SSLA, **12**, 179–199. Printed in the United States of America.

CONNECTIONISM AND UNIVERSALS OF SECOND LANGUAGE ACQUISITION

Michael Gasser
Indiana University

This article examines the implications of connectionist models of cognition for second language theory. Connectionism offers a challenge to the symbolic models which dominate cognitive science. In connectionist models all knowledge is embodied in a network of simple processing units joined by connections which are strengthened or weakened in response to regularities in input patterns. These models avoid the brittleness of symbolic approaches, and they exhibit rule-like behavior without explicit rules. A connectionist framework is proposed within which hypotheses about second language acquisition can be tested. Inputs and outputs are patterns of activation on units representing both form and meaning. Learning consists of the unsupervised association of pattern elements with one another. A network is first trained on a set of first language patterns and then exposed to a set of second language patterns with the same meanings. Several simulations of constituent-order transfer within this framework are discussed.

CONNECTIONISM

In the past ten years, cognitive science has seen the rapid rise of interest in *connectionist* models, theories of the mind based on the interaction of large numbers of simple neuron-like processing units. The approach has already reshaped the way many cognitive scientists think about mental representations, processing, and learning. Connectionism offers a challenge to traditional *symbolic* models of cognition. Despite the powerful appeal of symbols, rules, and logic, the traditional view suffers from a very unhuman-like brittleness. Linguistic and conceptual entities are assigned in all-or-none fashion to categories, rules typically apply in a fixed sequence, and

I am indebted to Roger Andersen, Kathleen Bardovi-Harlig, Evelyn Hatch, Robert Port, and two anonymous reviewers for comments on this article.

deviations from expected patterns are not handled well, if at all. In connectionist models the brittleness is avoided because there are no discrete symbols and rules as such; the entities that a connectionist system uses to characterize the world are fluid patterns of activation across portions of a network. In addition, connectionism puts the emphasis back on learning in cognitive science. In symbolic models it is often assumed that it is enough to characterize a particular point in the process of acquisition. Most connectionists do not agree; it is how the system progresses from one state to another that is most interesting, and connectionists have developed a variety of new network learning algorithms to be studied and applied to particular problem domains.

No subfield of cognitive science, including second language acquisition research, can afford to ignore the implications of this new approach. While it is premature to speak of a connectionist theory of linguistic behavior let alone second language acquisition, it is possible to outline what connectionism may have to offer the field of second language research. This is the purpose of this article.

Computational Approaches to Cognition

In order to understand connectionism, it is necessary to put it in its context as a *computational* approach to the study of the mind. Like other models in the fields that make up cognitive science, what connectionist models seek to do is to describe cognitive processing in computational terms, that is, in terms of data structures and the processes that operate on them, yielding outputs from given inputs. With respect to the study of linguistic behavior, computational approaches differ from most generative models in not making a fundamental distinction between competence and performance (Winograd, 1983). What is of interest in a computational model of language is comprehension, production, and the representations, linguistic and otherwise, which enable these processes. Usually such models are simulated in computer programs, though it is important to note that for most cognitive scientists the computer itself is nothing more than a tool to test the validity of the models.[1]

Processing in Connectionist Models

Current connectionist models, also referred to as *neural networks* and *parallel distributed processing* (PDP) models, are related to pioneering work by neuroscientists and computer scientists in the 1940s and 1950s (McCulloch & Pitts, 1943; Rosenblatt, 1962), who were interested in the computational power of networks of simple neuron-like processing units. The recent resurgence of interest in these models has been spurred by the discovery of new learning algorithms as well as by dissatisfaction with the achievements of classical symbolic models of cognition. Work continues on the formal properties of networks of various types as well as on applications of these networks to areas as diverse as the detection of explosives in airline baggage (Shea & Lin, 1989), discovery of lexical classes from word order (Elman, 1988), and the use of scripts in story understanding (Dolan & Dyer, 1989).

There is not space in this article to do more than introduce the basic concepts

involved in connectionist models. For more in-depth discussions, see Rumelhart, McClelland, and PDP Research Group (1986).

Most connectionist models share the following basic features:

1. The system's memory consists of a network of simple processing units joined by weighted connections. Each weight is a quantity determining the degree to which the unit at the source end of the connection activates or inhibits the unit at the destination end of the connection.
2. The behavior of units is based loosely on that of neurons. They sum the inputs they receive on connections and compute an activation, which is a function of the total input, and an output, which is a function of the activation. A unit's output is passed along its output connections to other units. The current pattern of activation on the units in the system corresponds to short-term memory in more traditional models, and inputs and outputs to the system take the form of patterns of activation over groups of input and output units.
3. The analogue of long-term memory in other models is the set of weights on the network connections. In learning models, these weights are adjusted as a consequence of processing.
4. Processing is parallel. In most traditional models, as in conventional computers, decisions and actions are made one at a time. In connectionist models, as in the brain, there is activity in many places simultaneously.
5. Control is distributed. Unlike traditional cognitive models, connectionist systems have no central executive whose job it is to determine which rule or rules are currently applicable and to execute them. In fact, there are no rules to be executed.

Connectionist models divide into two basic categories: *localist* approaches (e.g., Cottrell, 1989; Feldman & Ballard, 1982; Gasser, 1988; Waltz & Pollack, 1985), in which units represent particular concepts, such as BLUE,[2] GLASNOST, ELVIS-PRESLEY, INANIMATE, and TRANSITIVE-CLAUSE; and *distributed* approaches (e.g., McClelland, Rumelhart, & PDP Research Group, 1986; Kanerva, 1989; and Rumelhart et al., 1986), in which complex concepts are distributed over many units, and each unit participates in the representation of many concepts. Because it is the distributed models which have attracted the most attention, are better suited for learning, and have the most radical claims to make, I will focus on them in this article.

The interesting properties of (distributed) connectionist networks include the following:

1. Robustness, graceful degradation: The systems do not break down when inputs are incomplete or errorful, or even when a portion of the network is destroyed.
2. Graded representations: The concepts that the systems acquire and make use of bear little resemblance to the discrete categories of traditional models. Things belong to connectionist *categories* to varying degrees, the representations continually evolve as the system learns, and concepts are free to blend in intricate ways.
3. Fixed memory size: Because knowledge is shared in the system's connections, the addition of new knowledge does not necessarily require new units and connections.
4. Automatic generalization, rule-like behavior: As connectionist systems learn about specific patterns, they are also building the knowledge that will allow them to handle a range of similar patterns. That is, they are making generalizations, possibly at many different levels

of abstraction. Unlike the rules of traditional models, however, these generalizations do not appear explicitly in the network. Rather, they arise as needed during processing.

5. Interaction of multiple sources of knowledge: Connectionist systems work by integrating information in the form of the parallel spread of activation in many parts of the network at once. This approach lends itself to modeling in domains where decisions are made on the basis of diverse sorts of knowledge.

Learning in Connectionist Models

Connectionist models, at least those of the distributed type, are typically empiricist accounts of cognition, and learning plays a central role. I will return to a characterization of the sort of empiricism favored by connectionists later.

Most connectionist models implement one form or another of *pattern association*. A pattern associator is a network which learns to associate input–output pairs, where each input or output consists of a pattern of activation over a set of input or output units. For example, a pattern associator might represent the tendencies for particular odors, represented on the input units, to result in particular visual images, represented on the output units (McClelland, Rumelhart, & Hinton, 1986). The modeler may assign particular significance to individual input and output units (typically in the form of *microfeatures* such as ANIMATE, INVOLVING-CONTACT, and the like), or patterns may be assigned in an arbitrary fashion to the particular concepts that the system is to be given. Associations between inputs and outputs are usually mediated by one or more layers of *hidden units*. These units are hidden in two senses. First, they have no access to the environment (i.e., they are neither inputs nor outputs). Second, they are not assigned any significance by the designer of the network; they develop their significance as the network learns to associate inputs and outputs.

Once the network has been trained on a set of mappings, it should yield the (approximately) correct output given an input. In the odor-to-image example, presentation of a pattern representing an odor on the input units should result in the activation of the appropriate visual pattern on the output units. Most importantly, the network may be able to yield an appropriate output for an input pattern on which it has not been trained if the pattern is similar enough to one or more familiar patterns. (Traditional cognitive scientists would speak here of the extraction of a rule.) Furthermore, because of redundancy that is automatically built in as the network learns, an incomplete or degraded input pattern may also yield an appropriate output.

Probably the most familiar example of a connectionist pattern associator is the NETTALK system (Sejnowski & Rosenberg, 1987). In NETTALK the inputs represent the written forms of English words and the outputs represent phonological representations of English words. The network is trained on a set of spelling–pronunciation pairs, and on the basis of this training it is able to "pronounce" not only those words on which it has been trained, but also a large set of unfamiliar words. At no point is the network given any rule that would help it out; in fact, the network would not know what to do with a rule if given one. What the network is doing is looking for regularities in the orthography–phonology pairings that are presented to it. Most interestingly, in the process of learning, the network arrives at some of the phonologi-

cal categories familiar to linguists. Using cluster analysis to investigate what the hidden layer in NETTALK is actually representing, it has been found, for example, that certain units respond more to vowel letters and others to consonant letters. Again it must be emphasized that these categories were not given to the network initially; they were, rather, implicit in the patterns themselves.

An important subtype of pattern association is *auto-association*. Here a pattern is associated with itself; that is, outputs are meant to be identical to inputs. A network trained in this way can perform *pattern completion*: given a portion of an input, the network can return the complete pattern. For example, the input (and output) units might represent aspects of a visual scene. Once trained on enough of these patterns, the system could be given a portion of a scene or a scene containing some incorrect information on the input units and could then generate the complete, correct scene on the output units. What makes the pattern completion idea so appealing is that it does not matter which portion of a pattern such a network is given, as long as enough information is provided. That is, processing can proceed in any direction. I will argue how this permits an auto-associative system to perform both comprehension and production using the same units.

Connectionism Versus Behaviorism

Some critics of connectionism (Fodor & Pylyshyn, 1988; Pinker & Prince, 1988) contend that it is no more than a revival of behaviorism dressed up to look like neuroscience. It is true that connectionist models share with behaviorism a focus on the learning of stimulus–response (or "input–output") associations. The differences lie in the concern of connectionists with the internal representations that are constructed between the inputs from and the outputs to the environment, and with the specific mental processes that are involved in the construction of these representations (Rumelhart & McClelland, 1986b). In addition, many (though by no means all) connectionist models involve feedback connections which would not be possible in a strict stimulus–response framework; and connectionists are also increasingly concerned with the initial structure of the networks they work with; that is, with what could be thought of as innate "knowledge" of a sort.

SYMBOLS, RULES, AND LANGUAGE

Linguistics and the Symbolic Paradigm

With the exception of recent work within the connectionist paradigm, all of cognitive science belongs to what has been referred to as the *symbolic paradigm* (Fodor & Pylyshyn, 1988; Newell, 1980; Pinker & Prince, 1988). Despite vast differences, these approaches agree on the validity of the basic distinction between the "software" of the mind and the "hardware" of the brain. Much as a program or a programming language is (ideally) independent of the type of computer that it runs on, the mind's programs and programming language(s) are said to be describable in terms that do not make reference to neural structures or processes. Cognition, in the symbolic view,

consists primarily in the application of rules. A rule-based system, requires (a) a central program to direct the system, (b) a sophisticated facility for pattern matching to select rules appropriate for given contexts, and (c) *symbols*, that is, tokens which denote other tokens or full-blown structures in memory. Symbols in rules play the role of variables; that is, what they denote depends on the particular context in which the rule is applied.

Modern formal linguistic theories are no exception to this dominant tendency in cognitive science. Like rules in typical artificial intelligence (AI) systems, linguistic rules make reference to structures such as trees and require variables. And, although this fact is not emphasized except in computational linguistics, it must be assumed that there is a central control guiding the process (whether the process is comprehension, production, or "derivation") and a mechanism for pattern matching to determine which rules apply to the current state of the system.

It is important to recognize that these basic aspects unite approaches which have previously been seen as radically different accounts of linguistic knowledge and behavior; for example, generative linguistic theories on the one hand and the natural language processing research associated with Roger Schank and his colleagues (e.g., Schank & Abelson, 1977) on the other.

Connectionism and the Subsymbolic Paradigm

Connectionists reject the basic premises of symbolic cognitive science, in particular the notion that the behavior of neurons is not relevant in accounting for cognition. While they differ in the extent to which they take the functioning of real neurons seriously (and none of their models could be said to be faithful representations of neural processes), they hold that:

1. The nature of the brain constrains mental processes in important ways; in particular, the relative slowness of the primitive operations of neurons forces one to conclude that the brain makes use of massive parallelism in processing (Feldman & Ballard, 1982); and
2. A neurally inspired type of processing provides a better account of what is known about mental processes than symbolic processing does, for reasons discussed earlier.

The highly constrained form of processing that characterizes connectionist models must operate without the symbols, the symbolic pattern matcher, and the central program which are required for rule-based processing. How then are connectionist models to account for behavior which is apparently rule-governed? Research is still underway on the details of the answer to this question, but the usual response of connectionists is that rules and symbols are "emergent" phenomena; that is, they arise out of the complex interactions of more primitive elements and processes. For this reason, connectionist approaches are said to define the *subsymbolic paradigm*. The usual connectionist argument is that while a symbolic characterization may provide a useful description of a phenomenon, it is to be understood as an approximation in the same sense as a characterization of a physical phenomenon can be approximated in Newtonian terms, though it is more accurately understood in terms of relativity and quantum mechanics.

Consider the formation of the past tense in English, the best-known example of an apparently rule-governed behavior which connectionists have succeeded in modeling without explicit rules. If there is one thing that formal linguists would agree on, it is that adult native speakers of English have a set of rules for past tense formation. This has seemed to everyone the only interpretation for the results of Berko's (1958) classic experiments demonstrating that children as young as 5 could correctly generate past tense forms for nonsense verbs. Yet two series of simulations (Plunkett & Marchman, 1989; Rumelhart & McClelland, 1986a) have shown that a simple connectionist network can learn to generate both regular and irregular past tense forms from verb stems without any explicit rule and to go through some of the same sorts of stages that human learners do. Most interestingly, the model goes through the three stages in the typical "U-shaped curve":

1. Past tense forms, including many irregulars, are simply memorized.
2. The regular "rule" begins to be acquired, and many irregular past tenses, including some previously formed correctly, are now formed according to the regular rule.
3. Both regulars and irregulars are formed correctly.

The network's rule-like behavior is the result of the combination of many associations of specific stem features with specific past tense features. These associations are implemented in the weights on the connections joining input (stem) and output (past tense) units. Each rule involves many weights, and each weight typically participates in many rules.

What takes the place of the symbols of traditional models? Doesn't there need to be a place or an invariant pattern in memory for concepts such as LEG, FRY, and VELAR? The answer is that in a connectionist system it is not the case that something either is a LEG or is not a LEG; rather, things are more or less LEGs. There might be a set of units which would tend to be activated strongly when a leg is perceived or imagined, but it would not be possible to draw a clear boundary around this set of units, and each of these units would also participate in the representation of other concepts. Thus when (more-or-less) LEG is active, any number of other concepts will also be partially active. Note that this approach permits metaphor and analogy to be treated as natural processes rather than as peripheral phenomena that are not amenable to formalization.

There is another important way in which subsymbolic models differ from symbolic ones. Much of the thinking in traditional cognitive science has revolved around distinctions between the general and the specific. Psychologists and AI researchers typically distinguish between instances (tokens) and classes (types). Linguists distinguish between grammar, embodying the general aspects of a language, and lexicon, embodying the specific aspects; and between regular (general) processes and irregular (more specific) ones.

For connectionists, these distinctions are of degree rather than kind. A pattern of activation over a set of units might represent a token or a whole class, the difference being the number of units involved; that is, the extent to which the characterized entity is specified. A set of connection weights normally embodies at once a number of rules of different degrees of specificity. For example, in the past tense models referred to above, there is no clear-cut distinction between the way in which the

regular past tense rule works and the way in which an "irregular" rule (e.g., cut–cut, beat–beat) works. This is one of the major distinctions that has been brought out between this model and traditional accounts of morphological processes (Pinker & Prince, 1988).

While these aspects of the subsymbolic paradigm do not endear it to generative linguists, they are consistent with ideas that have arisen in recent years in *cognitive linguistics* (Fauconnier, 1985; Fillmore, 1988; Lakoff, 1987; Langacker, 1987). Cognitive linguists and other like-minded cognitive scientists emphasize the fluidity of concepts (Hofstadter, 1985; Lakoff, 1987; Rosch, 1978), the continuous nature of the differences between rules and exceptions (Harris, 1989; Langacker, 1987), and the relative importance of the more specific end of the spectrum. As we have seen, these are also features of connectionist models of cognition.

In sum, connectionists have the goal of modeling cognitive processes without the use of symbols such as JOHN and X-BAR, without explicit rules associating inputs with outputs, and without distinctions between general and specific concepts and processes.

It is by now clear that some form of connectionism will figure in a general model of human linguistic behavior. The only question is whether the role will be a minor one, relegated to "low-level" pattern matching tasks and the learning of exceptional behavior, or whether the connectionist account will supersede symbolic accounts, rendering them nothing more than neat approximations of the actual messy process. Thus far, connectionist research on language has been most convincing in demonstrating the extent to which models can extract regularities from linguistic input and in modeling the interactions of lexical, syntactic, and contextual information in parsing (Cottrell, 1989; Waltz & Pollack, 1985). But there are features of linguistic behavior which remain difficult for connectionists to simulate, and these are usually the features which are the easiest for symbolic approaches. A central concern is compositionality and the representation of part–whole hierarchies. Consider an example discussed by Touretzky (1989), himself an active connectionist researcher. If a system is trained to understand English NPs containing prepositional phrases by being exposed to phrases such as *the dog on the hood* and *the dog in the car*, how will that system then detect the anomaly in *the dog on the hood in the car*? This would seem to require a relatively sophisticated semantics, one which can, for example, build tentative representations of DOG ON HOOD and DOG IN CAR and then recognize that they are incompatible. Such intermediate representations are standard fare in symbolic approaches to parsing, but they do not seem to be implementable within existing connectionist systems. Touretzky (1989) suggests that connectionists need to build more initial structure into their models, making them in one sense more similar to traditional models while retaining their basic processing characteristics. The main point to be made here is that while connectionism does not yet have all of the answers, the range of possible architectures is only beginning to be explored.

CONNECTIONISM AND UNIVERSALS

It should be clear from what has been said that Universal Grammar (UG), at least as it is usually conceived, is not compatible with the connectionist framework. The princi-

ples and parameters of the UG of Government and Binding (GB) Theory, for example, are stated in terms of variables. For example, the principle that relates "θ-marking" to subcategorization states that "if α subcategorizes the position occupied by β, then α θ-marks β" (Sells, 1985). There is no way that variables such as the α and β in this rule can be "wired in" to the network at the outset.

At the same time, connectionism, with its focus on learning, tends to offer a strongly empiricist view of cognition, and the provision for innate mechanisms specific to particular domains, though not ruled out, is something which most connectionists would want to try only as a last resort. It appears that connectionist models may be able to shed new light on the nativism issue because of their ability to outperform previously conceived empirical systems (Walker, 1989). Two models relevant to language acquisition in particular are those of Hanson and Kegl (1987) and Elman (1988). Hanson and Kegl presented a network with a large corpus of English sentences in which syntactic categories replaced words. On this basis alone, the network was able to demonstrate a knowledge of a great deal of English grammar. The researchers used an auto-associative paradigm. If, following training, their network successfully reproduced an input pattern on its output units, this was counted as a positive grammaticality judgment. Sentences viewed as ungrammatical were often corrected by the network in its output. Most impressively, the network treated center-embedded structures as grammatical even though the only embedded sentences it had seen were right-embedded. Elman trained a network to perform a simple pronoun reference task, using sentences that had been thought to require the notion of c-command (Reinhart, 1983). Elman's model made no use of c-command or any other complex symbolic notion; indeed, there is probably no way his network could have been designed to incorporate such notions. While it is still possible that some form of innate linguistic "knowledge" will be required, as Hanson and Kegl (1987) argue, connectionist models should give us a better idea of what this is since we will know more about limits on the abilities of networks to extract regularities from input.

No one has proposed a set of connectionist universals for language, but we can consider the range of possibilities that are open. Two areas in which candidate universals might be considered are (a) the relative modifiability of particular connections, and (b) the architecture of the network. These are aspects of a system that one would want to assume the system is "born" with.

One view (Rumelhart & McClelland, 1986b) is that connections might have varying degrees of plasticity (though this is not a feature of most existing models). Some connections might start with fixed weights and others with weights that are modifiable to different degrees. The modifiability of connections might also decrease over time as one way of modeling the increased difficulty which adults experience with language learning. One can also imagine connections which wait to have their weights set on the basis of a relatively small set of inputs and then quickly become rigid. This might be one way to implement a sort of "parameter setting", though of a very different type than that envisioned by practitioners of GB because of the inability of the system to make direct reference to complex syntactic constructs.

Within the range of possible network architectures are those with modular subnetworks dedicated to particular functions. The modularity would derive from the sparse interconnections among the subnetworks and possibly from different learning

or activation rules for the units in the subnetworks. For example, there are several current connectionist approaches to handling temporal patterns (e.g., Elman, 1988; Jordan, 1986; Williams & Zipser, 1989). These require particular types of network connectivity and particular learning rules, which would need to be characteristic of the portion of the network concerned with linguistic form (and musical patterns) but not necessarily characteristic of the portion which is dedicated to, say, color recognition.

While provision for some modularity in connectionist models seems to be on the rise, it is unlikely that connectionists will accept the extreme position of some generativists (e.g., Fodor, 1983). One of the strong points of connectionism is its ability to model decisions made on the basis of the interaction of a variety of types of information. Thus connectionists tend to be interactionalists rather than modularists. This goes as well for the (non-)distinction between linguistic and non-linguistic cognitive behavior, and it is also in agreement with the views of cognitive linguists (Langacker, 1987).

While connectionism is not consistent with most generative notions of universals, it is not incompatible with various proposals for processing universals. The best-known of these are probably Slobin's "operating principles" (Slobin, 1973). Among these principles are a number that make reference to the sequencing of items, and recently connectionist models have been shown to provide powerful means of modeling the learning and processing of sequential patterns. One approach sets aside a group of units to serve as a kind of short-term memory (Anderson, Merrill, & Port, 1989; Elman, 1988; Jordan, 1986). Because this memory has a decay built into it, more recent items are remembered better than those which appeared further back. Thus, such a system is more likely to make use of sequential pieces that are not interrupted than of those that are. This is precisely the content of Slobin's Principle 4: Avoid interruption or rearrangement of linguistic units.

A CONNECTIONIST FRAMEWORK FOR SLA RESEARCH

There is as yet no generally agreed on "connectionist linguistics" (but see Lakoff, 1989, for a first cut). What I will propose in this section is a framework which is consistent with much connectionist thinking and also with the basic tenets of cognitive linguistics, the approach which is the most compatible with connectionism. The key idea is one of language processing as pattern completion, where a pattern includes features of all of the types which a learner can generalize over; that is, both features of linguistic form and of the context of the utterance. Pattern completion is implemented in auto-associative networks.

First Language Processing and Acquisition

The framework starts from the following basic tenets:

1. Knowledge of language consists of generalizations made over *linguistic pattern complexes* (LPCs), each consisting of features of form (morphosyntactic, phonological) and content

(semantic, pragmatic, contextual). In the connectionist implementation, LPCs appear as patterns of activation over a set of input/output units.

2. Associations between the form and content features that make up LPCs are mediated by a complex structured layer of hidden units which comprises the lexicon/grammar of the system. Patterns of activation over these units correspond to lexical entries as well as syntactic structures. Representations are distributed; that is, it is not possible to isolate a unit or set of units which reliably represent notions such as CLAUSE, SUBJECT, INITIAL-CONSONANT-CLUSTER, and MEANING-OF-THE-WORD-*TABLE*.
3. Language acquisition is an auto-associative process. The system is presented with partial or complete LPCs, and on this basis associations are built up between the microfeatures of LPCs (via the hidden units).
4. Language processing consists in the completion of partial LPCs. In comprehension, the system starts with most of the formal features of an input pattern and, because of context, usually some of the content features as well, and the task is to fill in the missing content. In production, the system starts with a goal in the form of a set of content features, and the task is to fill in the features specifying the form.

Of course, a number of questions need to be answered about the adequacy of this model. One that will strike some readers of this article is the need to demonstrate that the system can cope with the "poverty of the stimulus" problem, one of the major arguments for the generative approach. That is, it must be shown that the network, without the help of symbolic predispositions, can produce structures which it has never been exposed to and, at the same time, recognize as anomalous other structures which it has not been exposed to. However, as argued by Walker (1989), the question of the adequacy for language acquisition of systems constrained only by general cognitive mechanisms is an empirical question, not one that can be decided by the armchair theorizing that has been thought to suffice.

Second Language Processing and Acquisition

There are three ways in which second language acquisition may differ from first language acquisition. The latter two are relevant in particular for adult learners.

1. L1 patterns may transfer to L2 (and vice versa).
2. Neurophysiological changes or cognitive developments not related specifically to language may limit the learner's ability to acquire language or may predispose the learner to particular acquisition strategies.
3. Contextual factors, such as the acquisition setting or the communicative demands placed on the learner, may affect acquisition.

Transfer is precisely what connectionist models are good at. Once a network has learned an association of a pattern P1 with a pattern P2, when it is presented with a new pattern P3, this will tend to activate a pattern that is similar to P2 just to the extent that P3 is similar to P1. Thus, the connectionist framework provides an excellent means of testing various notions about the operation of transfer in SLA. What claims would be made about L1-to-L2 (or L2-to-L1) transfer within this framework? I

will assume first of all that the primitives of form and content are the same across languages; that is, that the basic network units over which input form and content are represented are the same for L1 and L2. The claim then would be that overlap of any type between L1 and L2 should be the basis of transfer. However, the details of how transfer would actually operate would need to be tested through simulation for different sets of circumstances. That is, other than these very general points, no specific claims are made regarding transfer on the basis of features of connectionist models. In the next section I describe a simple model which has been implemented to test some of the transfer possibilities.

It is less clear at this stage how connectionist models would handle the second and third aspects of second language acquisition. As noted, it is possible to model changes in neuronal plasticity with networks, but no one has yet suggested how Piagetian stages would emerge from connectionist processing or how a monitoring facility would be implemented. These belong to the realm of "higher-level" cognitive processes, which connectionists in general want to see emerge from lower-level properties but which are only beginning to be investigated within this framework. Factors involved in interaction between language users are even more remote from current models, since they would seem to require a relatively complete characterization of the separate cognitive systems.

Example Simulations

A set of simple simulations was run to investigate the efficacy of using networks to study transfer in SLA. The network for the simulations is an auto-associator in which input patterns are mapped to identical output patterns via a layer of hidden units.[3] Each input pattern is intended to represent a simple clause consisting of two words, a subject and a verb. The input and output layers consist of three groups, a pair of language units, representing the language being learned or processed; a set of form units, representing the words and their positions in the clause; and a set of content units, representing the word meanings and their roles in the proposition denoted by the clause. The roles for this example are simply AGENT and PROCESS; that is, in the sentence *Mary sleeps*, the concept MARY (the meaning of the word *Mary*) plays the AGENT role, and the concept SLEEP (the meaning of the word *sleep*) plays the PROCESS role. Following the terminology standard in AI models, I will refer to MARY as the *filler* of the role AGENT. Figure 1 shows the basic structure of the network. Small circles represent units, and heavy arrows signify complete connectivity between groups of units. That is, every input unit is connected to every hidden unit, and every hidden unit to every output unit.

Each input sentence consisted of two words, a subject and a verb, and each input pattern represented a sentence, its meaning, and its language. To create the input patterns, an arbitrary binary vector[4] of length 7 was assigned to each word and word meaning. For example, the word *John* might be assigned the vector [0110100] and the meaning JOHN the vector [1100001]. In the input layer of the network, 14 units represented the two words in the sentence, 7 for the word in initial position and 7 for the word in second (final) position. For the pattern representing the sentence *John*

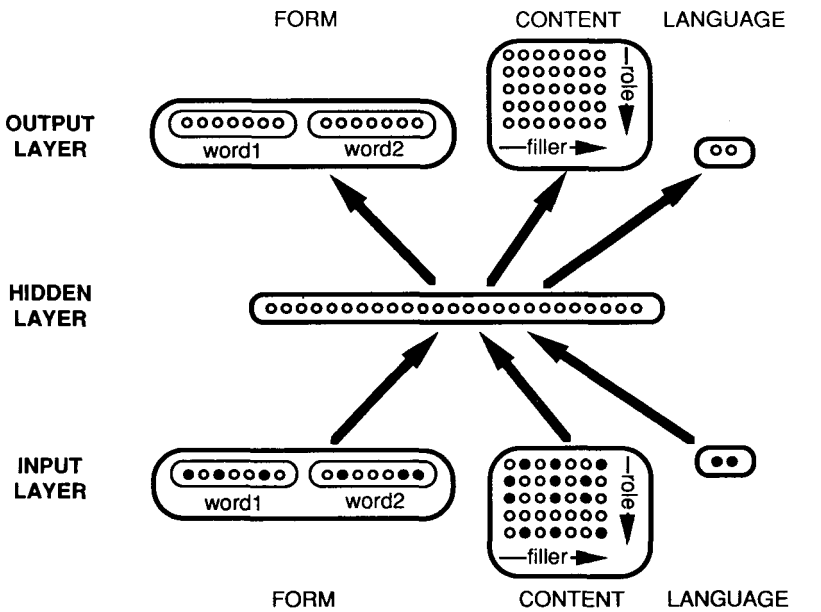

Figure 1. Architecture of network used in simulations.

sings, the units corresponding to the vector assigned to *John* are turned on or off in the 7 first-position units, and those corresponding to the vector assigned to *sings* are turned on or off in the 7 second-position units. For example, if the vector for *John* is [0110100], then the second, third, and fifth units among the 7 first-position units are turned on for a sentence beginning with the word *John*.[5]

The coding for the content units is somewhat different. Roles (only AGENT and PROCESS for this simulation) are assigned binary vectors of length 5, and there are 35 (7×5) content units, one for each *conjunction* of a role element and a filler element. Using this approach, it is possible to input more than one role-filler pair at the same time (Dolan & Smolensky, 1989). Thus, the pattern across the 35 content units can represent both the filler of the current AGENT and the filler of the current PROCESS. In addition there are two units for the language, either L1 or L2. L1 is assigned the vector [00] and L2 the vector [11].

Thus, a complete input pattern for the clause *John sings* as an L2 sentence consists of a pattern representing the word *John* on the first-position form units, a pattern representing the word *sings* on the second-position form units, a pattern representing JOHN as AGENT and SING as PROCESS on the content units, and a pattern representing L2 on the language units. This complete input pattern is shown on the input units in Figure 1, with the following assignment of vectors to tokens: *John*: [1010010], *sings*: [0100011], JOHN: [0101001], SING: [1001010], AGENT: [10001], PROCESS: [01100], and L2: [11]. Filled circles represent units that are on (activation 1) and empty circles those that are off (activation 0).

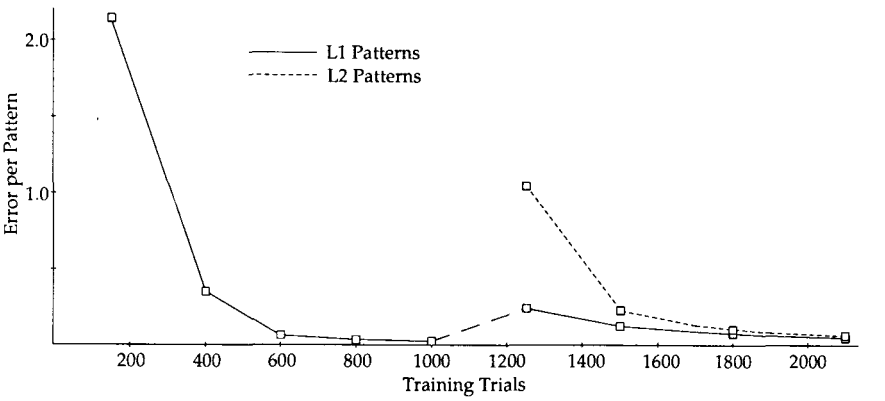

Figure 2. Sum-of-square errors for L1 and L2 patterns.

The system was trained using back-propagation (Rumelhart, Hinton, & Williams, 1986) to associate input patterns of this type with identical output patterns. For the simulations described here, 25 hidden units[6] separated the 51 (14+35+2) input and 51 output units. There was complete connectivity between the layers; thus, there were 2,550 ($2\times51\times25$) adjustable connections in all. A small set of words and meanings was used for the training patterns: 6 verbs and verb meanings, and 6 nouns and noun meanings. The training set consisted of randomly selected pairings of noun-verb, AGENT-PROCESS, except that a small set of combinations was never trained on. After training on 1,100 such patterns (many repeated, of course) the network had learned to map input patterns to themselves with a very small error rate. The error measure used here is the mean sum-of-squares error per pattern, that is, the sum of the squares of the errors made on each output unit. Data for this run are shown on the left side of Figure 2. The errors are the means over 200 (or in the case of the initial datum, 300) training trials. Thus, the point shown at 1,000 on the abscissa in the figure is the mean for the 900th to 1,099th trials.

Following training, the network was able to successfully complete partial input patterns. Input patterns in which the words were missing (with form units all set to 0.25, the mean activation value for all word patterns), corresponding to a production task, resulted in output patterns with the appropriate activation on the form units (mean sum-of-squares error per pattern about 0.27). Input patterns in which the content was missing, corresponding to a comprehension task, resulted in output patterns with appropriate activation on the content units (mean sum-of-squares error per pattern about 1.1). Results were only slightly worse for patterns which the network was not trained on. For example, though the network never saw the pattern for the sentence *John drinks*, it was able to correctly turn on the output units for the words *John* and *drinks* in the first and second position groups respectively when presented with an input pattern giving only the fillers JOHN and DRINK together with their roles.

To test transfer to a second language, a second set of input words was then generated. In the first 2 simulations, these bore no relation to the corresponding first language words. In the third and fourth simulations, each differed from the cor-

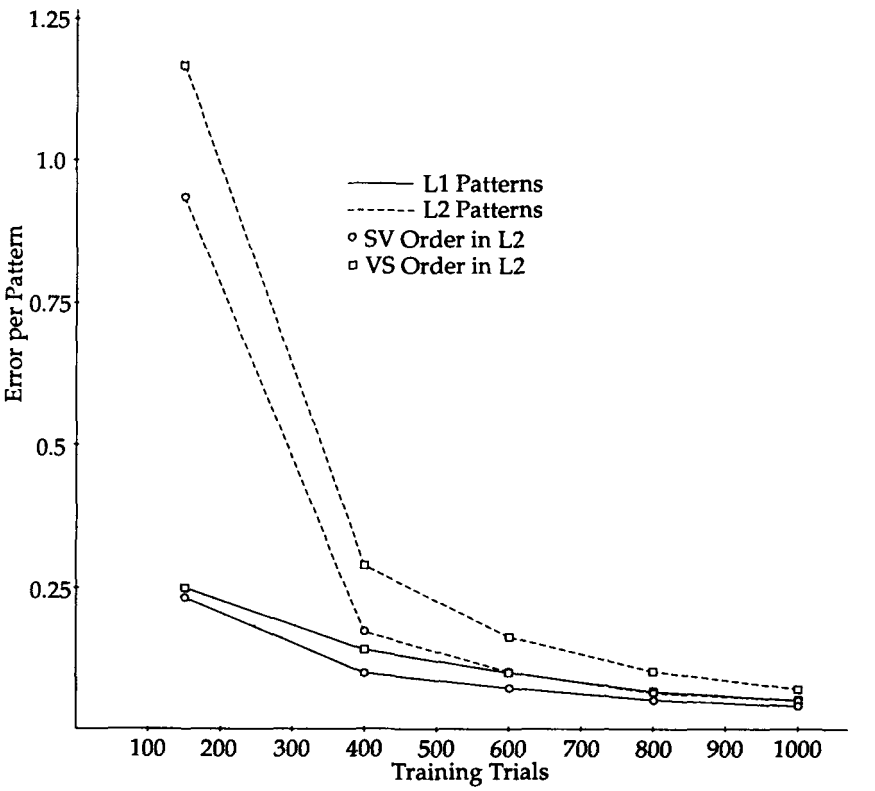

Figure 3. Sum-of-square errors for L1 and L2 patterns by L2 word order.

responding L1 word by only one bit. For example, in the latter simulations, the L1 vector for *sings* was [0101100] and the L2 vector for *sings* was [0101101]. After the network had been trained on the L1 patterns as described above, it was trained on both L1 and L2 patterns for a total of 2,200 more repetitions. In addition to the difference in lexical items, the L2 in some of the simulations differed in constituent order; that is, for these simulations the L2 had verb first and subject second. Of interest was the speed with which the system was able to learn the L2 patterns. The right side of Figure 2 shows pattern errors averaged over all four simulations. There are several things to notice here. First, the L1 patterns clearly suffer interference from the L2 patterns. Even after 1,100 additional training iterations, they do not recover their previous accuracy in any of the simulations. Not surprisingly, however, the L2 patterns remain less well known throughout. Second, though the L2 patterns are initially difficult for the network, they are not as difficult as the L1 patterns were when they were first presented to the network. This is true even for the simulation in which the L2 patterns differ most from the L1 patterns (see Figure 5).

Figures 3, 4, and 5 present detailed results for the portion of the simulations following the initial acquisition of L1 patterns alone. Figure 3 shows the effect of

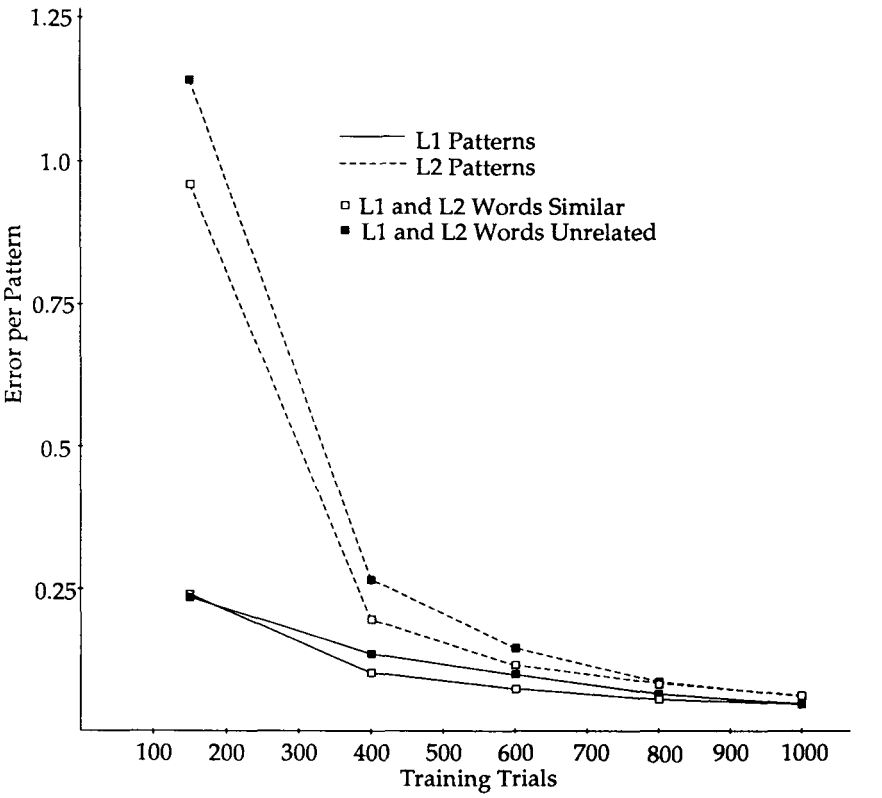

Figure 4. Sum-of-square errors for L1 and L2 patterns by word similarity.

word-order differences. It is more difficult for the network to learn the L2 patterns when the word order differs from the L1 patterns than when it is the same. This difficulty is reflected not only in the speed with which the L2 patterns are learned but also in the degree to which the L1 patterns are interfered with. The word order difference also seems to have less and less of an effect on mastery of the L2 (and interference with the L1) as learning continues.

Figure 4 shows the effect of word similarity. There is some evidence of an advantage when the L2 words are similar to the corresponding L1 words, but again this difference seems to disappear as more patterns are presented. Figure 5 shows the data for only the L2 patterns for all four simulations. This brings out what appears to be an interesting interaction between the two independent variables: the effect of different word order is greater with similar words than it is with unrelated words.

To what extent do these results agree with the facts of transfer? It is well known that the learning of a second language affects the knowledge of the first (Sharwood Smith, 1983), though this has not been the focus of much research. Note that the network is exposed to as many different L2 patterns as it has already learned in the

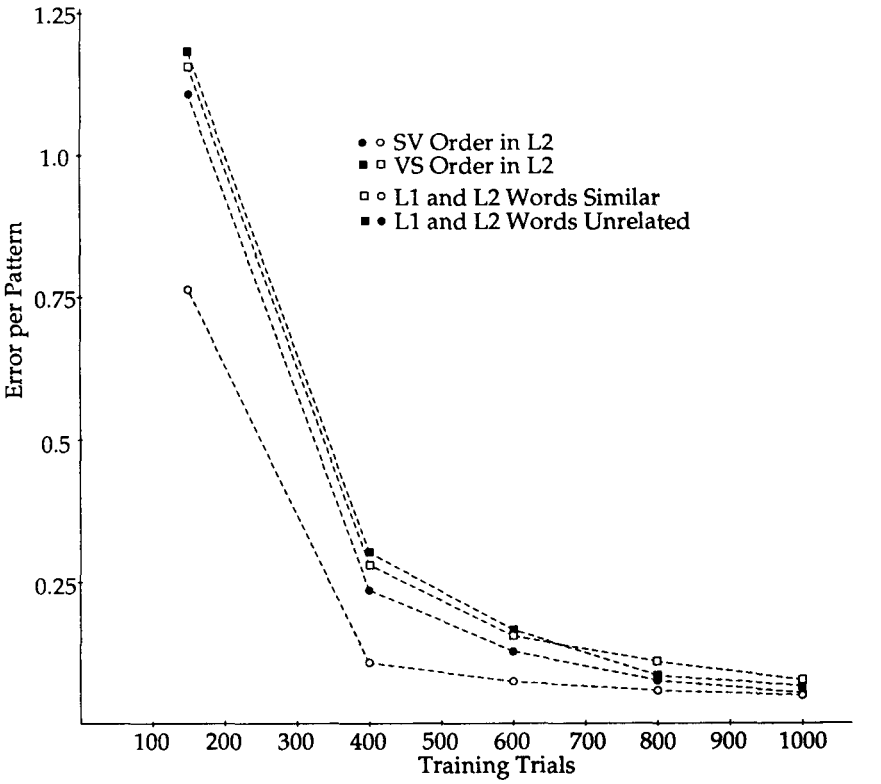

Figure 5. Sum-of-square errors for L2 patterns.

L1, a feature which is not characteristic of most second learning situations, and this exaggerates the effect of the L2 on the L1. With respect to word order and lexical form, a traditional contrastive analysis position would hypothesize that both types of differences would make the L2 harder to learn. More recently, it has been shown that the details of transfer are more complex. Rutherford (1983) argues that basic constituent order does not seem to transfer. However, he does not consider beginning learners, who would correspond to this network at the point where word order does seem to enter in. Flege (1987) found that, at the level of phonetics, similar patterns may be more difficult to learn than those that are completely unrelated to L1 patterns. For the network, by contrast, lexical forms which are more different are (somewhat) more difficult, at least in the early stages.[7] Clearly the issue of the relative ease of learning similar forms at phonetic, phonological, lexical, and syntactic levels is a crucial one, and one that connectionist networks are well suited for investigating.

Finally, there is work which shows that the degree of transfer from L1 to L2 depends on the extent to which the two languages are perceived as related (Kellerman, 1978). This seems to agree with one result here: the word order difference is

more significant when there is a relationship between the words in the two languages.

The task presented to this network is obviously a gross oversimplification of the second language learner's task. The network has no real semantics and no sense of time whatsoever. Both semantics (e.g., Harris, 1989; Lakoff, 1989) and temporal processing (e.g., Elman, 1988) are active areas of current connectionist research, and there are several ways in which the present model could be augmented to handle these aspects in a more plausible manner.

While these results should be regarded very tentatively, they point to a possible line of connectionist SLA research, one in which networks test our particular hypotheses about transfer and suggest what types of data are needed to flesh out the transfer picture. The main conclusion to be drawn from these simulations is that, even with this extremely simple model of the transfer process, it was impossible to predict precisely how the network would behave. Thus, simulations have an important role to play.

CONCLUSIONS

Thirty-three years ago Chomsky forced linguists to take seriously the limitations on the sort of input that learners have access to. Real language in use is a messy business, a fact that all applied linguists are certainly aware of. Chomsky argued that the learner must somehow take this limited input and construct a clean grammar of the language being learned, one that characterizes the competence of adult native speakers. The picture that the adult ends up with, he claimed, is one in which redundancy is minimized at all costs and in which neat lines are drawn between concepts, between components of language, and between language and everything else. It was a logical next step to posit a set of innate constraints which made the formidable task of the language learner possible.

Connectionism now offers a radical alternative to this view. What if the adult "grammar" is not a neat one after all? What if the best characterization of adult "performance" is one quite unlike the idealized picture that generative theory would have us believe in? Once we are willing to accept the possibility of an adult system in which redundancy is rampant, concepts are fluid, metaphor is a fundamental process, and exceptions are the rule, our picture of the learner and our research strategy change dramatically. Rather than focusing on innate constraints, our work seeks powerful ways of extracting regularities from the input. Using these techniques, learners are free to examine the input and decide for themselves whether and where lines are to be drawn.

Where does this leave SLA research? In a recent article, Frederick Newmeyer (1987) has comforting news for the field: generative linguistics, which once appeared to be in disarray, is converging on a number of points which can now guide future applied research. Yet, if we look beyond the narrow confines of generative linguistics, we see that this convergence is an illusion. The old questions about innateness, the mind and the body, and what it means to know are being asked again, and a new set of answers is being proposed. If radical connectionists are right, a great deal of

rethinking will be needed in SLA theory, as elsewhere in cognitive science. This may not be very comforting news, but there is compensation in the possibility that the end product may be a more elegant model of acquisition, one which allows it to be integrated into the rest of the mind, and perhaps even the brain.

NOTES

1. This is particularly true for connectionists, who, unlike many other cognitive scientists, do not view the mind as analogous to a digital computer (though connectionist models, like any other computational models, can be simulated on digital computers).

2. Throughout this article I will use small capital letters to denote conceptual entities, to be distinguished from words, which will appear in italics. Thus, BLUE is intended to be the meaning of *blue*.

3. This is not the only way to implement auto-association. Another possibility makes use of a single layer of units representing both inputs and outputs (Hinton & Sejnowski, 1986). However, learning is generally not as efficient in such models.

4. That is, an ordered list, each of whose elements is either 0 or 1.

5. A more realistic approach to sequencing, in which words appear within the same group of units at different times, is also possible (Elman, 1988).

6. There is as yet no simple way to determine how many hidden units a particular system needs. Thus, no particular significance should be attached to the number 25.

7. It is possible that this is a "floor effect," that is, that the network has reached a point at which further improvement is either impossible or very gradual, resulting in a minimization of differences that were significant at earlier stages.

REFERENCES

Anderson, S., Merrill, J., & Port, R. (1989). Dynamic speech categorization with recurrent networks. In D. Touretzky, G. Hinton, & T. Sejnowski (Eds.), *Proceedings of the 1988 Connectionist Models Summer School* (pp, 398–406). San Mateo, CA: Morgan Kauffman.

Berko, J. (1958). The child's learning of English morphology. *Word, 14*, 150–177.

Cottrell, G. (1989). *A connectionist approach to word sense disambiguation*. Los Altos, CA: Morgan Kaufmann.

Dolan, C. P., & Dyer, M. G. (1989). Parallel retrieval and application of conceptual knowledge. In D. Touretzky, G. Hinton, & T. Sejnowski (Eds.), *Proceedings of the 1988 Connectionist Models Summer School* (pp. 273–280). San Mateo, CA: Morgan Kaufmann.

Dolan, C. P., & Smolensky, P. (1989). Implementing a connectionist production system using tensor products. In D. Touretzky, G. Hinton, & T. Sejnowski (Eds.), *Proceedings of the 1988 Connectionist Models Summer School* (pp. 265–272). San Mateo, CA: Morgan Kaufmann.

Elman, J. L. (1988). *Finding structure in time* (Technical Report 8801). La Jolla: University of California, San Diego, Center for Research in Language.

Fauconnier, G. (1985). *Mental spaces*. Cambridge, MA: MIT Press.

Feldman, J. A., & Ballard, D. H. (1982). Connectionist models and their properties. *Cognitive Science, 6*, 205–254.

Fillmore, C. J. (1988). The mechanics of "Construction Grammar." *Proceedings of the 14th annual meeting of the Berkeley Linguistic Society* (pp. 35–55). Berkeley, CA: Berkeley Linguistic Society.

Flege, J. E. (1987). Effects of equivalence classification on the production of foreign language speech sounds. In A. James & J. Leather (Eds.), *Sound patterns in second language acquisition* (pp. 9–39). Dordrecht: Foris.

Fodor, J. A. (1983). *The modularity of mind*. Cambridge, MA: MIT Press.

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

Gasser, M. (1988). *A connectionist model of sentence generation in a first and second language* (Technical Report UCLA-AI-88-13). Los Angeles: University of California, Los Angeles, Computer Science Department.

Hanson, S. J., & Kegl, J. (1987). PARSNIP: A connectionist network that learns natural language grammar from exposure to natural language sentences. *Proceedings of the ninth annual conference of the Cognitive Science Society* (pp. 106–119). Hillsdale, NJ: Erlbaum.

Harris, C. L. (1989). *Connectionist explorations in cognitive linguistics*. Unpublished manuscript, Program in Cognitive Science, University of California, San Diego.

Hinton, G. E., & Sejnowski, T. J. (1986). Learning and relearning in Boltzmann machines. In D. E. Rumelhart, J. L. McClelland, & PDP Research Group (Eds.), *Parallel Distributed Processing. Explorations in the microstructures of cognition: Vol. 1. Foundations* (pp. 282–317). Cambridge, MA: MIT Press.

Hofstadter, D. R. (1985). Variations on a theme as the crux of creativity. In D. R. Hofstadter, *Metamagical themas* (pp. 232–259). New York: Basic Books.

Jordan, M. I. (1986). Attractor dynamics and parallelism in a connectionist sequential machine. *Proceedings of the Eighth Annual Conference of the Cognitive Science Society* (pp. 531–546). Hillsdale, NJ: Erlbaum.

Kanerva, P. (1989). *Sparse distributed memory*. Cambridge, MA: MIT Press.

Kellerman, E. (1978). Giving learners a break: Native speaker intuitions as a source of predictions about transferability. *Working Papers on Bilingualism, 15*, 59–92.

Lakoff, G. (1987). *Women, fire, and dangerous things: What categories reveal about the mind*. Chicago: University of Chicago Press.

Lakoff, G. (1989). A suggestion for a linguistics with connectionist foundations. In D. Touretzky, G. Hinton, & T. Sejnowski (Eds.), *Proceedings of the 1988 Connectionist Models Summer School* (pp. 301–314). San Mateo, CA: Morgan Kaufmann.

Langacker, R. W. (1987). *Foundations of cognitive grammar* (Volume 1). Stanford, CA: Stanford University Press.

McClelland, J. L., Rumelhart, D. E., & Hinton, G. E. (1986). The appeal of parallel distributed processing. In D. E., Rumelhart, J. L. McClelland, & the PDP Research Group (Eds.), *Parallel distributed processing. Explorations in the microstructures of cognition: Vol. 1. Foundations* (pp. 3–44). Cambridge, MA: MIT Press.

McClelland, J. L., Rumelhart, D. E., & PDP Research Group (Eds.). (1986). *Parallel distributed processing. Explorations in the microstructures of cognition: Vol. 2. Psychological and biological models*. Cambridge, MA: MIT Press.

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics, 5*, 115–133.

Newell, A. (1980). Physical symbol systems. *Cognitive Science, 4*, 135–183.

Newmeyer, F. J. (1987). The current convergence in linguistic theory: Some implications for second language acquisition research. *Second Language Research, 3*, 1–19.

Pinker, S., & Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition, 28*, 73–193.

Plunkett, K., & Marchman, V. (1989). *Pattern association in a back propagation network: Implications for child language acquisition* (Technical Report 8902). La Jolla: University of California, San Diego, Center for Research in Language.

Reinhart, T. (1983). *Anaphora and semantic interpretation*. Chicago: University of Chicago Press.

Rosch, E. (1978). Principles of categorization. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and categorization* (pp. 27–48). Hillsdale, NJ: Erlbaum.

Rosenblatt, F. (1962). *Principles of neurodynamics*. New York: Spartan.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart, J. L. McClelland, & the PDP Research Group (Eds.), *Parallel distributed processing. Explorations in the microstructures of cognition: Vol. 1. Foundations* (pp. 319–362). Cambridge, MA: MIT Press.

Rumelhart, D. E., & McClelland, J. L. (1986a). On learning the past tenses of English verbs. In J. L. McClelland, D. E. Rumelhart, & the PDP Research Group (Eds.), *Parallel distributed processing. Explorations in the microstructures of cognition: Vol. 2. Psychological and biological models* (pp. 216–271). Cambridge, MA: MIT Press.

Rumelhart, D. E., & McClelland, J. L. (1986b). PDP models and general issues in cognitive science. In D. E. Rumelhart, J. L. McClelland, & the PDP Research Group (Eds.), *Parallel distributed processing. Explorations in the microstructures of cognition: Vol. 1. Foundations* (pp. 110–149). Cambridge, MA: MIT Press.

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

Rutherford, W. E. (1983). Language typology and language transfer. In S. M. Gass & L. Selinker (Eds.), *Language transfer in language learning* (pp. 358–370). Rowley, MA: Newbury House.

Schank, R. C., & Abelson, R. (1977). *Scripts, plans, goals and understanding*. Hillsdale, NJ: Erlbaum.

Sejnowski, T. J., & Rosenberg, C. R. (1987). Parallel networks that learn to pronounce English text. *Complex Systems, 1*, 145–168.

Sells, P. (1985). *Lectures on contemporary syntactic theories*. Stanford, CA: Center for the Study of Language and Information.

Sharwood Smith, M. A. (1983). On first language loss in the second language acquirer: Problems of transfer. In S. M. Gass & L. Selinker (Eds.), *Language transfer in language learning* (pp. 222–231). Rowley, MA: Newbury House.

Shea, P. M., & Lin, V. (1989). Detection of explosives in checked airline baggage using an artificial neural system. *Proceedings of the First International Joint Conference on Neural Networks, 2*, 31–34.

Slobin, D. I. (1973). Cognitive prerequisites for the development of grammar. In C. A. Ferguson & D. I. Slobin (Eds.), *Studies of child language development* (pp. 175–208). New York: Holt, Rinehart and Winston.

Touretzky, D. S. (1989). Connectionism and PP attachment. In D. Touretzky, G. Hinton, & T. Sejnowski (Eds.), *Proceedings of the 1988 Connectionist Models Summer School* (pp. 325–332). San Mateo, CA: Morgan Kaufmann.

Walker, V. (1989). Competent scientist meets the empiricist mind. *Center for Research in Language Newsletter, 3*, 5–17.

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

Williams, R. J., & Zipser, D. (1989). A learning algorithm for continually running fully recurrent neural networks. *Neural Computation, 1*, 270–280.

Winograd, T. (1983). *Language as a cognitive process: Vol. 1. Syntax*. Reading, MA: Addison-Wesley.