EXCHANGE

On the Nature of Connectionist Conceptualizations and Connectionist Explanations

Foong-Ha Yap
Kanto Gakuin University
Yasuhiro Shirai
Daito Bunka University

In previous issues of *IAL* (cf. Fantuzzi, 1992, 1993), it was argued that connectionist explanations are too vague to qualify as theories of cognitive functions. Much of the argument hinges on the claim that hidden unit activation patterns of connectionist networks are currently too difficult to analyze, and that such opacity renders connectionist accounts virtually ineffective. However, recent attempts at analyzing connectionist hidden units using statistical techniques such as hierarchical clustering and principal component analysis reveal that connectionist networks easily yield categories which we traditionally associate with constituent structures (Elman, 1990). In this paper, we will focus on Elman's statistical analyses of the hidden unit activation patterns in his simple recurrent network on sentence prediction, first to highlight the feasibility of such analyses, and then to show how connectionist explanations contribute to the development of effective explanatory theories. In addition, in response to Fantuzzi's (1993) argument that connectionist conceptualizations do not qualify as connectionist explanations, we argue for the evolutionary nature of explanations as they relate to theory development, and for the role of conceptualizations as constructive intermediate explanations. Finally, to address Fantuzzi's (1993) criticism that pre-simulation connectionist conceptualizations are unproductive exercises akin to putting the cart before the horse, we provide examples to show that the relationship between

Issues in Applied Linguistics
ISSN 1050-4273

Vol. 5 No. 1 1994 173-194

connectionist conceptualization and connectionist simulation is essentially an interactive one, and we conclude by advocating the more tenable simulation-*and*-theory approach over an unnecessarily restrictive simulation-*then*-theory paradigm.

ON THE NATURE OF CONNECTIONIST EXPLANATIONS

It was argued in Fantuzzi (1992, 1993) that connectionist explanations are too vague to qualify as theories of cognitive functions. To illustrate with an example from our previous discussions, Fantuzzi (1992), citing McCloskey (1991), states:

> While Seidenberg & McClelland (1989) have provided an explicit computer simulation of a cognitive behavior, McCloskey argues that the underlying theory of human cognition remains vague: just general statements to the effect that representations are distributed and similar words are represented similarly. (Fantuzzi, 1992, p. 328)

According to McCloskey, among the questions that Seidenberg and McClelland's model of word recognition and naming would have to address are: What regularities and idiosyncrasies does the network encode in response to the pool of words it encounters? How does the network represent the acquired knowledge over a set of connection weights? And how is the appropriate knowledge brought into play in the appropriate context? For example, how does the network determine the appropriate phonological instantiation of *a* in the case of regular words like *hate,* exception words like *have,* and nonwords like *mave?* (McCloskey, 1991, p. 390).

In response to McCloskey's questions, Seidenberg (1993) points out that complex data sets like the ones connectionist models generate can be analyzed in many different ways in response to different research questions. In the case of the Seidenberg and McClelland model, for instance, the theoretical claims were about "the form in which knowledge is represented, not about the ways in which individual letters or sounds are encoded" (Seidenberg, 1993, p. 233). Theoretical issues, then, determined the kinds of analyses that were reported. Seidenberg

also points out that some of the questions McCloskey raises are already answerable, among them the question of the pronunciation of specific letters or letter patterns. What we need to do, according to Seidenberg, is to "observe the patterns of activations over the hidden units that occur for different spelling-sound correspondences" (p. 233). This is because the hidden units play a central role in helping the network develop appropriate internal representations capable of supporting the necessary (and often complex) input-output functions, which in this case involves grapheme-phoneme correspondences.

For those less familiar with dynamical systems[1] and dynamical explanations, the notion of "patterns of activations over hidden units" would, as Fantuzzi points out, mean nothing more than "general statements to the effect that representations are distributed and similar words are represented similarly"--connectionist jargon that appears to describe rather than explain complex cognitive functions. Indeed, Fantuzzi (1992) is not alone when she points out that "the dynamics of complex nonlinear connectionist systems are difficult to analyze, and thus understand" (p. 328; see also McCloskey, 1991; Pavel, 1990; Rager, 1990).

Notions such as "patterns of activations over hidden units" do, however, provide deeper insights into cognitive functions. In what follows, we will illustrate this point with the help of a specific connectionist model, namely Elman's (1990, 1993) simple recurrent network on sentence prediction. We first describe the model and the cognitive behavior it simulates, then proceed to analyze its hidden units to discover the nature of its internal representation.

Like most other connectionist networks, Elman's network consists of a layer of input units which is connected to a layer of output units via an intermediate layer of "hidden" units (so called because it represents the network's internal representation). In addition to this basic connectionist structure, Elman's network also has a recurrent context layer that gives it a dynamic memory (see Figure 1). Activations from the hidden (or internal representation) layer are thus propagated, not only to the output layer, but in this case also to the context layer, which then feeds the activations back to the hidden layer to influence subsequent activations within the

network's internal representation. In so doing, the recurrent context layer enables the network to encode prior context, or temporal information, and thus to "live in time" (Plunkett, 1993, p. 55).[2]

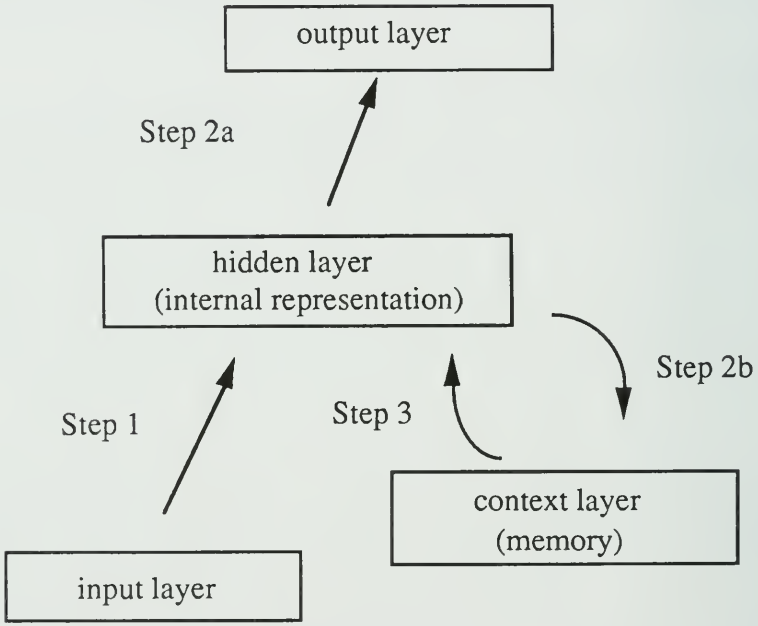

Figure 1. A recursive network that provides memory feedback via the context layer. Step 1 shows the input activating the hidden units to form an internal representation. Step 2a shows the hidden units generating an output. Step 2b shows the hidden units simultaneously updating the context layer with information about the internal representation. Step 3 shows the context layer supplying the hidden units with information about the network's prior internal representations. (Adapted from Elman, 1990, 1991, used with permission.)

Trained on a succession of words, presented one at a time, from an unbroken string of grammatically well-formed sentences, the network succeeded in abstracting several important properties

underlying human language. In the first simulation (Elman, 1990), the network learned to make lexical category distinctions between nouns and verbs on the basis of distributional evidence alone. In addition, it learned to make type-token distinctions, as well as grammatical role distinctions (such as subject versus object). In the second simulation (Elman, 1993), the network further learned to assign proper subject-verb agreement, even in the case of complex sentences that involved long-distance dependencies. Clearly, the network was encoding some fundamental linguistic properties. The question is *how*.

As emphasized earlier, notions such as "patterns of activations over hidden units" can be analyzed to provide important insights into a network's internal representation. Such analyses frequently make use of statistical techniques such as hierarchical clustering analysis and principal component analysis.[3] Elman (1990) employed the former, while Elman (1993), the latter.

Hierarchical clustering analysis is a technique that allows us to identify the similarity structure of the hidden unit activations. In effect, the technique permits us to use spatial organization at the hidden unit level to observe the categorization distinctions encoded within the network's internal representation (Elman, 1992). Principal component analysis, on the other hand, allows us to observe how hidden unit activations change over time. It does this by reducing the network's multi-dimensional internal representation to a set of more manageable 2 or 3-dimensional phase portraits, or graphs.

When Elman (1990) subjected the hidden unit activations in his network to a hierarchical clustering analysis, he obtained the tree shown in Figure 2. The tree reveals a *highly structured* internal representation. To begin with, *nouns* are clustered separately from *verbs,* indicating that the network has captured an important grammatical/syntactic distinction, namely that of lexical category. Moreover, within the noun cluster, *animates* form a separate category from *inanimates*. And further down the hierarchy, among the animates, *humans* are distinct from *non-humans,* while among the inanimates, distinctions are made between *breakables, edibles,* etc. These are semantic distinctions. What the cluster analysis reveals, then, is a hierarchy in which semantic clusters are nested within syntactic clusters (Elman, 1992; van Gelder, 1992). The analysis also reveals that verbs are further

differentiated into *those that require objects, those that optionalize objects* and *those that preclude objects*. These are differentiations involving verb argument structure, a linguistic property with strong semantic overtones, yet at the same time inextricably involved with grammatical/syntactic relations. What we see, then, is an internal representation in which the semantic and syntactic domains are intimately interlocked, such that both semantic and syntactic features are simultaneously instantiated. Here we get a glimpse into an internal representation in which both semantic and syntactic features intertwine to affect the course of language processing.[4] (See Figure 2).

What might it mean to have an internal representation in which the semantic and syntactic domains are intricately interwoven with each other? Before answering this question, let us first consider the results of a second, more elaborate cluster analysis on the same set of hidden unit activation patterns.[5] The analysis reveals the emergence of grammatical role distinctions such as subject versus object within the noun clusters. For example, within the BOY-cluster, *BOY-sleep* and *BOY-move-rock* occur closer to each other than to *woman-like-BOY* or *man-chase-BOY*. Indeed, barring a few exceptions, tokens of BOY-in-subject-role tend to occur closer to each other than to tokens of BOY-in-object-role. A similar distinction holds within other noun clusters. In fact, as Elman (1990) observes, "The differentiation is nonrandom" (p. 207).

The picture we get is one in which grammatical/syntactic distinctions are arranged in subset-superset relations. First, we have a lexical category distinction between nouns and verbs. Then, within the NOUN-cluster, we have a grammatical role distinction between subject-nouns and object-nouns. At the same time, we see a picture of grammatical role clusters nested within semantically-defined noun clusters (e.g., [+animate], [+human], [+masculine], etc.), which in turn are nested within lexical category clusters. The network does not instantiate a syntactic distinction devoid of semantic interpretation, nor vice-versa.

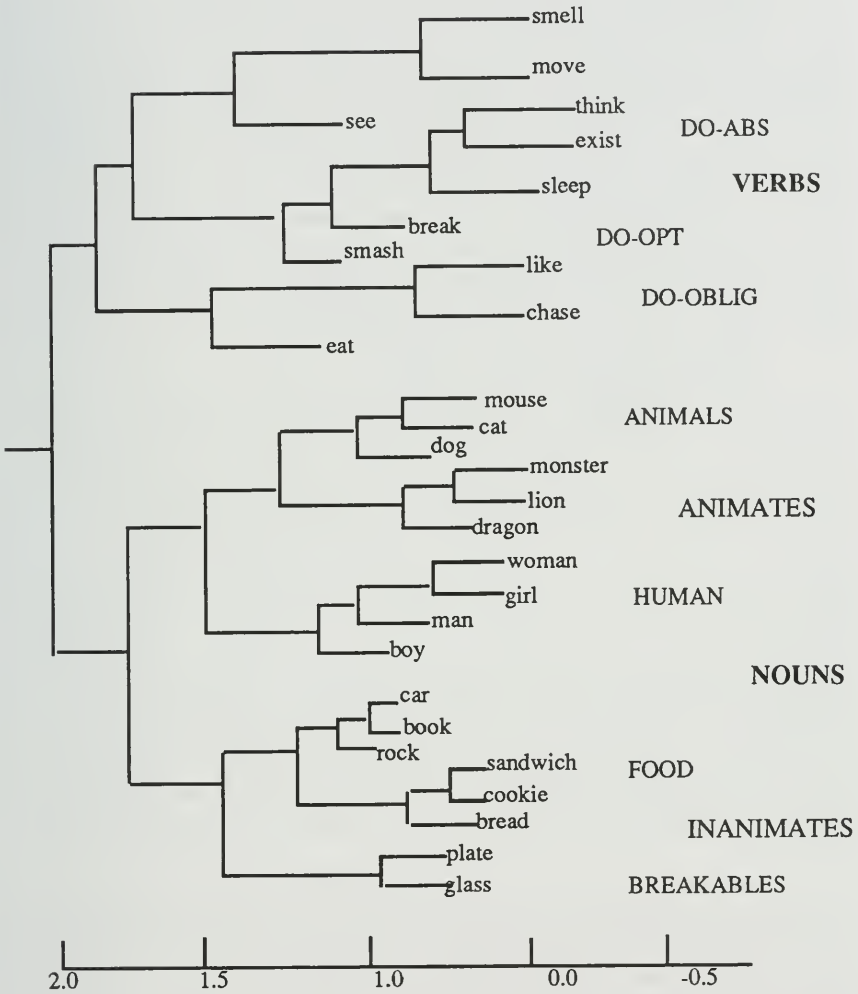

Figure 2. A hierarchical cluster diagram of the hidden unit activation vectors in a simple sentence prediction task. Labels indicate the inputs which produced the hidden unit vectors. Inputs were presented in context, and the hidden unit vectors averaged across multiple contexts. (Elman 1990, used with permission)

The second cluster analysis also reveals how the network preserves type-token relations while at the same time allowing each token to be instantiated in a context-sensitive manner. For example, the analysis shows that all BOY-tokens are closer to each other than they are to GIRL-tokens. This is because tokens of the same type have hidden unit activations that are very similar--certainly more similar to each other than to the activations of tokens of a different type. At the same time, no two tokens of the same type are exactly identical. *BOY-sleep,* for example, has a different internal representation from *BOY-move-rock* or *woman-like-BOY*. This is because each hidden unit activation pattern reflects its own unique prior context.

The above analysis is important because it shows that sensitivity to context "does not preclude the ability to capture generalizations" (Elman, 1991, p. 220). According to Elman, all the network needs to do is learn "to respond to contexts which are more abstractly defined" (p. 220). This is a simple task for the network. Given sufficient similarity in the hidden unit activation pattern, the network can easily abstract and generalize. At the same time, whatever is dissimilar automatically contributes to individual token identity. This use of context to establish generalizations about classes of items and at the same time to identify individual items is also significant from a processing perspective because it shows that types and tokens can be identified simultaneously, and without recourse to additional procedures such as indexing or binding operations which often feature prominently in traditional symbolic modeling (Elman, 1990).

The context-sensitivity of each token, in effect, means that its hidden unit activation pattern is subject to subtle adjustments as the token combines with other tokens (Elman, 1992). How the hidden units "accommodate" themselves as a token interacts with other tokens can be captured to some extent by phase-portraits, or graphs, obtained through principal component analysis. The graph in Figure 3 illustrates how hidden unit activation patterns are constrained by prior context. The graph shows the trajectories through state space for the sentences *boy* ***who boys chase*** *chases boy* and *boys* ***who boys chase*** *chase boy*. Note that the embedded clause for both sentences is the same. Nevertheless, the path taken by the vector representing the hidden unit activations for this

particular embedded clause varies slightly depending on the preceding context. In the first sentence, the prior context was a singular subject *boy*, while in the second sentence, the prior context was a plural subject *boys*. Each context produces its own expectations, and these expectations affect the pattern of hidden unit activations of the tokens that follow.

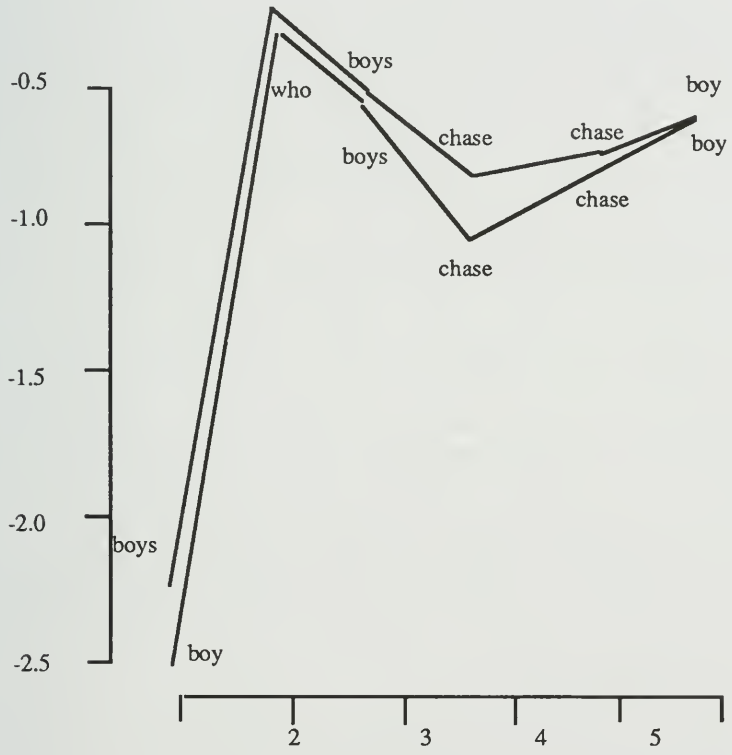

Figure 3. Plot of the movement through one dimension of the hidden unit activation space (the second principal component) as the successfully trained network processes the sentences *boy who boys chase chases boy* vs. *boys who boys chase chase boy*. The second principal component encodes the singular/plural distinction in the main clause subject. (Elman, 1993, used with permission).

Without going into further examples, due to space constraints, let us integrate the insights we have gained about the internal representation in Elman's network. This will help us address some of the questions raised earlier: What kinds of linguistic knowledge does the network encode? How is this knowledge represented in a distributed manner? And how is the appropriate knowledge brought into play in the appropriate context? From hierarchical clustering analyses, we know that the network encodes *constituent structure*. The constituents, however, do not have the discrete characteristics commonly associated with traditional symbolic representations. For example, instead of occupying a discrete location in memory, the constituents reflect their characteristics through a shared network of hidden units. Constituents are similar as well as dissimilar to other constituents in different ways, and this is reflected in the similarity structure of their hidden unit activation patterns. On the basis of their similarity and dissimilarity to other hidden unit activation patterns, their identity as individual tokens and their identity as members of different classes are recognized. Moreover, because hidden units are shared along many dimensions, a constituent often belongs to more than one type of category. For example, *boy* belongs to lexical category NOUN, semantic categories [+ANIMATE], [+HUMAN], [+MALE], etc., grammatical role category SUBJECT or OBJECT depending on its context, number category SINGULAR, which prompts verb agreement expectations in the event that *boy* is identified as a SUBJECT, and the list could go on. Information about the various category memberships is activated, not category by category, but all at the same time, and in conjunction with information about individual (or context-sensitive) token identity.

Nor are semantic and syntactic considerations delegated to autonomous domains. Because semantic categories are found nested within syntactic categories, and vice-versa, constituents are also bound to instantiate semantic and syntactic properties at one and the same time. This is possible because the internal representation of each constituent is a pattern of hidden unit activations, and not a single discrete hidden unit. In fact, by sharing hidden units, constituents are further able to exhibit finer-grained distinctions among category members when appropriate, thereby overriding the brittleness of all-or-none category

memberships. The question of when it is appropriate to make gradient distinctions and when it is appropriate to make all-or-none categorizations will, of course, be determined by the demands of the task--in other words, by the context.

Another important characteristic of the network's internal representation is that it is non-static. Whereas traditional symbolic representations tend to be "timeless" (Port & van Gelder, 1991), in the sense that they remain unchanged once learning or retrieval or production rules have acted upon them, network representations are dynamic, keenly sensitive to context, and thus at different points in time never quite the same, although similar enough for most categorization purposes. More importantly, as the trajectories derived from principal component analysis reveal, each pattern of hidden unit activations (with prior context encoded) constrains the next pattern of hidden unit activations. In other words, hidden unit activation patterns have "causal properties" or, as Elman (1991) puts it, "they are cues which guide the network through different grammatical states" (p. 221). In this sense, each constituent is dynamic and plays an active role in determining an entire composite representation (Port & van Gelder, 1991), be that composite representation a single sentence, a whole paragraph, or an entire discourse. In this sense, too, we see the emergence of grammatical relations--not in static, timeless fashion, but malleable to context and functioning in real time.

Clearly, the kind of internal representation found in Elman's network is very different from the kind conceived within the traditional symbolic paradigm. At the very least, the difference is significant enough to challenge us to give more serious thought to the temporal and integrative aspects of language representation. The arguments necessary to advance this new focus in language research have come from some innovative analyses into the hidden units of dynamic networks. Although some have argued that the explanations provided by these analyses are not explicit in the same way that symbolic explanations are explicit, it is important to bear in mind that network explanations are not so vague that they cannot broaden our conception of what counts as a viable scheme of linguistic representation within the human brain. With further research, some of these explanations may become more explicit, but it is also worth remembering that the so-called "vague explanations" mentioned earlier--those referred to as "general

statements to the effect that representations are distributed and similar words are represented similarly"--are the kinds of basic principles that keep re-surfacing across different domains. If, as proposed in Seidenberg (1993), one of the desiderata for theories of cognition is that they be able to show "how phenomena previously thought to be unrelated actually derive from a common underlying source" (p. 233), then connectionist explanations are way ahead in fostering the development of effective explanatory theories.

ON THE NATURE OF CONNECTIONIST CONCEPTUALIZATIONS

It is unfortunate that the value of conceptualization as explanation and as a precursor of theories has not always been fully appreciated. This undervaluation was highlighted in our previous exchange when Fantuzzi (1993) raised the question of whether conceptualization counts as explanation in SLA. Fantuzzi's own position is that "general conceptualizations are *not* explanations of linguistic behavior" (p. 309). Such a position, in our view, is much too restrictive. For while conceptualizations tend to focus on more abstract problems, or problems that are still fuzzily defined, which inevitably adds a general or notional flavor to them, to the extent that these conceptualizations help to clarify a problem and contribute to its solution, they nonetheless qualify as explanations.[6]

The tendency to undervaluate conceptualization crops up from time to time, and perhaps more frequently in the case of some paradigms than in others, but for a field as new as connectionism, unfamiliarity with the nature of connectionist conceptualizations may have been an additional factor. In this section, we will draw on an extended example found in van Gelder (1992) to illustrate how connectionist conceptualizations contribute to the development of explanatory theories.

As discussed earlier, Elman's work has shown how a network capitalizes on the similarity structure of hidden unit activations to encode constituent structure. The structure, as revealed by cluster analysis, is hierarchically nested, such that both syntactic and semantic considerations constrain the class of possible successors in a sentence. A logical question that follows

is how other considerations such as phonological and discourse-pragmatic factors constrain the composition of an utterance. To answer this question, we would need to look at how these other factors are encoded in the network's internal representation. We would also need to observe how these factors interact and integrate with other factors to produce complex cognitive behavior. One way to proceed is to build a phonologically-sensitive network (perhaps somewhat like NETtalk) and train it on a discourse-based training corpus. Obviously, such an ambitious project cannot proceed atheoretically: We would need to narrow our search space with a clearer definition of what we expect the network to do, and if possible to anticipate the answers the network will reveal. As in other paradigms, the anticipations serve as hypotheses to be confirmed, rejected, or modified. These anticipations are conceptualizations, and are important to the process of theory construction.

Conceptualizations are not conceived in a vacuum. Like our language representations, they are constrained by prior knowledge[7]. Van Gelder (1992), for example, recalls the nested hierarchy used to encode semantic and syntactic distinctions in Elman's network, and postulates that "a host of further subtle contextual and pragmatic factors" are likewise encoded in "an extraordinarily intricate hierarchical structure of regions within regions," only this time the "cascade of regions with differing functional significances" would have to be much deeper, perhaps infinitely so (pp. 183-84). But how, one might ask, do we envisage a hierarchically nested structure that is infinitely deep? Are there similar constructs in other fields that can serve as metaphors to help us visualize something that we can only begin to vaguely imagine?

In an earlier work, Pollack (1989) had identified one such construct (the classical Cantor Set) in the domain of fractal mathematics. As van Gelder (1992) explains, this set consists of points obtained in the following manner:

> Begin with the set of all points on the unit interval (the segment of the real number line between 0 and 1) and delete the middle third open subinterval--that is, every point between 1/3 and 2/3, although being careful not to delete these two points themselves. We are now left with two intervals, each 1/3 the length of the original. Now delete the middle third

> open subinterval of the remaining intervals, and so on ad infinitum. What remains is usually termed a 'dust'--an infinite set of points, systematically arranged in an infinitely deep hierarchy of clusters, such that between any two clusters at the same level there is a gap as big as those clusters. Thus, paradoxically, in this dust *the strict separation between clusters (and, eventually, points) is maintained at any level one cares to examine,* even though, given any distance, no matter how small, one can find an unbounded number of clusters that are less than that distance apart. (pp. 184-85, emphasis added)

The spatial organization in the Cantor Set reminds us of the hierarchical cluster found in Elman's network analysis. Without forgetting that a network's internal representations are in fact *highly distributed*, "not merely in the obvious sense that they are patterns taking place over a large set of hidden units, but--primarily--because they encode the relevant information about the input in a *superimposed* fashion" (van Gelder, 1992, p. 182, emphasis added), we can now go on to consider another important issue: How might fractal structures like the Cantor Set inform us about the *possible behavior* of deep nested hierarchies which encode not only semantic and syntactic constraints but other subtle contextual and discourse-pragmatic ones as well?

As discussed in van Gelder (1992), an important property of the Cantor Set is that clusters at the same level are divided by wide spaces, and these spaces give each cluster its distinctive identity. This property of distinctiveness is maintained even at the level of individual fractal points, no matter how infinitesimally close they might be to one another. In connectionist terms, this would mean that hidden unit activation patterns that are almost congruent can still exhibit subtle but distinctive properties. As van Gelder puts it, within a dynamical system like Elman's network, "[t]wo points in neighboring clusters would have much in common, but they would also have differences of functional significance that could eventually be very important for the future direction of processing in that system" (p. 192). For networks whose nested hierarchies must be *very deep* in order to encode numerous other constraints besides semantic and syntactic ones, the distance between neighboring points must be infinitesimally small, and this

would mean that even very slight variations in hidden unit activation patterns could make a subtle but significant difference in the processing output.

Such fine sensitivity to subtle changes would make the behavior of these networks appear chaotic at times (van Gelder, 1992), but one needs to remember that this chaos-like behavior is in fact entrenched in a type of internal representation that at the same time inherently honors type-token distinctions. Thus it appears that these networks, with their deep Cantor Set-like nested hierarchies, are well-suited to the paradoxical task of "capturing the order and regularity inherent in linguistic systems" and at the same time responding appropriately (and this could mean drastically) "to small changes in word order, intonation, or pragmatic context" (van Gelder, 1992, p. 193).

What van Gelder's conceptualization has done, then, is to provide us with a clearer idea of what an internal representation that has to encode numerous linguistic differentiations might look like and how such a representation might be expected to behave. In so doing, his conceptualization helps to more clearly define a possible line of simulation research that moves us closer toward an explanatory theory of language representation. Conceptualizations like van Gelder's are made possible by findings from previous simulation efforts, but are also enriched by attempts to seek out compatibilities with other fields. In fact, van Gelder points out that a possible consequence of drawing metaphors from a different domain like fractal mathematics is that we might discover the same set of principles (in this case, mathematically expressed) underlying not only cognitive functions but numerous other natural phenomena as well. It is in this more general but also more unifying sense, then, that conceptualizations often contribute to the development of effective explanatory theories.[8]

SIMULATION-THEN-THEORY, OR SIMULATION-AND-THEORY

Attempts are sometimes made to characterize a paradigm according to its research style. Thus, we come across observations about a paradigm being either theory-then-research oriented, or

vice-versa. Taken to the extreme, however, such characterizations can be counter-productive. One such example is found in the last exchange, when Fantuzzi (1993), observing that connectionist explanations are essentially "built bottom-up from a working model" (p. 296), argues that Shirai and Yap (1993), in attempting to come up with a general connectionist framework to account for language transfer phenomena prior to running actual simulations, "have the relationship backwards" (p. 303). While there is little doubt that general conceptualizations stand to benefit from the findings of specific simulations, it would be a mistake to assume that connectionist theorizing *cannot* proceed ahead of specific simulations.

As pointed out earlier, conceptualizations are rarely (if ever) conceived in a vacuum. For example, Shirai (1992) examined evidence from language transfer studies and discussed basic principles extracted from numerous connectionist simulation efforts. The purpose of general conceptualizations of this kind was to define a possible line of research that can help these conceptualizations gain greater specificity.

That connectionist research is not strictly bound to a simulation-then-theory approach is evident from the following illustration. In van Gelder (1992), we trace an interesting historical development: Armed with insights from exploratory forays into neuroscience and dynamical computational modeling, Paul Churchland (1986) postulated that "the brain represents various aspects of reality as a *position* in a suitable *state space"* (van Gelder, 1992, p. 179). As van Gelder notes, at the time, Churchland's statement was seen as an exciting but hopelessly bold speculation, yet in a matter of just a few years an explosion of connectionist simulations such as Elman's work had begun to rapidly clear some of the mysteries surrounding Churchland's postulate. And, as we have seen earlier in this paper, Elman's simulation efforts were followed by van Gelder's own bold conceptualization, which although is bound to invite criticism yet at the same time is sure to find its way into the theoretical underpinnings (whether explicit or implicit) of some researcher's simulation effort. What we see, then, is a principle of interaction and integration at work, not just at the level of hidden units within

a network's internal representation, but also at the level of conceptualization and simulation (i.e., in the way we as researchers go about formulating and formalizing our hypotheses and theories).

CONCLUSION

Thus far we have shown that connectionist explanations are not so vague that they cannot contribute to our understanding of how language is represented and processed. Nevertheless, since connectionism as a field is young, many questions remain to be explored. For instance, we have yet to address the question of how symbolic representations and processing might emerge from a subsymbolic (i.e., connectionist) architecture. As yet we still do not know if a hybrid symbolic-connectionist system holds the answer, or if the solution merely lies within a more sophisticated network. Nevertheless, the search for an answer is on, and philosophers and researchers like Clark and Karmiloff-Smith (1993) are proposing some interesting answers. Drawing on language acquisition and child development studies, and the results of relevant connectionist simulations, Clark and Karmiloff-Smith sketch a scenario in which a connectionist network makes use of cluster analysis to induce constituent categories (as discussed earlier), then submit these categories through another kind of statistical procedure known as "skeletonization" (cf. Mozer & Smolensky, 1989) to yield a more abstract version of each category. Through a series of skeletonization procedures, the network will eventually obtain explicit (i.e., symbolic) representations that are manipulable for novel combinations and transportable to different domains. In this way, Clark and Karmiloff-Smith argue, the connectionist network retains its sensitivity to contextual nuances by exploiting its implicit (i.e., subsymbolic, or distributed) representations, and at the same time acquires the flexibility to manipulate more explicit representations obtained through the process of skeletonization.

As with many other connectionist-based accounts, the sketch proposed by Clark and Karmiloff-Smith requires an understanding of how connectionist networks behave, but it also requires something more. It requires powers of perception,

reasoning and imagination for a researcher to integrate what (s)he knows about connectionist principles, cognitive principles, language acquisition principles, neurobiological principles, etc. The path to more plausible and more effective explanatory theories of cognitive functions requires that we recognize the value of broad conceptualizations in addition to explicit explanations. And it also requires that we conceptualize, and theorize, not only at the end of each simulation, but all the time.

ACKNOWLEDGMENTS

We are grateful for comments on an earlier version of this paper by Kevin Gregg and the *IAL* reviewer, neither of whom are responsible for any errors or inconsistencies that may remain. The writing of this paper is partially supported by a Grant-Aid for Scientific Research from Japanese Ministry of Education (no. 06851070) to Yasuhiro Shirai.

NOTES

[1] A system is said to be dynamical when its behavior evolves over time. An excellent example is the human brain with its non-static flow of neural excitations and inhibitions. Another good example is the connectionist network with its ever-changing pattern of hidden unit activations. Each dynamical system can be studied as a closed system in which their state at any given time can be captured in terms of values of a set of parameters. The values of these parameters change in interdependent ways as the system evolves and often it is possible to capture the changing behavior of the system by means of equations. As van Gelder (1991) points out, these equations can be used to address some important research questions, for example, how the system got to be in the state it is in, and what states it will move on into next. Van Gelder goes on to point out, however, that "dynamical explanations may proceed without making explicit use of the equations governing the system" (p. 500), as often happens when researchers do not have all the equations for a complex system, or when researchers do have the equations but prefer to use some more perspicuous way of explaining how the system works. Van Gelder cites the use of state trajectories in connectionist research (discussed a little later in the main text) as an example when explanations of particular features of a network's behavior can proceed without adverting to the full equations which formally govern the behavior of the network.

[2] It is worth noting that previous connectionist simulations often did not incorporate the temporal aspect of cognitive functions. By adding a temporal dimension to his simulation, Elman (1990, 1993) was able not only to explore the spatial

organization of network representation; he was able to analyze its dynamical properties as well.

[3] Elman (1991) mentions several other techniques for network analysis, among them weight matrix decomposition (McMillan & Smolensky, 1988), skeletonization (Mozer & Smolensky, 1989) and contribution analysis (Sanger, 1989).

[4] The intimate interaction between syntactic and semantic features sometimes makes it difficult to identify which particular features make the biggest difference in processing outcome. For example, Fantuzzi (1993), citing Kim, Pinker, Prince and Prasada (1991), argues for the role of syntax in determining whether a novel verb tends to be regularized for past tense, while Harris (1992, 1993) argues for the greater role of semantics.

Kim et al.'s work is on past tense naturalness ratings. They attributed the difference in naturalness ratings of past-tense forms to each verb's derivational status: a novel verb derived from a noun is more natural with regular past, while that derived from a verb is more natural with irregular past. They thus attribute the difference in naturalness ratings to syntax. Harris (1992, 1993), however, has proposed an alternative account, claiming that naturalness with irregular past is dependent on how much of a word's original meaning is preserved in the derived verb. Since most of the denominal verbs used by Kim et al. were semantically very distant from the original noun, it is only natural that they did not show high ratings for irregular past. In fact, when controlled for semantic similarity, denominal verbs and deverbal verbs did not show a statistically significant difference (Harris, 1993). This also shows that syntax and semantics are closely intertwined; the phenomenon attributed primarily to grammatical status (noun vs. verb) turned out to be more dependent on semantic factors (see also Stemberger (1993) for the strong effect of phonological factors in determining the regularizability of a verb).

It should also be pointed out that Marcus, Brinkman, Clahsen, Wiese, Woest and Pinker's (1993) argument (also cited by Fantuzzi) that connectionism cannot handle minority default plurals such as those seen in German and Arabic is also countered by Plunkett (1994), who argues that the minority default is problematic only for single-layered networks such as used in Rumelhart and McClelland (1986), and not problematic for multi-layered networks that have intermediate structures, and in fact his network successfully learned Arabic plurals.

[5] Whereas the first cluster analysis *averaged* all instantiations of the hidden unit patterns for each lexical item into a single vector, the second cluster analysis charted *every instantiation* of these patterns. Thus, whereas the first analysis involved just 29 vectors (one per lexical item), the second analysis ended up with 27, 354 vectors (one for every lexical item in a different context). The overall structure of the trees from both analyses are the same, but finer distinctions are visible in the second, more elaborate tree (Elman, 1992).

[6] It should be emphasized here that we are not advocating vague explanations nor vague theories. Contrary to Fantuzzi's claim, Shirai and Yap (1993) did not maintain that vagueness is all that can be expected from connectionist models. The point we made was that many linguistic phenomena are "beyond precise description by categorical rules" (p. 122), and thus any attempt to capture these phenomena in *categorical* terms would at best result in something "vague" (in the sense of "approximate" or "imprecise" rather than in the sense of "unclear"). In fact, we underscored that a partiality for categorical descriptions is what limits the effectiveness of classical/symbolic explanations, while connectionist explanations do a better job with their soft constraints (or "soft rules").

[7] Note that the terms prior context and prior knowledge are easily interchangeable. Although not discussed in this paper, prior context/knowledge need not always take the form of temporal information that finds its way into the internal representation from the external environment. It is conceivable that complex systems like the human brain may consist of several networks interacting together, in which case prior context/knowledge for a particular network can take the form of information previously stored within a neighboring network.

[8] It is in this same spirit that Shirai (1992) attempts to articulate a general framework for language transfer phenomena from the perspective of a different (in this case, connectionist) domain.

REFERENCES

Churchland, P. (1986). Some reductive strategies in cognitive neurobiology. *Mind, 95*, 279-309.

Clark, A., & Karmiloff-Smith, A. (1993). The cognizer's innards: A psychological and philosophical perspective on the development of thought. *Mind and Language, 8*, 487-519.

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

Elman, J. L. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning, 7*, 195-225.

Elman, J. L. (1992). Grammatical structure and distributed representations. In S. Davis (Ed.), *Connectionism: Theory and practice* (pp. 138-178). New York: Oxford University Press.

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

Fantuzzi, C. (1992). Connectionism: Explanation or implementation? *Issues in Applied Linguistics, 3*, 319-340.

Fantuzzi, C. (1993). Does conceptualization equal explanation in SLA? *Issues in Applied Linguistics, 4*, 295-314.

Harris, C. L. (1992). Understanding English past-tense formation: The shared meaning hypothesis. *Proceedings of the 14th Annual Meeting of the Cognitive Science Society* (pp. 100-105). Hillsdale, NJ: Lawrence Erlbaum.

Harris, C. L. (1993). Using old words in new ways: The effect of argument structure, form class and affixation. *CLS 29: (Vol. 2) Papers from parasession on the correspondence of conceptual, semantic and grammatical representations* (pp. 139-153). Chicago: Chicago Linguistic Society.

Kim, J. J., Pinker, S., Prince, A., & Prasada, S. (1991). Why no mere mortal has ever flown out to center field. *Cognitive Science, 15*, 173-218.

Marcus, G. F., Brinkman, U., Clahsen, H., Wiese, R., Woest, A., & Pinker, S. (1993). *German inflection: The exception that proves the rule* (Occasional Paper No. 47). Cambridge, MA: The Center for Cognitive Science, MIT.

McCloskey, M. (1991). Networks and theories: The place of connectionism in cognitive science. *Psychological Science, 2,* 387-394.

McMillan, C., & Smolensky, P. (1988). *Analyzing a connectionist model as a system of soft rules* (Technical Report No. CU-CS-303-88). University of Colorado, Boulder, Department of Computer Science.

Mozer, M. C., & Smolensky, P. (1989). *Skeletonization: A technique or trimming the fat from a network via relevance assessment* (Technical Report No. CU-CS-421-89). University of Colorado, Boulder, Department of Computer Science.

Pavel, M. (1990). Learning from learned network. *Behavioral and Brain Sciences, 13,* 503-504.

Plunkett, K. (1993). *Connectionist models summer school (Lecture Note).* Department of Experimental Psychology: Oxford University.

Plunkett, K. (1994, April). *Learning the Arabic plural: The case for minority default mappings in connectionist nets.* Paper presented at the Workshop on Cognitive Models of Language Acquisition, Tilburg University, Netherland.

Pollack, J. B. (1989). Towards a fractal basis for artificial intelligence. *Advances in Neural Information Processing Systems: Proceedings of the NIPS Conference.* Cited in van Gelder (1992).

Port, R. F., & van Gelder, T. (1991). Representing aspects of language. *Proceedings of the 13th Annual Meeting of Cognitive Science Society* (pp. 487-492). Hillsdale, NJ: Lawrence Erlbaum.

Rager, J. E. (1990). The analysis of learning needs to be deeper. *Behavioral and Brain Sciences, 13,* 505-506.

Rumelhart, D. E., & McClelland, J. L. (1986). On learning the past tense of English verbs. In J. L. McClelland, D. E. Rumelhart, & T. P. R. Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition. Volume 2: Psychological and biological models* (pp. 216-271). Cambridge, MA: MIT Press.

Sanger, D. (1989). *Contribution analysis: A technique for assigning responsibilities to hidden units in connectionist networks* (Technical Report No. CU-CS-435-89). University of Colorado, Boulder, Department of Computer Science.

Seidenberg, M. S. (1993). Connectionist models and cognitive theory. *Psychological Science, 4,* 228-235.

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review, 96,* 523-568.

Shirai, Y. (1992). Conditions on transfer: A connectionist approach. *Issues in Applied Linguistics, 3,* 91-120.

Shirai, Y., & Yap, F.-H. (1993). In defense of connectionism. *Issues in Applied Linguistics, 4,* 119-133.

Stemberger, J. P. (1993). Vowel dominance in overregularizations. *Journal of Child Language, 20,* 503-521.

van Gelder, T. (1991). Connectionism and dynamical explanation. *Proceedings of the 13th Annual Meeting of Cognitive Science Society* (pp. 499-503). Hillsdale, NJ: Lawrence Erlbaum.

van Gelder, T. (1992). Making conceptual space. In S. Davis (Ed.), *Connectionism: Theory and practice* (pp. 179-194) . New York: Oxford University Press.

Foong-ha Yap holds MA degrees in English and TESL, from Loma Linda University and UCLA respectively. Currently she teaches English at Kanto Gakuin University in Kanagawa, Japan. Her research interests include cognitive models of language learning, and neurobiological perspectives on language acquisition.

Yasuhiro Shirai has a Ph.D. in applied linguistics from UCLA, and is now an associate professor in the Department of English at Daito Bunka University in Tokyo, Japan, where he teaches linguistics, first language acquisition, and EFL. His research interests include crosslinguistic acquisition of tense-aspect morphology and cognitive models of L1/L2 acquisition.