To Appear in Macwhinney, B. (ed.) Emergentist Approaches to Language: proceedings of the 28th Carnegie symposium on cognition. Erlbaum.

The Emergence of Grammaticality in Connectionist Networks

Joseph Allen
Department of Linguistics
University of Southern California

Mark S. Seidenberg
Neuroscience Program
University of Southern California

Introduction

Linguistic theory in the generative tradition is based on a small number of simple but important observations about human languages and how they are acquired. First, the structure of language is extremely complex– so complex that it is often argued that it would be impossible to learn without prior knowledge as to its general character (Chomsky, 1965). Second, children learn languages rapidly and seemingly effortlessly. Although clearly limited with respect to other sorts of cognitive tasks, every normal child raised under normal circumstances learns the basic syntax of language within a few years of birth. Third, the world's languages exhibit structural commonalities–so-called linguistic universals. Together, these observations have led many researchers to the conclusion that language involves domain-specific forms of knowledge that are largely innate. In the generative approach, the faculty of mind dedicated to language is called *linguistic competence*. A generative grammar is a formal description of this faculty, in the form of a system that generates the set of possible sentences of a given language, and thereby bestows on its possessor the ability to distinguish between grammatical and ungrammatical utterances. Grammars developed within this tradition (which we will call the standard approach) typically consist of primitives, operations, and principles intended to describe the knowledge of an idealized speaker/hearer in a homogeneous speech community. In this approach cognitive representations are hierarchically structured sets of symbols and cognitive processes are operations on them.

Although the standard approach has been very successful in promoting the discovery of descriptive generalizations about linguistic structure and variation, it presents several problems when considered as the basis for a theory of how language is acquired and used. These problems arise from the competence-performance distinction that is one of the foundational assumptions of the approach. The distinction between what people know about language and what they do with that

This research was supported by grant MH PO1-47566 from the National Institute of Mental Health and a Research Scientist Development Award to MSS (NIMH KO2-01188), which we gratefully acknowledge. We thank Maryellen MacDonald and David Plaut for comments on an earlier draft.

knowledge is easy to recognize. However, the relationship between competence *grammars* and performance is more complex.

One issue concerns the systematic ambiguity in the field regarding the extent to which competence grammar should figure in accounts of performance. Chomsky has often suggested that competence grammars describe procedures for relating different levels of representation, but are not characterizations of the computations involved in using language. In Chomsky (1995) he reiterated this view: “The ordering of operations [in grammatical theory] is abstract, expressing postulated properties of the language faculty of the brain, with no temporal interpretation implied.” However, many researchers have pursued a more literal-minded interpretation of grammar as the basis for accounts of how language is acquired, used, or impaired as a consequence of brain injury. In acquisition, a well known example is the work of Borer and Wexler (1992), in which acquisition phenomena are characterized in terms of the maturation of principles ascribed to Universal Grammar such as the bi-uniqueness relations and A-bar chains. Within this approach acquisition is characterized as movement along a trajectory from not knowing to knowing rules of grammar (Gold, 1967). In the area of language processing, Frazier and Fodor (1978) developed a theory of parsing based on heuristics applying to grammatical representations developed within generative theory. In neurolinguistics, Grodzinsky (1995) argues for an account of agrammatic aphasia in which patients fail to represent traces, a particular aspect of grammatical theory. Uncertainty about the relationship between competence grammar and performance has existed throughout the history of generative linguistics (see Fodor, Bever, & Garret, 1974; Berwick & Weinberg, 1984; Bresnan, 1978).

A second problem created by the competence-performance distinction is that it motivates disregarding data that may actually be crucial to understanding basic characteristics of language. The competence approach excludes performance mishaps such as false starts, hesitations, and errors, but also more central aspects of linguistic performance. It is assumed, for example, that language should be characterized independently of the perceptual and motor systems employed in language use; memory capacities that limit the complexity of utterances that can be produced or understood; and reasoning capacities used in comprehending text or discourse. The competence theory also systematically excludes information about statistical and probabilistic aspects of language; the fact that “that” is used more often than “than,” for example, or that the word “the” is followed more often by a noun than a verb are not seen as relevant to this deeper characterization of linguistic knowledge. However, recent studies have emphasized the important roles these aspects of language and cognition play in acquisition and processing (MacDonald, Pearlmutter, & Seidenberg, 1994; Trueswell, Tanenhaus, & Kello, 1993; Kelly, 1992; Saffran, Aslin, & Newport, 1996).

On the standard view the child is learning a rule system that governs the relationships among abstract linguistic entities. The child attends to the structure of utterances guided by innate capacities in order to set language specific parameters. Poverty of the stimulus arguments are used to suggest that knowledge of language is underdetermined by evidence available to language learners and must therefore be attributable to innate Universal Grammar (Chomsky, 1981). Statistical and probabilistic properties of the input are presumed to play no role in this process and their influence is excluded from generative accounts of acquisition, suggesting that children ignore these aspects of input. Excluding the use of these factors from theories of acquistion is seen as positive, in that it avoids a possible combinatorial explosion of intercorrelations among linguistic properties that would make acquisition difficult. The fallacy in this argument is that the statistical and probabilistic aspects of language might actually facilitate acquisition. Allowing that children attend to all aspects of linguistic input–even speech errors–is not a problem because the low frequency of particular errors means that they will not seriously influence the statistical model of language developed in the course of learning.

The apparent complexity of language and its uniqueness vis a vis other aspects of cognition, which are taken as major discoveries of the standard approach, may derive in part from the fact that these “performance” factors are not available to enter into explanations of linguistic structure. If in fact the properties of the language faculty are to some extent determined by a combination of general neural information processing procedures applied to the unique types of tasks that language processing represents, then an approach to the characterization of the language faculty that excludes reference to these factors runs the risk of mischaracterizing the nature of linguistic cognition in a fundamental way.

A third issue concerns the role of performance data in deriving the competence theory itself. The mapping between competence grammar and performance is at best complex, as we have noted; it is also largely unknown. A problem arises because the primary data on which the standard approach relies–grammaticality judgments–are themselves performance data (e.g. Bever, 1972). The methodology of the standard approach holds that properties of the hypothesized language faculty can be identified on the basis of experts’ intuitive judgments of the well-formedness of utterances. However, the relationship between grammaticality judgment and the structure of the grammar is no more transparent than between other aspects of competence and performance.

On the standard view, a grammatical sentence is one that is generated by the competence grammar. This definition entails that every sentence is either grammatical (generated by the grammar) or not. The metaphor here is that of a Turing machine that recognizes some strings but not others as members of a language. A grammaticality judgment, in contrast, is a particular way of querying one’s grammatical knowledge. Among non-experts (i.e., non-linguists), performance on this task is affected by the memory limitations, distractions, shifts of attention and interest, errors, false starts and hesitations characteristic of other aspects of performance. For these informants, linguistic competence is only one factor in the judgment process.

For linguists, using grammaticality judgments to infer properties of the underlying computational system can only be justified if they are able to abstract away from these “grammatically irrelevant” distractions. The notion that linguists are partly in the business of looking beyond actual behavior (determined by a mix of performance and competence) to discover true underlying competence is suggested by Grimshaw and Rosen (1990), who argue against equating subjects’ performance on a judgment task with grammatical knowledge: “To determine properties of the underlying system requires inferential reasoning, sometimes of a highly abstract sort.” (p. 188). Linguists assume that they are capable of reasoning from intuitions about grammaticality to underlying competence. This type of reasoning obviously requires awareness of the types of factors that influence grammaticality judgments. The problem with this logic is that no general theory of how grammaticality judgments are made has ever been proposed. Considering the enormous number of performance factors that have been identified as potentially influencing the judgment process, and how poorly they are understood, it is not surprising that a careful review of the evidence leads Schutze (1996) to conclude that “it is hard to dispute the general conclusion that metalinguistic behavior is not a direct reflection of linguistic competence.”[1]

[1] Degrees of ungrammaticality have long been recognized in the standard linguistics literature (e.g. Chomsky, 1961, 1965). One way of reconciling the paradox entailed by a competence grammar with graded judgment data is to have the grammar assign degrees of badness to strings that violate grammatical principles. Another is to say that the judgment process itself results in graded judgments because it includes non-syntactic information (Bever, 1972). A third possibility

Given the three issues we have noted–the uncertainties about whether competence grammar should figure in accounts of performance, the exclusion of data concerning statistical and probabilistic aspects of language, and the difficulty involved in "reverse-engineering" grammar from performance data–it is quite possible that the formalisms of the standard approach really are only metaphorically related to the brain processes involved in producing and comprehending language. This alternative is sometimes recognized in the literature, but rarely taken seriously. For example, Schutze (1996) concedes that

> It is conceivable that competence in this sense of a statically represented knowledge does not exist. It could be that a given string is generated or its status computed when necessary, and that the demands of the particular situation determine how the computation is carried out, e.g., by some sort of comparison to prototypical sentence structure stored in memory. Since such a scenario would demand a major rethinking of the goals of the field of linguistics, I will not deal with it further.

The remainder of this article represents a step toward just such a rethinking of the linguistic endeavor.

An Alternative Framework

In recent years, a framework has begun to develop that differs significantly from the standard approach with respect to what it means to know a language. The goal of this work is not to devise primitives and principles that describe the set of sentences an idealized speaker/hearer would accept, but rather to make explicit the experiential and constitutional factors that account for the development of knowledge structures underlying linguistic performance. Whereas the standard approach is committed to the uniqueness of linguistic representations vis a vis other cognitive domains, and to the existence of representations whose fundamental character is shaped by the repertoire of innate ideas, the alternative view sees cognitive representations as one component of a system that includes both the organism and its environment. Cognitive processes are taken to be the manipulation of representations such that the organism is able to interact successfully with its environment (van Gelder, in press). Linguistic representations emerge as a function of the interplay among several factors, including the physical components of the human brain that are active during language processing (and their characteristic manner of processing information), the tasks such components are engaged in, and characteristics of the language signals to which they are exposed, particularly their statistical aspects. This view has arisen contemporaneously with and partly as a consequence of connectionism, which has provided novel views of both the nature of mental representation and the ways in which such representations might be learned.

A consequence of this move away from a commitment to the uniqueness of linguistic representations is a renewed interest in the possibility of relating factors typically considered nonlinguistic to linguistic regularities. For example, regularities in the sound system could be seen as arising out of a complex set of conspiracies and compromises among factors affecting production such as the shape of the articulators; constraints arising from the serial nature of language; and

is that constructions are underlyingly grammatical to a degree (Lakeoff, 1973). The existence of all of these possibilities simultaneously makes distinguishing the effects of grammatical knowledge on judgment data from those of processing difficult (Clark & Haviland, 1974), and the assumption that it is possible to determine the properties of an underlying grammar from judgment data even more problematic.

efficiency, i.e., the need to minimize effort expended while simultaneously remaining as communicative as possible. This perspective is beginning to be applied most productively to phonology (e.g. Maddiesson, 1997; Browman & Goldstein, 1989), and has the potential for being applied productively to other aspects of language behavior. Similarly, because there is no a priori commitment to describing knowledge of language in terms of formal primitives, functional considerations are not excluded from entering into explanations of what knowledge of language consists of (Bates & MacWhinney, 1982).

The alternative framework also entails a different view of the nature of language acquisition (Allen, 1997b; Seidenberg, Allen, & Christiansen, 1997; Seidenberg, 1997). On the standard approach, to know a language is to know the rules that define a computational system that generates the set of sentences in that language. It follows that to learn a language is to learn the rules of this computational system. The child's task is to identify the grammar (the rule set) that characterizes the target language. This identification paradigm has played a central role in linguistic theories of acquisition (Gold, 1967; Wexler & Hamburger, 1973; Wexler & Cullicover, 1980).

We view the task of learning a language differently. The task that children are engaged in is learning to use language. In the course of mastering this task, they develop various types of knowledge representations that allow communication to proceed. These knowledge representations are shaped by many factors, including non-linguistic ones, which should, on our view, provide the primitives of a theory of linguistic knowledge. The primary function of this knowledge is producing and comprehending utterances, whether grammatical or otherwise. A by-product of this knowledge is the capacity to distinguish grammatical from ungrammatical sentences.

As an analogy, consider the problem of learning to read. The beginning reader's problem is to learn how to read words. There are various models of how the knowledge relevant to this task is acquired (e.g. Coltheart, Curtis, Atkins, & Haller, 1993; Seidenberg & McClelland, 1989). Once acquired this knowledge can be used to perform many other tasks, including the many tasks that psychologists have used in studying language and cognition. One such task is lexical decision: judging whether a stimulus is a word or not. Even young readers can reliably determine that BOOK is a word but NUST is not. Note, however, that the task confronting the beginning reader is not learning to make lexical decisions. By the same token, the task confronting the language learner is not learning to distinguish well- and ill-formed utterances. In both cases, knowledge that is acquired for other purposes can eventually be used to perform these secondary (metalinguistic) tasks. Such tasks may provide a useful way of assessing peoples' knowledge but should not be construed as the goal of acquisition.

This perspective shares with Chomsky the view that the competence grammar is only metaphorically related to acquisition and processing. However, on our view it is also only indirectly related to the knowledge that underlies these and other aspects of language use. Knowledge of language is construed as one or more neural networks that are engaged in producing and comprehending utterances. Grammars represent high-level, idealized descriptions of the behavior of these networks that abstract away from the computational principles that actually govern their behavior. Grammatical theory has enormous utility as a framework for discovering and framing descriptive generalizations about languages and performing comparisons across languages, but it does not provide an accurate representation of the way knowledge of language is represented in the mind of the language-user.

Grammaticality Judgments

The approach that we have briefly summarized is beginning to be applied to a range of questions about acquisition, processing, and breakdown following brain injury (Plaut, McClelland, Seidenberg, & Patterson, 1995; MacDonald et al., 1994; Seidenberg, 1997). Here we want to return to the concept of grammaticality and to the task of making grammaticality judgments, both of which are central to the standard approach. We have suggested that knowledge of language is not a set of rules for generating sentences and that the child's task is not grammar acquisition. We therefore owe an account of how it is that people can nonetheless make grammaticality judgments.

The capacity to make these judgments emerges out of the ability to process language normally. The task requires informants to establish criteria for deciding whether to call a sentence grammatical or ungrammatical. One important property of the task is that different decision criteria may be used depending on the properties of the sentences being judged. Thus, judging the utterance **the the the the the* as ungrammatical may not rely on the same information as judging **the boy tried Bill to go* or **the boy fell the chair*. The first sentence can be judged on the basis of whether it even potentially conforms to an interpretable object; superficial properties such as the absence of nouns and verbs provide a reliable basis for deciding that it does not. These criteria will be not sufficient for the second and third sentences, which require using other sorts of information. A second point is that for a broad range of sentence types, judgments may be reliably cued by local statistical information concerning subsequences of words. For example, recognizing that *The boy fell the chair* is an instance of the pattern *NP fell NP* may be sufficient to make a decision based on properties of the verb FALL. People can obviously make use of information derived from a complete analysis of the utterance if it is required for further processing (as in formulating a correction, for example), but this level of analysis is not required in many cases.

One reason to believe that judgments of well-formedness reflect statistical information is that in many cases, such judgments reflect ambiguity resolution procedures that also rely on this information. For example, so called "garden paths" can arise when a word has two meanings, one of which is very frequent and one relatively rare. In a sentence such as *The horse raced past the barn fell*, RACED is used much more frequently as a transitive verb than as a past participle in a reduced passive. In such cases, the frequent meaning quickly dominates the rarer meaning, often to the point that the reader is unaware of the alternate structural interpretation. Such sentences are often judged ungrammatical by speakers who fail to recompute the relationships among the lexical items in the sentence after an initial parse. Statistical information of this type (the relative frequency with which a lexical item is used in one way rather than another) is thus required to account for the conditions under which judgments of well-formedness are made.

The usual argument against this approach to grammaticality is that there are sentences containing low probability sequences of words that can nonetheless be judged as grammatical (e.g., *Colorless green ideas sleep furiously*). The treatment of such sentences turns on the levels over which sequential statistics are being computed. Although the ungrammaticality of many sentences can be determined by detecting local anomalies defined over sequences of lexemes, others may depend on statistics involving other types of information. Assume, for example, that comprehension involves computing high level semantic types of words; for example, that a DOG is a thing and that PUSHING is an action. This information would provide the basis for deciding that the sequence *Colorless green ideas sleep furiously* is acceptable because each of the local (high level) semantic sequences PROPERTY PROPERTY THINGS ACTION MANNER is quite normal English. The sequence

Ideas colorless sleep furiously green would be rejected on this basis because the sequence THINGS PROPERTY ACTION MANNER PROPERTY does not occur.

Given this account of grammaticality judgment, the fact that the standard approach excludes most of this statistical information is important. If grammaticality judgments can be based on statistical information derived from experience with the target language, then it cannot be assumed that the task requires computing the kinds of representations assumed within the standard approach.

Agrammatism

We can now use this account to explore some puzzling data concerning apparent dissociations between knowledge of grammar and the capacity to make grammaticality judgments. The speech of agrammatic aphasics (Broca-type patients with lesions in the anterior portion of the dominant hemisphere) is typically restricted to telegraphic utterances that rely heavily on open class lexical items. This production impairment is frequently accompanied by impaired comprehension: Broca's aphasics tend to experience difficulty on comprehension tasks when syntax alone furnishes critical aspects of meaning (Caramazza & Zurif, 1976; Saffran, Schwartz, & Marin, 1980). Linguists have been interested in this form of aphasia because it was thought to provide another kind of evidence bearing on the nature of linguistic competence, specifically the existence for a syntactic module that can be selectively impaired.

The work of Linebarger, Schwartz, and Saffran (1983) raised important questions about the interpretation of agrammatic behavior. They described four agrammatic patients who exhibited comprehension difficulties but retained the ability to judge the grammaticality of many sentences. These results are important because they represent a dissociation between grammaticality judgment and other aspects of performance.

Attempts to relate agrammatic comprehension to syntactic theory assume that a failure to structurally represent a sentence causes a failure to comprehend that sentence. If grammaticality judgments represent evaluations over syntactic representations, then the failure to syntactically represent a sentence should also affect the ability to make appropriate grammaticality judgments. Linebarger et al.'s data provided evidence against the claim that agrammatism represents a selective loss of syntactic capacity, in that patients who performed at chance levels on comprehension tasks performed at high levels when judging the grammaticality of similar sentence types.

These findings generated considerable controversy. One response was the formulation of revised theories that attempted to maintain the idea that "agrammatism" has a grammatical basis (e.g., Grodzinsky, 1990; Hildebrandt, Caplan, & Evans, 1987; Zurif & Grodzinsky, 1983). However, these proposals have run into other problems (Tesak & Hummer, 1994; Druks & Marshall, 1991; Milekic, Boskovic, Crain, & Shankweiler, 1995). In the model presented below, we develop an alternative account in which a "syntactic" processing deficit is created by damaging parts of a neural network that computes from form to meaning and from meaning to form.

A second issue concerns the assumption that the grammaticality judgment task provides direct evidence about a person's syntactic knowledge. We have suggested that grammaticality judgments in many cases do not require evaluations of complete syntactic representations, but instead can be based on how well an utterance conforms (sometimes quite locally) to statistical regularities, acquired in the course of learning, and generally excluded on the standard approach from descriptions of language competence. Knowledge of such regularities might provide the basis for making well formedness judgments even when normal comprehension processes are significantly impaired. In the next section we present a simulation model that exhibits just this outcome.

Simulating Grammaticality Judgments

We now describe a connectionist model of grammaticality judgments that provides a basis for differentiating between several classes of grammatical and ungrammatical utterances and, when damaged, exhibits partial retention of this capacity. The model learned to perform two mappings. Given a sequence of words as input, it computed their semantic representations. This form to meaning mapping is an analogue of comprehension. Conversely, given an input sequence of meanings, the model computed the appropriate words, the meaning to form mapping involved in production. Our hypothesis about grammaticality judgment is that it involves computing the meaning of a sentence and then passing that derived representation through the production system. The mismatch between the form presented as input and the form computed on the backward pass through production provides the basis for judging grammaticality. In the implemented model this mismatch was quantified as the distance between relevant vectors. We assume that if these differences are large enough, subjects can set a decision criterion that allows them to distinguish the grammatical and ungrammatical utterances, as in the standard signal detection paradigm, although this decision process was not explicitly modeled.

Put simply, the judgment process is modeled by querying the network for its version of an input sentence. Given a particular input utterance, would the model have said it the same way? This is accomplished by processing the input sentence normally, computing as far as possible the corresponding meaning, generating a sentence that corresponds to that meaning, and then measuring how far apart the input and output forms are.

This way of implementing well-formedness judgments was inspired by a view of grammaticality in which a grammatical structure is seen as one which best satisfies the various constraints developed over the course of learning (e.g. Smolensky, 1986). These constraints reflect the interaction of innate constraints (whether linguistic or non-linguistic) and the input to which the learner has been exposed. It follows that an ungrammatical structure is one that is suboptimal, meaning that there is some other structure that better satisfies the relevant set of constraints given a particular input. As an example, let us take the input to a sentence generating system (production) to be a conceptual representation. On this view, the form produced on the basis of this conceptual representation will be that which best satisfies the multiple constraints that make up the speaker's knowledge of form–meaning relationships.

It follows that the grammaticality of an utterance is defined with respect to a particular meaning. Unlike the Turing machine metaphor of string recognition, the grammaticality of an utterance cannot be defined with respect to the form of that sentence alone, but must make reference to the meaning that gave rise to it. It further follows that an (absolutely) ungrammatical utterance is one to which no meaning maps. Note that this is not equivalent to saying that an ungrammatical utterance is one that maps to no meaning, because presumably there is always *some* semantic representation that best satisfies the constraints given the ungrammatical utterance as input. If we now take the input to the meta task of grammaticality judgment to be a sentence, and we generate a hypothetical space of all possible meaning candidates (comprehension), there will always be some best (semantic) candidate, even for (absolutely) ungrammatical utterances. On the other hand if we take the semantic output generated by that ungrammatical input and map it back to form (production) we will not get the sentence form that we started with, if it is the case that no meaning maps to that form.

The hypothesis, then, is that a mismatch between the form that is the input to the comprehension system and the form produced on the basis of what was comprehended could be used as the basis for detecting ungrammaticality. We assume that if these differences are large enough, subjects can set a decision criterion that allows them to distinguish the grammatical and ungrammatical utterances, as in the standard signal detection paradigm, although this decision process was not explicitly modeled.

Network Implementation

The network used in these simulations was trained on a series of utterances like those given by Linebarger et al. to their agrammatic subjects. As a consequence of training on the form to meaning and meaning to form mappings, the network developed a type of symmetric knowledge, i.e., both that form *a* entailed meaning *b* and that meaning *b* entailed form *a*. Because both mappings had a shared computational substrate, these two skills were not independent of one another. After training, the network was evaluated by supplying either novel forms or novel meanings and recording the network's behavior. In the course of training the network developed sensitivity to the statistical properties of the sentences to which it was exposed, and as a consequence behaved differently when provided with grammatical and ungrammatical versions of these utterances.

When normal processing was disrupted by "damaging" the network, it exhibited behaviors seen in agrammatic patients such as a failure to produce high frequency items that are low in semantic content (function words) and impaired comprehension (i.e., failure to activate the correct sequence of semantic representations for a given lexical input sequence). Although impaired in these ways, the damaged network retained the ability to distinguish between many grammatical and ungrammatical utterances.

Architecture

The architecture used in the simulations is shown in Figure 1, and consisted of three main layers. The semantic layer consisted of 297 units which served to represent the semantics of an utterance (see *Representation* below). This layer was connected to itself via a set of 15 cleanup units.

The pathway from the semantic units to the cleanup units and back to semantics allow for the semantic units to interact with one another during processing. The purpose of the layer of cleanup units along this pathway is to allow for interactions to develop among semantic units during processing. By providing an intermediate layer of units (the clean up units) along the pathway from semantics to semantics, it becomes possible to encode in the weights of these pathways a more complex set of relationships among semantic units. For example, EXCLUSIVE-OR relationships among sets of semantic units become learnable when a cleanup layer is used to connect the semantic layer to itself, whereas only linearly separable relationships (e.g., AND or OR) would be learnable if the semantic layer were to be connected to itself directly without an intermediate layer. The pathways between the semantic units and the cleanup units thus allow for combinations of semantic features to influence the patterns that develop over time on the semantic units.

Futhermore, in processing an exemplar through time, the semantic units and their associated cleanup units serve to form an attractor network, where an initial activity on the semantic layer may be coerced over time toward the nearest fixed point attractor developed during training[2]. The

[2] If a separate dimension is assigned to each unit, then each fixed point attractor corresponds to a particular point in a space whose dimensionality is defined by the number of units in the vector. The position of this point is determined by

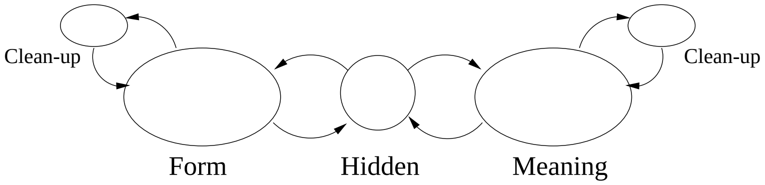

Figure 1. Architecture of the grammaticality judgment model. Arrows represent full connectivity between layers.

semantic layer was also connected to a hidden layer consisting of 50 units. These hidden units were connected to each form unit, each semantic unit, and to each other. The 97 form units were also connected to each other via a set of 15 cleanup units, allowing fixed point attractors to develop in the form representation as well. All connection sets were fully connected, and all weights were initially set to random values between -1 and 1.

Network Dynamics

The implemented network developed a sensitivity to the characteristics of sequences of words in an utterance. Our strategy for accomplishing this was in some ways quite similar to that used in simple recurrent networks (Elman, 1990), in which sequential dependencies are developed by representing sequences through time. Our network differs from the standard simple recurrent network in several ways, however. First, we exploited the advantages provided by the continuous activation function described in Pearlmutter (1989), in which the state of a unit changes smoothly over time in response to input from other units. This approach significantly improves the ability of networks to “reach back in time,” that is, to develop sensitivity to longer sequences than is possible in standard discrete time nets. This continuous approach is approximated by dividing the normal time steps of discrete back prop through time (Williams & Zipser, 1990) into ticks of some shorter duration. An infinite number of such ticks would represent truly continuous activation. The number of time steps per tick (called the integration constant) changes the grain at which activation is propagated and error injected into the network. Details of the implementation are given in the Appendix.

Second, unlike a simple recurrent network that freezes a set of weights (copy back connections) from the hidden units to the “context” units, all connection sets in this network were trainable. In this sense, the simple recurrent net represents a strict subset of the weight values that our network can take on. Like the simple recurrent network, however, this network does not suffer from the problem associated with providing a distinct set of units and connections for each distinct sequential element (letter, phoneme, word, etc.) in a representation, where the set of weights encoding knowledge about an element in one position is completely independent of the weights encoding knowledge about the same word in a different position. Rather, in this network information derived from experience about an element occurring at time t is available to the network when that element occurs at time $t^{+}_{-}n$.

For purposes of the simulation, we defined an exemplar as a sequence of states, each representing either a word or a word’s semantics. Under the version of continuous back propagation utilized here, the network does a forward pass on the entire string (all of the words) integrating

the activation value of the units. The set of patterns that are attracted to any of the fixed points in the course of processing form the basin of attraction for that fixed point.

activity up, and remembering its state for the whole utterance at every tick. On the backward pass, error is injected for each tick based on the integration constant, the error associated with each unit for each tick, and on what flowed backward from the following tick.

The targets for each utterance thus form a *trajectory* that the network attempts to learn to follow. For semantic targets, this trajectory is the sequence of points defining the semantic values of each word in the utterance. For form targets, this trajectory is the sequence of points representing the individual words of the utterance.

Use of this system allows the network to develop sensitivity to the sequences of state transitions defined by the training utterances. At word n, information about word $n-1$, $n-2$, etc. is available to the network in the form of the state of the target, hidden, and cleanup units when the processing of word n begins. Recurrent connections allow the state of the hidden units at any time to be influenced by their own state at previous times. The network learns to rely on regularities in these sequences to the extent that they minimize error.

Each utterance in the training set was presented for 65 "seconds," with an integration constant of .2 (5 ticks per second). Each word was presented for a window of 3 seconds, thus 15 ticks. Inputs were clamped only for the initial 7 ticks of the word's window, and feedback was given only on the final 8 ticks of each 3 second window. This offset between the time at which the input is clamped and the time at which the target is required forces the network to depend on its current state as well as its input. Thus, for example, activity on the semantic vector corresponding to the semantics of the word (e.g. CAR) was made to depend not only on the input from the formal representation of CAR, but also on the state of the network before and after the time that the form vector for the word CAR was clamped as input.

This technique forces the network to use information earlier in the sequence to begin to activate what it is predictable about the next item order to minimize error. Given the attractor network implemented via clean-up units at the meaning layer, if the prediction can get the semantic vector into the right neighborhood, the actual word presented will sharpen the representation quickly. Of course, only parts of the next word can be predicted: (e.g., the ENTITY feature after a determiner is a good guess). But if a wrong prediction is made, it takes the network a long time to recover, because both the current state of the semantic attractor and the identity of the current word have an effect on the semantic output.

Representation

Meaning. It is notoriously difficult to represent the semantics of propositions. It is even more difficult, if not impossible, to represent the semantics of propositions without a system for binding arguments and roles. In order to simplify the simulations, the semantics of utterances were represented by sequences of word level semantic representations. As a consequence, many relationships like coreference, binding, predication, and a host of others relevant to the semantics of propositions (whether semantic or syntactically represented) are not captured by this approach.

This simplification means that our model does not represent phrasal and propositional level relationships among words such as *subject of predicate* or *object of verb*. Although we assume that a good deal of knowledge concerning the formal expression of these higher level structures is also emergent from form-meaning pairings available in the learner's environment, the technical challenges involved in modeling such knowledge are considerable. This "role filler" (or "binding") problem arises in many cognitive domains, and has received considerable attention elsewhere e.g., in the area of vision Hummel and Biederman (1992), von der Malsburg and Schneider (1986); for phrase structure Omlin and Giles (1995); for grammatical category assignment Elman (1990); for verb argument structure Allen (1997a). Although the techniques used in these approaches vary, much of this work suggests that temporal dynamics of processing will play a crucial role in understanding how such relationships are represented. The approach adopted here is compatible with this general view, in that both this work and that focused on binding assume that understanding cognitive representation will involve the analysis of dynamical systems that change through time. In short, our model does not deal with all aspects of language but its limitations are not relevant to the idea that is our main focus, that many grammatical and ungrammatical utterances can be distinguished using much simpler and more local types of knowledge representations than are required for complete syntactic analysis or semantic interpretation.

Features for the semantic representations of words were based on the semantic hierarchy associated with each word from the Wordnet database (Miller, 1990), and then augmented by hand. Although Wordnet includes entries for many open class items, features for the closed class words in the training set were developed by hand. An example of the set of features used for the word HOUSE appears in 1.

(1) HOUSE: house housing lodging structure construction artifact object physicalobject entity

Since pronouns, prepositions, and determiners do not appear in the Wordnet database, these items were given semantic features that represented their relationship both to each other and to other words in the training set. For example, the pronoun *he* was given the features SINGULAR, MALE, and ANIMATE. In addition, features such as PLURAL were added to words such as *men, them* and *ducks* to distinguish between plural and singular versions of the same word. We used these representations because they form a series of hierarchies, with some features (e.g. ENTITY) applying to many different words, and others (e.g. VEHICLE) applying to many fewer. The semantic representations thus have an internal structure that the network can take advantage of during learning. Units higher on the hierarchy tend to be positively correlated with those lower on the hierarchy, and to develop positive weights between them. As a consequence, units higher on the hierarchy will tend to activate those below them, and those lower on the hierarchy will tend to activate those above them. In contrast, units at similar levels tend to inhibit each other.

The semantics of each word, then, was represented as the state of a space whose dimensionality was defined by the number of units (297) in the semantic representation, and the semantics of an utterance was represented by a series of such states.

Form. The forms of utterances were represented as a series of words presented over time. Words were represented locally, that is, each word was represented by a single unit. The vector representing each word was thus extremely simple, consisting of a single unit being on and all other form units being off for the appropriate time steps. The form of an utterance was presented to the network by activating the units representing each word of the utterance in sequence. There were 97 distinct words used in the example sets, so the form layer consisted of 97 units.

Training and Testing Materials

Agrammatic performance on ten sentence types was reported by Linebarger et al. (1983). These types formed the basis of the training and testing sets used in the simulations. Grammatical and ungrammatical versions of these ten types are listed in Table 1.

Table 1: Grammatical and ungrammatical examples of sentence types used for training and test corpora.

Type	Example
I. Strict subcategorization	
grammatical	He came to my house at noon.
ungrammatical	*He came my house at noon.
II. Particle movement	
grammatical	She went up the stairs in a hurry.
ungrammatical	*She went the stairs up in a hurry.
III. Sub-aux inversion	
grammatical	Did the old man enjoy the view?
ungrammatical	*Did the old man enjoying the view?
IV. Empty elements	
grammatical	Frank was expected to get the job.
ungrammatical	*The job was expected Frank to get.
V. Tag questions, Pronoun Agreement	
grammatical	The little boy fell down, didn't he?
ungrammatical	*The little boy fell down, didn't it?
VI. Left Branch condition	
grammatical	Which old man did you invite to the party?
ungrammatical	*Which old did you invite man to the party?
VII. Gapless relatives	
grammatical	Mary ate the bread that I baked.
ungrammatical	*Mary ate the bread that I baked a cake.
VIII. Phrase structure	
grammatical	The paper was full of mistakes.
ungrammatical	*The paper was full mistakes.
IX. Reflexive agreement	
grammatical	I helped myself to the birthday cake.
ungrammatical	*I helped themselves to the birthday cake.
X. Tag questions, aux copying	
grammatical	John is very tall, isn't he?
ungrammatical	*John is very tall, doesn't he?

A training and testing corpus was developed by providing *partial paradigms* for each of the sentence types listed above. Twenty sentences were created for each of the ten sentence types for a total of 200 utterances. The partial paradigm for each type was created by replacing individual words with others that might be used grammatically in those positions. For example, one of the sentence types was a reflexive (*The little boy cut himself while playing*). The paradigm created for this sentence type included:

- The little boy cut himself while playing.
- The little girl cut herself at noon.
- The big boy cut himself while playing.
- A little boy fell while running.
- An old man cut himself while shaving.

Half of this training set was used to train the network, and half was used to assess learning after training. The training corpus consisted of 100 utterances and 665 words (97 types). The mean number of words per utterance in the training corpus was 6.65, ans the mean number of content words per utterance was 3.1. The testing corpus consisted of 100 utterances and 652 words (97 types). The mean number of words per utterance in the testing corpus was 6.52, and the mean number of function words per utterance was 2.8. Two examples were developed per utterance, one for each mapping.

Finally, an ungrammatical corpus was developed by creating ungrammatical versions of each of the sentence types listed above. Each of the ungrammatical utterances deviated from the grammatical version in the way picked out by the category name. For example, the ungrammatical phrase structure utterances were all examples of phrase structure violations (e.g. *He came to my town*→ **He came my town*). The ungrammatical corpus consisted of 100 utterances and 653 words (97 types). The mean number of words per utterance in the ungrammatical corpus was 6.53, and the mean number of function words per utterance was 2.8.

Training the network consisted of presenting two types of trials. The model was trained on grammatical sentences only. On form to meaning trials, the unit representing each word in the utterance was activated in sequence. The task of the network was then to compute the correct semantic representation of each word in the sequence. On meaning to form trials, the requirements were reversed. Word meanings were presented one at a time to the network, and the task of the model was to compute the formal trajectory that constituted the representation of the utterance by activating the appropriate word units in the right sequence at the right time. The network was trained for 25000 iterations, where an iteration consisted of a presentation and feedback on either a form to meaning example or a meaning to form example. After 25000 iterations, the network produced the correct trajectories for all utterances on which it had been trained.

After training, the model's performance on three types of tasks was assessed under two different conditions. The three tasks were a comprehension task, a production task and a grammaticality judgment task. The comprehension and production tasks are assessments of the model's ability to handle the primary task of form to meaning and meaning to form mappings under normal and impaired conditions. The grammaticality judgment task is a test of the model's ability to discriminate two types of stimuli.. In the NORMAL condition, performance of the undamaged model was assessed. In the IMPAIRED condition, 10% of connections between the semantic and hidden units were lesioned by setting their weights to 0. This represents an impairment to the network's ability to successfully perform the mappings on which it was trained.

Results

Normal Comprehension

The model's ability to produce the correct semantic representations for novel utterances was tested by supplying the 100 novel utterance forms of the testing corpus to the network and recording activation of the semantic vector at the center of the target period (tick 11). The results are shown in the first columns of Tables 2 and 3. Table 2 shows the proportion of words correctly identified by the network. These figures were computed as follows. The semantic vector computed by the network 11 ticks after the form of a word was presented was compared with the vector representing the semantic target that word. If the computed vector was both closer to the target vector than any other word's vector and each unit of the computed vector was within .2 of its target, the word was considered recognized. The Euclidean distance between the computed and target vectors for each sentence type is shown in Table 3. Together these figures give an overall view of the network's performance on the comprehension tasks. The first column of Table 2 shows that the normal network is easily able to accommodate novel utterances. The range of identification is between 88% and 100% for comprehension in the normal network. Thus, although the network had not been trained on the sentence *A little boy fell at noon* it had no trouble producing the correct vector for each word at the correct time step.

Table 2: Percentage of words correctly comprehended or produced for normal and impaired network.

	Comprehension		Production	
Sentence Type	Normal	Impaired	Normal	Impaired
Subcategorization	1.00	0.66	0.93	0.76
Particle Movement	0.93	0.36	0.99	0.47
Inversion	1.00	0.37	0.93	0.41
Empty Elements	0.91	0.45	0.94	0.59
Tag Questions (PN)	0.94	0.41	0.86	0.46
Left Branch Condition	0.99	0.45	0.88	0.56
Gapless Relatives	0.98	0.49	0.93	0.51
Phrase Structure	0.94	0.34	0.97	0.51
Reflexive Agreement	0.90	0.54	0.97	0.54
Tag Questions (Aux)	0.88	0.31	0.89	0.57
Mean	0.95	0.43	0.93	0.53

These results show that under normal conditions computing the correct semantics for sequences of novel grammatical utterances is a simple problem for the network. The ability to recognize the elements of novel grammatical sequences is facilitated by the fact that the same weights are being used for words regardless of a word's position in the utterance. Thus, regardless of whether the network had been exposed to BOY in the third position of an utterance, the weights from the unit representing BOY are still those used when BOY appears in this position in a novel utterance.

In the impaired condition comprehension performance is significantly worse. The second columns of Tables 2 and 3 show that when damaged, the network is less likely to produce the correct

Table 3: Distances between target and computed vectors for Normal and Impaired production and comprehension.

	Comprehension		Production	
Sentence Type	Normal	Impaired	Normal	Impaired
Subcategorization	0.97	1.77	0.64	0.87
Particle Movement	0.87	1.86	0.68	0.92
Inversion	1.01	2.41	0.75	0.94
Empty Elements	1.11	1.98	0.75	1.00
Tag Questions (PN)	1.06	2.21	0.65	1.06
Left Branch Condition	1.00	2.34	0.61	0.83
Gapless Relatives	1.00	2.21	0.77	1.02
Phrase Structure	1.10	2.10	0.79	0.95
Reflexive Agreement	1.06	1.81	0.68	0.91
Tag Questions (Aux)	1.00	2.04	0.60	0.94
Mean	1.01	2.07	0.69	0.94

word's semantics, and that the average distance between the correct vector and that produced by the network is higher than in the undamaged network. (All differences between columns 1 and 2 of Tables 1 and 2 are significant at $p < .05$ or lower.)

Production

The model's ability produce the correct formal representation for novel utterances was then tested by supplying 100 novel meaning sequences representing the testing corpus. As in the comprehension task, ten sentences of each type were presented. The results are shown in the third columns of Tables 2 and 3. As in the comprehension task, for each word, the form vector that was computed by the network 11 ticks after the semantics of a word was presented was compared with the vector representing the formal target for that word. If the computed form vector was closer to the target vector than to any other formal vector, and the activation of each unit was within .2 of its target, the correct word was considered produced. As in the comprehension task, this task is fairly straightforward for the normal network, and performance was quite high, ranging from 86% to 99% words correctly produced.

Impairment to the network also significantly affects its ability to compute the correct sequence of words. The impaired model's ability produce the correct formal representation for novel meanings sequences was tested as above. The results are shown in the final columns of Tables 2 and 3. Again, under damaged conditions, the proportion of words correctly produced is lower, and the mean distance between target and computed vectors is higher, than under normal conditions. All differences between columns three and four of Tables 1 and 2 are significant at or below the .05 level.

An interesting aspect of the production tests on the impaired model was the differential impairment on grammatical morphemes as a consequence of their semantic "shallowness". As can be seen in Table 4, closed class words are more likely to fail to be produced, and to be further from their targets, than open class words. Why are function and content words differentially affected by damage to the connections between hidden and semantic representations? Activation of the correct semantic pattern for a word relies both on the word input to the model and on the semantic attractors which move the initial representation to its target. Because the hierarchy is typically deeper for content than for function words, content words are more resilient to damage to the system. The influence of semantic representations on agrammatic production may be an additional factor to those already recognized concerning why closed-class items may be impaired when, for other reasons, they might be expected to be easy to produce (e.g., Stemberger, 1985).

Table 4: Impaired Network: Content versus Function word production.

Sentence Type	OC	CC
Subcategorization	0.81	0.69
Particle Movement	0.60	0.35
Inversion	0.57	0.25
Empty Elements	0.63	0.56
Tag Questions (PN)	0.55	0.38
Left Branch Condition	0.74	0.46
Gapless Relatives	0.52	0.50
Phrase Structure	0.73	0.30
Reflexive Agreement	0.65	0.47
Tag Questions (Aux)	0.63	0.51
Mean	0.64	.44

Grammaticality Judgments

The network was trained by interleaving form to meaning and meaning to form exemplars. This interleaved training caused the network to develop knowledge of the probable contingencies among elements in sequences of both form and meaning. The dynamics of the grammaticality judgment task rely on the following property of the trained network: when a formal pattern is supplied to the network, the semantic pattern associated with it is activated because of the form to meaning connections. Activation then flows back to the form vector along the normal meaning to form path. This activation results in the recreation of the original form vector several ticks after it is released. Thus the form that was presented to the network is normally reproduced as activation flows back to the form layer. However, when the form of an utterance deviates from the type the network is familiar with, the computed semantics deviate from normal, and as a consequence, the form that is created deviates from the form presented. We simulate the meta-linguistic notion of grammaticality as the accurate reproduction of a supplied form, measured in terms of distance. The results show that ungrammatical utterances of the type used in Linebarger et al.'s study produce more deviant recreations of the input than novel grammatical sentences do.

Ten ungrammatical versions of each sentence type were presented to the both impaired and normal networks. Although impairment to the network significantly disrupts the ability of the model to compute the correct meanings of novel forms and the correct forms of novel meanings, the ability to distinguish grammatical from ungrammatical utterances is retained for 7 of the ten utterance types.

Figures 2 and 3 show the mean distance between the form vector supplied and that produced 11 time ticks after the onset of each word for normal and impaired networks. For example, the first set of bars in Figure 2 shows that the normal network (re-)produced vectors with a mean distance of

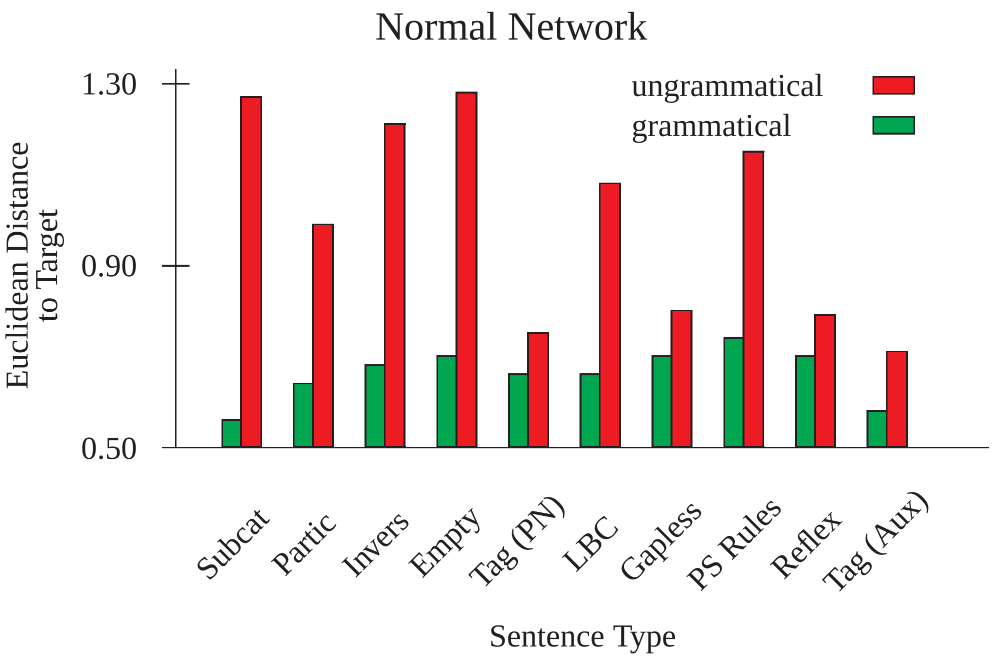

Figure 2. Mean distance between form supplied to network and form computed by network for grammatical and ungrammatical novel utterances for normal network

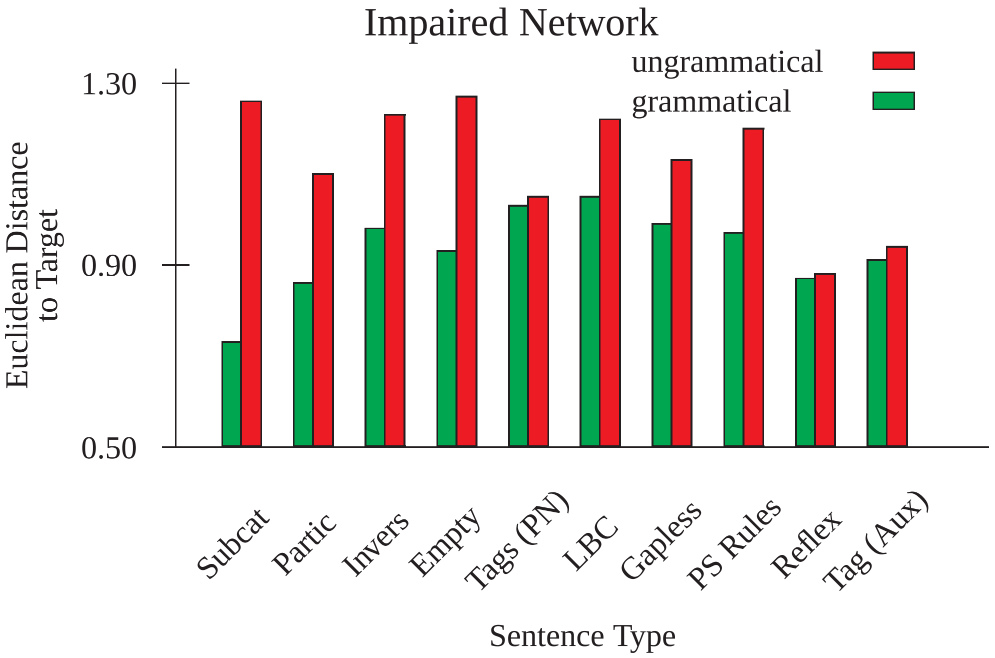

Figure 3. Mean distance between form supplied to network and form computed by network for grammatical (G) and ungrammatical (U) novel utterances for impaired network

.56 from those supplied on novel grammatical versions of the subcategorization sentences like *He left my house at noon*, but (re-)produced vectors with a mean distance of 1.27 from that supplied on novel ungrammatical sentences such as *He left to my house at noon*. For the normal network 7 of the 10 sentence types produced significant differences between grammatical and ungrammatical distances at or below the .05 level. The sentence types on which the network did not detect ungrammaticalities by this measure were the two types of tag questions and reflexive agreement. Figure 3 shows these distances for the grammatical and ungrammatical utterances produced by the impaired network. As seen in the first set of bars, the impaired network (re-)produced vectors with a mean distance of .73 from those supplied on novel grammatical versions of the subcategorization sentences, and vectors with a mean distance of 1.26 from that supplied on novel ungrammatical subcategorization sentences. Like the normal network, in the impaired network 7 of the 10 sentence types produced significant differences between grammatical and ungrammatical distances at or below the .05 level. The sentence types on which the network did not exhibit distinctions between grammatical and ungrammatical utterances by this measure were the same types as before.

Interestingly, Linebarger et al.'s patient data exhibit essentially the same pattern as the impaired simulation. Although the patients were able to judge the grammaticality of most types of sentences, they were impaired on the same three sentence types as the model. For the seven sentence types the patients were able to judge correctly, Linebarger (1989) reports performance with a range of 81.2-90.4% correct. For the other three sentence types, the patients performed at chance levels (Tag questions (Aux) 62.1%; Tag questions (PN) 63.7%; Reflexives 64.2%).

Figure 4 illustrates a comparison between the processing of the grammatical and ungrammatical versions of an utterance of the subcategorization type in the normal network. The utterances differ with respect to the subcategorization frames of the verbs. The verb LEFT does not subcategorize for the preposition TO, but the verb WENT does. The distance between presented and calculated values of the form vector at tick 11 are plotted for each word of the utterance. At the point of ungrammaticality, the distance between what is presented and what is computed rises. Although the continuations of the sentences are identical, the network continues to produce formal vectors that deviate from their targets more than in the grammatical case. This effect shows the impact of sequential processing in the network.

The opposite case is illustrated in Figure 5, where the verb LEFT is used correctly, but the verb WENT (which is consistently used with TO in the training set) is used in a violation of its "subcategorization frame". Again the network responds to this non-canonical sequence by producing vectors that continue to deviate from their targets for the next two words.

Why does this result obtain? Although the comprehension and production results reported above are consistent with the idea that the network was only responding on a word by word basis, its performance actually relies on more than merely a local mapping between the current form and meaning pair. Because the network was encouraged to develop a reliance on its current state as well as its current input, anomalous sequences such as *went the store* produce state trajectories in the semantic units that do not correspond to the regularities on which the network has come to rely. As a consequence anomalous local sequences tend to produce anomalous semantics, and anomalous semantics produce formal vectors that deviate from the form supplied. This result is partly brought about by the use of continuous time training. Since targets are supplied prior to the time at which clamping the current word form can activate the correct units on their own (because of the built-in rise time), the network learns to rely on information that is available, namely material prior in the sequence. Since only some parts of the prior sequence are reliable, the network learns to take

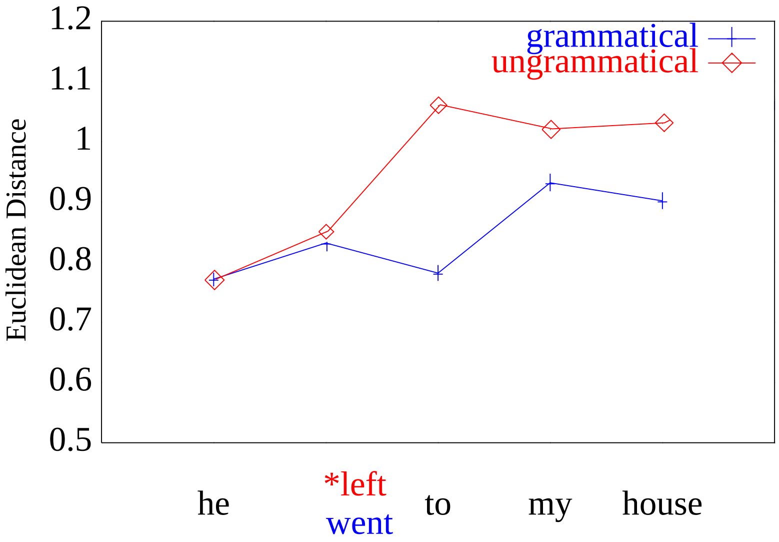

Figure 4. distance between form presented and form reproduced for a single sentence from the subcategorization set.

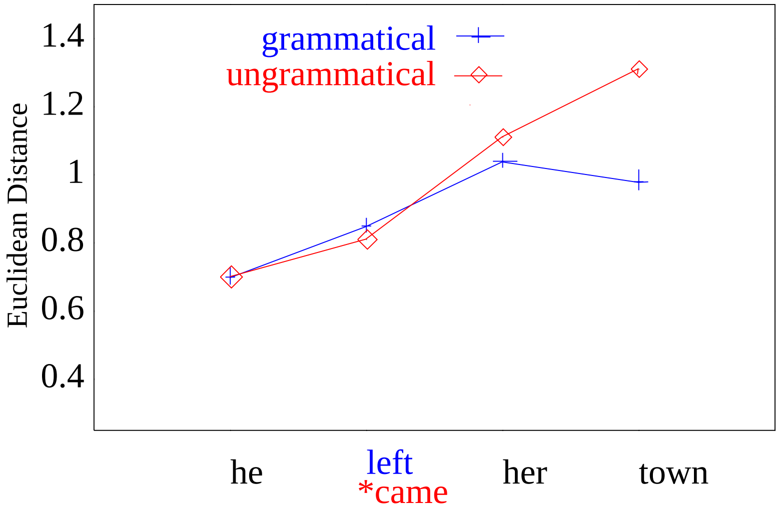

Figure 5. distance between form presented and form reproduced for a single sentence from the subcategorization set.

advantage of those regularities and ignore other aspects of its input.

A related issue concerns the fact that in our results the absolute value for the distance between grammatical and ungrammatical utterances varies between sentence types. Is it reasonable to suppose that different cutoffs are required for different sentence types? Although no single line can distinguish between grammatical and ungrammatical versions across sentence types, it isn't clear what the significance of such a line would be. Our basic theory is that an ungrammatical utterance results in a deviation from the normal course of processing, which we measure in terms of a comparison between grammatical and ungrammatical versions. Our method assumes a sensitivity to this distinction, and not to an absolute level of difference.

Colorless Green Ideas

Earlier we noted the existence of sentences such as *Colorless green ideas sleep furiously*, which contain low probability sequences of words that can nonetheless be judged as grammatical and suggested that the treatment of such sentences turns on the levels over which sequential statistics are being computed. We claimed that a sentence such as the *Colorless* one might be rated as more acceptable than a random permutation of words based on sequential regularities in the high level semantic properties of these items.

The network we have presented was designed to be sensitive to statistical regularities in lexical and semantic sequences simultaneously. The network was sensitive to sequences of lexical items because the input form of both the training task and the grammaticality judgement task consisted of local representations of lexical items. At the same time the network was sensitive to the sequences of the semantic representation of words, in that processing involves computation of these semantic representations.

In order to demonstrate that the network is sensitive to both of these levels simultaneoulsy, we tested the network under four conditions that manipulated two factors: the transitional probabilities between words and the transitional probabilities of the semantic types that the words represented.

In the first condition (HH) the network was presented with sentences in which the transitional probabilities between both lexical items and semantic types were high. These are normal sentences. The second condition (LH) consisted of sentences in which the transitional probabilities between words was low, but the transitional probabilities between semantic types was high. This type of sentence corresponds to ”Colorless ideas sleep furiously” in which the semantic sub-sequences [property property entity act manner] are consistent with semantic sequences that appeared in the training set, but the bigram frequencies of the words (e.g. colorless-green, green-ideas, etc.) were low or zero in the training corpus. The third condition (HL) consisted of sentences in which the transitional probabilities between words were high, but the transitional probabilities between semantic types were low. This condition is possible because there are sequences with high word transitional probabilities such as *mother cares*, where the semantic type represented by the first word in the pair predicts a different semantic continuation more strongly than that of the second word. The fourth condition (LL) consisted of sentences in which both types of transitional probabilities were low. This condition corresponds to random sequences of words such as *sleep ideas green furiously colorless*. Ten sentences of each type matched in terms of the overall frequencies of the words they contained were presented to the network.

Table 5 provides the mean transitional probabilities between words (wtp), the mean transitional probabilities between semantic types (stp), and an example sentence for each condition. The semantic type for lexical items was determined by taking the most frequent semantic feature that

Table 5: Example stimuli and mean word (wtp) and semantic (stp) transitional probabilities for four conditions in the grammaticalilty judgment task.

Condition	Example	wtp	stp	score
HH	which girl did you invite	.33	.32	.68
LH	which street invited a cake on time	.02	.26	.77
HL	my mother was expected to arrive	.30	.07	.80
LL	on invited cake street time the which	.00	.00	1.10

appeared in the word. For example, the representation for the word *mother* included the following features:

MOTHER: FEMALEPARENT PARENT FEMALE RELATIVE HUMAN LIVINGTHING ORGANISM ANIMATE ENTITY

The most frequent of these features in the training vocabulary is the entity bit. The semantic type of the word *mother* was thus taken to be entity.

The final column of Table 5 gives the mean grammaticality judgment of the network in these four conditions. As in the earlier grammaticality judgment task, these scores reflect the mean euclidean distance between the vector representing the form supplied to the network and that computed by the network for each sentence type. The LL condition, corresponding to a random sequence of words, is the sentence type that yields the largest deviation, as would be expected if it is the least grammatical. The HH sentence types yielded the least deviation. The other two conditions result in scores that are intermediate in value. The differences between the random word condition (LL) and the other three conditions were significant below the .01 level. The differences between the HH, LH and LH conditions are not significant, although this may reflect the relatively small number of test sentences in condition.

In summary, the model yielded graded performance on the four types of sentences and pseudosentences. The model can be seen as defining a metric in which sentences differ in degree of grammaticality. The largest differences were between the random word condition (LL) and the other 3 conditions, providing a basis for treating the LL items as ungrammatical and the other stimuli–including the model's version of a ”colorless green ideas” sentence–as grammatical. The model also suggests that it should be harder to judge sentences of the HL and LH types as grammatical than the HH items, an observation that appears to be consistent with human performance.

Discussion

The simulations presented here represent a step in the development of the alternative framework we described in the introduction. The implemented model illustrates how knowledge of language can be represented in a network rather than a grammar. The network acquired this knowledge in the course of learning to comprehend and produce utterances. The implemented model is clearly limited in scope, addressing only a fragment of the grammar of one language, but these results invite further investigations along the same lines and we have by no means approached the limit of what can be represented in such systems (for related work, see Chater & Christiansen, in press).

We also took a step toward developing a theory of how grammaticality judgments are made. In the absence of such a theory, linguists have interpreted performance on the task in different ways. Sometimes it is assumed that the judgments of native speakers, children, or aphasic patients more or less directly reflect the state of their grammatical knowledge. Sometimes it is noted that factors outside the scope of grammatical theory can influence decisions, but what is involved in filtering out these performance factors and whether this can be achieved in a consistent manner are unclear.

Our account of grammaticality judgment has three main features. First, we note that the capacity to perform the task emerges in the course of acquiring a language but does not play a central role in the acquisition process itself. In this sense it is like being able to make lexical decisions. Second, there are no absolute criteria for making such decisions; the criteria that people use vary depending on the nature of the sentences they are being asked to judge. Third, making the decision involves generating an error signal based on discrepancies between the sentence presented and what the linguistic system computes. In our model we generated this error signal by passing the computed semantic representation back through production. This was undoubtedly a simplification insofar as other error signals could be derived from the model and these are probably relevant to performance under some circumstances. For example, Plaut (1997) has described how anomalous patterns of semantic activation can provide a basis for making a lexical (word-nonword) decision, and it is easy to imagine the same kind of mechanism being used to judge grammaticality.[3] The results suggest that for a fairly broad range of sentence structures, local anomalies provide a sufficient basis for making correct responses. This means that it cannot be assumed that decisions necessarily require deriving a full syntactic representation of the utterance.

The method we used to implement well-formedness judgments was inspired by a view in which a grammatical structure is one which best satisfies the various probabilistic constraints encoded over the course of learning. On this view, a sentence is that which best satisfies the constraints that make up the speaker's knowledge of language specific form to meaning and meaning to form relationships given a particular semantic intention. The acceptability of an utterance on this view is defined with respect to a particular meaning. This account differs in kind, of course, from the view that a structure may be ill-formed solely on the basis of the syntactic features of its lexical items.

Implementing these ideas provided the basis for addressing questions concerning the bases of aphasia and the nature of grammaticality judgments raised by Linebarger et al. (1983)'s study of agrammatic patients. Damaging the network impaired its performance on the tasks on which it was trained, yet it was still able to distinguish between grammatical and ungrammatical representations of several sentence types. These results provide a basis for explaining how Linebarger et al's patients could perform above chance on such sentences even when their comprehension was significantly impaired. Given the simplicity of the input data that the model had to work with, the fit between the model and the Linebarger et al. (1983) data was quite good. Although there was some variation among the patients, overall the patients were impaired on the same types of sentences as the model. This outcome suggests that our explanation for the basis of grammaticality judgments is a viable one.

The sentences that both the network and patients could judge correctly are ones containing local sequential anomalies. The three sentence types on which both network and patients failed to distinguish between grammatical and ungrammatical versions were the ones for which these local

[3] Plaut (1997) uses a measure called *stress*, based on the entropy of sets of units. This measure reflects how far unit activation deviates from 0.5, in that the stress of a unit is 0 when its state is 0.5 and approaches 1 as its state approaches either 0 or 1. In a model simulating grapheme to semantic mappings, the target semantic patterns for words were binary, and thus they showed maximum stress. Because nonwords shared structure with sets of words that had conflicting semantic features, nonwords typically failed to drive semantic units as strongly as words did, producing semantic patterns with much lower average stress.

anomalies are not readily apparent. Examples of the ungrammatical versions of the ten sentence types are reproduced in Table 6. Sentence type I includes the local sequence CAME MY. Type II contains the sequence WENT THE STAIRS. Type III includes MAN ENJOYING. Type IV includes FRANK TO GET. Type VI includes WHICH OLD DID, and Type VII includes FULL MISTAKES. None of these sequences are consistent with the types of lexical-semantic sequences that appear in either the training set or in the novel grammatical testing set[4].

There were three sentence types on which the model was unable to detect differences between grammatical and ungrammatical versions. These were the same sentence types that Linebarger et al.'s patients had the most difficulty with. In both cases the basis for impaired performance on these items is unclear. There are a number of reasons why the network might have been unable to detect these types of ungrammaticalities. One possibility is that although these sentences, like the others, contain sequential anomalies, they are not sufficiently local. That is, although all ten sentence types involve violations of lexical and/or semantic sequences, the distances over which the anomalies are defined are too long in these three cases for the current architecture to pick up. The sentence in Table 6 illustrating Type V, for example, requires holding information about BOY for five lexical items prior to processing IT. Similarly, the example shown for Type X requires holding information about the auxiliary for 4 lexical items. This possibility could be addressed by conducting a larger scale simulation involving more sentences and adjusting the number of units in the network.

A second possibility is that the differences between the grammatical and ungrammatical forms involve kinds of dependencies that our simple network does not encode. The model does not encode all of the information on which grammaticality judgments can be made, and it is likely that many ungrammatical sentence types will require access to such information. A third possibility is that the poorer performance on these three sentence types derives from the fact that there happened to be less overlap between them and the other sentence types in the corpus. For example, the knowledge that the model brings to bear on sequences such AS CAME MY relies on exposure to all of the other sequences involving verbs in the training set. In contrast, knowledge concerning the relationship between reflexives and antecedents in the network comes only from exposure to sentence Type IX. Poorer performance on Type IX might simply reflect exposure to fewer relevant examples. Again, this possibility can be addressed in larger-scale simulations of the same type we have explored.

Although additional research is required in order to determine which of these factors is relevant to the model's performance, it is clear that there are two general factors limit the model's performance. First, the network was only given access to a fraction of the information that enters into the formation of the dynamic representations that underlie language behavior. It is likely that in humans, performance on the sentence types that we tested benefits from exposure to a broad range of other structures not included in the training set. Second, the model's architecture limits its capacity to represent important aspects of the semantics of utterances. For example, although we represent the semantics of propositions as a trajectory of semantic values, it is clearly the phrases

[4] Obviously these sequences are only anomalous relative to the training set: simple two word sequences such as FULL MISTAKES are less anomalous relative to the language as a whole (we can imagine sentences such as *Full mistakes are penalized less than partial mistakes* for example) and therefore would not be expected to trigger a judgment of ungrammaticality by themselves. Given the knowledge of the average human speaker, the specific sequence types that provide the basis for deciding that an utterance is ungrammatical will in many cases differ from those that the model is sensitive to, but the same principles will apply. For example, anomalies may be defined over longer stretches such as *was full mistakes*, where the use of was forces a particular interpretation of full, and with that interpretation in hand, the use of mistakes becomes an anomalous sequence.

Table 6: Ungrammatical example sentences.

Type	Example
I Strict subcategorization	*He came my house at noon.
II Particle Movement	*She went the stairs up in a hurry.
III Sub-aux inversion	*Did the old man enjoying the view?
IV Empty Elements	*The job was expected Frank to get.
V Tag Questions (PN)	*The little boy fell down, didn't it?
VI Left Branch	*Which old did you invite man to the party?
VI Gapless relatives	*Mary ate the bread that I baked a cake.
VIII PhraseStructure	*The paper was full mistakes.
IX Reflexive agreement	*I helped themselves to the birthday cake.
X Tag Questions (AUX)	*John is very tall, doesn't he?

of language that refer to conceptual units. Similarly, propositions have semantic characteristics that are compositional, that is, built up out of the semantics of the phrases and clauses that make up the form of a proposition. There are all sorts of semantic relationships that occur across multi-word windows, including co-indexation, predication, dependencies, thematic role binding, and others. In many cases grammaticality judgments are made on the basis of more information than is provided by the sequential regularities of semantic sequences that we were able to represent in our network.

In closing, we suggest that this model illustrates an approach to thinking about language acquisition, processing, and breakdown that shows considerable promise. Given the simplicity of the model's architecture and the limited corpus on which it was trained, it seems quite surprising that it was able to develop a basis for performing the grammaticality judgment task at levels comparable to normal and aphasic subjects. The claim that subjects can base their grammaticality judgments on statistical cues such as sequential probabilities of words clearly differs from the view that grammaticality judgments reflect access to principles of grammar. These differences can be seen clearly by considering Linebarger's (1989) discussion of the various bases on which sentence Type IV (Empty elements) might be judged ungrammatical:

"We might reject "Frank thought was going to get a job" for any number of reasons. If the empty category is PRO, then it violates the requirement that PRO be ungoverned, so we might reject it as a violation of the binding theory. Or we might take the empty category to be an NP trace of Frank, assigning [...] the D-Structure "__ thought Frank was going to get the job"; under this analysis [the utterance] represents, inter alia, a violation of the theta criterion since the moved NP is now assigned two theta roles. Recognition of any of these principles might trigger a rejection. On the other hand, perhaps the sentence is ultimately rejected because the grammar, – by disallowing PRO and NP trace in this position, provides us with no NP for the verb phrase 'get the job' to be predicated of, and the sentence simply 'makes no sense' unless it expresses who it is that is expected to get the job."

We take our results to indicate that many grammaticality judgments may be made on the basis of knowledge of sequential regularities of the type that humans apparently cannot help but absorb in the course of language learning (Saffran et al., 1996). The degree to which this approach can be extended to other aspects of linguistic structure is an important question that remains to be answered.

References

Allen, J. (1997a). *Argument structures without lexical entries.* Unpublished doctoral dissertation, University of Southern California.

Allen, J. (1997b). Probabilistic constraints in the acquisition of verb argument structures. In *Proceedings of the third conference on generative approaches to language acquisition.* Edinburgh.

Bates, E., & MacWhinney, B. (1982). A functionalist approach to language development. In E. Wanner & L. Gleitman (Eds.), *Language acquisition: The state of the art.* Canbridge, England: Cambridge University Press.

Berwick, R., & Weinberg, A. (1984). *The grammatical basis of linguistic performance.* MIT Press.

Bever, T. G. (1972). The limits of intuition. *Foundations of Language*, *8*, 411-412.

Borer, H., & Wexler, K. (1992). Bi-unique relations and the maturation of grammatical principles. *Natural Language & Linguistic Theory*, *10*(2).

Bresnan, J. (1978). A realistic transformational grammar. In M. Halle, J. Bresnan, & G. Miller (Eds.), *Linguistic theory and psychological reality* (p. 1-59). Cambridge, MA: MIT Press.

Browman, C., & Goldstein, L. (1989). Aticulatory gestures as phonological units. *Phonology*, *6*, 151-206.

Caramazza, A., & Zurif, E. (1976). Dissociation of algorithmic and heuristic processes in language comprehension. *Brain and Language*, *3*, 572-582.

Chater, N., & Christiansen, M. H. (in press). Connectionism and natural language processing. In S. Garrod & M. Pickering (Eds.), *Language processing.* University College London Press.

Chomsky, N. (1961). Some methodological remarks on generative grammar. *Word*, *17*, 219-239.

Chomsky, N. (1965). *Aspects of the theory of syntax.* Cambridge, MA: MIT Press.

Chomsky, N. (1981). *Lectures on government and binding.* Foris: Dordrecht.

Chomsky, N. (1995). Categories and transformations. In *The minimalist program* (p. 219-234). Cambridge, MA: MIT Press.

Clark, H. H., & Haviland, S. E. (1974). Psychological processes as linguistic explanation. In D. Cohen (Ed.), (p. 91-124). Washington D.C.: Hemisphere.

Coltheart, M., Curtis, B., Atkins, P., & Haller, M. (1993). Models of reading aloud: Dual-route and parallel-distributed-processing approaches. *Psychological Review*, *100*(4), 589-608.

Druks, J., & Marshall, J. C. (1991). Agrammatism: An analysis and critique, with new evidence from four hebrew-speaking aphasic patients. *Cognitive Neuropsychology*, *8*(6), 415-433.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*, 179–211.

Fodor, J., Bever, T., & Garret, M. (1974). *The psychology of language.* New York: McGraw Hill.

Frazier, L., & Fodor, J. (1978). The sausage machine: A new two stage parsing model. *Cognition*, *6*, 291-326.

Gold, E. M. (1967). Language identification in the limit. *Information and Control*, *10*(16), 447-74.

Grimshaw, J., & Rosen, S. T. (1990). Knowledge and obedience: The developmental status of the binding theory. *Linguistic Inquiry*, *21*(2), 187-222.

Grodzinsky, Y. (1990). *Theoretical perspectives on language deficits.* Cambridge, Mass: MIT Press.

Grodzinsky, Y. (1995). Trace deletion, theta-roles, and cognitive strategies. *Brain and language*, *51*(3), 469-497.

Hildebrandt, N., Caplan, D., & Evans, K. (1987). The man_i left t_i without a trace: A case study of aphasic processing of empty categories. *Cognitive Neuropsychology*, *4*(3), 257-302.

Hummel, J., & Biederman, I. (1992). Dynamic binding in a neural network for shape-recognition. *Psychological Review*, *99*(3), 480-517.

Kelly, M. H. (1992). Using sound to solve syntactic problems: The role of phonology in grammatical category assignments. *Psychological Review*, *99*(2), 349-364.

Lakeoff, G. (1973). Fuzzy grammar and the performance competence terminology game. In C. Corum, C. T. Smith-Stark, & A. Weiser (Eds.), *Papers from the ninth regional meeting.* Chicago.

Linebarger, M. C. (1989). Neuropsychological evidence for linguistic modularity. In G. N. Carlson & M. K. tanenhaus (Eds.), *Linguistic structure in language processing* (p. 197-238). Norwell, MA: Kluwer Academic.

Linebarger, M. C., Schwartz, M. F., & Saffran, E. M. (1983). Sensitivity to grammatical structure in so-called agrammatic aphasics. *Cognition*, *13*, 361-392.

MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological Review*, *101*(4), 676-703.

Maddiesson, I. (1997). Phonetic universals. In W. Hardcastle & J. Laver (Eds.), *The handbook of phonetic sciences.* Blackwell.

Milekic, S., Boskovic, Z., Crain, S., & Shankweiler, D. (1995). Comprehension of nonlexical categories in agrammatism. *Journal of psycholinguistic research*, *24*(4), 299-311.

Miller, G. (1990). Wordnet: An on-line lexical database. *International Journal of Lexicography*, *3*(4).

Omlin, C. W., & Giles, C. L. (1995). Extraction of rules from discrete-time recurrent neural networks. *Neural Networks*, *to appear*.

Pearlmutter, B. A. (1989). Learning state space trajectories in recurrent neural networks. *Neural Computation*, *1*, 263-269.

Plaut, D., McClelland, J., Seidenberg, M., & Patterson, K. (1995). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review*.

Plaut, D. C. (1997). Structure and function in the lexical system: Insights from distributed models of word reading and lexical decision. *Language and Cognitive Processes*, *12*.

Saffran, E., Schwartz, M., & Marin, O. S. M. (1980). The word order problem in agrammatism. *Brain and Language*, *10*, 249-262.

Saffran, J., Aslin, R., & Newport, E. (1996). Statistical learning by 8-month old infants. *Science*(5294), 1926-1928.

Schutze, C. T. (1996). *The empirical base of linguistics.* Chicago: University of Chicago Press.

Seidenberg, M. (1997). Language acquisition and use - learning and applying probabilistic constraints. *Science*, *275*(5306), 1599-1603.

Seidenberg, M., Allen, J., & Christiansen, M. (1997). Probabilistic constraints in acquisition. In *Proceedings of the third conference on generative approaches to language acquisition.* Edinburgh.

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review*, *96*(4), 523-568.

Smolensky, P. (1986). Information processing in dynamical systems: Foundations of harmony theory. In J. McClelland & D. Rumelhart (Eds.), *Parrallel distributed processing: Explorations in the microstructure of cognition. volume 2 psychological and biological models.* (p. 194-281). Cambridge, MA: MIT Press.

Stemberger, J. P. (1985). Bound morpheme loss errors in normal and agrammatic speech:one mechanism or two? *Brain & Language*, *25*(2), 246-256.

Tesak, J., & Hummer, P. (1994). A note on prepositions in agrammatism. *Brain & Language*, *46*(3), 463-468.

Trueswell, J., Tanenhaus, M. K., & Kello, C. (1993). Verb specific constraints in sentence processing: Separating effects of lexical preference from garden-paths. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *19*(3), 528-553.

van Gelder, T. (in press). The dynamical hypothesis in cognitive science. *Behavioral and Brain Sciences*.

von der Malsburg, C., & Schneider, W. (1986). A neural cocktail party processor. *Biological Cybernetics*, *54*, 29-40.

Wexler, K., & Cullicover, P. (1980). *Formal principles of language acquisition.* Cambridge, MA: MIT Press.

Wexler, K., & Hamburger, H. (1973). On the sufficiency of surface data for the learning of transformational languages. In K. Hintikka, J. Moravcsik, & P. Suppes (Eds.), *Approaches to natural languages.* Boston: D. Reidel.

Williams, R., & Zipser, D. (1990). *Gradient based learning algorithms for recurrent connectionist networks* (Technical Report NU-CCS-90-9). College of Computer Science, Northeastern University.

Zurif, E., & Grodzinsky, Y. (1983). Sensitivity to grammatical structure in agrammatic aphasics: A reply to linebarger, schwartz and saffran. *Cognition*, *15*, 207-213.

Appendix

The continuous approach to activation is approximated by dividing the normal time steps of discrete back prop through time (Williams & Zipser, 1990) into ticks of some shorter duration. An infinite number of such ticks would represent truly continuous activation. The number of time steps per tick (called the integration constant) changes the grain at which activation is propagated and error injected into the network.

Under this approach to approximating continuous time, the instantaneous change in the activation of each unit in a network is dependent both on its current state and on the input it is receiving from other units. Rather than the more commonly used discrete activation function then, change in the activity of units in the network was governed by the formula given in equation 2

$$\tau \frac{\partial y_i}{\partial t} = -y_i + \sigma(x_i) \qquad (2)$$

$$\sigma(x) = (1 + e^{-x})^{-1} \qquad (3)$$

where $\sigma(x)$ is the output of the normal sigmoidal activation function applied to inputs to unit y (seen in equation 3), and y_i is the state of $unit_i$. The final parameter τ is a time constant, also normally ranging between 0 and 1, which multiplicatively alters the rate at which units rise in activation. A value of τ close to 0 will mean that a unit rises in activation very slowly, and a value of 1 would mean that the unit would rise in activity at the rate of $1 - e^{-t}$, where t is the number of time steps at which input is provided at a constant rate. In all cases, there is some rise time associated with the activity of a unit.

The activation function described in equation 2 defines a leaky integrator in which the closer a unit's activation is to its goal output (defined by the output of the standard sigmoidal transformation of equation 3), the more slowly it approaches its target. Use of this system allows us to vary targets continuously over the course of an example, and to train the network to be sensitive both to the current state of its units and to the inputs it is currently processing.

In order to apply back propagation through time to targets with continuous units, the backward propagation of error must also be made continuous. The network was thus trained using a variant of back propagation through time adapted for continuous units (Pearlmutter, 1989) shown in equation 4. After a forward pass, weights are updated in the direction and to a magnitude made dependent on how much a small change in their values would affect error in the units to which they are connected. More concretely, the change in weight from unit i to unit j is made proportionate to the partial derivative of the overall error with respect to that weight.

$$\Delta w_{i,j} = -\epsilon \frac{\partial E}{\partial w_{i,j}} \qquad (4)$$

where ϵ is a small constant (the learning rate, set at .1 in our simulations), and

$$\frac{\partial E}{\partial w_{i,j}} = \frac{1}{\tau_j} \int_{t_i}^{t_0} y_i \sigma(x) z_j \, dt \qquad (5)$$

where z is defined by the differential equation 6:

$$\frac{dz}{dt} = \frac{1}{\tau_i} z_i - e_i - \sum j \frac{1}{\tau_j} w_{ij} \sigma_t(x_j) z_j \qquad (6)$$

Importantly, in this version of back prop, the τ values of equation 2 were also a trainable parameter of the network, and were also made sensitive to how minute changes in τ at time t would affect error rates, holding everything else constant, as in 7

$$\Delta\tau = -\mu \frac{\partial E}{\partial \tau_i} \tag{7}$$

where μ is another small constant (set at .005 in our simulations), and

$$\frac{\partial E}{\partial \tau} = -\tau \int_{t_0}^{t_i} z \frac{\partial y_i}{\partial t} dt \tag{8}$$

The τ values for all units in the network were initially set to 1, but (only) those of the hidden units and the clean up units were trained, and thus allowed to take on values that tended to minimize error in the network. In particular, some units could ramp up quickly while others ramp up more slowly. This aspect of the training regime is what allows the network to reach back somewhat further in time than the more standard discrete back propagation training regimes.