Cognition, 50 (1994) 385–401
0010-0277/94/$07.00

Language and connectionism: the developing interface

Mark S. Seidenberg*
Neuroscience Program, University of Southern California, Los Angeles, CA 90089-2520, USA

Abstract

After a difficult initial period in which connectionism was perceived as either irrelevant or antithetical to linguistic theory, connectionist concepts are now beginning to be brought to bear on basic issues concerning the structure, acquisition, and processing of language, both normal and disordered. This article describes some potential points of further contact between connectionism and linguistic theory. I consider how connectionist concepts may be relevant to issues concerning the representation of linguistic knowledge; the role of a priori constraints on acquisition; and the poverty of the stimulus argument. I then discuss whether these models contribute to the development of explanatory theories of language.

Introduction

During the "linguistics meets Frankenstein" era, many theoretical linguists viewed connectionist as the return of radical empiricism and were at pains to establish that it had nothing significant to contribute to understanding linguistic phenomena. At best it could address some side issues about how grammar might be realized in the brain (Pinker & Prince, 1988). The important work of Prince and Smolensky (in press) represents a second phase in which connectionist concepts have begun to be recognized, but only insofar as they can be assimilated into the standard generative program of research. Thus, Prince and Smolensky

*E-mail marks@neuro.usc.edu
The author is grateful to Maryellen MacDonald, David Corina, Kevin Russell, and Jay McClelland for helpful discussions. Research supported by NIMH grant MH47566.

SSDI 0010-0277(93)00599-3

allow certain aspects of connectionist models to figure centrally in what is otherwise a standard competence theory (see also some of the papers in Goldsmith, 1993). My feeling is that now that the genie is out of the bottle, it will be hard to prevent people from discovering other ways in which connectionism contributes to understanding language structure, acquisition, and use. In order to facilitate a rapid and orderly transition to this third phase, I offer the following speculations about areas of contact between linguistic theory and connectionism. No actual simulations or real results are presented, only speculations concerning future research. As always, there are no guarantees that things will work out as predicted, and please: no wagering.

1. Representations

An early assertion about connectionist models was that they only afford a trivial, associationistic type of representation, one already known to be incapable of explaining linguistic knowledge and other interesting aspects of cognition (Fodor & Pylyshyn, 1988; Lachter & Bever, 1988; Pinker & Prince, 1988). Whatever the limits of the approach turn out to be, it is clear that even very simple nets generate quite interesting abstract representations that are linguistically relevant.[1]

In order to illustrate this point, consider, as a "Gedanken simulation", a standard feedforward network in which there is an input layer consisting of units representing phonetic features (voiced, labial, etc.), a hidden layer, whose functions are not prespecified, and an output layer identical to the input layer. We construct this so-called "autoassociative" network so that there are fewer hidden units than input/output units. All units as adjacent levels are interconnected. Given a pattern on the input units, the model's task is to recreate the pattern on the output units. We train the model using a standard algorithm such as backpropagation. The training set consists of input/output patterns corresponding to consonants in English. In essence, we train the network on a phonetic feature matrix of the sort found in every introductory linguistic textbook (e.g., Fromkin & Rodman, 1978, p. 79). As always, training involves finding an appropriate set of weights. Given the number of features and patterns involved, this is not a difficult problem. Training is complete when the model successfully recreates every input with a sufficiently small degree of error.

Once the model has been trained, each input pattern will generate a pattern over the hidden units. What are these patterns? They are *abstract underlying representations*, created by the network in the course of solving the problem we

[1]See, for example, Hare and Elman (1992), Corina (in press), and Gupta and Mozer (1993).

have set for it. What are the properties of these representations? They are determined by the nature of the input/output representation (which we took from phonetic theory), the inventory of input patterns (a fact about English) and the architecture of the model. The last was decided on the basis of a computational theory of network performance and a theory about the task to be performed. Because the number of hidden units is smaller than the number of input/output units, the network cannot perform the task by simply copying the input representation to the hidden units and then to the output layer. Rather, we have chosen an architecture that forces the network to compress the input representation and then expand it again (see Cottrell, Munro, & Zipser, 1988, for discussion). The network accomplishes this by exploiting the considerable redundancy in the input representation. The standard matrix of binary features, in contrast, does not represent the higher-order relationships among the features – for example, the fact that two features cannot co-occur or that certain features can only co-occur in conjunction with another feature. The architecture of the network forces it to encode these relationships; indeed, the model can only perform the task if it discovers and represents those relationships, which can only be done if there is an interlevel of hidden units. In achieving this economical representation, the model must encode both general aspects of the distribution of phonetic features across segments and idiosyncratic aspects of individual segments.

The knowledge that the network acquires can be used in performing different tasks. For example, we can ask it to judge the well-formedness of segments. Although the model is trained on the set of English consonants, the input layer is capable of representing consonants that do not happen to occur in the language. What happens if we present a vector of features that forms a legal consonant in Spanish, say, but not English? The model will attempt to assimilate this input pattern to English, producing an output pattern that is more like an English segment than the input itself. Whereas the model will be able, with some difficulty, to interpret a segment that is well formed but did not happen to be part of the training inventory, it will not be able to perform this assimilation for an input vector consisting of an impossible combination of features (e.g., both back and bilabial). In short, the model will differentiate between the actual phonemes of a language, possible phonemes, and impossible phonemes.

We could then imagine several ways of extending this inquiry;

(a) We replicate the simulation using different training sets consisting of the phonemes of Spanish, Hebrew, French, etc. We compare the hidden unit representations that are created in these cases. We determine how the model represents language-specific versus universal aspects of phonology.

(b) We replicate the simulation using vowels. We determine which sets of vowels are easier or harder to learn, or unlearnable. We then see if these results

are consistent with facts about the cross-linguistic distributions of vowels. (This suggestion is due to Kevin Russell.)

(c) Finally, we could determine how the hidden unit representation in these simulations relate to the representations proposed in theories such as Clements (1985) and Archangeli (1988). One possibility is that these theories provide good approximations of the underlying representations that the model develops.

2. Learnability and a priori constraints

The Chomskyan theory of language acquisition emphasizes the role of the child's biological endowment: the child is born with capacities that make language learning possible and determine the course of acquisition. Connectionism is often held to be incompatible with this approach insofar as it employs powerful learning mechanisms. Although these learning algorithms do in fact have important implications for understanding language acquisition (see below), it would be a mistake to overlook the role of *a priori* constraints in connectionist networks. Any given model's capacity to learn is quite strictly determined by its architecture and other aspects of its initial configuration. To the extent that these starting conditions are made to reflect capacities that the child brings to bear on mastering a task, the network can be seen as providing a way to explore how biological (or other types of) constraints govern acquisition. In particular one could explore arguments to the effect that languages would not be learnable under the conditions that they are unless biological constraints contributed to the acquisition process in specific ways. Although I am not aware of any serious attempts to use connectionist networks to make learnability arguments, I think that the point can be established by considering some recent work on a seemingly unrelated topic: memory.

In a well-known article, McCloskey and Cohen (1989) suggested that simple feedforward networks exhibit a property that makes them unsuitable as models of human learning and memory. Take an autoassociative net such as the one discussed above and train it on a set of arbitrary patterns. For the purposes of this discussion assume that the patterns are nonsense syllable such as NAK, TOB, and SIM. As before, the network's task is to encode each input pattern and recreate it on the output units, having passed through a smaller number of interlevel hidden units. We train the model until it can perform this task for ten patterns (i.e., until the model has "memorized" them). We then train the model on a different set of ten patterns until it has memorized this second set. Finally, we retest the model on the first ten patterns to see what it remembers about them. McCloskey and Cohen observed that backpropagation nets exhibit considerable forgetting under these circumstances, because the changes to the weights that are relevant to learning the second set of patterns interfere with retention of the first set. The

networks were said to exhibit a "catastrophic" degree of interference – an unattractive property insofar as people's memories do not seem vulnerable in this way. For example, adding words to one's vocabulary does not result in extreme unlearning of words acquired earlier.

McCloskey and Cohen's simulations (which concerned both list learning and learning simple arithmetic problems) and later work by Ratcliff (1990) seem to indicate that a certain class of connectionist networks has an intrinsic property that is not very human-like. However, the occurrence of these "catastrophic interference" effects depends on some critical features of the simulations. First, there is the nature of the training regime. Catastrophic interference occurs when training is strictly blocked: first one set of patterns is learned and then a completely different second set is trained without any re-exposure to members of the initial set. It is hard to imagine any interesting aspect of learning in the real world that has this character. In learning to add, for example, the child is not taught all and only problems involving the number 1 until performance is perfect, then all and only the problems involving 2 until performance is perfect on them, then the problems involving 3, and so on, although that is how McCloskey and Cohen trained their network. The interfering effects of later learning on earlier learning are greatly reduced merely by relaxing this strict blocking of training, for example, by providing occasional retraining on earlier problems or using overlapping sets of problems (Hetherington & Seidenberg, 1989). These ways of training the network also correspond better to how children actually learn things like arithmetic, suggesting that having a better theory of how the human skill is acquired leads to better simulations of this behavior.[2]

A second property of the McCloskey and Cohen simulations is relevant to the question of *a priori* constraints. Consider the model that learned lists of nonsense syllables. This model was actually being asked to perform several tasks simultaneously. The task of primary interest was memorizing the patterns in the lists. However, the model also had to learn about structure of the stimulus patterns, for example, the fact that they consisted of syllables, and the fact that the syllables consisted of certain phonemes. McCloskey and Cohen's model was utterly tabula rasa; it was initialized with random weights and knew nothing about the phonological space from which the syllables were drawn. It was then supposed to memorize a set of patterns that happened to be CVCs consisting of certain phonemes. Consider how this situation differs from that of the subject in the 1950s verbal learning experiment on which the simulation is based; the subject brought to the experiment thorough knowledge of the set of phonemes in English

[2]I can think of only one case in which human learning is a strictly blocked as in the McCloskey and Cohen simulations: that would be verbal learning experiments in which subjects have to learn lists of nonsense syllables, random paired-associates, and other meaningless stimuli. Of course, the fact that this kind of verbal learning research is largely irrelevant to how people learn language or arithmetic or anything else of consequence is what brought it virtually to an end about 25 years ago.

and the phonotactics of syllabic structure. In contrast to the model, the subject only had to solve one problem, memorizing the particular nonsense syllables that were presented.

Two graduate students in my laboratory, Ken McRae and Phil Hetherington, showed very nicely that catastrophic interference is eliminated by separating the task of learning about the structure of the stimulus space from the task of learning the specific stimuli in the experiment. McRae and Hetherington (1993) initially taught their network about the structure of syllables *in general*. Once the network was pre-wired with this phonotactic knowledge, they trained it on the lists of nonsense syllables, which it was able to learn without exhibiting massive retroactive interference. Here again, having a better theory of how a task is learned allows one to build a better simulation.

The important point to note is that this solution to the catastrophic interference problem involved accurately assessing the state of the system *at the start of the learning process* (call this S_1). The model was able to learn and retain the lists only when it was *not* tabula rasa at S_1; rather, it already possessed knowledge of phonological structure. McRae and Hetherington happened to imbue the model with this knowledge by pre-training it on a set of phonologically well-formed nonsense syllables. Assume, for the sake of argument, that the child happens to be born with this knowledge of phonological structure. Under this interpretation, the simulation illustrates how innate structure of a specific sort is a precondition for acquiring a specific kind of knowledge.[3]

One can imagine other cases in which the initial state of the system (model/person) restricts the types of knowledge that can be acquired. Consider again the phonological simulations in section 1. What if – as seems likely – the range of phonological segments that people can learn to discriminate is constrained by facts about auditory perception and articulation? We now configure the phonological acquisition network so as to reflect these innate constraints (for example, by pruning connections between units that would otherwise produce impossible combinations of features). The model no longer has to learn about what a possible segment is; it only has to learn the structure of the segments in the language to which it is exposed. Constrained in the appropriate ways, the model would not be susceptible to learning impossible segments.

In summary, connectionist models are not merely compatible with the idea that

[3]Note that although it is important to understand the factors that determine the extent of interference in backpropagation nets, there is no reason to demand that such nets capture all aspects of human memory, especially in light of neurophysiological and neuropsychological evidence suggesting that human memory fractionates into components that may be subserved by different kinds of learning (see Gluck & Granger, 1993; McNaughton, McClelland, & O'Reilly, 1993, for discussion). A task such as learning a sequence of nonsense syllables may involve a more episodic, explicit type of learning than the gradual, implicit learning achieved by backpropagation networks. I am grateful to Dave Corina for reminding me of this point.

the initial state of the organism affects what can be learned; they provide a way to rigorously assess exactly where such constraints are necessary. Given an adequate theory of the structure of language, the input to the child, and the child's learning capacities, it should be possible to determine what kinds of innate constraints make language learnable under the conditions that they are. This moves us far beyond the return of the empiricist dybbuk that many linguists had envisaged (see Seidenberg, 1992, for further discussion).

3. The poverty of the "poverty of the stimulus argument" argument

The poverty of the stimulus argument is part of the foundation of modern generative linguistics. Summarizing familiar material very briefly, the argument runs as follows (see Atkinson, 1992; Chomsky, 1965; Lightfoot, 1982).

(a) The grammar cannot be learned on the basis of experience because the child is exposed to both grammatical and ungrammatical utterances. As Lightfoot (1982, p. 16) notes, "If only 5 percent of the expressions the child hears are of this [ill-formed] type, there will be a significant problem in generalizing to the set of grammatical sentences of the language because the pseudosentences do not come labeled as defective."

(b) The child encounters a finite set of utterances but comes to be able to comprehend and produce an infinite range of sentences. Also, children are exposed to different samples of utterances but converge on the same grammar.

(c) Children come to know things about language for which there is no direct evidence in the input. Thus, for example, they know that "Who does John believe that kissed Mary" is ungrammatical without being exposed to anything remotely like this structure.

(d) The theory that Universal Grammar (UG) is part of the child's biological endowment provides an explanation for how languages can be learned; no alternative theory that is even descriptively adequate has ever been proposed.

Further reason to assume the correctness of the poverty of the stimulus argument is provided by the observation that:

(e) The assumption that none of the important aspects of linguistic structure are learned is a strong theory and developing strong theories to whatever extent they can be is how science progresses.

Belief in the essential correctness of this argument is so complete that it has become axiomatic. Thus, for example, Pesetzky (1990) observed:

> The theory of grammar, if it is to shed light on language acquisition, should put as small a burden as possible on learning from experience and as great a burden as possible on general principles of UG. While it is clear the the role of experience is not zero, the null hypothesis should place it as close to zero as possible. (p. 2)

Developments within connectionism have not demonstrated that the poverty of the stimulus argument is incorrect. What they have done is suggest that the validity of particular parts of the argument need to be examined very carefully – something that would not have seemed worth doing even a few years ago. Insofar as any significant change in the validity of the poverty of the stimulus argument would have vast implications for syntactic theory and the practice of theoretical linguistics, it is worth considering the potential relevance of some of these recent developments.

That language acquisition depends on the child's biological endowment, the nature of the input to the child and the child's capacity to learn is a fact – one that every candidate theory must accommodate. Theories differ, notoriously, in terms of the weights they assign to these components. Connectionism has the potential to alter these weights because it entails some new ideas (and the rediscovery of some old ideas) about learning. For example, the vast amount of research on various learning algorithms calls into question the intuitions underlying (a). It is demonstrably true that networks can learn to solve problems in the face of inconsistent training. For example, a model could learn to categorize objects as animate or inanimate even if it were given incorrect feedback on 5% of the trials. This robustness in the face of inconsistencies in the input is thought to be an attractive property of such nets. More interesting, however, is that analysis of these learning systems suggests that at least some inconsistencies in feedback or network behavior might actually *facilitate* finding the solution to a problem. It is known, for example, that adding noise to the activation function can be useful in helping a network escape local optima (Hinton, 1989; Zipser, 1991). Or as learnability theory would have it, noise helps the network escape from incorrect generalizations without direct negative evidence. These results are certainly inconsistent with the intuition underlying (a). The implication is not that (a) is necessarily wrong or that another theory of language acquisition is necessarily correct. The fact that some networks can solve some problems in the face of a certain amount of noise does not entail that they can solve the language acquisition problem in the face of exactly the circumstances that confront the child. How such networks will fare with regard to language acquisition issues is not something that can be decided on the basis of inuitions such as (a), however.

The situation is similar with respect to (b). We do not have any networks that generate infinite sets of sentences. We do have networks that clearly demonstrate how nets develop knowledge representations that allow them to generalize beyond the patterns to which they are exposed. Thus, for example, Daugherty and Seidenberg (1992) described a simple feedforward net that mapped from the

present tense of an English verb to the past tense. The model was trained on 350 verbs, including both "rule-governed" items and exceptions (e.g., BAKE–BAKED, TAKE–TOOK). It learned the training set and was able to generalize to novel cases at a very high level of accuracy (e.g., given NUST, it said NUSTED). If we replicate the simulation using a different training corpus (i.e., expose the model to a different subset of verbs), it will converge on the same knowledge. Again, the argument is not that (b) is necessarily wrong. Rather, it is that it is based on intuitions that are called into question by these developments outside of linguistic theory.[4]

Point (c) is widely assumed on the basis of famous studies such as Brown and Hanlon (1970) and famous anecdotes such as the one reported by McNeill (1966; "nobody don't likes me"). But, there is a growing body of research devoted to uncovering systematic aspects of parental speech to children (e.g., Fernald, 1984, 1989; Hirsch-Pasek, Treiman, & Schneiderman, 1984; Kelly, 1992). I am reminded here of the fate of Piaget's theory that children at certain stages in development lack cognitive capacities such as object permanence or conservation of quantity. Most of these claims have been systematically refuted by 15 years of careful empirical studies (see, for example, Baillargeon, 1987; Gelman & Baillergeon, 1983). Obtaining the critical behavioral evidence demanded ingenious methodological innovations that permitted researchers to detect regularities in children's behavior that had gone undetected in earlier research. We are now seeing the same sort of approach being used to assess the claim that parental speech contains little in the way of useful clues to grammatical structure. What role, if any, these recently discovered aspects of parental speech to children play in acquisition is largely unknown. Connectionist networks, which are able to induce far more structure from far noisier input than the poverty of the stimulus argument assumes is possible, suggest that even partial, intermittent cues to grammatical structure could be exploited, however. As before, the claim is not that (c) is wrong, only that the prior probabilities regarding its correctness have shifted considerably, warranting appropriate changes in the course of research.[5]

With regard to (d), I would claim that connectionism has indeed greatly expanded the range of potential explanatory devices. And (e) is a perfectly reasonable strategy but one that might be more appropriate for other kinds of problems. If it were unethical to study parents' speech to children, for example,

[4]The Seidenberg and McClelland (1989) model also illustrated this ability to generalize beyond the training set; based on exposure to 2900 monosyllabic words, it correctly generalized to nonwords such as FIKE and NUST, although it performed worse than people on difficult items such as FAIJE and JINJE. A subsequent version of the model utilizing an improved phonological representation (Plaut, McClelland, & Seidenberg, 1992) pronounces even the difficult items at a rate comparable to people.

[5]Of course, everyone gets to play this game; thus, other language acquisition research is directed at showing that even very young children possess certain kinds of linguistic capacities, which can only be observed by using relatively subtle experimental methods (e.g., Chien & Wexler, 1991).

or if there were no methods available for this purpose, the "strongest theory" heuristic might provide the only way to deduce the role of parental input. It is harder to justify when the only boundaries on what can be learned about the input to the child are self-imposed.

In short, these developments suggest that the theory of grammar, if it is to shed light on language acquisition, should assign exactly as much of the burden to experience as the facts about the input to the child and the child's ability to learn dictate, not the amount assumed by the poverty of the stimulus argument. I realize that many people find it difficult to distinguish the claim that recent research raises questions about the role of experience in language acquisition from rabid anti-nativism (I know because these people attend my talks). As I have tried to stress, it is not that there is no biological endowment relevant to language (see, for example, Seidenberg & Petitto, 1987, for a proposal concerning one component of this endowment) or even that specific nativist claims (e.g., about the inventory of parameters) are wrong. The problem isn't that the poverty of the stimulus argument is provably incorrect; rather, it is that while it originally had the positive effect of orienting the field to critical issues – ones that existing approaches had not acknowledged to be important – it now serves exactly the opposite function, directing attention away from issues that need to be addressed.[6]

4. Explanatory value

Thus far I have discussed some ways in which connectionist concepts may be relevant to some central issues in the study of language: the nature of linguistic representations, and how the child's acquisition of language is affected by innate capacities and by experience. I now want to consider how these concepts may contribute to developing explanatory theories.

The article by Pinker and Prince (1988) in this journal articulated a specific view of the relationship between connectionist models and linguistic theory. According to this view, the important generalizations about linguistic phenomena are captured by symbolic theories in which rules figure centrally (i.e., by grammatical theories of the standard sort). The function of connectionist

[6]A considerable amount of research on sentence processing indicates that people encode detailed information about the statistical properties of language and that this information plays an essential role in comprehension (MacDonald, Pearlmutter, & Seidenberg, in press). For example, people's representations of language include information argument structures for verbs (MacDonald, 1993), the frequencies with which nouns are used as the heads of NPs (MacDonald, in press), and the frequency with which *that* is used as a complementizer at the beginning of a sentence versus a later position (Juliano & Tanenhaus, 1993). How this knowledge is acquired and its role in the acquisition of complex syntactic knowledge are not known. The fact that children do not hear sentences including the sequence "verb that verb [past]" could provide evidence relevant to judging the grammaticality of "Who does John believe that kissed Mary?", however.

networks is to show how grammatical knowledge is represented in neural machinery. Such a neural network could be said to implement a grammatical theory, but would add nothing to the grammar's capacity to capture generalizations about language. The phenomena used to illustrate this theory concerned past tense morphology in English (see also Kim, Pinker, Prince, & Prasada, 1991; Marcus et al., 1992; Pinker, 1991; Prasada & Pinker, 1993).

An alternative view holds that the Pinker and Prince account is backwards (Seidenberg, 1992). The connectionist network is not in fact an implementation of a rule-based theory because its behavior deviates from what would be expected on the basis of such theories. These deviations represent novel predictions about behavior, which will prove to be correct or incorrect but are not identical to those of the rule-based theory. The function of the rule-based theory is to provide a folk psychological account of the network's behavior. This level of description is useful but misses certain generalizations that can only be stated at the level of network behavior. This view can also be illustrated with respect to past tense morphology (Daugherty & Seidenberg, 1992; in press; Daugherty, MacDonald, Petersen, & Seidenberg, 1993; MacWhinney & Leinbach, 1991; Plunkett & Marchman, 1993; Seidenberg, 1992). Summarizing this work briefly, it suggests the following:

(a) Pinker and colleagues have taken certain types of phenomena as evidence for the rule-governed basis of the past tense and suggested that they are inconsistent with properties of neural nets, but they are wrong insofar as the cited phenomena are quite compatible with known properties of relatively simple nets. Thus, for example, the fact the frequency affects the generation of irregular past tenses was taken as evidence that they are listed in the lexicon and the fact that frequency does not affect the generation of regular past tenses as evidence that they are generated by rule (Prasada & Pinker, 1993), but both effects are generated by simple feedforward networks (Daugherty & Seidenberg, 1992; Seidenberg & McClelland, 1989). Similarly, Marcus et al. (1993) assert that certain facts about the German plural point to the involvement of a rule and cannot be accommodated by neural nets, but both claims are called into question by Daugherty and Hare (1993), who describe a net that solves this type of problem.[7] Pinker and colleagues have identified numerous important facts about inflectional morphology, but the conclusion that these facts are incompatible with connectionist networks does not follow.

(b) The problem with connectionist treatments of inflectional morphology

[7]Marcus et al. observe that there are rule-governed German plurals, but they are less frequent than the "irregular" plurals. Because connectionist networks induce rules on the basis of frequency of exposure to patterns, they assert that they cannot learn this kind of low-frequency default rule. The analysis is flawed, however, because the irregular German plurals fall into semi-productive phonological patterns that also influence what a network learns. Daugherty and Hare demonstrate this in a simulation of some phenomena concerning historical change in the past tense system of English.

cited by Pinker and colleagues concern the specific model proposed by Rumelhart and McClelland (1986); they are not inherent in the broader framework and establish no bounds on what can be done within it.

(c) The principles that govern the past tense are not about the past tense in particular or about language in general; rather they are about the way in which knowledge is represented in certain types of neural nets (Seidenberg, 1992). The strongest evidence for this claim is the observation that the same principles have been observed to govern behavior in a completely unrelated domain: learning the correspondences between spelling and pronunciation in English (Seidenberg & McClelland, 1989). In both domains people's knowledge includes information about the frequency and consistency of mappings between codes (present and past tense; spelling and pronunciation); these factors also affect generalization to novel instances; there is a specific pattern of interaction between the frequency and consistency factors; the relevant knowledge can be represented in terms of multilayer networks employing distributed representations and weighted connections between units. The view that the past tense is a paradigmatic example of a linguistic subsystem (Pinker, 1991) misses these generalizations.[8]

In summary, the past tense has provided a domain in which to illustrate how connectionist models contribute to the development of explanatory theories. The models simulate detailed aspects of human behavior (e.g., in generating the past tense) and make novel predictions that have been confirmed in subsequent studies. The models are based on independently motivated principles concerning knowledge representation and learning. The principles that govern such models are not specific to the past tense or to language. Whereas saying that the network "implements" the rule-based theory ignores crucial ways in which the two accounts differ, saying that the rule-based theory provides a folk psychological account of the behavior of the net seems quite accurate.

Of course, the issues concerning the proper treatment of the past tense are far from settled; many technical issues remain unresolved. However, if it turns out that inflectional morphology represents a set of phenomena for which neural networks happen to be well suited, it could be argued that nothing substantive

[8]The phenomena in both domains are consistent with Kiparsky's (1982) Elsewhere condition governing the application of rules; importantly, the neural network models of the phenomena behave in an analogous manner. The Elsewhere condition imposes an ordering on rule application, such that a rule with a more limited range of application is applied before one with a broader range. This is sometimes thought to reflect a deep property of language, but it is observed in other domains. Moreover, this behavior falls out of neural networks very easily (see Corina, in press, for discussion). The Elsewhere condition is relevant to the past tense if we think of irregular cases such as SING–SANG/RING–RANG as ones in which there is a rule with a more limited range than the default rule. Models such as the one developed by Daugherty and Seidenberg (1992) produce these irregular past tenses where appropriate, but otherwise default to the rule. Presented with SING, the model produces SANG as the past tense, but presented with BING it will say BINGED. This is an automatic consequence of how knowledge is represented in such networks, not something that has to be built in.

follows about the nature of language. The past tense in English is a rather simple system with characteristics that are not shared by more interesting aspects of language. In particular, it involves a single morphological rule and while there are exceptions to it, they are finite in number and can be learned by rote, at least in principle. Thus, Pinker's (1991) account of the past tense involves two components: the rule and a separate "associative net" for representing the exceptions. That past tense formation is rule governed is claimed as a major discovery, but this could not fail to be true given that the rule does not have to apply in all cases because a second mechanism is invoked to explain the exceptions to it (see Seidenberg, 1992, for discussion). There are two types of data to be explained (rule-governed cases and exceptions) and two explanatory devices in the theory (a rule and an associative net).[9] These observations again suggest that the past tense might not provide the ideal example of a rule-governed linguistic phenomenon; it therefore might not be hugely surprising to discover that a neural network might also accommodate these phenomena. But what about other aspects of language? Clearly, no one thinks of knowledge of syntax as involving a set of rules and a list of exceptions to them that can be learned by rote. Perhaps the place to draw the line is not between connectionism and inflectional morphology but rather between connectionism and syntax.

Perhaps, but it seems to me that some of the same issues arise with regard to syntax. In Pinker's morphological theory, it is possible to formulate a rule because there is the option of excluding the exceptions to it. In syntactic theory, it is possible to formulate generalizations because there is the option of excluding unclear cases. Thus, Chomsky has suggested that one of the functions of a proper grammatical theory should be to disambiguate cases where grammaticality intuitions are unclear. Moreover, he reminds us that there are reasons to retain a grammatical theory in the face of apparent counterexamples. Given these additional degrees of freedom, it would be surprising if one could *not* detect regularities and describe them in terms of rules. One reason syntactic theory can be so "strong" is because there is a high level of tolerance for deviant phenomena.

These issues to not arise in the kind of connectionist research I have described. The goal of this research is not to develop a competence grammar. Rather, it is to understand how language is used in performing tasks such as speaking and comprehending. As in the standard approach, this goal entails understanding how

[9]The same issue arises with regard to Coltheart et al.'s (1993) recent observations about the need for rules in generating the pronunciations of words from print. They describe an algorithm that is able to extract rules governing spelling–sound correspondence on the basis of exposure to 2900 words. Success in this endeavor is built into the algorithm, because the 20% of the corpus whose pronunciations cannot be generated by the rules are treated separately (i.e., excluded). As in the Pinker model, there is no independent evidence that these are all and only the words that people treat as exceptions to the rules.

innate structures and experience contribute to the acquisition of these capacities. Linguistic representations do not reflect generalizations derived from samples of adult utterances drawn from within and across languages; rather, they develop in the service of solving the acquisition problem. Once such knowledge is acquired, it can be called upon in performing other tasks, such as judging the well-formedness of utterances. Thus, the phonological acquisition device described earlier was designed to solve a particular problem; it was preconfigured with certain innate capacities; in the course of solving the problem it developed abstract underlying representations. Once the model has mastered the primary task, it can be used to judge the well-formedness of phonological segments, though that was not its original purpose.[10]

At this point we could also choose to develop a theory of how the model performs this task. There are two ways such a theory might be developed. One would be to treat the model as a black box and formulate generalizations based on its behavioral output (e.g., in the well-formedness task). This would involve trying to infer the structure of what is in the box from behavioral evidence of a specific sort; this is the standard linguistic approach. The other alternative would be to develop the theory by examining how the model actually works. This would involve looking inside the box and drawing on knowledge of how other such boxes function in order to identify general principles governing its behavior.

In summary, I think there is reason to believe that the connectionist approach has more to offer than merely showing how grammars are represented in the brain. It faces enormous conceptual and computational problems, but any interesting approach would. If nothing else, rigorously determining what types of linguistic phenomena can and cannot be explained within this approach seems to be an important thing to do.

Why it will never work: the compleat connectionist critic

Unlike implemented models, thought simulations and speculation are not subject to simple refutation and so in the interest of fairness and balance, I probably should also provide the appropriate critical analyses of these views. What follows are all-purpose complaints that have already proven to have wide applicability.

(a) "The models will never be able to do *this*" [where "this" is behave in accord with some favored aspect of language]. Thus, no connectionist model will

[10]I should stress that this net is merely an example used to illustrate some general points, not a serious claim about how children acquire phonological representations. I see the child's problem as having to develop internal representations that mediate between auditory and articulatory speech codes; that is not a simple autoencoder problem.

be able to generate both rule-governed cases and exceptions with good generalization (Coltheart, Curtis, Atkins, & Haller, 1993; Prasada & Pinker, 1993), or induce a low-frequency default rule (Marcus et al., 1993). The risk in invoking (a) is that, in the absence of any formal proof of its validity, someone could develop such models tomorrow (see, for example, Daugherty & Hare, 1993; Plant, McClelland, & Seidenberg, 1992).

(b) "The models will always do *that*" [where "that" is exhibit a kind of behavior that is grossly unlike that of people]. Thus, backpropagation nets always exhibit catastrophic interference (McCloskey & Cohen, 1989). The risk here is that someone will develop models that do not (see, for example, McRae & Hetherington, 1993; Murre, 1992).

(c) "The models can do anything." Connectionist models are so powerful that they can solve any problem, both ones that people solve and ones that they can't. Some care should be exercised to avoid invoking (c) in conjunction with (a) or (b). Moreover, the question is really whether they exhibit appropriate sorts of behaviors when constrained in exactly the correct, independently motivated ways. This remains to be determined.

References

Archangeli, D. (1988). Aspects of underspecification theory. *Phonology*, *5*, 183–207.

Atkinson, M. (1992). *Children's syntax*. Oxford: Blackwell.

Baillargeon, R., (1987). Object permanence in $3\frac{1}{2}$ and $4\frac{1}{2}$ month old infants. *Developmental Psychology*, *23*, 655–664.

Brown, R., & Hanlon, C. (1970). Derivational complexity and order of acquisition in child speech. In J.R. Hayes (Ed.), *Cognition and the development of language*. New York: Wiley.

Chien, Y.-C., & Wexler, K. (1991). Children's knowledge of locality conditions in binding as evidence for the modularity of syntax and pragmatics. *Language Acquisition*, *1*, 225–295.

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

Clements, G.N. (1985). On the geometry of phonological features. *Phonology Yearbook*, *2*, 225–252.

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

Corina, D.P. (in press). The induction of prosodic constraints: Implications for phonological theory and mental representation. In S. Lima (Ed.), *The reality of linguistic rules*. Philadelphia: John Benjamins.

Cottrell, G.W., Munro, P., & Zipser, D. (1988). Image compression by back propagation: An example of extensional programming. In N.E. Sharkey (Ed.), *Advances in cognitive science* (Vol. 3). Norwood, NJ: Ablex.

Daugherty, K., & Hare, M. (1993). What's in a rule? The past tense by some other name might be called a connectionist net. In M. Mozer, P. Smolensky, D. Touretzky, J. Elman, & A. Weigand (Eds.), *Proceedings of the 1993 Connectionist Models Summer School*. Hillsdale, NJ: Erlbaum.

Daugherty, K., MacDonald, M., Petersen, A., & Seidenberg, M. (1993). Why no mere mortal has ever flown out to left field, but often people say they do. *Proceedings of the 15th Annual Meeting of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.

Daugherty, K., & Seidenberg, M.S. (1992). Rules or connections? The past tense revisited. *Proceedings of the 14th Annual Meeting of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.

Daugherty, K., & Seidenberg, M.S. (in press). Beyond rules and exceptions: A connectionist approach to inflectional morphology. In R. Corrigan, G. Iverson, & S. Lima (Eds.), *The reality of linguistic rules*. Philadelphia: John Benjamins Press.

Fernald, A. (1984). The perceptual and affective salience of mothers' speech to infants. In L. Feagans, C. Garvey, & R. Golinkoff (Eds.), *The origins and growth of communication*. Norwood, NJ: Ablex.

Fernald, A. (1989). Intonation and communication intent in mothers' speech to infants: Is the melody the message? *Developmental Psychology, 60*, 1497–1510.

Fodor, J., & Pylyshyn, Z. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition, 28*, 3–71.

Fromkin, V., & Rodman, R. (1978). *An introduction to language*. New York: Holt, Rinehart & Winston.

Gelman, R., & Baillergeon, R. (1983). A review of some Piagetian concepts. in P.H. Mussen (Ed), *Handbook of child psychology: Vol. 3. Cognitive development* (pp. 167–230). New York: Wiley.

Gluck, M., & Granger, R. (1993). Computational models of the neural bases of learning and memory. *Annual Review of Neuroscience, 16*, 667–706.

Goldsmith, J. (1993). *The last phonological rule*. Chicago: University of Chicago Press.

Gupta, P., & Mozer, M. (1993). Exploring the nature and development of phonological representations. *Proceedings of the 15th Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.

Hare, M., & Elman, J. (1992). A connectionist account of English inflectional morphology: Evidence from language change. *Proceedings of the 14th Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.

Hetherington, P., & Seidenberg, M.S. (1989). Is there "catastrophic inference" in the connectionist networks? *Proceedings of the 11th Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.

Hinton, G. (1989). Connectionist learning procedures. *Artificial Intelligence, 40*, 185–234.

Hirsh-Pasek, K., Treiman, R., & Schneiderman, M. (1984). Brown and Hanlon revisited: Mothers' sensitivity to ungrammatical forms. *Journal of Child Language, 11*, 81–88.

Juliano, C., & Tanenhaus, M.K. (1993). Contingent frequency effects in syntactic ambiguity resolution. *Proceedings of the 15th Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.

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

Kim, J.J., Pinker, S., Prince, A., & Prasada, S. (1991). Why no mere mortal has ever flown out to center field. *Cognitive Science, 15*, 73–218.

Kiparsky, P. (1982). *Explanation in phonology*. Dordrecht: Foris.

Lachter, J., & Bever, T.G. (1988). The relation between linguistic structure and theories of language learning: A constructive critique of some connectionist learning models. *Cognition, 28*, 195–247.

Lightfoot, D. (1982). *The language lottery*. Cambridge, MA: MIT Press.

MacDonald, M.C. (1993). The interaction of lexical and syntactic ambiguity. *Journal of Memory and Language, 32*, 692–715.

MacDonald, M.C. (in press). Probabilistic constraints and syntactic ambiguity resolution. *Language and Cognitive Processes*.

MacDonald, M., Pearlmutter, N., & Seidenberg, M.S. (in press). Syntactic ambiguity resolution as lexical ambiguity resolution. In C. Clifton, L. Frazier, & K. Rayner (Eds.), *Perspectives on sentence processing*. Hillsdale, NJ: Erlbaum.

MacWhinney, B., & Leinbach, J. (1991). Implementations are not conceptualizations: Revising the verb learning model. *Cognition, 40*, 121–157.

Marcus, G., Brinkman, U., Clahsen, H., Wiese, R., Woest, A., & Pinker, S. (1993). *German inflection: The exception that proves the rule*. MIT Department of Brain and Cognitive Sciences Occasional Paper #47.

Marcus, G., Pinker, S., Ullman, M., Hollander, M., Rosen, T., & Xu, F. (1992). Overregularization in language acquisition. *Monographs of the SRCD*. Chicago: University of Chicago Press.

McClelland, J.L., McNaughton, B.L., & O'Reilly, R.C. (1991). Why do we have a special learning system in the hippocampus? *Bulletin of the Psychonomic Society*, Abstract 580, *31*, 404.

McCloskey, M., & Cohen, N.J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. In G.H. Bower (Ed.), *The psychology of learning and motivation*. New York: Academic Press.

McNeill, D. (1966). Developmental psycholinguistics. In F. Smith & G.A. Miller, (Eds.), *The genesis of language*. Cambridge, MA: MIT Press.

McRae, K., & Hetherington, P. (1993). Catastrophic interference is eliminated in pre-trained nets. *Proceedings of the 15th Annual Meeting of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.

Murre, J.M.J. (1992). *Learning and categorization in modular neural networks*. Hillsdale, NJ: Erlbaum.

Pesetzky, D. (1990). *Experiencer predicates and universal alignment principles*. Manuscript, MIT Department of Linguistics and Philosophy.

Pinker, S. (1991). Rules of language. *Science*, *253*, 530–534.

Pinker, S., & Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, *28*, 73–194.

Plaut, D., McClelland, J., & Seidenberg, M.S. (1992). *Word pronunciation in reading: Are two routes really necessary*? Paper presented at the annual meeting of the Psychonomic Society.

Plunkett, K., & Marchman, V. (1993). From rote learning to system building. *Cognition*, *48*, 21–69.

Prasada, S., & Pinker, S. (1993). Generalization of regular and irregular morphological patterns. *Language and Cognitive Processes*, *8*, 1–56.

Prince, A., & Smolensky, P. (in press). *Optimality theory*. Cambridge, MA: MIT Press.

Ratcliff, R. (1990). Connectionist models of recognition memory: Constraints imposed by learning and forgetting functions. *Psychological Review*, *97*, 285–308.

Rumelhart, D.E., & McClelland, J.L. (1986). On learning the past tenses of English verbs. In J. McClelland & D. Rumelhart (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. 2). Cambridge, MA: MIT Press.

Seidenberg, M.S. (1992). Connectionism without tears. In S. Davis (Ed.), *Connectionism: Advances in theory and practice*. Oxford: Oxford University Press.

Seidenberg, M.S., & McClelland, J.L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review*, *96*, 447–452.

Seidenberg, M.S., & Petitto, L.A. (1987). Communication, symbolic communication, and language. *Journal of Experimental Psychology: General*, *116*, 279–287.

Zipser, D. (1991). Recurrent network model of the neural mechanism of short-term active memory. *Neural Computation*, *3*, 179–193.