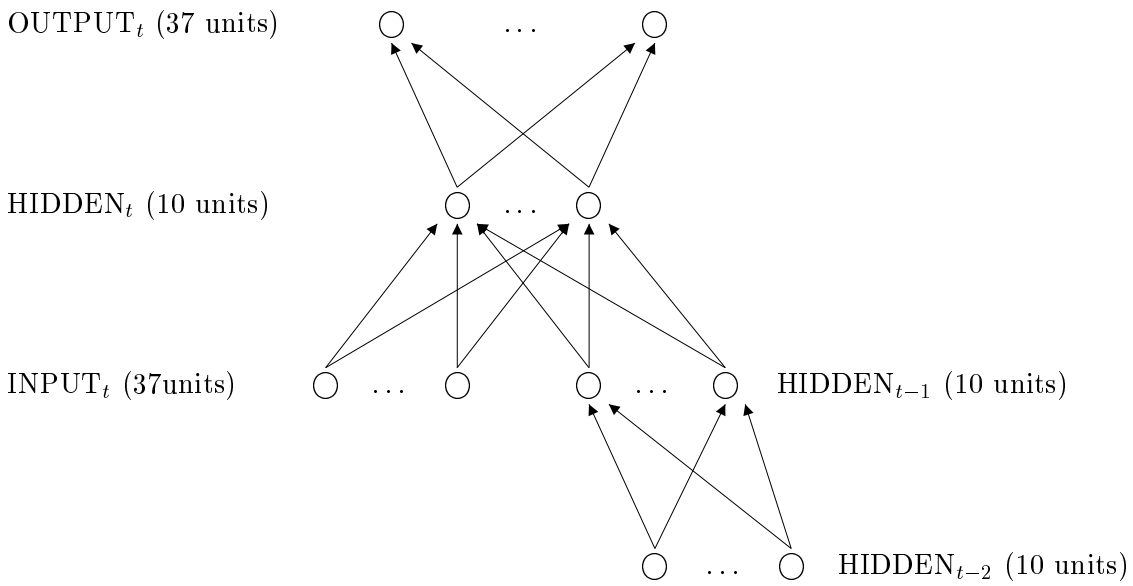
Figure 4. Global view of the visitation set for the thematic bias simulation.

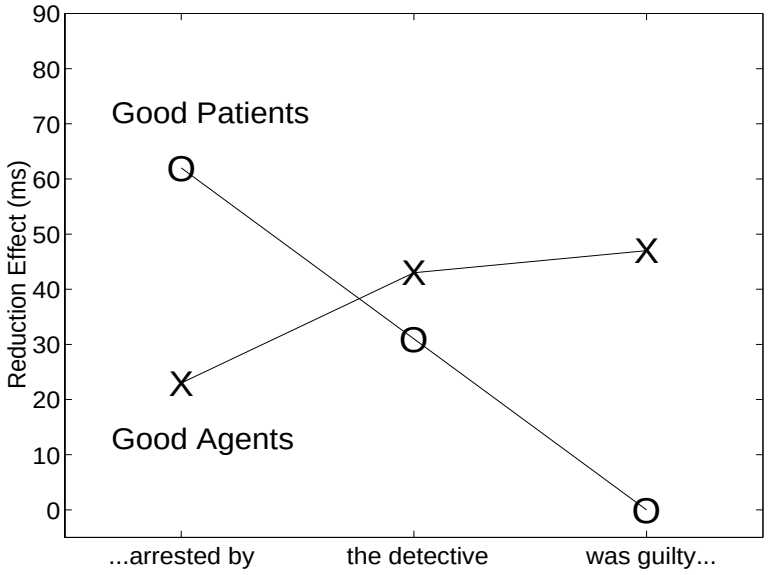
Figure 5. Three trajectories in the "V" region. (The label "yc-vc" identifies the starting point of the trajectory that ensued when "vc" had been presented on the input layer after "yc". The numbers '1', '2', '3', etc. proceeding from this label indicate the trajectory itself. The other labels have similar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

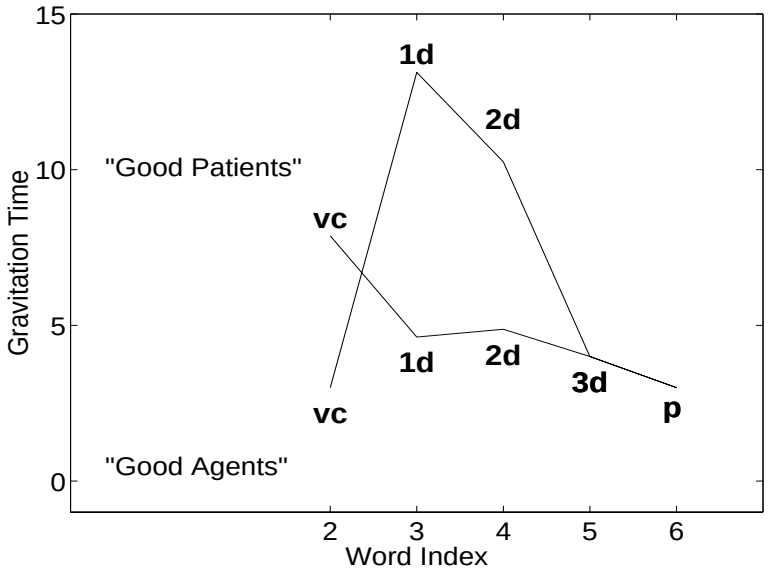
Figure 6. Three trajectories in the "1" region. (See previous figure for explanation of labels.)

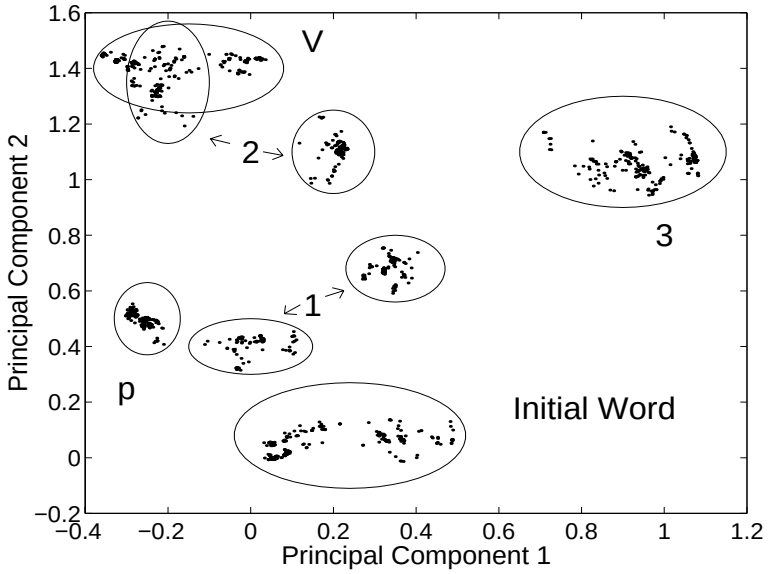
Figure 7. Trajectories for a semantic anomaly (labeled 'vc-1d' and a syntactic anomaly (labeled '1a-p'). The semantic anomaly occurs at the word '1d' in the sentence 'xc-vc-1d-2d-3d-p'. The syntactic anomaly occurs at the word 'p' in the string 'xb-va-1a-p'.

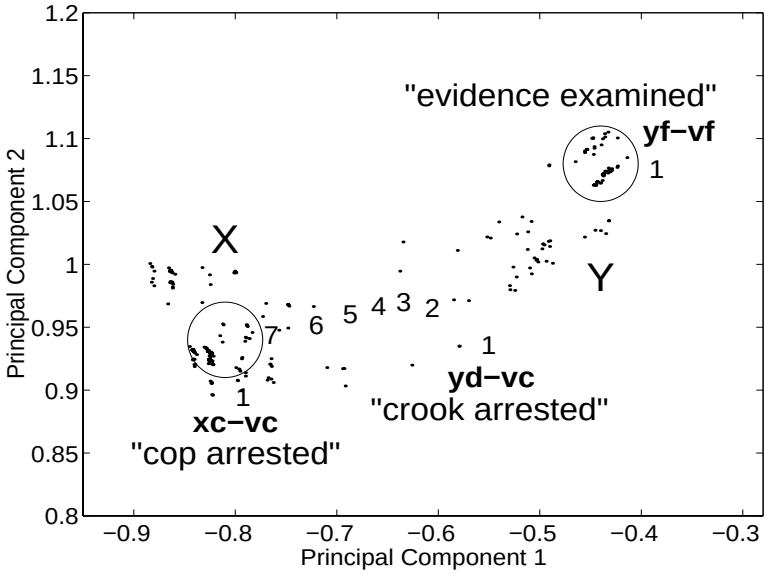
Figure 8. Velocity profiles for 20 semantically and 20 syntactically anomalous transitions. The profile is pictured for either the word at which the anomaly occurred or the word following this word, whichever had a longer gravitation time.

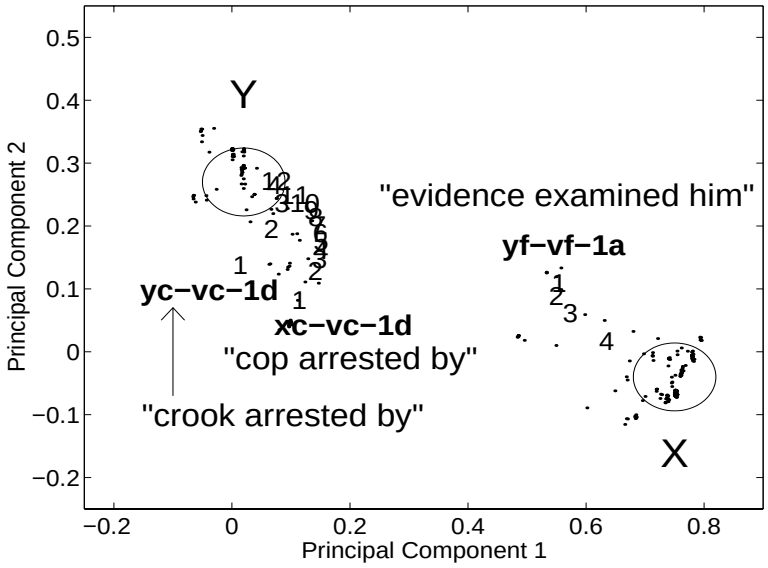

OUTPUT$_t$ (37 units)
HIDDEN$_t$ (10 units)
INPUT$_t$ (37units)
HIDDEN$_{t-1}$ (10 units)
HIDDEN$_{t-2}$ (10 units)

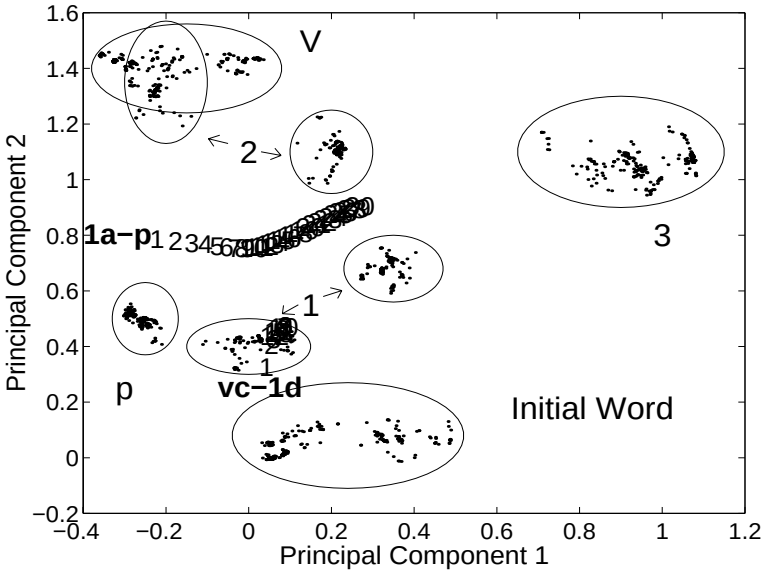

90
80
70
60
50
40
30
20
10
0
Reduction Effect (ms)
Good Patients
Good Agents
...arrested by
the detective
was guilty...


15
10
5
0
Gravitation Time
"Good Patients"
"Good Agents"
1d
2d
vc
1d
2d
3d
vc
p
2
3
4
5
6
Word Index

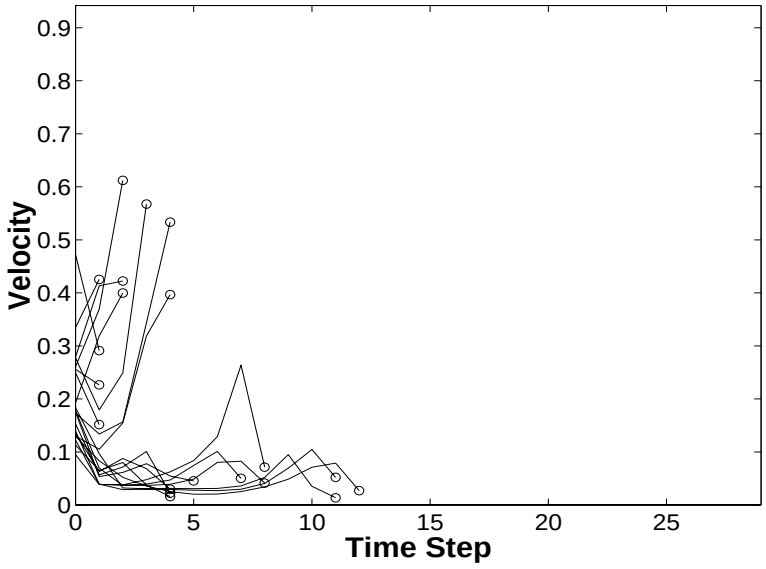

V
2
3
1
p
Initial Word
Principal Component 2
Principal Component 1
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.2
0
0.2
0.4
0.6
0.8
1
1.2


"evidence examined"
yf−vf
1
X
Y
7
6
5
4
3
2
1
1
yd−vc
"crook arrested"
xc−vc
"cop arrested"
Principal Component 2
Principal Component 1
1.2
1.15
1.1
1.05
1
0.95
0.9
0.85
0.8
−0.9
−0.8
−0.7
−0.6
−0.5
−0.4
−0.3


Y
X
"evidence examined him"
yf−vf−1a
yc−vc−1d
xc−vc−1d
"cop arrested by"
"crook arrested by"
Principal Component 2
Principal Component 1
0.5
0.4
0.3
0.2
0.1
0
−0.1
−0.2
−0.2
0
0.2
0.4
0.6
0.8


V
2
3
1a−p
1
p
vc−1d
Initial Word
Principal Component 1
Principal Component 2

Semantically anomalous transitions.



Syntactically anomalous transitions.