Learning of rules that have high-frequency exceptions: New empirical data and a hybrid connectionist model

John K. Kruschke and **Michael A. Erickson**
Dept. of Psychology and Cognitive Science Program
Indiana University, Bloomington, IN 47405-1301
kruschke@indiana.edu

Abstract

Theorists of human learning, in domains as various as category learning and language acquisition, have grappled with the issue of whether learners induce rules or remember exemplars, or both. In this article we present new data that reflect both rule induction and exemplar encoding, and we present a new connectionist model that specifies one way in which rule-based and exemplar-based mechanisms might interact. Our empirical study was motivated by analogy to past tense acquisition, and specifically by the previous work of Palermo and Howe (1970). Human subjects learned to categorize items, most of which could be classified by a simple rule, except for a few frequently recurring exceptions. The modeling was motivated by the idea of combining an exemplar-based module (ALCOVE, Kruschke, 1992) and a rule-based module in a connectionist architecture, and allowing the system to learn which module should be responsible for which instances, using the competitive gating mechanism introduced by Jacobs, Jordan, Nowlan, and Hinton (1991). We report quantitative fits of the model to the learning data.

Introduction

Theorists of human learning, in domains as various as category learning and language acquisition, have grappled with the issue of whether learners induce rules or remember exemplars, or both. In the field of language acquisition, this issue has been highlighted by debate over the adequacy of certain connectionist models, that have no explicit rules, to account for the acquisition of the past tense of English verbs (e.g. Rumelhart & McClelland, 1986; Pinker & Prince, 1988; Plunkett & Marchman, 1991; MacWhinney & Leinbach, 1991; Ling & Marinov, 1993). Pinker (1991; Prasada & Pinker 1993) argued that a satisfactory explanation of past tense acquisition and production requires *both* rules — to account for aspects of regular verbs — *and* exemplar memory — to account for aspects of irregular verbs. In the field of category learning, it has been found that people can learn a classification using different strategies, such that in some situations their learning and performance is best described by a rule, and in other situations it is best described by similarity to the training instances (e.g. Allen & Brooks, 1991; Nosofsky, Clark, & Shin, 1989; Palmeri & Nosofsky, 1994; Regehr & Brooks, 1993). Shanks and St. John (1994) argued that category learning is subserved by two separate and dissociable systems, one for rule induction and one for instance encoding. In this article we present new data that we believe reflect both rule induction and exemplar encoding, and we also present a new connectionist model that specifies one way in which rule-based and exemplar-based mechanisms might interact.

Our empirical study was motivated by analogy to past tense acquisition, and specifically by the previous work of Palermo (Palermo & Eberhart, 1968; Palermo & Howe, 1970). Human subjects learned to categorize items, most of which could be classified by a simple rule, except for a few frequently recurring exceptions. The results showed an increase in the proportion of over-generalization errors as training progressed, and little evidence of regressions in performance (U-shaped learning) on the exceptions. Importantly, we also found that many subjects learned the rule instances earlier in training than the exceptions. The latter result turns out to be the one most difficult to capture by models using instance encoding alone.

The modeling was motivated by the idea of combining exemplar-based and rule-based modules in a connectionist architecture, and allowing the system to learn which module should be responsible for which instances. A modified version of ALCOVE (Kruschke, 1992) served as the exemplar module, linear-threshold nodes (perceptrons) constituted the rule module, and the competitive gating between the modules used a mechanism introduced by Jacobs et al. (1991). We report quantitative fits of the model to the learning data.

Human Learning

In the field of language acquisition, proponents of rule-based theories often adduce the phenomenon of three-stage, or U-shaped, learning of high-frequency irregulars. For example, when learning the English past tense, some children exhibit three stages of acquisition (Ervin, 1964): First, they learn a few high-frequency irregular verbs, such as *go–went*. Second, they learn many regular verbs, such as *walk–walked*, and at the same time often over-generalize the regular suffix, producing forms such as *goed* or *wented*. Third, they relearn the proper forms of the irregulars and acquire a larger vocabulary of lower frequency verbs, both regular and irregular. The acquisition of high-frequency irregular verbs is often described as "U-shaped," because a plot of accuracy as a function of time would show a dip during the second stage. The over-generalization in stage 2 can be accounted for by positing rule induction: The learner has induced and overapplied a rule that was not present earlier in learning.

Recent research (Marcus, Pinker, Ullman, Hollander, Rosen, & Xu, 1992; Plunkett & Marchman, 1991) suggests that the U-shape might be more subtle than initially suggested. It is difficult accurately to measure performance on various verbs, insofar as only a relatively small sample of a child's speech and linguistic environment can be recorded and ana-

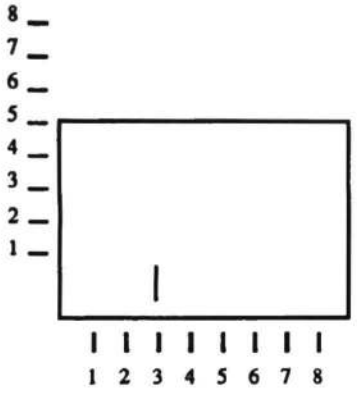

Figure 1: Example of a stimulus used in the category learning experiment.

lyzed. The difficulty of experimental control in natural language led Palermo (Palermo & Eberhart, 1968; Palermo & Howe, 1970) to create laboratory analogues of the language learning situation.

We make no claims about the relation of category learning and language learning (cf. Palermo, 1971; Palermo & Eberhart, 1971), but we are interested in the empirical question of whether carefully controlled category learning can exhibit U-shaped learning curves on high-frequency exceptions to rules, and we are interested in determining whether formal, exemplar-based models can account for the detailed laboratory data.

Method

Subjects learned to classify simple geometric forms into one of six categories. Stimuli were presented on a computer screen, and consisted of a rectangle that could have one of eight heights, and an internal line segment that could have one of eight lateral positions (see Figure 1). On each trial, the subject was shown a stimulus and prompted to make a classification decision. The subject pressed the key corresponding to the category label of their choice, and then the correct label was displayed. In the initial trials, the subjects were just guessing, but after several trials they began to learn which stimuli corresponded to which category labels.

We chose a category structure that approximately followed the design of Palermo and Howe (1970). There were a total of $8 \times 8 = 64$ possible stimuli[1], most of which could be classified by the following simple rule: If the height is 5 or more, then it's in category R1, otherwise it's in category R2. There were four exceptions to the rule, chosen randomly for each subject, constrained so that no two exceptions had the same height or lateral position. In every block of 22 trials, there were 6 distinct instances of rule-based category R1, 6 distinct instances of rule-based category R2, 1 occurrence of the exception category E1, 2 recurrences of the exception E2, 3 recurrences of the exception E3, and 4 recurrences of the exception E4.

[1]Following Palermo and Howe (1970), eight stimuli were omitted from the design, for which the height value equaled the position value. Thus there were actually 56 possible stimuli.

The rule we used, which divided the heights into two equal regions, was simpler than the rule used by Palermo and Howe (1970). Their study included three rule categories that divided the stimulus space into alternating "stripes;" e.g., heights 1, 4 and 6 were category R1, heights 2, 5 and 8 were category R2, etc. We simplified the rule structure because we found in pilot experiments that most subjects could not reliably learn the more complicated structure in the time available (about 1.5 hrs).

Subjects were trained until they performed perfectly for 4 consecutive blocks, or for a maximum of 55 blocks. The experiment lasted about 1.5 hours. Subjects were volunteers from an introductory psychology course at Indiana University, who received partial course credit for their participation.

Results

Unlike Palermo and Howe (1970), we found many subjects who learned the rule-based categories before the exceptions. The fact that Palermo and Howe (1970) found no subjects who learned the rule first can probably be attributed to the more difficult rule used in their study. We decided that a subject had learned the exceptions first if her first best block on exceptions came before her first best block on rules. For example, suppose a given subject has several blocks in which she got 10 out of 10 exceptions correct. The first block in which she achieves that performance is her first best block for exceptions. Suppose the same subject achieves at most 11 correct responses out of 12 rule exemplars in a block. The first block in which she achieves that performance is her first best block for rules. We tried several other methods for dividing subjects into rule-first and exception-first groups, such as overall proportion correct, etc., and they agreed on nearly all the subjects. Of the 70 subjects in the experiment, 49 were classified as exceptions-first, and the remaining 21 were classified as rules-first. (The rules-first group also included ties.) The proportion of correct responses, for each category type, is shown as a function of training block in Figure 2. The solid line in each panel shows performance on the rule exemplars, and the dashed lines show performance on the exceptions. The four dashed lines in each panel correspond to the four different exceptions, with higher frequency exceptions learned better than lower frequency exceptions.

There was no strong evidence of U-shaped learning on the exceptions; the learning curves for the exceptions in Figure 2 show no dramatic drop after performance on the rule cases rises. We also aligned individual subjects' learning curves to their first best block on exceptions, and found a small regression in performance on the two lowest-frequency exceptions. We feel additional replications are required to put much weight on those results, however.

We considered the possibility that absolute performance on the exceptions would not decline, but the proportion of errors on exceptions that are overgeneralization errors might increase suddenly when the rule begins to be learned. For both the exceptions-first and rules-first subjects, there was a gradual, not sudden, increase in the proportion of over-generalization errors throughout the course of learning.

In some respects, then, we failed to find the touchstone empirical phenomena that we sought: We did not find dramatic U-shaped learning on the exceptions, nor did we find a sudden increase in the proportion of over-generalization errors. As

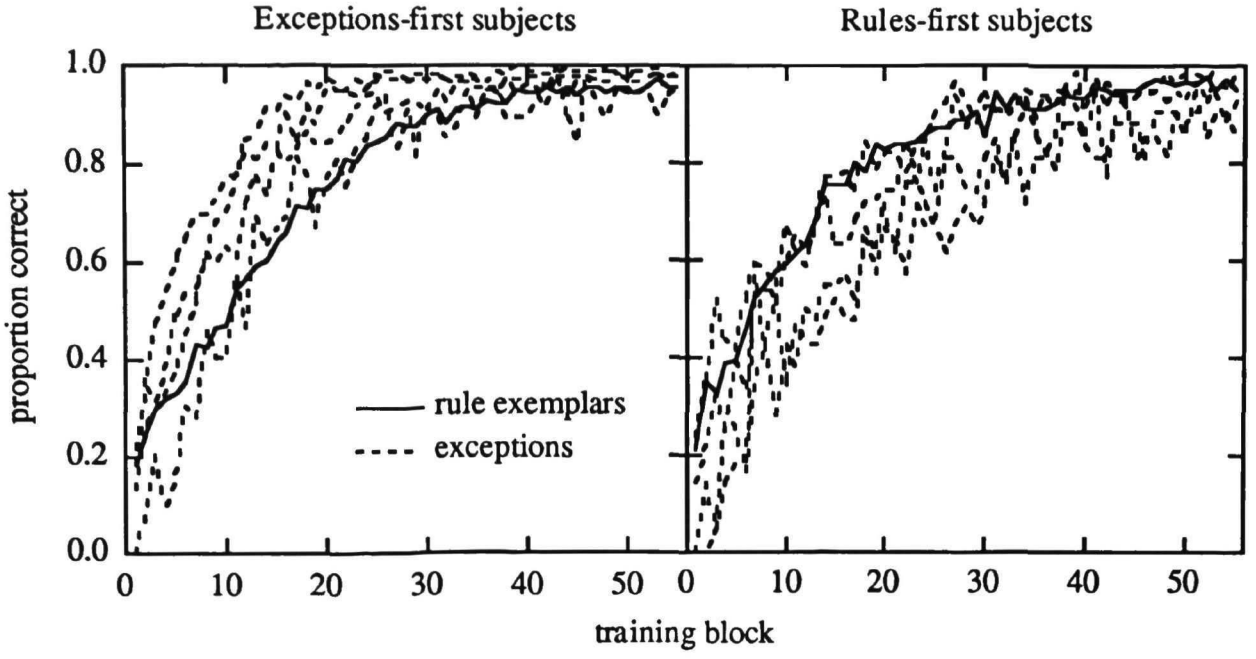

Figure 2: Human learning data.

those two phenomena were thought to be the key challenges to an exemplar-based account, it might seem that an exemplar-based model could handle the results readily. On the contrary, the unexpected result, that many subjects learned the rules first, turns out to be impossible to capture for at least one exemplar-based model.

A Hybrid Connectionist Model

An exemplar-based connectionist model named ALCOVE (Kruschke, 1992, 1993a, 1993b) previously has been shown to account for a variety of phenomena in human category learning. In particular, Kruschke (1992) demonstrated that ALCOVE could exhibit U-shaped learning of high-frequency exceptions to a rule, although in that demonstration there were no human learning data available for quantitative fits. The experiment reported above provides quantitative data from a relevant experiment.

ALCOVE is a feed-forward connectionist network that maps stimuli to category labels. It combines the classification scheme of the generalized context model (Nosofsky, 1986) with the learning mechanism of backpropagation (Rumelhart, Hinton, & Williams, 1986), and thereby formalizes three principles (Kruschke, 1993b): First, its hidden nodes correspond to individual training exemplars, such that the activation of a hidden node represents the similarity of the current input to the exemplar represented by the node. (Figure 3 shows the activation function of an exemplar node in ALCOVE.) Second, each input dimension is gated by an attention strength that reflects the learned relevance of the dimension for the current category distinctions. For example, in the categories used in the experiment described here, only the height of the rectangles is relevant for distinguishing the two rule-based categories, but both dimensions are relevant for distinguishing the exceptions. Third, the dimensional attention strengths, and the association weights between exemplars and categories, are learned via gradient descent on an error measure.

ALCOVE suffers two main problems when fit to the data in Figure 2: First, it cannot learn the rules as quickly as the high-frequency exceptions, unlike the rules-first subjects. The model suffers because it cannot generalize across rule exemplars extensively and rapidly enough; the exemplar-based similarity function is too localized for rapid extrapolation to distant exemplars. A second shortcoming of the model is that it is much too sensitive to the relative frequencies of the exceptions, so that the spread between the learning curves of the four exceptions is too large.

We are not claiming that no possible exemplar-based model could fit these data. Rather, we fit one promising model to the data and found it lacking. As described in the introduction, other evidence also points to inadequacies in purely exemplar-based models of human category learning.

When a model fails to account for a set of data, the theorist has two choices: Either re-formulate the existing principles embodied in the model, or incorporate new principles (Kruschke, 1993b). We believe that solutions to the aforementioned failures require new principles, not just reformulated old principles. One of the added principles is rule-based representation, used in conjunction with exemplar-based representation.

Model architecture

To accommodate the rapidity with which some subjects learn the rule-based instances, we conjoined to ALCOVE a module of nodes that represented rules. We wanted the nodes to represent simple rules such as "if the height is greater than value V, then it's in category X." A natural way to do

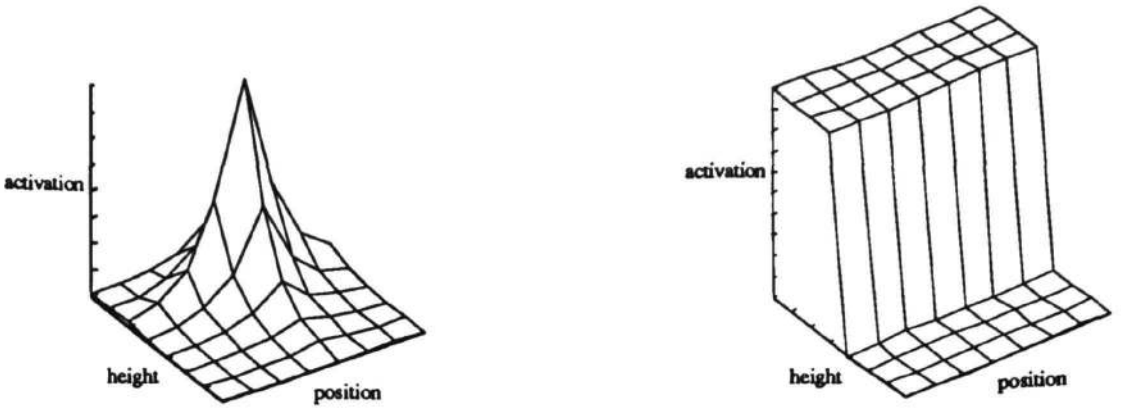

Figure 3: Receptive fields of an exemplar node (left) and a rule node (right).

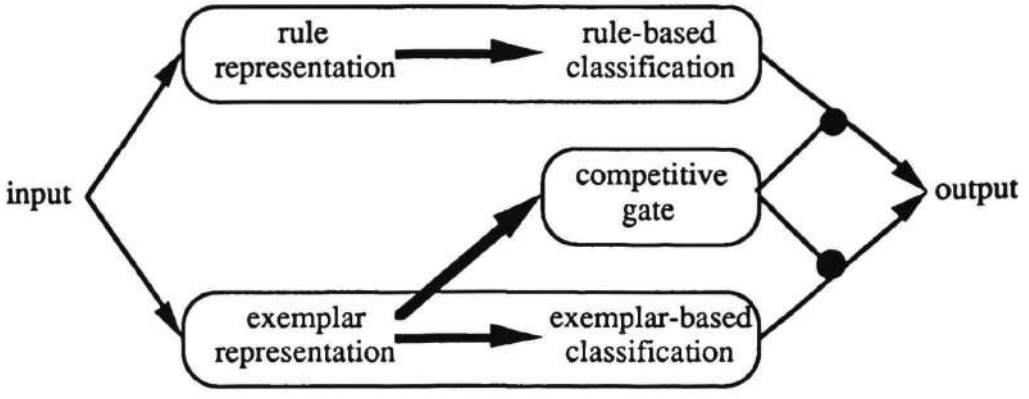

Figure 4: Architecture of the model. Heavy arrows denote connections with learned weights.

that is with linear-threshold nodes (perceptrons) aligned to the dimensional axes of the input space. Figure 3 illustrates the difference between the two types of representation: In ALCOVE's exemplar-based representation, each node is activated by a limited region of the input space, whereas in the rule-based representation, each node is activated by an entire half-space. Thus, we have two modules of nodes that embody different basis functions. The underlying notion is that the different representational schemes used by human learners can be implemented by different basis functions in a connectionist network.

Whereas previous researchers have suggested that human category learning uses both rules and exemplars, the difficult task of determining how those subsystems interact is yet to be worked out (cf. Shanks & St. John, 1994). In our model, we used the competitive gating scheme of Jacobs et al. (1991) to govern the learning of the exemplar and rule modules, as illustrated in Figure 4. Each rule node was connected to a set of category nodes, and each exemplar node was connected to a distinct set of category nodes. The final output was determined by randomly selecting either the rule-based or exemplar-based classification according to probabilities given by a competitive gating node. The gating node was connected to the exemplar representation only, so that the exemplar nodes could "veto" the more general rule nodes, when necessary. The category nodes in both modules, and the gating node, were simple linear summators (as in ALCOVE). All the connection weights to the category nodes and the gating node were initialized to zero and adjusted by gradient descent on the error function described in Jacobs et al. (1991).

The model had six, freely-estimated parameters, including the following five: the learning rate for the rule-based category nodes, the learning rate for the exemplar-based category nodes, the learning rate for the gating node, the fixed bias of the gating node, and the fixed receptive field diameter of the exemplar nodes. There was no dimensional attention learning in the exemplar module; i.e., the attention learning rate in ALCOVE was set to zero, because it was assumed that selective attention to dimensions would be accomplished by the competition between dimension-specific rules.

One other new principle was introduced into the model, to accommodate the relatively small spread between the learning curves of the different-frequency exceptions. Connections from exemplars to category nodes suffered a refractory period after learning. That is, if a given connection weight was adjusted on a certain trial, then its effective learning rate imme-

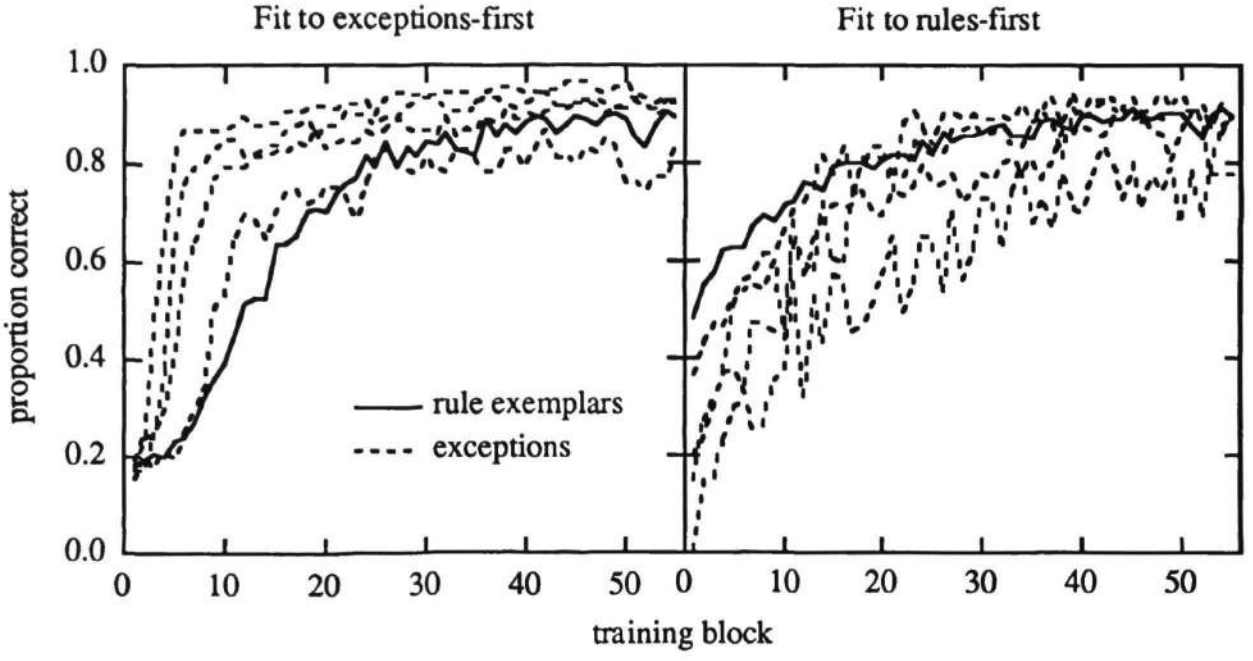

Figure 5: Fits of the model to the human learning data in Figure 2.

diately plunged after that trial and gradually recharged over subsequent trials. The recharge rate was the sixth parameter of the model. The ramifications of this modification are extensive: It might help account for a variety of phenomena in human learning, such as the effects of massed vs. distributed learning, etc. This particular modification is not the focus of this paper, however, so we will not discuss it further.

Fit to human data

Figure 5 shows the predictions of the model. The best-fitting parameter values reflected our intuitions about the subjects: For the exceptions-first group, the learning rate for the exemplar module was high (1.03) and the learning rate for the rule module was very low (0.0!), whereas for the the rules-first group, the opposite was true, with the exemplar-module learning rate low (0.41) and the rule-module learning rate high (1.27). The human learners in the exceptions-first group showed a gradual but robust increase in the proportion of over-generalization errors throughout training, and so did the model, despite the disuse of the rule module. In that case, the model showed overgeneralization entirely because of exemplar similarity: Each exception was surrounded by rule instances, so when an error was made, it tended to be an over-generalization error.

Summary and Conclusion

In this article we have emphasized two main points: First, a novel empirical phenomenon that required us to seek a hybrid, rule and exemplar, model was that some subjects learned the rule before they learned the exceptions, whereas other subjects learned the (high-frequency) exceptions before they learned the rule. Second, we described a candidate connectionist architecture for integrating rule-based and exemplar-based modules. It introduces the combination of ideas that different representational schemes for human categorization may be implemented by different basis functions in connectionist networks, and those basis functions can compete to learn the categories. The model also used a novel method of refractory learning rates on the connection weights.

The proposed architecture is not intended, in its present form, as a comprehensive model of human category learning in general. Rather, it is intended to demonstrate the veracity of the approach. Future models could include additional basis-function modules for representing prototypes as, e.g., radial basis functions with adaptive receptive fields, and decision boundaries as, e.g., polynomial basis functions. More complicated schemes for gating the modules might also need to be developed. We believe that one way to achieve significant progress on the theoretical issue of how rules and exemplars interact in human learning is to formulate specific models and test them with detailed quantitative data, as we have described here.

Acknowledgments

This research was supported in part by NIMH FIRST Award 1-R29-MH51572-01 to Kruschke, and by Indiana University Cognitive Science Program Fellowships to Erickson. Thanks to three anonymous reviewers for helpful comments. For their help in conducting related experiments we thank Kevin Adamick, Rebecca Baldwin, Shawn Fleck, Gilbert Hyatt, Don Lyon, Timothy Magill, Anthony Masterson, Jane Ream, Jennifer Ressler, and Michael Rush.

References

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

Ervin, S. M. (1964). Imitation and structural change in children's language. In Lenneberg, E. G. (Ed.), *New directions in the study of language*. MIT Press, Cambridge, MA.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., & Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural Computation, 3*, 79–87.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review, 99*, 22–44.

Kruschke, J. K. (1993a). Human category learning: Implications for backpropagation models. *Connection Science, 5*, 3–36.

Kruschke, J. K. (1993b). Three principles for models of category learning. In Nakamura, G. V., Taraban, R., & Medin, D. L. (Eds.), *Categorization by Humans and Machines: The Psychology of Learning and Motivation*, Vol. 29, pp. 57–90. Academic Press, San Diego.

Ling, C. X., & Marinov, M. (1993). Answering the connectionist challenge: A symbolic model of learning the past tenses of English verbs. *Cognition, 49*, 235–290.

MacWhinney, B., & Leinbach, J. (1991). Implementations are not conceptualizations: Revising the verb model. *Cognition, 40*, 121–157.

Marcus, G. F., Pinker, S., Ullman, M., Hollander, M., Rosen, T. J., & Xu, F. (1992). *Overregularization in Language Acquisition*, Vol. 57(4). University of Chicago Press, Chicago. Serial No. 228.

Nosofsky, R. M. (1986). Attention, similarity and the identification-categorization relationship. *Journal of Experimental Psychology: General, 115*, 39–57.

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

Palermo, D. S. (1971). On learning to talk: Are principles derived from the learning laboratory applicable?. In *The ontogenesis of grammar*, pp. 41–62. Academic Press, New York.

Palermo, D. S., & Eberhart, V. L. (1968). On the learning of morphological rules: An experimental analogy. *Journal of Verbal Learning and Verbal Behavior, 7*, 337–344.

Palermo, D. S., & Eberhart, V. L. (1971). On the learning of morphological rules: A reply to Slobin. In *The ontogenesis of grammar*, pp. 225–229. Academic Press, New York.

Palermo, D. S., & Howe, Jr., H. E. (1970). An experimental analogy to the learning of past tense inflection rules. *Journal of Verbal Learning and Verbal Behavior, 9*, 410–416.

Palmeri, T. J., & Nosofsky, R. M. (1994). Recognition memory for exceptions. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 00*, 00–00. in press.

Pinker, S. (1991). Rules of language. *Science, 253*, 530–535.

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. (1991). U-shaped learning and frequency effects in a multi-layered perceptron: Implications for child language acquisition. *Cognition, 38*, 43–102.

Prasada, S., & Pinker, S. (1993). Generalisation of regular and irregular morphological patterns. *Language and Cognitive Processes, 8*, 1–56.

Regehr, G., & Brooks, L. R. (1993). Perceptual manifestations of an analytic structure: the priority of holistic individuation. *Journal of Experimental Psychology: General, 122*, 92–114.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature, 323*, 533–536.

Rumelhart, D. E., & McClelland, J. L. (1986). On learning the past tenses of english verbs. In McClelland, J. L., & Rumelhart, D. E. (Eds.), *Parallel Distributed Processing*, Vol. 2, chap. 18, pp. 216–271. MIT Press, Cambridge, MA.

Shanks, D. R., & St. John, M. F. (1994). Characteristics of dissociable human learning systems. *Behavioral and Brain Sciences, 00*, 000–000. in press.