Constituent similarity and systematicity: The limits of first-order connectionism

Steven Phillips

Information Science Division

Electrotechnical Laboratory

1-1-4 Umezono, Tsukuba, 305-8568, Japan

stevep@etl.go.jp

Running head: Constituent similarity and systematicity

Author's address: Information Science Division, Electrotechnical Laboratory, 1-1-4 Umezono, Tsukuba, Ibaraki 305-8568 Japan. Email: stevep@etl.go.jp. Tel: +81 298 543315. Fax: +81 298 545857.

ABSTRACT

Standard feedforward and recurrent networks cannot support strong systematicity when constituents are presented as local input/output vectors (Phillips, 1998). To explain systematicity connectionists must either: (1) develop alternative models; or (2) justify the assumption of similar (non-local) constituent representations prior to the learning task. I show that the second commonly presumed option cannot account for systematicity, in general. This option, termed *first-order connectionism*, relies upon established spatial relationships between *common-class* constituents to account for systematic generalization: Inferences (functions) learned over, e.g., cats extend systematically to dogs by virtue of both being nouns with similar internal representations so that the function learned to make inferences employing one simultaneously has the capacity to make inferences employing the other. But, humans generalize beyond common-class constituents. Cross-category generalization (e.g., inferences that require treating mango as a colour, rather than a fruit) makes having had the necessary common context to learn similar constituent representations highly unlikely. At best, the constituent similarity proposal encodes for one binary relationship between any two constituents, at any one time. It cannot account for inferences, such as *transverse patterning* that require identifying and applying one of many possible binary constituent relationships that is contingent on a third constituent (i.e., ternary relationship). Connectionists are, therefore, left with the first option which amounts to developing models with the symbol-like capacity to explicitly represent constituent relations independent of constituent contents, such as in tensor-related models. However, rather just simply implementing symbol systems, I suggest reconciling connectionist and classical frameworks to overcome their individual limitations.

Key words. Systematicity, constituent, similarity, novelty, relation, connectionism, classicism, transverse patterning

INTRODUCTION

Networks generalize. But does the degree of generalization account for, even in principle, the sort of generalization evident in humans? Generalization is the distribution of functional capacity over system states. Since generalization is a function of specific computational mechanisms, different architectures exhibit characteristically different types of generalization. Therefore, a careful characterization of the distribution of cognitive behaviours in humans can provide insight into the underlying cognitive architecture.

For example, in general, association and relation based architectures differ on how they access constituents within a complex representation. Associative systems are basically *uni-directional*: constituent A permits retrieval of constituent B, but not the reverse; whereas relational systems are *omni-directional*: reverse access is also implied (Phillips, Halford, & Wilson, 1995). In an associative system, addition may be represented by uni-directional links (e.g., $1{+}2 \rightarrow 3$). The sum of two numbers is retrieved by matching the left side of the link to retrieve the right side constituent. But, in this state, subtraction cannot be accomplished without creating extra links (e.g, $3\text{-}2 \rightarrow 1$). In a relational system, both addition and subtraction are accomplished from the same internal state that represents, in this case, the triple $R_{+}(1,2,3)$: $R_{+}(1,2,?) \rightarrow 3$; and $R_{+}(?,2,3) \rightarrow 1$. The associative architecture distributes the capacities for addition and subtraction across two different representational states; whereas these capacities are localized to the same state in the relational architecture. Suppose, for the purpose of this example, it is never the case that subjects with the capacity to perform addition do not have the capacity to perform subtraction. Such an observation would rule out the associative, but not the relational architecture.

This example, though illustrative, captures the essence of the *systematicity* argument Fodor and Pylyshyn (1988) raised against connectionism. They argued that while human cognitive capacity is organized around groups of "related" behaviours (i.e., capability over

any one behaviour is systematicity related to any other behaviour within the same group), connectionist models are indifferent to this organization. Since the basic resources and processes of connectionist networks are characteristically nodes for representing concepts and activation flow along weighted connections for making inferences, a network can be equally configured so as to capture, or not the relevant distribution of cognitive behaviours. Thus, according to Fodor and Pylyshyn, connectionism does not explain the systematicity property of human cognition.

Acceptance of their argument depends on what one takes to be a group of related behaviours; and what one takes to be the basic components of a connectionist cognitive architecture. Clearly, regarding a connectionist architecture as a collection of nodes and links is too fine a characterization, since it affords no aggregation of behaviours over states - each state uniquely realizes each behaviour. When a learning function is considered as part of the architecture, there are numerous examples of networks exhibiting generalization - multiple inference capacities acquired from a single state (weight) change. The issue is whether the sort of generalization exhibited by these networks corresponds to that of humans, which can only be addressed by characterization of generalization in humans.

Hadley (1994a) introduced the notion of *strong systematicity* to characterize generalization in language. A network is said to demonstrate strong systematicity if it processes sentences with constituents (words) in novel syntactic positions and at novel levels of embedding (e.g., *Mary loves John* and *Mary knows Sue loves John*, where *John* only appeared in the agent position in the training set). When processing includes determining a word's semantic role (e.g., agent, patient, etc) it is termed *strong semantic systematicity* (Hadley & Hayward, 1997).[1]

By this characterization, networks demonstrating significant degrees of generalization in

[1]Here, both terms are called *strong systematicity*, since the results presented apply equally.

cognitive domains, for example, the feedforward network (Brousse, 1991), and simple recurrent network (Elman, 1990) did not appear to satisfy strong systematicity. A review of the training procedures in six models suggested that in all likelihood generalization was achieved from a training set containing all constituents in all allowable positions (Hadley, 1994a). While this observation does not constitute a proof, subsequent simulation and analysis showed that feedforward and a variety of recurrent networks (including the simple recurrent network) cannot demonstrate strong systematicity under local input/output vector representations[2] for constituents (Phillips, 1995, 1998). Essentially, the weights responsible for the construction/extraction of constituents in different positions are trained independently, necessitating a training set that includes each constituent in each position. (Closely related to strong systematicity is the *universal generalization* criterion, requiring generalization to constituents not appearing anywhere in the training set. Marcus, 1998a, 1998b, devised this criterion also from a characterization of language learning, and showed it cannot be satisfied by feedforward and simple recurrent networks under local constituent representations.[3])

Although strong systematicity and universality results (above) show feedforward and recurrent networks do not support the same distribution of behaviours as humans, even in principle,[4] these conclusions assume local input/output constituents representations. As explained in Phillips (1998), while some constituents may share similarity by virtue of belonging to common categories (e.g., *John* and *Mary* may share a common internal representation because they are both known proper nouns), it cannot be *just* by virtue of surface similarity (i.e., sensory appearance) that two constituents share similar input/output representations. In particular, constituents belonging to multiple categories can have identical

[2]That is, one component with value 1 and the rest zero (e.g., 1000, 0100, etc).

[3]Universal generalization requires at least as many correct test cases as strong systematicity, but with fewer training examples (e.g., *John* no longer appears in the training set). Therefore, the lack of strong systematicity results in Phillips (1995, 1998) also apply to universal generalization.

[4]The problem is architectural and not one of parameter tuning.

surface features, but their category membership and therefore their purported similarity to other constituents can only be deduced from context. If constituent similarity is *not* sufficient to account for systematicity, then explaining systematicity requires demonstrating "appropriate" generalization using local constituent representations as a control measure.

The connectionist is, therefore, left with two choices in explaining systematicity, either: (1) develop alternative networks capable of demonstrating strong systematicity (or, universal generalization) over local input/output constituent representations; or (2) explain why similar (non-local) constituent representations, sufficient for systematicity, are available to these networks prior to the learning task. The goal of this paper is to seriously evaluate the second option.

Constituent similarity assumption

A number of connectionists have made use of the idea that generalization could be facilitated by prior learning on related tasks. For example, Chalmers (1990) trained a recursive auto-associative memory (RAAM) network (Pollack, 1990) to auto-associate active and passive sentences. The internal (hidden unit) representations of active and passive sentences generated by the trained RAAM were then used as the inputs and target outputs for a feedforward network on a passive-to-active sentence transformation task. Similarly, Chrisman (1991) used internal representations of English and Spanish sentences as the basis for a Spanish-to-English translation task.

Niklasson and van Gelder (1994) argued that this idea of using previously learned internal representations provides for an, in principle, explanation of strong systematicity. Rather than using arbitrary input/output representations a network makes use of internal constituent representations learned from other tasks. Constituents learned in similar contexts would result in similar internal representations. If the relevant constituents are sufficiently similar,

generalization on new tasks can be achieved without requiring training of all constituents in all positions. In this way, a network demonstrates strong systematicity *relative* to a target task. (Niklasson & Boden, 1999, also applied the same idea to claim strong semantic systematicity on an *attribute inheritance* task).

To demonstrate their point, a feedforward network was trained to make transformations of logical formulas (e.g., $P \rightarrow Q \Leftrightarrow \sim P \vee Q$). The input/output representations for each formula were taken from the hidden unit activations of a RAAM trained to encode/decode each formula. Constituents (e.g., P, Q, $\rightarrow$, etc) were presented to the trained RAAM and the hidden unit activation after the last constituent was presented was taken as the encoding for that formula. So, for example, a RAAM encodes the logical expression $P \rightarrow Q$ as: 1. $[P][\rightarrow] \longmapsto [P \rightarrow]$; and 2. $[P \rightarrow][Q] \longmapsto [P \rightarrow Q]$, where $[.][.] \longmapsto [.]$ is a mapping from input and previous hidden vectors to a new hidden vector. The RAAM encodes expression $\sim P \vee Q$ as: 1. $[\sim][P] \longmapsto [\sim P]$; 2. $[\sim P][\vee] \longmapsto [\sim P \vee]$; and 3. $[\sim P \vee][Q] \longmapsto [\sim P \vee Q]$. The feedforward network maps the encoded representation $[P \rightarrow Q]$ to the encoded representation of its transformation $[\sim P \vee Q]$. The RAAM then decodes the transformation into its constituents. The feedforward network was tested on formulas containing a novel constituent S, for which it produced the correct transformation. Thus, Niklasson and van Gelder claimed a demonstration of strong systematicity.[5]

The claim for strong systematicity is, however, contentious. While the simulation results are not contested, what is problematic is the cognitive plausibility of their training procedure. Hadley (1994b) raised several problems, including (among others): the use of syntactic markers to denote the role of all trained and novel constituents; exhaustive training on all other constituents, including all of their combinations; and the fortuitous encoding

[5]Since S did not appear anywhere in either transformation or encoding/decoding training sets, it is also a claim for satisfying Marcus's universal generalization criterion.

of S, permitting its internal representation to lie between known constituents P and Q.[6] The extensive use of such additional information casts doubts on whether their "in principle" demonstration of strong systematicity could ever accommodate a cognitively plausible explanation, leading Hadley to conclude that it was a *borderline* case.

Borderline cases are, by definition, indecisive. On one hand, they suggest further refinement will establish their correctness. On the other hand, building upon a faulty foundation makes subsequent work futile. The main purpose of this paper is to determine whether the principle of prior constituent similarity, established by pre-task learning or otherwise, can account for the sorts of systematic generalizations found in human cognition. There are two main results (next two sections, respectively): (1) While constituent similarity *might* account for strong (syntactic) systematicity over *regularly* novel constituents by learning in previously similar contexts, it is highly unlikely that the necessary similar contexts occur for *radically* novel constituents. (2) Furthermore, constituent similarity, at best, can only encode one binary relation between any two constituents at any one time. It cannot address generalization based on a single relation drawn from multiple possible binary relations, contingent on a third constituent (i.e., ternary relation). A more recent proposal to include *context similarity* is also investigated. The implications of these results are discussed in the context of connectionist architectures that explicitly represent and process constituent relations.

CONSTITUENT SIMILARITY AND WITHIN CONSTITUENT CLASS RELATIONS

Similarity is a fluid concept. The similarity of the same pair of objects or events can vary between subjects within the same context; and within subjects across different contexts. But, at least for the purposes of connectionist modeling the concept is clear. Similarity between

[6]Their result is, at first, surprising since weights from the input unit encoding S were never trained, and so kept their initial random values. But, see also Phillips (1998) for an explanation.

two constituents refers to some measure of the "difference" between the constituents' vector representations. The dot product is a common measure, but it need not be the only one. In what follows, I will refer to constituent similarity as *any prior spatial relationship between two constituents*. The effect that similarity has on learning will depend on the activation function. A combination of similar constituents and activation function are chosen so as to maximize the likelihood of demonstrating systematicity. In the event that systematicity is not demonstrated, it is therefore unlikely to be demonstration by some other combination. After tracing a series of problems with this example and suggested replies, I argue that it cannot sustain a cognitively plausible explanation for systematicity.

Chair-colour example: Category-based similarity

In the target *chair-colour* inference task, subjects learn (know) that if *John painted the chair red*, then *The colour of the chair is red*. Assuming subjects can make this inference, it is reasonable to expect that they can make all the related inferences involving other colours, such as *green* and *blue*, even though it is unlikely that subjects would have been asked these specific inferences before. Constituents *green* and *blue* are called *regularly* novel constituents because they are new to the target task, but belong to the same category. A possible common pretext for learning the colours may simply be naming them (e.g., *This colour is red*).

The target task has the form: *John painted the chair X. What colour is the chair? X.* Assuming a basic capacity to segment input into words, the inference has two parts: (1) context - *What colour is the chair? John painted the chair*; and (2) constituent - *X*. The network is required to map the context and constituent to the same constituent (Figure 1(a)). In this case, the network makes use of internal representations for the constituents learned from a pre-task, say, *colour-identification* (Figure 1(b)).

Suppose there are N colours for which the appropriate internal representations were

learned. How many training examples on the target task are required to establish generalization? The number of training patterns depends on the number of weights in the context/constituent-to-constituent mapping, which in turn depends on the number of units representing context and constituent vectors. More compact (similar) internal representations for colour implies fewer units and weights, and therefore greater generalization. Assuming m context and n constituent units, the total number of weights is $(m+n)n$. Using linear activation functions, $n^2 + mn$ training patterns are required to configure weights to guarantee generalization to the remaining $N - n^2 + mn$ patterns.[7] Thus, in principle, provided the number of internal units representing context and constituent is small, the network will demonstrate strong systematicity (and universal generalization).

Figure 1 is a characterization of this sort of generalization. In the pre-task, there is no prior similarity between constituents. Constituents activate different regions in the input space, and are therefore mapped by different input-to-hidden and hidden-to-output weights. Thus, there is no generalization during the pre-task: learning that *red* stands for the colour **red**, does not say anything about the colour **blue**. However, pre-task training results in similar internal representations for the various colours: the same group of hidden units become active in the presence of a colour. Therefore, when this representation is used as the input/output for the target task, the same group of weights are being trained for all colours - training on some colours in the target task implies generalization to the other colours.

Insert Figure 1 about here

---

[7] In a linear system, parameters (in this case, weights) are uniquely determinable when there is at least one independent equation (generated by a training pattern) for each unknown parameter.

Problem 1: Cross-category generalization

The capacity to generalize depends on their being an appropriate pre-task that learns similar representations. In the case of colours, one can envisage that learning the names of colours (e.g., *This colour is blue*; etc) forms the basis for similarity. But, the presumption of common prior context does not so easily fit with *radically* novel constituents (i.e., constituents taken from other categories). It is less likely that one would have also learnt, say, *mango* as an instance of colour. Therefore, it is less likely for there to be any similarity between the internal representations for *mango* and the standard colours.

If there is no previously established similarity then there is no chance of generalization to such examples. Suppose the internal representation of *mango* does not lie on any of the dimensions for colour. That is, the internal representations for colours are activations over other units (Figure 2(b)). Now, while it is reasonable to expect the weights linking the corresponding input/output units for this constituent may have been trained on a prior categorization task (say, *This fruit is a mango*), there would be no such training for the new target task context (i.e., *John painted the chair ...*). Since the weights are untrained, no generalization is expected (Figure 2(a)).

Insert Figure 2 about here

Reply 1: Super-category similarity

The reply to this problem is to note that while the specific prior contexts of these constituents differ: one identifying a colour category and the other fruit, they both share the common abstract context of object identification. Even though two constituents may differ in their base category, they ultimately share a common super-category. Thus, the same in prin-

ciple argument for strong systematicity based on categories also applies to super-categories.

Problem 2: Non-category generalization

Ultimately, the appeal to prior training supposes every possible constituent in place of X has appeared in some, albeit remotely related, pre-task. Yet, it is still possible to generalize over constituents for which subjects have no prior experience: nonsense and foreign words are two examples. Indeed, it is this capacity that motivated Marcus's universality criterion and subsequent rejection of feedforward and recurrent networks.

For example, a perfectly meaningful yet unexperienced constituent occurs in the example *John painted the chair aka.* English-only speaking subjects would have no difficulty inferring the chair's colour, despite not knowing that *aka* is also known as *red*, in Japanese. There is no case of prior experience with this constituent, and therefore no grounds for category based similarity. As with the *mango* example, constituents represented on different units require additional training.

Reply 2: Constituent's constituent similarity

Perhaps the constituent similarity solution could be saved by noting that even nonsense and foreign words consist of recognizable components (e.g., common graphemes). In the previous two problems, the constituent was considered *atomic*, consisting of no other structure for the purposes of the target task (i.e., it is not by virtue of the word *green* being composed of the letters g, r, e and n that makes it stand in for the concept **green**). But, in the case of foreign words, one is still able operate on them without their semantic content, by syntactic decomposition. While one may not know the word *aka*, it is still composed upon known graphemes. By this reading, it is the constituent's constituents that share similarity, and so become the basis of generalization.

Problem 3: One-shot generalization

If one regards this type of constituent as itself compositional, then the same question of generalization that was applied at the level of word combinations (i.e., sentences) can also be applied at the level of grapheme combinations (i.e., words). In other words, how many grapheme combinations are necessary to achieve generalization to nonsense and foreign words? If nonsense and foreign words are represented over a different vector space from regular words (because they are treated syntactically, not semantically), then at least one word per dimension is required, for the same reasons as for regular words in the *chair-colour* task. The picture that this revision paints is one where subjects generalize to nonsense and foreign words only after training on other nonsense and foreign words in the same position. There is no chance of *one-shot* generalization over meaningless words.

Reply 3.1: Empirical issue

A final recourse one might offer then, as Niklasson and van Gelder do in their defence against Hadley's (1994b) criticisms, is to maintain that, in principle, constituent similarity can account for systematicity and more precise degrees of generalization are empirical issues. That is to say, the problem lies not with the principle *per se*, but aligning the principle to the relevant to be acquired data through a suitable implementation. But, as will be shown in the next section, there are generalizations for which this explanation cannot account for even in principle. However, before examining these cases, there is one other more recent alternative to be considered.

Reply 3.2: Contextual similarity

Boden and Niklasson (in press) have extended constituent similarity to be influenced by contextual similarity. Contextual similarity makes use of information in the current

context as a means of updating the representation for the novel constituent, rather than just relying on its representation learned from past contexts. Thus, the radical constituent *aka* is understood as a colour because of its surrounding context *John painted the chair aka*, not because of any established similarity to other colours. The architecture for this proposal uses the same combination of encoder and transformation networks, but error from the transformation network is also used to update weights in the encoder network.

Boden and Niklasson trained an encoder to learn internal representations for constituents *Ernie* and *Bo* as instances of bird and fish, respectively. A transformation network encoded facts *Ernie can fly* and *Bo cannot fly* by weights that mapped representations of *Ernie* and *Bo* (from the encoder network) to *true* and *false*, respectively. The two networks were given new fact *Jack can fly*, containing novel constituent *Jack*. Initially, weights that mapped the external representation of *Jack* (a unique local vector) to its internal representation in the encoder network were random, resulting in erroneous output when mapped by the transformation network where the target output was *true*. When this error was reduced by backpropagation from the transformation network to the encoder network the representation for *Jack* became similar to *Ernie* (see Boden & Niklasson, in press, Figure 11).

Although they did not test the method on novel inferences, one can see how the method might be applied here. An encoder network learns internal representations for the standard colours. An *assertion* network (i.e., another RAAM) learns statements such as *John painted the chair red* and *The colour of the chair is red* using the representations for colours generated by the encoder network. A transformation network learns inferences between these internally represented assertions so that *John painted the chair X* maps to *The colour of the chair is X*. For radical constituent *aka*, the assertion that *John painted the chair aka* is the current context provided by the assertion network. The error generated in encoding and decoding this assertion is used to update the encoder/decoder weights responsible for mapping

between the external and internal representations of *aka*. In doing so, it is supposed that the representation for *aka* becomes similar to the known colours, in much the same way as for the Boden and Niklasson example. Once error has been reduced, and consequently, similarity established, the new internal representation for *aka* can be used by the transformation network to infer the colour of the chair.

Whether or not this proposal can be extended to such inferences is not the concern here. Rather, of interest is the underlying principle. While generalization depended on the radical constituent being similar to normal constituents, similarity was achieved by current contextual information. It had nothing to do with the constituent's prior internal representation, which was in fact just random, or arbitrary.[8] It suggests that constituent similarity is not the crucial property that permits generalization, in this case. A point to which I will return later.

Summary

Why can I make inferences about constituents in novel contexts? The constituent similarity proposal says such constituents inherit similar representations to other constituents through learning in other common contexts. Dog is a noun learned from prior experience and so attains all the appropriate inferences afforded to nouns via its representational similarity to those nouns.

While this proposal seems intuitively plausible for regularly novel words, it becomes increasingly implausible for more radically novel words because the necessary common contexts for learning the required similarities are less likely to co-occur. The picture that emerges from the analysis of this proposal is that it relies upon relationships within a class (set) of constituents, while ignoring critical relationships between constituent classes.

[8]In general, it did not depend on its external representation either. When the non-zero dimension was randomly set between 0 and 1, the encoder still learned a similar representation to *Ernie* in 87% of 30 trials.

CONSTITUENT SIMILARITY AND BETWEEN CONSTITUENT CLASS RELATIONS

The *chair-colour* task is an instance of a general class of identity functions $I(X) = X$. The constituent similarity based solution attains generalization because the spatial relationship between any two constituents x_i and x_j in the domain is the same spatial relationship for the same two constituents in the co-domain. Thus, the basis for this solution is a common *within* constituent domain/co-domain (i.e., X) relationship. But, logically, the identity function is a relationship *between* a domain and a co-domain set of constituents. It just so happens that these two relationships coincide for this task. In tasks where these two relationships do not coincide, as exemplified in this section, constituent similarity cannot be relied upon as the basis for generalization.

Transverse patterning is an example of a stimulus-response task that depends on between constituent relations. A task instance, or problem set consists of three unique patterns (e.g., strings, shapes, etc) A, B and C, such that A predicts B; B predicts C; and C predicts A. Once the transverse patterning task structure has been learned from the first few problem sets, subjects only require one of the three stimulus-response pairs to predict the remaining two, for any new transverse patterning problem set.

There has been some recent debate as to whether subjects are in fact systematic over such logically structured tasks. van Gelder and Niklasson (1994) referred to data showing that even scientists do not always correctly use logical inference rules, such as *modus tollens* [i.e., (P implies Q) and (not Q), therefore (not P)]. Therefore, models of cognitive architecture are not required to be as systematic as these symbolic/logic rules imply. However, this apparent lack of systematicity may merely reflect interference from other sources. Under controlled conditions, subjects consistently make inferences implied by the underlying logical rules (Halford, Bain, Maybery, & Andrews, 1998). Indeed, such tasks are ideal tests for

systematicity in connectionist networks (Phillips & Halford, 1997; Phillips, 1999). (See, also, Bringsjorg, Noel, & Bringsjord, 1998, who showed that such logical errors are overcome with education, and therefore do not imply cognitive limitations.) The assumption that subjects can make the stated inferences for transverse patterning is reasonable when one considers that it is a simpler task (three constituents and one binary relationship) than the task used in Halford et al. (1998), which consisted of four constituents and two binary relationships, where subjects demonstrated perfect generalization by the fourth task instance.[9]

The constituent similarity proposal says that subjects systematically make correct inferences because prior learning on shapes has established suitable similarity between their internal (vector) representations. But for transverse patterning this explanation cannot be the case. Suppose constituents are taken from a general set of shapes, and the first stimulus-response pair for a new transverse patterning task is *triangle-square*. So that transverse patterning is directly comparable with the *chair-colour* task of the previous section, it is rewritten as *Associate triangle* $\rightarrow$ *square* (i.e., there is a context part followed by a constituent, predicting another constituent). The function mapping shapes to shapes is updated in accordance with this stimulus-response pair. Now, suppose the second stimulus is a *square*. According to the constituent similarity proposal, a network might infer the equivalent of, for example, a *square* is like a *triangle* (squares have *one-more-side* than triangles) just as a *pentagon* is like a *square* (pentagons have *one-more-side* than squares). Therefore, the predicted response is *pentagon*. But, clearly, the response depends on the yet unpresented third shape. If, instead, a *circle* was the third constituent, then square should predict *circle*. It is at this point the constituent similarity proposal clearly fails. *A priori*, for any measurement function, by definition, there is only one spatial relationship between any two constituents in any one state. There are not sufficient spatial relationships within the one

[9] Astur and Sutherland (1998) reported human subjects solved single transfer pattern problem sets. However, they did not examine generalization across problem sets.

state to accommodate the other possible responses that are contingent on the third stimulus. Alternatively, multiple units could make available different measurement functions simultaneously. But, that only introduces the new problem of deciding which measurement function should be relied upon. Of course, subjects can use the fact that *circle* is the only remaining constituent, and therefore the correct response. However, this knowledge cannot come from observing the first stimulus-response pair and any prior spatial relationship between the first and second stimuli. To accommodate the second response, either: the stimulus-to-response function must be *constructed* (updated) so as to map *square* to *circle*, and not *pentagon*, in the case where the necessary input similarity between *triangle* and *square* was not already in place. Or, the appropriate stimulus-to-response function must be *selected* using the information that *square* maps to *circle*, in the case where all possible measure functions were already in place. Either way, construction and selection require at least two training examples to learn the task, whereas only one example is required by subjects. Just as in the hypothetical association-based architecture (see Introduction), capacity is distributed across more than one representational state, whereas it is localized to one state in humans. The problem can be summed up by saying that transverse patterning consists of ternary relational information (Halford, 1993), whereas constituent similarity spatially encodes at best only binary relational information. Therefore, constituent similarity is ruled out as the principle underlying systematicity.

The implications of this analysis apply to connectionist models derived from the standard function approximation framework: That is, models whose state (usually determined by learning) encodes knowledge as a function from an input space to an output space, and whose generalizations on subsequent (test) inputs *in that state* are based on their similarity to prior (training) inputs relative to the encoded function (i.e., by extrapolation, or interpolation of the encoded function to test inputs). As such, the implications apply to feedforward and

recurrent networks trained and tested this way. The implications apply generally, since no assumptions were made regarding the input/output representations, connectivity, number of hidden units or layers, or the activation/error functions. Thus, it covers the majority of connectionist models used to date.

However, it was assumed that weights remained fixed during testing. Therefore, this analysis does not necessarily apply to those networks whose states are permitted to change on test input. But, as we shall see, these networks do not achieve systematicity via the constituent similarity principle. They rely on what might be called *relation approximation*.

CONTEXT (RELATIONAL) SIMILARITY AND BETWEEN CONSTITUENT CLASS RELATIONS

Boden and Niklasson's context similarity proposal differs from constituent similarity in that the current context is permitted to update the internal representation of a novel constituent *before* that constituent is used in a novel inference. This permits the novel constituent to be *aligned* to a known constituent before making an inference. The purpose of this section is to investigate how context similarity *could* also be applied to the transverse patterning task to see what implications this might have for constituent similarity. However, this will not be a defence of context similarity. Consequently, I will present a solution to transverse patterning, but I will not argue whether this solution is reliably learnable, or even cognitively plausible.

For transverse patterning, two networks are assumed: an auto-associative three-layered feedforward encoder/decoder network for mapping task elements A, B and C to internal representations, and recovering them; and a two-layered feedforward transformation network for representing the transverse patterning inferences (i.e., A predicts B, etc). The encoder/decoder network has two hidden units,[10] therefore the transformation network has

[10] Assuming sigmoid activation functions, two units are necessary and sufficient.

two input and two output units (Figure 3(a)).[11]

Assuming unique local input/output vectors for A, B and C, the input-to-hidden weights must encode these elements as hidden unit activation vectors A_h, B_h and C_h (respectively) so that two constraints are satisfied: (1) They must be decodable by the hidden-to-output weights; and (2) they must be transformable by the two-layered network so that encoded input A_h maps to encoded output B_h, B_h maps to C_h, and C_h maps to A_h. Geometrically, the transformation is satisfied by a rotation of the input space onto the output space by 120°. This rotation is realized by the weight matrix $(-\frac{1}{2}\ -\frac{\sqrt{3}}{2}\ \frac{\sqrt{3}}{2}\ \frac{1}{2})$, assuming for simplicity no biases and the identity activation function at the output units.[12] Suppose the transformation network has learned these, or similar weights from previous task instances. If the encoder/decoder learns an internal representation as depicted in Figure 3(b), generalization to the remaining two inferences is possible. With the transformation weights in place, the single training transformation $A \rightarrow B$ generates two training points for the encoder/decoder, indicated by solid squares labeled A (left activation space) and $B(A)$ (right space). In addition, two more training points B and C (left space) are generated, because mapping $A \rightarrow B$ is learned *in the context of* there being only three elements in each task instance. Hyperplanes are arranged to partition the space according to these four training points. In this case, when given the two test inferences, B and C are correctly mapped (by rotation) to the regions encoding C and A (dashed arrows), respectively.

However, even with the minimum number of hidden units, there is still considerable representational freedom so that even with suitable transformation weights there is an arrangement of hyperplanes that satisfies the encoding constraint and the single training inference constraint, but does not correctly infer one of the test cases (see Figure 3(c)).

---

[11] There is only one encoder/decoder, but two are shown to reflect the state before and after transformation.
[12] Sigmoids could also be used with the effect of rotating and shrinking the output space.

Insert Figure 3 about here

The point, though, is not whether context similarity can be made to solve this task, but what it says about the constituent similarity proposal. The constituent representations are almost arbitrary,[13] and generalization does not depend on whatever prior internal representation they may have had. Therefore, the extension to include context similarity may work, but not because of any contribution from constituent similarity.

The similarity at work here is between the transformation weight matrices for the different task instances. These weights must be the same or similar so that knowledge acquired in earlier task instances can be applied to facilitate generalization. These weights encode the relational structure of the task, independent of specific task elements. As such, this sort of architecture is derived from what could be called relation approximation. The properties of relational knowledge have been defined in Halford, Wilson, and Phillips (1998, sect. 2.2), and some of their implications for connectionist models have been analyzed in Phillips and Halford (1997) and Phillips (1999).[14] The problem with first-order connectionism is *not* that it relies on similarity, but that it relies on *constituent* similarity.

DISCUSSION

Within the same task paradigm (context/constituent-to-constituent inference), the constituent similarity proposal has gone from apparently plausible for systematic generalization across *natural* contexts to hopelessly inadequate for cross-category or category irrelevant generalizations. Two important problems arose with the constituent similarity proposal. First, it presupposes an unlikely amount of prior training to establish appropriate similarity

[13] Up to the point that they must be sufficiently distinct so as not to be transformed to the same region.

[14] For example, while this technique of *weight sharing* between subnetworks permits some degree of transfer between isomorphic task instances consisting of unique elements (see, e.g., Hinton, 1990), other mechanisms, such as enforcing one-to-one correspondence (alignment) between task and structure elements are needed to achieve degrees of transfer closer to subjects.

between constituents to guarantee generalization. Second, even when all possible spatial relationships between constituents are in place prior to the task, a single training example does not provide sufficient information to identify the right spatial relationship to achieve generalization. Constituent similarity alone fails to support the requisite degree of systematicity. Systematicity *may* be achieved by augmenting it with context similarity. But this achievement is at the expense of abdicating responsibility for generalization. In this regard, constituent similarity succeeds by not relying on constituent similarity.

The critical difference between failure and success lies in the relationships between constituents. In tasks that make use of the same within category constituent relationship in pre-tasks and target tasks, as in the *chair-colour* example for regularly novel constituents, constituent similarity (i.e., prior spatial relationships between constituents) may suffice to explain systematic generalization. But, in tasks involving multiple possible constituent relationships, as in the case of transverse patterning, constituent similarity does not suffice, since at best it can only code for binary relationships whereas ternary relationships are required.

Implicit and explicit relations

The problem of using spatial relationships to encode logical relationships makes contact with Kirsh's (1990) distinction between *implicit* and *explicit* information: Information is explicitly represented when the time required to access that information is constant in the size of the input, otherwise it is implicit. By this definition, the oddness or evenness of a number is explicitly represented since it only requires checking the number's last digit regardless of the size of the number (i.e., time is a constant function of the number of digits). But, primeness is implicitly represented since the number of possible factors, and therefore the number of steps, increases with the size of the number. Analogously, spatial relationships between constituents encode logical relationships implicitly because more training examples

are required to access those relationships. (Strictly, constant complexity means the resource [e.g., time, training examples] does not change, to within a constant factor, with the size of the function [problem]. But here I simply mean that operating on explicit information requires less resources than implicit information, since the size of the task does not change.)

Coding, or labeling is a common approach to making information explicit. Primeness, for example, can be made explicit by simply appending the letter 'p' to prime numbers only. Thus, regardless of the size of the number, one only need check the last character. Similarly, relationships between constituents are made explicit by affixing a name for that relationship. One only need check the name of the relation regardless of the number constituents that may partake in that relation. This technique is the basis for explicit representation of relations in a different class of connectionist architectures.

Connectionist alternatives: Explicit representation of relations

Tasks requiring operation over multiple possible binary relations, or equivalently ternary relations, are not isolated exceptions. In fact, according to one theory, they constitute a critical stage in cognitive development (Halford, 1993; Halford et al., 1998). Children before the age of five, said to be limited to binary relations, have great difficulty on tasks requiring ternary relations, such as transitive inference. Thus, they form an important and significant part of the cognitive repetoir, and connectionist architectures must be capable of modeling them if they are to claim a complete foundation for cognition.

Halford, Wilson, Guo, Gayler, Wiles, and Stewart (1994) showed how relationships between constituents are made explicit (i.e., readily accessible) using tensor networks developed by Smolensky (1990). In a tensor network, primeness is made explicit by binding (taking the outer product of) vectors representing numbers and a vector representing the unary relation *is-prime*. Prime numbers belong to the set: Prime(N) = {2, 3, 5, ...}. They are represented

by the rank two tensor (matrix): $T_P = \vec{P} \otimes \vec{2} + \vec{P} \otimes \vec{3} + \ldots$, where $\otimes$ is the outer product operator. Primeness is explicit in that it is determined by a one step *inner product* ($\odot$) operation: $T_P \odot \vec{7} = \vec{P}$, under the condition of orthonormality between constituents 2, 3, etc. Similar procedures apply to higher arity relations (e.g., binary, ternary), and have been used as the basis for analogical reasoning. For example, simple analogies of the form *A is to B as C is to ?* rely on explicitly identifying and applying a common relationship,[15] as in *Mare is to foal as cat is to ? (kitten)*. Assuming a rank three tensor that explicitly links the relationship *Mother-of* ($\vec{M}$) to the relevant constituents: $T = \vec{M} \otimes \vec{m} \otimes \vec{f} + \vec{M} \otimes \vec{c} \otimes \vec{k}$, the missing value is retrieved in a two step process: 1. $T \odot \vec{m} \odot \vec{f} = \vec{M}$; 2. $\vec{M} \odot \vec{c} \odot T = \vec{k}$ (Halford et al., 1994).

The explicit representation of constituent relations also appears in other cognitive models of, for example, analogy (Hummel & Holyoak, 1997) and reasoning (Shastri & Ajjanagadde, 1993). In particular, Hadley and Hayward (1997) proposed an alternative Hebbian-based learning model that demonstrated strong semantic systematicity. Constituent relations are not represented by spatial relationships between constituents in these models. In fact, they generally work with mutually orthogonal constituents. Instead, relations are explicitly represented by distinct units, together with mechanisms for binding/accessing constituents to/from their relations.

Reconciliation and the future of connectionism

Explicitly representing relations is not the final word on the issue of representation. Just providing for the same sort of property as symbol systems (i.e., arbitrary binding) suggests inheriting the same sorts of problems. If classicism accounts for capacity, then it has less to say about performance, as all instances of a common structure are regarded equal. Yet, the generalizations that *are* made under specific conditions can be affected by

[15] Also referred to as *relational similarity* (Goldstone, Medin, & Genter, 1991).

such structure-independent factors as frequency effects, which connectionism is better placed to explain. So, while sensitivity to context in connectionist models undergeneralizes on the issue of capacity, sensitivity to structure in classical models overgeneralizes on the issue of performance. Networks that provide for arbitrary binding, in a sense, remove the blinkers only to be blinded by the light.

However, the apparent paradox only arises when one considers the two principles (connectionist similarity and classicist structure sensitivity) as being in conflict, rather than cooperation. The possibility of a mixed connectionist-classicist theory has been suggested by Harnad (1990), and also acknowledged by Fodor (1997). And, a number of connectionists have experimented with integrating neural and symbolic models. See, for example, the collection by Sun and Alexandre (1997), though, none of these models specifically addresses the problem of systematicity. While it is beyond the scope of this paper to address the implications of mixed theories (and models), which raise new issues (e.g., under what conditions are either applied?), it is still useful to suggest what such a model might look like for the connectionist.

As argued, connectionist representations must have both context-sensitive and context-independent components.[16] The context-sensitive component, as exemplified by feedforward and recurrent networks, identifies the relevant context-dependent structure, or relation predicate. The context-independent component, as exemplified by tensor and related networks is used to make arbitrary bindings between the identified context and novel constituents. The apparent paradox is resolved for radically novel constituents/relations by relying on *normal*[17] constituents and their relationships to carry the relevant contextual information. Novel constituents are bound to structural elements regardless. So long as there is sufficient contextual information carried by the other constituents, a network can demonstrate

[16] Boden and Niklasson (in press) have also argued for these two components.

[17] Constituents and their relationships appearing in their usual contexts (e.g., mango as a fruit).

strong systematicity/universal generalization to unseen constituents. One example is the integration of tensor and recurrent networks (Phillips, 1994),[18] another is a type of structured hebbian network (Hadley & Hayward, 1997).

Representation combines context-sensitive and context-independent components. The relationship between these two components is illustrated for the normal and radical colours (Figure 4). Normal colour constituents (e.g., *red*, *blue*) are learned as colours. Thus, their presence triggers the colour predicate. This predicate remains active as the context for subsequent context-independent processing. The radical constituent *mango* does not itself trigger the predicate, since it was never learned as an instance of a colour, in contrast to the constituent similarity proposal. However, colour remains active by the prior appearance of normal colours to provide a link to *mango* via the binding units. It is at this point the network breaks away from reliance on constituent spatial similarity. Any vector value for *mango* is sufficient for binding, and unbinding. By the property of outer and inner products: $\vec{C} \otimes \vec{m}$ (binding); and $\vec{C} \odot \vec{C} \otimes \vec{m} = \vec{m}$ (unbinding), any novel constituent is linkable to colour, provided the colour predicate vector ($\vec{C}$) is orthonormal to other possible predicates. Importantly, though, *mango* or other novel constituent is only bound (interpreted) as a colour in the present context, instantiated by previous normal colour constituents. In principle, prior context restricts arbitrary binding of novel constituents.

Insert Figure 4 about here

Of course, then one must explain why normal constituents trigger the right predicate/structure and not others, since constituents normally participate in many possible structures, in any general sense.[19] But the problem of choosing one structure from many must be less severe

[18] Though this network does not address embedding.

[19] Models of analogy only address this problem in highly circumscribed domains.

than the problem of having no structure to choose from at all.

Put simply, connectionist networks require additional units that explicitly encode relational knowledge. This claim converges with Clark and Thornton's (1997) arguments that solving *type-2* (i.e., relational) problems requires some form of input *representational redescription* (Karmiloff-Smith, 1992). They tentatively suggested *incremental learning* (Elman, 1993): the idea of incrementally increasing the number of training patterns, or retainable weights. But, the approach I have been critiquing here shares common roots with this idea. Consequently, redescription cannot just be recoding constituent similarities for the reasons already put forth. Additional sources for recoding are needed.

CONCLUSION

The main purpose of this paper was to evaluate the constituent similarity proposal. The main conclusion is that while constituent similarity may support some types of generalization, it cannot account for systematicity in its entirety, irrespective of the type of constituent representation used (e.g., local, distributed). Firstly, humans generalize to radically novel constituents that are highly unlikely to have the necessary similarity to known constituents. Secondly, constituent similarity at best encodes binary spatial relationships, which do not suffice for tasks involving ternary relationships, as exemplified in the transverse patterning task. This work extends the results of Phillips (1995, 1998) and Marcus (1998a, 1998b) from the case of local constituent input/output vectors to include the case of distributed constituent input/output vectors.

For some (classicists), these conclusions may seem like stating the obvious. But that is because the classicist framework presupposes independence between constituent representation and syntactic relations. Connectionism, generally, does not make this assumption. Without a clear correspondence between the two frameworks, implications from one do not

immediately carry over to the other. Therefore, to make the limitation of constituent similarity obvious the same connectionist paradigm was used in taking the proposal to its logical extension where it clearly failed.[20] While these points may appear obvious (in hindsight), they are not trivial: Any network that relies on constituent (*first order*) similarity will ultimately fail on logically structured tasks, or what one might term relational (*second-order*) similarity, where the similarity is not between constituents, but between the relations over them (or, what Shepard & Chipman, 1970, called *second-order isomorphism*).

But, making use of relational similarity requires, by definition, the explicit representation of those relations. At this point, standard function approximation style connectionism[21] gives way to a more symbol/relation-approximation style connectionism that explicitly represents constituent relations permitting the arbitrary binding of variables to values. However, to *only* implement a symbol system is to likely inherit the same sorts of problems. I have suggested that the most promising way forward for connectionism is by integration of context-sensitive (e.g., feedforward) networks and context-insensitive (e.g., tensor) networks. The critical question then becomes how.

It is perhaps ironic that advocates of explicit representations in cognitive development (Karmiloff-Smith, 1992) should consider the sorts of connectionist models that rely on, as shown here, implicit representations (Elman, Bates, Johnson, Karmiloff-Smith, Parisi, & Plunkett, 1996). But, just as Karmiloff-Smith has argued for explicit representation in cognitive development, I have been arguing for explicit representations in the development of connectionist cognitive models. Just like the five-year-olds, connectionist models are in for a new stage of development.

[20] By contrast, for English past tense acquisition, where the same basic principle is clouded by a more complex domain, the issue is still being debated (see Marslen-Wilson & Tyler, 1998, for a recent review).

[21] See also Hadley (1999) for very general arguments against universal function approximators (e.g., feedforward networks) as models of higher cognition.

Acknowledgements

I thank Lars Niklasson and two anonymous reviewers for their thoughtful comments, which have helped improve the presentation of this work.

REFERENCES

Astur, R. S., & Sutherland, R. J. (1998). Configural learning in humans: The transverse patterning problem. *Psychobiology*, *26*(3), 176–182.

Boden, M., & Niklasson, L. (in press). Semantic systematicity and context in connectionist networks. *Connection Science*, *12*(1).

Bringsjorg, S., Noel, R., & Bringsjord, E. (1998). In defence of logical minds. In M. A. Gernsbacher & S. J. Derry (Eds.), *Proceedings of the Twentieth Annual Meeting of the Cognitive Science Society*, pp. 173–178. Erlbaum Erlbaum.

Brousse, O. J. (1991). *Generativity and systematicity in neural network combinatorial learning.* Ph.D. thesis, University of Colorado, Boulder, CO.

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

Chrisman, L. (1991). Learning recursive distributed representations for holistic computation. *Connection Science*, *3*, 345–366.

Clark, A., & Thornton, C. (1997). Trading spaces: Computation, representation and the limits of uniformed learning. *Behavioral and Brain Sciences*, *20*, 57–90.

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

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

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

Fodor, J. A. (1997). Connectionism and the problem of systematicity (continued): Why Smolensky's solution still doesn't work. *Cognition*, *63*(1), 109–119.

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

Goldstone, R. L., Medin, D., & Genter, D. (1991). Relational similarity and the non-independence of features in similarity judgements. *Cognitive Psychology*, *23*, 222–264.

Hadley, R. F. (1994a). Systematicity in connectionist language learning. *Mind and Language*, *9*(3), 247–272.

Hadley, R. F. (1994b). Systematicity revisited: Reply to Chater and Christiansen and Niklasson and van Gelder. *Mind and Language*, *9*(4), 431–444.

Hadley, R. F. (1999). Cognition and the computational power of connectionist networks. In M. Hahn & S. C. Stoness (Eds.), *Proceedings of the Twentyfirst Annual Conference of the Cognitive Science Society*, pp. 196–201. Hillsdale, NJ: Lawrence Erlbaum.

Hadley, R. F., & Hayward, M. B. (1997). Strong semantic systematicity from Hebbian connectionist learning. *Minds and Machines*, *7*, 1–37.

Halford, G. S. (1993). *Children's understanding: The development of mental models.* Hillsdale, NJ: Lawrence Erlbaum.

Halford, G. S., Bain, J. D., Maybery, M. T., & Andrews, G. (1998). Induction of relational schemas: Common processes in reasoning and complex learning. *Cognitive Psychology*, *35*, 201–245.

Halford, G. S., Wilson, W. H., Guo, J., Gayler, R. W., Wiles, J., & Stewart, J. E. M. (1994). Connectionist implications for processing capacity limitations in analogies. In K. J. Holyoak & J. Barnden (Eds.), *Advances in Connectionist and Neural Network theory*, chap. 7. Norwood, NJ: Ablex.

Halford, G. S., Wilson, W. H., & Phillips, S. (1998). Processing capacity defined by relational complexity: Implications for comparative, developmental, and cognitive psychology. *Behavioral and Brain Sciences*, *21*(6), 803–831.

Harnad, S. (1990). The symbol grounding problem. *Physica D*, *42*, 335–346.

Hinton, G. E. (1990). Mapping part-whole hierarchies in connectionist networks. *Artificial Intelligence*, *46*(1-2), 47–76.

Hummel, J. E., & Holyoak, K. J. (1997). Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review*, *104*(3), 427–466.

Karmiloff-Smith, A. (1992). *Beyond Modularity: A developmental perspective on cognitive science*. Cambridge, MA: MIT Press/Bradford Books.

Kirsh, D. (1990). When is information explicitly represented?. In P. Hanson (Ed.), *Information, Language and Cognition: Vancouver Studies in Cognitive Science*. Vancouver, BC: UBC Press.

Marcus, G. (1998a). Can connectionism save constructivism?. *Cognition*, *66*, 153–182.

Marcus, G. (1998b). Rethinking eliminative connectionism. *Cognitive Psychology*, *37*, 243–282.

Marslen-Wilson, W., & Tyler, L. K. (1998). Rules, representations, and the English past tense. *Trends in Cognitive Sciences*, *2*(11), 428–435.

Niklasson, L., & Boden (1999). Content, context and connectionist networks. In M. Hahn & S. C. Stoness (Eds.), *Proceedings of the Twentyfirst Annual Conference of the Cognitive Science Society*, pp. 474–479. Hillsdale, NJ: Lawrence Erlbaum.

Niklasson, L., & van Gelder, T. (1994). Systematicity and connectionist language learning. *Mind and Language*, *9*(3), 288–302.

Phillips, S. (1994). Strong systematicity within a connectionist framework: The tensor-recurrent network. In A. Ram & K. Eiselt (Eds.), *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society*, pp. 723–727. Lawrence Erlbaum.

Phillips, S. (1995). *Connectionism and the problem of systematicity.* Ph.D. thesis, The University of Queensland, Department of Computer Science, Brisbane, Australia. Available from http://www.etl.go.jp/~stevep.

Phillips, S. (1998). Are feedforward and recurrent networks systematic? Analysis and implications for a connectionist cognitive architecture. *Connection Science*, *10*(2), 137–160.

Phillips, S. (1999). Systematic minds, unsystematic models: Learning transfer in humans and networks. *Minds and Machines*, *9*(3), 383–398.

Phillips, S., & Halford, G. S. (1997). Systematicity: Psychological evidence with connectionist implications. In M. G. Shafto & P. Langley (Eds.), *Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society*, pp. 614–619.

Phillips, S., Halford, G. S., & Wilson, W. H. (1995). The processing of associations versus the processing of relations and symbols: A systematic comparison. In J. D. Moore & J. F. Lehman (Eds.), *Proceedings of the Seventeenth Annual Conference of the Cognitive Science Society*, pp. 688–691.

Pollack, J. B. (1990). Recursive distributed representations. *Artificial Intelligence*, *46*, 77–105.

Shastri, L., & Ajjanagadde, V. (1993). From simple associations to systematic reasoning: A connectionist representation of rules, variables and dynamic binding using temporal synchrony. *Behavioral and Brain Sciences*, *16*(3), 417–494.

Shepard, R., & Chipman, S. (1970). Second-order isomorphism of internal representations: Shapes of states. *Cognitive Psychology*, *1*, 1–17.

Smolensky, P. (1990). Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence*, *46*, 159–216.

Sun, R., & Alexandre, F. (Eds.). (1997). *Connectionist-symbolic integration: From unified to hybrid approaches*. Mahwah, NJ: Lawrence Erlbaum.

van Gelder, T., & Niklasson, L. (1994). Classicism and cognitive architecture. In A. Ram & K. Eiselt (Eds.), *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society*, pp. 905–909. Lawrence Erlbaum.

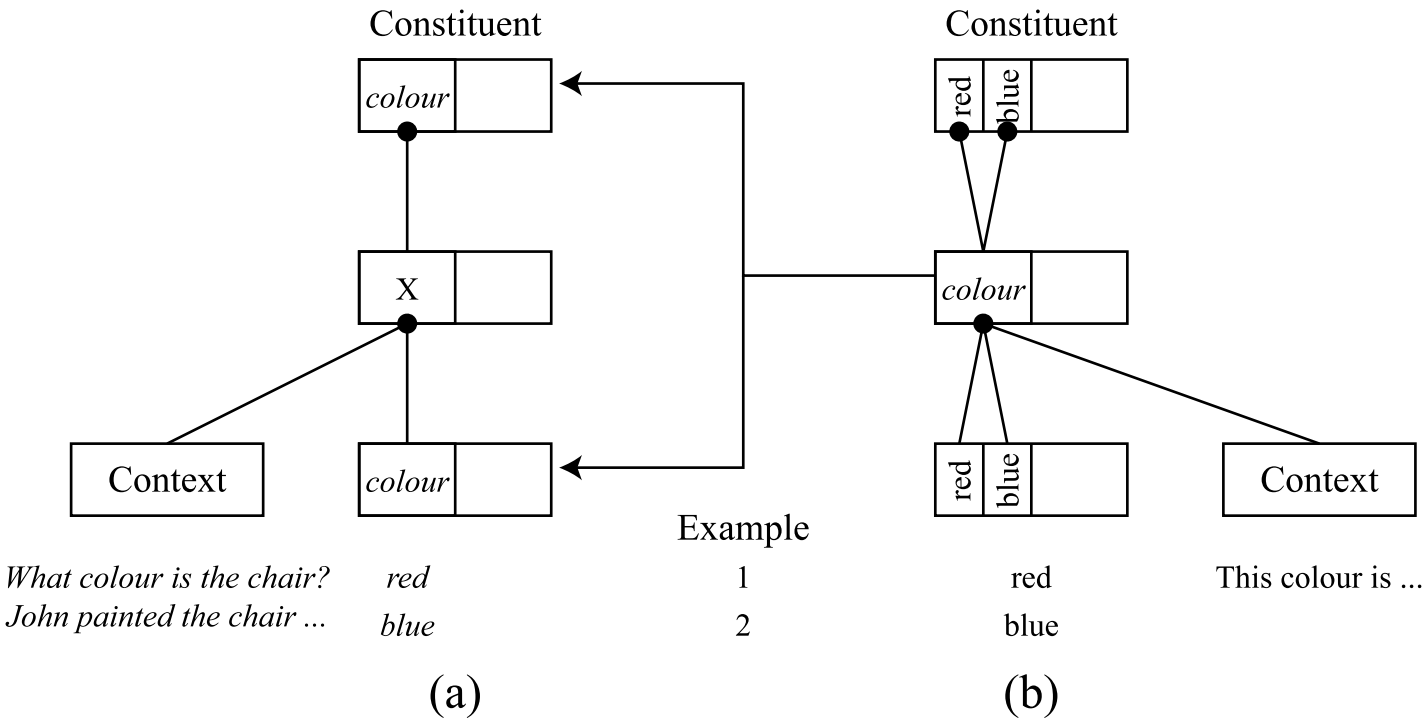
Figure Captions

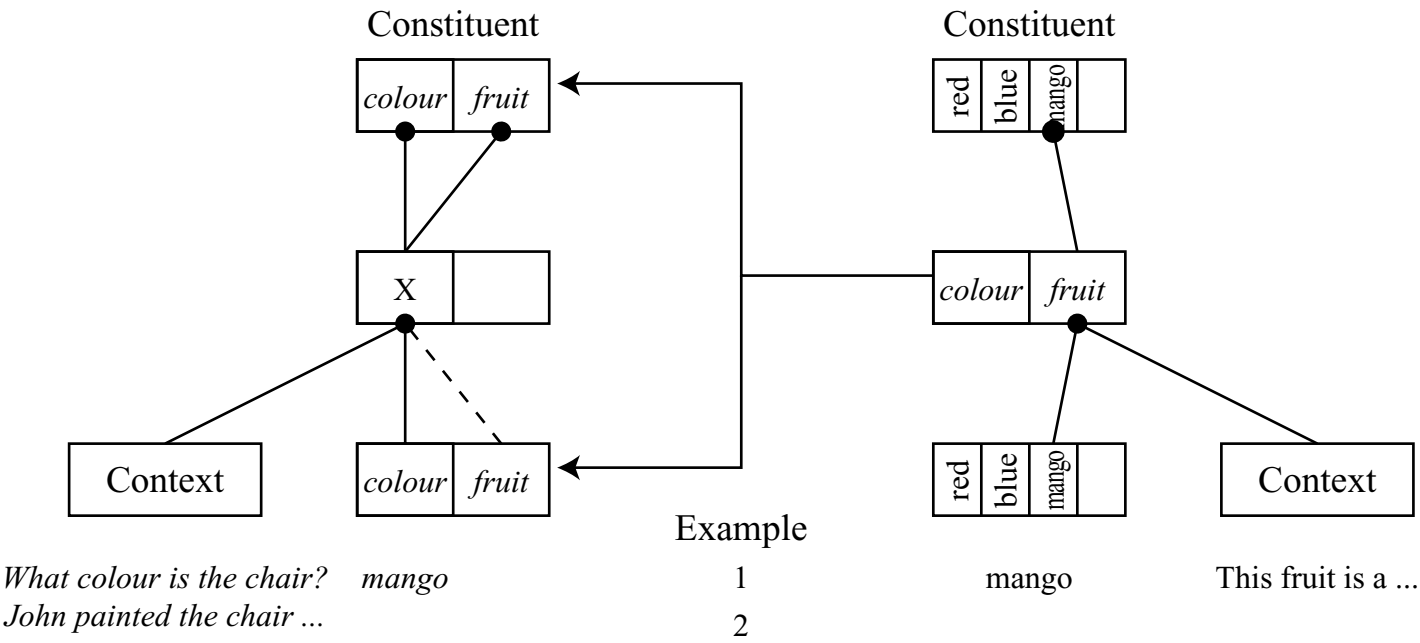
Figure 1. A network for learning the *chair-colour* task (a), using internal representations for colour constituents learned from task (b).

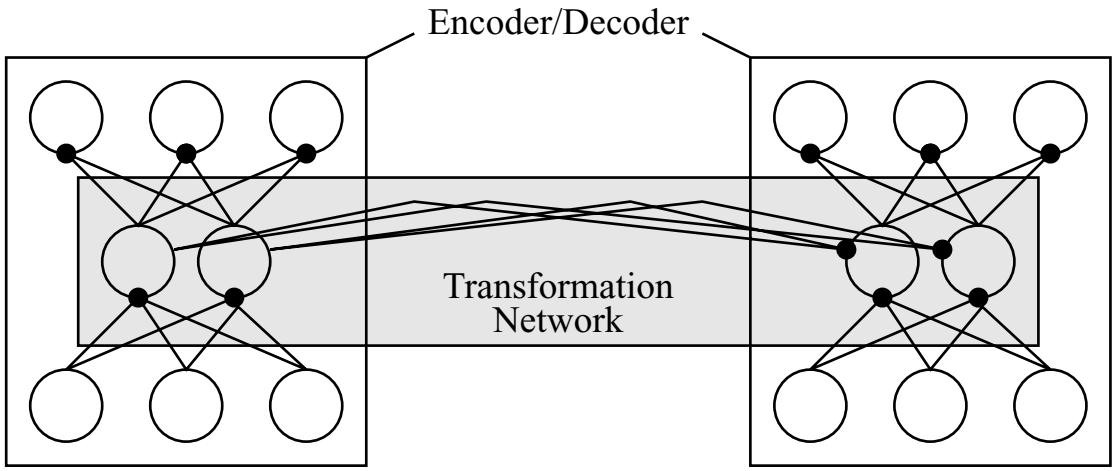
Figure 2. A network for learning the *chair-colour* task (a), using internal representations for fruit constituents learned from task (b). Dashed connections indicate untrained weights.

Figure 3 Encoder/decoder and transformation networks for transverse patterning (a). Hidden unit activations (squares) and output unit hyperplanes (thick lines) for the encoder/decoder network resulting in no generalization errors (b); and one generalization error (c). Solid squares and arrows indicate training points and transformations; empty squares and dashed arrows indicate test points and transformations, respectively. Star indicates test inference error.

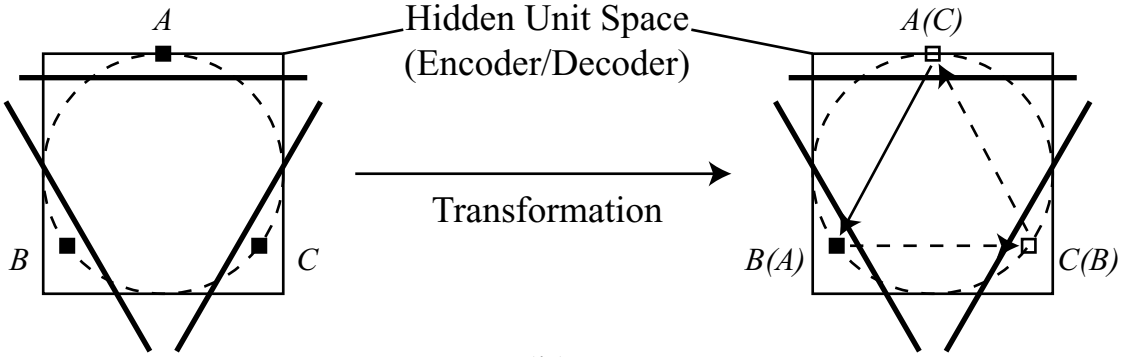
Figure 4. Scheme for integrating context-sensitive and context-independent network components. Shared circles indicate currently active units. Thin lines indicate connections; medium and thick lines indicate previous and current activation flow, respectively.


Constituent
Constituent
colour
red
blue
X
colour
Context
colour
red
blue
Context
Example
What colour is the chair?
John painted the chair ...
red
blue
1
2
red
blue
This colour is ...
(a)
(b)

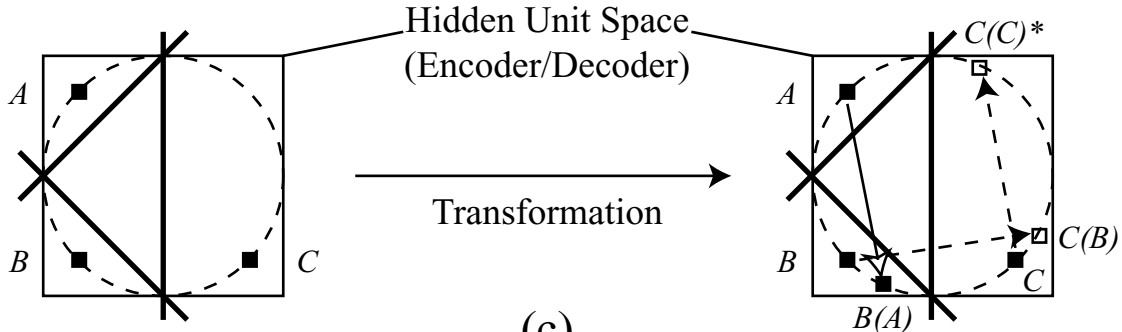

Constituent
colour
fruit
X
Context
colour
fruit
Constituent
red
blue
mango
colour
fruit
red
blue
mango
Context
Example
What colour is the chair?
John painted the chair ...
mango
1
2
mango
This fruit is a ...
(a)
(b)


Encoder/Decoder
Transformation
Network

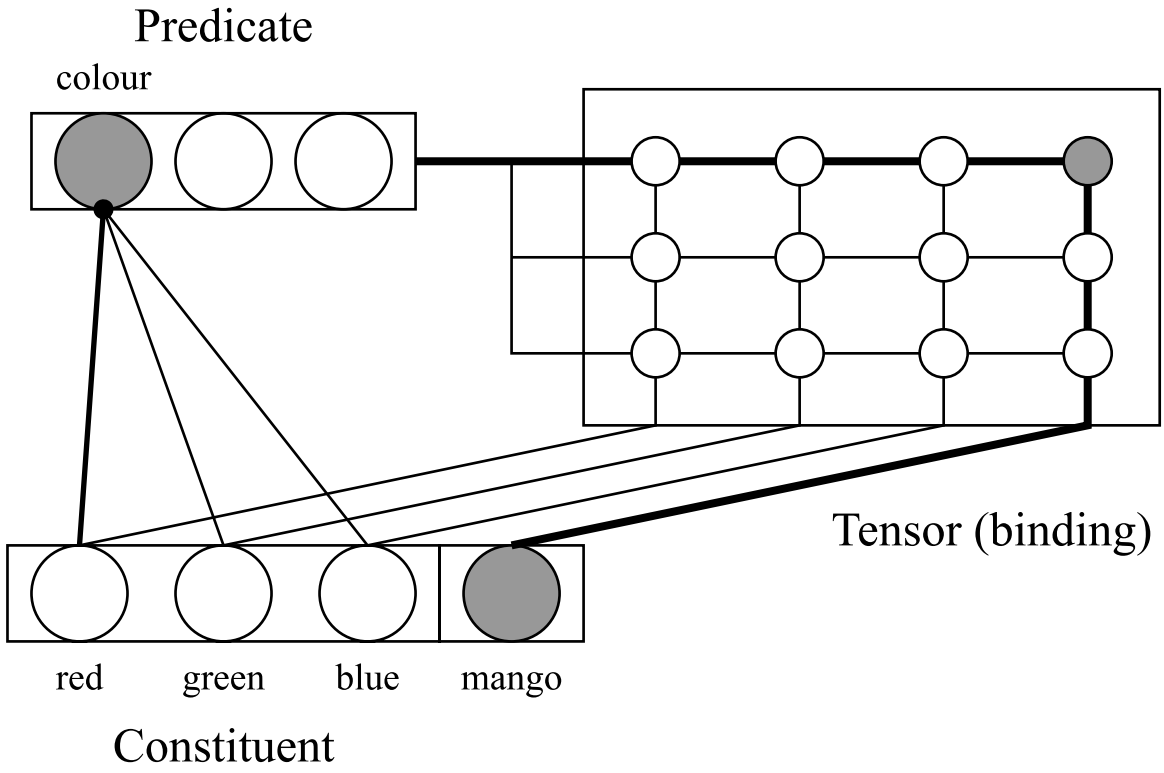
(a)


A
Hidden Unit Space
(Encoder/Decoder)
A(C)
Transformation
B
C
B(A)
C(B)

(b)


Hidden Unit Space
(Encoder/Decoder)
C(C)*
Transformation
A
B
C
A
B
C
C(B)
B(A)

(c)


Predicate
colour
Tensor (binding)
red
green
blue
mango
Constituent