A Dynamical Connectionist Account of Conceptual Change

Athanassios Raftopoulos (raftop@ucy.ac.cy),
Andreas Demetriou (ademet@ucy.ac.cy)
Department of Educational Sciences, University of Cyprus
P.O. Box 20537, 1678 Nicosia, Cyprus.

Abstract

Conceptual change can be accounted for at various levels of explanation. The cognitive level (Marr's computational level), the representational (Marr's "algorithmic"), and the implementational level. In this paper, we offer a dynamical account of types of conceptual change at the representational level. Our aim is to show that some classes of neural models can implement the types of change that we have proposed elsewhere. First we briefly describe at the cognitive level certain types of change that purport to account for some of the kinds of conceptual change. Then we lay forth the framework of dynamical connectionism; we discuss the representational level realizations of the cognitive level and claim that these can be depicted as points in the system's activational landscape. We offer, third, a dynamical account of some types change and we claim that conceptual change can be modeled as a process of modification, appearance of new and disappearance of attractors and/or basins of attraction that shape the system's landscape. Finally, we discuss the kinds of mechanisms at the representational level that could produce the types of change observed at the cognitive level, as modeled by means of dynamic connectionism.

Introduction

Conceptual change can be accounted for at various levels of explanation. Following Marr (1980), one can distinguish between three levels: the *computational,* the *algorithmic*, and the *implementational* level of explanation of cognitive systems. We prefer the term "cognitive" to "computational", and the term "representational" to "algorithmic", since there are accounts of cognition that deny the algorithmic nature of mental operations.

At the cognitive level, one can discuss cognitive operations that apply to information-processing content (such as addition and subtraction), operations that apply to structures as wholes, such as differentiation or coalescence (Carey, 1985; Chi, 1992), or, conceptual combination, generalization and abduction, and hypothesis formation (Thagard, 1992). This level addresses the issue of the functions computed by the information processing system.

At the representational level one can examine the algorithmic processes that realize conceptual change at the cognitive level by transforming representations, such as Newell and Simon's (1972) "problem behavior graph" in production systems. In the connectionist paradigm one can study the processes of the emergence of new attractors, and repositioning of points realizing representational states in high-dimensional state spaces (Horgan and Tienson, 1996), or the changes in the connection weights and network structure (Elman et al., 1996; Schultz et. al., 1995; Plunkett & Sinha, 1992).

In this paper, we will discuss a theory of different types of cognitive change and their implementation at the representation level. Our aim is to show how certain classes of neural networks could implement some of the types of change that the authors have proposed (Demetriou and Raftopoulos, 1999). First, we will summarize these types of change. In the second part we will sketch the framework for the dynamics of change, relying on the dynamical interpretation of connectionist networks to explore possible means of modeling the stipulated types. In the third part we offer a dynamical account of some types change and we claim that conceptual change can be modeled as a process of modification, appearance of new and disappearance of attractors and/or basins of attraction that shape the system's landscape. Finally, we discuss the kinds of mechanisms at the representational level that could produce the types of cognitive change.

To that end we will employ neural networks whose behavior can be viewed as falling under one or the other of our kinds of change, and describe the behavior that neural networks should exhibit if they are to implement type of change.

Types of Change

Demetriou and Raftopoulos (1999) previously published a theory of conceptual change that addresses the issue of how a learning system makes the transition from one state to another. The theory provides a detailed analysis of the types of change that are observed both in cognitive development and during learning. The types of change are summarized in Figure 1.

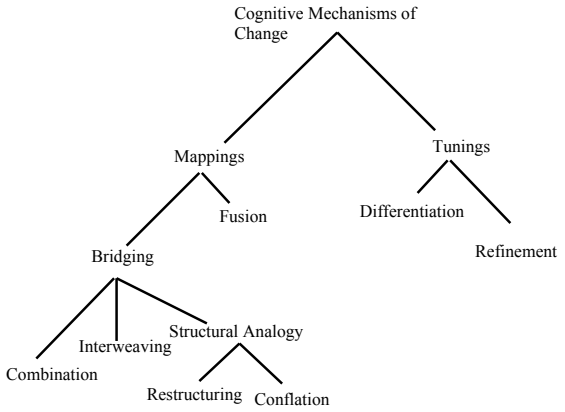

Fig. 1. The types of change

We will briefly present here *combination*, and *fusion. Bridging* is a class of types of change, whose unifying feature is that (a) two or more existing structures are brought together to bear on each other and form a more complex structure, and (b) after bridging the constituent structures retain their functional autonomy, even though they may have been modified. The blended structures may remain unaltered and the resulting structure(s) may retain the characteristics of the constituent structures (as when "striped" and "apple" are combined to produce "striped apple". In this case, the type of change is *combination*.

Fusion differs from bridging in that the mapped structures do not retain their relative autonomy after the mapping; instead, they fuse to one of the existing structures, or form a new structure. An example would be the fusion of retrieval and counting strategies, which are involved in simple operations of addition and subtraction performed by children aged 4-6. After fusion, around the age 6-7, the predominant strategy is retrieval by rote memory (Siegler, 1996).

The Representational Level

We discuss here the way representations can be modeled as properties of cognitive systems. At this level one examines the mathematical implementation of the cognitive level. In other words, we examine the way cognitive states are represented and how they are transformed and processed by means of operations performed on data structures.

These transformations can be either algorithmic (determined by a set of rules that apply to discrete static symbols that are the representations of the system) or dynamical (determined by mathematical relations that apply to continuous variables and specify their interrelations and evolution in time). This is why we call these transformative processes mathematical-state transitions; they describe the way the system moves between points in its state-space. We will address the issue of change from the perspective of connectionist theory interpreted in a dynamical way. Thus, will assume that a cognitive system is associated with a dynamical system physically realized by a neural network.

Neural Networks as Dynamical Systems

Recurrent neural networks (Elman, 1990) with distributed representations and continuous activation levels can naturally be construed in a dynamical way. They can be described by means of the evolution of the activation values of their units over time. To be able to model growth and avoid problems of lifelong (mainly catastrophic interference), one needs to consider a special class of networks, namely adaptive or generative networks. These networks can modify their structure during learning by adding or deleting nodes and can change their learning rates.

The number of units of the network determines the number of dimensions of the state-space associated with the system. Their activation values constitute the actual position in the state-space of the system. Adding a time-dependent parameter yields the phase-space of the system. Both in state- and phase-space, one can represent all the possible states that a system can take in time. Hence, in the connectionist account, the states of a cognitive system are depicted by the sets of activation values of the units that distributively encode these states.

These activation values are the variables of the dynamical system and their temporal variation constitutes the internal dynamics of the system. In addition to the state-space of a system, an external control space is also defined. The external space contains the real-value control parameters that control the behavior of the system, i.e., the connection weights, biases, thresholds, and, in networks in whose structural properties are implemented as real-value parameters (Raijmakers et. al., 1996), the structure of the system. In dynamical systems the fast internal dynamics is often accompanied by a slow external dynamics. The external dynamics consist of the temporal paths in the external control space. The external dynamics consist of the network's learning dynamics (the various learning rules) and the dynamics that determine structural changes, such as the rules for inserting nodes in cascade correlation and growing radial basis function networks.

When the network receives input, activation spreads from the input units to the rest of the network. Each pattern of activation values defines a vector or a point, within the activational space of the system whose coordinates are the activation values of the pattern. The activation rules determine the state transitions that specify the internal dynamics of the system, i.e., the functions of the evolution of the system in time. Thus, the behavior of such a system is depicted as a trajectory between points in the activational state space.

The activation rules, the number of units, the pattern of their connectivity, and the learning rate(s) of the network determine the architecture of the system. These factors are determined by its long-term history of experiences, since the class of networks discussed here may modify either their patterns of connectivity, as they learn, by adding nodes, deleting nodes, and sharpening their connections, or their biases and learning rates. The activation vectors and the behavior of the system evolve as a result of the synergies among the architecture of the network, the input it receives, and the previous activity of the network, under the control of the external dynamics.

The behavior of the system is a collective effect of cooperation and competition, (Kelso, 1995). The competition is due to the effort of the system to retain its current state in the face of incoming information. If this information cannot be assimilated by the system, then the weights of a network change and the network may alter its structure to accommodate the new input.

The activation states, in which a network may settle into after it is provided with an input signal, are the attractors of the system. These are the regions in state-space toward which the system evolves in time. The points in state-space from which the system evolves toward a certain attractor lie within the basin of attraction of this particular attractor. Thus, the inputs that land within the basin of attraction of an attractor will be transformed by the connectivity of the network so that they end up at this attractor where the system will settle.

Networks in which the outputs change over time until the pattern of activation of the system settles into one of several states, depending upon the input, are called attractor networks. The sets of possible states into which the system can settle are the attractors. If the network is used to model cognitive behavior, then the attractors can be construed as realizing cognitive states to which the system moves from other cognitive states that lie within the attractor's basin of attraction.

The signal of the input is transformed as it moves through the hidden units into an attractor pattern as follows: a given input moves the system into an initial state realized by an initial point. This input feeds the system with an activation that spreads causing the units of the system to change their states. The processing may take several steps, as the signal is recycled through the recurrent connections in the network. Since any pattern of activity of the units corresponds to a point in activation space, these changes correspond to a movement of the initial point in this state space. When the network settles, this point arrives at the attractor that lies at the bottom of the basin in which the initial point had landed. In this sense, the inputs fed into the system are the initial conditions of the dynamic system. Similarly to a dynamical system that settles into a mode depending on its initial conditions, a neural network settles into the attractor state in whose basin of attraction the input falls.

For instance, in a semantic network meanings of words are represented as patterns of activity over a large number of semantic features. However, only some of the combinations of semantic features are features of objects. The patterns that correspond to these combinations are the attractors of the network, which are points in the state space corresponding to the semantic features of the prototype of the object signified by the word. These attractors are the meanings of words.

The concepts "attractor" and "basin of attraction" suggest a way of simulating the classical notion of symbol. The attractor basins that emerge as the network interacts with specific inputs might be construed to have symbolic-like properties, in that inputs with small variations that fall within the same attractor basin are pulled toward the same attractor (or cognitive state) of the system. Thus, various inputs (tokens) give rise to the same stable point of attraction, the attractor (type), which in this sense offers a dynamical analog of the classical symbol (Elman, 1995).

The dynamical "symbols", unlike the symbols of classical cognitivism, are dynamic and fluid rather than static and context independent. The dynamic properties result from the dynamical nature of the activations of associative patterns of units. As the network learns and develops, the connection strengths continuously change. The same happens when new units emerge and old units "die" and the system reconfigures to maintain its knowledge and skills. All these cause changes in the original pattern in which an attractor/symbol was created in the first place, and as a result, subsequent activations differ. The same effects are caused by the different contexts in which the "symbol" may be activated. This happens because connections from the differing contextual features bias the activation of the units of the original pattern in different ways emphasizing some feature of the pattern or other. Thus, the attractor/symbol is almost never instantiated with the same activation values of the units that realize it.

The activational state-space of a network is a high-dimensional mathematical landscape. The state transitions in such a system are trajectories from one point on to another. Attractors correspond to cognitive states and the activation pattern that realizes each state is a vector, or a point. Thus, cognitive states are realized by points on this landscape. Since the distributed encoding of a cognitive state does not involve all units of the system, there will be points on the activational landscape that will realize more than one cognitive states (the set of coordinates of a point may satisfy the partial coordinates given by several activational vectors).

During the phase of activation-value changes the system passes through various possible outputs. All these outputs can be viewed as lying on an energy surface. When the system passes through a certain output-state whose energy is not lower than the energies of the neighboring states, it goes through another phase of activation-value changes in order to reduce the energy of the output state. When it reaches a point at which all the neighboring states have higher energies, it settles.

These states of minimum local energy are the attractors and can be construed as valley bottoms on energy surfaces. Thus, attractors should be distinguished from the networks' outputs in general. Not all outputs are settling points. Attractors form a subset of the set of outputs of a network, in that they are those outputs at which the system can settle. When the input of the system is such that the activation state of the system lies within the walls of the valley, the system will settle at the attractor at its bottom. Hence, the valley is the basin of attraction that leads to the specific attractor-state of minimum energy. Since the network has many attractors and basins of attraction, their relative position shapes the relief of the activational landscape of the system.

Modeling the Dynamics of Cognitive Change

In this theoretical framework, cognitive change results from the molding of the activational landscape, as a result of changes in the weights and the architecture of the network,

as the network attempts to accommodate new input signals. The molding may result either in the emergence of new, and/or disappearance of old, attractors, or in the reshaping of the basins of attraction. This process corresponds to a trajectory on the activational landscape. The idea that change is to be modeled by means of transitions in the state space of a dynamic system is at the heart of dynamical theories of cognition. Transitions in the state space of a dynamic system substitute for the algorithmic syntactically governed transitions of cognitivism.

The relief of the landscape determines the trajectories that are allowed, and the possible transformations among cognitive states. Cognitive change, thus, depends on the activational landscape of the system that learns. When information enters, the system tends to assimilate it within the existing framework of knowledge, which, in neural networks, is determined by the connection weights and the architecture of the network, which, in their turn, distribute the points that realize cognitive states on the network's landscape. We have posited certain types of cognitive change. In what follows we will sketch their dynamic realization at the representational level.

Combination

This type of change involves the combination of structures in such a way that the existing attractors and the landscape's relief (their basins of attraction) of the system are not affected. The new structure is superimposed, as it were, on the constituent structures. Consider the networks that simulate learning to pronounce words and non-words (Plaut et. al. 1996). These networks learn the pronunciation of both regular and irregular words, by building the appropriate attractors. The attractors of regular words consist of componential attractors, in which case the basin of attraction is the intersection of the sub-basins of attraction of the componential attractors. The exception words have their own attractors with a lesser degree of componentiality. Combination explains the ability of the network to learn the pronunciation of words and non-words, in that this knowledge is the result of the combination of the sub-knowledge encoded by the componential attractors, as is shown in Figure 2.

In this figure only two componential attractors are depicted, for onset and the vowel in the reduced two-dimensional activation space of the phonemic units of the network. The basins of attraction for the word "by" and the non-word "dy" are the intersections of the sub-basins for pronunciation of *b*, *d*, and *y*, that is, the regions in the state space in which these sub-basins overlap. The black circle is the attractor for the word *by*, and the striped circle is the attractor for the non-word *dy*. The trained network learns to pronounce words by applying its knowledge regarding the pronunciation of the parts of the words (and of the role of context in pronunciation when it comes to exception words). The reduced componentiality of the exception words is depicted by means of a deformation of the intersection of the salient attractors for the onset *d* and the vowel *o*. The componential attractors and their basins of attraction remain unaltered.

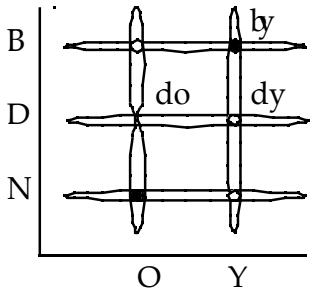

Figure 2. Componential attractors

After the new pronunciation is learned, the two basins change their relative positions so that they intersect. Their intersection (i.e., the pattern corresponding to both sets of features) forms a new basin, which is the area in which the two basins overlap. The appearance of a new basin of attraction represents the learning of a new concept. The new basin of attraction is superimposed onto the two intersecting basins. The basins of attraction (sub-basins) and the attractors do not change. Whatever input was falling within one of the two basins before learning, still does so after the network has learned the new concept. The only change after learning is that some inputs fall within both the new basin and the old basins of attraction, This is a result of the superimposition of the new basin of attraction onto the two sub-basins.

Fusion

Stable structures within the neural net can be thought of as attractor states. Thus, the activation pattern of the structure attracts all other activation patterns that are similar enough with it (that is, all activation patterns that fall within the basin of attraction of the attractor). As a network learns, a new attractor state may emerge, which swallows the attractors that existed before. This is what happens in fusion. The two initial basins of attraction are also swallowed by the new one, so that all patterns that were falling within the one or the other now fall within the new basin of attraction. The system undergoes a phase transition that can be described as a reverse Hopf bifurcation (Figure 3), in which two stable states (bistability) are fused and disappear, and one stable state emerges (unistability).

Figure 4 displays the phase transitions associated with the fusion of "counting from one" and "memory retrieval" strategies (used by 4-6 year old children in simple arithmetic tasks) to the "memory retrieval" strategy that becomes predominant between 6 and 7 years of age.

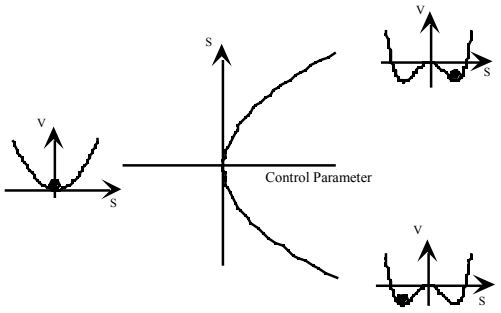

Figure 3. Fusion as an inverse Hopf bifurcation

The generative networks designed by Schultz et. al., (1995) to model a series of cognitive tasks simulate the variability of the strategies available to children. Networks at some stage of their training in the balance-beam tasks may "employ" two different strategies to solve the same problem and, as training continues, progress to using reliably the more advanced strategy. These networks implement "fusion", by moving from bistability to unistability.

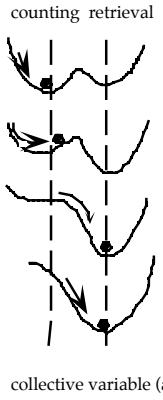

Figure 4. Fusion of two counting strategies

Strong and Weak Cognitive Change

In this context of dynamical connectionism, cognitive change consists in changes in connection weights, the structure, and the learning rates of the network. In connectionist networks an individual's state of knowledge is determined by the weights of the hidden units. Cognitive representational change is regarded as the individual's actual path through the space of possible synaptic configurations, that is, as the transformations of the weight vector in an n-dimensional weight space, where *n* is the number of the weights.

The appearance of novel cognitive states, and thus, the appearance of new attractors and the disappearance of old ones, implies a change of relief in the network's landscape (molding). Since the relief depends on the structure of the network, that is, the number of nodes, their connectivity, and the activation functions, its molding is the result of changes in the structure of the network. Networks evolve as a result of the system's effort to adapt to a new environment, by superimposing new representations to old ones. Thus, the system modifies the "knowing assumptions" that do not fit in.

The account of cognitive change at the representational level allows us to recast the discussion regarding strong and weak representational change in terms of dynamic systems theory. Whether a cognitive change is weak or strong depends on whether the new structure increases the representational resources of the system. Since representations are points in the state space of the system, the expressive contents of the system correspond to such points. If the relief of the landscape is such that the system cannot settle at a content realizing point, that is, if this point is not a possible attractor state, the content that is realized by this point is not within the expressive capabilities of the system. When changes in the relief render this point an attractor, the change is strong; it results in an increase in representational power.

But the mere appearance of an attractor does not necessarily imply that a radical change has taken place, that is, that this is a novel attractor state. This is so if the content realizing point that appears as a new attractor was in fact expressible within the system; that is, if the system could have settled at that point, even if it had not done so, up to that time. When the structures "striped" and "apple" are combined an attractor state "appears and the system acquires the new concept of "striped apple". This "new" attractor is a region in the state space, which realizes the content "striped apple" and is superimposed on the attractors of "apple" and "striped". But this is not a novel attractor, because this content was already within the expressive power of the system, since the relief of the landscape was such that the system could have settled if fed with the appropriate input at this point. In other words, the "new" attractor was situated at a local energy minimum in which the system could have settled if it had been fed with the appropriate input (the experience of a striped apple). The attractor appears without the landscape being molded and this attractor is just the sum of information expressed by the other attractors, which remain intact. In this case, the ensuing change is weak.

Weak change refers to changes in the semantic content of representations, which broaden their field of application but do not increase the expressive capabilities of the system. Attractors are merely repositioned in the landscape, which means that the activation patterns that define them do change. Reposition of any content-realizing point is accompanied by changes in the activation values that constitute the point's activation pattern, and changes in its spatial relations with other content realizing points. Since semantic information in dynamic systems is captured by the relative positions of content realizing points, repositioning is accompanied by semantic change.

This scenario does not apply to the case of fusion. No mere intersection of existing basins of attraction or any simple repositioning, could accommodate the salient input. A reshaping of attractor basins is required, as well as the

disappearance of an older attractor and the emergence of a novel one. These changes mould the landscape.

Thus, when new information is learned with repositioning of attractors and basins of attraction, and attractors are preserved (though the slope of the basins may change, with some becoming steeper and others becoming less steep), the resulting change is weak. Updating the connection weights seems to suffice for this. If the change in weights does not suffice for learning, the landscape is molded by changes in the network's structure (Horgan and Tienson, 1996). This may induce the appearance of new attractors; since the attractors are points on the landscape, the appearance of new cognitive states realizing points on this landscape, and the disappearance of old constitute strong changes, since the content-expressive power of the system increases. This process may require structural, i.e., qualitative, change.

Mechanisms of Change

At the cognitive level, the main Piagetian mechanisms driving conceptual changes are assimilation, accommodation, and equilibration. It is time now to consider the mechanisms driving change at the representational level. In each of the types of change discussed previously the processes that lead to the change are the same, always reducing to quantitative and qualitative changes in connection weights and the architectural structure of the network. These processes cause the repositioning of existing attractors, the disappearance of old ones, the appearance of new ones, and changes in the basins of attraction that shape the relief of the landscape. It could hardly be otherwise. In connectionism the computational mechanisms are domain general, statistical learning mechanisms, based on brain-style computation, that is, (a) on the spreading of the activation of each unit to other units, (b) on the modification of the connection weights, and (c) on the modification of the network structure.

McClelland (1989) argued that Piagetian "assimilation" corresponds to the activation spread in a network when a signal is presented to the input units and propagates through the network causing the activation of its units. The alteration of the weights, as a result of the network's learning, models Piaget's "accommodation", that is, the change that the network undergoes trying to fit in new experiences. Shultz, et al., (1995), and others, have proposed networks that adapt their structure as they learn by increasing their hidden units to accommodate the demands of the task. They offer a variation of McClelland's account that is suited better for networks that can modify their structure. The quantitative phase of error reduction and weight change may correspond to Piaget's "assimilation" of information in a pre-existing structure, whereas the qualitative structural change corresponds to Piaget's "accommodation" of the system. Quantitative change renders possible knowledge acquisition within a fixed representational framework, whereas qualitative change allows an increase in representational power.

References

Carey, S. (1985). *Conceptual change in childhood*. Cambridge, MA: The MIT Press.

Chi, M. T. H. (1992). Conceptual Change within and across Ontological Categories. In R. Giere (Ed.), *Cognitive models of science*. Minnesota University Press, 112-136.

Demetriou, A., & Raftopoulos, A. (1999). Modeling the developing mind: From structure to change. *Developmental Review, 19*, 319-368.

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

Elman, J. L. (1995). Language as a dynamic system. In R. F. Port & T. Van Gelder (Eds.), *Mind As motion: exploration in the dynamics of cognition.* Cambridge, MA: The MIT Press.

Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996). *Rethinking innateness: a connectionist perspective on development.* Cambridge, MA: The MIT Press.

Horgan, T., & Tienson, J. (1996). *Connectionism and the philosophy of psychology*. Cambridge, MA: The MIT Press.

Kelso, S. (1995). *Dynamic patterns: the self organization of brain and behavior.* Cambridge MA: The MIT Press.

Marr, D. (1982). *Vision: A computational investigation into human representation and processing of visual information.* San Francisco, CA: Freeman.

McClelland, J. L. (1989). Parallel distributed processing: implications for cognition and development. In R. G. M. Morris (Ed.), *Parallel distributed processing: Implications for psychology and neurobiology.* Oxford: Oxford University Press.

Newell, A., & Simon, H. A. (1972). *Human Problem Solving*. Englewood Cliffs, NJ: Prentice Hall.

Plunkett, K., & Sinha, C. (1992). Connectionism and developmental theory. *British Journal of Developmental Psychology, 10*, 209-254.

Raijmakers, M. E. J., van der Maas, H. L. J., & Molenaar, P. C. M. (1996a). Numerical bifurcation analysis of distance-dependent on-center off-surround shunting neural networks. *Biological Cybernetics, 75*, 495-507.

Shultz, T. R, Schmidt, W. C., Buckingham, D., & Mareschal, D. (1995). Modeling cognitive development with a generative connectionist algorithm. In T. J. Simon & G. S. Halford (Eds.), *Developing cognitive competence: new approaches to process modeling*. Hillsdale, NJ: Erlbaum.

Siegler, R. S. (1996). *Emerging minds*. Oxford: Oxford University Press.

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review, 103*(1), 56-115.

Thagard, P. (1992). *Conceptual Revolutions.* Princeton, NJ: The Princeton University Press.