Proceedings of the Twelfth Annual Conference of the Cognitive Science Society, pp. 340-347.

Why Fodor and Pylyshyn Were Wrong: The Simplest Refutation

David J. Chalmers
Center for Research on Concepts and Cognition
Indiana University

Abstract

This paper offers both a theoretical and an experimental perspective on the relationship between connectionist and Classical (symbol-processing) models. Firstly, a serious flaw in Fodor and Pylyshyn's argument against connectionism is pointed out: if, in fact, a part of their argument is valid, then it establishes a conclusion quite different from that which they intend, a conclusion which is demonstrably false. The source of this flaw is traced to an underestimation of the differences between localist and distributed representation. It has been claimed that distributed representations cannot support systematic operations, or that if they can, then they will be mere implementations of traditional ideas. This paper presents experimental evidence against this conclusion: distributed representations can be used to support direct structure-sensitive operations, in a manner quite unlike the Classical approach. Finally, it is argued that even if Fodor and Pylyshyn's argument that connectionist models of compositionality must be mere implementations were correct, then this would still not be a serious argument against connectionism as a theory of mind.

Introduction

The trenchant critique by Fodor and Pylyshyn (1988) threw a scare into the field of connectionism, at least for a moment. Two distinguished figures, from the right side of the tracks, were bringing the full force of their experience with the computational approach to cognition to bear on this young, innocent field. It was enough to get anybody worried for a while. But after the initial flurry, connectionists gradually settled down to the view that while Fodor and Pylyshyn had posed a challenge for the field, it was certainly not an unanswerable one. A spate of "refutations" quickly followed. These generally took two forms: argument (e.g. Clark 1989, Smolensky 1987, van Gelder 1990), or counterexample (Elman 1990, Pollack 1990, Smolensky 1990). (One is reminded of Nietzsche's observation: "It is not the least charm of a theory that it is refutable.")

The point of this paper is to offer a few observations on the whole business. The primary purpose is to offer a particularly simple refutation of Fodor and Pylyshyn's argument that I do not believe has been presented elsewhere. Straightforward considerations about the structure of their argument will show that it cannot have succeeded in its intended purpose. Furthermore, simple as these considerations are, they lead into deeper issues about just *why* their argument was wrong, and about the vital properties of connectionist models that were not taken into account. In particular, the role of *distributed representation* will be gone into. The ability of distributed representations to support structure-sensitive operations will be demonstrated with some experimental results. Finally, this will lead into the issue of the possible *implementation* of Classical ideas by connectionist models, and expose the shortsightedness of some of Fodor and Pylyshyn's claims here.

Refutation

Recall the major thrust of Fodor and Pylyshyn's argument: that connectionist models cannot admit of a compositional semantics. Or, more accurately, not unless they are an implementation of a Classical architecture. Manifestations of compositional semantics are certainly ubiquitous in our thought, particularly in our language, through its *compositionality* (the meaning of "the girl loves John" is a function of the meaning of its constituent parts, "the girl", "loves", and "John"),

and its *systematicity* (the ability to think "John loves the girl" is tied to the ability to think "the girl loves John"). So if connectionism cannot handle compositional semantics, then that's a problem for connectionism.

The refutation of F&P's argument can be stated in one sentence, then explained. *If F&P's argument is correct as it is presented, then it implies that no connectionist network can support a compositional semantics; not even a connectionist implementation of a Turing Machine, or of a Language of Thought.* But this is a problem for F&P, as it is well-known that connectionist networks can be used to implement Turing Machines (or at least Turing Machines with arbitrarily large but finite tape), and it is well-known that Turing Machines can be used to support a compositional semantics. Furthermore, the *human brain* is like a connectionist network in many ways, and the human brain certainly supports a compositional semantics. So if F&P's argument really establishes that *no* connectionist network can support a compositional semantics, then it establishes a false conclusion. So, applying the contrapositive of the italicized sentence above, F&P's argument is not correct as it stands.

Of course, Fodor and Pylyshyn do not *want* to imply such a conclusion. Indeed, they take great care to point out that the best future for connectionism will lie in using it as an *implementation* strategy. Connectionist implementations of Classical systems will certainly support a compositional semantics, if not in a particularly interesting way. Well and good; of course they must say such a thing: it may be slightly embarrassing that the brain is made of neurons and not directly out of symbolic structures, but it is a *fact*, and as a fact it must be dealt with. But what they *say* is one thing. Their actual *argument* is a different matter.

The substantive argument in F&P's paper, that connectionist models cannot support a compositional semantics, takes up only a few pages (pp. 15-28). This starts with a simple localist connectionist network (that is, a network with one node representing one concept). F&P show that this network cannot possibly possess a compositional semantics, and argue that this applies equally to networks with distributed semantics (that is, a network with one concept being represented over many nodes). Therefore, the argument concludes, it is impossible for the semantics of a connectionist network to be compositional, whether these semantics are localist or distributed.

There is something very strange about this conclusion. It is plainly false; it is universally recognized that *some* connectionist networks have compositional semantics: namely, connectionist implementations of Classical architectures. So why are these not excluded from the argument? Going through the argument, the reader expects that at any point soon, there will be an *escape clause* — a clause showing why the argument as it stands does not apply to connectionist implementations of Classical architectures. But this clause never appears; nothing close to it, in fact. F&P are left in the improbable position of having "proved" that even connectionist implementations of Classical models have no compositional semantics. Faced with such a situation, we can only conclude that the argument is defective. Supporters of F&P might argue that the flaw simply lies in the lack of an escape clause, which can easily be supplied; but no such escape clause is in evidence, and the onus lies with these people to provide it. In the meantime, we can conclude that the defect lies elsewhere: very likely, in the generalization from localist to distributed semantics. More on this in a moment, after an analogy.

Say a mad scientist comes up to us with a "proof" that the Earth is the only inhabited planet in the universe. She runs through an impressive *a priori* argument, showing why it is impossible that the right kinds of biochemicals could be assembled in the right way, that the requisite organizational complexity could not arise, and so on. She concludes: life could not have arisen on any planet in this universe. But then, of course, it is an obvious fact that life arose on Earth. "That's OK," she answers, "that suits me fine. We knew that already. So what I've established is that life cannot have arisen anywhere but Earth." Now this will strike us as ad hoc, and as

extremely poor logic. Their main argument never mentioned Earth; there was no *escape clause* showing just why the argument doesn't apply to Earth. To modify the conclusions of one's argument by considerations *external* to the argument is to admit that the argument is faulty. ("Mars is inhabited? OK, our argument demonstrates that life cannot have arisen anywhere but Earth or Mars.") If the argument can be fixed so that Earth is excluded from its force, very likely other planets will be excluded also. Analogously: if F&P's argument can possibly be fixed up so that it excludes Classical implementations from the scope of its conclusion, then the same fixes will probably exclude many other connectionist models too.

Refuting Fodor and Pylyshyn in Four Easy Steps

All this has been a long-winded way of making the following simple argument:

(1) In F&P's argument that no connectionist models can have compositional semantics, there is no escape clause excluding certain models (such as Classical implementations) from the force of the conclusion. (By observation.)

(2) If F&P's argument is correct as it stands, then it establishes that *no* connectionist model can have compositional semantics. (From (1).)

(3) But some connectionist models obviously *do* have compositional semantics; namely, connectionist implementations of classical models. (By observation, accepted by all.)

(4) Therefore, F&P's argument is not correct as it stands. (From (2), (3).)

Summing things up: Let C denote the class of all possible connectionist models, together with all possible associated semantics. Let FP denote the subset of C of models whose semantics are not compositional. Let L denote the subset of C consisting of models with localist semantics. Let IMP denote the subset of C consisting of connectionist implementations of Classical models. The conclusion that F&P *want* to establish is that FP = C – IMP.

In their argument, F&P first establish that L ≤ FP. (Here "≤" denotes set inclusion.) Let us grant them this, though some might argue. They then argue that it makes no difference whether the semantics are localist or distributed. Now, clearly the two possibilities of localist and distributed semantics exhaust the set C, so this argument, if correct, establishes that FP = C. But this is plainly false, as IMP < C but it is not the case that IMP < FP.

We may conclude that *all* F&P have established is that L ≤ FP ≤ C – IMP. The step in the argument that generalizes to *all* distributed semantics is plainly defective. Although F&P would like to hold that it generalizes to all distributed semantics *except* those used to implement Classical models, the burden rests with them to show that this is the case. The conclusion established is a much weaker statement than FP = C – IMP. As things stand, it is just as likely that FP = L as that FP = C – IMP, though no doubt the truth lies somewhere in the middle.

Localist and Distributed Representation

So far, we have given a simple logical demonstration that F&P's argument must be flawed. It remains to precisely locate the weak spot in the argument. Fortunately, this is not hard to do. To find this, we must think about just *why* certain models, implementations and possibly others, slip through the argument's net. By now, no doubt, supporters of F&P are lining up in droves, waiting to say: "But of *course* the argument doesn't apply to implementations of Classical models. Implementations are *different* — the representations of Classical symbols in such a network will not exist at the level of the *node*, but at a much higher level. These symbols will be

able to combine compositionally and autonomously." To such a person we might reply "Congratulations! You have just discovered the power of distributed representation."

Many connectionists have noted that the small localist network that F&P used as their chief example was most unrepresentative of the connectionist endeavour of a whole. When one asks what is the deepest philosophical commitment of the connectionist movement, the answer is surely this: the rejection of the atomic symbol as the bearer of meaning. Connectionists feel that atomic tokens simply do not carry enough information with them to be useful in modeling human cognition. Rather, distributed, subdivisible, malleable representations are the cornerstone of the connectionist endeavour. For this reason, localist networks are regarded by many connectionists as not really connectionist at all. These networks employ precisely the traditional notion of atomic symbols, with a new twist added by connecting all of these by associative links. (We might thus call localist connectionism "symbolic AI with soft constraint satisfaction.")

The use of a localist network by F&P, then, betrays a lack of understanding of the connectionist endeavor. They believe that nothing depends on the localist/distributed distinction; the connectionist, on the other hand, believes that everything depends on it. To F&P, a connectionist distributed representation is just a spread-out version of a single node (this comes out clearly in the footnote to p. 15). To the connectionist, a group of nodes functioning separately has functional properties far beyond those of an isolated unit. Small differences in the activity of a subset of nodes can make subtle or unsubtle differences to later processing, in a way that no single node can manage. A group of nodes carries far more *information* than a single node, and as such to the connectionist is a far more likely candidate for for semantic interpretation. And most importantly, a distributed representation has a great deal of internal structure. (The point that Fodor and Pylyshyn underestimate the power of distribution is by no means original. It was first made by Smolensky (1987).)

Before moving on, we should briefly examine F&P's demonstration of why their argument applies equally to localist and distributed networks. This will be brief, as the relevant material is brief. On the bottom of p. 15, we find

> To simplify the argument, we assume a more 'localist' approach, in which each semantically interpreted node corresponds to a single Connectionist unit; but nothing relevant to this discussion is changed if these nodes actually consist of patterns over a cluster of units.

No argument to be found there. And later (p. 19)

> To claim that a node is neurally distributed is presumably to claim that its states of activation correspond to changes in neural activity — to aggregates of neural 'units' — rather than to activations of single neurons. The important point is that nodes that are distributed in this sense can perfectly well be syntactically and semantically atomic: Complex spatially-distributed implementation in no way implies constituent structure.

No-one will begrudge F&P this passage. As it stands, it is perfectly true. But it would only be interesting as argument if the last two sentences changed so that the "can" became a "must" and the "in no way implies" became "forbids". But it is precisely this that F&P cannot establish. We can conclude that their argument against distributed representation (and this is the extent of it) is weak. F&P go on to argue against connectionist models whose semantics are "distributed over microfeatures". But, as elsewhere, the kinds of semantics they consider bear little resemblance to those found anywhere in connectionism. This is the fundamental flaw in F&P's argument: lack of imagination in considering the possible ways in which distributed representations can carry semantics. It is a different variety of distributed semantics that would be carried by a connectionist implementation of a Turing Machine (and this, then, accounts for the logical flaw

detailed above.) And it is a different variety again of distributed semantics that can yield connectionist models of compositionality in important new ways.

It is no accident that three of the most prominent *counterexamples* to F&P's argument — the models of Elman, Pollack, and Smolensky — all use distributed representation in an essential way. Smolensky's tensor-product architecture simply could not work in a localist framework. Its multidimensional tensor representations are by their nature spread over many nodes. Elman's implicit structure which develops in a recurrent network could also not succeed in a localist framework — the many subtle adjustments needed for various syntactic distinctions to develop could not be made. And Pollack's Recursive Auto-Associative Memory has a deep commitment to distribution — if it were one-concept-to-one-node, then its recursive encoding scheme could never get off the ground.

Structure-Sensitive Operations on Distributed Representations

The Classicist might now reply: "All this talk of distributed representations is all very well. Maybe you can *encode* compositional information into such a representation. But can you *use* it?" This point is initially plausible. If the structural information is present but cannot be processed, then it is useless. The Classicist might hold that connectionist compositional structure might be buried too deeply, too implicitly, to be accessed in a useful way. Indeed, in a recent paper, Fodor and McLaughlin (1990) argue that to support structure-sensitive processing, a compositional representation must be a concatenation of explicit tokens of the original constituent parts. If this argument is correct, then connectionist representations that represent structure only in a distributed, implicit way will not have the causal power to support structure-sensitivity.

One obvious reply that the connectionist might make is that clearly *some* structure-sensitive operations can be supported by such representations: namely, the operation of extraction of the original constituents. Both Smolensky's and Pollack's models, for instance, include decodal processes that go from a compositional representation back to its parts. This reply, while valid, is not very interesting. If structure-sensitive processing must always proceed through an initial stage of decomposition into constituents, then what we are dealing with is essentially a connectionist implementation of a Classical symbol processing. In such processing, distributed representation is used as a mere implementational technique.

Fortunately, this is not always the case. In fact, distributed representations of compositional structure *can* be operated on directly, without proceeding through an extraction stage. This offers the promise of a connectionist approach to compositionality that is in no sense an implementation of the Classical notion. (It should be noted that Pollack and Smolensky have addressed this issue briefly in their models, but in a more limited way than outlined below.)

I have performed a series of experiments demonstrating the possibility of effective structure-sensitive operations on distributed representations. I can only outline them very briefly here; they are presented in more detail in (Chalmers, 1990). The experiments used a Recursive Auto-Associative Memory (RAAM; see Pollack 1988, 1990) to encode syntactically structured

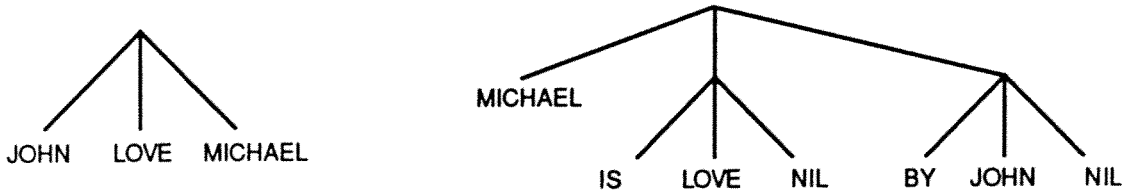

Figure 1. Examples of sentences to be represented.

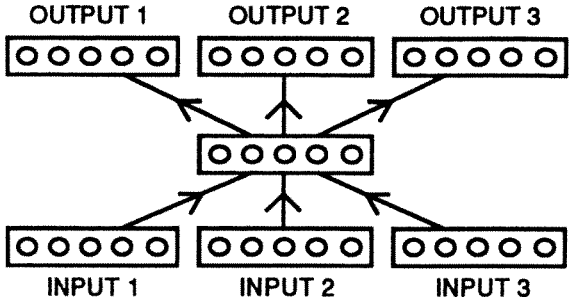

Figure 2. The basis of the RAAM network.

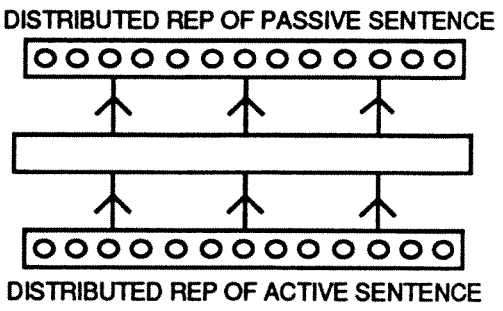

Figure 3. The Transformation network.

representations of sentences in distributed form. Following this, a back-propagation network learned to perform syntactic transformations directly from one encoded representation to another.

The sentences represented were all of similar syntactic form to "John loves Michael" (active) or "Michael is loved by John" (passive). Five different names/verbs were used as fillers for each slot of subject, verb or object, giving 125 possible sentences of each type altogether. These sentences were assigned syntactic structure as shown in Figure 1. A RAAM network was trained to encode 125 sentences of each kind into a distributed form. (Pollack 1990 gives details of the RAAM architecture.) This is done by assigning each word a primitive localist representation (over 13 units), and then training a 39-13-39 backpropagation network (Figure 2) to auto-associate on the three leaves descending from every internal "node" (in the trees in Figure 1).

This gives us a 13-node distributed representation of the three leaves. Where necessary, this 13-node distributed representation is repropagated as part of the input to the 39-13-39 network, leading to higher-order structures being encoded. Eventually, we have a distributed representation of the entire tree. This process can be used, in principle, to encode any tree of valence 3 recursively.

The RAAM network learned to represent all 250 sentences satisfactorily, so that the distributed encodings of each sentence could be decoded back to the original sentence. These distributed representations were then used in modeling the process of syntactic transformation. In particular, the transformation of *passivization* was modeled: that is, the passing from sentences like "John loves Michael" to sentences like "Michael is loved by John". (No commitment to any particular linguistic paradigm is being made here; syntactic transformations are simply being used as a clear example of the kind of structure-sensitive operations with which connectionist models are supposed to have difficulty.)

150 of the encoded distributed representations (75 active and and the corresponding 75 passive sentences) were randomly selected for the training of the Transformation Network. This was a simple 13-13-13 backpropagation network (Figure 4), which took a representation of an active ("John loves Michael") sentence as input, and was trained to produce a representation of the corresponding passivized sentence ("Michael is loved by John") as output.

Training proceeded satisfactorily. The interesting part was the test of generalization, to see if the network was truly sensitive to the syntactic structure encoded in the distributed forms. The Transformation network was tested on the 100 remaining sentences from the original corpus. The 50 active sentences were encoded by the RAAM and fed to the Transformation network, yielding a 13-node output pattern. This was fed to the RAAM network for decoding. In all 50 cases, the output pattern decoded to the correct passivized sentence. Thus, not only was the Transformation network able to be trained to optimal performance, but the generalization rate on

new sentences of the same form was 100%. The reverse transformation was also modeled (from passive to active). Performance was equally good, with a generalization rate of 100%.

These results establish without doubt that it is possible for connectionist networks to model structure-sensitive operations directly upon distributed representations. This bears on the arguments at hand in two ways.

(1) It demonstrates that not only can compositional structure be *encoded* in distributed form, but that the structure implicitly present within the distributed form can be *used* directly for further processing. This provides a direct counterexample to the Fodor and McLaughlin argument. Despite the lack of explicit concatenative structure in the RAAM representations, they support structure-sensitive processing anyway.

(2) It demonstrates the possibility of structure-sensitive operations in connectionist models which are in no sense implementations of Classical algorithms. To see this, note that when a structure-sensitive operations is being performed upon a Classical compositional representation, all processing *must* first proceed through a step of explicit decomposition, with particular tokens being explicitly extracted. In the connectionist model above, the transformation operation takes place without ever having to extract those constituent parts. Instead, the operation is direct and holistic.

The Relationship Between the Approaches

A argument made frequently by Fodor and Pylyshyn is that connectionists have two choices: either (1) ignore the facts of compositionality and systematicity, and thus have a defective theory of mind, or (2) accept compositionality and systematicity, in which case connectionism merely becomes a strategy for implementing Classical models. The following passage is typical:

> ...if you need structure in mental representations anyway to account for the productivity and systematicity of minds, why not postulate mental processes that are structure sensitive to account for the coherence of mental processes? Why not be a Classicist, in short? [p. 67]

This argument is rather curious. It is not only that it contradicts the evidence, demonstrated above, that connectionism might model structure-sensitive processes in a non-Classical way. There is also a deeply-embedded false assumption here: the assumption that *compositionality is all there is.*

To see the role that this assumption plays, shift the temporal position of the debate back a few decades. Let us imagine two traditional behaviorists, Fido and Pavlovian, who are rather distressed at the current turn of events. The revolutionary "cognitivists" have recently appeared on the screen, and are doing their best to undermine the basic assumptions of decades of solid research in psychology. Our behaviorists have difficulty grasping the idea of this movement. They express their bewilderment as follows: "Surely you all recognize that Classical Conditioning is a fact of human nature. The empirical evidence is overwhelming. But your cognitivist ideas do not take it sufficiently into account. There is no guarantee of stimulus-response association in your models as they stand. It seems to us that you have two choices: either (1) ignore the facts of Classical Conditioning, and therefore have a defective theory of mind, or (2) accept Conditioning and stimulus-response association, in which case cognitivism merely becomes a strategy for implementing the Behaviorist agenda."

Presumably a cognitivist (such as Fodor or Pylyshyn) would quickly see the flaw in this argument. To be sure, Conditioning is an empirical fact, and any complete theory must account for it. But it's certainly not the *only* fact, or even the most important fact, about the human mind. The cognitivists may pursue their own research agenda, making progress in many areas, and

paying as much or as little attention to Conditioning as they like. Eventually they will have to come up with some explanation of the phenomenon, and who knows, it may well end up looking much like the Behaviorist story, *as far as conditioning is concerned.* But this doesn't mean that the cognitivist theory of *mind* looks much like the behaviorist theory overall, for the simple reason that *conditioning is only one part of the story.*

Similarly, compositionality is only one part of the story. Connectionists are free to pursue their own agenda, explaining various aspects of the mind as they see fit. Sooner or later, they will have to explain how compositionality fits into the picture. The story that connectionism tells about compositionality may prove quite similar to the Classical picture, or it may prove different. But even if it proves similar, this diminishes the status of connectionism not at all. The fact that connectionism might implement Classical theories of *compositionality* does not imply that connectionism would be implementing Classical theories of *mind.* Compositionality is just one aspect of the mind, after all. (Aspects of cognition for which compositionality seems relatively unimportant include: perception, categorization, motor control, memory, similarity judgments, association, attention, and much more. Even within language processing, compositionality is only part of the story, albeit an important part.)

Behaviorism was very good at explaining conditioning, but it had a problem: it was *only* good at explaining conditioning. Fodor and Pylyshyn's Classicism is good at explaining compositionality and compositional semantics, but it's not necessarily good at explaining much else. Both conditioning and compositionality are only small aspects of the mind; it seems to be an illusion of perspective that led to behaviorists and Classicists putting so much respective emphasis on them.

Fodor and Pylyshyn's arguments establish that compositionality *exists*, but for their arguments above to succeed, they would need to establish a rather stronger claim: that compositionality is *everything*. Such a claim is obviously false, so connectionism can go on happily trying to explain those areas of the mind that it chooses to. If the connectionist story about compositionality ends up looking a little like the Classical story, then well and good — it implies that the Classicists haven't been wasting their time completely all these years, and there may be room for a healthy amount of ecumenicism. In the meantime, preemptive relegation of either approach to a subsidiary role is probably a bad idea.

Acknowledgements: Thanks to Indiana University for support, and to Bob French, Liane Gabora and Doug Hofstadter for comments.

References

Chalmers, D. J. (1990). Syntactic transformations on distributed representations. *Connection Science*, 2: 53-62.

Clark, A. (1989). *Microcognition.* Cambridge, MA: MIT Press.

Elman, J. L. (1990). Structured representations and connectionist models. In Gerald Altmann (ed.), *Computational and Psycholinguistic Approaches to Speech Processing.* New York: Academic Press.

Fodor, J. A.,& Pylyshyn, Z. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28: 3-71.

Fodor, J.A., & McLaughlin, B. P. (1990). Connectionism and the problm of systematicity: Why Smolensky's solution doesn't work. *Cognition*, 35: 183-204.

Pollack, J. B. (1988). Recursive auto-associative memory: Devising compositional distributed representations. In *Proceedings of the Tenth Annual Conference of the Cognitive Science Society.* Montreal, Canada, pp. 33-39.

Pollack, J. B. (1990). Recursive distributed representations. *Artificial Intelligence*, forthcoming.

Smolensky, P. (1987). The constituent structure of connectionist mental states: A reply to Fodor and Pylyshyn. *Southern Journal of Philosophy*, 26: 137-163.

Smolensky, P. (1990). Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence*, forthcoming.

Van Gelder, T. (1990). Compositionality: A connectionist variation on a Classical theme. *Cognitive Science*, 14.