CHISELING COMPETENCE: A CONNECTIONIST REVISION OF CHOMSKY'S LANGUAGE ACQUISITION DEVICE

By

Michael Pittman Walker

Adviser: Dr. Mark W. Risjord

A thesis submitted to the Faculty of Emory College
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Arts with Honors

Department of Philosophy

1999

Contents

Introduction 3

I. Philosophical Interest in the Problem of Language Acquisition 5
The problem of language acquisition 9

II. Artificial Intelligence and Language Acquisition 18
Connectionism and neural networks 21
Characteristics of neural networks 24
Neural network language processing: Case studies 28

III. Consequences for the Language Acquisition Device 34
Conclusions on language acquisition in a connectionist system 35

Postscript 39
References 1

Introduction

Language comprehension is a fascinating and complicated trait. Traces of language use and understanding can be found in young children, yet other aspects of their development seem to be in a far more primitive state. The mysterious process of language acquisition has only recently been explained in any successful manner, and the details of developing a linguistic comprehension are vague, if even understood at all.

It is my purpose to explore the mysteries of language acquisition from a philosophical standpoint, and suggest a method by which one can better understand the details of the process. Specifically, I begin with the foundations of the philosophical argument, grounded in the rational-empirical knowledge debate. These sometimes opposing views have much to offer in understanding how the mind can acquire linguistic abilities in such a timely fashion. The poverty of linguistic stimulus seems to suggest that humans carry innate linguistic abilities. It is the poverty of the stimulus that frames the interests of the paper.

Noam Chomsky's device for language acquisition provides insight into how a mind might overcome the poverty problem, but it leaves much to be explained—namely, how a brain might specifically carry a general device for linguistic comprehension, without any explicit knowledge or evidence of the device. I propose that connectionist theory offers a clear insight into this problem. Connectionist architectures, currently an interest in artificial intelligence, offer a solution to the poverty of the stimulus. In my thesis, *I wish to show that the brain is essentially connectionist, and that the language acquisition device is implicitly included in the connectionist structure of the nervous*

system. By viewing the mind as a connectionist device, we can gain a clearer understanding of (1) how the poverty of the stimulus is overcome, and (2) how our knowledge of language emerges from the combination of linguistic data and the innately connectionist structure of the brain.

I. Philosophical Interest in the Problem of Language Acquisition

The possibility of innate linguistic universals is really part of a larger debate conducted between philosophers since ancient times. Plato's Meno argued that one might have an innate knowledge of geometrical truths without being aware of such knowledge. But the presence of innate knowledge is difficult to demonstrate, as the mere possibility of innate truths does not guarantee their existence. Empiricists like Locke, Berkeley, and Hume argue that experience alone leads to knowledge of truth, whereas the rationalist tradition, defended by Plato, Descartes, Leibniz and Kant, among others, defend the principle of innate ideas (Stich 145).

An examination of Locke's An Essay Concerning Human Understanding in comparison with Leibniz's New Essays gives a good philosophical picture of the ensuing debate. The two positions held by the philosophers reveal a split in the philosophy of mind. Generally speaking, Locke's empirical approach to human understanding yields a system of knowledge based wholly on experience. Leibniz holds fast to the rationalist belief that a human can know certain things innately. For Leibniz, the senses only aid in revealing these preexisting truths.

Leibniz draws a clear line between two types of truths: eternal, or necessary truths, and contingent truths. The former accounts for innate knowledge, whereas the latter provides freedom from necessity, and, for us, freedom of the will. A truth is eternal if it is true through the principle of contradiction; namely, that anything requiring a contradiction (p & -p) must be false, and likewise, that anything that would produce a contradiction if it were not the case must be true (Leibniz, Philosophical Essays 217).

Eternal truths are necessary because they are logically impossible to be false. This validity provides a truth that is built into the structure of reason itself. Thus, the rational brain must hold in it these eternal truths by way of its inherently rational nature, and the truths can be considered innate.

The mind applies the principle of contradiction to reveal eternal truths about the world. The fact that all the interior angles of a triangle sum to 180 degrees is undeniable by the principle of contradiction, and we consider it to be a necessary truth. "An apple is not fire" (Stich 60) is impossible to disprove, and must be true due to our innate knowledge. Leibniz compares these truths to

> a block of marble which has veins, rather than a block of marble completely uniform, or of blank tablets—that is to say, of what is called among philosophers a *tabula rasa*. For if the soul were like these blank tablets, truths would be in us as the figure of Hercules is in a piece of marble when the marble is completely indifferent as to whether it receives this figure or some other. But if there were veins in the block which indicated the figure of Hercules rather than other figures, this block would be more determined thereto, and Hercules would be, as it were, in a fashion innate in it, although it would be necessary to labor to discover these veins, to clear them by polishing and by cutting away what prevents them from appearing. Thus it is that ideas and truths are innate in us, as inclinations…and not as actions. (Leibniz New Essays, 52)

This poetic passage from the preface of the New Essays explains how the principle of contradiction is innately embedded in truths we seem to discover through experience. Even before we encounter apples and fires, the idea that an apple is not fire is innately true to us. We use our senses to reveal the differences between the two, and the principle of contradiction ensures that the two can never be equal. If this truth is the figure of Hercules, then the veins mark the presence of the principle of contradiction, and the

senses only reveal what is innately guaranteed to be true by chipping away confusion between the innate truth and any falsities.

Locke however, is hesitant to believe that one could possess knowledge of things in the mind without being consciously aware of the knowledge. The infinity of necessary truths derived from the principle of contradiction seem to require some kind of reasoning to be discovered, and might not be innately given to the mind itself. Instead, Locke claims that the mind has an unlimited capacity for knowledge, but it must acquire the knowledge on its own, through experience and reflection upon that experience.

Knowledge of general and abstract truths, of the type Leibniz might consider innate, is, in Locke's opinion, only attainable through experience. The only existing things in the world are *particulars*, such that all encounters with worldly things are particular existing entities, and that no general idea or abstract concept can exist outside of a mind (Locke 16). Through the senses, particular thoughts "furnish the yet empty cabinet" (Stich 52) of the mind with sense-data. Over time, these particular thoughts are labeled with names, and, in degrees, are categorized and associated as they become more familiar. The mind separates them "from the circumstances of time and place, and any other ideas that may determine them to this or that particular existence" (Locke 17). Finally, an abstracted idea of a collection of particular ideas is formed, known to Locke to be a *general idea*. These general ideas are the materials by which the mind can use language, for instance, and create new sentences (Stich 52).

Locke's empirical theory of knowledge acquisition still allows for the possibility of unlimited knowledge, but simply does not insist that the knowledge needs to be innate. All concepts are general ideas, formed by the abstraction of particular ideas. A collection

of these particular ideas is sufficient information to yield an abstracted idea. Because the mind has the capability to hold any type of particular idea (as they are simple to grasp), the mind also has the capability to acquire any part of the infinite number of general ideas that can be abstracted from these particular ideas. Thus, a mind can formulate any logically possible concept, particular or general, given the appropriate experience (Stich 146).

For Locke, the mind can gain ideas in only two ways. Experience with the world lends the mind empirical data that gives us the *sensation* of things. Sensation, then, is the only way that we know the world, and is the method by which all particular ideas are gathered. The second method, *reflection*, is "the perception of our own mind as it operates in us" (Stich 64), from which we derive the activities of the mind. Reflection allows the mind to associate particular ideas, and make abstractions that form all general ideas. Locke concludes that the necessity of reflection for the formation of general ideas is "the reason why it is pretty late before most children get ideas of the operations of their own minds; and some have not any very clear or perfect ideas of the greatest part of them all their lives" (Stich 66).

Locke's observation reveals one apparent contradiction in the empiricist hypothesis of knowledge acquisition: if general concepts must be learned through reflection, how could a mind come to have so many in such a short period of time? How, for instance, might a child come to possess an understanding of language so great that it can use and understand the language fluently within a few years of its life? Actually, this particular question fuels the debate between linguists defending empirical and rational theories of language acquisition. The problem, known as the *poverty of the stimulus*, is a

philosophical one as much as it is a problem of linguistic theory, and deserves further examination.

The problem of language acquisition

Let the discussion of opposing rationalist and empiricist views of knowledge be narrowed to the problem of language acquisition. Noam Chomsky's 1965 book, Aspects of the Theory of Syntax, can be considered a rationalist's approach to language acquisition, and proposes a theory of language acquisition that has been the subject of debate since the work was published. Chomsky's general approach accounts for the poverty of the stimulus by proposing that any language user must have an innate capability to learn language. This ability allows a speaker to form a correct and complete system of rules for generating sentences with the aid of the few examples the speaker encounters within the first few years of life. Just as Leibniz's block of marble held the veins that suggest the features of the Hercules figure, Chomsky's innate device suggests a particular understanding of language by which a child might become a fluent speaker of the language.

First, one must establish exactly what it means to be a fluent and capable language user, especially since Chomsky has a slightly unconventional approach to the subject. Chomsky defines the speaker-hearer's tacit knowledge of his language as *competence* (Chomsky 4). This knowledge of language allows the speaker to produce and understand utterances in that language. Thus, competence in the language is required before any speaker can successfully have the ability to use the language (or, have the ability to *perform* in the language). Chomsky's approach is mentalistic in that he is "concerned

with discovering a mental reality underlying actual behavior" (Chomsky 4). The internal language, or "I-language", is what Chomsky considers important, because the performance is merely the (often ill-formed) product of the I-language competence.

Linguistic competence can be sufficiently described by a *grammar*. An adequate grammar, operating with a given vocabulary, should generate all and only the sentences of the described language (Lyons 18), where a language is a set of sentences, and sentences are made up of lexical primes. Because the grammar generates sentences in the language explicitly, without the further aid of an intelligent operator to guide its generation of sentences, such a competence is described by a *generative grammar* (Chomsky 4).

Most importantly, Chomsky believes that a linguistic competence—the underlying system of rules—is "chosen" by some innate device. An internal system of rules is required before any performative skills can develop in the language user, such that a competence "provides the basis for actual use of language" (Chomsky 9). Otherwise, Chomsky argues, the user would have no means of understanding or using the language, as there would exist no rules to guide the child (Chomsky 8, 9). However, the amount of language data does not provide enough stimuli on its own to account for the complexity and completeness of the competence achieved by the fluent speaker. Thus, a child must have some ability to determine a linguistic competence before any performance abilities can be acquired. Chomsky names it the "language acquisition device" (Chomsky 30-32).

Three separate notions in Chomsky's theory of language acquisition must be further explored. Firstly, the poverty of the stimulus must be considered. Does the linguistic data provide enough information to form an adequate representation of the language, or is an

innate device really required to provide this adequacy? Secondly, the language acquisition device itself must be better explained. What is this so-called language acquisition device, and how does it function? Chomsky gives many reasons why such a device is necessarily innate, but makes it hard to picture how the device actually works. Finally, a discussion of Chomsky's notions of competence and performance in language acquisition must be further examined. Why and how would a tacit knowledge of rules exist previous to any use, or performance, of a language? Chomsky's theory of competence and performance is essential to the reasoning behind the language acquisition device, but the theory itself may need to be revised before an empiricist might accept any type of language acquisition device as possible.

Chomsky supposes that a child encounters very little linguistic data before gaining a competent grasp of the rules of the language, and the child probably understands even less of what it has encountered. We can divide this problem of underdetermination into two "paradoxes": (1) the production of new sentences, and (2) the problem of obscured linguistic data.

In the first paradox, one must consider how a child can construct an infinite number of sentences from a finite set of examples (Routledge 330, Chomsky 24). The apparent problem is resolved by Chomsky by way of generative grammars. The grammar by which a language user constructs sentences consists of a system of rules and a vocabulary of lexical primes. The system of rules Chomsky proposes allows for recursive construction of sentences in the language, and, naturally, could produce an infinite number of well-formed sentences (Putnam 137). Chomsky concludes that the competence

that allows for such a construction of sentences might be innately derived, with the aid of a language acquisition device.

However, exactly what must be innate is not clear. Hilary Putnam, in his critique of Chomsky's innateness hypothesis, suggests that Chomsky's assumptions may have no relevance to innate knowledge at all. He concedes that a human would need innate intellectual equipment to maintain a list of lexical primes, and to formulate a system of rules. It would be absurd to think otherwise—how else might a human learn anything, had no device for learning be given? But there is no reason to believe that "a particular mighty arbitrary set of grammars Σ is 'built in' to the brain" (Putnam 138), or that humans do not already have the natural capacity to learn and retain a list of phonemes. The ability to generate an infinite number of valid sentences from a finite amount of examples, then, might be attributed to the way humans learn, rather than to specific, innate grammatical tools.

The remaining paradox contributing to the problem of the poverty of the stimulus have to do with learning language. In only a few years, no one could possibly take often ill-formed sentences and, without much explicit guidance, devise a system of language using the simple mind of a four-year-old. Yet a child certainly does just this, and with relative ease, argues Chomsky.

A child's grasp of language despite its deficient evidence is truly amazing, but one must consider how deficient the evidence really is for the child. In second-language acquisition, a college student might take an immersion class in a foreign language, where s/he is bombarded with the language about three hours a week. Putnam argues that it would take the average adult about four years, or 600 hours of instruction and reading to

gain performance in the new language (Putnam 141). A child, however, encounters its native language even more frequently; 600 hours of direct-method language listening could pass in less than one year. Yet after four years of learning, the child has a small vocabulary compared to any mature student, and the child's grammatical mistakes are still numerous. Perhaps the "ease" in language learning is not really so easy for us, but is instead a process that holds many errors and misunderstandings.

There are many opposing opinions on the degree by which the poverty of the stimulus restricts language learning, most of which cannot be easily resolved. Chomsky's solution is inexorably tied to the language acquisition device, as the device provides exactly what stimulus will not provide. But what is this device, and what does it really do?

Chomsky describes a device that aids in forming the linguistic theory of a language user as having:

> (i) an enumeration of the class s_1, s_2, … of possible sentences
>
> (ii) an enumeration of the class SD_1, SD_2, … of possible structural descriptions
>
> (iii) an enumeration of the class G_1, G_2, … of possible generative grammars
>
> (iv) specification of a function *f* such that $SD_{f(i,j)}$ is the structural description assigned to sentence s_i by grammar G_j, for arbitrary i,j
>
> (v) specification of a function *m* such that *m*(i) is an integer associated with the grammar G_i as its value [by which alternative grammars can be evaluated]
>
> (Chomsky 30-31)

In other words, given a set of primary linguistic data, the device examines which of its possible generative grammars G_1, G_2, … will be "compatible". The grammar must

satisfy rule (iv), such that every sentence in the primary linguistic data has a functional correspondence to a particular structural description SD_i. In this way, each sentence in the primary list of sentences is described by a particular grammar, operating with a set of structural descriptions for each sentence structure in the grammar. There may exist multiple grammars that satisfy these requirements and handle the primary data with relative success, but the best one will be chosen according to rule (v). The language acquisition device has a function specifically available to compare and evaluate the success of each possible grammar, until one is chosen to be the best possible grammar.

When a list of possible grammars is determined by the device, one can say that the device has made available "a *class* of generative grammars containing, for each language…a grammar that (by means of (iv)) assigns structural descriptions to sentences in accordance with the linguistic competence of the speaker" (Chomsky 34, italics added). We can call this list of generative grammars *descriptively adequate,* in that a class of potentially working grammars is available to the speaker. However, this leaves a class of grammars, of which some might be more appropriate than others to provide a linguistic theory according to the data examined.

Rule (v) ensures that the best generative grammar is picked for the job. Thus, Chomsky's device can be described as *explanatorily adequate* in that it provides "a principled basis for selecting…[a grammar chosen by a] well-defined evaluation measure" (Chomsky 34). This grammar should be able to interpret any sentence in the language, including those that lie outside of the primary linguistic data. In this sense, the language acquisition device has constructed a tacit competence of the language, by which a child might be able to perform in the language. The work done by the device provides

far more understanding of the language than any amount of sensory-based observation could provide, Chomsky asserts, and the underlying linguistic competence "is in no sense an 'inductive generalization' from these [primary linguistic] data" (Chomsky 33).

The formation of a linguistic competence under Chomsky's approach expresses a rationalist method of language acquisition. The competence necessary for the use and understanding of a language is not, in itself, innately selected to be the working rules of the language user. On the contrary, the language acquisition device actively chooses the proper competence, given the outside influence of linguistic data. However, neither the data nor the device will adequately provide a grammar for competence on its own. Leibniz insisted that "the senses, although necessary for all our actual knowledge, are not sufficient to give it all to us, since the senses never give us anything but examples" (Leibniz, New Essays 42-43). Likewise, Chomsky believes that the sensory data about language does not provide sufficient information for the formation of a linguistic competence. On the other hand, a language acquisition device, complete with a list of generative grammars, would be useless without primary linguistic data to guide its decisions on an appropriate grammar. The necessity of both innate tools and empirical data suggests a rationalist explanation of language acquisition where one can consider the linguistic competence a synthesis, or an *emergent* property of both innate and sensory perceptions.

Chomsky's rationalist approach does not force him to assert that linguistic competence is innate. Instead, competence is "constructed" (Chomsky 32), and can be considered an emergent product of innate and empirical influence. To complete the previously suggested analogy, the emergence of linguistic competence resembles the

emergence of Hercules from a block of veined marble. The veins are the suggestions of grammars made by the language acquisition device, innately found in the block. The figure of competence is finally revealed as the primary linguistic data chisels away a clearer picture of the grammar, until a competence is formed that is explanatorily adequate.

One interesting and essential question remains to be answered in Chomsky's explanation of language acquisition: Exactly how do innate principles and empirical data work together to form an emergent theory of language? Chomsky offers rules (iv) and (v) as an algorithmic method by which competence can emerge, but the methods themselves are vague and primarily unexplained. Given that the user has neither the competence nor the performance capabilities for language, how would the device take a string of input sounds and decipher it into phrases, word boundaries, or syntactical elements? The jump from complete ignorance of the meaning of these sounds to a tacit understanding of the underlying theory of language still seems outrageous without an explanation of the more complicated behavior of this language acquisition device. Therefore, to make the device plausible, one must explain *how such a device might operate in finding a grammar that provides explanatory adequacy for a language*.

The language acquisition device thus described suggests a sort of algorithm for finding a particular grammar by which a linguistic competence might form. A mathematical method to generate a list of potential grammars, as described by (iv), and a similarly precise method of deciding which grammar is best, as described by (v), sounds appealing, to say the least! It is hard, however, to determine what sort of system is capable of this decision-making. The intent of this paper is to present such a system--a

system that could account for the details of the decision-making that goes on in the language acquisition device. This system can take primary linguistic data and decide upon an explanatorily adequate linguistic competence. Most interestingly, a promising model of this system is currently found in artificial intelligence research.

II. Artificial Intelligence and Language Acquisition

Over the last sixty years, the development of artificial intelligence has fostered a debate between rationalist and empiricist scientists that resembles the previously introduced spilt in linguistic theory. The two schools of research attempted to create intelligent machines with different philosophical attitudes about how knowledge can be represented in the brain, and the consequent systems proposed by the groups had fundamental differences. A theory of linguistic competence is nearly impossible to be produced by any computational system that can be proposed by the first, traditional school of AI. By appealing to the other, there remains a possibility that linguistic competence can be explained through the interaction of innate learning devices and empirical data, and this system will yield us a successful and detailed model of a language acquisition device.

The first group proposed an atomistic representation of knowledge inspired by the rationalist influence of Descartes, Leibniz, Kant, and early Wittgenstein. Originally led by Rand Corporation researchers Allen Newell and Herbert Simon, they proposed a *physical symbol system* for knowledge (Dreyfus and Dreyfus 310). This system encodes information in the form of binary digits, and manipulates these symbols according to a system of rules. Computers, by their nature, are very adept at manipulating symbols. All that remains for these scientists, then, is to represent logical primitives of knowledge and the rules of the mind that would manipulate them, and a intelligent system would have been created.

The symbol system theorists attempt a formalized system of rules and symbols that will model the behavior of the mind. The external behavior should mimic intelligent

thought, but any differences between the internal structure of the formalized system and the internal workings of the brain are of little consequence to the design. Facts about the world are abstracted into essential elements and then placed in a symbol system that can easily manipulate those elements, producing a composite output from these essential elements.

This approach to artificial intelligence has produced stunning preliminary results[1], and gained the support of the greater scientific community. However, as the development of this "traditional" AI continues, and as the scope of the machine's intelligence is required to increase, more difficult problems arise. *Common-sense understanding*, which is successfully used by humans to solve novel situations, has proven to be extremely difficult to formalize in a symbol system. These systems can manipulate lists of knowledge units to solve complicated problems, but are very unsuccessful at applying the units to a similar situation with slightly new conditions. This occurs because the systems are simply manipulating a list of bit symbols according to instructions provided by other lists. When a new problem presents itself, the machine is unable to extrapolate a new rule or new element outside of its given set. Results we derive with the help of common-sense are impossible to program into a list of rules, as there could potentially be an infinite number of slightly different results following from an infinite number of slightly different situations. Even Marvin Minsky, an veteran advocate of the physical symbol system,

[1] Weizenbaum's Eliza program (1966), Winograd's Shrdlu (1972), and more recently, Deep Blue, have been the source of amazement when introduced. Eliza was designed to communicate with humans as a computer psychotherapist, and although it used a very primitive method of communication by today's standards, it was once declared "ready for clinical use" in the *Journal of Nervous and Mental Disease* (1966). Shrdlu's skill in manipulating "objects" on a computer screen according to a human's instruction was quite impressive, as is Deep Blue's brute-force method of calculating chess moves, which lead it to a

admits that “a ‘minimal’ common-sense system must ‘know’ something about cause-effect, time, purpose, locality, process, and types of knowledge…[and] we still know far too little about the contents and structure of common-sense knowledge” (Minsky 124).

The inability of physical symbol systems to make common sense decisions produces a strong limitation on its ability to use and understand language. Symbol systems require a knowledge base of logical simples by which it can break apart and manipulate input, known as the *data model* of knowledge. For a machine to learn language effectively, the knowledge base should contain linguistic primes, structural descriptions, and rules designed to contain a particular competence for a language. However, a finite enumeration of these rules would not be sufficient for a linguistic competence. Consider, for instance, the ability to generate and understand an infinite number of sentences, given a finite number of words—the first paradox of language acquisition. This ability requires a speaker to make common-sense analogies to similar contexts, such that one might derive a new meaning out of comparison with older ones. A knowledge base of words, sentences, and structural descriptions would fail to provide this power of “know-how”, or the ability to expand rules in a novel way. Thus, the data model of knowledge fails to provide an adequate explanation for the poverty of the stimulus. It cannot provide a successful competence because common-sense understanding can not be successfully formalized as a set of rules. Instead, “know-how” is a skill of “knowing what to do in a vast number of special cases” (Dreyfus 325), and requires the assistance of an intuition only gained through learning and experience. A symbolic representation cannot

---

world chess championship, but both lack the ability to expand their knowledge outside of the micro-world in which they inhabit.

correctly capture a skill based on experience rather than rules, and, consequently, no symbolic representation will successfully capture a complete competence.

Connectionism and neural networks

Just as Wittgenstein abandoned a rational atomization of language and its rules in his later Philosophical Investigations (1953), so have many computer scientists. In response, researchers have turned back to *connectionism*, a field in artificial intelligence once described to be "generally disappointing…[and one which] in no part of the field have the discoveries made so far produced the major impact that was then promised" (Dreyfus 314). However, the structural approach to AI that connectionism provides has made it an appealing outlet for research once again. Whereas a physical symbol system attempts a formalization of rules that should model the external behavior of the mind, a connectionist system focuses on the internal workings of the brain. Based on neuroscience more than philosophical tradition, connectionism simulates the *structure* of the brain, rather than its *behavior*. Connectionist scholars argue that "it is both easier and more profitable to axiomatize the *physical system* and then investigate this system analytically to determine its behavior, than to axiomatize the *behavior* and then design a physical system by techniques of logical synthesis" (Rosenblatt 386).

In a connectionist system, the structure of the brain is modeled by establishing an interconnected system of processing units. These units are meant to be analogous to neurons in the brain. The connections leading into a unit carry a certain connection strengths, by which another unit might send a communicating output signal. If the collective strength of all connected units ever reaches the threshold level designated to the

unit in question, then this unit will, in turn, “fire” to all neurons to which its output is connected (Copeland 208-210). In this way, complex networks of units can be built into systems that operate in a similar fashion to neurons in the brain.

These “neural networks” have developed into more sophisticated systems that can be used for AI research. As with neurons, one can allow nodes to collect input from both other nodes and from the outside world (Elman 50). The connections between nodes can be unidirectional (known as *feedforward* networks), bidirectional (Elman 51), or *recurrent*, allowing connections from the nodes to themselves (Elman 73-74). Recurrent connections allow the network to reflect external input it received, as well as its previous internal state. Different types of node organization add further complexity to the networks.

Adjusting the connection strengths of nodes in the network allows the system to “learn” a certain type of behavior. By altering the right connections between sets of nodes, the network is taught to respond appropriately to its input. This method of teaching is not new at all; in fact, Donald Hebb suggested that human nervous systems learn in the same way:

> When an axion of cell A is near enough to excite a cell B and repeatedly or persistently takes part in firing it, some *growth process* or metabolic change takes place in one or both cells such that A’s efficiency, as one of the cells firing [to] B, is increased. (Hebb 1949 62, italics added)

A natural growth process strengthens the connection between the two neurons through patterns of frequent activity, according to Hebb. The more times a neuron fires to another, the stronger the connection between the two becomes. This system of Hebbian learning is implemented in multi-layered neural networks, where hidden layers are

adjusted to produce the desired output, given a specified input. Hidden units allow networks to internally represent inputs within the system (Elman 64), and their adjustment is a way of teaching the network about the input, as well as how to react to the input. The method by which adjustments are made to hidden layers, teaching the layers about the input, is known as *backpropagation of error*.

Backpropagation, in its simplest sense, compares the actual output and desired output of a network, and adjusts the network to produce an output with less error. Most backpropagation networks begin with a series of node layers, each with randomly assigned connection weights. An input pattern will work its way through the network, activating some nodes and skipping others, until some output exits the system. The output will inevitably be far from what is desired of the network, but it is precisely this difference between desired and actual outputs that allows the backpropagation process to succeed. In accordance with the rules of Hebbian learning, every node found in the hidden layers is slightly adjusted, proportional to its error in producing the wrong output and its influence on the rest of the network (Elman 69). A node that greatly influenced many other nodes to lead to the wrong output will be more severely adjusted than others, but it is important to note that none of the nodes are changed dramatically. Instead, the slight adjustment of all nodes in the layers allows the network to correctly produce the output pattern as desired without dramatically changing the overall behavior of the system.

The overall goal of backpropagation is to find a set of connection weights that allow the network to correctly produce output for a large set of inputs. Thus, the slight adjustment of many nodes in the network is an invaluable method of instruction for the

network. By keeping the basic behavior of the network the same, we can repeat the process of backpropagation for many sets of input-output pairs, and build a repetoire for the network. The connection weights are refined through this repetition until the system can produce the correct output for any input in our set of input-output pairs. Backpropagation completed, the network has been successfully trained.

Characteristics of neural networks

One can observe unique and valuable behaviors resulting from a well-trained neural network of the type described. We will examine two characteristics of a trained network that have important implications for any philosophical understanding of the way the networks work. In particular, we turn to the distributed method of storage and the ability to generalize input into categories.

Most neural networks are modeled after a parallel and distributed design first established by Hinton, McClelland, and Rumelhart (1986). The method of parallel distributed processing, or PDP, is modeled after the behavior of the brain. A *distributed representation* holds each unit of information in a system by some pattern of activity distributed over many computing elements, as opposed to using only one computing element to represent each unit (Hinton 248). In the case of backpropagation, the adjusting of connection weights is such a distributed pattern of activity. This holistic approach to storage keeps information spread throughout the system, making the global activity of the system the focus, rather than any particular node. One cannot “point” to the location of a certain unit of knowledge, as it is the entire system alone that can be considered to have the knowledge.

A distributed system provides power in areas where localized systems are lacking, but it is deficient in just what the von-Neumann machines provide. PDP networks have trouble storing a large amount of arbitrary associations, as demonstrated in the increasing difficulty found in teaching a well-trained system a new fact (Hinton 249). However, the distributed storage techniques used in neural networks allow units of knowledge to be stored in a *subsymbolic* manner (Sharkey 177). In contrast to the approach of physical symbol system machines, "the currency of [distributed] systems is not symbols, but excitation and inhibition" (Copeland, 219). A distributed system encodes knowledge in the strengths of its connection weights. The knowledge is reflected in the improved structure of the network, and can be observed in the resulting behavior of the network. Distributed networks avoid using the data model of knowledge, which consequently avoids some of the problems encountered in the formalization of knowledge, namely, the common-sense problem.[2] The distributed approach to data representation provides one

[2] Some might argue that a subsymbolic representation of knowledge does not really add any new power to the system. According to the Church-Turing thesis (1936), any algorithmic process can be simulated by a universal Turing machine (Lewis and Papadimitriou 245-250). We can describe a PDP system to be algorithmically calculable, as the connection weights and thresholds could be enumerated in a mathematical set, and then algorithmically manipulated to simulate the output of the network. Any PDP network could be simulated by a Turing machine (or any physical symbol system) in theory, and it should follow that any power of subsymbolic representation could be contained within a symbol system.

Although it is undeniable that the universal Turing machine is a powerful device, there are certain limitations that keep it from being equivalent to a PDP network. Neural networks are massively parallel, allowing multiple computations to occur simultaneously. Turing machines can theoretically process the same calculations sequentially, but the speed and efficiency would decrease exponentially. A physical symbol system simulation of a massively parallel network may not be polynomially bounded or practically feasible (see Lewis and Papadimitriou 275-277, Churchland 232, Turing (1950)).

Even aside from its decided advantage in speed and practicality, the behavior of PDP networks can only be *simulated* by a physical symbol system, which further differentiates it from traditional approaches to AI. The neural network in theory and the network in simulation are two different entities. A simulation of a network will produce the same output as the simulated entity, but the internal states of the two machines can differ completely. Their internal differences keep us from calling the two machines equivalent. Because simulation of a PDP network may not exhibit the same internal behavior as the actual network, it may not contain the same structural power found in the original network.

useful tool to aid in resolving the first paradox of the poverty argument, and we will find the other necessary tool in the generalization of input.

Of all the advantages offered by neural networks, the ability to produce the correct output for a new and unseen input is among the most intriguing. By its nature, backpropagation learning promotes *generalization* in the network (Hinton 252). That is, a properly trained network will have connection strengths carefully adjusted to reproduce the correct output for a large number of inputs. The inputs are usually of the same type, so the network needs to be adjusted less and less as it forms a better picture of its input class, and assigns more accurate weights to handle it. So the input class becomes "embedded" in the structure of the network through backpropagation. Consequently, *the network forms a generalized picture of its input class*.

The ability to generalize allows the network to then determine the correct output for a new but similar input. "Neural networks are...a kind of analogy machine" (Elman 59), where similar inputs will yield similar outputs. The network learns to extrapolate outside of its instruction set, and *it can now produce novel output, given a new input pattern*. The process of finding similarities among a group closely resembles the way in which a human acquires knowledge. We take a number of facts and use them to make more general claims about things. For example, a person can abstract the general idea of a tree by observing similarities in the different times that person has seen different instances of trees in the world. The abstraction is extremely powerful because the person can then classify other objects to be trees by way of interpolation, and even generate a picture of a tree never before seen through extrapolation of the general idea.

The generalizations made in neural networks should seem curiously similar to the general ideas found in Locke's explanation of knowledge and knowledge acquisition. Locke held that general ideas were abstracted from particular ones through a process of reflection, and that these general ideas allowed for novel and creative thoughts. It is therefore tempting to consider the generalizations of neural networks to be equivalent to these general ideas, and likewise, assume that backpropagation methods of learning result in the same gains as reflective thought does for Locke. The methods are indeed similar, as neural networks form their generalizations from a collection of empirical data, but the general ideas formed by Locke's reflection are necessarily composed of elementary ideas. In contrast, neural networks can form associations from any type or degree of similarity at all (Elman 59)[3].

One can now offer a more detailed explanation for the first paradox in the poverty argument proposed by Chomsky. A connectionist system could theoretically be trained with any type of input set—even with linguistic data resembling the set of words encountered by a child in the early stages of linguistic development. A distributed system of data storage, coupled with a system of Hebbian learning, successfully forms generalizations of input data. A *connectionist language processor* would form generalizations of linguistic classes necessary for the creation of new sentences in the

[3] Because generalizations are formed from any similarity found in the input patterns, the power of generalization may sometimes become more of a hindrance than a help to the network. Mistakes made in the backpropagation of a learning set might mislead a network to make false generalizations about the data. As more of these false generalizations are made, it becomes more difficult to override them with further backpropagation trials (see Elman 130). Thus, the backpropagation process is both crucially important for learning and a potential threat to the integrity of the system, as it might alter its generalizations in damaging ways. Careful execution of the learning procedure is necessary for any neural network to be successful in generating the correct output based on general ideas rather than specific examples or rules, just as accurate reflection reveals accurate general ideas in Locke's theory.

language. It would not need any special preparation of nodes before training; the nonsensical data it processes would become more clear as the generalizations emerged from backpropagation trials. The emergent categorizations could even come from a limited and obscured amount of linguistic data. This generalization might provide a grammar sufficient for competence in the language, and provide a solution to the second paradox—the problem of limited linguistic data. Also, the network could receive a finite amount of linguistic data, and produce a potentially infinite set of sentences, given a successful generalization. The power of abstraction allows the network to use a common-sense approach to extrapolate new sentences, and effectively solves the remaining piece of the paradox.

The connectionist language processor sounds very exciting in theoretical terms, but how feasible are these accomplishments? One must turn to recent research to find specific evidence of such language processing.

Neural network language processing: Case studies

Rumelhart and McClelland's work in neural network verb conjugations is just one of many projects on language processing conducted in the last fifteen years. Their particular work enjoyed a restricted success, considering the fact that the network used neither hidden layers to increase the level of possible training nor a true backpropagation process, which leaves some residual error by the nature of the problem[4]. More notably, the input sets were adjusted for each learning trial in a way that would completely guide

its development. The adjustments generated a successful machine, but the input sets contained an overwhelming number of irregular verbs in the outset, which is far from the nature of linguistic data most people encounter when learning English[5].

Other past-tense networks have provided even more successful results without altering the input set quite as radically by using feedforward networks and hidden unit layers. Plunkett and Marchman (1991) fed a hidden layers network an input set where irregular verbs were present more often than regular verbs, but each verb was presented only once during a backpropagation trial. The network was "given an opportunity to establish [irregular verb patterns] before the regular verb patterns became too firmly established" (Elman 139), avoiding overregulation errors that would lead to an unsuccessful network. Later, the two researchers used a small input set for the first few trials, and increased the size as the sophistication of the network increased, much like a child would encounter more verbs as the capacity for observing elements of language develops. This 1993 work yielded a pattern of growth in conjugation of regular and irregular verbs that was not only successful, but also mirrored the growth of a child's vocabulary (Elman 144-145).

Other explorations of neural network language processing have used a general connectionist architecture to distinguish words and word types. Jeffery Elman (1990) constructed a recurrent network without any pre-calculated representational constraints to

---

[4] Instead, they used a one-layer perceptron for the network and the Perceptron Convergence Rule for its learning algorithm, which can only be used for one-layer networks. See the Plunkett and Marchman (1991) analysis of Rumelhart and McClelland's work for a more detailed explanation.

[5] See Pinker and Prince (1988) for more discussion.

distinguish word-boundaries in a continuous stream of phonemes (elementary units of words). The network was given strings such as:

Manyyearsagoaboyandgirllivedbytheseatheyplayedhappily.

One phoneme at a time, the network processes the data and guesses the next phoneme. Because the network is recurrent and has hidden units, it was able to keep track of previous phonemes, which aided in the decision process (Elman 120). Specifically, "context units" communicate with the input and output units in a recurrent fashion to influence the decision based on the context of the current phoneme in the word string (Sharkey 175). The influence of the context units works with the connection strengths of the network to guess the next phoneme. Of course, these connection strengths are tuned by backpropagation of error. Thus, the network bases its phoneme decision on both the history of the phonemes found in previous words and on the particular context of the sentence.

As expected, the error of the network was highest at the beginning of each new word, and decreased as it reached the word-boundary. One can consider the error to be "a measure of the level of confidence with which the network is making its prediction" (Elman 121), and so the network would be expected to gain "confidence" as each word became more familiar. The context units greatly aided in cases of lexical segmentation, where "aboy" could be separated into "a" and "boy" due to contextual memory of the input. The input itself was simply a list of arbitrarily different phonemes, and the system did not receive any special assistance in learning from the nature of the input, unlike the Rumelhart and McClelland (1986) past-tense network. Elman found that the network

required "a long initial period" for learning, but it was most successful at determining the next phoneme after the training was complete (vowel or consonant, etc.).

The most intriguing part of Elman's work came out of the fully trained network. Eventually, he found that the network

> "learns to predict, not necessarily the actual phoneme, but the correct *category* of phoneme (vowel or consonant, etc.). Thus, the network progressively moves from processing mere surface regularities to *representing something more abstract*, but without this being built in as a pre-specified phonemic or other linguistic constraint." (Elman 123, italics added)

Despite the fact that the network started with only general architectural constraints, it was trained to create language-specific representational constraints over time. Eventually, a phonemic categorization emerged from training. The network's repeated experience with phonemes, in combination with backpropagation learning, allowed a generalization of the specific input data. This abstraction of the data allowed the network to categorize the input in broader phonemic categories. The network can then categorize new input in terms of its new categories, without having the categories embedded in the architecture from the outset. Clearly, connectionism has scored an (arguably large) victory: *the desired behavior of phonemic categorization has emerged simply from the nature of the system and its training methods*, whereas traditional AI has only attempted to program this behavior into the system through fabricated rules.

The word-boundaries network is not alone in generating such emergent behavior. Elman (1990) set up a similar network to parse lexical items in short sentences, and found

similar results in the trained network. By a statistical analysis[6], Elman found that the network had

> "developed hidden unit representations for the input patterns that reflected information about the possible sequential ordering of the inputs, e.g. the net knew that the lexical category VERB followed the lexical category NOUN" (Sharkey 174).

Again, a categorization had arisen from training that allowed the network to represent both token and type information about its input set. This particular network actually subdivided the noun category into groups of animate and inanimate objects, and still further, into human and non-human animates (Sharkey 175)! A sophisticated level of abstraction emerged, guided only by the structure of the network and its learning process.

A separate study by Plunkett, Sinha, Moller, and Strandsby (1992) demonstrates the ability of a neural network to recognize obscured input patterns as variations on specific prototypes. In their work, an auto-associative network was trained to recognize 32 input patterns based on training set of 192 distortions of these patterns (Elman 125). With the proper training, the network was able to abstract the prototypes from the distortions it was given. Eventually, Plunkett found that the performance on the prototypes, which the network has never encountered, exceeded the performance on the distortions on which it had been trained (see Fig. 1). Despite an obscured data set, successful abstractions of the prototypes emerged, and these abstractions allow the network to overcome the deviations in the input sets.

---

[6] Specifically, a hierarchical cluster analysis displays the relationships between the representations in the network in a spacially-oriented way (Sharkey 170-172, 174). See Elman (1990) for his analysis.

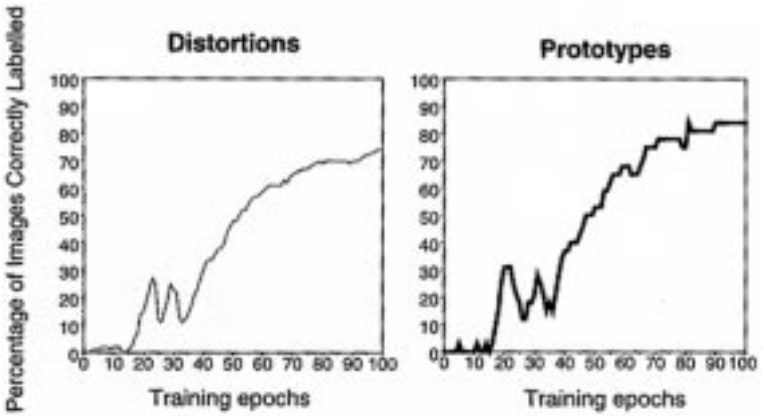

Fig. 1: The Plunkett et al. (1992) auto-associative network was trained only on distortions, but eventually grew to associate the prototypes even more successfully than the original input. Note the higher success rate on prototypes that occurred later in the training cycles.

Certainly, the evidence found in research supports the possibility that a connectionist network could overcome the poverty of the stimulus in acquiring linguistic competence. If limited and obscured data can be abstracted into generalized forms, then ill-formed linguistic data may not restrict a connectionist system's ability to form a linguistic competence after all. By the nature of the system, slight discrepancies in the input data are ignored in favor of a more general understanding of the input set. Additionally, Elman's studies revealed the emergence of linguistically specific categorizations from a non-specific network. The evidence suggests that a connectionist language processor can successfully establish a linguistic competence from what seems to be a poverty of linguistic data. The emergence of such a competence has, in the case of Plunkett and Marchman's verb recognition network, even been shown to mirror the development of children's knowledge of language. Not only does this evidence offer a explanation of linguistic competence found in artificial intelligence, but it offers an explanation that is completely relevant to our understanding of human language acquisition.

III. Consequences for the Language Acquisition Device

Given their similar structure to the brain, how valuable are connectionist language processors in explaining language acquisition? It is easy to assume that the two systems are essentially the same, but a further examination is necessary before any claims about the two are made.

Firstly, it should be mentioned that neural structure in human brains is far more complex than any neural network created to date (Copeland 222). PDP neural "nodes" only resemble the neurons in the nervous system in a basic sense, and do not provide the functionality of any particular type of neuron. Instead, they best resemble an abstracted hybrid of real neurons, where the capabilities of many neurons (such as inhibition or exhibition, which are controlled by separate neurons in the brain) are collapsed into one "node". However, these nodes may or may not add any new power to the system's operation, and likewise, may or may not lose any functionality. Therefore, it cannot be concluded that the node representation used in neural networks harmfully alters the system in a way that would separate the two.

From the opposite direction, many of the principles of the nervous system that differentiates it from connectionist networks have not been proven to be essential. For instance, the brain has a highly organized geometry, arranging neurons and dendrites in very specific ways (Copeland 224). However, not enough is known about the brain to show that this geometry is essential to a nervous system, and we should not conclude that connectionist systems are deficient in any essential way.

Ultimately, the mentioned similarities between connectionism and the nervous system are not even *necessary* to make claims about the operations of the language

acquisition device. One only needs the necessary features of a connectionist network to be present in the nervous system. If these features are present, then we can consider the brain to be a connectionist network as well. Thus, any conclusions about language acquisition for connectionist networks could then apply to the nervous system.

What are these essential characteristics of connectionist networks? They are the inherent structure of the system and the method of learning; or more specifically, a distributed, recurrent architecture, and a Hebbian process of learning. As neuroscience discovers more about the nature of the brain, it is becoming clear that the nervous system naturally stores information in a distributed manner. A distributed representation is the only way to account for the content-addressability of human memory, and its graceful degradation[7]—properties that cannot be reproduced by localized representations (Copeland 222). And although the process of learning has not been proved to involve the strengthening of connection weights, it is very probable that the brain implements a version of Hebbian learning. The brain operates in a connectionist way; thus, a connectionist understanding of language acquisition offers a logically plausible understanding of any human language acquisition device.

Conclusions on language acquisition in a connectionist system

It is unreasonable and incorrect to assume that a set of grammars Σ is *explicitly* provided by the brain. Although Chomsky's theory of language acquisition may lead

[7] Memory in the human brain has been shown to "degrade gracefully" such that the loss of functionality in one section of the brain may not seriously effect memory in any way, given that the area is not overwhelmingly large. See Wood (1978), Nelville et al. (1991), and King and Kutas (1995) for a discussion of lesion studies, ERP studies, and distributed memory in the human brain.

many linguists to such a conclusion, it is not necessarily what Chomsky had in mind, and is due to a misconception in linguistics. It seems natural for a linguist to describe the cognitive system as a language tool governed by rules, as much of language is described by rules of syntax, grammar, etc. in common practice. However, language is really a method of communication that *can be described by rules*. Similarly, the grammars can be described by specific sets G_1, G_2, …, but, in a connectionist system, they are only *implicitly* present. The possibility of the establishment of a particular grammar G_i is present *simply by the nature of distributed representations in the brain*.

We have observed that as a distributed network grows in the learning process, an emergent property of that development is the categorization of relationships in the network. The categorization can overcome slight discrepancies in its training data in forming these categorizations. This connectionist method of abstraction allows neural networks to form general concepts that, in turn, allow the network to interpret and produce novel input. In a connectionist language processor, it is precisely these emergent general concepts that can be described as the specific grammar G_i in Chomsky's theory that satisfies rules (iv) and (v), and produces a competence in the language.

Most interestingly, the ability to abstract and categorize is provided by the *structure* of the system itself—namely, its connectionist characteristics. The distributed, Hebbian learning method used in the brain implicitly includes the possibility for the correct emergent grammar. In fact, it is the structure of the brain holds the infinite enumeration of grammars G_1, G_2, … by way of the infinite configurations of a distributed network that holds information in the (infinitely modifiable) strengths of its connections.

This might lead to the conclusion that a connectionist system has an innately specified device that provides an *implicit* representation of grammars, and that device is simply the network itself. This conclusion is accurate, but incomplete in describing the process of language acquisition, and might lead to the inaccurate assumption that the tacit knowledge of language is *innately* (albeit implicitly) specified in the brain.

Although the innateness of the structure of the brain is a prerequisite for language acquisition, it is merely *descriptively adequate* for producing any linguistic competence. The brain satisfies rule (iv), providing an enumeration of grammars through its connectionist structure, but it requires experience to be explanatorily adequate. Chomsky clearly states that this process requires the presence of linguistic data. Only through experience with linguistic data can the connectionist brain "choose" a specific grammar for the language it has encountered. And this choice is really the emergent product of a distributed system using Hebbian learning to adapt to the data. Thus the connectionist system itself is *not explicitly designed to acquire language*; instead, the ability to acquire language falls conveniently from its structure, when offered linguistic input.

Revisiting Leibniz's poetic analogy once again, one can consider the brain a block of marble containing a connectionist method of growth and learning in its veins. No specific group of veins are intended for language acquisition or explicitly contain the figure of the perfect grammar, but the veins implicitly supply opportunity for many linguistic grammars. Through an often ill-guided and erroneous process of chiseling, the figure of a particularly successful grammar begins to emerge. The veins guide each stroke, but do not dictate the emergent figure. Instead, they make up for slight inadequacies in the artist's attempts, aiming towards a generally smooth cut. Any rough

strokes thrown against the veins will be resisted, and the resistance grows as the emergent figure becomes better defined, and more structurally sound. Finally, the deficiencies in the veins themselves must be polished away if the artist is to aim for a perfect competence. Linguistic data aims to chisel away a particular competence, and the veins provided by connectionism allow the marble to fall into an explanatorily adequate one.

As Locke suggested, general ideas form from reflection upon sense-data, although such a reflection would occur unconsciously, as it is due to the structure of the brain itself. Likewise, the competence that emerges from such an unconscious reflection will be a tacit one. In this sense, the linguistic competence is in complete accordance with Chomsky's theory of language acquisition.

We can now consider the process of constructing a linguistic competence an emergent property of the greater system of empirical data and innately specified structure—a synthesis of rationalist and empiricist representations of knowledge. In fact, the process is simultaneously responsible for an emergent linguistic *performance*, and this is no coincidence. As general ideas develop through abstraction, the connectionist system becomes more adept at the use and comprehension of language, as shown by the experiments reviewed earlier. The linguistic performance is a reflection of the linguistic competence that is developing. Equivalently, the linguistic competence that can be described by rules is a reflection of the performance that has formed. Linguistic competence and performance are really two external observations of the same internal phenomenon: language acquisition.

Postscript

The connectionist understanding of language acquisition is a relatively new one, even considering the age of the field from which connectionism was originally established. Frank Rosenblatt's work in the 1950s was soon dismissed from artificial intelligence research because the results did not produce the "intuition, insight, and learning [that were once] exclusive possessions of humans" (Rosenblatt 6). It took almost twenty years for a resurgence of interest, arguably spawned by a lack of symbol system success, to bring connectionism back into the light.

Despite the modest gains in early connectionist networks, the theoretical concept has always been a powerful one. The lack of research in a philosophical understanding of connectionism and language acquisition is, in my opinion, really due to the lack of interdisciplinary efforts. Only within the past ten years have any attempts been made to explain language acquisition in a connectionist way, and most of them are more recent than this. I hope to see more collaboration between artificial intelligence, linguistics, and philosophy in the future. An interdisciplinary approach to a traditionally linguistic problem has certainly added a new approach to understanding language acquisition, and its value should be taken seriously.

References

Chomsky, Noam. Aspects of the Theory of Syntax. Cambridge: MIT Press, 1965.

Copeland, Jack. Artificial Intelligence: A Philosophical Introduction. Oxford: Blackwell, 1993.

Dreyfus, H. L. and S. E. Dreyfus. 1988. "Making a Mind Versus Modelling the Brain: Artificial Intelligence Back at a Branch-Point." In The Philosophy of Artificial Intelligence, ed. Margaret A. Boden, 309-333. Oxford: Oxford University Press, 1990.

Dreyfus, H.L. What Computers Still Can't Do: A Critique of Artificial Reason. Cambridge: MIT Press, 1992.

Elman, J.L. "Finding Structure in Time." Cognitive Science, 14 (1990): 179-211.

Elman, J.L. "Learning and Development in Neural Networks: The Importance of Starting Small." Cognition, 48: 71-99, 1993.

Elman et al. Rethinking Innateness: A Connectionist Perspective on Development. Cambridge: MIT Press, 1996.

Harman, G. H. "Psychological Aspects of the Theory of Syntax." The Journal of Philosophy 64 (1967): 75-87.

Hebb, D.O. The Organization of Behavior: A Neuropsychological Theory. New York: Wiley, 1949.

Hinton, G.E. et al. "Distributed Representations." In D.E. Rumelhart and J.E. McClelland (eds.) Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol.1, Foundations. Cambridge: MIT Press, 1986.

Katz, J.J. The Philosophy of Language. New York: Harper and Row, 1965.

Leibniz, G.W. New Essays on Human Understanding. Translated by Peter Remnant and Jonathan Bennett. Cambridge: Cambridge University Press, 1981.

_____. Philosophical Essays. Translated by Roger Ariew and Daniel Garber. Indianapolis: Hackett Publishing, 1989.

Lewis, H.R. and C.H. Papadimitriou. Elements of the Theory of Computation. New Jersey: Prentice-Hall, 1998.

Locke, John. An Essay Concerning Human Understanding, ed. Peter H. Nidditch. Oxford: Clarendon Press, 1979.

Lyons, John. "On Competence and Performance and Related Notions." In Performance and Competence In Second Language Acquisition, ed. Gillian Brown et al., 11-32. Cambridge: Cambridge University Press, 1996.

Minsky, M. "A Framework for Representing Knowledge." In Mind Design, ed. J. Haugeland, 95-128. Cambridge: MIT Press, 1981.

Plunkett, K, and V. Marchman. "U-Shaped Learning and Frequency Effects in a Multi-Layered Perceptron: Inplications for Child Language Acquisition." Cognition, 28: 43-102, 1991.

Pulnkett, K. et al. "Symbol Grounding of the Emergence of Symbols? Vocabulary Growth in Children and a Connectionist Net." Connection Science, 4 (1992): 3-4, 293-312.

Putnam, Hilary. "The 'Innateness Hypothesis' and Explanatory Models in Linguistics." In Innate Ideas, ed. S.P. Stich, 133-144. Berkeley: University of California Press, 1975.

Rosenblatt, F. "Strategic Approaches to the Study of Brain Models." In Principles of Self-Organization, ed. H von Foerster. Elmsford: Pergamon Press, 1962.

"Noam Chomsky." In Routledge Encyclopedia of Philosophy, Vol. 1, ed. Edward Craig et al. London: Routledge, 1998.

Shank, R.C. Explanation Patterns: Understanding Mechanically and Creatively. Hillsdale: Lawrence Erlbaum Associates, 1986.

Sharkey, N.E. "Fundamental Issues for Connectionist Language Processing." In Performance and Competence In Second Language Acquisition, ed. Gillian Brown et al., 155-184. Cambridge: Cambridge University Press, 1996.

Stainton, R.J. Philosophical Perspectives on Language. Ontario: Broadview Press, 1996.

Stich, S.P., ed. Innate Ideas. Berkeley: University of California Press, 1975.

Turing, A.M. "Computing Machinery and Intelligence." Mind 59 (1950): 433-460.

Wittgenstein, Ludwig. Philosophical Investigations. Oxford: Blackwell, 1968.