The Psychological Record, 2001, 51, 409-428

A TRANSFER OF SEQUENCE FUNCTION VIA EQUIVALENCE IN A CONNECTIONIST NETWORK

FIONA LYDDY
University of Wales Institute, Cardiff

DERMOT BARNES-HOLMES
National University of Ireland, Maynooth

PETER J. HAMPSON
University of the West of England, Bristol

Connectionist networks may provide useful models of stimulus equivalence and transfer of function phenomena. Such models have been applied to a range of behavioral tasks and have demonstrated transfers of function via equivalence relations following appropriate training, with networks accurately simulating the behavior of human subjects. In the current study, a connectionist network was pretrained on a series of equivalence and sequence tasks to simulate the preexperimental experience of an adult subject. It was then exposed to the equivalent of six conditional discriminations, and was tested for the formation of three 3-member equivalence classes (corresponding to A1-A2-A3, B1-B2-B3, C1-C2-C3). It was subsequently trained to produce a pair of four part sequences (corresponding to B1→B2→Ct1→B3 and B3→B2→Ct2→B1, where Ct1 and Ct2 represented contextual cues) before being tested for transfer, through equivalence, of the sequence responses to the C stimuli. Following appropriate pretraining, the network showed the formation of three equivalence classes and a transfer of sequence function to the nontrained C stimuli (producing the novel sequences C1→C2→Ct1→C3 and C3→C2→Ct2→C1). A control network, which was not exposed to conditional discrimination training, failed to demonstrate equivalence and the transfer of sequence function, as predicted by findings from experimental demonstrations with human participants. Network performance was analyzed as a function of amount of pretraining and a number of psychologically plausible training methods are presented. The data suggest that connectionist networks may provide accurate and plausible models of stimulus equivalence and transfer of function phenomena in natural language.

Connectionist networks have been used to model many aspects of human cognitive processing (including natural language syntax; e.g., see Plunkett & Marchman, 1993; Rumelhart & McClelland, 1986). However, the plausibility of the simulations has sometimes been questioned (e.g., see Massaro, 1988; Pinker & Prince, 1988). One problem stems from the difficulty in defining tasks in such a way that they can be simply and abstractly represented by a connectionist network and still maintain their plausibility as simulations of natural language. As analogies of human

Reprint requests may be sent to Fiona Lyddy, Department of Psychology, University of Wales Institute, Llandaff Campus, Cardiff CF5 2YB, Wales.

language, the task sets developed to explore equivalence class formation and transfers of function are abstract, yet plausible, and may be well suited to connectionist simulation. Studies of derived relational responding use highly specific, well defined analogies of language tasks that can be readily translated into input patterns for a connectionist network. Furthermore, they tend to be confined to a specific domain (for example, investigating just one aspect of sequence responding). Target or goal performance is well specified and network performance can be compared to that of human participants on very similar abstract tasks, increasing the plausibility of the simulation. In this way, the true potential of connectionist models of language might be explored using tasks that are not biased against network capabilities in the way that less clearly defined language tasks may be. Furthermore, if connectionist networks can provide plausible simulations of behavioral data, they may provide a useful method of testing issues that are not easily addressed using human subjects (see Dougher & Markham, 1994, for a discussion of such issues).

An example of such a research agenda is provided by Cullinan, Barnes, Hampson, and Lyddy (1994). In an experimental demonstration, undergraduate subjects were trained on a set of two response sequences, B1→B2 and B2→B3, and they were tested for the emergence of a novel, untrained three-response sequence B1→B2→B3. The subjects were then trained on a set of six related conditional discriminations using a matching-to-sample procedure (i.e., A1-B1, A1-C1, A2-B2, A2-C2, A3-B3, A3-C3) and subsequently demonstrated the formation of three equivalence classes (A1-B1-C1, A2-B2-C2, A3-B3-C3). They were later tested to determine whether the trained (B1→B2 and B2→B3) and emergent (B1→B2→B3) response sequences would transfer to the nontrained C stimuli via equivalence relations (C1→C2, C2→C3, C1→C2→C3). Subjects who had received the conditional discrimination training showed this transfer of sequence function, whereas control subjects, who did not receive conditional discrimination training, failed to show this effect.

In the second part of the study, a connectionist model based on the earlier work of Barnes and Hampson (1993) was used to simulate these data. As in the Barnes and Hampson simulation, the model was provided with extensive pretraining designed to simulate specific aspects of the preexperimental verbal history of a typical adult subject. The model was then trained on a set of tasks corresponding to the training of six conditional discriminations and the two 2-response sequences in the human experiment. When tested, the model produced the nonexplicitly trained three-response sequence and, via equivalence relations, the derived two- and three-response sequences. In effect, the model's performance closely paralleled the response patterns observed for the experimental subjects. A control model, which did not receive the conditional discrimination training, failed to show the transfer of function, simulating the data obtained from the human control group.

In addition, the model's potential for simulating more subtle features

of the task was apparent. The results suggested that equivalence was only demonstrated on the test (experimental) stimuli following sufficient training across multiple-exemplars, of functionally equivalent tasks, in the preexperimental history of the network. This history of multiple-exemplar training in equivalence responding is normally assumed to be provided by the experimental subjects' verbal community prior to their demonstrating the emergence of nontrained equivalence relations in the typical experiment (e.g., Hayes, 1991). Controlling this preexperimental history of multiple-exemplar training in humans is not feasible, however. Connectionist networks could address this problem insofar as they may be used to simulate the preexperimental conditions of derived language performances that can not be readily manipulated using human subjects. By manipulating aspects of the tasks presented to the network, the early emergence of equivalence and transfer of function could be charted in a highly controlled manner. In addition, the network can easily be deprived of stimulation to simulate an impoverished environment or lesioned to simulate physical damage to the system. As a tool for behavioral research, it may provide a flexible new methodology to complement more traditional methods.

The Cullinan et al. (1994) findings were also of theoretical interest, highlighting stimulus equivalence and transfer of function phenomena as possible mechanisms underlying complex syntax. Based on the data, it was argued that the derived transfer effects shown, by the model and by the human subjects, may parallel the emergence of three-word and more complex multiword sequences in natural language. Specifically, it was argued that the simple syntax of two-word utterances may give rise to three-word utterances and, via the formation of equivalence classes, to novel two- and three-word utterances. In this way, a transfer of sequence function through equivalence may allow the production of novel utterances and could form the basis of a behavioral interpretation of the generativity of language (Barnes-Holmes, Barnes-Holmes, Roche, Healy, Lyddy, Cullinan, & Hayes, 2001). In order to test the explanatory potential of equivalence and transfer of function as processes underlying syntactic development, more complex aspects of syntax could initially be modeled in a connectionist network. Features of syntax might therefore be represented in terms of multiple input-output mappings learned by a network, corresponding to a highly controlled history of reinforcement underlying stimulus equivalence and transfer of function in language-able humans. The potential contribution of connectionist modeling to behavior analysis is reciprocated by the provision of structured tasks for simulation. The well-defined tasks typically employed in stimulus equivalence and transfer of function studies may lend themselves to a connectionist account of natural language.

In the current study, the kinds of experimental tasks developed to examine equivalence and transfer of function phenomena formed the basis of connectionist tasks sets, in order to examine the emergence of complex sequence responses through equivalence in the highly controlled context of a computational model. The specific goal here was

to address the role which stimulus equivalence and transfers of function might play in the acquisition of sequence responses of the type evident in active and passive form sentences. If the lexical categories (such as nouns and verbs) in sentences come to participate in equivalence classes, a transfer of the positional information in active-passive sentence counterparts may occur, through equivalence, in the absence of further instruction. In this way, the components of sentences may come to participate in equivalence classes that allow their sequence functions to transfer to new stimuli, thus generating novel utterances. A network similar to that of Cullinan et al. (1994) was designed to learn to produce response sequences analogous to active and passive voice sentences. In a simple utterance, there is an action generally associated with the subject of the sentence, and there is often an object of the action. The subject of the sentence may be the origin or agent of the action, for example, the sentence 'The boy is eating' consists of an agent-verb construction. The action may also relate to an object, often the patient or recipient of the action; for example, 'The boy is eating the apple' contains the additional recipient of the action. A simple three-component active-voice sentence may consist of an agent, a verb in active form, and a patient. The formulation of the passive-voice counterpart differs in two ways. There is an inversion of word order to patient-action-agent and there is a passivized verb form (passive form verb plus auxiliary plus 'by'). For example, the active-form sentence 'the boy is eating the apple' would become 'the apple *is being* eaten *by* the boy.' These syntactic categories reflect an abstract relationship with the verb in the sentence, rather than a direct relationship to its meaning. The sentences are clearly related, they have a very similar meaning, although their surface formulations differ. Active- and passive-voice sentence pairs have the same basic semantic content, although the form used may produce a difference in emphasis. Because they have the same meaning yet different structures, active and passive sentences have been manipulated in many studies of language processing. It has been argued that studying such sentences may allow one to control for the effect of meaning while manipulating structural information, and therefore to examine the autonomy of syntactic and semantic processing (e.g., see Fodor, Bever, & Garrett, 1974; Garrett, 1990; Slobin, 1966). In connectionist task sets, meaning can be stripped away altogether, leaving the structural component as a pure sequence string.

In the current study, a connectionist network was trained on a series of input-output mappings corresponding to a set of conditional discriminations (A1-B1, A1-C1, A2-B2, A2-C2, A3-B3, A3-C3) and was tested for the formation of three equivalence classes (A1-B1-C1, A2-B2-C2, A3-B3-C3). It was then trained on a sequence analogy of an active-voice sentence (B1→B2→Ct1→B3, where Ct1 is a flag or contextual cue for active form) and its passive form counterpart (B3→B2→Ct2→B1, where Ct2 is a flag for passive form). For the network to demonstrate a transfer of sequence function through equivalence, these sequence

responses had to transfer to the nontrained C stimuli, eliciting the derived sequences C1→C2→Ct1→C3 and C3→C2→Ct2→C1 on testing. This derived response and the formation of equivalence classes was predicted to emerge only following appropriate explicit training on similar task sets (simulating the preexperimental history of human subjects). To this end, the network was pretrained on a number of analogous task sets (i.e., multiple-exemplars), before being exposed to the critical test material. Network performance was analyzed as a function of the amount of pretraining that the network was exposed to.

In previous simulations of behavioral data, interference of old learning by new stimulus sets became apparent. This phenomenon is particularly problematic for psychologically plausible training methods (see McCloskey & Cohen, 1989) and was evident in the Cullinan et al. (1994) model. In the current study, a number of network training manipulations were conducted in order to address the issue of information loss from early training sets, and to increase the plausibility of the simulation for comparison to human data.

In summary, the goals of the current study are to demonstrate the formation of an equivalence class and a transfer of sequence function via equivalence in a connectionist network, to explore the role of stimulus equivalence and transfer of function phenomena in syntax-like sequence processing, and to highlight the potential benefits of using connectionist models of verbal behavior.

Method

The Task Set

A connectionist task set or pattern set was designed to simulate experimental and control conditions of the training and testing procedure shown in Figure 1. The procedure presents sample-comparison items as they might be presented to a human subject in order to demonstrate the formation of equivalence classes, and a series of sequence and transfer of function tasks. The patterns presented to the connectionist network parallel these tasks. The patterns are labeled as p01-06 (conditional discrimination training), p07-09 (equivalence testing), p10-11 (sequence training), and p12-13 (testing for transfer of sequence function). The network was firstly trained on six training patterns (Figure 1, p01-p06) corresponding to a set of six conditional discriminations (A1-B1, A1-C1, A2-B2, A2-C2, A3-B3, A3-C3). For each of the six patterns, one of the A stimuli was presented as a sample and the network was trained to produce as output the appropriate B or C stimulus from three available comparisons (see Figure 1). The network was then trained to produce two sequences of responses (B1→B2→Ct1→B3 and B3→B2→Ct2→B1, i.e., p10 and p11). It was tested for the formation of three equivalence classes (A1-B1-C1, A2-B2-C2, A3-B3-C3) on patterns p07, 08, and 09 and was tested for the transfer of the sequence responses to the non-trained C stimuli (i.e. tested for the production of C1→C2→Ct1→C3 and C3→C2→

Ct2→C1) on p12 and p13. These patterns are presented in four stages in Figure 1, corresponding to the training and testing stages that a human subject might be exposed to. All training patterns were presented together

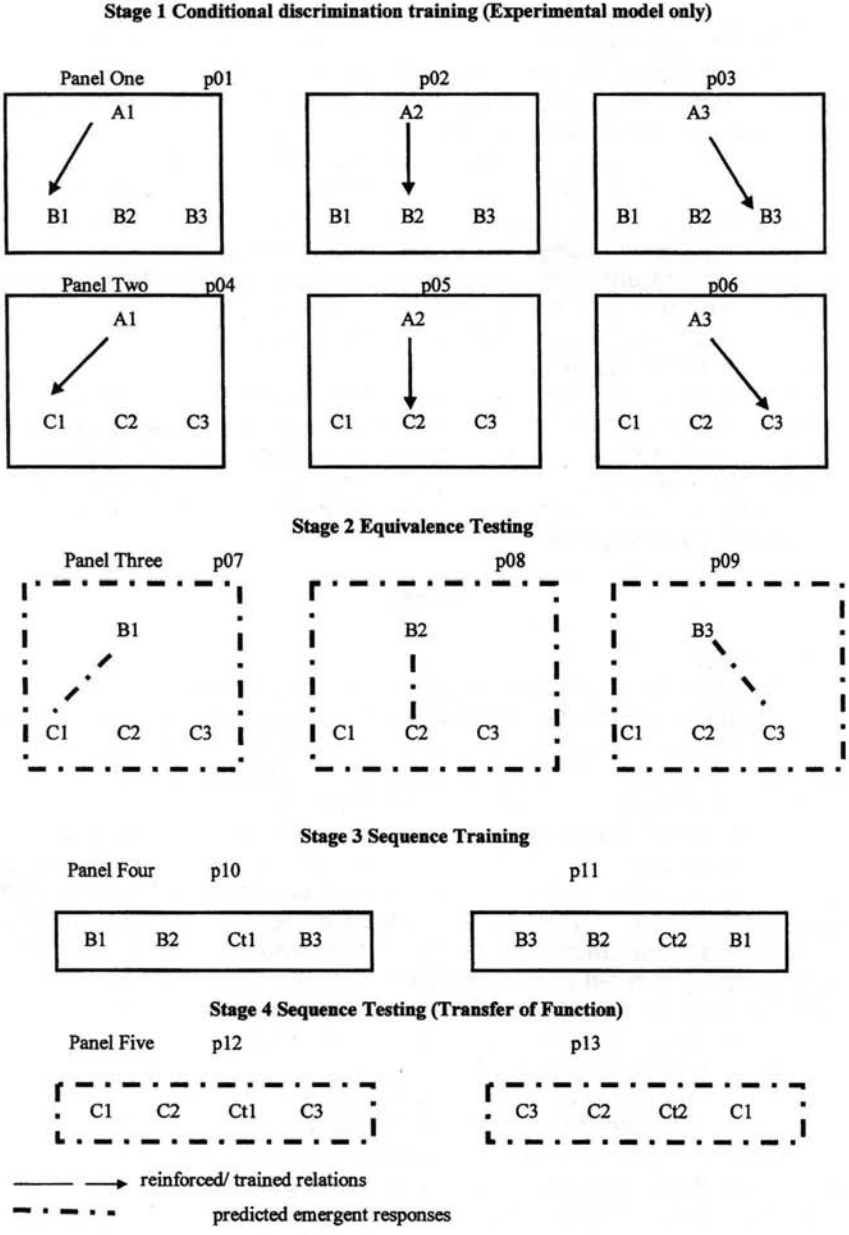

Figure 1. Procedure for training and testing stimulus relations.

and two test stages were administered, one to test for equivalence and another to test for transfer of the sequence responses.

Network Architecture

The network was constructed using software from McClelland and Rumelhart (1988), with 88 units arranged in three layers (see Figure 2). The architecture was based on the Barnes and Hampson (1993) model designed to simulate arbitrarily applicable relational responding. The input layer of 56 units included four sets of nine units representing four stimulus sets. The nine units of the fourth set correspond to the experimental stimuli (A1, A2, A3, B1, B2, B3, C1, C2, C3) shown in Figure 2. The previous three sets can be thought of as analogous sets of stimuli (such

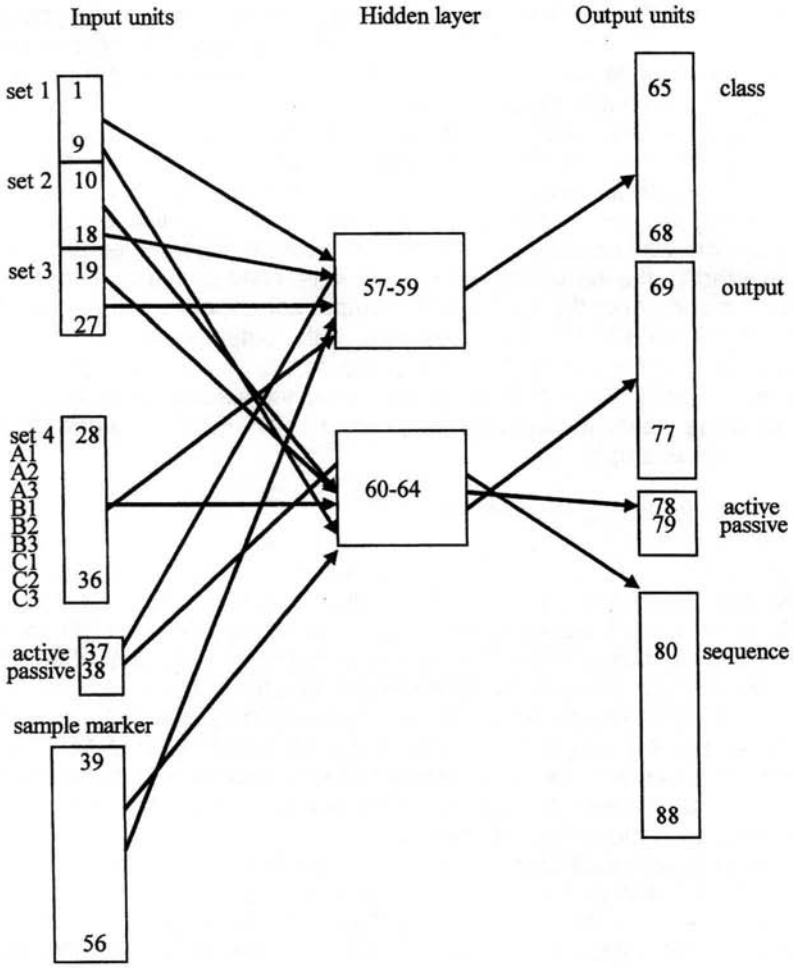

Figure 2. Network architecture.

as the set X1, X2, X3, Y1, Y2, Y3, Z1, Z2, Z3) used to pretrain the network on similar relations. Exposure to these sets was designed to simulate the preexperimental experience of equivalence relations that language-able subjects glean from their everyday exposure to natural language, without providing the network with information about the content of the to-be-tested set. The network was trained on all the patterns of these analogous sets, and on the fourth critical set it was exposed to the training patterns only, before performance on the test patterns was assessed.

Two input units flagged the input as being active (Ct1) or passive (Ct2) form, as contextual cues might be used to flag different sequence responses for human subjects. An additional 18 units made up the sample marker, which indicated whether a stimulus was a sample or a comparison item (for example, see Figure 1; p01 presented as input A1, B1, B2, and B3, where A1 was marked as a sample with three possible comparisons [B1, B2, and B3]). The input layer was fully connected to eight hidden units, which do the computational work in the network, in order for an output to be generated.

Three of the hidden units were connected to four output units denoting class of output or stimulus identity (i.e., these identified the output stimulus as from set 1, 2, 3, or 4, where sets 1 to 3 were the pretraining sets encountered in full and set 4 was the critical set on which the network was partially trained and then tested). Five hidden units were connected to the remaining 20 output units. Nine of these output units gave the content of the output (set 4 output corresponded to A1, A2, A3, B1, B2, B3, C1, C2, C3). Two units marked the output as active or passive type. The last nine units gave the sequence of the output. These units should produce output only during the sequence stages of the procedure, to allow the distinction between the sequence B1-B2-Ct1-B3 and B3-B2-Ct2-B1, for example.

Training and Testing Procedures

Training utilized the standard backpropagation of error learning algorithm (see Rumelhart, Hinton, & Williams, 1986). Backpropagation uses the discrepancy between the actual and target (goal) outputs to determine weight adjustments (thus modulating the strength of the connection between units). This is a supervised learning algorithm, as a teacher signal is present during training to flag the correct response. This is not always considered to be a psychologically plausible model of learning, but it is well suited to simulating the current type of experiment in which a subject would be presented with feedback designating the response as correct or incorrect. The teacher signal is used only for comparison in the simulator's test mode.

The patterns that were presented to the network corresponded to the training and test stimuli shown in Figure 1 and described above. The network patterns that most closely represent those given in Figure 1 belong to the critical fourth set, of which the nontrained relations were used to test the network. There were analogous patterns for three other

sets; set 1 patterns are referred to below as p1/01-1/13, set 2 as p2/01-2/13, and set 3 as p3/01-3/13. These sets corresponded in structure to the patterns in Figure 1, though their specific content differed. The pretraining simulated the preexperimental experience that human subjects bring to the experimental situation, and network performance was assessed as a function of the amount of pretraining received. The units of the network corresponded to the stimuli as shown in Figure 2, and input to the network in the form of a training or test pattern consisted of marking a unit as on (1) or off (0), depending on whether that unit was active or not in a given pattern. For example, in p01, the sample A1 was active as were the comparisons B1, B2, and B3, the output stating the class as 4 was active, and the output B1 as the target response was active (these units were given an activation of 1). Units corresponding to all other stimuli were not active (0), and so a string of 1 or 0 entries represented this particular pattern to the network. The same format was used to construct the other training patterns. The network was therefore gradually trained to produce a corresponding string of 1 or 0 outputs and by comparing the output to the target, its training performance was tracked. On testing, patterns were presented as for training, but in test mode no further modification of network activation took place (i.e., no further learning occurred).

Experimental and control conditions were simulated. The control model received no conditional discrimination training (that is, it was not exposed to patterns 01 to 06; see Figure 1). It would therefore be predicted to fail to show equivalence or a transfer of function. An additional training manipulation was also added. Previous models (e.g., Cullinan et al., 1994) have encountered problems of data loss when using methods other than batch mode of presentation. Under a batch training regime, training patterns are presented to the network at once (here, all three pretraining sets plus the critical fourth set would be presented together in a single batch). This is not very plausible in psychological terms; a comparison would be reexposing human subjects to equivalence relations across previously learned sets (from their preexperimental histories) whenever they were trained and tested on a new set. Clearly, this is not comparable to the typical equivalence experiment. However, when previously learned sets are not presented to a network, data may be lost from earlier learned sets. In effect, the network will abstract the structural information that is invariant across task sets, but may lose the specific stimulus domain content (presenting sets 1 then 4 would result in set 4 overwriting set 1). This is not entirely implausible as a model of human learning; subjects may fail to recall content specific information following the learning of new and similar information. However, compared to human information loss, the connectionist network's data loss is severe (this type of data loss has been called 'catastrophic interference'; see McCloskey & Cohen, 1989). For connectionist networks to be useful models of stimulus equivalence, plausibility needs to be maximized while the scale of data loss is reduced. A number of sequential training variants

are reported here in order to provide the most plausible simulation of the experimental procedures. The standard batch training regime, which presents all patterns to the network in a single training batch, is also reported for comparison.

1. Batch training method. During batch training conditions, all training sets were presented to the model together in a single batch. Several procedures were implemented, with varying amounts of pretraining, in order to assess the network's performance as a function of amount of experience with analogous sets. There were three levels of pretraining, though all manipulations involved presentation of only the training stages of the critical task set 4 (shown in Figure 1) and testing of the nontrained patterns from that set. Level 1 was the full training condition. The model was exposed to all of the patterns on sets 1, 2, and 3, and the training patterns of the critical set 4. The network was then tested on the test patterns of this fourth set. Level 2 was an intermediate condition, presenting sets 1 and 2 as well as the training patterns of set 4. Level 3 was the minimum pretraining set, exposing the model to set 1 with the training patterns of set 4. These three levels were implemented in an experimental and a control model, for comparison at each level of pretraining. Each manipulation was carried out five times using randomly generated initial weights each time, and the resulting data were averaged across the five runs.

It was predicted that network error would decrease as a function of amount of pretraining on analogous pattern sets. Following suitable pretraining, the experimental model was predicted to show formation of equivalence classes and a transfer of sequence function through these equivalence relations in the absence of further explicit training. The control model was trained on the sequence patterns, but not the conditional discrimination training patterns, on the critical fourth set. Both the control and experimental models were pretrained in full on up to three sets; the difference between the control and experimental models was whether or not the conditional discrimination training was given on the final fourth set. This corresponds to similar manipulations with human participants (see Cullinan et al., 1994). It was predicted that the control model would fail to demonstrate a transfer of sequence function on the test set, irrespective of the amount of pretraining. All training procedures required the network to reach an error criterion of .05 (producing < 5% error) on the training set before being tested on the nontrained patterns of the critical fourth set.

2. Sequential training methods. A series of sequential training methods was employed in order to more closely approximate human exposure to equivalence relations. Here, training sets were presented in stages, and the network was trained to produce less than 5% error before progressing to the subsequent stage. Although sequential training is more psychologically plausible than the batch model, it can be subject to retroactive interference, in that there may be overwriting of earlier training sets. For example, presenting set 1, then 2, then 3 and 4 may lead to the

loss of the specific content of the earlier sets (while domain invariant information may be retained). To overcome this effect, several additional training procedures were implemented which involved the rehearsal of old pattern sets as the new sets were presented. The sets of initial weights used in the batch regime were also used here under each training procedure, and the data were then averaged across five runs. A standard sequential procedure without incorporating rehearsal is also reported. The specific sequential training schedules were as follows.

(i) Standard sequential training. The network was exposed to sets 1 and 2, then sets 2 and 3, then set 3 and the training set 4. It was tested on the test patterns of set 4, and on sets 1 and 2 to assess the degree of interference.

(ii) Sequential training with complete (full set) rehearsal. The first of the sequential training methods with rehearsal involved the presentation of sets 1 and 2 first. Then the network was exposed to sets 2 and 3, as well as one third the amount of total exposure to sets 1 and 2 (i.e., Sets 2 and 3: Set 1 = 2:1). The network was then exposed to sets 3 and 4 as well as one third the total exposure to sets 1 and 2. In this way old learning was rehearsed while priority was given to the new patterns. The nontrained stages of set 4, and the complete sets 1 and 2 were tested.

(iii) Sequential training with restricted (full set) rehearsal. The second rehearsal procedure presented pattern sets as in training method (i), however the ratio was further reduced to 4:1 from 2:1. Therefore, old learning accounted for one fifth of all training under this new regime. Results from this method were to be compared with those from method (ii), and therefore the same test sets were used.

(iv) Sequential training with restricted (selected pattern) rehearsal. This method involved selecting eight representative patterns from the previously encountered pattern sets, rather than re-presenting the full sets. In this way, representative training patterns from the previously encountered sets were rehearsed while priority was given to new information. Of the rehearsed patterns, each individual pattern represented a training or testing configuration (see Figure 1) from the total of 15 patterns in each set, with patterns selected as representative of the panels. The network was firstly trained on sets 1 and 2. Then the network was given sets 2 and 3 four times for every once it was exposed to the selected patterns from the previously learned set 1 (patterns 1/01, 1/03, 1/04, 1/06, 1/07, 1/09, 1/10, 1/13). Following completion of training on these patterns, the network was then exposed to sets 3 and 4 four times for every one exposure to the patterns from sets 1 and 2 (patterns 1/02, 1/03, 1/04, 1/05, 1/08, 1/09, 1/11, 1/12 from set 1 and patterns 2/01, 2/03, 2/04, 2/06, 2/07, 2/09, 2/10, 2/13 from set 2). The network was tested on the unrehearsed patterns from this last exposure to sets 1 and 2 as well as the critical fourth set test.

(v) Comparison method for restricted (selected pattern) rehearsal. The purpose of training method (v) was purely to provide a comparison for method (iv) training. The network was trained on the last stage only of method (iv). That is, the network was given sets 3 and 4, four times for

every one exposure to the patterns from sets 1 and 2 (patterns 1/02, 1/03, 1/04, 1/05, 1/08, 1/09, 1/11, 1/12 from set 1 and patterns 2/01, 2/03, 2/04, 2/06, 2/07, 2/09, 2/10, 2/13 from set 2). Network performance was tested as in method (iv) and compared to the training method (iv) results in order to ensure that previous learning was having an effect, rather than just the last pattern sweep.

(vi) Sequential training with minimal (selected pattern) rehearsal. This method further restricted the amount of previously encountered patterns that were rehearsed. Only one pattern from each training panel (see Figure 1) was presented. The network was trained on sets 1 and 2. Following convergence to criterion, the network was then exposed to sets 2 and 3 four times for every one exposure to set 1 patterns 1/1, 1/4, 1/9, 1/10, and 1/13. The network was then further exposed to sets 3 and 4, four times for every one exposure to the selected patterns (from set 1, the patterns 1/2, 1/5, 1/8, 1/11, and 1/12, and from set 2, the patterns 2/01, 2/04, 2/09, 2/10, 2/13). Following training, the network was tested on the test patterns from set 4, as well as the patterns from sets 1 and 2 not presented in the final sweep (i.e., patterns 1/01, 1/03, 1/04, 1/06, 1/07, 1/09, 1/10, 1/13, and 2/02, 2/03, 2/05, 2/06, 2/07, 2/08, 2/11, and 2/12).

Network performance was compared across these training regimes. Following training to criterion performance on a regime, the network was

Table 1

Summary of Training Schedules and Test Patterns

Training Method	Training Sequence	Test Patterns
Batch method	1, 2, 3, 4*	Sets 4**, 1, 2
(i) Standard sequential training	1+2 2+3 3+4*	Sets 4**, 1, 2
(ii) Sequential training with complete (full set) rehearsal	1+2 (2+3)+(1) = 2:1 (3+4*)+(1+2) = 2:1	Sets 4**, 1, 2
(iii) Sequential training with restricted (full set) rehearsal	1+2 (2+3)+(1) = 4:1 (3+4*)+(1+2) = 4:1	Sets 4**, 1, 2
(iv) Sequential training with restricted (selected pattern) rehearsal	1+2 (2+3)+(1/01, 1/03, 1/04, 1/06, 1/07, 1/09, 1/10, 1/13) = 4:1 (3+4*)+(1/02, 1/03, 1/04, 1/05, 1/08, 1/09, 1/11, 1/12, 2/01, 2/03, 2/04, 2/06, 2/07, 2/09, 2/10, 2/13) = 4:1	Set 4** 1/01, 1/06, 1/07, 1/10, 1/13 2/02, 2/05, 2/08, 2/11, 2/12
(v) Comparison method for restricted (selected pattern) rehearsal	(3+4*)+(1/02, 1/03, 1/04, 1/05, 1/08, 1/09, 1/11, 1/12, 2/01, 2/03, 2/04, 2/06, 2/07, 2/09, 2/10, 2/13) = 4:1	Set 4** 1/01, 1/06, 1/07, 1/10, 1/13 2/02, 2/05, 2/08, 2/11, 2/12
(vi) Sequential training with minimal (selected pattern) rehearsal	1+2 (2+3)+(1/01,1/04, 1/09, 1/10, 1/13) = 4:1 (3+4*)+(1/2, 1/5, 1/8, 1/11, 1/12, 2/01, 2/04, 2/09, 2/10, 2/13) = 4:1	Set 4** 1/01, 1/03, 1/04, 1/06, 1/07, 1/09, 1/10, 1/13 2/02, 2/03, 2/05, 2/06, 2/07, 2/08, 2/11, 2/12

* training stages only, ** test stages only.

tested on the nontrained patterns of the fourth training set (that is, the test for equivalence and the test for transfer of the sequence function, see Figure 1). The training schedules and the corresponding tests of network performance are summarized in Table 1.

Results

Network performance was measured as total sum of squared error (TSS) scores and as discrete error scores (DES). The TSS is a measure of the total error across the pattern set (i.e., the accumulated disparity between the actual and target outputs). This measure of global error provides an overall representation of network performance but may not reflect differences in performance on individual patterns. A network with small errors consistently accumulating across pattern sets might have a similar TSS to a network with poor performance on just one individual pattern. In the first case, the network may have learned the patterns reasonably accurately; in the latter, failure to acquire one of the patterns might be taken to suggest that the network had failed to complete the task. Furthermore, because TSS accumulate over a number of patterns in a set, larger pattern sets tend to produce higher TSS than smaller sets. Therefore, an additional measure was used. The DES measures error at the level of the pattern and here consists of two types of discrete error. Noise error was defined as activation of ≥20% of a unit where the target activation is 0. Under-activation errors were defined as failure of a unit to achieve activation ≥80% where the target activation is 1. Thus DES = 3 would indicate that three units in the test sweep of a particular pattern produced an error.

1. Batch training results. The results for the batch training procedures are shown in Tables 2 and 3. The TSS and DES have been averaged across five runs.

The equivalence test performance (see Table 2) shows that the experimental model improved as a function of amount of pretraining, as

Table 2

Mean Batch Training Performance on Equivalence Test

Condition	Experimental		Control	
	TSS	DES	TSS	DES
Level 1 Full training	.0412	0	2.574	4.4
Level 2 Intermediate training	.7338	.2	3.052	6.6
Level 3 Minimum training	.12534	.6	5.296	11.4

predicted. The control model, which received no conditional discrimination training on set 4, failed to demonstrate equivalence. There are consistent differences between experimental and control conditions across stages.

On the sequence test (see Table 3), the experimental model's error decreases as a function of amount of pretraining, with the fully trained model yielding a TSS of .00536, with no DES. The fully trained control model

Table 3

Mean Batch Training Performance on Sequence Test

Conditiol	Experimental		Control	
	TSS	DES	TSS	DES
Level 1 Full training	.00536	0	2.469	4.8
Level 2 Intermediate training	.0081	0	3.11084	5.8
Level 3 Minimum training	.13578	.4	4.0218	5.2

produced an average of 4.8 errors and failed to improve significantly as a function of training. The control model consistently produced higher DES and TSS than the experimental model at each training level.

Overall, the results of the batch training conditions show a general decrease in error as a function of amount of pretraining, with the experimental model producing low error overall (see Figure 3). Given appropriate training, the experimental model demonstrated the formation of equivalence classes and a transfer of sequence function through equivalence. By contrast, and as predicted, the control model failed to show this effect.

2. Sequence training results.

(i) Standard sequential training. The network was trained gradually by

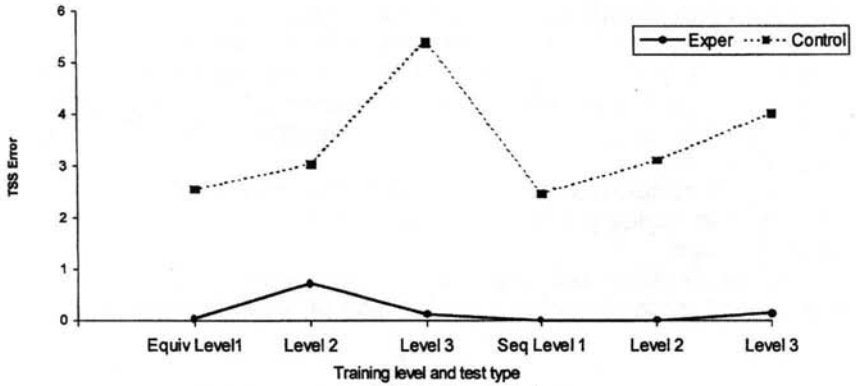

Figure 3. Summary of batch training performance by amount of pretraining and test type.

the stages outlined above. After training to criterion on all three stages, the network was tested on the patterns from set 4 that had not been previously trained, and it yielded a TSS error score of .01974 when averaged across the five runs. No DES were produced. Whereas performance on set 3 was also good (as it had been encountered in full during the last sweep), analysis of sets 1 and 2 showed interference by later learning. Testing on the full pattern set 1 (across 13 patterns) gave a mean TSS of 21.87, with DES at 28.6. The network produced errors of stimulus class or domain, giving the class output as sets 3 and 4 (the last stage encountered). Set 2 testing gave a TSS of 12.1, with DES at 24.8 (the TSS for set 2 is lower than set 1 because the network has been

exposed to set 2 on more of the sequential stages). Errors of class were apparent, with set 4 output here. In some cases, the network succeeded in producing the correct class 2 response at above 50% activation, but it continued to produce a class 4 response at a higher level. Performance on sets 1 and 2 was not derived; the model had been exposed to these sets in full and was therefore being tested for retention of those items. However output was still well below criterion. While the domain invariant information was retained, the domain specific aspects of the old information was lost, as new information was presented to the network.

(ii) Sequential training with complete (full set) rehearsal. The model was tested on both the equivalence and sequence tests of set 4, and was also tested on sets 1 and 2. It averaged a TSS of .01754 when tested on the critical set 4, and a TSS of .0302 over the entire sets 1 and 2. No DES were recorded. The model therefore retained the data from earlier sets as well as learning the new information.

(iii) Sequential training with restricted (full set) rehearsal. The network was exposed to patterns as in method (ii) above, with a ratio of 4:1. A TSS of .0222 was produced on set 4, and of .02962 on sets 1 and 2. There was no significant difference in performance between training in methods (ii) and (iii), on either test set 4, t (4) = 1.248, p = .28, or sets 1 and 2, t = .703, p = .5. Accurate output was therefore produced after only one fifth of training involving rehearsal.

(iv) Sequential training with restricted (selected pattern) rehearsal. Testing on set 4, the mean TSS was .028 with a mean DES of 0.2 (i.e., one error occurred on one of the five runs). Testing on the patterns from set 1 gave a mean TSS of .00928 (no DES), whereas testing on the patterns from set 2 gave a TSS of .00986 (also no DES). The model therefore performed well on the critical set 4, as well as retaining the previously learned patterns on sets 1 and 2, having been reexposed to only eight patterns from sets 1 and 2.

(v) Comparison method for restricted (selected pattern) rehearsal. For comparison to method (iv), the network was trained only on the final stage of that method, in order to ensure that previous stages were affecting the final test result. This time testing on set 4 gave a TSS of .1523 and a DES of .8. This was not found to be significantly different from method (iv). Testing on the selected patterns from set 1 gave a TSS of .05966 (DES = 0.2). This is a significantly higher error score than in training method (iii), t = 3.2441, $p \leq .05$. Testing on set 2 yielded a TSS of .42324 (DES 1.2) which is again a significantly higher error than training method (iii), t = 2.14665, $p \leq .05$. There was also a significant difference between DES scores for these two conditions, t = 2.057983, $p \leq .05$. Therefore, although there was no significant difference when testing on the novel patterns from set 4, the network performed better on sets 1 and 2 in training method (iv). It can be concluded that the network retained previous learning due to rehearsal and that this effect was not merely due to generalization across a pattern set [if that were the case, there should be no significant difference in performance on methods (iv) and (v)].

(vi) Sequential training with minimal (selected pattern) rehearsal. The training regime of method (iv) was further restricted such that less than 40% of the training patterns were rehearsed. The model was then exposed to three test phases. The first consisted of the test patterns from set 4, on which the model produced TSS of .01762. It was then tested on the patterns omitted from set 1 in the final training sweep, and it produced a TSS of .02822. Testing on the set 2 omitted patterns produced a TSS of .02486. The network performed well on set 4 test, while also producing an adequate performance on sets 1 and 2. No DES were produced. More error was produced than on method (iv), but it remained well below criterion.

As these data are not all directly comparable because of the differing test sizes, the TSS scores were adjusted for comparison across the procedures, by accounting for the varying test set sizes. These data are

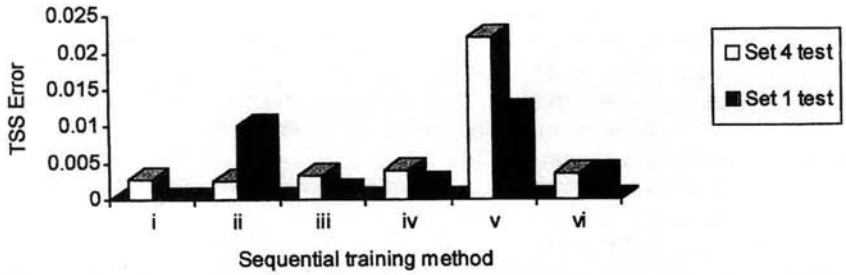

Figure 4. Sequential training performance on sets 1 and 4 with data adjusted for pattern set size.

shown in Figure 4. Error from the standard procedure on set 1 testing (in which large-scale interference took place) went off the scale and was therefore not included here (at a mean TSS error of 1.458). These adjusted figures give a clearer indication of how the TSS scores change with training. The disparity between performance on methods (iv) and (v) is apparent, and it shows that earlier presented sets have an effect on network output. Method (vi) performance was low in error on both sets 1 and 4 despite the limited training on the old sets. These data show that relatively small amounts of rehearsed information presented with new patterns may overcome the interference effect. Even if performance on set 4 was slightly poorer here, correct output was more equally distributed across all sets to which the network had been exposed, and criterion performance was reached at all stages.

The sequential training regime could therefore be improved by presenting some of the previously learned sets' individual patterns along with the new information. This was found to improve performance significantly, even with rehearsal of a relatively small proportion of old patterns (e.g., 38.5% of a set as one fifth of the entire information set). Method (vi) showed that further limiting the training of old patterns to just one pattern from each training panel did not significantly impair performance; the network retained the old information, as well as learning the new. When rehearsal was used, previously learned patterns were not

completely overwritten; method (v) produced significantly higher error than method (iv), showing that the exposure to the final stage alone was not sufficient for successful performance. The basis of rehearsing old learning is consistent with psychological findings, therefore adding to the plausibility of the sequential training method.

Discussion

The connectionist network demonstrated the formation of three 3-member equivalence classes and a transfer of sequence function following appropriate training. A control model, which did not receive conditional discrimination training, failed to show this effect. Although the control model had been exposed to equivalence and sequence tasks, equivalence training on the critical fourth set was also needed in order to perform well on the novel task. These results could be predicted from experiments with human subjects. In fact, the experimental model behaved very much as one might expect a human subject to behave on exposure to the experimental stimuli, suggesting that connectionist networks may provide useful and plausible models of complex human behavior.

In the experimental model, performance was seen to improve as a function of the amount of pretraining provided on analogous training sets, which may parallel the role of developmental experience in the emergence of equivalence and transfer of function in humans (Lipkens, Hayes, & Hayes, 1993). The control afforded by this manipulation of information input to the system has great potential for exploring issues that are not easily examined with human subjects, such as the emergence of equivalence in the early stages of language acquisition. The network's exposure to equivalence relations can be tracked precisely, and the experimental emergence of equivalence could be explored as a function of 'preexperimental' experience in a controlled model.

The sequential training trials presented here incorporated aspects of earlier learning as a form of rehearsal for the network. As new patterns were presented, earlier patterns were also rehearsed, with the emphasis on the new input. The inclusion of some of the previously learned patterns along with the new data allowed retention of the old learning, as well as acquisition of the new, even when the amount of exposure to the earlier learning was significantly reduced. Such rehearsal methods greatly increase the plausibility of the simulation, paralleling the sequential manner in which knowledge of equivalence and transfer of function responses presumably accumulates preexperimentally in humans. In addition, because of the nature of the task being simulated, the supervised learning regime used in the network provided a plausible simulation of the human experiments in this area, and it suggested that connectionist networks might be usefully employed alongside such experiments (as in the Cullinan et al., 1994, study) to investigate more complex aspects of verbal behavior.

The results also show that, in a connectionist network, a transfer of function through equivalence can account for the types of response

sequences that may underlie active and passive voice sentences. Once equivalence classes of lexical categories have formed, their sequence functions in a sentence may transfer through equivalence to enable the generation of novel sequences. Although the pattern sets presented to the network are merely analogies of sequences in natural language and do not reflect the true complexity of human syntax, nevertheless they may provide insight into the productivity of sequential responses, by modeling transfers of sequence function via equivalence.

The current data speak directly to some of the key issues currently being debated in the area of derived stimulus relations. Some researchers have argued that multiple-exemplar training may be one of the fundamentally important processes involved in the acquisition of derived relational responding and language skills more generally (e.g., Hayes, 1994). Others, however, have suggested that the concept of multiple-exemplar training is rather simplistic and can not function as an explanation for derived relational responding (e.g., Lowe & Horne, 1996; Sidman, 1994). The various manipulations that were undertaken in the context of the current connectionist network, however, suggest that the process of multiple-exemplar training is far from simplistic. For example, under the sequential training manipulations, five different levels of rehearsal were employed in the current study, and in general these proved to be functionally significant for the network. More specifically, we found that even relatively limited exposure to 'old learning' embedded in 'new learning' significantly improved the overall performance of the network. Thus the extent and the design of the multiple-exemplar training appears to be an important variable for generating derived relational responding in a connectionist network. Whether this also proves to be the case for human or even nonhuman populations remains to be seen (cf. Schusterman & Kastak, 1993). In any case, the current study indicates that the topic of multiple-exemplar training may be an important one, in connectionist, human, and nonhuman research.

The tasks simulated here allowed the network to be tested on a highly specific and well-defined yet plausible behavioral task. In this context the model provided a useful and psychologically plausible simulation of stimulus equivalence and transfer of function phenomena. The results suggest that using connectionist networks and behavior analytic tasks together has potential for advancing the contribution of both approaches to the study of human language and cognition.

References

BARNES, D., & HAMPSON, P. J. (1993). Stimulus equivalence and connectionism: Implications for behavior analysis and cognitive science. *The Psychological Record*, 43, 617-638.

BARNES-HOLMES, Y., BARNES-HOLMES, D., ROCHE, B., HEALY, O., LYDDY, F., CULLINAN, V., & HAYES, S. C. (2001). Psychological development. In S. C. Hayes, D. Barnes-Holmes, & B. Roche (Eds.), *Relational frame theory: A post-Skinnerian account of human language and cognition.* New York: Plenum Press.

CULLINAN, V., BARNES, D., HAMPSON, P. J., & LYDDY, F. (1994). A transfer of explicitly and non-explicitly trained sequence responses through equivalence relations: An experimental demonstration and connectionist model. *The Psychological Record*, 44, 559-586.

DOUGHER, M. J., & MARKHAM, M. R. (1994). Stimulus equivalence, functional equivalence and the transfer of function. In S. C. Hayes, L. J. Hayes, M. Soto, & K. Ono (Eds.), *Behavior analysis of language and cognition.* Reno, NV: Context Press.

FODOR, J. A., BEVER, T. G., & GARRETT, M. F. (1974). *The psychology of language.* New York: McGraw Hill.

GARRETT, M. F. (1990). Sentence processing. In D. N. Osherson & H. Lasnik (Eds.), *An invitation to cognitive science.* Cambridge, MA: MIT Press.

HAYES, S. C. (1991). A relational control theory of stimulus equivalence. In L. J. Hayes & P. N. Chase (Eds.), *Dialogues on verbal behavior.* Reno, NV: Context Press.

HAYES, S. C. (1994). Relational Frame Theory: A functional approach to verbal events. In S. C. Hayes, L. J. Hayes, M. Soto, & K. Ono (Eds.), *Behavior analysis of language and cognition.* Reno, NV: Context Press.

LIPKENS, R., HAYES, S. C., & HAYES, L. J. (1993). Longitudinal study of derived stimulus relations in an infant. *Journal of Experimental Child Psychology*, 56, 201-239.

LOWE, C. F., & HORNE, P. J. (1996). Reflections on naming and other symbolic behavior. *Journal of the Experimental Analysis of Behavior*, 65, 315-340.

MASSARO, D. W. (1988). Some criticisms of connectionist models of human performance. *Journal of Memory and Language*, 27, 213-234.

MCCLELLAND, J., & RUMELHART, D. (1988). *Explorations in parallel distributed processing: A handbook of models, programs, and exercises.* Cambridge, MA: MIT Press.

MCCLOSKEY, M., & COHEN, N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. *The Psychology of Learning and Motivation*, 24, 109-165.

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

PLUNKETT, K., & MARCHMAN, V. (1993). From rote learning to system building: Acquiring verb morphology in children. *Cognition*, 48, 21-69.

RUMELHART, D. E., HINTON, G., & WILLIAMS, R. (1986). Learning internal representations by back-propagating errors. *Nature*, 323, 533-536.

RUMELHART, D. E., & MCCLELLAND, J. L. (1986). On learning the past tense of English verbs. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing, Volume 2: Psychological and biological models.* Cambridge, MA: MIT Press.

SCHUSTERMAN, R. J., & KASTAK, D. (1993). A California sea lion (*Zalophus californianus*) is capable of forming equivalence relations. *The Psychological Record*, 43, 823-840.

SIDMAN, M. (1994). *Equivalence relations and behavior: A research story.* Boston, MA: Authors Cooperative, Inc., Publishers.

SLOBIN, D. I. (1966). Grammatical transformations and sentence comprehension in childhood and adulthood. *Journal of Verbal Learning and Verbal Behavior*, 5, 219-227.