Connectionist models of cognition

Michael S. C. Thomas and James L. McClelland

1. Introduction

In this chapter, we review computer models of cognition that have focused on the use of neural networks. These architectures were inspired by research into how computation works in the brain and subsequent work has produced models of cognition with a distinctive flavor. Processing is characterized by patterns of activation across simple processing units connected together into complex networks. Knowledge is stored in the strength of the connections between units. It is for this reason that this approach to understanding cognition has gained the name of *connectionism*.

2. Background

Over the last twenty years, connectionist modeling has formed an influential approach to the computational study of cognition. It is distinguished by its appeal to principles of neural computation to inspire the primitives that are included in its cognitive level models. Also known as artificial neural network (ANN) or parallel distributed processing (PDP) models, connectionism has been applied to a diverse range of cognitive abilities, including models of memory, attention, perception, action, language, concept formation, and reasoning (see, e.g., Houghton, 2005). While many of these models seek to capture adult function, connectionism places an emphasis on learning internal representations. This has led to an increasing focus on developmental phenomena and the origins of knowledge. Although, at its heart, connectionism comprises a set of computational formalisms, it has spurred vigorous theoretical debate regarding the nature of cognition. Some theorists have reacted by dismissing connectionism as mere implementation of pre-existing verbal theories of cognition, while others have viewed it as a candidate to replace the Classical Computational

Theory of Mind and as carrying profound implications for the way human knowledge is acquired and represented; still others have viewed connectionism as a sub-class of statistical models involved in universal function approximation and data clustering.

In this chapter, we begin by placing connectionism in its historical context, leading up to its formalization in Rumelhart and McClelland's two-volume *Parallel Distributed Processing* (1986) written in combination with members of the Parallel Distributed Processing Research Group. We then discuss three important early models that illustrate some of the key properties of connectionist systems and indicate how the novel theoretical contributions of these models arose from their key computational properties. These three models are the Interactive Activation model of letter recognition (McClelland & Rumelhart, 1981; Rumelhart and McClelland, 1982), Rumelhart and McClelland's model of the acquisition of the English past tense (1986), and Elman's simple recurrent network for finding structure in time (1991). We finish by considering how twenty-five years of connectionist modeling has influenced wider theories of cognition.

2.1 Historical context

Connectionist models draw inspiration from the notion that the information processing properties of neural systems should influence our theories of cognition. The possible role of neurons in generating the mind was first considered not long after the existence of the nerve cell was accepted in the latter half of the 19th Century (Aizawa, 2004). Early neural network theorizing can therefore be found in some of the associationist theories of mental processes prevalent at the time (e.g., Freud, 1895; James, 1890; Meynert, 1884; Spencer, 1872). However, this line of theorizing was quelled when Lashley presented data appearing to show that the performance of the

brain degraded gracefully depending only on the quantity of damage. This argued against the specific involvement of neurons in particular cognitive processes (see, e.g., Lashley, 1929).

In the 1930s and 40s, there was a resurgence of interest in using mathematical techniques to characterize the behavior of networks of nerve cells (e.g., Rashevksy, 1935). This culminated in the work of McCulloch and Pitts (1943) who characterized the function of simple networks of binary threshold neurons in terms of logical operations. In his 1949 book *The Organization of Behavior*, Donald Hebb proposed a cell assembly theory of cognition, including the idea that specific synaptic changes might underlie psychological principles of learning. A decade later, Rosenblatt (1958, 1962) formulated a learning rule for two-layered neural networks, demonstrating mathematically that the *perceptron convergence rule* could adjust the weights connecting an input layer and an output layer of simple neurons to allow the network to associate arbitrary binary patterns. With this rule, learning converged on the set of connection values necessary to acquire any two-layer-computable function relating a set of input-output patterns. Unfortunately, Minsky and Papert (1969) demonstrated that the set of two-layer computable functions was somewhat limited – that is, these simple artificial neural networks were not particularly powerful devices. While more computationally powerful networks could be described, there was no algorithm to learn the connection weights of these systems. Such networks required the postulation of additional internal or 'hidden' processing units, which could adopt intermediate representational states in the mapping between input and output patterns. An algorithm (backpropagation) able to learn these states was discovered independently several times. A key paper by Rumelhart, Hinton and Williams (1986) demonstrated

the usefulness of networks trained using backpropagation for addressing key computational and cognitive challenges facing neural networks.

In the 1970s, serial processing and the Von Neumann computer metaphor dominated cognitive psychology. Nevertheless, a number of researchers continued to work on the computational properties of neural systems. Some of the key themes identified by these researchers include the role of competition in processing and learning (e.g., Grossberg, 1976; Kohonen, 1984), the properties of distributed representations (e.g., Anderson, 1977; Hinton & Anderson, 1981), and the possibility of content addressable memory in networks with attractor states, formalized using the mathematics of statistical physics (Hopfield, 1982). A fuller characterization of the many historical influences in the development of connectionism can be found in Rumelhart and McClelland (1986, chapter 1), Bechtel and Abrahamsen (1991), McLeod, Plunkett, and Rolls (1998), and O'Reilly and Munakata (2000). Figure 1 depicts a selective schematic of this history and demonstrates the multiple types of neural network system that have latterly come to be used in building models of cognition. While diverse, they are unified on the one hand by the proposal that cognition comprises processes of constraint satisfaction, energy minimization and pattern recognition, and on the other that adaptive processes construct the microstructure of these systems, primarily by adjusting the strengths of connections among the neuron-like processing units involved in a computation.

Insert Figure 1 about here

2.2 Key properties of connectionist models

Connectionism starts with the following inspiration from neural systems: computations will be carried out by a set of simple processing units operating in parallel and affecting each others' activation states via a network of weighted connections. Rumelhart, Hinton and McClelland (1986) identified seven key features that would define a general framework for connectionist processing.

The first feature is the set of processing units u_i. In a cognitive model, these may be intended to represent individual concepts (such as letters or words), or they may simply be abstract elements over which meaningful patterns can be defined. Processing units are often distinguished into input, output, and hidden units. In associative networks, input and output units have states that are defined by the task being modeled (at least during training), while hidden units are free parameters whose states may be determined as necessary by the learning algorithm.

The second feature is a state of activation *(a)* at a given time *(t)*. The state of a set of units is usually represented by a vector of real numbers *a(t)*. These may be binary or continuous numbers, bounded or unbounded. A frequent assumption is that the activation level of simple processing units will vary continuously between the values 0 and 1.

The third feature is a pattern of connectivity. The strength of the connection between any two units will determine the extent to which the activation state of one unit can affect the activation state of another unit at a subsequent time point. The strength of the connections between unit *i* and unit *j* can be represented by a matrix *W* of weight values w_{ij}. Multiple matrices may be specified for a given network if there are connections of different types. For example, one matrix may specify excitatory connections between units and a second may specify inhibitory connections. Potentially, the weight matrix allows every unit to be connected to every other unit in

the network. Typically, units are arranged into layers (e.g., input, hidden, output) and layers of units are fully connected to each other. For example, in a three-layer feedforward architecture where activation passes in a single direction from input to output, the input layer would be fully connected to the hidden layer and the hidden layer would be fully connected to the output layer.

The fourth feature is a rule for propagating activation states throughout the network. This rule takes the vector *a(t)* of output values for the processing units sending activation and combines it with the connectivity matrix *W* to produce a summed or net input into each receiving unit. The net input to a receiving unit is produced by multiplying the vector and matrix together, so that

$$net_i = W \times a(t) = \sum_j w_{ij} a_j \qquad (1)$$

The fifth feature is an activation rule to specify how the net inputs to a given unit are combined to produce its new activation state. The function *F* derives the new activation state

$$a_i(t+1) = F(net_i(t)) \qquad (2)$$

For example, *F* might be a threshold so that the unit becomes active only if the net input exceeds a given value. Other possibilities include linear, Gaussian, and sigmoid functions, depending on the network type. Sigmoid is perhaps the most common, operating as a smoothed threshold function that is also differentiable. It is often important that the activation function be differentiable because learning seeks to improve a performance metric that is assessed via the activation state while learning itself can only operate on the connection weights. The effect of weight changes on the performance metric therefore depends to some extent on the activation function, and the learning algorithm encodes this fact by including the derivative of that function (see below).

The sixth key feature of connectionist models is the algorithm for modifying the patterns of connectivity as a function of experience. Virtually all learning rules for PDP models can be considered a variant of the Hebbian learning rule (Hebb, 1949). The essential idea is that a weight between two units should be altered in proportion to the units' correlated activity. For example, if a unit u_i receives input from another unit u_j, then if both are highly active, the weight w_{ij} from u_j to u_i should be strengthened. In its simplest version, the rule is

$$\Delta w_{ij} = \eta \, a_i a_j \tag{3}$$

where η is the constant of proportionality known as the learning rate. Where an external target activation $t_i(t)$ is available for a unit i at time t, this algorithm is modified by replacing a_i with a term depicting the disparity of unit u_i's current activation state $a_i(t)$ from its desired activation state $t_i(t)$ at time t, so forming the delta rule:

$$\Delta w_{ij} = \eta \, (t_i(t) - a_i(t)) a_j \tag{4}$$

However, when hidden units are included in networks, no target activation is available for these internal parameters. The weights to such units may be modified by variants of the Hebbian learning algorithm (e.g., Contrastive Hebbian; Hinton, 1989; see Xie & Seung, 2003) or by the backpropagation of error signals from the output layer.

Backpropagation makes it possible to determine, for each connection weight in the network, what effect a change in its value would have on the overall network error. The policy for changing the strengths of connections is simply to adjust each weight in the direction (up or down) that would tend to reduce the error, by an amount proportional to the size of the effect the adjustment will have. If there are multiple layers of hidden units remote from the output layer, this process can be followed iteratively: first error derivatives are computed for the hidden layer nearest the output

layer; from these, derivatives are computed for the next deepest layer into the network, and so forth. On this basis, the backpropagation algorithm serves to modify the pattern of weights in powerful multilayer networks. It alters the weights to each deeper layer of units in such a way as to reduce the error on the output units (see Rumelhart, Hinton, & Williams, 1986, for the derivation). We can formulate the weight change algorithm by analogy to the delta rule in shown in equation 4. For each deeper layer in the network, we modify the central term that represents the disparity between the actual and target activation of the units. Assuming u_i, u_h, and u_o are input, hidden, and output units in a 3-layer feedforward network, the algorithm for changing the weight from hidden to output unit is:

$$\Delta w_{oh} = \eta \left(t_o - a_o\right) F'\left(net_o\right) a_h \qquad (5)$$

where $F'\left(net\right)$ is the derivative of the activation function of the units (e.g., for the sigmoid activation function, $F'\left(net_o\right) = a_o\left(1 - a_o\right)$). The term $(t_o - a_o)$ is proportional to the negative of the partial derivative of the network's overall error with respect to the activation of the output unit, where the error E is given by $E = \sum_o \left(t_o - a_o\right)^2$.

The derived error term for a unit at the hidden layer is based on the derivative of the hidden unit's activation function, times the sum across all the connections from that hidden unit to the output later of the error term on each output unit weighted by the derivative of the output unit's activation function $\left(t_o - a_o\right) F'\left(net_o\right)$ times the weight connecting the hidden unit to the output unit:

$$F'\left(net_h\right) \sum_o \left(t_o - a_o\right) F'\left(net_o\right) w_{oh} \qquad (6)$$

The algorithm for changing the weights from the input to the hidden layer is therefore:

$$\Delta w_{hi} = \eta \;\; F'\left(net_h\right) \sum_o \left(t_o - a_o\right) F'\left(net_o\right) w_{oh} \; a_i \qquad (7)$$

It is interesting that the above computation can be construed as a backward pass through the network, similar in spirit to the forward pass that computes activations in that it involves propagation of signals across weighted connections, this time from the output layer back toward the input. The backward pass, however, involves the propagation of error derivatives rather than activations.

It should be emphasized that a very wide range of variants and extensions of Hebbian and error-correcting algorithms have been introduced in the connectionist learning literature. Most importantly, several variants of backpropagation have been developed for training recurrent networks (Williams & Zipser, 1995); and several algorithms (including the Contrastive Hebbian Learning algorithm and O'Reilly's 1998 LEABRA algorithm) have addressed some of the concerns that have been raised regarding the biological plausibility of backpropagation construed in its most literal form (O'Reilly & Munakata, 2000).

The last general feature of connectionist networks is a representation of the environment with respect to the system. This is assumed to consist of a set of externally provided events or a function for generating such events. An event may be a single pattern, such as a visual input; an ensemble of related patterns, such as the spelling of a word and its corresponding sound and/or meaning; or a sequence of inputs, such as the words in a sentence. A range of policies have been used for specifying the order of presentation of the patterns, including sweeping through the full set to random sampling with replacement. The selection of patterns to present may vary over the course of training but is often fixed. Where a target output is linked to each input, this is usually assumed to be simultaneously available. Two points are of note in the translation between PDP network and cognitive model. First, a representational scheme must be defined to map between the cognitive domain of

interest and a set of vectors depicting the relevant informational states or mappings for that domain. Second, in many cases, connectionist models are addressed to aspects of higher-level cognition, where it is assumed that the information of relevance is more abstract than sensory or motor codes. This has meant that the models often leave out details of the transduction of sensory and motor signals, using input and output representations that are already somewhat abstract. We hold the view that the same principles at work in higher-level cognition are also at work in perceptual and motor systems, and indeed there is also considerable connectionist work addressing issues of perception and action, though these will not be the focus of the present article.

2.3 Neural plausibility

It is a historical fact that most connectionist modelers have drawn their inspiration from the computational properties of neural systems. However, it has become a point of controversy whether these 'brain-like' systems are indeed neurally plausible. If they are not, should they instead be viewed as a class of statistical functional approximators? And if so, shouldn't the ability of these models to simulate patterns of human behavior be assessed in the context of the large number of free parameters they contain (e.g., in the weight matrix) (Green, 1998)?

Neural plausibility should not be the primary focus for a consideration of connectionism. The advantage of connectionism, according to its proponents, is that it provides *better theories of cognition*. Nevertheless, we will deal briefly with this issue since it pertains to the origins of connectionist cognitive theory. In this area, two sorts of criticism have been leveled at connectionist models. The first is to maintain that many connectionist models either include properties that are not neurally plausible and/or omit other properties that neural systems appear to have. Some connectionist

researchers have responded to this first criticism by endeavoring to show how features of connectionist systems might in fact be realized in the neural machinery of the brain. For example, the backward propagation of error across the same connections that carry activation signals is generally viewed as biologically implausible. However, a number of authors have shown that the difference between activations computed using standard feedforward connections and those computed using standard return connections can be used to derive the crucial error derivatives required by backpropagation (Hinton & McClelland, 1988; O'Reilly, 1996). It is widely held that connections run bi-directionally in the brain, as required for this scheme to work. Under this view, backpropagation may be shorthand for a Hebbian-based algorithm that uses bi-directional connections to spread error signals throughout a network (Xie & Seung, 2003).

Other connectionist researchers have responded to the first criticism by stressing the cognitive nature of current connectionist models. Most of the work in developmental neuroscience addresses behavior at levels no higher than cellular and local networks, whereas cognitive models must make contact with the human behavior studied in psychology. Some simplification is therefore warranted, with neural plausibility compromised under the working assumption that the simplified models share the same flavor of computation as actual neural systems. Connectionist models have succeeding in stimulating a great deal of progress in cognitive theory – and sometimes generating radically different proposals to the previously prevailing symbolic theory – just given the set of basic computational features outlined in the preceding section.

The second type of criticism leveled at connectionism questions why, as Davies (2005) puts it, connectionist models should be reckoned any more plausible as

putative descriptions of cognitive processes just because they are 'brain-like'. Under this view, there is independence between levels of description because a given cognitive level theory might be implemented in multiple ways in different hardware. Therefore the details of the hardware (in this case, the brain) need not concern the cognitive theory. This functionalist approach, most clearly stated in Marr's three levels of description (computational, algorithmic, and implementational; see Marr, 1982) has been repeatedly challenged (see, e.g., Rumelhart & McClelland, 1985; Mareschal et al., 2007). The challenge to Marr goes as follows. While, according to computational theory, there may be a principled independence between a computer program and the particular substrate on which it is implemented, in practical terms, different sorts of computation are easier or harder to implement on a given substrate. Since computations have to be delivered in real time as the individual reacts with his or her environment, in the first instance cognitive level theories should be constrained by the computational primitives that are most easily implemented on the available hardware; human cognition should be shaped by the processes that work best in the brain.

The relation of connectionist models to symbolic models has also proved controversial. A full consideration of this issue is beyond the scope of the current chapter. Suffice to say that because the connectionist approach now includes a diverse family of models, there is no single answer to this question. Smolensky (1988) argued that connectionist models exist at a lower (but still cognitive) level of description than symbolic cognitive theories, a level that he called the *sub-symbolic*. Connectionist models have sometimes been put forward as a way to implement symbolic production systems on neural architectures (e.g., Touretzky & Hinton, 1988). At other times, connectionist researchers have argued that their models represent a qualitatively

different form of computation: while under certain circumstances, connectionist models might produce behavior approximating symbolic processes, it is held that human behavior, too, only approximates the characteristics of symbolic systems rather than directly implementing them. Furthermore, connectionist systems incorporate additional properties characteristic of human cognition, such as content addressable memory, context-sensitive processing, and graceful degradation under damage or noise. Under this view, symbolic theories are approximate descriptions rather than actual characterizations of human cognition. Connectionist theories should replace them because they both capture subtle differences between human behavior and symbolic characterizations, and because they provide a specification of the underlying causal mechanisms (van Gelder, 1991).

This strong position has prompted criticisms that in their current form, connectionist models are insufficiently powerful to account for certain aspects of human cognition – in particular those areas best characterized by symbolic, syntactically driven computations (Fodor & Pylyshyn, 1988; Marcus, 2001). Again, however, the characterization of human cognition in such terms is highly controversial; close scrutiny of relevant aspects of language – the ground on which the dispute has largely been focused – lends support to the view that the systematicity assumed by proponents of symbolic approaches is overstated, and that the actual characteristics of language are well matched to the characteristics of connectionist systems (Bybee & McClelland, 2005; McClelland, Plaut, Gotts & Maia, 2003). In the end, it may be difficult to make principled distinctions between symbolic and connectionist models. At a fine scale, one might argue that two units in a network represent variables and the connection between them specifies a symbolic rule linking these variables. One might also argue that a production system in which rules are

allowed to fire probabilistically and in parallel begins to approximate a connectionist system.

2.4 The relationship between connectionist models and Bayesian inference

Since the early 1980s, it has been apparent that there are strong links between the calculations carried out in connectionist models and key elements of Bayesian calculations. The state of the early literature on this point was reviewed in McClelland (1998). There it was noted, first of all, that units can be viewed as playing the role of probabilistic hypotheses; that weights and biases play the role of conditional probability relations between hypotheses and prior probabilities, respectively; and that if connection weights and biases have the correct values, the logistic activation function sets the activation of a unit to its posterior probability given the evidence represented on its inputs. A second and more important observation is that, in stochastic neural networks (Boltzmann Machines and Continuous Diffusion Networks; Hinton & Sejnowski, 1986; Movellan & McClelland, 1993) a network's state over all of its units can represent a constellation of hypotheses about an input; and (if the weights and the biases are set correctly) that the probability of finding the network in a particular state is monotonically related to the probability that the state is the correct interpretation of the input. The exact nature of the relation depends on a parameter called temperature; if set to one, the probability that the network will be found in a particular state exactly matches its posterior probability. When temperature is gradually reduced to zero, the network will end up in the most probable state, thus performing optimal perceptual inference (Hinton & Sejnowski, 1983). It is also known that backpropagation can learn weights that allow Bayes-optimal estimation of outputs given inputs (MacKay, 1993) and that the Boltzmann machine learning

algorithm (Ackley, Hinton, & Sejnowski, 1986; Movellan & McClelland, 1993) can learn to produce correct conditional distributions of outputs given inputs. The algorithm is slow but there has been recent progress producing substantial speedups that achieve outstanding performance on benchmark data sets (Hinton & Salakhutdinov, 2006).

3. Three illustrative models

In this section, we outline three of the landmark models in the emergence of connectionist theories of cognition. The models serve to illustrate the key principles of connectionism and demonstrate how these principles are relevant to explaining behavior in ways that are different from other prior approaches. The contribution of these models was twofold: they were better suited than alternative approaches to capturing the actual characteristics of human cognition, usually on the basis of their context sensitive processing properties; and compared to existing accounts, they offered a sharper set of tools to drive theoretical progress and to stimulate empirical data collection. Each of these models significantly advanced its field.

3.1 An interactive activation model of context effects in letter perception (McClelland & Rumelhart, 1981, 1982)

The interactive activation model of letter perception illustrates two interrelated ideas. The first is that connectionist models naturally capture a graded constraint satisfaction process in which the influences of many different types of information are simultaneously integrated in determining, for example, the identity of a letter in a word. The second idea is that the computation of a perceptual representation of the current input (in this case, a word) involves the simultaneous and mutual influence of representations at *multiple levels of abstraction* – this is a core idea of parallel distributed processing.

The interactive activation model addressed itself to a puzzle in word recognition. By the late 1970s, it had long been known that people were better at recognizing letters presented in words than letters presented in random letter sequences. Reicher (1969) demonstrated that this was not the result of tending to

guess letters that would make letter strings into words. He presented target letters either in words, unpronounceable nonwords, or on their own. The stimuli were then followed by a pattern mask, after which participants were presented with a forced choice between two letters in a given position. Importantly, both alternatives were equally plausible. Thus, the participant might be presented with WOOD and asked whether the third letter was O or R. As expected, forced-choice performance was more accurate for letters in words than for letters in nonwords or presented on their own. Moreover, the benefit of surrounding context was also conferred by pronounceable pseudowords (e.g., recognizing the P in SPET) compared to random letter strings, suggesting that subjects were able to bring to bear rules regarding the orthographic legality of letter strings during recognition.

Rumelhart and McClelland took the contextual advantage of words and pseudowords on letter recognition to indicate the operation of *top-down* processing. Previous theories had put forward the idea that letter and word recognition might be construed in terms of detectors which collect evidence consistent with the presence of their assigned letter or word in the input (Morton, 1969; Selfridge, 1959). Influenced by these theories, Rumelhart and McClelland built a computational simulation in which the perception of letters resulted from excitatory and inhibitory interactions of detectors for visual features. Importantly, the detectors were organized into different layers for letter features, letters and words, and detectors could influence each other both in a bottom-up and a top-down manner.

Figure 2 illustrates the structure of the Interactive Activation (IA) model, both at the macro level (left) and for a small section of the model at a finer level (right). The explicit motivation for the structure of the IA was neural: '[We] have adopted the approach of formulating the model in terms similar to the way in which such a

process might actually be carried out in a neural or neural-like system' (McClelland & Rumelhart, 1981, p.387). There were three main assumptions of the IA model: (1) perceptual processing takes place in a system in which there are several levels of processing, each of which forms a representation of the input at a different level of abstraction; (2) visual perception involves parallel processing, both of the four letters in each word and of all levels of abstraction simultaneously; (3) perception is an interactive process in which conceptually driven and data driven processing provide multiple, simultaneously acting constraints that combine to determine what is perceived.

The activation states of the system were simulated by a sequence of discrete time steps. Each unit combined its activation on the previous time step, its excitatory influences, its inhibitory influences, and a decay factor to determine its activation on the next time step. Connectivity was set at unitary values and along the following principles: in each layer, mutually exclusive alternatives should inhibit each other. For each unit in a layer, it excited all units with which it was consistent and inhibited all those with which it was inconsistent in layer immediately above. Thus in Figure 2, the 1st-position W letter unit has an excitatory connection to the WEED word unit but an inhibitory connection to the SEED and FEED word units. Similarly, a unit excited all units with which it was consistent and inhibited all those with which it was inconsistent in the layer immediately below. However, in the final implementation, top-down word-to-letter inhibition and within-layer letter-to-letter inhibition were set to zero (gray arrows, Figure 2).

Insert Figure 2 about here

The model was constructed to recognize letters in 4-letter strings. The full set of possible letters was duplicated for each letter position, and a set of 1,179 word units created to represent the corpus of 4-letter words. Word units were given base rate activation states at the beginning of processing to reflect their different frequencies. A trial began by clamping the feature units to the appropriate states to represent a letter string, and then observing the dynamic change in activation through the network. Conditions were included to allow the simulation of stimulus masking and degraded stimulus quality. Finally, a probabilistic response mechanism was added to generate responses from the letter level, based on the relative activation states of the letter pool in each position.

The model successfully captured the greater accuracy of letter detection for letters appearing in words and pseudowords compared to random strings or in isolation. Moreover, it simulated a variety of empirical findings on the effect of masking and stimulus quality, and of changing the timing of the availability of context. The results on the contextual effects of pseudowords are particularly interesting, since the model only contains word units and letter units and has no explicit representation of orthographic rules. Let us say on a given trial, the subject is required to recognize the 2nd letter in the string SPET. In this case, the string will produce bottom-up excitation of the word units for SPAT, SPIT, and SPOT, which each share three letters. In turn, the word units will propagation top-down activation reinforcing activation of the letter P and so facilitating its recognition. Were this letter to be presented in the string XPQJ, no word units could offer similar top-down activation, hence the relative facilitation of the pseudoword. Interestingly, although these top-down 'gang' effects produced facilitation of letters contained in orthographically legal nonword strings, the model demonstrated that they also

produced facilitation in orthographically illegal, unpronounceable letter strings such as SPCT. Here, the same gang of SPAT, SPIT, and SPOT produce top-down support. Rumelhart and McClelland (1982) reported empirical support for this novel prediction. Therefore, although the model behaved *as if it contained orthographic rules influencing recognition*, it did not in fact do so, because continued contextual facilitation could be demonstrated for strings that had gang support but violated the orthographic rules.

There are two specific points to note regarding the IA model. First, this early connectionist model was not adaptive – connectivity was set by hand. While the model's behavior was shaped by the statistical properties of the language it processed, these properties were built into the structure of the system, in terms of the frequency of occurrence of letters and letter combinations in the words. Second, the idea of bottom-up excitation followed by competition amongst mutually exclusive possibilities is a strategy familiar in Bayesian approaches to cognition. In that sense, the IA bears similarity to more recent probability theory based approaches to perception.

What happened next?

Subsequent work saw the principles of the IA model extended to the recognition of spoken words (the TRACE model: McClelland & Elman, 1986) and more recently to bilingual speakers where two languages must be incorporated in a single representational system (see Thomas & van Heuven, 2005, for review). The architecture was applied to other domains where multiple constraints were thought to operate during perception, for example in face recognition (Burton, Bruce, & Johnston, 1990). Within language, more complex architectures have tried to recast the

principles of the IA model in developmental settings, such as Plaut and Kello's (1999) model of the emergence of phonology from the interplay of speech comprehension and production.

The more general lesson to draw from the interactive activation model is the demonstration of multiple influences (feature, letter, and word-level knowledge) working simultaneously and in parallel to shape the response of the system; and the somewhat surprising finding that a massively parallel constraint satisfaction process of this form can appear to behave as if it contains rules (in this case, orthographic) when no such rules are included in the processing structure. At the time, the model brought into question whether it was necessary to postulate rules as processing structures to explain regularities in human behavior. This skepticism was brought into sharper focus by our next example.

3.2 On learning the past tense of English verbs (Rumelhart & McClelland, 1986)

Rumelhart and McClelland's (1986) model of English past tense formation marked the real emergence of the PDP framework. Where the IA model used localist coding, the past tense model employed distributed coding. Where the IA model had handwired connection weights, the past tense model learned its weights via repeated exposure to a problem domain. However, the models share two common themes. Once more, the behavior of the past model will be driven by the statistics of the problem domain, albeit these will be carved into the model by training rather than sculpted by the modelers. Perhaps more importantly, we see a return to the idea that a connectionist system can exhibit rule-following behavior without containing rules as causal processing structures; but in this case, the rule-following behavior will be the product of learning and will accommodate a proportion of exception patterns that do

not follow the general rule. The key point that the past tense model illustrates is how (approximate) conformity to the regularities of language – and even a tendency to produce new regular forms (e.g., regularizations like 'thinked' or past tenses for novel verbs like 'wugged') – can arise in a connectionist network without an explicit representation of a linguistic rule.

The English past tense is characterized by a predominant regularity in which the majority of verbs form their past tenses by the addition of one of three allomorphs of the '-ed' suffix to the base stem (walk/walked, end/ended, chase/chased). However, there is a small but significant group of verbs which form their past tense in different ways, including changing internal vowels (swim/swam), changing word final consonants (build/built), changing both internal vowels and final consonants (think/thought), an arbitrary relation of stem to past tense (go/went), and verbs which have a past tense form identical to the stem (hit/hit). These so-called irregular verbs often come in small groups sharing a family resemblance (sleep/slept, creep/crept, leap/leapt) and usually have high token frequencies (see Pinker, 1999, for further details).

During the acquisition of the English past tense, children show a characteristic U-shaped developmental profile at different times for individual irregular verbs. Initially they use the correct past tense of a small number of high frequency regular and irregular verbs. Latterly, they sometimes produce 'overregularized' past tense forms for a small fraction of their irregular verbs (e.g., thinked) (Marcus, Pinker, Ullman, Hollander, Rosen, & Xu, 1992), along with other, less frequent errors (Xu & Pinker, 1995). They are also able to extend the past tense 'rule' to novel verbs (e.g., wug-wugged). Finally, in older children, performance approaches ceiling on both regular and irregular verbs (Berko, 1958; Ervin, 1964; Kuczaj, 1977).

In the early 1980s, it was held that this pattern of behavior represented the operation of two developmental mechanisms (Pinker, 1984). One of these was symbolic and served to learn the regular past tense 'rule', while the other was associative and served to learn the exceptions to the rule. The extended phase of overregularization errors corresponded to difficulties in integrating the two mechanisms, specifically a failure of the associative mechanism to block the function of the symbolic mechanism. That the child comes to the language acquisition situation armed with these two mechanisms (one of them full of blank rules) was an a priori commitment of the developmental theory.

By contrast, Rumelhart and McClelland (1986) proposed that a single network that does not distinguish between regular and irregular past tenses is sufficient to learn past tense formation. The architecture of their model is shown in Figure 3. A phoneme-based representation of the verb root was recoded into a more distributed, coarser (more blurred) format, which they called 'Wickelfeatures'. The stated aim of this recoding was to produce a representation that (a) permitted differentiation of all of the root forms of English and their past tenses, and (b) provided a natural basis for generalizations to emerge about what aspects of a present tense correspond to what aspects of a past tense. This format involved representing verbs over 460 processing units. A two-layer network was then used to associate the Wickelfeature representations of the verb root and past tense form. A final decoding network was then used to derive the closest phoneme-based rendition of the past tense form and reveal the model's response (the decoding part of the model was somewhat restricted by computer processing limitations of the machines available at the time).

The connection weights in the two-layer network were initially randomized. The model was then trained in three phases, in each case using the delta rule to update

the connection weights after each verb root / past tense pair was presented (see Section 1.2). In Phase 1, the network was trained on 10 high frequency verbs, 2 regular and 8 irregular, in line with the greater proportion of irregular verbs amongst the most frequent verbs in English. Phase 1 lasted for 10 presentations of the full training set (or 'epochs'). In Phase 2, the network was trained on 410 medium frequency verbs, 334 regular and 76 irregular, for a further 190 epochs. In Phase 3, no further training took place, but 86 lower frequency verbs were presented to the network to test its ability to generalize its knowledge of the past tense domain to novel verbs.

Insert Figure 3 about here

There were four key results for this model. First, it succeeded in learning both regular and irregular past tense mappings in a single network that made no reference to the distinction between regular and irregular verbs. Second, it captured the overall pattern of faster acquisition for regular verbs than irregular verbs, a predominant feature of children's past tense acquisition. Third, the model captured the U-shaped profile of development: an early phase of accurate performance on a small set of regular and irregular verbs, followed by a phase of overregularization of the irregular forms, and finally recovery for the irregular verbs and performance approaching ceiling on both verb types. Fourth, when the model was presented with the low-frequency verbs on which it had not been trained, it was able to generalize the past tense rule to a substantial proportion of them, as if it had indeed learned a rule. Additionally, the model captured more fine-grained developmental patterns for subsets of regular and irregular verbs, and generated several novel predictions.

Rumelhart and McClelland explained the generalization abilities of the network in terms of the *superpositional* memory of the two-layer network. All the associations between the distributed encodings of verb root and past tense forms must be stored across the single matrix of connection weights. As a result, similar patterns blend into one another and reinforce each other. Generalization is contingent on the similarity of verbs at input. Were the verbs to be presented using an orthogonal, localist scheme (e.g., 420 units, 1 per verb), then there would be no similarity between the verbs, no blending of mappings, no generalization, and therefore no regularization of novel verbs. As the authors state, 'it is the statistical relationships among the base forms themselves that determine the pattern of responding. The network merely reflects the statistics of the featural representations of the verb forms' (p. 267). Based on the model's successful simulation of the profile of language development in this domain and, compared to the dual mechanism model, its more parsimonious a priori commitments, Rumelhart and McClelland viewed their work on past tense morphology as a step towards a revised understanding of language knowledge, language acquisition, and linguistic information processing in general.

The past tense model stimulated a great deal of subsequent debate, not least because of its profound implications for theories of language development (no rules!). The model was initially subjected to concentrated criticism. Some of this was overstated – for instance, the use of domain-general learning *principles* (such as distributed representation, parallel processing, and the delta rule) to acquire the past tense in a single network was interpreted as a claim that all of language acquisition could be captured by the operation of a single domain-general learning *mechanism*. Such an absurd claim could be summarily dismissed. However, as it stood, the model made no such claim: its generality was in the processing principles. The model itself

represented a domain-specific system dedicated to learning a small part of language. Nevertheless, a number of the criticisms were more telling: the Wickelfeature representational format was not psycholinguistically realistic; the generalization performance of the model was relatively poor; the U-shaped developmental profile appeared to be a result of abrupt changes in the composition of the training set; and the actual response of the model was hard to discern because of problems in decoding the Wickelfeature output into a phoneme string (Pinker & Prince, 1988).

The criticisms and following rejoinders were interesting in a number of ways. First, there was a stark contrast between the precise, computationally implemented connectionist model of past tense formation and the verbally specified dual-mechanism theory (e.g., Marcus, Pinker, Ullman, Hollander, Rosen, & Xu, 1992). The implementation made simplifications but was readily evaluated against quantitative behavioral evidence; it made predictions and it could be falsified. The verbal theory by contrast was vague – it was hard to know how or whether it would work or exactly what behaviors its predicted (see Thomas et al., 2006, for discussion). Therefore it could only be evaluated on loose qualitative grounds. Second, the model stimulated a great deal of new multidisciplinary research in the area. Today, inflectional morphology (of which past tense is a part) is one of the most studied aspects of language processing in children, in adults, in second language learners, in adults with acquired brain damage, in children and adults with neurogenetic disorders, and in children with language impairments, using psycholinguistic methods, event-related potential measures of brain activity, functional magnetic resonance imaging, and behavioral genetics . . . This rush of science illustrates the essential role of computational modeling in driving forward theories of human cognition. Third, further modifications and improvements to the past tense model have highlighted how

researchers go about the difficult task of understanding which parts of their model represent the key theoretical claims and which are implementational details. Simplification is inherent to modeling but successful modeling relies on making the *right* simplifications to focus on the process of interest. For example, in subsequent models, the Wickelfeature representation was replaced by more plausible phonemic representations based on articulatory features; the recoding/two-layer-network/decoding component of the network (the dotted rectangle in Figure 3) that was trained with the delta rule was replaced by a three-layer feedforward network trained with the backpropagation algorithm; and the U-shaped developmental profile was demonstrated in connectionist networks trained with a smoothly growing training set of verbs or even with a fixed set of verbs (see, e.g., Plunkett & Marchman, 1991, 1993, 1996).

What happened next?

The English past tense model prompted further work within inflectional morphology in other languages (e.g., pluralization in German: Goebel & Indefrey, 2000; pluralization in Arabic: Plunkett & Nakisa, 1999), as well as models that explored the possible causes of deficits in acquired and developmental disorders such as aphasia, Specific Language Impairment and Williams syndrome (e.g., Hoeffner & McClelland, 1993; Joanisse & Seidenberg, 1999; Thomas & Karmiloff-Smith, 2003a; Thomas, 2005). The idea that rule-following behavior could emerge in a developing system that also had to accommodate exceptions to the rules was also successfully pursued via connectionist modeling in the domain of reading (e.g., Plaut et al., 1996). This led to work that also considered various forms of acquired and developmental dyslexia.

For the past tense itself, there remains much interest in the topic as a crucible to test theories of language development. However, in some senses the debate between connectionist and dual-mechanism accounts has ground to a halt. There is much evidence from child development, adult cognitive neuropsychology, developmental neuropsychology, and functional brain imaging to suggest partial dissociations between performance on regular and irregular inflection under various conditions. Both connectionist and dual-mechanism models have been modified: the connectionist model to include the influence of lexical-semantics as well as verb root phonology in driving the production of the past tense form (Joanisse & Seidenberg, 1999; Thomas & Karmiloff-Smith, 2003a); the dual-mechanism model to suppose that regular verbs might also be stored in the associative mechanism, thereby introducing partial redundancy of function (Pinker, 1999). Both approaches now accept that performance on regular and irregular past tenses partly indexes different things – in the connectionist account, different underlying knowledge, in the dual-mechanism account, different underlying processes. In the connectionist theory, performance on regular verbs indexes reliance on knowledge about phonological regularities while performance on irregular verbs indexes reliance on lexical-semantic knowledge. In the dual-mechanism theory, performance on regular verbs indexes a dedicated symbolic processing mechanism implementing the regular 'rule' while performance on irregular verbs indexes an associative memory device storing information about the past tense forms of specific verbs. Both approaches claim to account for the available empirical evidence. However, to date, the dual-mechanism remains unimplemented, so its claim is weaker.

How does one distinguish between two theories that (a) both claim to explain the data but (b) contain different representational assumptions? Putting aside the

different level of detail of the two theories, the answer is that it depends on one's preference for consistency with other disciplines. The dual-mechanism theory declares consistency with linguistics – if rules are required to characterize other aspects of language performance (such as syntax), then one might as well include them in a model of past tense formation. The connectionist theory declares consistency with neuroscience – if the language system is going to be implemented in the brain, then one might as well employ a computational formulism based on how neural networks function.

Finally, we return to the more general connectionist principle illustrated by the past tense model. So long as there are regularities in the statistical structure of a problem domain, a massively parallel constraint satisfaction system can learn these regularities and extend them to novel situations. Moreover, as with humans, the behavior of the system is flexible and context sensitive – it can accommodate regularities and exceptions within a single processing structure.

3.3 Finding Structure in Time (Elman, 1990)

In this section, we introduce the notion of the simple recurrent network and its application to language. As with past tense, the key point of the model will be to show how conformity to regularities of language can arise without an explicit representation of a linguistic rule. Moreover, the following simulations will demonstrate how learning can lead to the discovery of useful internal representations that capture conceptual and linguistic structure on the basis of the co-occurrences of words in sentences.

The IA model exemplified connectionism's commitment to parallelism: all of the letters of the word presented to the network were recognized in parallel and

processing occurred simultaneously at different levels of abstraction. But not all processing can be carried out in this way. Some human behaviors intrinsically revolve around temporal sequences. Language, action planning, goal-directed behavior, and reasoning about causality are examples of domains that rely on events occurring in sequences. How has connectionism addressed the processing of temporally unfolding events? One solution was offered in the TRACE model of spoken word recognition (McClelland & Elman, 1986) where a word was specified as a sequence of phonemes. In that case, the architecture of the system was duplicated for each time slice and the duplicates wired together. This allowed constraints to operate over items in the sequence to influence recognition. In other models, a related approach was used to convert a temporally extended representation into a spatially extended one. For example, in the past tense model, all the phonemes of a verb were presented across the input layer. This could be viewed as a sequence if one assumed that the representation of the first phoneme represents time slice *t*, the representation of the second phoneme represents time slice *t+1*, and so on. As part of a comprehension system, this approach assumes a buffer that can take sequences and convert them to a spatial vector. However, this solution is fairly limited, as it necessarily pre-commits to the size of the sequences that can be processed at once (i.e., the size of the input layer).

Elman (1990, 1991) offered an alternative and more flexible approach to processing sequences, proposing an architecture that has been extremely influential and much used since. Elman drew on the work of Jordan (1986) who had proposed a model that could learn to associate a 'plan' (i.e., a single input vector) with a series of 'actions' (i.e., a sequence of output vectors). Jordan's model contained recurrent connections permitting the hidden units to 'see' the network's previous output (via a

set of 'state' input units that are given a copy of the previous output). The facility for the network to shape its next output according to its previous response constitutes a kind of memory. Elman's innovation was to build a recurrent facility into the internal units of the network, allowing it to compute statistical relationships across sequences of inputs and outputs. To achieve this, first time is discretized into a number of slices. On time step *t*, an input is presented to the network and causes a pattern of activation on hidden and output layers. On time step *t+1*, the next input in the sequence of events is presented to the network. However, crucially, a copy of the activation of the hidden units on time step *t* is transmitted to a set of internal 'context' units. This activation vector is also fed to the hidden units on time step *t+1*. Figure 4 shows the architecture, known as the *simple recurrent network (SRN)*. It is usually trained with the backpropagation algorithm (see Section 2.3) as a multi-layer feedforward network, ignoring the origin of the information on the context layer.

Insert Figure 4 about here

Each input to the SRN is therefore processed in the context of what came before, but in a way subtly more powerful than the Jordan network. The input at *t+1* is processed in the context of the activity produced on the hidden units by the input at time *t.* Now consider the next time step. The input at time *t+2* will be processed along with activity from the context layer that is shaped by *two* influences:

(the input at *t+1* (shaped by the input at *t*))

The input at time *t+3* will be processed along with activity from the context layer that is shaped by *three* influences:

(the input at *t+2* (shaped by the input at *t+1* (shaped by the input at *t*)))

The recursive flavor of the information contained in the context layer means that each new input is processed in the context of the *full history* of previous inputs. This permits the network to learn statistical relationships across sequences of inputs or, in other words, to find structure in time.

In his original paper of 1990, Elman demonstrated the powerful properties of the SRN with two examples. In the first, the network was presented with a sequence of letters made up of concatenated words, e.g.:

MANYYEARSAGOABOYANDGIRLLIVEDBYTHESEATHEYPLAYEDHAPPIL

Each letter was represented by a distributed binary code over 5 input units. The network was trained to predict the next letter in the sentence for 200 sentences constructed from a lexicon of 15 words. There were 1,270 words and 4,963 letters. Since each word appeared in many sentences, the network was not particularly successful at predicting the next letter when it got to the end of each word, but within a word it was able to predict the sequences of letters. Using the accuracy of prediction as a measure, one could therefore identify which sequences in the letter string were words: they were the sequences of good prediction bounded by high prediction errors. The ability to extract words was of course subject to the ambiguities inherent in the training set (e.g., for *the* and *they*, there is ambiguity after the 3rd letter). Elman suggested that if the letter strings are taken to be analogous to the speech sounds available to the infant, the SRN demonstrates a possible mechanism to extract words from the continuous stream of sound that is present in infant-directed speech. Elman's work has contributed to the increasing interest in the statistical learning abilities of

young children in language and cognitive development (see, e.g., Saffran, Newport, & Aslin, 1996).

In the second example, Elman created a set of 10,000 sentences by combining a lexicon of 29 words and a set of short sentence frames (noun + [transitive] verb + noun; noun + [intransitive] verb). There was a separate input and output unit for each word and the SRN was trained to predict the next word in the sentence. During training, the network's output came to approximate the transitional probabilities between the words in the sentences – that is, it could predict the next word in the sentences as much as this was possible. Following the first noun, the verb units would be more active as the possible next word, and verbs that tended to be associated with this particular noun would be more active than those that did not. At this point, Elman examined the similarity structure of the internal representations to discover how the network was achieving its prediction ability. He found that the internal representations were sensitive to the difference between nouns and verbs, and within verbs, to the difference between transitive and intransitive verbs. Moreover, the network was also sensitive to a range of semantic distinctions: not only were the internal states induced by nouns split into animate and inanimate, but the pattern for 'woman' was most similar to 'girl', and that for 'man' was most similar to 'boy'. The network had learnt to structure its internal representations according to a mix of syntactic and semantic information because these information states were the best way to predict how sentences would unfold. Elman concluded that the representations induced by connectionist networks need not be flat but could include hierarchical encodings of category structure.

Based on his finding, Elman also argued that the SRN was able to induce representations of entities that varied according to their context of use. This contrasts

with classical symbolic representations that retain their identity irrespective of the combinations into which they are put, a property called 'compositionality'. This claim is perhaps better illustrated by a second paper Elman published two years later called 'The importance of starting small' (1993). In this later paper, Elman explored whether rule-based mechanisms are required to explain certain aspects of language performance, such as syntax. He focused on 'long-range dependencies', which are links between words that depend only on their syntactic relationship in the sentence and, importantly, not on their separation in a sequence of words. For example, in English, the subject and main verb of a sentence must agree in number. If the noun is singular, so must be the verb; if the noun is plural, so must be the verb. Thus, in the sentence 'The **boy chases** the cat', *boy* and *chases* must both be singular. But this is also true in the sentence 'The **boy** whom the boys chase **chases** the cat'. In the second sentence, the subject and verb are further apart in the sequence of words but their relationship is the same; moreover, the words are now separated by plural tokens of the same lexical items. Rule-based representations of syntax were thought to be necessary to encode these long-distance relationships because, through the recursive nature of syntax, the words that have to agree in a sentence can be arbitrarily far apart.

Using an SRN trained on the same prediction task as that outlined above but now with more complex sentences, Elman (1993) demonstrated that the network was able to learn these long-range dependencies even across the separation of multiple phrases. If *boy* was the subject of the sentence, when the network came to predict the main verb *chase* as the next word, it predicted that it should be in the singular. The method by which the network achieved this ability is of particular interest. Once more, Elman explored the similarity structure in the hidden unit representations, using principal component analyses to identify the salient dimensions of similarity across

which activation states were varying. This enabled him to reduce the high dimensionality of the internal states (150 hidden units were used) to a manageable number in order to visualize processing. Elman was then able to plot the *trajectories* of activation as the network altered its internal state in response to each subsequent input. Figure 5 depicts these trajectories as the network processes different multi-phrase sentences, plotted with reference to particular dimensions of principal component space. This figure demonstrates that the network adopted similar states in response to particular lexical items (e.g., tokens of *boy*, *who*, *chases*), but that it modified the pattern slightly according to the grammatical status of the word. In Figure 5(a), the second principal component appears to encode singularity/plurality. Figure 5(b) traces the network's state as it processes two embedded relative clauses containing iterations of the same words. Each clause exhibits a related but slightly shifted triangular trajectory to encode its role in the syntactic structure.

The importance of this model is that it prompts a different way to understand the processing of sentences. Previously one would view symbols as possessing fixed identities and as being bound into particular grammatical roles via a syntactic construction. In the connectionist system, sentences are represented by trajectories through activation space in which the activation pattern for each word is subtly shifted according to the context of its usage. The implication is that the property of compositionality at the heart of the classical symbolic computational approach may not be necessary to process language.

Insert Figure 5 about here

Elman (1993) also used this model to investigate a possible advantage to learning that could be gained by initially restricting the complexity of the training set. At the start of training, the network had its memory reset (its context layer wiped) after every third or fourth word. This window was then increased in stages up to 6-7 words across training. The manipulation was intended to capture maturational changes in working memory in children. Elman (1993) reported that *starting small* enhanced learning by allowing the network to build simpler internal representations that were later useful for unpacking the structure of more complex sentences (see Rohde & Plaut, 1999, for discussion and further simulations). This idea resonated with developmental psychologists in its demonstration of the way in which learning and maturation might interact in constructing cognition. It is an idea that could turn out to be a key principle in the organization of cognitive development (Elman et al., 1996).

What happened next?

Elman's simulations with the SRN and the prediction task produced striking results. The ability of the network to induce structured representations containing grammatical and semantic information from word sequences prompted the view that associative statistical learning mechanisms might play a much more central role in language acquisition. This innovation was especially welcome given that symbolic theories of sentence processing do not offer a ready account of language development. Indeed, they are largely identified with the nativist view that little in syntax develops. However, one limitation of the above simulations is that the prediction task does not learn any categorizations over the input set. While the simulations demonstrate that information important for language comprehension and production can be induced

from word sequences, neither task is performed. The learned distinction between nouns and verbs apparent in the hidden unit representations is tied up with carrying out the prediction task. But to perform comprehension, for example, the SRN would need to learn categorizations from the word sequences, such as deciding which noun was the agent and which noun was the patient in a sentence, irrespective of whether the sentence was presented in the active ('the dog chases the cat') or passive voice ('the cat is chased by the dog'). These types of computations are more complex and the network's solutions typically more impenetrable. While SRNs have borne the promise of an inherently developmental connectionist theory of parsing, progress on a full model has been slow (see Christiansen & Chater, 2001). Parsing is a complex problem – it is not even clear what the output should be for a model of sentence comprehension. Should it be some intermediate depiction of agent-patient role assignments, some compound representation of roles and semantics, or a constantly updating mental model that processes each sentence in the context of the emerging discourse? Connectionist models of parsing await greater constraints from psycholinguistic evidence.

Nevertheless, some interesting preliminary findings have emerged. For example, some of the grammatical sentences that the SRN finds the hardest to predict are also the sentences that humans find the hardest to understand (e.g., center embedded structures like 'the mouse the cat the dog bit chased ate the cheese') (Weckerly & Elman, 1992). These are sequences that place maximal load on encoding information in the network's internal recurrent loop, suggesting that recurrence may be a key computational primitive in language processing. Moreover, when the prediction task is replaced by a comprehension task (such as predicting the agent/patient status of the nouns in the sentence), the results are again suggestive.

Rather than building a syntactic structure for the whole sentence as a symbolic parser might, the network focuses on the predictability of lexical cues for identifying various syntactic structures (consistent with Bates and MacWhinney's Competition model of language development; Bates & MacWhinney, 1989). The salience of lexical cues that each syntactic structure exploits and the processing load that each structure places on the recurrent loop makes them differentially vulnerable under damage. Here, neuropsychological findings from language breakdown and developmental language disorders have tended to follow the predictions of the connectionist account in the relative impairments that each syntactic construction should show (Dick et al., 2001; 2004; Thomas & Redington, 2004).

For more recent work and discussion of the use of SRNs in syntax processing, see Mayberry, Crocker, and Knoeferle (2005), Miikkulainen and Mayberry (1999), Morris, Cottrell, and Elman (2000), Rohde (2002), and Sharkey, Sharkey, and Jackson (2000). Lastly, the impact of SRNs has not been restricted to language. These models have been usefully applied to other areas of cognition where sequential information is important. For example, Botvinick and Plaut (2004) have shown how this architecture can capture the control of routine sequences of actions without the need for schema hierarchies.

In sum, then, Elman's work demonstrates how simple connectionist architectures can learn statistical regularities over temporal sequences. These systems may indeed be sufficient to produce many of the behaviors that linguists have described with grammatical rules. However, in the connectionist system, the underlying primitives are context-sensitive representations of words and trajectories of activation through recurrent circuits.

4. Related models

Before considering the wider impact of connectionism on theories of cognition, we should note a number of other related approaches.

4.1 Cascade-correlation and incremental neural network algorithms.

Backpropagation networks specify input and output representations, while in self-organizing networks only the inputs are specified. These networks therefore include some number of internal processing units whose activation states are determined by the learning algorithm. The number of internal units and their organization (e.g., into layers) plays an important role in determining the complexity of the problems or categories that the network can learn. In pattern associator networks, too few units and the network will fail to learn; in self-organizing networks, too few output units and the network will fail to provide good discrimination between the categories in the training set. How does the modeler select in advance the appropriate number of internal units? Indeed, for a cognitive model, should this be a decision that the modeler gets to make?

For pattern associator networks, the cascade correlation algorithm (Fahlman & Lebiere, 1990) addresses this problem by starting with a network that has no hidden units and then adding in these resources during learning as it becomes necessary in order to carry on improving on the task. New hidden units are added with weights from the input layer tailored so that the unit's activation correlates with network error – i.e., the new unit responds to parts of the problem on which the network is currently doing poorly. New hidden units can also take input from existing hidden units, thereby creating detectors for higher order features in the problem space.

The cascade correlation algorithm has been widely used for studying cognitive development (Mareschal & Shultz, 1996; Shultz, 2003; Westermann, 1998), for example in simulating children's performance in Piagetian reasoning tasks (see Section 5.2). The algorithm makes links with the *constructivist* approach to development (Quartz, 1993; Quartz & Sejnowski, 1997), which argues that increases in the complexity of children's cognitive abilities are best explained by the recruitment of additional neurocomputational resources with age and experience. Related models that also use this 'incremental' approach to building network architectures can be found in the work of Carpenter and Grossberg (Adaptive Resonance Theory; e.g., Carpenter & Grossberg, 1987a,b) and in the work of Love and colleagues (e.g., Love, Medin, & Gureckis, 2004).

4.2 Mixture of experts models

The preceding sections assume that only a single architecture is available to learn each problem. However, it may be that multiple architectures are available to learn a given problem, each with different computational properties. Which architecture will end up learning the problem? Moreover, what if a cognitive domain can be broken down into different parts, for example in the way that the English past tense problem comprises regular and irregular verbs – could different computational components end up learning the different parts of the problem? The mixture-of-experts approach considers ways in which learning could take place in just such a system with multiple components available (Jacobs, Jordan, Nowlan, & Hinton, 1991). In these models, functionally specialized structures can emerge as a result of learning, in the circumstance where the computational properties of the different components happen

to line up with the demands presented by different parts of the problem domain (so called *structure-function correspondences*).

During learning, mixture-of-experts algorithms typically permit the multiple components to compete with each other to deliver the correct output for each input pattern. The best performer is then assigned the pattern and allowed to learn it. The involvement of each component during functioning is controlled by a gating mechanism. Mixture-of-experts models are one of several approaches that seek to explain the origin of functionally specialized processing components in the cognitive system (see Elman et al., 1996; Jacobs, 1999; Thomas & Richardson, 2006, for discussion). An example of the application of mixture of experts can be found in a developmental model of face and object recognition, where different 'expert' mechanisms come to specialize in processing visual inputs that correspond to faces and those that correspond to objects (Dailey & Cottrell, 1999). The emergence of this functional specialization can be demonstrated by damaging each expert in turn and showing a double dissociation between face and object recognition in the two components of the model (see Section 5.3). Similarly, Thomas and Karmiloff-Smith (2002) showed how a mixture-of-experts model of English past tense could produce emergent specialization of separate mechanisms to regular and irregular verbs, respectively (see also Westermann, 1998, for related work with a constructivist network).

4.3 Hybrid models

The success of mixture-of-experts models suggests that when two or more components are combined within a model, it can be advantageous for the computational properties of the components to differ. Where the properties of the

components are radically different, for example involving the combination of symbolic (rule-based) and connectionist (associative, similarity-based) architectures, the models are sometimes referred to as 'hybrid'. The use of hybrid models is inspired by the observation that some aspects of human cognition seem better described by rules (e.g., syntax, reasoning) while some seem better described by similarity (e.g., perception, memory). We have previously encountered the debate between symbolic and connectionist approaches (see Section 2.3) and the proposal that connectionist architectures may serve to implement symbolic processes (e.g., Touretzky & Hinton, 1988). The hybrid systems approach takes the alternative view that connectionist and symbolic processing principles should be combined within the same model, taking advantage of the strengths of each computational formalism. A discussion of this approach can be found in Sun (2002a, b). Example models include CONSYDERR (Sun, 1995) and CLARION (Sun & Peterson, 1998) and ACT-R (Anderson & Lebiere, 1998).

An alternative to a truly hybrid approach is to develop a multi-part connectionist architecture that has components that employ different representational formats. For example, in a purely connectionist system, one component might employ distributed representations that permit different degrees of similarity between activation patterns, while a second component employs localist representations in which there is no similarity between different representations. Behavior is then driven by the interplay between two associative components that employ different similarity structures. One example of the hybrid model in this weaker sense is the complementary learning systems model of McClelland, McNaughton and O'Reilly (1995), which employs localist representations to encode individual episodic memories but distributed representations to encode general semantic memories.

Hybrid developmental models may offer new ways to conceive of the acquisition of concepts. For example, the cognitive domain of *number* may be viewed as hybrid in the sense that it combines the similarity-based representations of quantity and the localist representations of number facts (such as the order of number labels in counting) and object individuation. Carey and Sarnecka (2006) argue that a hybrid multi-component system of this nature could acquire the concept of *positive integers* even though such concept could not be acquired by any single component of the system on its own.

4.4 Bayesian Graphical Models

The use of Bayesian methods of inference in graphical models, including causal graphical models, has recently been embraced by a number of cognitive scientists (Chater, Tenenbaum & Yuille, 2006; Gopnik et al, 2004). This approach stresses how it may be possible to combine prior knowledge in the form of a set of explicit alternative graph structures and constraints on the complexity of such structures with Bayesian methods of inference to select the best type of representation of a particular data set (e.g., lists of facts about many different animals); and within that, to select the best specific instantiation of a representation of that type (Tenenbaum, Griffiths, & Kemp, 2006). These models are useful contributions to our understanding, particularly because they allow explicit exploration of the role of prior knowledge in the selection of a representation of the structure present in each data set. It should be recognized, however, that such models are offered as characterizations of learning at Marr's "Computational Level" and as such they do not specify the representations and processes that are actually employed when people learn. These models do raise questions for connectionist research that does address such questions, however.

Specifically, the work provides a benchmark against which connectionist approaches might be tested for their success in learning to represent the structure from a data set, and in using such a structure to make inferences consistent with optimal performance according to a Bayesian approach within a graphical model framework. More substantively, the work raises questions about whether or not optimization depends on the explicit representation of alternative structured representations, or whether an approximation to such structured representations can arise without their pre-specification. For an initial examination of these issues as they arise in the context of causal inference, see McClelland and Thompson (2007).

5. Connectionist influences on cognitive theory

Connectionism offers an *explanation* of human cognition because instances of behavior in particular cognitive domains can be explained with respect to set of general principles (parallel distributed processing) and the conditions of the specific domains. However, from the accumulation of successful models, it is also possible to discern a wider influence of connectionism on the nature of theorizing about cognition, and this is perhaps a truer reflection of its impact. How has connectionism made us think differently about cognition?

5.1 Knowledge versus processing

One area where connectionism has changed the basic nature of theorizing is memory. According to the old model of memory based on the classical computational metaphor, the information in long-term memory (e.g., on the hard disk) has to be moved into working memory (the CPU) for it to be operated on, and the long-term memories are laid down via a domain-general buffer of short-term memory (RAM). In this type of system, it is relatively easy to shift informational content between different systems, back and forth between central processing and short and long-term stores. Computation is predicated on variables: the same binary string can readily be instantiated in different memory registers or encoded onto a permanent medium.

By contrast, knowledge is hard to move about in connectionist networks because it is encoded in the weights. For example, in the past tense model, knowledge of the past tense rule 'add –ed' is distributed across the weight matrix of the connections between input and output layers. The difficulty in portability of knowledge is inherent in the principles of connectionism – Hebbian learning alters connection strengths to reinforce desirable activation states in connected units, tying

knowledge to structure. If we start from the premise that knowledge will be very difficult to move about in our information processing system, what kind of cognitive architecture do we end up with? There are four main themes.

First, we need to distinguish between two different ways in which knowledge can be encoded: *active* and *latent* representations (Munakata & McClelland, 2003). Latent knowledge corresponds to the information stored in the connection weights from accumulated experience. By contrast, active knowledge is information contained in the current activation states of the system. Clearly the two are related, since the activation states are constrained by the connection weights. But, particularly in recurrent networks, there can be subtle differences. Active states contain a trace of recent events (how things are at the moment) while latent knowledge represents a history of experience (how things tend to be). Differences in the ability to maintain the active states (e.g., in the strength of recurrent circuits) can produce errors in behavior where the system lapses into more typical ways of behaving (Munakata, 1998; Morton & Munakata, 2002).

Second, if information does need to be moved around the system, for example from a more instance-based (episodic) system to a more general (semantic) system, this will require special structures and special (potentially time consuming) processes. Thus McClelland, McNaughton, and O'Reilly (1995) proposed a dialogue between separate stores in hippocampus and neocortex to gradually transfer knowledge from episodic to semantic memory. French, Ans and Rousset (2001) proposed a special method to transfer knowledge between two memory systems: internally generated noise produces 'pseudopatterns' from one system that contain the central tendencies of its knowledge; the second memory system is then trained with this extracted knowledge to effect the transfer.

Third, information will be processed in the same substrate where it is stored. Therefore, long-term memories will be active structures and will perform computations on content. An external strategic control system plays the role of differentially activating the knowledge in this long-term system that is relevant to the current context. In anatomical terms, this distinction broadly corresponds to frontal/anterior (strategic control) and posterior (long-term) cortex. The design means, somewhat counter-intuitively, that the control system has no content. Rather, the control system contains placeholders that serve to activate different regions of the long-term system. The control system may contain plans (sequences of placeholders) and it may be involved in learning abstract concepts (using a placeholder to temporarily co-activate previously unrelated portions of long-term knowledge while Hebbian learning builds an association between them) but it does not contain content in the sense of a domain-general working memory. The study of frontal systems then becomes an exploration of the activation dynamics of these placeholders and their involvement in learning (see, e.g., work by Davelaar & Usher, 2002; Haarmann & Usher, 2001; O'Reilly, Braver, & Cohen, 1999; Usher & McClelland, 2001).

Similarly, connectionist research has explored how activity in the control system can be used to modulate the efficiency of processing elsewhere in the system, for instance to implemented selective attention. For example, Cohen, Dunbar, and McClelland (1990) demonstrated how task units could be used to differentially modulate word naming and color naming processing channels in a model of the color-word Stroop task. In this model, latent knowledge interacted with the operation of task control, so that it was harder to selectively attend to color naming and ignore information from the more practiced word-naming channel than vice versa. This work was later extended to demonstrate how deficits in the strategic control system (pre-

frontal cortex) could lead to problems in selective attention in disorders like schizophrenia (Cohen & Servan-Schreiber, 1992).

Lastly, the connectionist perspective on memory alters how we conceive of *domain generality* in processing systems. It is unlikely that there are any domain-general processing systems that serve as a 'Jack of all trades', i.e., that can move between representing the content of multiple domains. However, there may be domain-general systems that are involved in modulating many disparate processes without taking on the content of those systems, what we might call a system with 'a finger in every pie'. Meanwhile, short-term or working memory (as exemplified by the active representations contained in the recurrent loop of a network) is likely to exist as a devolved panoply of discrete systems, each with its own content-specific loop. For example, research in the neuropsychology of language now tends to support the existence of separate working memories for phonological, semantic, and syntactic information (see MacDonald & Christiansen, 2002, for discussion of these arguments).

5.2 Cognitive development

A key feature of PDP models is the use of a learning algorithm for modifying the patterns of connectivity as a function of experience. Compared to symbolic, rule-based computational models, this has made them a more sympathetic formalism for studying cognitive development (Elman et al., 1996). The combination of domain-general processing principles, domain-specific architectural constraints, and structured training environments has enabled connectionist models to give accounts of a range of developmental phenomena. These include infant category development,

language acquisition and reasoning in children (see Mareschal & Thomas, 2007, for a recent review).

Connectionism has become aligned with a resurgence of interest in statistical learning, and a more careful consideration of the information available in the child's environment that may feed their cognitive development. One central debates revolves around how children can become 'cleverer' as they get older, appearing to progress through qualitatively different stages of reasoning. Connectionist modeling of the development of children's reasoning was able to demonstrate that continuous incremental changes in the weight matrix driven by algorithms such as backpropagation can result in non-linear changes in surface behavior, suggesting that the stages apparent in behavior may not necessarily be reflected in changes in the underlying mechanism (e.g., McClelland, 1989). Other connectionists have argued that algorithms able to supplement the computational resources of the network as part of learning may also provide an explanation for the emergence of more complex forms of behavior with age (e.g., cascade correlation; see Shultz, 2003).

The key contribution of connectionist models in the area of developmental psychology has been to specify detailed, implemented models of transition mechanisms that demonstrate how the child can move between producing different patterns of behavior. This was a crucial addition to a field that has accumulated vast amounts of empirical data cataloguing what children are able to do at different ages. The specification of mechanism is also important to counter some strongly empiricist views that simply identifying statistical information in the environment suffices as an explanation of development; instead, it is necessary to show how a mechanism could use this statistical information to acquire some cognitive capacity. Moreover, when connectionist models are applied to development, it often becomes apparent that

passive statistical structure is not the key factor; rather, the relevant statistics are in the transformation of the statistical structure of the environment to the output or the behavior that is relevant to the child, thereby appealing to notions like the regularity, consistency, and frequency of input-output mappings.

Recent connectionist approaches to development have begun to explore how the computational formalisms may change our understanding of the nature of the knowledge that children acquire. For example, Mareschal et al. (2007) argue that many mental representations of knowledge are partial (i.e., capture only some task relevant dimensions); the existence of explicit language may blind us to the fact that there could be a limited role for truly abstract knowledge in the normal operation of the cognitive system (see Westermann et al., 2007). Current work also explores the computational basis of critical or sensitive periods in development, uncovering the mechanisms by which the ability to learn appears to reduce with age (e.g., McClelland et al., 1999; Thomas & Johnson, 2006).

5.3 The study of acquired disorders in cognitive neuropsychology

Traditional cognitive neuropsychology of the 1980s was predicated on the assumption of underlying modular structure, i.e., that the cognitive system comprises a set of independently functioning components. Patterns of selective cognitive impairment after acquired brain damage could then be used to construct models of normal cognitive function. The traditional models comprised box-and-arrow diagrams that sketched out rough versions of cognitive architecture, informed both by the patterns of possible selective deficit (which bits can fail independently) and by a task analysis of what the cognitive system probably has to do.

In the initial formulation of cognitive neuropsychology, caution was advised in attempting to infer cognitive architecture from behavioral deficits, since a given pattern of deficits might be consistent with a number of underlying architectures (Shallice, 1988). It is in this capacity that connectionist models have been extremely useful. They have both forced more detailed specification of proposed cognitive models via implementation and also permitted assessment of the range of deficits that can be generated by damaging these models in various ways. For example, models of reading have demonstrated that the ability to decode written words into spoken words and recover their meanings can be learned in a connectionist network; and when this network is damaged by, say, lesioning connection weights or removing hidden units, various patterns of acquired dyslexia can be simulated (e.g., Plaut et al., 1996; Plaut & Shallice, 1994). Connectionist models of acquired deficits have grown to be an influential aspect of cognitive neuropsychology and have been applied to domains such as language, memory, semantics, and vision (see Cohen, Johnstone & Plunkett, 2000, for examples).

Several ideas have gained their first or clearest grounding via connectionist modeling. One of these ideas is that patterns of breakdown can arise from the statistics of the problem space (i.e., the mapping between input and output) rather than from structural distinctions in the processing system. In particular, connectionist models have shed light on a principal inferential tool of cognitive neuropsychology, the *double dissociation*. The line of reasoning argues that if in one patient, ability A can be lost while ability B is intact, and in a second patient, ability B can be lost while ability A is intact, then the two abilities may be generated by independent underlying mechanisms. In a connectionist model of category-specific impairments of semantic memory, Devlin et al. (1997) demonstrated that a single undifferentiated network

trained to produce two behaviors could show a double dissociation between them simply as a consequence of different levels of damage. This can arise because the mappings associated with the two behaviors lead them to have different sensitivity to damage. For a small level of damage, performance on A may fall off quickly while performance on B declines more slowly; for a high level of damage, A may be more robust than B. The reverse pattern of relative deficits implies nothing about structure.

Connectionist researchers have often set out to demonstrate that, more generally, double dissociation methodology is a flawed form of inference, on the grounds that such dissociations arise relatively easily from parallel distributed architectures where function is spread across the whole mechanism (e.g., Plunkett & Bandelow, 2006; Plunkett & Juola, 1998). However, on the whole, when connectionist models show robust double dissociations between two behaviors (for equivalent levels of damage applied to various parts of the network and over many replications), it does tend to be because different internal processing structures (units or layers or weights) or different parts of the input layer or different parts of the output layer are differentially important for driving the two behaviors – that is, there is specialization of function. Connectionism models of breakdown have, therefore, tended to support the traditional inferences. Crucially, however, connectionist models have greatly improved our understanding of what modularity might look like in a neurocomputational system: a partial rather than an absolute property; a property that is the consequence of a developmental process where emergent specialization is driven by *structure-function correspondences* (the ability of certain parts of a computational structure to learn certain kinds of computation better than other kinds; see Section 4.2); and a property that must now be complemented by concepts such as

division of labor, degeneracy, interactivity, and redundancy (see Thomas & Karmiloff-Smith, 2002a; Thomas et al., 2006, for discussion).

5.4 The origins of individual variability and developmental disorders

In addition to their role in studying acquired disorders, the fact that many connectionist models learn their cognitive abilities makes them an ideal framework within which to study *developmental disorders*, such as autism, dyslexia, and specific language impairment (Mareschal et al., 2007; Joanisse & Seidenberg, 2003; Thomas & Karmiloff-Smith, 2002b, 2003a, 2005). Where models of normal cognitive development seek to study the 'average' child, models of atypical development explore how developmental profiles may be disrupted. Connectionist models contain a number of constraints (architecture, activation dynamics, input and output representations, learning algorithm, training regime) that determine the efficiency and outcome of learning. Manipulations to these constraints produce candidate explanations for impairments found in developmental disorders or to the impairments caused by exposure to atypical environments such as in cases of deprivation.

In the 1980s and 1990s, many theories of developmental deficits employed the same explanatory framework as adult cognitive neuropsychology. There was search for specific developmental deficits or dissociations, which were then explained in terms of the failure of individual modules to development. However, as Karmiloff-Smith (1998) and Bishop (1997) pointed out, most of the developmental deficits were actually being explained with reference to non-developmental, static, and sometimes adult models of normal cognitive structure. Karmiloff-Smith (1998) argued that the causes of developmental deficits of a genetic origin are likely to lie in changes to low-level neurocomputational properties that only exert their influence on cognition via an

extended atypical developmental process (see also Elman et al., 1996). Connectionist models provide the ideal forum to explore the thesis that an understanding of the constraints on the developmental process is essential for generating accounts of developmental deficits.

The study of atypical variability also prompts a consideration of what causes variability *within the normal range*, otherwise known as individual differences or intelligence. Are differences in intelligence caused by variation in the same computational parameters that can cause disorders? Are some developmental disorders just the extreme lower end of the normal distribution or are they qualitatively different conditions? What computational parameter settings are able to produce above average performance? Connectionism has begun to take advantage of the accumulated body of models of normal development to consider the wider question of cognitive variation in parameterized computational models (Thomas & Karmiloff-Smith, 2003b).

5.5 Future directions

The preceding sections indicate the range and depth of influence of connectionism on contemporary theories of cognition. Where will connectionism go next? Necessarily, connectionism began with simple models of individual cognitive processes, focusing on those domains of particular theoretical interest. This piecemeal approach generated explanations of individual cognitive abilities using bespoke networks, each containing its own pre-determined representations and architecture. In the future, one avenue to pursue is how these models fit together in the larger cognitive system – for example, to explain how the past tense network described in Section 3.2 might link up with the sentence-processing model described in Section 3.3 to process past tenses as they

arise in sentences. A further issue is to address the developmental origin of the architectures that are postulated. What processes specify the parts of the cognitive system to perform the various functions and how do these sub-systems talk to each other, both across development and in the adult state? Improvements in computational power will aid more complex modeling endeavors. Nevertheless, it is worth bearing in mind that increasing complexity creates a tension with the primary goals of modeling – simplification and understanding. It is essential that we understand why more complicated models function as they do or they will merely become interesting artifacts (see Elman, 2005; Thomas, 2004, for further discussion).

In terms of its relation with other disciplines, a number of future influences on connectionism are discernible. Connectionism will be affected by the increasing appeal to *Bayesian probability theory* in human reasoning. In Bayesian theory, new data are used to update existing estimates of the most likely model of the world. Work has already begun to relate connectionist and Bayesian accounts, for example in the domain of causal reasoning in children (McClelland & Thompson, 2007). In some cases, connectionism may offer alternative explanations of the same behavior, in others it may be viewed as an implementation of a Bayesian account (see Section 3.1). Connectionism will continue to have a close relation to *neuroscience*, perhaps seeking to build more neural constraints into its computational assumptions (O'Reilly & Munakata, 2000). Many of the new findings in cognitive neuroscience are influenced by *functional brain imaging techniques*. It will be important, therefore, for connectionism to make contact with these data, either via systems-level modeling of the interaction between sub-networks in task performance, or in exploring the implications of the subtraction methodology as a tool for assessing the behavior of distributed interactive systems. The increasing influence of brain imaging foregrounds

the relation of cognition to the neural substrate; it depends on how seriously one takes the neural plausibility of connectionist models as to whether an increased focus on the substrate will have particular implications for connectionism over and above any other theory of cognition.

Connectionist approaches to individual differences and developmental disorders suggest that this modeling approach has more to offer in considering the computational causes of variability. Research in *behavioral genetics* argues that a significant proportion of behavioral variability is genetic in origin (Bishop, 2006; Plomin, Owen & McGuffin, 1994). However, the neurodevelopmental mechanisms by which genes produce such variation are largely unknown. While connectionist cognitive models are not neural, the fact that they incorporate neurally inspired properties may allow them to build links between behavior (where variability is measured) and the substrate on which genetic effects act. In the future, connectionism may therefore help to rectify a major shortcoming in our attempts to understand the relation of the human genome to the human mind – the omission of the cognitive level.

6 Conclusions

In this chapter, we have considered the contribution of connectionist modeling to our understanding of cognition. Connectionism was placed in the historical context of 19th century associative theories of mental processes and 20th century attempts to understand the computations carried out by networks of neurons. The key properties of connectionist networks were then reviewed and particular emphasis placed on the use of learning to build the microstructure of these models. The core connectionist themes include the following: (1) that processing is simultaneously influenced by multiple sources of information at different levels of abstraction, operating via soft constraint satisfaction; (2) that representations are spread across multiple simple processing units operating in parallel; (3) that representations are graded, context-sensitive, and the emergent product of adaptive processes; (4) that computation is similarity-based and driven by the statistical structure of problem domains, but it can nevertheless produce rule-following behavior. We illustrated the connectionist approach via three landmarks models, the Interactive Activation model of letter recognition (McClelland & Rumelhart, 1981), the past tense model (Rumelhart & McClelland, 1986), and simple recurrent networks for finding structure in time (Elman, 1990). Apart from its body of successful individual models, connectionist theory has had a widespread influence on cognitive theorizing, and this influence was illustrated by considering connectionist contributions to our understanding of memory, cognitive development, acquired cognitive impairments, and developmental deficits. Finally, we peeked into the future of connectionism, arguing that its relationships with other fields in the cognitive sciences are likely to guide its future contribution to understanding the mechanistic basis of thought.

Acknowledgements

This work was supported by British Academy Grant SG – 40400 and UK Medical Research Council Grant G0300188 to Michael Thomas, and National Institute of Mental Health Centre Grant P50 MH64445, James L. McClelland, Director.

References

Ackley, D. H., Hinton, G. E., & Sejnowski, T. J. (1985). A learning algorithm for Boltzmann machines. *Cognitive Science, 9*, 147-169.

Aizawa, K. (2004). 'History of connectionism'. In C. Eliasmith (Ed.), *Dictionary of Philosophy of Mind*. October 4, 2006, http://philosophy.uwaterloo.ca/MindDict/connectionismhistory.html

Anderson, J. & Rosenfeld, E. (1988). *Neurocomputing: foundations of research*. MIT Press: Cambridge, MA.

Anderson, J. A. (1977). Neural models with cognitive implications. In D. LaBerge & S. J. Samuels (Eds.), *Basic processes in reading perception and comprehension* (pp. 27-90). Hillsdale, NJ: Erlbaum.

Anderson, J., & Lebiere, C. (1998). *The atomic components of thought*. Lawrence Erlbaum Associates, Mahwah, NJ.

Bates, E. & MacWhinney, B. (1989). Functionalism and the competition model. In B. MacWhinney and E. Bates (Eds.), *The crosslinguistic study of language processing* (pp. 3-37). New York: Cambridge University Press.

Bechtel, W. & Abrahamsen, A. (1991). *Connectionism and the mind*. Blackwell, Oxford.

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

Bishop, D. V. M. (1997). Cognitive neuropsychology and developmental disorders: Uncomfortable bedfellows. *Quarterly Journal of Experimental Psychology*, *50A*, 899-923.

Bishop, D. V. M. (2006). Developmental cognitive genetics: how psychology can inform genetics and vice versa. *Quarterly Journal of Experimental Psychology*, *59(7)*, 1153-1168.

Botvinick, M. & Plaut, D. C. (2004). Doing without schema hierarchies: A recurrent connectionist approach to normal and impaired routine sequential action. *Psychological Review*, *111*, 395-429.

Burton, A. M., Bruce, V., & Johnston, R.A. (1990). Understanding face recognition with an interactive activation model. *British Journal of Psychology*, *81*, 361–380.

Bybee, J. and McClelland, J. L. (2005). Alternatives to the combinatorial paradigm of linguistic theory based on domain general principles of human cognition. *The Linguistic Review, 22(2-4)*, 381-410.

Carey, S. & Sarnecka, B. W. (2006). The development of human conceptual representations: A case study. In Y. Munakata & M. H. Johnson (Eds.), *Processes of change in brain and cognitive development: Attention and Performance XXI*, (pp. 473-496). Oxford: Oxford University Press.

Carpenter, G. A., & Grossberg, S. (1987a). A massively parallel architecture for a self-organizing neural pattern recognition machine. *Computer Vision, Graphics and Image Processing*, *37*, 54-115.

Carpenter, G. A., & Grossberg, S. (1987b). ART2: Self-organization of stable category recognition codes for analog input patterns. *Applied Optics*, *26*, 4919-4930.

Chater, N., Tenenbaum, J. B., and Yuille, A. (2006). Probabilistic models of cognition: Conceptual foundations. *Trends in Cognitive Sciences 10*(7), 287-291

Christiansen, M. H. & Chater, N. (2001). *Connectionist psycholinguistics*. Westport, CT.: Ablex.

Cohen, G., Johnstone, R. A., & Plunkett, K. (2000). *Exploring cognition: damaged brains and neural networks*. Psychology Press: Hove, Sussex.

Cohen, J. D., Dunbar, K., & McClelland, J. L. (1990). On the control of automatic processes: A parallel distributed processing account of the Stroop effect. *Psychological Review*, *97*, 332-361

Cohen, J. D., Servan-Schreiber, D. (1992). Context, cortex, and dopamine: A connectionist approach to behavior and biology in schizophrenia. *Psychological Review*, *99*, 45-77.

Dailey, M. N., & Cottrell, G. W. (1999). Organization of face and object recognition in modular neural networks. *Neural Networks*, *12*, 1053-1074.

Davelaar, E. J., & Usher, M. (2002). An activation-based theory of immediate item memory. In J. A. Bullinaria, & W. Lowe (Eds.), *Proceedings of the Seventh Neural Computation and Psychology Workshop: Connectionist Models of Cognition and Perception*. Singapore: World Scientific.

Davies, M. (2005). Cognitive science. In F. Jackson & M. Smith (Eds.), *The Oxford Handbook of Contemporary Philosophy*. Oxford: Oxford University Press.

Devlin, J., Gonnerman, L., Andersen, E., & Seidenberg, M.S. (1997). Category specific semantic deficits in focal and widespread brain damage: A computational account. *Journal of Cognitive Neuroscience*, *10*, 77-94.

Dick, F., Bates, E., Wulfeck, B., Aydelott, J., Dronkers, N., & Gernsbacher, M.A. 2001. Language deficits, localization, and grammar: Evidence for a distributive model of language breakdown in aphasic patients and neurologically intact individuals. *Psychological Review*, *108(3)*, 759-788.

Dick, F., Wulfeck, B., Krupa-Kwiatkowski, & Bates (2004). The development of complex sentence interpretation in typically developing children compared with children with specific language impairments or early unilateral focal lesions. *Developmental Science*, *7(3)*, 360-377.

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

Elman, J. L. (1991) Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning*, *7*, 195-224

Elman, J. L. (1993). Learning and development in neural networks: the importance of starting small. *Cognition*, *48*, 71-99.

Elman, J. L. (2005). Connectionist models of cognitive development: Where next? *Trends in Cognitive Sciences*, *9*, 111-117.

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: MIT Press:

Ervin, S. M. (1964). Imitation and structural change in children's language. In E. H. Lenneberg (Ed.), *New directions in the study of language*. Cambridge, MA: MIT Press.

Fahlman, S. & Lebiere, C. (1990). The cascade correlation learning architecture. In D. Touretzky (Ed.), *Advances in neural information processing 2* (pp. 524-532). Morgan Kauffman, Los Altos, CA.

Feldman, J. A. (1981). A connectionist model of visual memory. In G. E. Hinton & J. A. Anderson (Eds.), *Parallel models of associative memory* (pp. 49-81). Hillsdale, NJ: Erlbaum.

Fodor, J. A. & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *78*, 3-71.

French, R. M., Ans, B., & Rousset, S. (2001). Pseudopatterns and dual-network memory models: Advantages and shortcomings. In R. French & J. Sougné (Eds.), *Connectionist Models of Learning, Development and Evolution* (pp. 13-22). London: Springer.

Freud, S. (1895). Project for a scientific psychology. In J. Strachey (Ed.), *The standard edition of the complete psychological works of Sigmund Freud*. London: The Hogarth Press and the Institute of Psycho-Analysis.

Goebel, R., & Indefrey, P. (2000). A recurrent network with short-term memory capacity learning the German –s plural. In P. Broeder & J. Murre (Eds.), *Models of language acquisition: Inductive and deductive approaches* (pp. 177-200). Oxford: Oxford University Press.

Gopnik, Alison, Clark Glymour, David M. Sobel, Laura E. Schulz, Tamar Kushnir, & David Danks. 2004. "A Theory of Causal Learning in Children: Causal Maps and Bayes Nets." *Psychological Review*, 111 (1): 3-32.

Green, D. C. (1998). Are connectionist models theories of cognition? *Psycoloquy*, *9(4)*.

Grossberg, S. (1976). Adaptive pattern classification and universal recoding: Parallel development and coding of neural feature detectors. *Biological Cybernetics*, *23*, 121-134.

Haarmann, H. & Usher, M. (2001). Maintenance of semantic information in capacity limited item short-term memory. *Psychonomic Bulletin & Review*, *8*, 568-578.

Hebb, D. O. (1949). *The organization of behavior: A neuropsychological approach*. New York: John Wiley & Sons.

Hinton, G. E. (1989). Deterministic Boltzmann learning performs steepest descent in weight-space. *Neural Computation*, *1*, 143-150.

Hinton, G. E. & Anderson J. A. (1981). *Parallel models of associative memory*. Hillsdale, NJ: Erlbaum.

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science*, *Vol. 313. no. 5786*, 504 - 507.

Hinton, G. E. & Sejnowski, T. J. (1983, June). Optimal perceptual inference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* Washington, DC.

Hinton, G. E. & Sejnowksi, T. (1986). Learning and relearning in Boltzmann machines. In D. Rumelhart & J. McClelland (Eds.), *Parallel distributed processing, Vol. 1* (pp. 282-317). MIT Press: Cambridge, MA.

Hoeffner, J. H. & McClelland, J. L. (1993). Can a perceptual processing deficit explain the impairment of inflectional morphology in developmental dysphasia? A computational investigation. In E. V. Clark (Ed.), *Proceedings of the 25th Child language research forum* (pp. 38–49). Stanford, CA: Center for the Study of Language and Information.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Science USA*, *79*, 2554-2558.

Houghton, G. (2005). *Connectionist models in cognitive psychology*. Hove: Psychology Press.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., & Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural Computation*, *3*, 79-87.

Jacobs, R.A. (1999). Computational studies of the development of functionally specialized neural modules. *Trends in Cognitive Sciences*, *3*, 31-38.

James, W. (1890). *Principles of psychology*. New York, NY: Holt.

Joanisse, M. F., & Seidenberg, M. S. (1999). Impairments in verb morphology following brain injury: A connectionist model. *Proceedings of the National Academy of Science USA*, *96*, 7592-7597.

Joanisse, M. F., & Seidenberg, M. S. (2003). Phonology and syntax in specific language impairment: Evidence from a connectionist model. *Brain and Language*, *86*, 40-56.

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

Juola, P. & Plunkett, K. (1998). Why double dissociations don't mean much. In *Proceedings of the 20th Annual Conference of the Cognitive Science Society* (pp. 561-566). Hillsdale, NJ: Erlbaum.

Karmiloff-Smith, A. (1998). Development itself is the key to understanding developmental disorders. *Trends in Cognitive Sciences*, *2*, 389-398.

Kohonen, T. (1984). *Self-organization and associative memory*. Berlin: Springer-Verlag.

Kuczaj, S. A. (1977). The acquisition of regular and irregular past tense forms. *Journal of Verbal Learning and Verbal Behavior*, *16*, 589-600.

Lashley, K. S. (1929). *Brain mechanisms and intelligence: A quantitative study of injuries to the brain*. New York: Dover Publications, Inc.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review*, *111*, 309-332.

MacDonald, M. C., & Christiansen, M. H. (2002). Reassessing working memory: A comment on Just & Carpenter (1992) and Waters & Caplan (1996). *Psychological Review*, *109*, 35-54.

MacKay, D. J. (1992). A practical Bayesian framework for backpropagation networks. *Neural Computation, 4,* 448--472.

Marcus, G. F. (2001). *The algebraic mind: integrating connectionism and cognitive science*. Cambridge, MA: MIT Press

Marcus, G. F., Brinkmann, U., Clahsen, H., Wiese, R., Woest, A., & Pinker, S. (1995). German inflection: the exception that proves the rule. *Cognitive Psychology*, *29*, 189-256.

Marcus, G., Pinker, S., Ullman, M., Hollander, J., Rosen, T. & Xu, F. (1992). Overregularisation in language acquisition. *Monographs of the Society for Research in Child Development*, *57* (Serial No. 228).

Mareschal, D. & Shultz, T. R. (1996). Generative connectionist architectures and constructivist cognitive development. *Cognitive Development*, *11*, 571-605.

Mareschal, D. & Thomas, M. S. C. (in press). Computational modeling in developmental psychology. *IEEE Transactions on Evolutionary Computation (Special Issue on Autonomous Mental Development)*.

Mareschal, D., Johnson, M., Sirios, S., Spratling, M., Thomas, M. S. C., & Westermann, G. (2007). *Neuroconstructivism: How the brain constructs cognition*. Oxford: Oxford University Press.

Marr, D. & Poggio, T. (1976). Cooperative computation of stereo disparity. *Science*, 194, 283-287.

Marshall R. M., Crocker, M., & Knoeferle, P. (2005). A connectionist model of sentence comprehension in visual worlds. In *Proceedings of the 27th Annual Conference of the Cognitive Science Society* (COGSCI-05, Stresa, Italy), Mahwah, NJ: Erlbaum.

Marr, D. (1982). *Vision*. San Francisco: W.H. Freeman.

McClelland, J. L. (1981). Retrieving general and specific information from stored knowledge of specifics. *Proceedings of the Third Annual Meeting of the*

Cognitive Science Society (pp. 170-172). Hillsdale, NJ: Lawrence Erlbaum Associates.

McClelland, J. L. (1989). Parallel distributed processing: Implications for cognition and development. In M. G. M. Morris (Ed.), *Parallel distributed processing, implications for psychology and neurobiology* (pp. 8-45). Oxford: Clarendon Press.

McClelland, J. L. (1998). Connectionist models and Bayesian inference. In M. Oaksford & N. Chater (Eds.), *Rational Models of Cognition.* Oxford: Oxford University Press. 21-53.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, *18*, 1-86.

McClelland, J. L., McNaughton, B. L., and O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, *102*, 419-457.

McClelland, J. L., Plaut, D. C., Gotts, S. J. and Maia, T. V. (2003). Developing a domain-general framework for cognition: What is the best approach? Commentary on a target article by Anderson and Lebiere. *Behavioral and Brain Sciences, 22,* 611-614.

McClelland, J. L. & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: Part 1. An account of basic findings. *Psychological Review*, *88(5)*, 375-405.

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

McClelland, J. L., Thomas, A. G., McCandliss, B. D., & Fiez, J. A. (1999). Understanding failures of learning: Hebbian learning, competition for representation space, and some preliminary data. In J. A. Reggia, E. Ruppin & D. Glanzman (Eds.), *Disorders of brain, behavior, and cognition: The neurocomputational perspective* (pp. 75-80). Elsevier: Oxford.

McClelland, J. L. and Thompson, R. M. (2007). Using Domain-General Principles to Explain Children's Causal Reasoning Abilities. *Developmental Science, 10*, 333-356.

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, 5, 115-133. Reprinted in Anderson & Rosenfield (1988).

McLeod, P., Plunkett, K. & Rolls, E. T. (1998). *Introduction to connectionist modelling of cognitive processes*. Oxford: Oxford University Press

Meynert, T. (1884). *Psychiatry: A clinical treatise on diseases of the forebrain. Part I. The Anatomy, Physiology and Chemistry of the Brain.* Trans. B. Sachs. New York: G.P. Putnam's Sons.

Miikkulainen, R. & Mayberry, M. R. (1999). Disambiguation and Grammar as Emergent Soft Constraints. In B. MacWhinney (Ed.), *Emergence of language*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Minsky, M., & Papert, S. (1988). *Perceptrons: An introduction to computational geometry*. Cambridge, MA: MIT Press.

Morris, W., Cottrell, G., & Elman, J. (2000). A connectionist simulation of the empirical acquisition of grammatical relations. In Wermter, S. & Sun, R., (Eds.), *Hybrid neural systems*. Springer Verlag, Heidelberg.

Morton, J. (1969). Interaction of information in word recognition. *Psychological Review*, *76*, 165-178.

Morton, J. B., & Munakata, Y. (2002). Active versus latent representations: A neural network model of perseveration, dissociation, and decalage in childhood. *Developmental Psychobiology*, *40*, 255-265.

Movellan, J. R., & McClelland, J. L. (1993). Learning continuous probability distributions with symmetric diffusion networks. *Cognitive Science, 17,* 463-496.

Munakata, Y. (1998). Infant perseveration and implications for object permanence theories: A PDP model of the AB task. *Developmental Science*, *1*, 161-184.

Munakata, Y., & McClelland, J. L. (2003). Connectionist models of development. *Developmental Science*, *6*, 413-429.

O'Reilly, R.C. (1998). Six Principles for Biologically-Based Computational Models of Cortical Cognition. *Trends in Cognitive Sciences, 2,* 455-462.

O'Reilly, R. C. & Munakata, Y. (2000). *Computational explorations in cognitive neuroscience: Understanding the mind by simulating the brain*. Cambridge: MIT Press.

O'Reilly, R. C., Braver, T. S. & Cohen, J. D. (1999). A Biologically based computational model of working memory. In A. Miyake. & P. Shah (Eds.), *Models of working memory: Mechanisms of active maintenance and executive control*. New York: Cambridge University Press.

Petersen, C. & Anderson, J. (1987). A mean field theory learning algorithm for neural networks. *Complex systems*, *1*, 995-1019.

Pinker, S. (1984). *Language learnability and language development*. Cambridge, MA: Harvard University Press.

Pinker, S. (1999). *Words and rules*. London: Weidenfeld & Nicolson

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

Plaut, D. C., & Kello, C. T. (1999). The emergence of phonology from the interplay of speech comprehension and production: A distributed connectionist approach. In B. MacWhinney (Ed.), *The emergence of language* (pp. 381-415). Mahwah, NJ: Erlbaum.

Plaut, D. C., & Shallice, T. (1993). Deep dyslexia: A case study of connectionist neuropsychology. *Cognitive Neuropsychology*, *10*, 377-500.

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

Plaut, D., & McClelland, J. L. (1993). Generalization with componential attractors: Word and nonword reading in an attractor network. In *Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society* (pp. 824—829). Hillsdale, NJ: Lawrence Erlbaum Associates.

Plomin, R., Owen, M. J., & McGuffin, P. (1994). The genetic basis of complex human behaviors. *Science*, *264*, 1733-1739.

Plunkett, K. & Bandelow, S. (2006). Stochastic approaches to understanding dissociations in inflectional morphology. *Brain and Language*, *98*, 194-209.

Plunkett, K., & Marchman, V. (1991). U-shaped learning and frequency effects in a multi-layered perceptron: Implications for child language acquisition. *Cognition*, *38*, 1-60.

Plunkett, K., & Marchman, V. (1993). From rote learning to system building: acquiring verb morphology in children and connectionist nets. *Cognition*, *48*, 21-69.

Plunkett, K., & Marchman, V. (1996). Learning from a connectionist model of the English past tense. *Cognition*, *61*, 299-308.

Plunkett, K., & Nakisa, R. (1997). A connectionist model of the Arabic plural system. *Language and Cognitive Processes*, *12*, 807-836.

Quartz, S. R. (1993). Neural networks, nativism, and the plausibility of constructivism. *Cognition*, *48*, 223–242.

Quartz, S. R. & Sejnowski, T. J. (1997). The neural basis of cognitive development: A constructivist manifesto. *Behavioral and Brain Sciences*, *20*, 537–596.

Rashevsky, N. (1935). Outline of a physico-mathematical theory of the brain. *Journal of General Psychology*, *13*, 82-112.

Reicher, G. M. (1969). Perceptual recognition as a function of meaningfulness of stimulus material. *Journal of Experimental Psychology*, *81*, 274-280.

Rohde, D.L.T. (2002). *A Connectionist model of sentence comprehension and production.* Unpublished PhD thesis, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA.

Rohde, D. L. T., & Plaut, D. C. (1999). Language acquisition in the absence of explicit negative evidence: How important is starting small? *Cognition*, *72*, 67-109.

Rosenblatt, F. (1958). The Perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*, 386–408.

Rosenblatt, F. (1962). *Principles of neurodynamics: Perceptrons and the theory of brain mechanisms*. Washington, D.C.: Spartan Books

Rumelhart, D. E. & McClelland, J. L. (1982). An interactive activation model of context effects in letter perception: Part 2. The contextual enhancement effect and some tests and extensions of the model. *Psychological Review*, *89*, 60-94.

Rumelhart, D. E. & McClelland, J. L. (1986). On learning the past tense of English verbs. In J. L. McClelland, D. E. Rumelhart & the PDP Research Group (Eds.) *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 2: Psychological and Biological Models* (pp. 216-271). Cambridge, MA: MIT Press.

Rumelhart, D. E. and McClelland, J. L. (1985). Levels indeed! *Journal of Experimental Psychology General*, *114(2)*, 193-197.

Rumelhart, D. E., Hinton, G. E. & Williams, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart, J. L. McClelland & the PDP Research Group, *Parallel Distributed Processing: Explorations in the microstructure of cognition, Volume 1: Foundations* (pp. 318-362). MIT Press: Cambridge, MA.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart, J. L. McClelland and The PDP Research Group, *Parallel distributed processing: Explorations in the microstructure of cognition. Vol. 1: Foundations* (pp. 318-362). Cambridge, MA: MIT Press.

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

Rumelhart, D. E., Smolensky, P., McClelland, J. L., & Hinton, G. E. (1986). Schemata and sequential thought processes in PDP models. In J. L. McClelland

& D. E. Rumelhart (Eds.), *Parallel distributed processing, Vol. 2* (pp. 7-57). MIT Press: Cambridge, Mass.

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word segmentation: The role of distributional cues. *Journal of Memory and Language*, *35*, 606-621.

Selfridge, O. G. (1959). Pandemonium: A paradigm for learning. In *Symposium on the Mechanization of Thought Processes, London: HMSO*.

Shallice, T. (1988). *From neuropsychology to mental structure*. Cambridge, UK: Cambridge University Press.

Sharkey, N., Sharkey, A., & Jackson, S. (2000). Are SRNs sufficient for modelling language acquisition. In P. Broeder & J. Murre (Eds.), *Models of language acquisition: Inductive and deductive approaches* (pp. 33-54). Oxford: Oxford University Press.

Shultz, T. R. (2003). *Computational developmental psychology*, MIT Press: Cambridge, Mass.

Smolensky, P. (1988). On the proper treatment of connectionism. *Behavioral and Brain Sciences*, *11*, 1-74.

Spencer, Herbert. (1872). *Principles of psychology (3rd Edition)*. London: Longman, Brown, Green, & Longmans.

Sun, R. (1995). Robust reasoning: integrating rule-based and similarity-based reasoning. *Artificial Intelligence*, *75*, 241-295.

Sun, R. (2002a). Hybrid systems and connectionist implementationalism. In *Encyclopedia of Cognitive Science* (pp. 697-703). Macmillan Publishing Company (Nature Publishing Group).

Sun, R. (2002b). Hybrid connectionist symbolic systems. In M. Arbib (Ed.), *Handbook of Brain Theories and Neural Networks (2nd Edition)*, (pp. 543-547). MIT Press, Cambridge, MA.

Sun, R., & Peterson, T. (1998). Autonomous learning of sequential tasks: experiments and analyses. *IEEE Transactions on Neural Networks*, *9(6)*, 1217-1234.

Thomas, M. S. C. (2005). Characterising compensation. *Cortex*, *41(3)*, 434-442.

Tenenbaum, J. B., Griffiths, T. L., and Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. Trends in Cognitive Sciences 10(7), 309-318.

Thomas, M. S. C., & Johnson, M. H. (2006). The computational modelling of sensitive periods. *Developmental Psychobiology*, *48(4)*, 337-344.

Thomas, M. S. C., & Karmiloff-Smith, A. (2002a). Are developmental disorders like cases of adult brain damage? Implications from connectionist modelling. *Behavioral and Brain Sciences*, *25(6)*, 727-788.

Thomas, M. S. C., & Karmiloff-Smith, A. (2002b). Modelling typical and atypical cognitive development. In U. Goswami (Ed.), *Handbook of Childhood Development* (pp. 575-599). Blackwells Publishers.

Thomas, M. S. C., & Karmiloff-Smith, A. (2003a). Modeling language acquisition in atypical phenotypes. *Psychological Review*, *110(4)*, 647-682.

Thomas, M. S. C., & Karmiloff-Smith, A. (2003b). Connectionist models of development, developmental disorders and individual differences. In R. J. Sternberg, J. Lautrey, & T. Lubart (Eds.), *Models of intelligence: International perspectives*, (p. 133-150). American Psychological Association.

Thomas, M. S. C., & Redington, M. (2004). Modelling atypical syntax processing. In W. Sakas (Ed.), *Proceedings of the First Workshop on Psycho-computational*

models of human language acquisition at the 20th International Conference on Computational Linguistics. Pp. 85-92.

Thomas, M. S. C., & Richardson, F. (2006). Atypical representational change: Conditions for the emergence of atypical modularity. In Y. Munakata & M. H. Johnson (Eds.), *Processes of change in brain and cognitive development: Attention and Performance XXI*, (pp. 315-347). Oxford: Oxford University Press.

Thomas, M. S. C., & Van Heuven, W. (2005). Computational models of bilingual comprehension. In J. F. Kroll & A. M. B. De Groot (Eds.), *Handbook of Bilingualism: Psycholinguistic Approaches* (pp. 202-225). Oxford: Oxford University Press.

Thomas, M. S. C., Forrester, N. A., & Richardson, F. M. (2006). What is modularity good for? In *Proceedings of The 28th Annual Conference of the Cognitive Science Society* (p. 2240-2245), July 26-29, Vancouver, BC, Canada.

Touretzky, D. S., & Hinton, G. E. (1988). A distributed connectionist production system. *Cognitive Science*, *12*, 423-466.

Usher, M., & McClelland, J. L. (2001). On the time course of perceptual choice: The leaky competing accumulator model. *Psychological Review*, *108*, 550-592.

van Gelder, T. (1991). Classical questions, radical answers: Connectionism and the structure of mental representations. In T. Horgan & J. Tienson (Eds.), *Connectionism and the philosophy of mind*. Dordrecht: Kluwer Academic Publishers.

Williams, R. J. and Zipser, D. (1995). Gradient-based learning algorithms for recurrent networks and their computational complexity. In: Y. Chauvin and D.

E. Rumelhart (Eds.) *Back-propagation: Theory, Architectures and Applications*, Hillsdale, NJ: Erlbaum.

Weckerly, J., & Elman, J. L. (1992). A PDP approach to processing center-embedded sentences. In *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.

Westermann, G. (1998). Emergent modularity and U-shaped learning in a constructivist neural network learning the English past tense. In *Proceedings of the 20th Annual Conference of the Cognitive Science Society*, (pp. 1130–1135), Erlbaum, Hillsdale, NJ.

Westermann, G., Mareschal, D., Johnson, M. H., Sirois, S., Spratling, M. W., & Thomas, M. S. C. (2007). Neuroconstructivism. *Developmental Science*, *10*, 75-83.

Xie, X., & Seung, H. S. (2003). Equivalence of backpropagation and contrastive Hebbian learning in a layered network. *Neural Computation*, *15*, 441-454.

Xu, F., & Pinker, S. (1995). Weird past tense forms. *Journal of Child Language*, *22*, 531-556.

Figure Captions

Figure 1. A simplified schematic showing the historical evolution of neural network architectures. Simple binary networks (McCulloch & Pitts, 1943) are followed by 2-layer feedforward networks (perceptrons; Rosenblatt, 1958). Three subtypes then emerge: 3-layer feedforward networks (Rumelhart & McClelland, 1986), competitive or self-organizing networks (e.g., Grossberg, 1976; Kohonen, 1984), and interactive networks (Hopfield, 1982; Hinton & Sejnowksi, 1986). Adaptive interactive networks have precursors in detector theories of perception (Logogen: Morton, 1969; Pandemonium: Selfridge, 1955) and in handwired interactive models (IA: McClelland & Rumelhart, 1981; IAC: McClelland, 1981; Stereopsis: Marr & Poggio, 1976; Necker cube: Feldman, 1981, Rumelhart et al., 1986). Feedforward pattern associators have produced multiple subtypes: for capturing temporally extended activation states, cascade networks in which states monotonically asymptote (e.g., Cohen, Dunbar, & McClelland, 1990) and attractor networks in which states cycle into stable configurations (e.g., Plaut & McClelland, 1993); for processing sequential information, recurrent networks (Jordan, 1986; Elman, 1991); for systems that alter their structure as part of learning, constructivist networks (e.g., cascade correlation: Fahlman & Lebiere, 1990; Shultz, 2003).

Figure 2. Interactive Activation model of context effects in letter recognition (McClelland & Rumelhart, 1981, 1982). Pointed arrows are excitatory connections, circular headed arrows are inhibitory connections. Left: macro view (connections in gray were set to zero in implemented model). Right: micro view for the connections from the feature level to the first letter position for the letters S, W, and F (only

excitatory connections shown) and from the first letter position to the word units SEED, WEED, and FEED (all connections shown).

Figure 3. Two-layer network for learning the mapping between the verb roots and past tense forms of English verbs (Rumelhart & McClelland, 1986). Phonological representations of verbs are initially encoded into a coarse, distributed 'Wickelfeature' representation. Past tenses are decoded from the Wickelfeature representation back to the phonological form. Later connectionist models replaced the dotted area with a three-layer feedforward backpropagation network (e.g., Plunkett & Marchman, 1991, 1993).

Figure 4. Elman's simple recurrent network architecture for finding structure in time (Elman, 1991, 1993). Connections between input and hidden, context and hidden, and hidden and output layers are trainable. Sequences are applied to the network element by element in discrete time steps; the context layer contains a copy of the hidden unit activations on the previous time step transmitted by fixed, 1-to-1 connections.

Figure 5. Trajectory of internal activation states as the SRN processes sentences (Elman, 1993). The data show positions according to the dimensions of a principal components analysis (PCA) carried out on hidden unit activations for the whole training set. Words are indexed by their position in the sequence but represent activation of the same input unit for each word. (a) PCA values for the 2nd principal component as the SRN processes two sentences, '*Boy who boys chase chases boy'* or '*Boys who boys chase chase boy'*; (b) PCA values for the 1st and 11th principal components as the SRN processes '*Boy chases boy who chases boy who chases boy*'.

Figure 1.

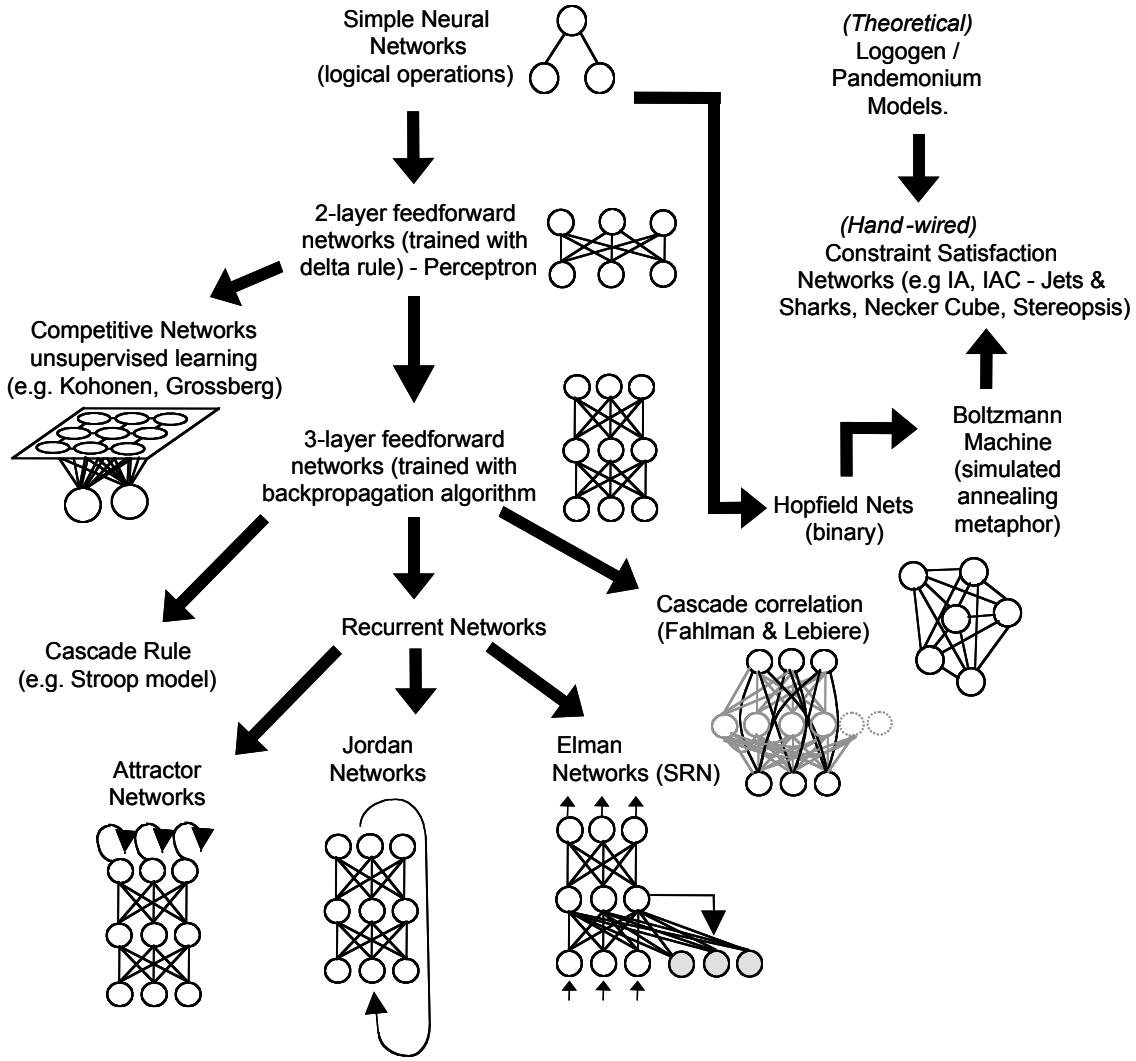

Simple Neural Networks (logical operations)
(Theoretical) Logogen / Pandemonium Models.
2-layer feedforward networks (trained with delta rule) - Perceptron
(Hand-wired) Constraint Satisfaction Networks (e.g IA, IAC - Jets & Sharks, Necker Cube, Stereopsis)
Competitive Networks unsupervised learning (e.g. Kohonen, Grossberg)
3-layer feedforward networks (trained with backpropagation algorithm
Boltzmann Machine (simulated annealing metaphor)
Hopfield Nets (binary)
Cascade correlation (Fahlman & Lebiere)
Recurrent Networks
Cascade Rule (e.g. Stroop model)
Attractor Networks
Jordan Networks
Elman Networks (SRN)

Figure 2

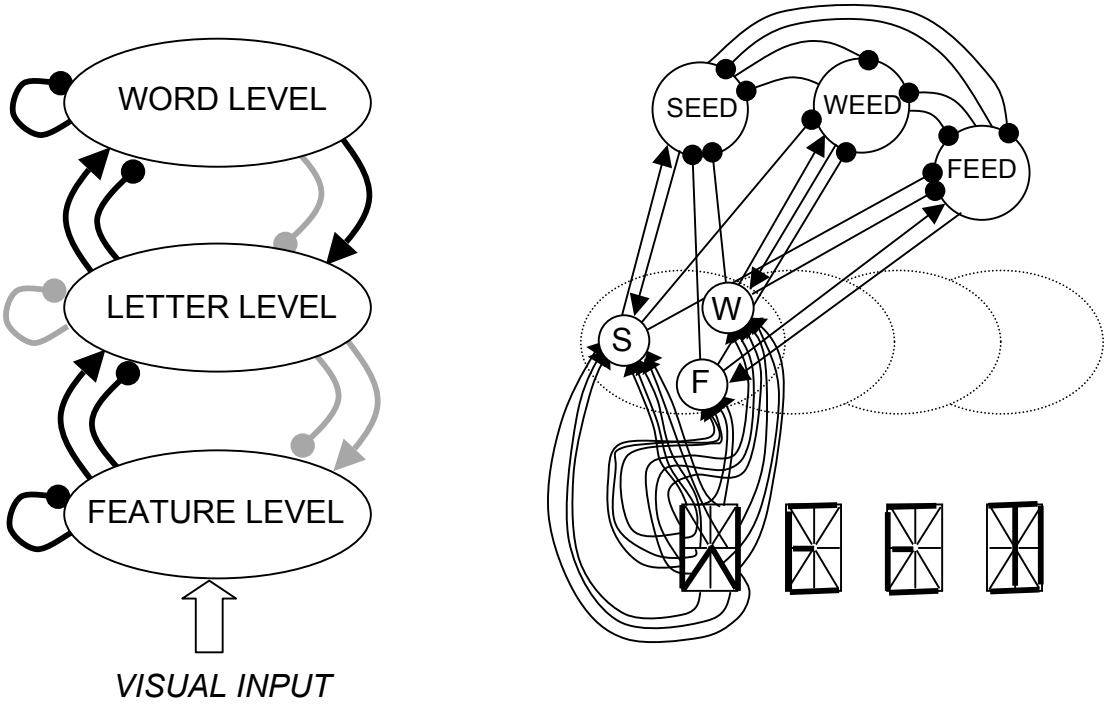

WORD LEVEL
LETTER LEVEL
FEATURE LEVEL
VISUAL INPUT
SEED
WEED
FEED
S
W
F

Figure 3

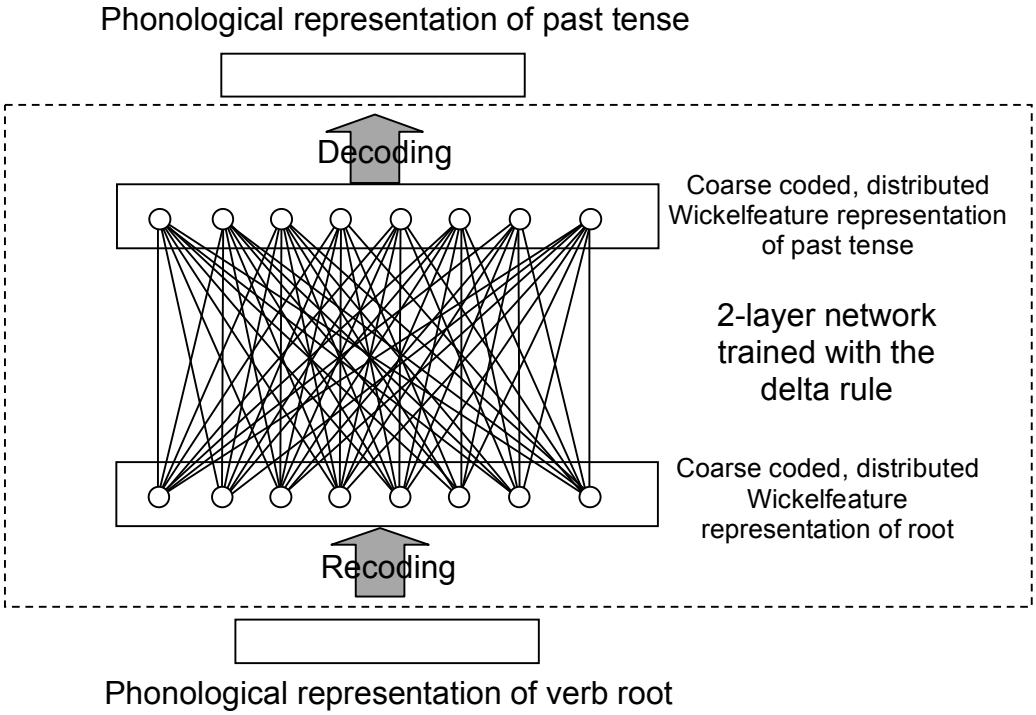

Phonological representation of past tense
Decoding
Coarse coded, distributed Wickelfeature representation of past tense
2-layer network trained with the delta rule
Coarse coded, distributed Wickelfeature representation of root
Recoding
Phonological representation of verb root

Figure 4

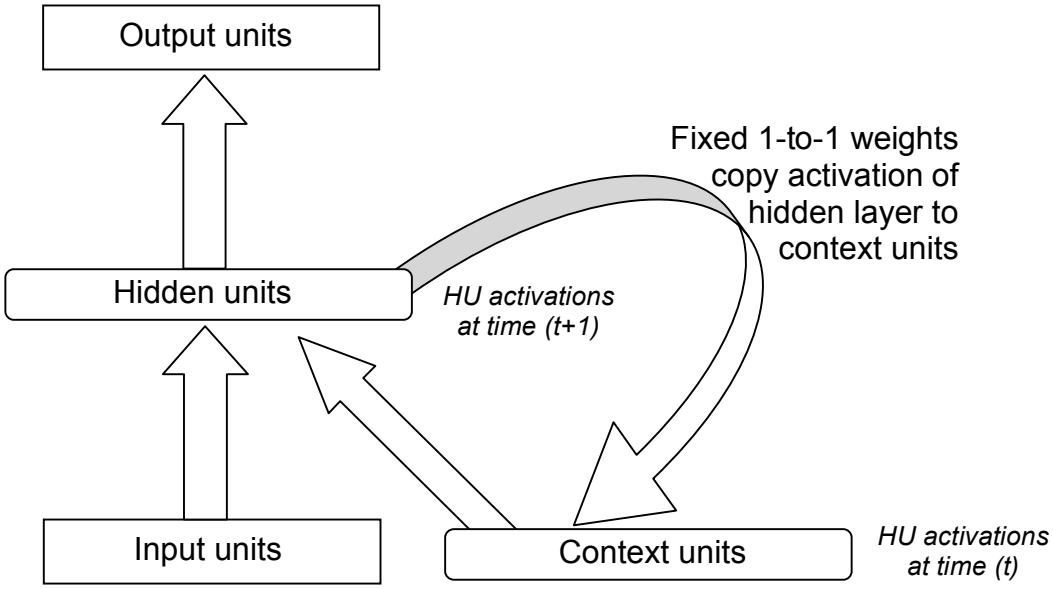
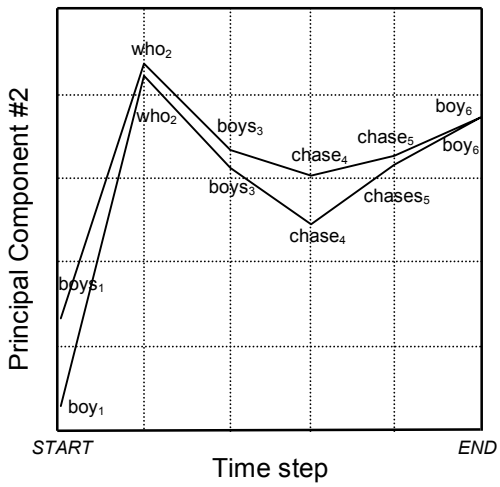
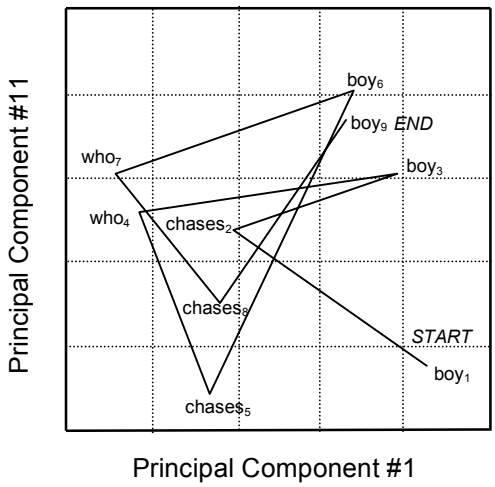

Output units
Fixed 1-to-1 weights
copy activation of
hidden layer to
context units
Hidden units
HU activations
at time (t+1)
Input units
Context units
HU activations
at time (t)

Figure 5

(a)


Principal Component #2
who2
who2
boys3
chase4
chase5
boy6
boy6
boys3
chases5
chase4
boys1
boy1
START
END
Time step

(b)


Principal Component #11
boy6
boy9 END
who7
boy3
who4
chases2
chases3
START
boy1
chases5
Principal Component #1