Modeling Interference Effects In Instructed Category Learning

David C. Noelle and **Garrison W. Cottrell**
Computer Science & Engineering 0114
University of California, San Diego
La Jolla, CA 92093-0114
{dnoelle,gary}@cs.ucsd.edu

Abstract

Category learning is often seen as a process of inductive generalization from a set of class-labeled exemplars. Human learners, however, often receive direct instruction concerning the structure of a category before being presented with examples. Such explicit knowledge may often be smoothly integrated with knowledge garnered by exposure to instances, but some interference effects have been observed. Specifically, errors in instructed rule following may sometimes arise after the repeated presentation of correctly labeled exemplars. Despite perfect consistency between instance labels and the provided rule, such inductive training can drive categorization behavior away from rule following and towards a more prototype-based or instance-based pattern. In this paper we present a general connectionist model of instructed category learning which captures this kind of interference effect. We model instruction as a sequence of inputs to a network which transforms such advice into a modulating force on classification behavior. Exemplar-based learning is modeled in the usual way: as weight modification via backpropagation. The proposed architecture allows these two sources of information to interact in a psychologically plausible manner. Simulation results are provided on a simple instructed category learning task, and these results are compared with human performance on the same task.

Introduction

Investigations into concept formation have often focused on the learning of category structure solely through exposure to labeled exemplars. Indeed, much of the success of connectionist learning models may be attributed to their ability to perform exactly this sort of statistical induction. Human learners often need not rely solely on exemplars, however, to formulate an understanding of a concept. The presence of language allows us to learn from the direct instruction provided by others. Such explicit advice may simply direct our attention to relevant features, or it may actually spell out necessary and/or sufficient conditions for membership in a category. Learning "by being told" may facilitate an otherwise exemplar-based learning task, enabling a multistrategy approach integrating induction and instruction.

Some forms of advice may be seen as explicitly providing categorization rules to the learner. When viewed in this way, the integration of exemplar-based induction and direct instruction begins to resemble another form of integration discussed in the category learning literature. Sometimes human subjects presented with labeled exemplars appear to induce explicit classification rules, and sometimes induced category structures are better captured by prototype-based or instance-based representations. Much evidence has been gathered suggesting that these two kinds of category representation result from two dissociable learning mechanisms (Shanks & St. John, 1994), so a question arises as to how these two processes are integrated in common learning tasks. In terms of instructed category learning, this question becomes one of how categorization rules provided by direct instruction interact with exemplar-based knowledge to form a learned category structure.

This work focuses on one particular interference phenomenon which has manifested itself in instructed category learning studies (Allen & Brooks, 1991; Nosofsky et al., 1989). Specifically, subjects initially provided with an explicit classification rule may begin to violate that rule after the presentation of correctly labeled exemplars. While none of the presented examples contradict the instructed rule in any way, similarity between instances drives subject behavior away from rule following and towards a more prototype-based or instance-based pattern. It appears, in these experiments, as if an exemplar-based inductive learning process is directly interfering with an instruction-based rule application process.

The goal of this paper is to demonstrate that our previously proposed connectionist model of instructed learning (Noelle & Cottrell, 1995) may capture and account for this interference effect. To this end, we discuss one psychological experiment in which this phenomenon has appeared, review our modeling framework for instructed learning, and present the results of applying our model to the discussed experimental domain.

An Interference Effect

In the category learning literature a debate has raged over the internal representation of inductively learned categories. Some view these as simple sentential rules which are acquired in response to experience. Others view categories as regions in some feature space, with region boundaries determined by some similarity measure and a collection of remembered exemplars or prototypes. In order to test these two competing views, Nosofsky, Clark, and Shin (1989) designed and conducted a number of elegant experiments. They used a set of stimuli with two pertinent continuous features – the size of a circle and the angle of rotation of a radial line. These objects may be plotted as points in a two dimensional feature space, as in Figure 1. In that diagram, each letter corresponds to a potential stimulus object, and the letters enclosed by small polygons are objects which were presented as labeled training exemplars. The two enclosing shapes, triangles and squares, correspond to category labels for two disjoint categories. We

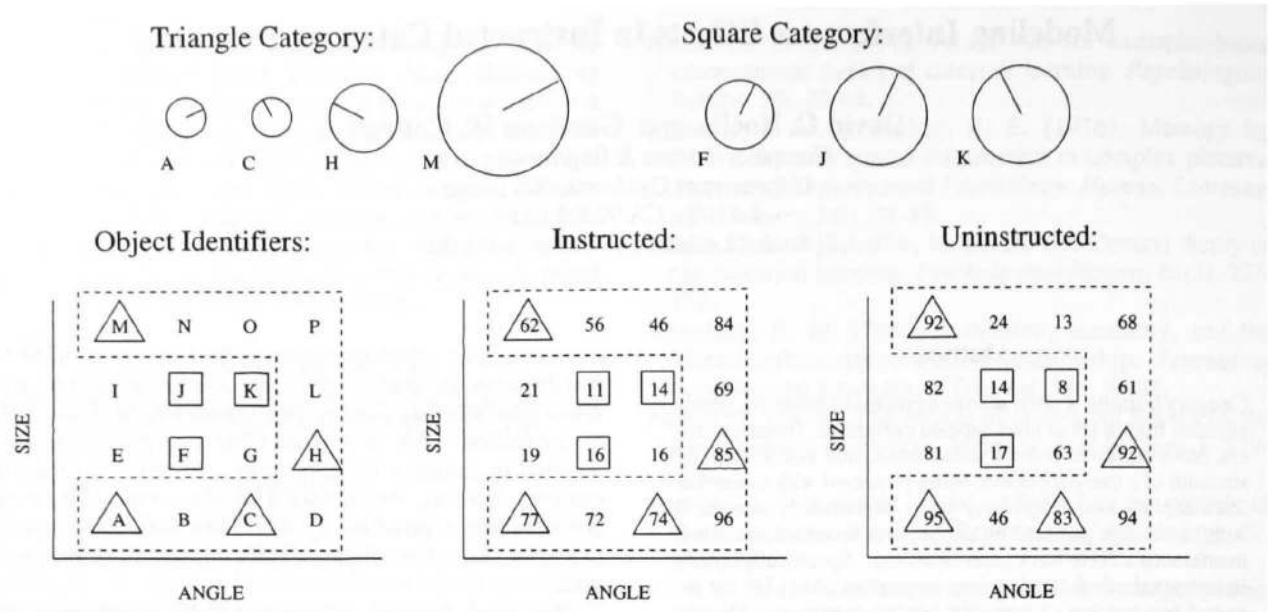

Figure 1: Two Categories: Training Exemplars, Feature Space, And Frequency Of Classification Into The "Triangle" Category For Both Instructed And Uninstructed Subjects

may imagine that distance in this feature space corresponds to perceived "similarity" between objects. How should object "P" be classified? If similarity is used to define categories, object "P" will be placed in the "square" category, due to its close proximity to exemplar "K". If some form of minimum description length sentential rule is used instead (e.g., a rule for the triangle class is "`tiny or huge or` 155° `rotation`" – shown as a dotted box), then this object will be labeled as a "triangle". Thus, classification behavior on novel objects can inform explorations into the structure of internal category representations.

In one experiment, Nosofsky and his colleagues compared subjects who were explicitly told a sentential categorization rule before being exposed to the training exemplars with subjects who depended solely on the exemplars to formulate category boundaries. In general, the explicitly instructed subjects exhibited rule governed classification behavior, whereas the uninstructed subjects matched a similarity-based model. However, instructed subjects sometimes *deviated* from their rule-based behavior when classifying objects highly similar to training exemplars from the opposite category.

The situation depicted in Figure 1 produced the most striking results. Subjects were instructed to classify objects as being members of the "triangle" category if and only if they fit the given disjunctive rule. Following this instruction, the subjects received 300 random presentations of the seven training exemplars, and they were asked to classify each of them. After each selection, the subjects were told the true category of the training exemplar, and the next object was presented. Upon completion of this training period, the subjects were tested on the entire collection of 16 objects. The final mean frequencies of classifying objects as members of the "triangle" category are displayed as percentages in Figure 1, along with the same frequencies for subjects who received *no* explicit classification rule before being presented with labeled exemplars. Of particular note here is object "O", which instructed subjects tended to place in the "square" category more often than not, despite the fact that such classification violated the given rule. Note the low frequency with which *uninstructed* subjects identified this object as a member of the "triangle" category. It would seem that the same inductive process which caused this response in the uninstructed subjects has interfered with the rule processing of the instructed subjects. Our goal is to show that an interference effect of this type may be explained by our connectionist model.

A Connectionist Model

The modeling approach we advocate here is based on our connectionist model of learning "by being told" (Noelle & Cottrell, 1995). This model arises from the recognition that the weight update techniques typically used for inductive learning in artificial neural networks are simply too slow to account for the high speed of behavior change which occurs in response to direct instruction. Activation propagation in such networks, on the other hand, is quite fast. We suggest that instructed learning is properly seen as a process in which presented advice *pushes* the activation state of part of the cognitive system into a novel basin of attraction – a stable region of activation space which encodes the proper operationalization of the given advice. Such novel attractors, corresponding to newly received instructions, come into existence through the componential interaction of basins sculpted via past experience with the instructional language (Plaut & McClelland, 1993). Under this view, advice is seen as a sequence of input activity, presented to a network which transforms such sequences into appropriate behavior.

Our proposed general architecture is shown in Figure 2, on the left. The boxes in that diagram represent layers of sigmoidal processing elements and arrows represent complete interconnections between layers. Categorization rules are en-

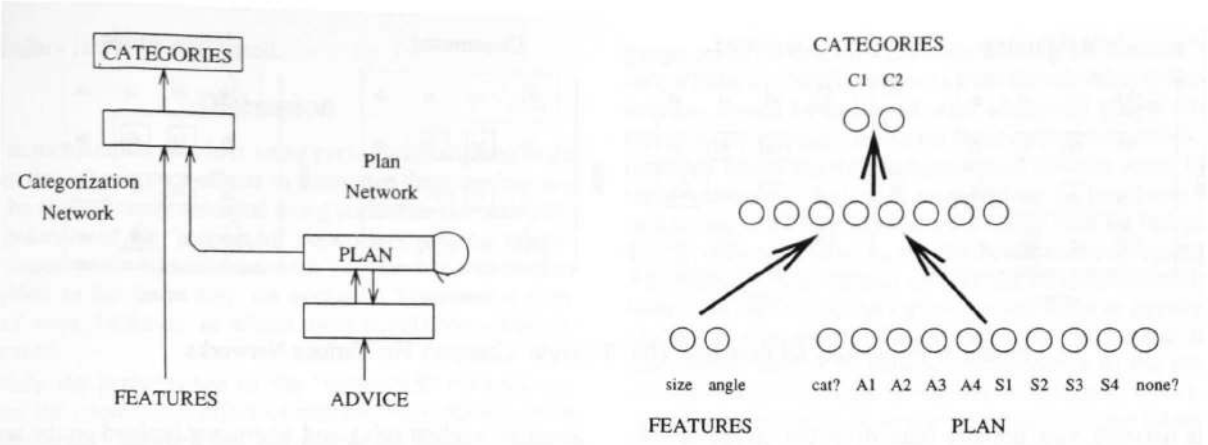

Figure 2: Instructable Network: General Architecture & Isolated Categorization Network

coded as sequences of instruction tokens which are presented, one at a time, at the *advice* input layer. Activation is then propagated through the recurrent "Plan Network" to produce a stable pattern of activation at the *plan* layer. The *plan* is then used to modulate the behavior of a simple feed-forward "Categorization Network", causing the mapping from stimulus *features* to output *categories* to exactly match the rule specified by the input advice.

Initially, this network must be inductively trained to understand the language of instruction. This may be accomplished using a version of *backpropagation through time* (BPTT) (Rumelhart et al., 1986) in which error is backpropagated for only a single time step, much as is done for Simple Recurrent Networks (Elman, 1990). A training regimen similar to that used by the architecturally similar *Sentence Gestalt* network (St. John & McClelland, 1990) may be used, requiring an external error signal only at the final output layer. The output units encode categorization judgements for specific input stimuli in the context of instructions presented at the *advice* layer. By backpropagating error from the final category outputs in this way, the internal representation of instruction sequences, maintained at the *plan* layer, may be learned in the service of the categorization task (St. John, 1992). Once the network has learned to represent advice at the *plan* layer, no further weight modifications are needed to exhibit *immediate* behavior change in response to instruction.

This is the point at which induction and instruction become integrated in this model. While further inductive weight updates are not *needed* in order to exhibit instructed behavior, such inductive learning may commence nonetheless. Indeed, inductive weight modification in the "Categorization Network" provides a means for rule following and exemplar-based induction to interact. Modulating activation from the *plan* layer will bias the network to act in accordance with instructions, but further exemplar-based error feedback may modify weights so as to violate the instructed rule.

While the initial learning of an instructional language is an important component of our general model, the issue of interference effects between exemplar-based induction and instructed rule following is relatively orthogonal to how the instructional language is acquired. In order to focus, then, on interaction effects, we have modeled the instructed category learning task using the "Categorization Network" alone. We have fabricated an arbitrary representational format for the *plan* layer, allowing us to replace the "Plan Network" with the direct presentation of encoded categorization rules to the "Categorization Network". Some training in the instructional language is still necessary for the network to discover the "meaning" of our *plan* layer encoding, but this initial preparation is much more rapid than when the recurrent "Plan Network" must be trained simultaneously. Still, we assume that something similar to our *plan* layer representation could be generated by the "Plan Network" from linguistic input, given sufficient training.

The resulting categorization network is shown in Figure 2, on the right. Disjunctive categorization rules were encoded into a ten element *plan* using a bit-vector representation for the set of rule terms. The first *plan* unit was used to indicate if the given disjunctive rule was to describe the members of the "square" category or of the "triangle" category. The next eight units were used to encode the actual rule, with each unit corresponding to one of the four levels of "size" or to one of the four levels of "angle". If a given "size" unit was turned on, this implied that the rule covered all stimuli of that size, and a similar code was used over the "angle" units. Activating multiple units produced disjunctive rules, so the rule in Figure 1 required the activation of the three units: "size = tiny", "size = huge", and "angle = 155°". The last unit in the *plan* vector was used to signal the *absence* of any rule – the uninstructed case. When no rule was available, this last unit was turned on and the activity of the other nine *plan* units was set to a medial value (i.e., 0.5). Unlike the quasi-binary *plan* layer encoding, the two features of observed stimuli, size and angle, were presented to the network in a continuous fashion. One input unit was available for each of the features, and each unit could take on one of four ranked values: 0, $\frac{1}{3}$, $\frac{2}{3}$, or 1. The network possessed two output units – one for each category. The activation levels at these outputs where normalized into conditional class probabilities using Luce ratios (Luce, 1963). The hidden layer consisted of eight processing elements. The result was a network which produced probabilistic categorization judgements from two continuous features and an encoding of an instructed classification rule (or the absence of such a rule).

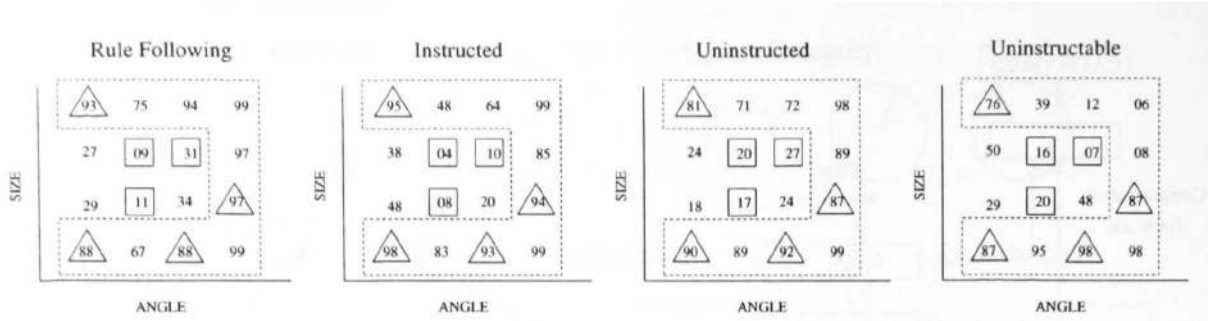

Figure 3: Probability Of Selecting The "Triangle" Category For Various Networks

This network was initially trained on the complete collection of possible disjunctive classification rules, in order to encourage the proper understanding of the rule encoding format. The eight inputs which corresponded to rule terms allowed for $2^8 = 256$ possible disjunctive clauses, and each of these could be used to describe either the "triangle" category or the "square" category, for a total of $256 \times 2 = 512$ possible classification rules. In addition, the network experienced each of these patterns *with the rule removed* (i.e., without instruction) so that it associated the "no rule" input unit with an uninformed 50/50 chance for either category. During this initial training phase all 16 stimulus objects were presented alongside each categorization rule, resulting in a total of $512 \times 2 \times 16 = 16384$ patterns used to train the network in the instructional language. These patterns were presented to the network in random orderings for 200 epochs (i.e., passes through the entire training set). Standard incremental (non-batch) backpropagation of mean squared error was used, with activation levels ranging between 0.0 and 1.0, a learning rate of 0.05, and no momentum. The termination at 200 epochs was arbitrary, being a point at which instruction following behavior was good, but not perfect.[1]

In order to observe interaction effects, this network was then exposed to further inductive training in a *particular classification* of the objects. Specifically, further training of the network involved *only* the seven training exemplars shown in Figure 1. This training was conducted separately under two conditions of instruction: *uninstructed* and *instructed*. In the *uninstructed* case, the input *plan* layer was set to the "no rule" configuration for the duration of this training. In the *instructed* case, the input *plan* layer was set to the disjunctive rule shown in Figure 1 for the duration of this training. In both cases, training on the seven exemplars was conducted using standard incremental (non-batch) backpropagation of mean squared error with a very high learning rate (0.5) for 300 epochs. The high learning rate was needed to produce significant learning over a small number of pattern presentations.

The classification performance was recorded on all 16 stimuli in four distinct network states: *Guessing* (immediately after initial training, given no rule to follow – should be guessing 50/50 category assignments), *Rule Following* (immediately after initial training, given the appropriate rule to follow), *Uninstructed* (trained on the seven exemplars, but given no explicit rule), and *Instructed* (trained on the seven exemplars with the appropriate rule at the *plan* layer). The basic hypothesis was that the behavior of the *instructed* network would deviate from the behavior of the initial *rule following* network in a manner which brought it closer to that of the *uninstructed* network.

As another baseline, a network of this kind was exposed to the seven training exemplars *without* any initial training in the instructional language. This "uninstructable" network was trained for 600 epochs at a very high learning rate (0.5) with the *plan* layer always set in the "no rule" configuration. This provided a view of how the category partitions would be formed if no knowledge of the instructional language was available at all.

Modeling Results

The results of the network simulations are summarized in Figure 3. That diagram displays the network predicted probabilities, expressed as percentages, of the given stimulus objects being in the "triangle" category. The only condition not appearing in this figure is the "guessing" case, under which the network consistently produced scores between 42% and 46%, showing a slight but consistent bias towards the "square" category.

An examination of the resulting probabilities reveals the phenomenon of interest in the behavior of these networks. The first case of interest is the difference between the "rule following" network and the "instructed" network. Recall that the difference between these is that the "instructed" network received specific training on the 7 exemplars. The classification probabilities for a number of the objects move markedly away from the "rule following" predictions as a result of exemplar-based training. For example, object "O", which showed the greatest change for the human subjects, drops from 94% to only 64% after exposure to the seven training items. A similar change occurs for object "N", which moves from strong "triangle-ness" to uncertainty. Object "L" provides yet another example, and object "I" shows a weaker trend from "square" to "triangle". One thing that is surprising about these probability changes, however, is that they are all *worse* than the probabilities generated by the "uninstructed" network. That is, our "uninstructed" network follows rules too well, compared to the human data. The output of the "uninstructable" network appears, as expected, much less "rule-like". This suggests that the behavior of the "uninstructed" network was overly shaped by its non-specific experience with the instructional language. That is, its representational

[1]This termination point is discussed further in our closing discussion.

vocabulary is overly rule-based.

Discussion

This demonstration provides some preliminary support to the notion that interference effects in instructed category learning may be appropriately modeled using connectionist networks. The behavior of the "instructed" network provided a reasonable match to the human data, with 7 of the 9 test exemplars classified in the same way, on average. There are a number of ways, however, in which these simulations could be improved.

While the performance of the "instructed" network displayed the interference effect of interest, the behavior of the "uninstructed" network did not match the human subjects data very well. The pattern of probabilities generated by this network more closely matched that of the "rule following" network than that of the "uninstructable" network. This difference is particularly striking for objects "L", "N", "O", and "P". It appears as if this network preferred categories describable as disjunctive rules.

It is not surprising, in retrospect, that the uninstructed network produced rule-like behavior, since *all* categories presented to the network during initial training in the instructional language possessed the disjunctive rule structure. During this initial training phase, hidden units were recruited to encode portions of the rule-structured categories, and these units were later put to use by the inductive learning process. While this resulted in an "uninstructed" network that produced behavior notably different than that of the uninstructed subjects, it is possible that we are modeling a certain class of subjects. Nosofsky and his colleagues noted that, ". . . there is support for the idea that some of the category learners did indeed adopt a simple rule-based strategy . . . ". Indeed, the human data may reflect an average over a bimodal distribution of subjects: those who, like the "uninstructed" network, appeared to induce a rule and others who, like the "uninstructable" network, used a more "similarity based" strategy. Notice that an average of our "uninstructed" and "uninstructable" results provides a somewhat better match to the uninstructed human data than the "uninstructed" network alone.

What is surprising – and this is something for which we have no explanation – is that exposure to the seven exemplars caused the instructed network to deviate from the rule it was given, while the uninstructed network found the correct rule and stuck to it. The difference between these networks rested only in the pattern of activation presented at the *plan* inputs during exemplar training. These inputs were fixed for all seven exemplars, so the inputs acted like a bias during exemplar exposure. In the instructed case, the rule pattern was on, and in the uninstructed case, the "no rule" pattern is on. It could be that this representation actually caused more interference for the network given a rule, because these inputs could be individually combinatorially combined with a *subset* of the possible rules, where the "uninstructed" network had a level playing field from which to select the perfect rule. Our current research is aimed at understanding this puzzle.

In any case, we would like to get something like the effect of combining the "uninstructable" network with the "uninstructed" network in a single network. This might be achieved in two ways. First, initial training on the instructional language could be interspersed with training on random "natural", similarity-based categories over the stimulus space. The network would be equipped with additional output units to represent the category labels for these natural categories. This modified initial training regime would allocate some hidden layer resources to the task of representing the similarity-based categories, and these hidden units could then be redeployed during inductive learning on the seven training exemplars. Alternatively, some hidden units could be architecturally isolated from the instruction inputs, forcing them to encode only similarity information. This would result in a "dual route" mechanism, similar in general configuration to the rule enhanced ALCOVE model (Kruschke & Erickson, 1994). Categorization instructions would be allowed to modulate only "rule-based" hidden units, leaving other hidden units unaffected by the structure of the instructional language. In future work we will examine both of these options, with the goal of producing a network capable of inducing both natural and rule-structured categories.

Another issue for future work involves how the initial instructional language training is controlled. We were required to limit the accuracy of the "rule following" network by limiting the initial training time to 200 epochs. This resulted in a network which followed rules well, but not perfectly. An ideal "rule following" network can be developed by training for 1000 epochs or more, but this is undesirable. The *lack* of perfection in rule following was necessary for this model to work. The reason for this is simple – the network that never makes a mistake has essentially lost plasticity. Interference effects would only arise if actual weight modifications occurred during the exemplar based training phase. Such weight modifications are contingent on a significant error signal. If the network follows rules perfectly, there will be no error signal and, thus, no interference effect from exemplar based training.

This "training to slight imperfection" does not seem very cognitively realistic. It also has unwanted side effects, like the slight "guessing" case bias towards the "square" category which was not corrected by epoch 200. Fortunately, there are a number of other methods that might allow us to learn the instructional language without destroying our error signal. We could introduce normally distributed noise into the network both during training and during regular use. This noise might be localized to the *plan* layer or to the stimuli features, or it might be injected into the net input of every processing element. This noise would generally be averaged out over the course of learning the instructional language, but it would still introduce a non-zero error signal for the exemplar based learning phase.

Conclusion

We have provided some empirical support for the use of our connectionist model of learning "by being told" as a model of instructed category learning. In particular, we have shown that an observed interference effect between instructed rule following and exemplar-based category learning arises naturally in this model. These results also suggest a connectionist mechanism through which both rule governed category structures and "natural" categories may be induced from examples. The observed mismatch between the "uninstructed" network and the uninstructed subjects suggests that we produced a

model that is overly biased towards rules. We are currently pursuing modifications to our training procedure to correct this bias.

Acknowledgements

The work of Nosofsky, Clark, & Shin (1989) was suggested to the authors by Mark St. John. Thanks are also due to the members of *Gary's & Eric's Unbelievable Research Unit (GEURU)* for their comments and suggestions on this work. We also extend our thanks to three anonymous reviewers for their helpful advice concerning the clear presentation of this research.

References

Allen, S. W. and Brooks, L. R. (1991). Specializing the operation of an explicit rule. *Journal of Experimental Psychology: General*, 120(1), 3–19.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, 14(2), 179–211.

Kruschke, J. K. and Erickson, M. A. (1994). Learning of rules that have high-frequency exceptions: New empirical data and a hybrid connectionist model. In Ram, A. and Eiselt, K. (Eds.), *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society* (pp. 514–519). Hillsdale, NJ: Lawrence Erlbaum Associates.

Luce, R. D. (1963). Detection and recognition. In Luce, R. D., Bush, R. R., and Galanter, E. (Eds.), *Handbook of Mathematical Psychology*, volume 1 (pp. 103-189). New York, NY: John Wiley & Sons.

Noelle, D. C. and Cottrell, G. W. (1995). A connectionist model of instruction following. In Moore, J. D. and Lehman, J. F. (Eds.), *Proceedings of the Seventeenth Annual Conference of the Cognitive Science Society* (pp. 369–374). Hillsdale, NJ: Lawrence Erlbaum Associates.

Nosofsky, R. M., Clark, S. E., and Shin, H. J. (1989). Rules and exemplars in categorization, identification, and recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 15(2), 282–304.

Plaut, D. C. and McClelland, J. L. (1993). Generalization with componential attractors: Word and nonword reading in an attractor network. In *Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society* (pp. 824–829). Hillsdale, NJ: Lawrence Erlbaum Associates.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986a). Learning internal representations by error propagation. In Rumelhart, D. E., McClelland, J. L., and the PDP Research Group (Eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* (ch. 8). Cambridge, MA: MIT Press.

Shanks, D. R. and St. John, M. F. (1994). Characteristics of dissociable human learning systems. *Behavioral and Brain Sciences*, 17(3), 367–447.

St. John, M. F. (1992). Leaning language in the service of a task. In *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society* (pp. 271–276). Hillsdale, NJ: Lawrence Erlbaum Associates.

St. John, M. F. and McClelland, J. L. (1990). Learning and applying contextual constraints in sentence comprehension. *Artificial Intelligence*, 46(1–2), 217–257.