SCHEMATA: BOOTSTRAPPING LANGUAGE ACQUISITION

Sean McLennan

Dr. Robert Port
L645: Language and Cognition
April 28, 1999

1. INTRODUCTION

It has been demonstrated that Genetic Algorithms (GAs) can perform extremely efficient searches of "solution space" in order to find optimal solutions to complex problems. John Holland (1975) details and David Goldberg (1989) further develops) a theory of schemata to show how even poor solutions hold implicit information about the targeted good solutions. This accounts in part for how GAs can do what they do.

However, it may be that the principles of Schema Theorem extend beyond GAs. One of the arguments against statistical approaches to language acquisition, for example, is that the amount of input required to learn complex linguistic structures would greatly outweigh that which is exhibited in reality. Schema Theorem may hold a key to understanding why this argument is invalid.

Thus, the purpose of this paper is to explain Schema Theorem and explore its potential as a more general principle of information processing by examining language acquisition in particular. An effort is made to restrict the discussion as much as possible; however, inevitably, the implications and arguments presented here will apply to and draw on areas of cognition other than language. This should be interpreted as the possible pervasiveness and usefulness of Schema Theorem.

Should the generalization of the principles of schemata to language acquisition prove valid, the consequence will be that dynamic, statistical models of language and language learning may be more efficient than is currently believed, lending support to the growing argument against assumed innateness.

Section 2. discusses Schema Theorem in depth and how it is applied to genetic algorithms. Section 3. describes how the same fundamentals can be applied to language acquisition, and Section 4. summarizes and suggests areas for further study.

2. GAs and Schema Theorem

Originally developed as an tool to study natural evolution, genetic algorithms have proven to be an efficient tool in computation. The working assumption that fuels their functionality is that nature's principle of "survival of the fittest" is an effective problem solver. For an in depth introduction of the mechanics of GAs, the reader is referred to Mitchell's (1996) excellent discussion; however, it is necessary to outline a few basics.

GAs consist of three typical elements: 1) a population of chromosomes (usually a bit-string of ones and zeros), 2) a fitness function, and 3) a process of mating chromosomes to produce new "offspring". The fitness function is a characterization of a problem to be solved and each chromosome represents a possible solution. Based on the fitness function, chromosomes are assigned a "fitness evaluation" depending on how well they solved the problem. The more fit chromosomes are "mated" by a process a crossover and / or mutation that produces "offspring" to comprise a new "generation" of chromosomes.

A simple example would be to search for the maximum value of a mathematical function, say, $f(x) = \sin(x)$. In this case the bit-string chromosomes represent real numbers. In the process of evaluation, the bit-string is converted to a decimal number, entered into the equation as x, and assigned a higher fitness rating the higher the resultant value. The chromosomes with the highest fitness ratings are mated producing the next generation. After several generations, a chromosome with 100% or near 100% fitness usually emerges.

It is not immediately clear how the search of solution space is guided by fitness. In a sense, we are not concerned with fit strings alone; we are concerned with the *similarities* between fit strings. Holland (1975) introduces the formal notion of "schemata" to explain. Schemata are essentially similarity templates that describe a set of chromosomes that share values in certain positions. To describe a schema we must add the wildcard "*" to our string notation. Thus, the schema *0 describes a subset of 2 chromosomes: {10, 00}; 1** describes the subset {100, 101, 110, 111}, etc. Of course, schemata with no *'s describe sets of 1 element — i.e. the notion of schemata subsumes individual chromosomes.

The total number of possible schemata given a chromosome length of l is 3^l since there are three possibilities at each position: 1, 0, or *. The chromosomes, whose values are set, are instantiations of 2^l schemata since each position may take its actual value or the wildcard. To see this, let's examine an example of a short chromosome length, $l = 3$, for which there are $3^3 = 27$ possible schemata. The chromosome 101 is an instantiation of $2^3 = 8$ of those 27 schemata: {***, 1**, *0*, **1, 10*, 1*1, *01, 101}.

The important insight is that a single chromosome, in fact, also represents a great number more schemata — i.e. *categories* — and thus in some sense, by judging the fitness of that individual, the fitness of each category is *also* judged. By the same token, the fitness of an individual chromosome is also in a sense a function of the fitnesses of each schema it represents. The fitness of a schema is defined as the average of the fitnesses of all instantiations of that schema in the population. Although this figure is never explicitly calculated in a GA it is implicitly calculated because individual chromosomes are members of a population. It can therefore be seen that in the process of selection, not only are relatively fit individuals selected and mated, but also relatively fit schemata.

Schemata, themselves, can be thought of as instantiations of other schemata. For example, *1**1* is as much an instantiation of ***1* and *1*** as the chromosome 010010. Thus, ***1* and *1*** can be regarded as “building blocks” of *1**1*. Let’s say that *1**1* is a solution to a given problem and any chromosome instantiating that schema would be evaluated as 100% fit. Chromosomes that are instantiations of *1*** and ***1* (but not *1**1*) and thus contain building blocks of the target schema would be evaluated with relatively high fitnesses, 50% say, increasing the likelihood they will be mated together. This in turn increases the likelihood that their building blocks will be combined to evolve the target solution.

Taken in a slightly different light, schemata can be thought of as hypotheses. The string 101 makes 8 implicit “hypotheses” about the solution to the problem; that a 1 in the first position is important to the solution; that a zero in the second position is important to the solutions; that a 1 in the first position *and* a zero in the second position is important to the solution; etc. The merit of those hypotheses is rated by the fitness function and by comparing the similarities between the chromosomes that have the highest ratings, the hypotheses with the greatest merit eventually emerge.

Implicit parallelism — this power to judge many categories by judging a single member — is the primary power of the GA. One chromosome implicitly represents a number of schemata, and a single evaluation of that chromosome implicitly evaluates all the associated schemata. In the process of crossover and mutation, relatively short schemata are not disrupted and are allowed to propagate through the population from generation to generation, guiding the search through solution space.

In summary, the primary ideas that we can extrapolate from Schema Theorem as it is applied to genetic algorithms and that are relevant to the thesis of this paper appear in (1).

(1) a. A single representation (ex. a chromosome) implicitly contains a huge amount of information about the categories to which it belongs.
b. Processes that act on those representations (ex. selection) can implicitly make use of all that information in parallel.
c. Valuable information (ex. a solution) can emerge from the repetition of the same process.

In the following sections, I propose that these principles can be extended to language learning.

3. GAs ➡ Language Acquisition

The task of learning language seems monumental; the amount of information that children must process is staggering and considering the time in which they acquire language, one must feel a sense of awe. Current generative linguistic theory holds that there exists an innate "Universal Grammar" and that language acquisition is a process of setting the values of innate parameters that may vary only in restricted ways. A large motivation behind this claim is the sheer amount of information children must assimilate in a short time to become competent speakers. By proposing that their options are limited by innate linguistic knowledge, the burden is greatly eased.

However, increasingly, work in psychology, cognitive science, and artificial intelligence is progressing towards a dynamical systems account of development and cognition (Thelen and Smith, 1994; Port and van Gelder, 1995; etc.) — an approach that argues against innateness. For the most part, linguists have ignored this trend and fail to see the benefits it has to offer.

Furthermore generativists, Miller and Chomsky (1963) in particular, argue that statistically based algorithms such as would be required by a dynamical systems account of language acquisition cannot account for the productive nature of language — i.e. that we can produce sentences we have never heard before. This is because (they claim) the amount of input data required by a statistical approach would greatly outweigh that exhibited in reality. However, several researchers have demonstrated that this is not the case. Chalmers (1990), for example, showed that connectionist networks could represent syntactic structures and learn syntactic transformations. Another researcher, Elman (1995), demonstrated that his SRN system could accurately predict subject and verb number agreement despite intervening clauses of various types. Evidence such as this very strongly suggests that the underlying assumptions upon which the theory of generative grammar is built are suspect.

In this vein, I propose to show that the principles of Schema Theorem as illustrated above can also be applied to language learning so as to "lessen the burden" placed on the language learner in a way that is consistent with a dynamical, statistical system that lacks innate structures.

3.1. Chromosomes ➜ Memories

Esther Thelen (1994: 33-34) in her arguments against innateness comments on the work of Newport (Johnson and Newport, 1989; Newport, 1990) and her studies in language acquisition.

> "Newport speculates that young children learn deep syntactic properties more readily than adults precisely because young children are cognitively 'deficient.' Newport suggests that when mature persons with all their cognitive resources try to learn a language, they attend to and remember all that they hear and the full range of meanings in context. Very young children are, however, cognitively deficient. They cannot hear, or remember, or think about it all. They only pick up bits and pieces of language."

It is an important insight that children are "cognitively deficient"; however, Newport's claim that children "cannot hear, or remember, or think about it all — they only pick up bits and pieces of language" requires further qualification. It is clearly not the case that infants have primitive perception and memory — in fact their senses are quite sophisticated and are capable of receiving and "storing" the same range of input that adults do. Thelen's own work points this out; in discussing Rovee-Collier's (1991) experimental results of a task in which babies learn that kicking moves a mobile that is tethered to their leg, she states:

> "Over time, [the memory of the task] faded, although simply seeing the mobile would reactivate it. Most important is that this action memory was highly specific to the training situation. If Rovee-Collier changed the mobile, or even the designs on the pads that lined the cribs in which infants originally learned the task, infants forgot that kicking a lot makes the mobile move more. The action memory was highly tied to the learning context."
>
> Thelen (1995: 96)

It's clear from these results that even infants have quite detailed sensory memories — even the patterns on the crib liners could affect the learning of a task. We can conclude then that "cognitively deficient" does not mean that infants can not perceive and process complex stimuli in their environment; it is that they cannot perceive and process the salient, discrete, and symbolic aspects of their environment. In the task above, the infant did not comprehend that the crib liners were not a salient aspect of the task — they remained unchanging and thus are as likely a factor as any other stimuli until experience demonstrates otherwise. Indeed, Rovee-Collier went on to show that if the child was trained on the task with several different pads, the same memory effects were not apparent (Thelen, 1995: 96).

What we can extrapolate from Rovee-Collier's results is that infants, in learning a novel task, "record" the experience in its totality — all visual, auditory, tactile, olfactory information, etc. Being "cognitively deficient" they have no understanding of what information in that experience is salient; it is all relatively new, and thus an unparsable, meaningless blur. Adults on the other hand, readily extract the salient features of a scene, utterance — any sensory input — and only attend that which is particularly important. All other information that is irrelevant is immediately recognized as such and discarded. This skill of determining salient and non-salient information is precisely what the infants are in the process of learning.

In the discussion of language acquisition presented here, it is important to define a few simple conventions so that it is possible to talk about phenomena that are very difficult to conceive of conceptually. Thus, I refer to the representation / storing / encoding of an experience as a "memory" without speculating on the precise nature of those representations. Let's also, for the sake of simplicity, assume that "experiences" and "memories" can be divided into discrete units, although this is not likely true.

Within the framework of Schema Theorem, I propose that "memories" are to language acquisition, what chromosomes are to GAs. Both encode information in some way and both are instantiations of more general categories — schemata. To see how this latter claim applies to memories, recall that schemata are implicitly represented in chromosomes because 1) chromosomes are members of a *population* and 2) a process (selection based on the fitness function) acts upon the *similarities* between individual chromosomes. Memories exist in a population since it is possible to experience the same (or very similar) sensory input, temporally displaced. That is, experiences can be repeated. A simple example would be drinking water; each act of drinking water (yesterday, this morning, this afternoon) creates a "memory" and each memory is very similar with respect to sensory input. The process that acts on that similarity, I argue, is learning.

3.2. Selection ➨ Learning

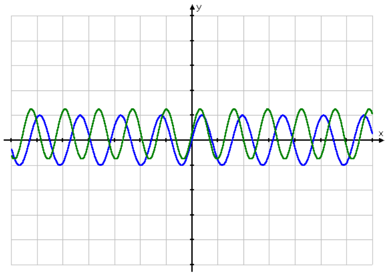
"Learning" is a rather broad concept that can be defined on many levels. "Learning" here, is to be understood on a neuronal level and I will take a simple characterization made by Rumelhart (1997: 213-214) for learning in connectionist systems:

> "Changing the processing of knowledge structure in a connectionist system involves modifying the patterns of interconnectivity. In principle this can involve three kinds of modification:
>
> (1) development of new connections;
> (2) loss of existing connections;
> (3) modification of strengths of connections that already exist.
>
> . . .
>
> "Virtually all learning rules for models of this type can be considered variants of the *Hebbian* learning rule, . . . if a unit u_i receives input from another unit u_j t a time when both units are highly active, then the weight w_{ij} to u_i from u_j should be *strengthened.*"

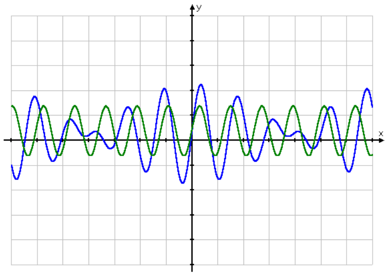
Since this definition is originally based on real brains (Hebb, 1949), it is at least analogous to learning that occurs in real neurons — that two highly active, connected cells develop a stronger excitatory relationship and thus repeated activations of the same neurons and groups of neurons result in gradually increasing strengths of activations over time. Additionally, associations between different areas of the brain can be learned through "convergence zones" — a neuronal grouping that directs the stimulation of anatomically distant regions (Damasio and Damasio, 1994).

Although unique memories are stored (by high-level convergence-zones) "cumulative memories" are also generalized (by low-level convergence zones). It is these "cumulative memories" that we are most concerned with. Memories can be though of as "cumulative" since each affects the activations and connections of the neurons across which they are distributed. They store experiences that, sharing similar sensory input, activate similar neurons (or similar groups of neurons) to similar degrees.

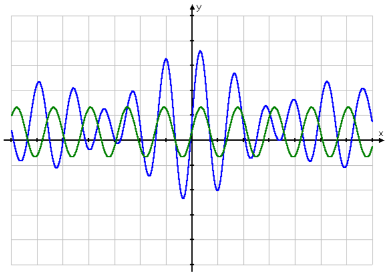
However, similar experiences are not *identical*. Areas of experiences that are most similar will cause the highest increase of activation — the highest degree of learning. To express this idea in a visual medium, let's pretend that it is possible to represent a sensory experience / memory — all its visual, auditory, tactile information etc — as a single waveform. Examine the graphs in (2) through (5).



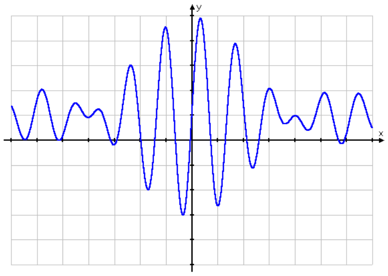
(2).



(3).



(4).



(5).

Each of the first three graphs has two wave-forms: the blue one corresponds to the collective learning of past memories, and the green one corresponds to a new, similar experience. Adding the waveforms together to create more complex waveforms is analogous to learning. If peaks and troughs correspond to strength of activation, by (5), we already see that certain areas are more highly activated than others. These areas correspond to the areas of the constituent waveforms (i.e. memories) that were "in phase" (i.e. most similar). Thus high areas of activation develop where experiences are most alike. This process is essentially the same as the development of feature detectors in connectionist networks described by Rumelhart and Zipser (1985) only on a much larger scale, and is reminiscent of the story told by Thelen (1995: 96-98).

Learning then, is a process that takes advantage of the similarities between individual memories. Although learning and selection are not comprable as processes — they are dramatically different both in nature and output — with regards to Schema Theorem, they play the same role. That is they are both processes that can take advantage of the implicit information about schemata that is present in individual representations.

It is still not clear what the nature of "language acquisition schemata" are; it is necessary to examine the final Schema Theorem principle defined in (1). This is the topic of the following section.

3.3. Solution ➧ Salience

One of the major tenets of generative theory is the presence of "features". And rightly so; there is a great deal of evidence that they exist. However, Chomsky and his followers claim that those features are *innate* — a claim that may be unwarranted.

Rumelhart and Zipser (1985) have already demonstrated that competitive, unsupervised learning in neural network models can develop "feature detectors". Although their studies were not on language in particular, there is no reason in principle why their results cannot be applied to linguistic features since they too must exist in neural tissue on some level. The implication is that discrete features need not be initially present in a neural system; they can emerge based solely on input.

In section 3.2. we saw how the process of learning created higher activations in areas where memories were most similar (cf. the graph in (5)). This basically amounts to learning the relevant characteristics of a given set of memories. A real-world example is useful at this point to illustrate this process; consider Rovee-Collier's experiments that were described in section 3.1.

When the infants experienced the visual stimuli of the mobile moving, at the same time, they experienced the motor sensation of kicking and the visual stimuli of the pattern on the crib liners.

The repetition of the experience over time, although continuous, can be thought of as multiple experiences, each creating memories that are used by the process of learning. This being a relatively new experience and the infant being cognitively immature, the infant cannot discard the crib liner pattern as being irrelevant. It is stable over the course of learning and thus remains a "peak"of activation — a salient characteristic of the experience. Thus, in learning the task, the infant learns associations between the leg kicks, the motion of the mobile, *and* the crib liner patterns. However, if the liners are changed either during training or later when the infant re-learns the task, it becomes clear that the crib liners are *not* salient characteristics of the experience. Presumably, it would be possible to witness the same phenomenon using a *any* stable sensory input — a tone, perhaps, or even a strong scent.

Thus, whereas selection acting upon chromosomes produces a solution, I would argue that learning acting upon memories produces salience in the form of defining features (in the informal sense of the word). Returning to the question of what schemata *are* in this system, we can use this example to try and define them.

A memory relatively unbiased by learning instantiates many hypotheses (schemata) about what features are salient and are thus to be associated. In learning the kicking task, for example, there are hypotheses about associations between the crib liner patterns and the movement of the mobile, the appearance of the mobile and the movement of the mobile, the leg-kicking and the appearance of the mobile, the leg-kicking and the pattern of the crib liners — every possible combination of sensory input. However, as learning progresses, only the most persistent, most stable stimuli will emerge as salient.

With respect to the principles in (1), this learning process directly parallels genetic algorithms; (1a.) representations — memories — are acted upon repeatedly by (1b.) a process — learning — producing (1c.) valuable results — salient information.

3.4. Applications to Phonetics and Phonology

Finally turning to the applications of Schema Theorem to language learning, consider the acquisition of phonological features (±voice, ±nasal, ±continuant, ±anterior, etc.). These features are considered to be "a given" in current linguistic thought — part of Universal Grammar. The child need only learn which are relevant to the language they are learning to speak.

Extrapolating from the arguments above, each utterance the child hears is encoded *in its entirety* as a memory. They remain, initially, completely unparasable and therefore unintelligible and unproducible. However, as being instantiations of schemata, each memory represents a multitude of hypotheses about salient features and associations. Note that "feature" here does not refer to {±voice, ±nasal, ±continuant, ±anterior, etc.}; it instead refers to *any* aspects of the waveform — specific frequencies, harmonics, timing, contours, voicing, etc. "Phonological features" in the formal sense of the word will naturally emerge as predicted by Rumelhart and Zipser's (1985) model and as seen above.

Furthermore, this process supports more than just the learning of phonological features; sound-sound and sound-position hypotheses are also being tested resulting in phonotactic constraints; sound-feature hypotheses are being tested resulting in phonological processes like voicing assimilation; feature-position hypotheses are being tested resulting in phonological processes like word-final devoicing; etc.

The power that we gain from Schema Theorem is that all these features and processes are being discovered *simultaneously, implicitly,* and in *parallel* in the process of learning. Every memory contributes to the learning of every feature / association / schemata that it instantiates. Given this idea, it is much easier to see how a limited amount of input can lead to learning extremely complicated, *productive* rules of language use.

3.5. Learning Reference and Beyond

Having seen the general example of Rovee-Collier's experiments with infants, it is not difficult to see how learning word reference is a subset of a broader category of learned associations. Let's examine a concrete example: think of a toddler that by now is considerably more cognitively sophisticated than an infant. They can efficiently "chop up" the world visually into objects, know that objects persist beyond direct sensory perception, and so on. They have also begun to "chop up" utterances into discrete units, and have begun to learn a few words.

Picture the toddler in a kitchen drinking milk — the recording of that experience as a memory includes the locale, the appearance, taste, texture of the milk and the cup, the sensation of swallowing, and perhaps, a care-giver saying the word "milk". The child does not understand the word, but perhaps can mimic it. That memory instantiates schemata in which the word "milk" is associated with the liquid, the taste of the liquid, the color of the liquid, the cup, the liquid *and* the cup, the sensation of swallowing, and so on. However, as the child experiences "milk" in more situations, the process described in the above sections takes effect. More memories are encoded, learning occurs, and gradually the correct hypotheses emerge.

It is important to note that within this framework, what the child is acquiring is *not* a direct index to an entity in the real-world, but instead an association to a collection of learned (and thereby distilled) memories. One of the predictions that this makes is that weaker associations will also be learned. For example, if the child usually drinks milk in the kitchen the word “milk” will also cause weak activation of memories of kitchens. Similarly, the child learns weak associations to milk-cartons, cups, cows and other objects and concepts related to “milk” on a scale of relative salience. Visually, weaker associations correspond to the medium-sized peaks and troughs in the graph in (5). It would seem these predictions are indeed psychologically attested.

Gradually, as the child builds a strong association between the word and collection of memories, convergence zones (Damasio and Damasio, 1984) begin to be built allowing for the stimulation of the relevant areas of the brain in the *absence* of all stimuli — i.e. when the child experiences milk, the memory of the word “milk” becomes activated without actually hearing it and vice versa.

Now suppose that the child also learns the word “water”. Color and taste easily distinguish them from each other, yet at the same time, they share a great deal of similarity in the other respects recorded in the memories of each — texture, movement, swallowing sensations, cups, kitchens, etc. Thus, a schema that is instantiated by both water and milk is reinforced simultaneously but separately (by developing differing convergence zones) from the schemata that are solely instantiated by the individual entities. It is the beginning of a category of memories that include water and milk, and will also include coke, juice, and beer when they are eventually learned. When the child hears the phrases “drink milk” and “drink water”, it is to that category that “drink” is associated, *as well as* the specific words. Consequently, the child can infer that newly learned words that are also instantiations of that category, “coke” for example, can also be appropriately used with “drink”. Thus we see the emergence of the productivity.

This same learned association helps bootstrap language acquisition in other ways as well. The more times “water”, “milk” etc. are used in similar contexts, the stronger evidence there is for the category that they are contained in. In a way, this is priming the category to be labeled facilitating the learning of the noun “drink” when the child is finally exposed to it. Additionally, when the child hears a new word in the absence of the real-world entity, say, “drink whiskey”, they can immediately infer properties of “whiskey” even though they have not experienced it directly.

Again, it is important to emphasize that all of these hypotheses / associations / schemata are being learned *implicitly in parallel*. Each time the child hears “drink milk”, not only are word references being learned, but also their phonological features are being refined and word categories like Verb and Noun are being learned. So, too, are intonation and discourse patterns, individual speaker characteristics — all are implicitly represented in the schemata instantiated in the memories of not just the word, but also the entire context within which the word was *experienced*.

4. Conclusion

This paper has proposed that the principles of Schema Theorem as defined by Holland (1975) to explain the efficiency of genetic algorithms can be distilled to those presented in (6) and generalized to apply to other phenomenon, particularly language acquisition.

(6) a. A single representation (ex. a chromosome) implicitly contains a huge amount of information about the categories to which it belongs.
b. Processes that act on those representations (ex. selection) can implicitly make use of all that information in parallel.
c. Valuable information (ex. a solution) can emerge from the repetition of the same process.

To illustrate that the principles in (6) can be generalized, an analogy was drawn between GAs and language acquisition; a summary of the relevant mappings with respect to Schema Theorem is presented in (7).

(7)

Schema Theorem		**GAs**		**Acquisition**
representation	➡	chromosomes	➡	memories
process	➡	selection	➡	learning
product	➡	solution	➡	salience

It has been demonstrated that a generalized Schema Theorem is at least feasible and should be explored in greater detail.

This approach makes many predictions about the nature of language learning. For example, it predicts: 1) a gross order of acquisition based on what is perceptively most salient; (phonology precedes syntax; concrete precedes abstract); 2) associations can be made between features on *any* level and other features on *any* level (intonation patterns and semantic readings; voicing and social register); and 3) second language learning is more difficult because salient information that is already learned interferes with learning new salient information. Further research could be carried out any number of directions, for example exploring the above predictions in clinical tests or cognitive models, or examining the applications of Schema Theorem of other domains aside from GAs and cognition.

The primary benefit of this hypothesis is that it provides a key to understanding how, contrary to the claims of generative linguists, statistical, dynamical models of language and cognition can, in fact, give rise to the complexity and productivity of human language based on the limited input that children receive. This lends strong support to the argument against innateness of linguistic knowledge, and by extension, innateness of cognition itself.

REFERENCES:

Chalmers, David. (1990). Syntactic Transformations on Distributed Representations. *Connection Science,* Vol. 2, Nos 1 & 2, 53-62.

Damasio, A. and H. Damasio. (1994). Cortical systems for retrieval of concrete knowledge: The convergence zone framework. In: Koch, C. and L. Davis (Eds.). *Large-Scale Neuronal Theories of the Brain*. Cambridge, MA: MIT Press.

Elman, Jeffery. (1995). Language as a dynamical system. In: Port et al.

Goldberg, David. (1989). *Genetic Algorithms in Search, Optimization, and Machine Learning*. New York, NY: Addison-Wesley Publishing Company, Inc.

Hebb, Donald. (1949). *The Organization of Behavior*. New York, NY: Wiley.

Holland, John. (1975). *Adaption in Natural and Artificial Systems.* Ann Arbor, MI: University of Michigan Press.

Johnson, J. and E. L. Newport. (1989). Critical period effects in second language learning: The influence of mturational state on the acquisition of English as a second language. *Cognitive Psychology, 21*, 60-99.

Miller, G.A. and Noam Chomsky. (1963). Finitary models of language users. In: R.D. Luce, R.R. Bush, and E. Galanter (Eds.), *Handbook of Mathematical Psychology*, Vol. 2. New York, NY: Wiley.

Mitchell, Melanie. (1996). *An Introduction to Genetic Algorithms.* Cambridge, MA: MIT Press.

Newport, E. L. (1990). Maturational constraints on language learning. *Cognitive Science, 14*, 11-29.

Port, Robert and Timothy van Gelder. (1995). *Mind as Motion: Explorations in the Dynamics of Cognition*. Cambridge, MA: MIT Press.

Rovee-Collier, C. (1991). The “memory system” of prelinguistic infants. *Annals of the New York Academy of Sciences, 608,* 517-536.

Rumelhart, David. (1997). The architecture of mind: A connectionist approach. In: Haugeland, John (Ed.) *Mind Design II.* Cambridge, MA: MIT Press.

Rumelhart, David and D. Zipser. (1985). Feature discovery by competitive learning. In: Rumelhart, D. and J. McClelland (Eds.). *Parallel Distributed Processes: Explorations in the Microstructure of Cognition, Vol 1.* Cambridge, MA: MIT Press.

Thelen, Ester and Linda Smith. (1994). *A Dynamic Systems Approach to the Development of Cognition and Action*. Cambridge, MA: MIT Press.

Thelen, Ester. (1995). Time-scale dynamics and the development of an embodied cognition. In: Port et al.