Mapping the Syntax/Semantics Coastline

Whitney Tabor
tabor@uconnvm.uconn.edu
Sean Hutchins
sonicmoose@aol.com

Department of Psychology
University of Connecticut
Storrs, CT 06269 USA

Abstract

A number of language processing studies indicate that violations of syntactic constraints are processed differently from violations of semantic constraints (Brain imaging: e.g., Ainsworth-Darnell et al., 1998; Ni et al., in press; Speeded grammaticality judgment: McElree & Griffith, 1995; Eye-tracking: Ni et al., 1998). Although these results are often taken as support for the view that the processor employs two separate modules for enforcing the two classes of constraints, we find (in keeping with Rohde & Plaut, 1999, and Tabor & Tanenhaus, 1999) that a nonmodular connectionist network can learn a quantitative distinction between the two types of constraints. But prior connectionist studies have been inexplicit about why the distinction arises. We argue that it stems from the distinct distributional correlates of the different types of information: syntax involves gross distinctions; semantics involves subtle ones. We also describe the Bramble Net, an attractor network which derives grammatical categories and models an approximation of the syntax/semantics distinction in qualitative terms. These results support Elman's (1990) suggestion that grammatical structures may arise by self-organization, rather than by hardwiring. They also help clarify what the grammatical structures are in a self-organizing connectionist network, and emphasize the usefulness of dynamical systems theory in grammatical explanation.

Introduction

Definition of syntax vs. semantics

By the distinction between syntax and semantics we mean the fundamental one that Chomsky (1957) identified when he contrasted (1a) with (1b).

(1) a. Colorless green ideas sleep furiously.
b. Furiously sleep ideas green colorless.

The modificational relationships between the words in (1b) are not evident to a native English speaker, and one cannot identify any coherent phrasal hierarchy. We thus label (1b) as syntactically anomalous. By contrast, native speakers have no trouble deciding on a parse tree for (1a), but the meanings of the complex phrases are odd and seemingly contradictory. We thus call (1a) semantically anomalous.

By employing some of the basic apparatus of Generative Grammar, we can make a finer characterization of the two types. If we assume that phrases are organized around grammatical heads which select the semantic attributes of their complements, then (1a) can be diagnosed as an amalgam of *selection violations*. *Subcategory errors* involve incorrect selection of an argument-structure constellation, typically of a verb, (e.g., in *Ermin put the book). *Agreement errors* involve inconsistencies between elements that are required to share a common feature like number or gender (e.g., *They eats.) We refer to other mistakes in the sequencing of categories (e.g. *See dog dog) as *category errors*. The last three types are standardly considered syntactic errors.

Evidence for the distinction

Drawing a fundamental distinction between syntax and semantics has several advantages.

First, it is only by factoring out the variation in sentence quality due to semantic contrast that it is possible to discern the simple approximation of the range of a language that its phrase structure rules provide (Chomsky, 1957). These rules receive independent justification from the observation that they permit a compositional treatment of meaning that largely accords with human judgment (Frege, 1892).

Second, several recent language processing studies indicate distinct processing responses to syntactic and semantic anomaly. McElree and Griffith (1995) used a speeded grammaticality judgment task to find out how quickly people could detect syntactic and semantic anomalies. They found that detection of syntactic anomaly (both subcategorization violation and category violation) rose above the level of chance about 100 ms. sooner than detection of semantic anomaly (selection violation). Ni et al. (1998) and Braze et al. (submitted) used an eye-tracker to monitor participants as they read sentences that were semantically (selection violation) and syntactically (agreement violation) anomalous. They found that readers slowed down at both kinds of anomalies, but for syntactic anomalies the distribution of their regressive eye movements spiked abruptly on the anomaly itself or shortly after, while for semantic anomalies it was strongly skewed toward the end of the sentence. Ainsworth-Darnell et al (1998), tied together many previous EEG studies by demonstrating independent responses to the two types of anomalies in individual participants. Ni et al. (in press) showed distinct regions of brain response to the two types using fMRI.

Models

The distinction between syntactic and semantic anomaly seems to be well supported both theoretically and empirically. It is therefore desirable to have a good understanding of how it is instantiated in mental representations. The standard view, coming from Generative Linguistic Theory, assigns separate modules the jobs of checking the two types of anomaly. But this model leaves open the question of how a learner decides whether to attribute an observed distributional systematicity to a syntactic or semantic module. For example, why is "Dogs moo" classified as a semantic anomaly, while "Dogs barks" is a syntactic one?

Connectionist models have exhibited an ability to glean both syntactic and semantic information from text data. Elman (1990, 1991) trained a Simple Recurrent Network or "SRN" on the task of predicting each next word in a simple, English-like corpus. He found that a hierarchical cluster analysis of the trained-network's hidden unit space contained clusters corresponding to both syntactic classes (Noun, Verb, and various transitivity classes of verbs) and semantic classes (Animate, Large, Edible, etc.). Rohde and Plaut (1999) studied a similar simulation and found that the inclusion of semantic-like lexical cooccurrence biases significantly enhanced the ability of the network to learn complex phrase structures. Moreover, the average lowest transition likelihoods in natural grammatical sentences were higher than the average lowest in grammatical but semantically odd (selection violation) sentences, which in turn were higher than the average lowest in ungrammatical sentences (including verb subcategorization, agreement, and other category sequencing violations). Allen & Seidenberg (in press) used a continuously settling recurrent network and included a bidirectional mapping from form to meaning. The resulting fixed point dynamics provided good generalization behavior.

These results indicate that connectionist networks can derive a distinction between syntactic and semantic structure, while encoding both in a common metric space. But the results raise many questions about what syntactic and semantic structure consist of in such self-organizing models. While, the resemblance of network cluster structures to linguistic categories is suggestive and the alignment of graded network properties with category levels (well-formed, semantically odd, ungrammatical) are encouraging, the findings do not provide much insight into why the resemblances hold or what general properties of the networks produce these results. We performed several additional simulations to better understand how connectionist networks represent syntactic and semantic structure.

Simulation 1

Following Elman (1991) and Rohde and Plaut (1999), we employed a SRN with three hidden layers, and recurrent connections only in the middle hidden layer. The 30 input units were clamped on or off, one at a time, with each unit uniquely coding the appearance of a particular word. The hidden units (10 in layer 2, 20 in 3, 10 in 4) had fixed sigmoid activation functions. The target at each point in time was an activation of 1 on the output unit corresponding to the next word in the training sequence. We wanted the outputs to converge on probability distributions over next words, so the output units as a group had the softmax (normalized exponential) activation function. We thus employed the multinomial cost function (Rumelhart et al, 1995) and the delta rule was used to adjust the hidden-to-output weights. The remaining feedforward units were trained using additional backpropagation (Rumelhart, Hinton, & Williams, 1986), and the recurrent connections were trained on the approximation to backpropagation through time (BPTT) in which the gradient is estimated on the basis of only a single previous time step of the hidden units (see Pearlmutter, 1995).

Table 1: The grammar for simulation 1. All productions have equal likelihood of being used. The lexical classes expand to between 1 and 4 individual lexical items.

S	→	N[human] V[eat] N[food] p
S	→	N[human] V[perceive] N[inanimate] p
S	→	N[human] V[destroy] N[breakable] p
S	→	N[human] V[cogitate] p
S	→	N[human] V[perceive] N[human] p
S	→	N[human] V[pursue] N[human] p
S	→	N[human] V[move] N[inanimate] p
S	→	N[human] V[move] p
S	→	N[animate] V[eat] N[food] p
S	→	N[animate] V[perceive] N[animate] p
S	→	N[animate] V[pursue] N[animate] p
S	→	N[animate] V[act-on] N[animate] p
S	→	N[animate] V[move] N[inanimate] p
S	→	N[animate] V[move] p
S	→	N[inanimate] V[move] p
S	→	N[aggressive] V[destroy] N[fragile] p
S	→	N[aggressive] V[eat] N[human] p
S	→	N[aggressive] V[eat] N[animate] p
S	→	N[aggressive] V[eat] N[food] p

We used probabilistic context free rewrite rules to construct a simple grammar similar to the one used by Elman 1990 for training a syntax network (Table 1). The grammar generated only nouns, verbs, and end-of-sentence markers ("periods"). The verbs were either transitive or intransitive. Both the nouns and verbs fell into a number of semantic classes (See Table 1). We defined a selectional violation to be a sentence in which a verb had the right transitivity, but the noun features were not consistent with the grammar (e.g., N[inanimate] V[eat] N[food]). We defined a subcategorization violation to be a sequence in which a strictly intransitive verb took an object, or a strictly transitive verb did not.

The grammar was used to generate strings of words at random. These were strung together end to end and presented to the network one word at a time. The network was trained with a learning rate of 0.01. Momentum was

Table 2: Means of the grammaticality measure. All within-language comparisons are significant ($p < .001$).

Language	Class	N	Mean	SD
SVO	Well-formed	662	-1.56	0.35
SVO	Sel Viol	2002	-4.18	1.08
SVO	Subcat Viol	1098	-5.21	1.27
SOV	Well-formed	662	-1.60	0.34
SOV	Sel Viol	2002	-5.37	1.66
SOV	Subcat Viol	1098	-6.81	0.82

Table 3: Distances to closest grammatical state. All within-language comparisons are significant ($p < .001$).

Language	Class	N	Dist	SD
SVO	Well-formed	662	0.040	0.029
SVO	Sel Viol	2002	0.176	0.206
SVO	Subcat Viol	1098	0.360	0.266
SOV	Well-formed	159	0.020	0.025
SOV	Sel Viol	1000	0.288	0.329
SOV	Subcat Viol	1000	0.625	0.394

not used.

The grammar was used to compute exact target distributions for every juncture between words in the training corpus (see Rohde & Plaut, 1999). The Kullback-Leibler divergence (E) between the network's output and the correct distribution was computed at each word in the training corpus ($E_w = \sum_i t_i \, ln \, t_i/o_i$ where t_i is the target for unit i and o_i is its output on word w). Training was stopped when the cumulative divergence error over a large sample of patterns was consistently small enough that we could conclude that the network was not conflating any of the target distributions with one another (approximately 1 million word presentations).

Rohde & Plaut (1999) studied a measure of sentence goodness based on the network's output predictions. They found that the mean goodness (log of the product of the two lowest output activation transitions) of normal grammatical sentences was higher than that of selection violation sentences, and the selection violation sentences, in turn, had a higher mean than syntactic violation sentences. Because our sentences were much shorter than theirs, we used a simplified version of their goodness measure (log of the single worst transition) and tested it on well-formed sentences, selection violations, and subcategorization violations. We also found a clear stratification (See the "SVO" rows in Table 2).

One of the consequences of defining syntactic category descriptions independently of semantic classifications is that category order is expected to be able to vary independently of the contrast between semantic and syntactic violation. Generative theory thus predicts that the distinction between selection and subcategorization will persevere across languages with different fundamental word orders. To see if the network made a similar separation, we tested it on the output of a grammar exactly like Grammar 1 except that the order of constituents was systematically Subject (Object) Verb (SOV) rather than Subject Verb (Object) (SVO). Indeed a similar relationship between goodness values obtained in the SOV case (Table 2).

A disadvantage of Rohde and Plaut's goodness measure is that it does not explicitly characterize the effects on processing of making a low-probability transition. The experiments of Ni et al. (1998) and Braze et al. (submitted) indicate that people react to the anomaly of a sentence at or after the anomalous word or words (in Rohde and Plaut's terms, after they have made a low-probability transition). We studied the response of the network to anomalies by examining the hidden unit representations. To do this, we presented a long sequence (2000 words) of grammar-generated words to the network and recorded the hidden unit states associated with each word. Tabor et al. (1997) called this kind of sample a *Visitation Set*. We then tested the network on ill-formed sentences by finding the hidden unit location visited following the transition with the lowest output activation over the course of the sentence (the *low-point*). Table 3 shows the mean distance in hidden unit space between the low-point and the nearest point in the Visitation set for samples of selection violation sentences and subcategorization violation sentences. For comparison, a new random sample of grammatical sentences was also tested against the visitation set.

The minimum distance measure parallels Rohde and Plaut's grammaticality measure, and points to a useful way of characterizing the effect of anomaly on the network: there is a subset of the hidden unit space that the network sticks to during grammatical processing. This subset is approximated by the Visitation Set. Selection violations throw the network off the track somewhat. Syntactic violations throw it off more substantially.

This geometrical contrast between the anomaly types has a simple explanation in terms of the distributional distinction between selection and subcategorization. Subcategorization refers to more abstract classes than selection. Thus more instances of training are involved in the development of subcategorization contrasts than in the development of selection contrasts, and subcategorization distinctions produce larger separations in hidden unit space. Violations are cases where the information provided by the current word clashes with the information provided by the preceding context. The network responds to such clashes by averaging the conflicting signals. In the case of selection violation, this averaging interpolates between nearby structures. In the case of syntactic violation, the averaging interpolates between widely separated, major clusters. As a result, syntactic violations tend to result in greater displacement from familiar territory. We hypothesize that the empirical results of McElree & Griffith (1995), Ni et al. (1998), and Braze et al. (1999), which found syntactic violations more readily detected than semantic, stem from this contrast: wildly divergent states are easier to distinguish from normal states than slightly divergent ones.

Simulation 2

Samples of geometric relationships in the SRN's hidden unit space do not make it clear what the network's total generalization behavior is, nor whether its coverage of a language can match that of symbolic phrase structure rules. Nor do relative distance measures alone explain the eye-tracking and brain-imaging results indicating qualitatively distinct responses to semantic and syntactic anomaly. Our previous work on sentence processing (Tabor et al, 1997; Tabor & Tanenhaus, 1999) suggests that the study of dynamical settling networks can clarify the structural principles underlying connectionist sequence-learning. We designed the Bramble Network (BRN) to explore this hypothesis. The BRN is similar to the simple version of the SRN that has one input layer, one recurrently connected hidden layer, and one output layer. But the BRN has two sets of recurrent connections in the hidden layer. One set, the discrete weights, works like the recurrent connections in the SRN, changing the hidden activations discretely every time a new word is read. The other set, the continuous weights, undergoes continuous settling according to Equation (1).

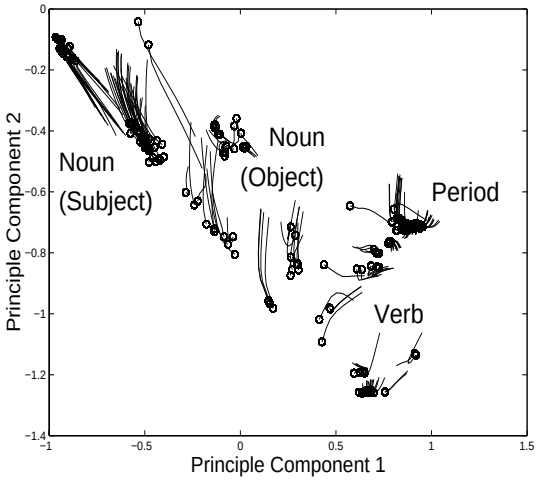
$$\frac{dv_i}{dt} = net_i - v_i \tag{1}$$

where v_i = unit state, $net_i = b_i + \sum_j w_{ij}\sigma(v_j)$, b_i = unit bias, w_{ij} =weight from j to i, and $\sigma(x) = tanh(x)$.

In the BRN, the input and context units are updated first. Then the input-to-hidden weights and the discrete hidden-to-hidden weights are used to compute an initial state of the hidden units. Continuous settling is carried out via the continuous weights among the hidden units. Finally, the hidden-to-output weights map the final state of the hidden units to the output.

The discrete weights in the BRN are updated just as in the SRN. We also assume that settling only occurs for brief periods of time (1 cycle) before the discrete weights are updated. This makes it easier for the network to discover dependencies across words. The continuous weights are updated according to a principle of stability maximization. That is, for continuous weights, we define the error on unit i as $E_i = (dv_i/dt)^2$ so that $dE_i/dw_{ij} = 2\sigma(v_j)(net_i - v_i)$. This equation says: change the weights in the direction that minimizes the magnitude of recent activation change. Continuous weight learning is applied only when the network has almost converged to a stable state. It thus moves the stable state in the direction of the initial state, causing bifurcations when widely separated initial states are associated with a single attractor. The overall effect is that the attractors of the continuous weights tend to track the centers of masses of clusters defined by the discrete weights (cf. Tabor, Juliano, & Tanenhaus, 1997). We found it most effective to train the network with a mixture of fast (1 cycle) discrete weight training and slow (approximating convergence) continuous weight training. A similar result was produced more quickly when we did all the discrete training first and then followed it with the continuous training. The simulation we report below used this batch technique.

Figure 1: Principal component projection of the visitation set for the Simulation 2 network.



As in Simulation 1, the network was trained on output from Grammar 1. In this case, we trained it directly on the output of the grammar for 200,000 words of discrete training (learning rate = 0.002, momentum = 0.9) and then 120,000 words of continuous training (learning rate = 0.05, no momentum). At this point, both discrete and continuous training had successfully distinguished the states of the grammar.

To gain insight into the organization of the trained BRN's processing, we saved the trajectories associated with a random sample of 200 words in sequence from Grammar 1. We performed Principal Component Analysis (Jolliffe, 1986) on this set of points in order to make the structure visible. The trajectories are graphed in Figure 1. (The two principal components shown account for 87% of the variance). Note that there are regions corresponding the major lexical classes (Noun, Verb, and Period). There are also discernible subclusters within the lexical classes. These correspond to both syntactic (e.g. Subject versus Object, Transitive vs. Intransitive) and semantic (e.g. Big vs. Small, Edible vs. Inedible) classes as well as some clusters whose determinants we have not yet ascertained.

We tested the network on the same sets of good and anomalous sentences that were used in Simulation 1. We defined convergence times for the network by using Euler integration to compute trajectories with $\Delta t = 0.05$, and stopping a trajectory when the distance between successive points on the trajectory passed below a threshold (0.005) or when a maximum of 200 steps was reached. The number of steps in the trajectory was taken as a model of reading difficulty. Table 4 shows mean convergence times for several string classes of interest.

When we designed this model, we expected convergence times to provide a good model of human reading times. This prediction is partially sustained in the contrast between normal sentences in their most familiar sequence (71.43) and selection violations (84.52), for much processing evidence supports the claim that readers slow

Table 4: Mean convergence times (MCT) for Simulation 2. All comparisons significant with $p < .001$ except between selection violations and the sample from all well-formed sentences.

Class	N	MCT
Well-formed (Randomly generated by grammar)	265	71.43
Well-formed (Randomly sampled from list of all well-formed strings)	220	83.69
Sel Viol	250	84.52
Subcat Viol	274	122.85
Syntactic Viol	251	155.13

Figure 2: The trajectories the network follows upon processing selection violations (solid lines) against a background of normal processing (dotted lines).

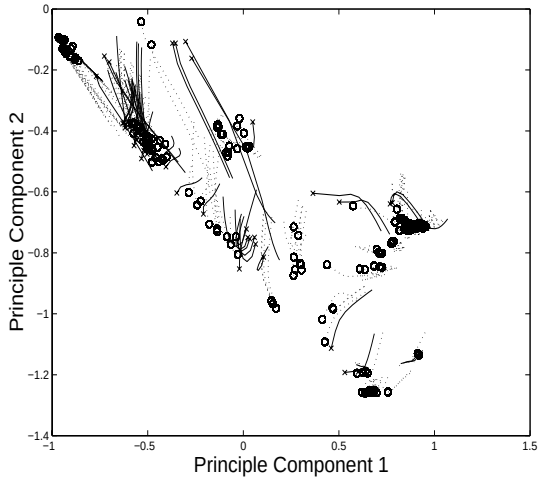


Figure 3: The trajectories the network follows upon processing subcategorization violations (solid lines).

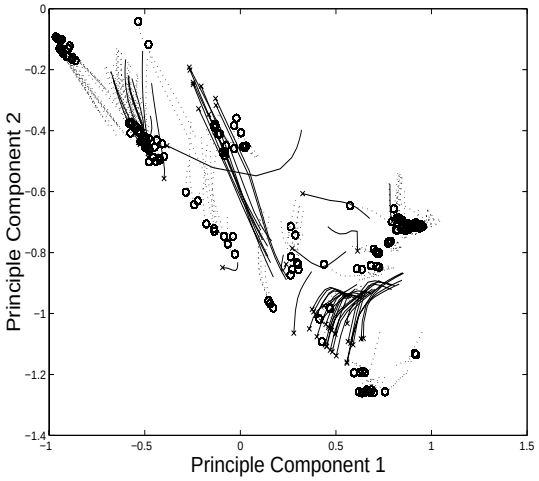
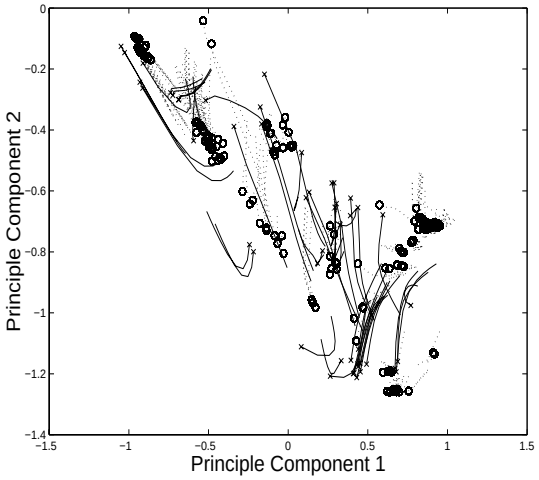


Figure 4: Figure 7. The trajectories the network follows upon processing category violations (solid lines).



down when they encounter less familiar sequences (see Jurafsky, 1996). In a loose sense, the model's very high reading times for syntactic anomalies are also consistent with empirical evidence, for Ni et al. (1998) and Braze et al. (submitted) found readers making substantial regressive eye movements at syntactic anomalies, which implies that they take quite a long time to read past the anomalies. However, it is not clear whether the BRN can predict the McElree and Griffith results showing fast detection of syntactic anomalies. It needs to be able to tell quickly when it's not in a familiar attractor basin. We leave this as a question for future work.

Figures 2–4 show a sample of selection violations, subcategorization violations, and category violations (trajectories end on the x's) against the background of normal processing (end on the o's). The sample of anomalous events was generated by picking the longest trajectory in each sentence. These graphs reveal an interesting structure around which the computation is organized. There appears to be a stable connected manifold (continuous structure that attracts nearby trajectories) running from the upper left of the figure to near the lower right.

There also appear to be pieces of connected manifolds extending to the various other regions where normal processing trajectories end. Perhaps the combination of these manifolds is the locus of grammatical processing. Even semantically anomalous transitions and subcategorization anomalies land by and large on this manifold, though the anomalous cases tend to land on different parts from the normal cases. By contrast, the category violations generally lead to attractors that are separate from the manifold. This suggests that the highly relativistic network model does make a qualitative distinction between types of sentences, and its distinction lines up approximately with current notions of syntactic vs. semantic structure. It is true that the dividing line seems to be different from that of standard linguistic theory, for it is between subcategorization and category error,

rather than between selection and subcategorization error. This difference may stem from a difference between our training grammar and natural language: in natural language, subcategorization constraints are generalizations over more populous classes of items than they are in Grammar 1.

Conclusions

These graphical results suggest an interesting possibility: the skeleton of a language may be a connected manifold in a dynamical system. Such a finding would be appealing because a connected manifold contains an infinity of points, more than we could ever observe. Thus, identifying such a skeleton could be a way of characterizing one aspect of the unbounded nature of linguistic generalization. Such an insight would be similar to the sort of insight that Generative Theory strives for when it posits a phrase structure or transformational architecture. The trouble with current Generative models, however, is that the steps leading to their creation are very controversial (witness the plethora of current syntactic theories), the data themselves are controversial (note the disagreement about grammaticality judgments), and much of the decision-making that goes into building models of specific parses is not made explicit (note the paucity of implemented parsers that employ modern syntactic theory). The dynamical connectionist approach may be an effective alternative, for it is based on a relatively uncontroversial mathematical theory, it uses performance data rather than competence data and thus does not depend on grammaticality judgments, and the process of choosing a parse is explicit. Moreover, unlike the natural language parsers that have been implemented for practical application, the connectionist theory makes contact with fundamental questions about the principles that underlie linguistic representation.

Acknowledgments

This work was supported by University of Connecticut Research Foundation Grant # 477138.

Ainsworth-Darnell, K., Shulman, H. and 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.

Allen, J. & Seidenberg, M.S. (in press). The emergence of grammaticality in connectionist networks. In B. Macwhinney (Ed.), *Emergentist Approaches to Language.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Braze, D., Shankweiler, D., Ni, W., & Palumbo, L.C. (1999). Readers' eye movements distinguish anomalies of form and content. Manuscript, Department of Psychology, University of Connecticut.

Chomsky, N. (1957). *Syntactic Structures.* The Hague: Mouton and Co.,

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.

Frege, G. (1892). ber Sinn und Bedeutung. Zeitschrift fr Philosophie und philosophische Kritik 100, 25-50.

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

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.

Ni, W., Constable, R.T., Mencl, W.E., Pugh, K.R., Fulbright, R.K., Shaywitz, S.E., Shaywitz, B.A., & Gore, J.C. (in press) An Event-related Neuroimaging Study Distinguishing From and Content in Sentence Processing. *Journal of Cognitive Neuroscience.*

Ni, W., Fodor, J.D., Crain, S., & Shankweiler, D. (1998). Anomaly detection: eye movement patterns. *Journal of Psycholinguistic Research, 27*(5), 515-539.

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

Rohde, D.L.T. & Plaut, D.C. (1999). Language acquisition in the absence of explicit negative evidence: How important is starting small? *Cognition, 72*, 67-109.

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

Rumelhart, D.E., Hinton, G.E., & Williams, R.J. (1986). Learning Internal Representations by Error Propagation. In *Parallel Distributed Processing, Volume I* (pp. 318-362). Cambridge, Massachusetts: MIT Press.

Tabor, W. & Tanenhaus, M. K. (1999). Dynamical Models of Sentence Processing. *Cognitive Science, 23*(4), 491-515.

Tabor, W., Juliano, C., and 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.