ELSEVIER

Cognition 66 (1998) 153–182

COGNITION

Can connectionism save constructivism?

Gary F. Marcus*

Department of Psychology, New York University, 6 Washington Place, New York NY 10003, USA

Received 3 July 1997; accepted 23 March 1998

Abstract

Constructivism is the Piagetian notion that learning leads the child to develop new types of representations. For example, on the Piagetian view, a child is born without knowing that objects persist in time even when they are occluded; through a process of learning, the child comes to know that objects persist in time. The trouble with this view has always been the lack of a concrete, computational account of how a learning mechanism could lead to such a change. Recently, however, in a book entitled *Rethinking Innateness*, Elman et al. (Elman, J.L., Bates, E., Johnson, M.H., Karmiloff-Smith, A., Parisi, D., Plunkett, K., 1996. *Rethinking Innateness*: A Connectionist Perspective on Development. Cambridge, MA: MIT Press) have claimed that connectionist models might provide an account of the development of new kinds of representations that would not depend on the existence of innate representations. I show that the models described in *Rethinking Innateness* depend on innately assumed representations and that they do not offer a genuine alternative to nativism. Moreover, I present simulation results which show that these models are incapable of deriving genuine abstract representations that are not presupposed. I then give a formal account of why the models fail to generalize in the ways that humans do. Thus, connectionism, at least in its current form, does not provide any support for constructivism. I conclude by sketching a possible alternative.

Keywords: Connectionism; Constructivism; Nativism

1. Introduction

In an ambitious new book entitled *Rethinking Innateness* (henceforth, *RI*), Elman et al. (1996) promise to use

* E-mail: gary.marcus@nyu.edu

0010-0277/98/$19.00
PII S0010-0277(98)00018-3

> existing connectionist models that simulate developmental phenomena in new and theoretical [sic] exciting ways... [to] *show how domain-specific representations can emerge from domain-general architectures and learning algorithms and how these can ultimately result in a process of modularization as the end product of development rather than its starting point*. (p. 115, emphasis in original).

If these strong claims were correct, it would be big news, for three reasons: it would provide the first ever computational account of Piagetian constructivism, it would undermine many arguments for nativism, and it would obviate the need for innate symbol-manipulating machinery. I treat these points in turn.

1.1. Relation to constructivism

First, the authors of *RI* see themselves as remedying the crucial weakness of constructivism. Constructivism is the Piagetian notion that learning leads the child to develop new types of representations. For example, a child might be born without having an abstract notion of *object*, and hence without knowing that all objects persist in time even when they are occluded. Instead, the child would develop the abstract notion of *persisting object* through a process of learning. This constructivist view contrasts with that of Spelke (1994) who proposes that abstract entities like *object*, *person*, and *place* are innate. Likewise, whereas many generative linguists hold that we are born with abstract categories like *noun* and *verb*, a constructivist might argue instead that these categories emerge through some sort of learning process, such as the one Karmiloff-Smith (1996) called 'representational redescription'.

Historically, the trouble with the constructivist view – that qualitatively new kinds of representations emerge – is that there has never been a concrete, computational account of how a learning mechanism could lead to constructivist 'representational redescriptions' (Fodor, 1975, 1981; Bloom and Wynn, 1994). For example, describing Piaget, Fodor (1975) noted that

> on [Piaget's] view, some concepts, like conservation of quantity, cannot be learned by the 'preoperational' child because characterizing the extension of the concepts presupposes algebraic operations not available in the preoperational logic. But if the child cannot so much as represent the conditions under which quantities are conserved, how in the world could he conceivably learn that those *are* the conditions under which quantities are conserved? Small wonder that Piaget gives so little by way of a detailed analysis of the processes of 'equilibration' which are supposed to effect stage-to-stage transitions. In fact, Piaget's account of equilibration is, so far as I can tell, *entirely* descriptive; there is simply no theory of the processes whereby equilibria are achieved. (p. 90, emphasis original).

This is where the authors of *RI* come in: they see connectionism as providing an

explicit computational account that would support constructivism. That these scholars take their view to be a vindication of Piaget is made plain here:

> constructivism [considered] development in terms of self-organizing emergent structures arising from the complex interactions between both organism and environment. We believe that the biological–connectionist perspective opens the door to a new framework for thinking about development which embodies some aspects of Piaget's, Werner's, and Vygotsky's constructivist intuitions, but *which goes beyond them and provides a formalized framework within which to generate empirically testable questions* (p. 114, emphasis added).

1.2. Relation to nativism

Second, although (as the authors of *RI* themselves stress) any system that can learn can do so only in virtue of innate machinery, the authors of *RI* see their work as providing an argument against 'representational nativism'

> [A]s far as representation-specific predispositions are concerned, they may only be specified at the subcortical level as little more than attention grabbers so that the organism ensures itself of massive experience of certain inputs prior to subsequent learning...at the cortical level, representations are not prespecified; at the psychological level representations *emerge* from the complex interactions of brain and environment and brain systems among themselves... (p. 108, emphasis in original).

Of course, virtually everyone would agree that the representations of many individual words and concepts are learned, but this seems to be a much stronger claim. Unfortunately, the *RI* claim that representations are not prespecified is difficult to evaluate, since the authors do not make explicit what they count as a representation. Still, sections like 'does anyone disagree?' make their targets clear: the views of scholars such as Chomsky, Pinker, Crain, Spelke, Carey, and Leslie, researchers who argue that some (though by no means all) aspects of mind are innate.

Note that although the authors of *RI* take a strong position against *any* innate representations, many constructivists do not take such an extreme position. Some more moderate constructivists are willing to acknowledge a great deal of innate machinery, including a variety of representations as well as the constructivist mechanisms themselves. Any evidence for the existence of innate representations would thus count only against the extreme position advocated in *RI*, but not necessarily against other, weaker versions of constructivism.

1.3. Relation to connectionism and symbol-manipulation

Third, the authors see themselves as using connectionism to provide an alternative

to theories of cognition that involve symbol-manipulation. Again, while the authors of *RI* endorse a version of connectionism that denies the existence of symbol-manipulation, not all connectionists take such a position. It is worth distinguishing, then, two kinds of connectionism (Fodor and Pylyshyn, 1988; Pinker and Prince, 1988; Marcus, 1997), 'eliminative connectionism', a school of connectionism that denies the existence of symbol-manipulating primitives and 'implementational connectionism', a school of connectionism that seeks to account for how symbols are instantiated in the brain.

The authors of *RI* plainly distance themselves from implementational connectionism when they write that 'we know that it is possible to build a connectionist network which implements a LISP-style rule system (Touretzky and Hinton, 1985). But when it comes down to it, it is probably easier to write LISP programs in LISP', further noting that they see connectionism as providing an alternative in which rules 'look very different than [in the] traditional symbolic' view (p. 103).

Still, while the *RI* position draws heavily on research in connectionism, it should be noted that the *RI* position – in which innate constraints, domain-specific structure, and symbol-manipulation are avoided – may not be representative of the connectionist community as a whole. Indeed, their claims would surprise many less radical connectionists. For example, Denker et al. (1987) argued (p. 877) that

> Since antiquity, man has dreamed of building a device that would 'learn from examples', 'form generalizations', and 'discover the rules' behind patterns in the data. Recent work has shown that a highly connected, layered network of simple analog processing elements can be astonishingly successful at this, in some cases. [But]... the solutions that humans seem to prefer are not the ones that the network chooses from the vast number of solutions available; indeed, the generalized delta method and similar learning procedures do not usually hold the 'human' solutions stable against perturbations.

Many connectionist researchers have stressed the importance of domain-specific structure (Burgess and Hitch, 1992; Regier, 1995; Shallice et al., 1995; Hartley and Houghton, 1996). Regier (1995), for example, advocates 'adaptive structured connectionism', a connectionist approach which is distinguished by 'its emphasis on the use of highly domain-specific and domain-motivated structure' (p71).

Likewise, many connectionists have stressed the importance of initial representations and biases (Seidenberg and McClelland, 1989; Kolen and Goel, 1991; Geman et al., 1992; Anderson, 1995; Clark and Thornton, 1997). For example, Geman et al. (1992) concluded that

> it is our opinion that categorization must be largely built in…and that identifying these mechanisms is at the same time more difficult and more fundamental than understanding learning *per se*. (p. 51).

Similarly, many connectionists concede the existence of the symbol-manipulation machinery that the authors of *RI* appear to oppose (Barnden, 1984; Touretzky and

Hinton, 1985; Sun, 1992; Shastri and Ajjanagadde, 1993; Smolensky, 1995; Hummel and Holyoak, 1997). Smolensky (1995, p. 226), for instance, aims to 'explain how symbolic computation is built upon neural computation'.

In short, although the *Rethinking Innateness* approach represents one particular use of connectionism that is aimed at eliminating innate domain-specific representations and symbol-manipulation, such radical claims are not an intrinsic part of connectionism, and other researchers have pursued more moderate approaches to connectionism incorporating the very notions that Elman et al. aim to deny.

1.4. Summary

The central argument of *RI* has two planks. First, representations and modules are not innate; second, new kinds of representations and new modules can be learned through connectionist architectures. The *RI* position seeks to provide a mechanism for constructivism while denying that representations, modules, and the machinery of symbol-manipulation are innate.

Before evaluating these arguments, let me again stress that the framework of connectionism does not itself rest on the status of the *RI* research program. The arguments given below should not be taken as arguments against connectionism per se, but only against the particular version of connectionist theory espoused in *RI*. (My own view is that connectionism *can* ultimately make profound contributions to our understanding of cognition, albeit embedded in a more moderate view that seeks to explain rather than deny the existence of innate representations, innate modules, and innate machinery for manipulating symbols.)

2. Arguments against representational nativism

Let us first consider the *RI* claim (which, as mentioned, is not intrinsic to constructivism) that there are no innate representations. Much of the argument for the *RI* view must depend on the plausibility of the models that they discuss; still, before discussing how these models fit with their view, it is worth first briefly considering their principal argument *against* the view that there are innate representations:

> the strongest and most specific form of constraint...is representational, expressed at the neural level in terms of direct constraints on fine-grain patterns of cortical connectivity. This is, we think, the only kind of neural mechanism capable of implementing the claim that detailed knowledge of (for example) grammar, physics or theory of mind are innately specified. (p. 360.)

This claim rests entirely on the assumption that the only way for a representation to be innate is for 'neurons to be born 'knowing' what kinds of representations they are destined to take on' (p. 27).

While the authors muster ample evidence (primarily about plasticity) against the

cell-predestination view, the problem is that this is a straw version of nativism, a view of embryology never endorsed by the authors that *RI* aims to critique. That representations are not prespecified in this way (with neurons born knowing their final destinations) is not controversial, since the genetic code itself is not a blueprint that specifies precisely which cell goes where (Dawkins, 1987).

Instead, just as the structure of the heart can be plausibly construed as being innate even though no individual cell is born 'knowing' that it will be a heart cell, the structure of some kinds of representations could be innate even though no individual neuron is born 'knowing' its fate. Rather, a cell might need only carry 'instructions' (I use the term very loosely) like 'if I am in the neighborhood of heart cell of type 1, become a heart cell of type 2', 'if I am in the neighborhood of neuron of type 1, become a neuron of type 2' or 'if I receive a chemical signal of type 7, activate the master gene that controls the operation of genes 43 through 57', and so forth.

Much of developmental biology seems to work this way. One piece of evidence that this sort of developmental mechanism plays a role in neural development comes from the research of the neuroscientists Katz and Shatz (1996), who suggest that 'early in development, internally generated spontaneous activity sculpts circuits on the basis of the brain's 'best guess' at the initial configuration of connections necessary for function and survival' (p. 1133), leading them to conclude that

> visual experience alone cannot account for many features of visual system development. In nonhuman primates, for example, ocular dominance columns in layer 4 begin to form in utero and are fully formed by birth. Thus, although visual experience can modify existing columns, initial formation of the stripes is independent of visual experience. Other features of cortical functional architecture, such as orientation tuning and orientation columns, are also present before any visual experience... (p. 1134.)

This picture of internally generated activity leading to initial configurations – before postnatal experience – seems perfectly consistent with claims of scholars like Pinker and Spelke who hold that innate constraints influence what can be learned. A set of such instructions could lead to richly differentiated structures (and representations) *without requiring that the fates of individual cells be prespecified and without requiring input from outside the developing embryo*. In this way, representations could be innate without having individual cells be predestined.

RI offers no evidence against this more nuanced view of representational nativism. At best, their argument against alternative versions of representational nativism seems to be a poverty of the imagination argument (i.e. the fact that nobody has yet identified a detailed account for how innate knowledge would be instantiated in the brain). But it is ironic that they seem to make this suggestion, since in defending the biological plausibility of their own models, *RI* correctly notes that 'There is obviously a great deal which remains unknown about the nervous systems and one would not want modeling to always remain several paces behind the current state of the science...' In short, at least for now, the mere fact that a given structure or developmental mechanism has not yet been identified tells us little, and we cannot

therefore use such a gap in our knowledge to rule out the possibility that representations are innate.

In sum, although current neurophysiological evidence described in *RI* militates against a straw man version of nativism, it leaves more sophisticated views of nativism largely untouched. For now arguments for their approach must rest then not on their critique of nativism, but instead on the plausibility of accounts in which representations are not innate.

3. The modern-day connectionist infant

Let us turn now to *RI*'s positive proposal. The specific view that is advocated by these scholars is what Karmiloff-Smith (1996) has dubbed 'the modern-day connectionist infant':

> an infant with several biases specified prior to processing the input; a recurrent network, an autoassociative network, and so forth, each of which is suited to a different type of input, but with weights and connections left random. This illustrates how the infant brain might start out with a number of architectural, computational, temporal and other biases or constraints, without building in representational content. (p. 2.)

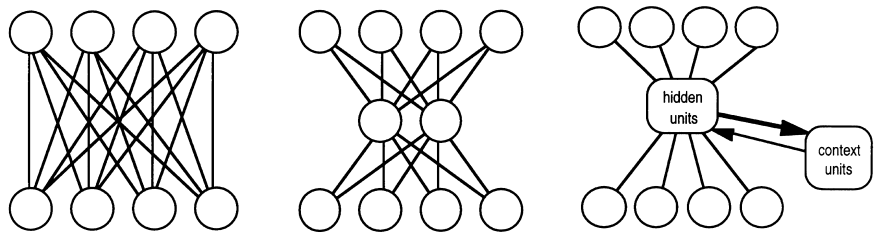
On this view,

> networks learn by slowly assimilating the input via tiny changes each time an input is processed as well as by accommodating...the progressively changing representations of the network to the particularities of the structure of the input. This retains Piaget's deep intuition that systems have to learn, and when they learn, this affects the way in which they subsequently process the input. (ibid., p. 2.)

These by now familiar models consist of a set of interconnected units, an encoding scheme, a training regime, and input/output scheme (for an introduction to connectionism, see Bechtel and Abrahamsen, 1991). Examples of how the units are interconnected are shown in Fig. 1. The way in which these networks work is that a 'teacher' (external to the model) provides a series of input–output pairs. The learning algorithm (typically the back-propagation algorithm) adjusts the weights on connections between nodes in a way that tends to minimize the difference between the model's output for a given input and the target output for that unit.

4. Evaluation of the *RI* proposal

To repeat, the key claims of the *RI* proposal seem to be these. First, there are no innate representations but there are learned (kinds of) representations (e.g. p. 108);



a. 2-layer pattern associator b. Single feedforward network c. Single recurrent network

Fig. 1. Sketches of some network architectures used in *Rethinking Innateness*. Input units are drawn on the bottom of each figure, output units on the top. Not all connections are shown. The ovals labeled as 'hidden units' and 'context units' contain multiple nodes. (a) Two-layer pattern associator. (b) Single feedforward network. (c) Single recurrent network.

second, there are few innate modules, but there are learned modules (e.g. p. 387). It follows, obviously, then, that the plausibility of their position rests on providing models that – lacking innate representations and innate modules – develop new representations and new modules. (Again, a weaker but still interesting version of constructivism that is not defended in *RI* would be one in which some representations and some modules were innate and others were learned.)

4.1. Representations

Because a full evaluation of these claims would obviously depend upon a clear definition of what counts as a representation – not provided in *RI* – we need to start by reconstructing their claim. A weak version of the claim that new representations are learned would be the simple claim that new instances of known types of representations are learned, such as new words (*fax*, *VCR*) or kinds of object; I presume that the authors of *RI* have something stronger in mind, since such a claim is uncontroversial. Likewise, virtually any learning system can combine already known primitives, forming say the concept of red square from the primitive concepts of red and square (Fodor, 1975), but I take it that such combinations are not new representations, but only combinations of already known representations. Thus, in what follows, I will assume that what the authors of *RI* mean when they claim that a new representation emerges is that what is learned is an abstractions like 'object', 'word' or 'noun', i.e. the very sorts of entities that *RI* denies are innate.

Patently, if a given model actually has innate representations, it cannot count as evidence for the 'lack of innate representations' view; likewise, if all of its representations are prespecified, it cannot count as a model that has learned new representations. Although the authors of *Rethinking Innateness* appear to believe the models that they describe do not have innate representations, what the authors of *RI* overlook is that the input and output encoding schemes used by their models *are* representations.

In each network, each input node (and each output node, see below) encodes –

hence, represents – something different[1]. In one model, one node is activated if and only if the input contains the speech segment /ba/, and another node is activated if and only if the input contains /pa/; in another model, this node turns on if and only if the speech stream contains the word *cat*, that one only if the input contains the word *dog*. Each input node is thus a feature detector that, from the perspective of the model, is innately given, innately responding to that which it represents. (For a similar point, see Bloom and Wynn, 1994.)

These pre-coded input representations are absolutely crucial towards determining what sorts of generalizations will be drawn by a given model (Lachter and Bever, 1988; Wexler, 1991; Geman et al., 1992). Providing a simpler or different innate encoding scheme often leads to completely different performance, sometimes to a model's total failure. As Kolen and Goel (1991) put it, 'the content of what is learned by the method of back propagation is strongly dependent on the initial abstractions present in the network' (p. 364).

Even more damning to the *RI* view is the fact that every *output* representation is prespecified. It is here that the grand dreams of empiricist learning – starting from raw sensation and bootstrapping all the way up to a full adult conceptual system – fall hardest. A real system of representational emergence would develop new concepts where there were none; the models of *RI*, however, do not deliver. Of course, the authors of *RI* cannot be expected to give a full account from sensation to adult concepts. Rather the problem here is that *RI* never give an account of how a single output representation might 'emerge'; instead, every single output representation in every single model is prespecified.

What the *RI* models learn, then, is not representations, but merely correspondences between prespecified sets of representations. Connectionist models, like all other known learning models, necessarily presuppose representation. Again, the fact that these models presuppose representations is of little consequence for constructivism in general, but it undermines the stronger claim made in *RI* that representations are not innate. The fact that the models that *RI* describe do not construct new representations means that the models that describe cannot provide any support for constructivism.

A newer connectionist model, the wake–sleep model of Hinton et al. (1995), at first glance seems to provide an account of how new output units could be developed, but probably turns out to be overly sensitive to variation in the input stimuli. In essence, this model, an 'unsupervised' learning model (see also Linsker, 1988), which is not cited in *RI*, tries to find the most efficient way of compressing some

[1]Among the things that are represented by input or output nodes in models advocated in *Rethinking Innateness* are the following: sequences of phonetic features, the presence or absence of a consonant or vowel in a word's slot, the presence or absence of particular phonetic features in particular positions, the presence or absence of a morphological feature in a given position, the presence or absence of a grammatical feature, sequences of letters, sequences of phonemes, graphemes in specified slots, phonemes in specified slots, the visibility or lack thereof of abstract objects, properties of objects, the distance of some weight on a balance-beam from a fulcrum, the number of weights on a particular side of a balance-beam, words, coordinates of cube vertices, coordinates of joint locations, the relative location of an object, the distance of one object from another object, the identity of an object, or the description of which of two tasks a model is participating in at any given moment.

set of input stimuli[2]. For example, in one demonstration, Hinton et al. presented a version of their model with a series of stimuli, half of which are composed entirely of vertical bars and half of which are composed entirely of horizontal bars. At the onset of training, the model has a set of (innately given) input units that represent pixels and a smaller set of not-yet-meaningful output units. Through a trial-and-error process, the model adjusts the weights feeding the output units such that most output units wind up encoding either a vertical line in a particular location (e.g. one unit encodes a vertical bar in the leftmost position) or horizontal line in a particular location. Of course these encodings are not genuine abstractions (e.g. the vertical-line-in-the-leftmost-position detector is activated when all or most of the pixels in the leftmost column are activated), but much more interesting is that the model can construct a node that responds to a 'vertical line in any location'. If the model could robustly construct such nodes in a wide range of environments, it might provide an account of how (some) new representations are formed. The construction of such a node appears, however, to be entirely dependent on a special property of the input stimuli: because the stimuli are composed entirely of vertical bars or else composed entirely of horizontal bars, the vertical bars are correlated with one another. (In other words, it's a quirk of the input data that knowing that status of one bar – whether it is present or absent – tells you something about what other bars can appear in the stimulus. In the real world, vertical and non-vertical bars co-occur routinely, in T-junctions, plus (+) signs, L's, and so forth.) If the stimuli were instead constructed such that the presence (or absence) of each bar was independent of the presence or absence each other bar (i.e. vertical bars would be just as likely to co-occur with horizontal bars as with vertical bars), it appears that the model would not construct a node that corresponded to 'vertical line in any location.' In short, although the wake–sleep model provides a way of finding useful combinations of input features, it may not be robust enough to discover abstract encodings that are robust across a broad range of input environments.

To take another example of the model's capabilities and limits, if the model were trained on a set of words that were represented in an array of pixels, the model would likely construct nodes corresponding to particular letters in particular positions (e.g. an 'r' in the second position). But because the presence of an 'r' in one position does not predict the presence of an 'r' in another position, it appears that the model would not construct a node corresponding to 'an 'r' in any position', suggesting that some other mechanism is necessary to account for how humans extract such features.

4.2. Modularity

Modules, according to the classic definition from Fodor (1983), are devices that are among other things, mandatory, unconscious, functionally distinct computa-

[2]This whole approach is very different from the *RI* models. Whereas the *RI* models are 'supervised' models that are trained on pairs of inputs and outputs, seeking to discover some mapping between inputs and outputs, the unsupervised wake–sleep model is provided only with a set of inputs; rather than learning some kind of mapping, the wake–sleep model tries to provide an account of how to segment some set of input stimuli into useful component parts.

tional input units that are informationally encapsulated. The authors of *RI* appear to accept roughly this view, writing that 'A module is a specialized, encapsulated mental organ that has evolved to handle specific information types of particular relevance to the species' (p.36). They note correctly (p. 101) that

> nothing intrinsic to the connectionist frame [sic] precludes modularity, and we have already made the point that some degree of organization and modular structure appears necessary if models are to be scaled up.

Are modules prespecified in the *RI* framework? Once again, the models that *RI* offers are far more prosaic than their claims. *RI* models already *are* modules; indeed, *RI* proposes a separate, informationally encapsulated model for essentially each task that the book tries to explain. Distinct models with distinct architectures and encoding schemes are proposed for modeling the past tense, vocabulary acquisition, object permanence, the balance-beam task, and so forth. Indeed, it is hard to see how the fault lines presumed by *RI* differ from the fault lines presumed by the most passionate advocate of modularity.

What's more, these architectures are always entirely presupposed; none of these models learn to divide itself into new modules. Although the authors of *RI* repeatedly advocate a position in which modules 'emerge', they never actually present a model in which modules emerge. Instead the authors point (p. 355) only to models like those proposed by Jacobs and Kosslyn (1994). But the modules in the Jacobs/Kosslyn network (there are three of them, two 'experts' and a 'gating network') do not 'emerge' – they are prespecified. Each module learns whatever task it is assigned, but each has its own internal structure that operates independently of each other module; each module has its own output representation, each modules has its own pre-specified wiring diagram, and so forth[3]. (Indeed, Jacobs and Kosslyn do not themselves claim that these modules emerge, so it is puzzling that the authors of *RI* make such a claim.) For now, there simply is no proposal for a mechanism by which new 'modules' can emerge.

4.3. Summary

The models proposed in *RI* do not match the theoretical claims made in *RI*. *RI* argues that the mind lacks innate representations and innately defined modules, and that new modules and representations 'emerge'. The models they present, however, are replete with innately defined modules and innately defined representations; neither new modules nor new representations emerge; hence connectionism has not yet provided a sound computational basis for constructivism.

[3]In principle, one could try to avoid having to prespecify the exact number of modules that are required for some task, by simply starting with a larger-than-needed number of expert networks, some of which would ultimately contribute nothing to the final overall set of networks. But the inventors of modular networks (Jacobs et al., 1991) write that 'we do not advocate providing the modular architecture with a large number of different expert networks. Rather, the experimenter should judiciously design a small set of potentially useful expert networks where the potential utility of an expert network is evaluated using domain knowledge'.

5. Representations within hidden units

We have already seen that within the models described in *RI*, input representations are innate and output representations are innate, so the only place left for new representations would be in the 'hidden units'. Indeed, in a discussion of Elman's Single Recurrent Network model (p. 95), a model that aims to learn the categories of nouns and verbs on the basic of distributional information, the authors of *RI* seem to suggest that new representations can be learned within hidden units, suggesting that the network 'forms hidden unit representations which places these two groups in different areas of activation space. These groups correspond to what we would call nouns and verbs'. Thus, for example, *RI* describes a set of data in which nouns tended to yield similar patterns of hidden unit activations. Thus nouns tend to appear in similar points in an *n*-dimensional 'hidden unit space'.

If *any* noun elicited the same pattern of hidden unit activity as any other noun, we would surely be justified in concluding that that pattern served as an abstract representation of the notion noun. But the real reason that the nouns that Elman studied elicited similar patterns of activity is less interesting: the training corpus that was generated by Elman's artificial grammar was designed such that each noun appeared in roughly the same set of contexts.

We can tease apart the issues of similarity of contexts and the notion of a newly developed abstract category of noun with a simple experiment: we can test the model on a novel noun that appears just once. In this case, it turns out that the novel noun *does not* elicit the same pattern of activation as other nouns. Instead, the novel noun appears in roughly the middle of the *n*-dimensional space, relatively far from all the other nouns, which tend to occupy a corner of the n-dimensional space. Thus, common patterns of hidden activity reflect *words that have appeared in similar circumstances*, not *words that belong in the same grammatical category*. The hidden units therefore cannot be said to have developed an abstract representation of the category *noun*[4]. This point is devastating to *RI*'s connectionist constructivist position, because the ability of connectionism to support constructivism would rest on finding a way for a network to develop an abstract representation of *noun*[5].

[4]One could argue that the hidden units form a category of 'the nouns seen so far', but the hidden units would respond differently to two different nouns if those nouns were presented in sufficiently different sets of contexts. Instead, the hidden units form a category of 'words that appeared in similar contexts' which, like Fodor's 'red and square' example, amounts to a conjunction of preexisting representations, rather than a genuinely new representation. (It is quite common in our everyday experience for two words to appear in differing contexts, as when one noun is frequent and occurs in a wide variety of contexts and another is extremely infrequent and occurs in only a handful of sentence frames – despite the vastly different contexts, we can use the infrequent word in the same set of grammatical contexts as the frequent word.)

[5]At this point, defenders of connectionism who are not concerned with constructivism might concede that current connectionist models do not yield a mechanism of constructivism, but wonder whether connectionism might instead provide a novel rendering of a view which we can call 'cognition without abstract representations'. Note, however, that if such a view were correct, there would be no need for constructivism, since if there were no abstract representations, a fortiori there would be no new abstract representations. (An important limitation on the *RI* implementation of the cognition-without-abstract-representation view is discussed in the next section.)

6. Towards a better understanding of how the *RI* models work

We have seen, then, that what *RI* models can do is to learn correspondences between observed sets of predefined features. What the *RI* models cannot do, however, is to generalize abstract relationships to features that did not appear in the input data.

Unpacking what this last sentence means is the burden of this section. We need to start by contrasting the *RI* approach with the standard symbol-manipulation view that the mind represents abstract relationships between categories. Grammars for instance, are traditionally taken to be sets of rules that range over variables. The rule: *a sentence is formed by concatenating a Noun Phrase and a Verb Phrase* allows us to form a sentence using anything that can instantiate the variables Noun Phrase and Verb Phrase respectively, regardless of how familiar the instantiations of those variables may be. For instance, we can combine the relatively novel Noun Phrase (say 'a lovesick young boy saddled with the peculiar name Dweezil') with an equally novel verb phrase ('became the subject of a perversely uncreative linguistic demonstration'). Since the rule is given in terms of relationships between variables, the rule is easily extended to unfamiliar items.

Likewise, at least by the time we are toddlers, we realize that, other things being equal, all objects persist in time, even if they are temporarily occluded from our view. In this way, knowledge of object permanence is thought to hold for all items in the class 'physical object', allowing us to extend our knowledge of object permanence to any kind of object, even one we have just seen for the first time.

It turns out that *RI* networks behave differently – their ability to generalize relationships to novel items depends on how familiar those items are. For example, suppose we train an *RI*-style model on the simple relationship 'sameness' or 'identity', using a model like the one shown in Fig. 2. This model is given two percepts, one (A) represented by the left bank of input nodes, and the other (B) represented by the right bank of units. The model is trained to produce the output '1' if the two inputs are the same, '0' if the two inputs are different.

Suppose that we use the standard back-propagation algorithm to train the model to respond 'same' if percepts A and B are both the sound /ba/, and likewise train the model to 'respond same' if the percepts A and B are both the sound /da/. Will the

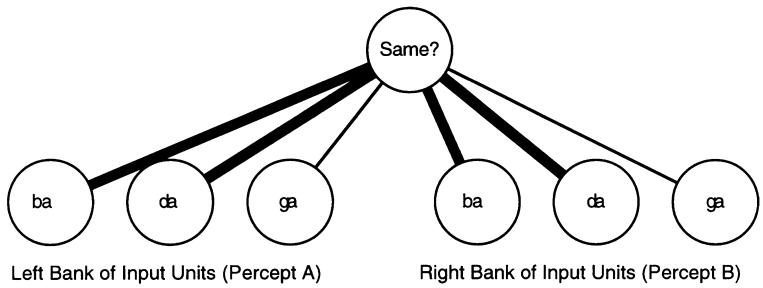


Fig. 2. A simple connectionist model which is trained to represent the relationship of 'sameness'.

network be able to generalize this relationship to the case in which both percepts are of the sound /ga/?

The answer, it turns out, is no: the network does not generalize the identity function in the same way as a person would. In fact, if we trained the model on 100 sets of inputs, the model would not be able to generalize the sameness function to the 101st set of inputs. The problem, in a nutshell, is that what the model learns about how to treat one set of inputs doesn't generalize to how the model treats the next set of inputs, a problem I will call 'featurewise independence'.

We can put this a bit more formally. With respect to some function that we train the model on, any item that consists entirely of features that have appeared in the training set lies within the *training space*. Any item that contains one or more features that have not been exemplified in training lies outside the *training space*. It turns out that none of what the model learns about items inside the training space extends to items outside the training space.

The reason that none of what the model learns about items inside the training space extends to items *outside* the training space turns out to follow directly from the mathematics of back-propagation (and also follows from the mathematics underlying many other popular training algorithms, such as Hebbian learning), a consequence of the fact that the learning of these models is in some sense strictly local, in two ways, one pertaining to input units, the other to output units.

Consider first the localism of input units. On any given trial, if a given input unit is activated at the level of zero (i.e. for any feature that is not present), the weights leading from that input unit to the hidden units (or output units) remain unchanged[6], regardless of the activity levels of the other input and output units. Thus if the input node representing /ga/ in the left bank of units is not activated, the model learns literally nothing about that node, regardless of what the rest of the training pair looks like[7].

Training independence applies even if we set up the network in a different way; for example, another typical way of setting up the relationship of sameness is shown in Fig. 3. Suppose that we train this network on the identity function for binary numbers. Suppose further that we train this model only on even numbers (e.g. 1010 = 10), and then test the model on odd numbers: (e.g. 1111 = 15).

What happens here is that, again because of featurewise independence, the model can't generalize 'sameness' to the odd numbers; for example, given input [1111] the network produces output [1110]. Here, not only is there a problem arising from the

[6]Proof: the equation that adjusts the weight of connection *ij* from input unit *i* to unit *j* is given as delta_{ij} = learning rate*error signal*activation of input unit *i*. Since, by assumption, the activation of input unit $i = 0$, and the weight-change equation includes that zero value as a multiplicative factor, delta_{ij} remains at zero in every trial, hence no learning takes place (note that this is true regardless of how the error signal is calculated).

[7]One might wonder what would happen if the absence of a feature were coded by some value besides 0.0. Indeed, if a given input unit is activated at some other level besides 0.0, the weights leading from that unit will change from trial to trial (unless, for example, the network has already reached a local minimum). Still, if the activity value of some input unit is not correlated (alone or in conjunction with other input units) with the output unit, it appears that the net effect is still the absence of learning, presumably because the weight adjustments for such units tend to cancel one another out.

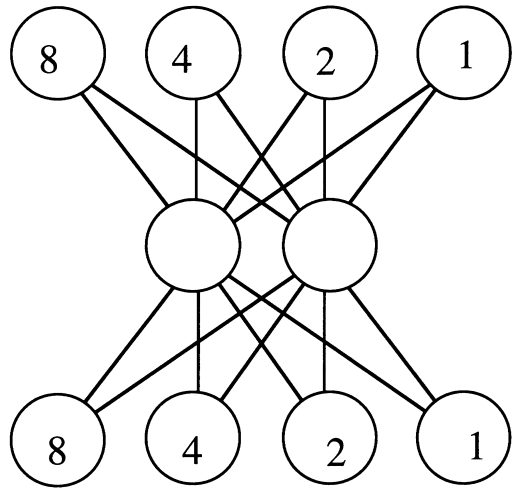


Fig. 3. A simple connectionist model that represents inputs and outputs as strings of binary digits.

fact that each input unit is trained independently, but an additional problem arising from the fact that each output unit is trained independently, hence what the model learns about the conditions for activating one output unit does not affect the conditions for activating another output unit[8].

Minimally, this limitation applies to all single network models that are trained by the back-propagation algorithm, the Hebbian algorithm, or other similar algorithms

[8]Proof: the computation of the weights feeding an output unit *j* are as follows. For a given input–output pair, the weight on the connection from a given (input or hidden) unit *a* to output unit *j* is adjusted according to the following equations, derived in Rumelhart and McClelland (1986): change in weight of connection from unit *a* to output unit *j* = learning rate × error signal × activation of input/hidden unit *a*, where error signal = (target for unit*j* − observed activation of output unit*j*)*[activation of unit*j**(1 − activation of unit*j*)]. Crucially, the equations that change the weights that feed a given output unit do not make reference to the activation levels of the other outputs or the weights feeding the other output units. Consequently, the set of weights connecting one output unit to its input units is entirely independent of the set of weights feeding all other output units, a limitation that we can call *output independence*. Note that strictly speaking, the output units exert a common influence on the hidden units that feed them. What is independent from unit to unit is the connections from the hidden units to the output unit. The output independence limitation is best understood by first considering how the back-propagation algorithm would apply to a two-layer network. Importantly, in these networks, no matter how felicitous our choice of 'input nodes', we must always *learn* the mapping between the input nodes and the output nodes. In the two-layer network, as guaranteed by the above equations, the training on the connections from the input nodes to a output unit *i* is completely independent of the training on the connections from the input nodes to output unit *j*. In a multilayer network the situation is made slightly more complex, but the same problem holds. Here, the weights from the *input* units to the *hidden* units are not adjusted independently, but crucially, the weights from the *hidden* units to the *output* units are still adjusted independently. To understand the implications of this, think of the hidden units as 'quasi-input' units. Under some circumstances, the mutual influence of output nodes upon the input-to-hidden-layer-connections may lead to felicitous encodings of these 'quasi-input' nodes, but no matter how felicitous the choice of 'quasi-input' units may be, the network must always *learn* the mapping between these 'quasi-input' nodes and the output nodes. The equations of back-propagation guarantee that this is done independently for each output node.

that represent their inputs with multiple sets of units working in parallel and that train each unit independently, a class that includes virtually all of the models described in *Rethinking Innateness*. Note that the localism (i.e. independence between units) that underlies these algorithms is a fundamental aspect of what makes these models different from traditional symbol-manipulating models. For example, in an influential book chapter (p. 214), McClelland and Rumelhart (1986) motivated their approach to connectionism by writing that

> The delta rule – and now the generalized delta rule – provide very simple mechanisms for extracting regularities from an ensemble of inputs without the aid of sophisticated generalization or rule-formulating mechanisms that oversee the performance of the processing system. These learning rules are completely local, in the sense that they change the connection between one unit and another on the basis of information that is locally available to the connection rather than on the basis of global information about overall performance.

Abandoning this localism would likely thus entail abandoning that which made connectionism seem like an interesting alternative to symbol-manipulation in the first place.

At first glance, the argument I have just given must seem like a bit of a trick, but this argument – essentially a sharpening of a claim in the connectionist literature that networks can interpolate but extrapolate – turns out to undermine most of the models discussed in *RI*. To see this, consider the following examples.

6.1. Single recurrent network models of syntax

First, consider again the widely-cited Single Recurrent Network of Elman (1991)[9]. This model was innately constructed to take in a sequence of words and predict the possible continuations to that sequence. For example, given *girls love the* ____, the model tries to predict (i.e. activate the nodes corresponding to) continuations such as *boys*, *girls*, *cats*, *dogs*, etc. The authors of *Rethinking Innateness* claim that models like this 'will spontaneously discover grammatical categories such as noun and verb' (*RI*, p. 6) and argue that 'new representations emerge from processing the input set' (*RI*, p. 129).

Yet what the model knows about one word does not necessarily generalize to the next. Because each new word must be represented by a new node[10], and because featurewise-independence guarantees that each input or output node is trained independently, the network cannot extend abstract grammatical relationships from one word to the next.

To illustrate this, I conducted a series of simple experiments (Marcus, 1997) designed to test the ability of the SRN to generalize simple abstract relations to

[9]Although the SRN is widely cited, in several areas of adult cognition in which serial structure is important the SRN has been explicitly rejected in favor of somewhat more structured connectionist models (Burgess and Hitch, 1992; Shallice et al., 1995; Hartley and Houghton, 1996).

novel words (full details are provided in Appendix A). For example, in one simulation, I trained an SRN on sentences drawn from two sentence frames: Sentences constructed from the first frame took the form 'the bee sniffs the X', where X was instantiated by 10 different words like 'rose', 'lily', 'tulip', and 'lilac'. Sentences constructed from the second frame took the form of a simple linking relationship, 'a Y is a Y', where Y was instantiated by nine of the ten instances of Y used in the first frame (e.g. *a rose is a rose*). Crucially, the word 'lilac' appeared in the first frame but not in the second frame.

The SRN mastered all the sentences on which it was trained. But it was unable to use the distributional information about *lilac*'s appearance in the first sentence frame to predict that *lilac* was a plausible continuation to the sentence fragment 'a lilac is a...' This failure is robust, unaffected by changes in the learning rate, the number of hidden layers, and the number of hidden units. In fact, the model can't generalize *any* linking relationship to novel words. A system that could simply label a given word as a noun could easily extend the frame 'an X is an X' to the word *lilac*, but a system that lacks such an abstract representation is unable to generalize this relationship in the way that humans do.

6.2. Object permanence

To take another example, let us consider the notion of object permanence. The nativist view that Spelke (1994) advocates is that initial knowledge about objects 'may emerge through maturation or be triggered by experience, but learning processes do not appear to shape it.' Instead, on this view, knowledge is expressed in terms of abstract entities such as objects.

The connectionist–constructivist alternative to this view is illustrated by a model of object permanence (cited favorably in *RI*) proposed by Munakata et al. (1998). This model does not start with innate knowledge about the spatiotemporal continuity of objects. Like Elman's model of syntax, the Munakata et al. model is based on

[10]Elman's model uses 'localist' output representations in which each output node represents a distinct word – from the perspective of the model words like 'cat' and 'dog' are therefore innate. Apropos my earlier point about how much a given model depends on a particular pre-specified input/output encoding, this seemingly innocuous design choice turns out to be crucial for the model. The 'localist' representation is crucial because it allows Elman to finesse the problem of not having categories like 'noun' – rather than predicting that the next word in the sequence '*the dogs ate the* ___' is a noun, the model can predict that the words *bone, bones, hamburgers, cat* and so forth are likely continuations. This is because the local encoding scheme allows each of these words to be activated without activating any other words. If words were instead represented through 'distributed representations' as when a given word is represented by a collection of bits of sound, the model could no longer keep the nouns separate from other words, because activating all the sounds that can appear in any noun would be tantamount to activating all the sounds that can appear in all words. Given such an output representation, the networks predictions cannot distinguish nouns from verbs. Thus the choice of encoding scheme (i.e. the innate representational scheme used for the input and output nodes) is, again, crucial to the operation of the model. (Likewise, the model is innately designed to be a device that performs a prediction task akin to a game of fill-in-the-blanks. That is, the model's innate design is such that all it can do is finish sequences. It cannot do any of things that language-understanding devices normally do, such as recover meanings from those strings, resolve anaphoric antecedents, or judge the grammaticality of sentences.)

simple recurrent network, but one in which each node represents a retina-like location instead of a word. The task of the model is to take as input a sequence of retina-like percepts, and to predict the next such percept. Each percept is a 7-unit wide by 2-units deep visual field. Munakata et al. showed that in scenario on which the network was trained, it can predict, at the moment in which the ball is occluded by the screen, that the ball will reappear once the screen has passed.

Interpreting these results, Munakata et al. make the following argument:

> The network's ability to form expectations is subserved by its connection weights. These weights are adjusted in the course of learning to make predictions from observed events. As the encoding weights from the input layer are adjusted, the network becomes increasingly able to represent occluded objects, not part of the input itself, as patterns of activity on the internal representation units. These patterns of activity thus provide a signal for an occluded objects continued existence.

To a naive reader, this must sound impressive, but it is deeply misleading – as I discovered in a series of simulations, the hidden units have not in fact formed an abstract, generalizable representation of object permanence. For example, in one test of this model, I trained the model (using the same configuration of nodes as Munakata et al.) on the occlusion relationship for an object in position 6; the model cannot predict the proper relationships when the object appears in position 5. To show that this problem persists even with substantially more training, I constructed 14 scenarios based on two directions of screen movement crossed with seven object locations; I trained the network (to the point of mastery) on 13 of those scenarios, and then tested the network on the 14th scenario. After 5000 training sequences (each consisting of six time steps), the model was able to correctly predict what would happen in each scenario on which it was trained. Still, despite this more extensive training, the model's predictions in the untrained scenario make clear that the network has not mastered object permanence. Even with this substantial training, the results for the 14th scenario showed that the net made peculiar predictions such as predicting that an object could be seen through an opaque screen. This model clearly has not genuinely abstracted the relations that underlie object occlusion and object permanence[11].

6.3. The past tense

The empirical domain that *RI* spends the most time discussing (18 pp. in Chapter 3 and scattered comments throughout the book) is the past tense. Unfortunately, this long discussion is also one of the least balanced in the book, since the section, although quite thorough in that it mentions six connectionist models of the past

[11]One alternative that the authors of *RI* endorse (p. 330) is a 'learning system that has an architectural bias such that it only develops representations of a certain kind, namely, those in which objects are extracted and represented on the basis of spatial-temporal continuity of component features'. But a system that is innately constrained such that it can *only* develop spatio-temporal continuity is not a way of defeating Spelke's position that such constraints are innate, but rather a way of covertly implementing it.

tense (Rumelhart and McClelland, 1986; Plunkett and Marchman, 1991, 1993; Daugherty and Seidenberg, 1992; Marchman, 1993; Hare and Elman, 1995), never mentions any criticism of their position that has been published since 1992 (Ling and Marinov, 1993; Prasada and Pinker, 1993; Kim et al., 1994; Marcus, 1995; Marcus et al., 1995). For a more recent discussion of connectionist models of inflection, see Marcus (1996).

The authors advocate a position in which '*regular and irregular verbs can behave quite differently even though represented and processed similarly in the same device*.' (p. 139, emphasis theirs.) In fact, the 'single mechanism' claim is a just promissory note. Critics of *RI*-style models like Pinker (1991) have noted that one argument for modeling the past tense with two mechanisms is that regular and irregular inflection appear to behave in systematically different ways in a wide variety of contexts. Yet no single connectionist model addresses even half of the phenomena Pinker outlines, and some of the phenomena Pinker describes have yet to be addressed. What's relevant here is that the response of connectionists has been to build a separate model for each problem raised by Pinker, e.g. one model for why denominal verbs receive regular inflection (Daugherty et al., 1993), another for handling defaults for low-frequency verbs (Hare et al., 1995), another to distinguish homonyms that different past tense forms (MacWhinney and Leinbach, 1991), and still another for handling a U-shaped developmental sequence (Plunkett and Marchman, 1993). These models differ from one another in their input representations, their output representations, and their training regimes; there is thus far no evidence that they can be put together in a single 'uniform' model, and to my knowledge no attempt has been made to do so.

This proliferation of past tense models also undermines another central claim of *RI*, '[t]he important lesson', they claim, 'is that some problems have natural good solutions; they have computational requirements which impose their own constraints on how the problem can be solved. ***Nature does not always need to provide the solution; it often suffices to make available the appropriate tools which can* then be recruited to solve the problems as they arise**.' (p. 78, italics and boldface in original). The implication of this passage is that the tools (i.e. innate machinery) need not be specified in much detail, because the richness of the environment will take up the slack. Note, however, that if the richness of the environment really was sufficient to obviate the need for carefully-structured learning devices, it would not be necessary to postulate so many different past tense models.

In any case, one point that critics of connectionism have raised, but that is unaddressed in *RI*, is the inability of 'single-mechanism' models like the ones championed in *RI* to generalize the regular inflection pattern (the *-ed* morpheme) to genuinely unfamiliar-sounding words. Humans appear to be able to extend the *-ed* suffixation process to novel words regardless of a word's similarity to stored examples, even to words that contain sounds unfamiliar in English, like *Jelsin out-gorbacheved Gorbachev* (Prasada and Pinker, 1993). In contrast, in a simulation using the Rumelhart and McClelland (RM) model, Prasada and Pinker found that although the RM model can apply generalizations to novel items that strongly resemble training items, it encountered difficulty when inflecting input words that lacked

resemblance to trained examples. The network, because it relies on feature-wise correlations rather than abstract representation of the *-ed* morpheme, tended to produce weird responses like *fraced* as the past tense of the novel word *slace*, *imin* as the past tense of *smeeb*, *bro* as the past tense of *ploanth*, and *freezled* as the past tense of *frilg*.

The story is essentially the same with respect to all the single-mechanism connectionist models of the past tense that are described in *RI*[12]. Given enough training, any of these models can master the training set, but none are able to generalize to novel words that contains pieces of phonology that have not appeared in training (Prasada and Pinker, 1993; Marcus, 1996, 1997).[13]

6.4. Balance-beams

As a final example, consider the balance-beam model of McClelland (1989) that is discussed in *RI*. In this model, a network is confronted with a version of the balance-beam problem, in which a seesaw contains ten pegs, divided into five equally-spaced pegs on either side of the fulcrum. The network is a simple feedforward network with a bank of 20 input units, four hidden units, and two output units. The output units represent the relative weight on the left and right sides of the balance-beam. The 20 input units are divided into two banks of ten, one for the left side of the beam, one for the right; each bank of ten is in turn subdivided into a bank of five units representing the number of weights on some peg and a bank of five units representing the distance that the weight-bearing peg is from the fulcrum.

McClelland trained the network on all 625 possible inputs (calculated as five codes for positions for the left object $\times$ five codes for the number of weights on the left side $\times$ five codes for positions of the right object $\times$ five codes for the number of weights on the right side). After sufficient training, the model is able to accurately produce outputs corresponding to whether the balance-beam would tip. If however, the model is trained only on problems with one, two or three weights on either side of the balance-beam, featurewise independence guarantees that the model cannot generalize correctly to problems in which one side contained four or five weights. Likewise, an expanded version of the model that was trained on problems with up to ten weights on either side could not distinguish a problem in which one side had 11 weights and the other had 12; it is plain the children's knowledge is much more freely generalized.

[12]It is possible to construct a structured connectionist model of the past tense that has distinct pathways for regular and irregular inflection, but such a model would not provide support for the sort of unstructured single-mechanism model that is advocated in *RI*. The important question is not whether one can build any connectionist model of the past tense, but rather, 'what architectural assumptions must be made in order to build an adequate model?'.

[13]'Categorization' networks like those discussed in Plunkett (1997) in which all a network has to do is choose which class a word belongs to (e.g. 'the regular class' versus 'the irregular class') evade this problem, but do so by assuming an external mechanism that combines the resulting output label ('regular') with the verb's stem in order to produce an inflected word. Hence these models effectively depend on external device that implements the very rule that earlier connectionist models aimed to eliminate, 'add -ed to the stem of any verb'.

6.5. Summary

I have discussed four of the primary models on which the authors of *RI* base their claims; the same problem pervades all of them: the models generalize well to items that overlap the training set, but the models are unable to generalize in the ways that humans to do items that include features that did not appear in the training set[14]. Similar arguments can be made for essentially all the connectionist models of higher level cognition that are proposed in *RI*. (These problems are not inherent to connectionism, but only to eliminative connectionism; other models that incorporate important aspects of symbol-manipulation fare better, see, for example, Holyoak and Hummel (1998)). In other words, it is plain that the models discussed in *RI*, which try to acquire aspects of the acquisition of syntax and of the notion of object permanence, cannot form an abstraction that is independent of the training examples; instead they rely on context-sensitive patterns of hidden unit activation that are not sufficiently abstract to support the sorts of generalizations that humans routinely make.

7. Cascade-correlation

Another connectionist approach used to defend constructivism and briefly discussed in *RI* is the cascade-correlation algorithm (Fahlman and Lebiere, 1990). For example, Mareschal and Shultz (1996) have recently proposed that cascade-correlation provides an 'escape from Fodor's (1980) critique of constructivist development'. In this section I show that the cascade-correlation architecture is vulnerable to essentially the same criticisms as the other models I described. This final section is somewhat technical; readers less interested in such details should feel free to skip to the concluding section.

7.1. Cascade correlation: background

A network that lacks hidden units cannot learn to represent functions that are not linearly separable, such as the simple logical function exclusive-or (XOR). XOR is a logical function that is true if and only if exactly one of its inputs is true. In a network, we can represent this by setting the Boolean value *true* to equal an activation level of 1.0. In an elegant proof, Minsky and Papert (1969) showed that there could be no set of weights that would allow a network that lacked hidden units to represent exclusive-or[15].

In contrast, networks that do have hidden units *can* represent functions like XOR;

[14]Note that even if inputs (or outputs) are encoded with 'distributed representations' in which each input (or output) consists of multiple features, some possible inputs may include features on which the model has not been trained. Because of training independence, the multilayer perceptrons described in *RI*-style would not be able to generalize to those untrained features in the ways that humans would.

[15]This proof assumes that an output unit computes a monotonically increasing function of the sum of its input.

indeed the fact that such networks can represent such functions is a major part of their popularity. The addition of hidden units allows a network to – given an arbitrarily large training sample – approximate arbitrary functions.

Cascade-correlation is an algorithm that, simplifying somewhat, is programmed to add hidden units dynamically, as necessary. The way the cascade-correlation works is roughly this:

1. the model starts with zero hidden units
2. the model trains its weights to do the best it can without hidden units
3. the model adds a hidden unit that helps the model perform better
4. the model trains its weights to do the best it can with the newly-supplemented model, repeating steps (3) and (4) until it manages to approximate the input/output function to some prespecified degree of accuracy.

The network essentially proceeds by adding epicycles to epicycles, as illustrated in Fig. 4. In this diagram, the model's goal is to learn the function described in Fig. 4a. It starts by positing a simple function (Fig. 4b); if that doesn't work, the model superimposes another function (Fig. 4c) to explain some of the remaining variance; if that doesn't work, the model can add yet another function (Fig. 4d), yielding the somewhat better approximation of the function shown in Fig. 4e. This process can be repeated indefinitely. Like a Taylor-series approximation, the model will eventually converge on the an adequate approximation, simply by adding enough lower-order wiggles. As Mareschal and Shultz (1996) astutely note, this ability to add hidden units guarantees that the models can always, given enough training examples, and enough time, approximate any function[16].

However, there is absolutely no guarantee that the solution that the model finds will be the same as the one human finds, that the model can find a solution given a realistic training regime, or that this model has anything to do with human developmental psychology. As Mareschal and Shultz concede 'although it can be shown theoretically that there exists a network that can reproduce any desired behavior, it is not clear that this network can be found, or the behavior learned by the network is in feasible time'. While it is empirically possible that human cognitive development depends on a function-approximation devices that functions by successively adding epicycles, to my knowledge no empirical evidence for that has thus far been proposed.

7.2. Cascade-correlation and constructivism

With these preliminaries in place, we can now consider the argument for con-

[16]Although progressively adding epicycles is a sure way to approximate any function to an arbitrary degree, the history of science shows us that adding epicycles does not necessarily yield the best description of the underlying function that nature uses. For example, to account for deviations from expected orbits, Ptolemy defended his geocentric astronomy by successively adding more and more epicycles – until the epicycles were eventually replaced by Copernicus's more elegant heliocentric view. The point here is that although an epicycle-adding process is guaranteed to converge on some solution, there is no guarantee that epicycle-adding is the best solution.

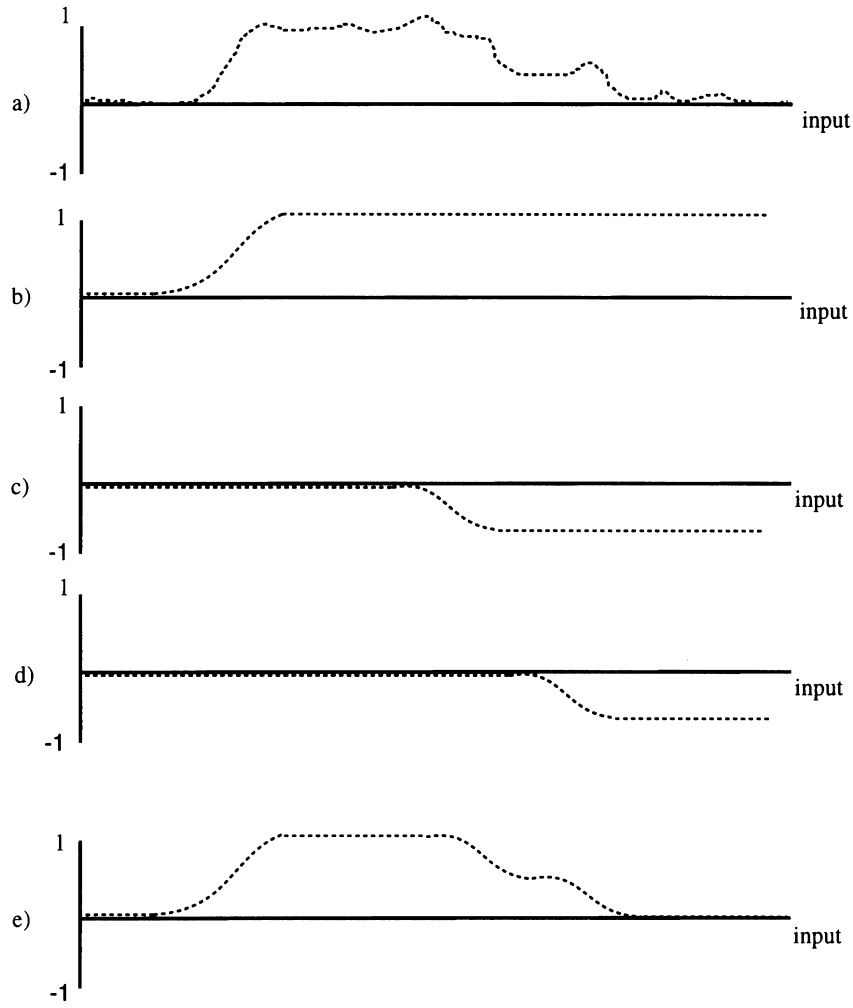


Fig. 4. The model aims to learn the function described in (a); see text for further explanation.

structivism put forward by Mareschal and Shultz (1996). At the initial state of a cascade-correlation model, the model has no hidden units; consequently, the model cannot represent, say, the exclusive-or function. If, however, after training, the model does not converge on a sufficiently adequate solution, the cascade-correlation algorithm ensures that the model will eventually add one or more hidden units, at which time the model will be able to represent exclusive-or. The idea, then, is that the initial 'baby' network cannot represent a function, XOR, that the 'adult' network can. Hence, they suggest, Fodor's claim that no device can construct new representations has been refuted.

This argument rests, however, on a confusion between the ordering of hypotheses

and the ability to construct new representations. What the cascade-correlation algorithm does is to force the model to consider hypotheses in which the function to be modeled is linearly separable before the model considers hypotheses in which the function is not linearly separable. The model does not however *learn* to posit functions that are non-linearly separable. Rather, the ability to posit such functions is *innately* given to the model – the algorithm specifies that (roughly) if linearly separable functions don't work, try positing functions that are not linearly separable. In other words, both the primitives and the combinatorial apparatus are, as always, innately given. Consequently, the cascade-correlation algorithm, like the back-propagation algorithm, does not provide a sound basis for constructivism.

More generally, models that use cascade-correlation are vulnerable to exactly the same criticisms as the other models discussed in this paper: they learn relationships between prespecified input and output representations, and they do not generalize in the ways that humans do to features that have not appeared in the training set.

8. Summary and discussion

To recap, the authors of *Rethinking Innateness* claim that neuroscientific evidence suggest that representations cannot be innate and further claim to have shown that

> domain-general architectures and learning algorithms can give rise to abstract domain-specific representations which ultimately become modularized as a the product of learning, not its starting point. (p. 170.)

In contrast to these claims, I have shown the following:

- The arguments against representational nativism are aimed at a straw version of nativism in which neurons are born knowing their final destinations.
- Each of the models discussed in *RI* presumes innate, domain-specific representations
- None of the models discussed in *RI* learns new kinds of representations
- Each of the models discussed in *RI* presumes one or more innate modules devoted to a single task
- None of the models discussed in *RI* uses learning to construct new modules
- Cognition depends on the existence of abstract representations; *RI*-style models lack the ability to form such representations and consequently do not generalize in some of the ways that humans do.

In sum, there is a gap between the ambitious promises made in *RI* – in which representations and modules will be learned rather than innate – and the more prosaic models that they actually describe, in which representations and modules are invariably prespecified, not learned. The fact that *RI* was unable to deliver models that are consistent with their stated goals does not doom constructivism, but their models do not provide any support for constructivism, either.

8.1. An alternative to constructivism

Some apparent cases of constructivism may turn out to be wrong, cases of competence conflated with performance. Rather than learning object permanence, for instance, children may be innately endowed with the (unconscious) knowledge that objects persist through time. Apparent dramatic changes in children's abilities might reflect increasing memory or processing resources on the part of the child (see Spelke and Newport (1998) for several examples of this sort of argument) or maturation of inhibitory mechanisms (Diamond, 1991). In some cases, however, children of two different ages may genuinely qualitatively differ in their understanding of the world. For instance, Carey (1985) argues that children are not innately endowed with a knowledge of biology. If Carey's arguments are correct, something that at least looks like a constructivist change must be occurring. I have argued that the kinds of models discussed in *Rethinking Innateness* could not account for such wholesale changes. What could?

One alternative possibility that may be worth future investigation is this: changes that prima facie appear to depend on dramatic, constructivist-like changes may instead involve either learning new rules that hold for all elements within some class, learning new classes over which rules can be restricted, or learning to extend or restrict a previously known rule to a new class of entities. On such a view, aside from any knowledge that might be innate, a child would minimally have a mechanism for forming rules, a mechanism for forming classes of items, and a way of representing and generalizing rules.

For example, consider what happens when a child playing 'Go Fish' for the first time learns the notion of a 'pair'. One way to learn the notion of a 'pair' is to compile an exhaustive list of specific items ('two aces form a pair; two twos form a pair; two threes form a pair, two fours form a pair, two fives form a pair, two sixes form a pair, two sevens form a pair, two eights form a pair, two nines form a pair, two tens form a pair, two jacks form a pair, two queens form a pair, and two kings form a pair'). A standard unstructured connectionist model like the one discussed above in Fig. 2 could serve as an implementation of such an approach.

There is good reason, however, to think that what we learn is more general than such a list: if we start playing a new game with a special deck containing four instances of a new card (call it a 'duke'), we can immediately infer that two dukes will count as a pair. This suggests that our representation of the concept of a pair might come in the form of an abstract rule, which is to say that it might be a relationship between two variables, say card *x* and card *y*, in which the value of some card *x* must equal the value of some card *y*. By representing the notion of a pair as relationship between variables, we could extend the relationship to new items.

Much of our knowledge of the world seems to come in the form of rules which we can extend to essentially arbitrary instances. If we know that Joe's last name is *N*, we can predict that (other things being equal) Joe's father will also have the last name *N*. We know that this rule can be extended to anything that can instantiate the variable *N*, which is to say that we know that this rule holds for all possible names.

My suggestion is that children's knowledge of domains such as biology is similar, which is to say that I am suggesting that children have the capacity to learn rules that hold for all entities in a domain, such as *living things* or *animals*. If a child represents a domain like animal, a child can then learn and represent a rule that holds for all animals. Conversely, if the child's notion of what counts as an *animal* changes (whether by being explicitly told that some new organism is an animal, or by some internal process of reanalyzing what it takes to be a animal), the domain to which a child applies some rule will automatically change. Constructing a theory of biology – in part a collection of rules – in the first place would thus depend on the initial construction of the class *living thing*.

In theories of language acquisition, these sorts of mechanisms (rules and classes) are taken for granted: once a 23-month-old child learns a rule about the syntax of nouns, for instance, she can automatically extend it to anything that she can identify as a noun (Tomasello and Olguin, 1993). Each rule that a child acquires must be restricted to some class of items. The classes may be narrow (plural nouns or transitive verbs) or broad (nouns). Acquiring language is a process of figuring out what the word classes are, which elements belong in which classes, and figuring out the relations between those classes. In the same way, much of cognitive development may be a process of figuring out what classes of entities there are, and what the relations are between those entities. Because rules and the domains that constrain rules are so tightly intertwined, changes in either might lead to dramatic changes in the child's behavior and in the child's understanding of the world.

Acknowledgements

I thank Tom Bever, Luca Bonatti, Chuck Clifton, Jerry Fodor, Giyoo Hatano, David Jensen, Steve Pinker, Bill Ramsey, Arnold Trehub and Zsofia Zvolenszky for helpful discussion. This research was partially supported by a Faculty Research grant from the University of Massachusetts.

Appendix A. Details of a test of the simple recurrent network

Method

The architecture I used in the experiment described in Section 6.1 was based on the architecture of Elman (1990). The network had 13 inputs units, 13 output units, a layer of 40 hidden units[17], and a layer of 40 context units; the weights from the hidden units to the context units were fixed at 1.0; all remaining weights were

[17]Transducers (i.e. additional hidden layers) were included in the models reported in Elman (Elman, 1991; Elman, 1993) but not the models reported in Elman (1990); since pilot testing indicated that transducer banks impaired the ability of the model to learn the training sets tested here, transducer banks were not included in the simulations reported here. Formal results presented in the training independence section prove that the limitations described in Section 6 do not depend on the presence or absence of the transducers.

randomly assigned and adjusted by the back-propagation algorithm in the course of training. The activation function of the context units was linear; for the remaining units, the activation function was sigmoidal. Each input node and each output node corresponded to a single word; on the input, a word was encoded as a single 1 embedded in a string of 0s; the output was interpreted as a likelihood vector (see Elman (1990) for details). The learning rate was 0.1, the momentum was 0.0. The contexts units were reset prior to the beginning of each training passage (e.g. before 'a rose is a rose').

There were two experimental conditions. In the 'full input' condition, the model was trained on 20 000 sentences that were randomly ordered repetitions of ten distinct instantiations of the frame 'a *q* is a *q*', and tested on those same ten sentences. In the 'partial input' condition, the network was trained on 20 000 sentences that were randomly-ordered repetitions of nine distinct instances of *x*, and tested on its ability to extend the relationship to the tenth instance. At the conclusion of training, learning was turned off (following Elman's procedures), and testing began. Responses were gathered at the point (defined by the grammar that produced the sentences) at which the variable should have recurred. For example, the model would be tested after the second 'a' in the sequence 'a rose is a...'.

If the network's most active unit at that point was the correct instantiation of the variable (in the example, the unit corresponding to *rose*) the response was scored as correct; otherwise, the network's response was scored as incorrect. (Since each output is a string of 0s with a single 1, the most activated output unit is necessarily the closest output pattern in Euclidean space.)

Each instantiation of the variable (i.e. the ten trained words and the ten novel words) was tested in a separate test run. Context units were reset at the beginning of each test run.

Results

For every trained item, the model successfully predicted the reappearance of the instantiation of the variable. For example, given the sentence fragment 'a house is a...', the model correctly predicted the continuation *house*. Thus, in the full-input condition, the model mastered the training set. The mean activation of the second instance of the variable, for trained items was 0.958; the mean of the other possible continuations was less than 0.050. For trained items in the partial-input condition, the mean activation of the correct continuation was 0.956; the mean of the other possible continuations was less than 0.050.

In contrast, although the model was trained on a grammar that consisted entirely of sentences like *a house is a house*, it could not use distributional cues to predict the reappearance of *blicket* in the frame *a blicket is a*...; the correct continuation was not activated (its activation level was less than 0.008, far less than the mean of the other (i.e. incorrect) continuations, 0.175, and even farther from the above-mentioned mean of 0.956 for the activations of the correct continuations for items in the training set). The results, then, provide an illustration of that fact that the SRN does not extend the generalization in a human-like way outside the training space.

References

Anderson, J.A., 1995. An Introduction to Neural Networks. MIT Press, Cambridge, MA.
Barnden, J.A., 1984. On short-term information processing in connectionist theories. Cognition and Brain Theory 7, 285–328.
Bechtel, W., Abrahamsen, A., 1991. Connectionism and Mind: an Introduction to Parallel Processing in Networks. Basil Blackwell, Cambridge, MA.
Bloom, P., Wynn, K., 1994. The real problem with constructivism. Behavioral and brain sciences 17, 707–708.
Burgess, N., Hitch, G.J., 1992. Toward a network model of the articulatory loop. Journal of memory and language 31, 429–460.
Carey, S., 1985. Constraints on semantic development. In: Mehler, J., Fox, R. (Eds.), Neonate Cognition: Beyond the Blooming Buzzing Confusion. Erlbaum, Hillsdale, NJ.
Clark, A., Thornton, C., 1997. Trading spaces: computation, representation, and the limits of uninformed learning. Behavioral and Brain Sciences 20, 57–90.
Daugherty, K., Seidenberg, M., 1992. Rules or connections? The past tense revisited. In: Proceedings of the 14th annual meeting of the cognitive science society. Erlbaum, Hillsdale, NJ, pp. 149–156.
Daugherty, K.G., MacDonald, M.C., Petersen, A.S., Seidenberg, M.S., 1993. Why no mere mortal has ever flown out to center field but people often say they do. In: Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society. Lawrence Erlbaum, Hillsdale, NJ, pp. 383–388.
Dawkins, R., 1987. The Blind Watchmaker. Norton, New York.
Denker, J., Schwartz, D., Wittner, B., Solla, S., Howard, R., Jackel, L., Hopfield, J., 1987. Large automatic learning, rule extraction, and generalization. Complex Systems 1, 877–892.
Diamond, A., 1991. Neuropsychological insights into the meaning of object concept development. In: Carey, S., Gelman, R. (Eds.), The Epigenesis of Mind. Lawrence Erlbaum, Hillsdale, NJ.
Elman, J.L., 1990. Finding structure in time. Cognitive Science 14, 179–121.
Elman, J.L., 1991. Distributed representations, simple recurrent networks, and grammatical structure. Machine Learning 7, 195–224.
Elman, J.L., 1993. The importance of starting small. Cognition 48, 71–99.
Elman, J.L., Bates, E., Johnson, M.H., Karmiloff-Smith, A., Parisi, D., Plunkett, K., 1996. *Rethinking Innateness*: a Connectionist Perspective on Development. MIT Press, Cambridge, MA.
Fahlman, S.E., Lebiere, C., 1990. The cascade-correlation learning architecture. In: Touretzky, D.S. (Ed.), Advances in Neural Information Processing Systems 2. Morgan Kaufmann, Los Angeles, pp. 38–51.
Fodor, J.A., 1975. The Language of Thought. T.Y. Crowell, New York.
Fodor, J.A., 1981. On the Present Status of the Innateness Controversy. In: Representations. MIT Press, Cambridge, MA, pp. 257–316.
Fodor, J.A., 1983. Modularity of Mind. MIT Press Cambridge, MA.
Fodor, J.A., Pylyshyn, Z., 1988. Connectionism and cognitive architecture: a critical analysis. Cognition 28, 3–71.
Geman, S., Bienstock, E., Doursat, R., 1992. Neural networks and the bias/variance dilemma. Neural computation 4, 1–58.
Hare, M., Elman, J., 1995. Learning and morphological change. Cognition 56, 61–98.
Hare, M., Elman, J., Daugherty, K., 1995. Default generalisation in connectionist networks. Language and Cognitive Processes 10, 601–630.
Hartley, T., Houghton, G., 1996. A linguistically constrained model of short-term memory for nonwords. Journal of memory and language 35, 1–31.
Hinton, G.E., Dayan, P., Frey, B., Neal, R.M., 1995. The wake–sleep algorithm for self-organizing neural networks. Science 268, 1158–1160.
Holyoak, K.J., Hummel, J.E., 1998. The proper treatment of symbols in a connectionist architecture. In: Deitrich, E. Markman, A. (Eds.), Cognitive Dynamics: Conceptual Change in Humans and Machines. MIT Press, Cambridge, MA, in press.
Hummel, J.E., Holyoak, K.J., 1997. Distributed representations of structure: a theory of analogical access and mapping. Psychological Review 104, 427–466.

Jacobs, R.A., Jordan, M.I., Barto, A.G., 1991. Task decomposition through competition in a modular connectionist architecture: the what and where vision tasks. Cognitive Science 15, 219–250.

Jacobs, R.A., Kosslyn, S., 1994. Encoding shape and spatial relations: the role of receptive field size in coordinating complementary representations. Cognitive Science 18, 361–386.

Karmiloff-Smith, A., 1996. The Connectionist Infant: Would Piaget Turn in His Grave? Society for Research in Child Development Newsletter, Fall 1996, 1–10.

Katz, L.C., Shatz, C.J., 1996. Synaptic activity and the construction of cortical circuits. Science 274, 1133–1138.

Kim, J.J., Marcus, G.F., Pinker, S., Hollander, M., Coppola, M., 1994. Sensitivity of children's inflection to grammatical structure. Journal of Child Language 21, 173–209.

Kolen, J.F., Goel, A.K., 1991. Learning in parallel distributed processing networks: computational complexity and information content. IEEE Transactions on Systems, Man, and Cybernetics 21, 359–367.

Lachter, J., Bever, T.G., 1988. The relation between linguistic structure and associative theories of language learning: a constructive critique of some connectionist learning models. Cognition 28, 195–247.

Ling, C.X., Marinov, M., 1993. Answering the connectionist challenge: a symbolic model of learning the past tense of English verbs. Cognition 49, 235–290.

Linsker, R., 1988. Self-organization in a perceptual network. Computer 21, 105–129.

MacWhinney, B., Leinbach, J., 1991. Implementations are not conceptualizations: revising the verb learning model. Cognition 40, 121–157.

Marchman, V., 1993. Constraints on plasticity in a connectionist model of the English past tense. Journal of Cognitive Neuroscience 5, 215–234.

Marcus, G.F., 1995. The acquisition of inflection in children and multilayered connectionist networks. Cognition 56, 271–279.

Marcus, G.F., 1996. What does it take to get a connectionist model to generalize a low-frequency default? Submitted.

Marcus, G.F., 1997. Rethinking eliminative connectionism. Submitted.

Marcus, G.F., Brinkmann, U., Clahsen, H., Wiese, R., Pinker, S., 1995. German inflection: the exception that proves the rule. Cognitive Psychology 29, 186–256.

Mareschal, D., Shultz, T.R., 1996. Generative connectionist networks and constructivist cognitive development. Cognitive Development 11(4), in press.

McClelland, J.L., 1989. Parallel distributed processing: implications for cognition and development. In: Morris, R.G.M. (Ed.), Parallel Distributed Processing: Implications for Psychology and Neurobiology. Oxford University Press, Oxford, pp. 9–45.

McClelland, J.L., Rumelhart, D.E., 1986. A distributed model of human learning and memory. In: McClelland, J.L., Rumelhart, D.E. and the PDP Research Group (Eds.), Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 2: Psychological and Biological Models. MIT Press, Cambridge, MA, pp. 170–215.

Minsky, M.L., Papert, S.A., 1969. Perceptrons. MIT Press, Cambridge, MA.

Munakata, Y., McClelland, J.L., Johnson, M.H., Siegler, R.S., 1998. Rethinking infant knowledge: toward an adaptive process account of successes and failures in object permanence tasks. Psychological Review, in press.

Pinker, S., 1991. Rules of language. Science 253, 530–555.

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

Plunkett, K., 1997. Theories of early language acquisition. Trends in Cognitive Sciences 1, 146–151.

Plunkett, K., Marchman, V., 1991. U-shaped learning and frequency effects in a multi-layered perceptron: implications for child language acquisition. Cognition 38, 43–102.

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

Prasada, S., Pinker, S., 1993. Similarity-based and rule-based generalizations in inflectional morphology. Language and Cognitive Processes 8, 1–56.

Regier, T., 1995. A model of the human capacity for categorizing spatial relations. Cognitive Linguistics 6, 63–88.

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

Seidenberg, M.S., McClelland, J.L., 1989. A distributed, developmental model of word recognition and naming. Psychological Review 96, 523–568.

Shallice, T., Glasspool, D.W., Houghton, G., 1995. Can neuropsychological evidence inform connectionist modeling? Analyses of spelling. Language and Cognitive Processes 1995, 195–225.

Shastri, L., Ajjanagadde, V., 1993. From simple associations to systematic reasoning: a connectionist representation of rules, variables, and dynamic bindings using temporal synchrony. Behavioral and Brain Sciences 16, 417–494.

Smolensky, P., 1995. Reply: constituent structure and explanation in an integrated connectionist/symbolic cognitive architecture. In: Macdonald, C., Macdonald, G. (Eds.), Connectionism: Debates on Psychological Explanation. Basil Blackwell, Oxford.

Spelke, E.S., 1994. Initial knowledge: six suggestions. Cognition 50, 431–445.

Spelke, E.S., Newport, E.L., 1998. Nativism, empiricism, and the development of knowledge. In: Lerner, R. (Ed.), Handbook of Child Psychology, Vol. 1: Theories of Development. Wiley, New York, in press.

Sun, R., 1992. On variable binding in connectionist networks. Connection Science 4, 93–124.

Tomasello, M., Olguin, R., 1993. Twenty-three-month-old children have a grammatical category of noun. Cognitive Development 8, 451–464.

Touretzky, D.S., Hinton, G.E., 1985. Symbols among the neurons. In: Proceedings IJCAI-85, Los Angeles, CA.

Wexler, K., 1991. On the arguments from the poverty of the stimulus. In: Kasher, A. (Ed.), The Chomskyan Turn. Basil Blackwell, Oxford.