

The Forgetting Algorithm: How Fragmentary Knowledge of Exemplars Can Abstract Knowledge

Robert C. Mathews
Louisiana State University

The position of Perruchet and Pacteau (1990, 1991) concerning what is retained about past exemplars when subjects implicitly learn an artificial grammar oscillates between two extremes (bigram information versus intact exemplars). On the other hand, THYOS, the model proposed by Mathews, Druhan, and Roussel (1989), on the basis of classifier systems, is capable of finding optimal sets of features to retain through rule competition. It is argued that people implicitly generate abstract rules that cannot be explained in terms of retention of bigrams or larger chunks without retention of spatial information and that THYOS integrates findings associated with prototype and exemplar models of induction.

A couple of years ago, while I was puzzling over what type of nonconscious process might be capable of learning artificial grammars, I was struck by a very simple but powerful idea: Partial memories of exemplars can be used as rules to identify valid strings. Whereas an exact memory of an exemplar matches only itself among the set of all possible strings one might see, fragments or selected features from the same exemplar identify or match a whole set of possible strings. For example, remembering that I saw an exemplar that began with "SCT" allows me to choose other strings on subsequent trials that share the same features. Alternately, I could remember whole exemplars and choose future strings based on their relative similarity to my remembered exemplars. However, hidden in this notion is a host of problems concerning how similarity is to be computed (see Medin, 1990).

The simple idea of making rules out of partial memories for exemplars, which we call the forgetting algorithm (Mathews, Druhan, & Roussel, 1989) contains no hidden complexities. Each fragmentary memory or "rule" either matches a future candidate string or it does not. No need for considering partial matches. Further, the set of strings that is selected by such a rule is likely to contain more valid strings than a randomly selected set of choices. Thus, rules generated by this forgetting algorithm will often constitute partially valid rules for selecting valid strings (see Mathews, Druhan, & Roussel, 1989).

The forgetting algorithm seems like the type of mechanism that one might expect to find in the cognitive unconscious: It is blessedly simple. What could be easier than forgetting part of an experience to create a new rule? It requires very little processing resources beyond those necessary to encode the experience. Finally, when inserted into a system capable of

using lots of partially valid rules and adjusting their strength based on feedback from the task, these simple rules are capable of performing complex tasks (Goldberg 1988; Holland, Holyoak, Nisbett, & Thagard, 1986). We have incorporated this idea into a model of implicit learning based on the Holland et al. (1986) theory that we call THYOS. This model has been found to fit human data well in a variety of experiments on artificial grammar learning (Druhan & Mathews, 1989; Mathews, Druhan, & Roussel, 1989; Roussel, Mathews, & Druhan, 1990).

Part of the beauty of the forgetting algorithm is that it suggests that it might be more adaptive to forget features of experienced episodes rather than retain all their features. It forces us to reevaluate what memory researchers have traditionally considered a down side of human memory, namely interference or forgetting. Perhaps forgetting is not a loss of information or, in the words of Perruchet and Pacteau (1991), a "failure to integrate," but an adaptive system for capitalizing on common features of experiences to optimize performance in similar situations. This turnabout for viewing forgetting positively seems to be a very difficult pill to swallow for many cognitive psychologists. The typical reaction to this idea is "Wait a minute, forgetting is forgetting, it is not a *real* abstraction process." This unwillingness to look favorably on forgetting is evident in Perruchet and Pacteau's (1991) comment on the notion that fragments of exemplars are abstract rules: "The knowledge that can be qualified as abstract in a formal sense may result from a failure to integrate or unify the components of the displayed strings, rather than from an active process of feature extraction from a primarily unitary representation" (p. 113). But why integrate if the fragments give people more power to generalize across different episodes having common family resemblances?

Of course, not all features of experienced episodes are equally valuable for making good responses in the future. Some rules based on partial memories will turn out to be absolutely worthless. Even the good rules that emerge from the forgetting algorithm are typically only partially valid as a guide for future performance. It is only when large numbers of partially valid rules generated by such algorithms are allowed to function *in concert* that sophisticated behavior can

Research for this article was supported in part by National Science Foundation Grant BNS-8509493 to Robert C. Mathews, Ray R. Buss, and William B. Stanley, and in part by Louisiana Board of Regents Grant 86-LBR(21)-021-10 through the Louisiana Quality Education Support Fund to Robert C. Mathews.

Correspondence concerning this article should be addressed to Robert C. Mathews, Department of Psychology, Louisiana State University, Baton Rouge, Louisiana 70803-5501.

emerge. Thus, a system is needed to select higher validity rules and to resolve conflicts when rules disagree on their course of action. THYOS uses the classifier system architecture of Holland et al. (1986). Classifier systems use simple local procedures to adjust strengths of rules based on feedback and to resolve conflicts among sets of rules. Because these mechanisms do not require centralized control, THYOS is an appropriate model for nonconscious learning.

The Perruchet and Pacteau (1990, 1991) view seems to oscillate between two extremes concerning the nature of knowledge acquired about artificial grammars. The gist of their earlier article (1990) is that subjects only retain knowledge about tiny fragments of valid strings, such as knowledge of valid bigrams or letter pairs that can occur in valid strings. They argued that such bigram knowledge is almost sufficient to account for subjects' performance on their string discrimination task (however, see Servan-Schreiber & Anderson, 1990, for evidence that higher level chunks are involved). On the other hand, in their more recent article, when pressed to account for successful identification of valid strings in experiments in which subjects were transferred to a different letter set (one that had no bigrams in common with the old letter set), they switched to an exemplar model. Exemplar models suggest that intact exemplars are retrieved from memory and subjects respond to new strings through analogies to retrieved exemplars. It seems odd that subjects' knowledge would either consist of tiny fragments (bigrams) or complete exemplars. Why not anything in between, such as the trigrams or specific beginnings associated with specific endings? Classifier systems are capable of automatically seeking out and using chunk sizes of optimal validity for the task at hand (see Goldberg, 1988). We suspect that subjects' implicit memories are similarly adaptive to task demands, resulting in chunk sizes that vary from task to task without conscious control of the process.

Although we are perfectly at home with the idea that rules such as "look for strings that begin with S" or "look for strings that end in VV" are abstract rules (see Mathews, 1990), other researchers do not consider such rules to be the type of abstraction they associate with "active abstraction processes." Therefore, I would like to describe another type of abstraction that occurs implicitly with the biconditional grammar (Mathews, Buss, et al., 1989). In the biconditional grammar, any of the six letters in the letter set can occur in the first four positions, but these letters determine precisely what additional letters must follow in corresponding positions in the right half of the string (see Figure 1). There are three letter correspondence rules that specify which pairs of letters go together in corresponding serial positions in each half of the string (S goes with V, C goes with P, and T goes with X).

Because any sequence of four letters can occur in either half of valid strings generated by the biconditional grammar (although that sequence determines what must be in the other half of the string), all possible pairs of adjacent letters or bigrams can occur in valid strings. Thus, this grammar cannot be learned through memory of bigrams or higher order chunks of adjacent letters (Servan-Schreiber & Anderson, 1990) without reference to where they occur in a string (the spatial position that chunks must occur in). Yet subjects can learn

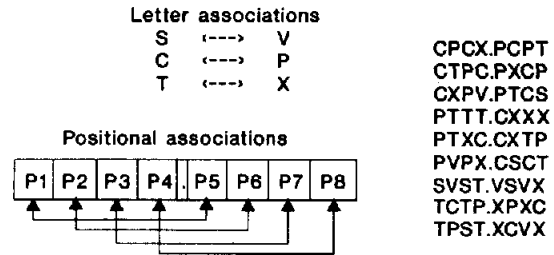


Figure 1. Illustration of the letter and positional association rules and several examples of valid strings generated by the biconditional grammar.

to select valid strings with greater than chance accuracy through implicit training. In subsequent experiments conducted in our lab using the biconditional grammar, we have found well above chance performance on a multiple choice string discrimination task following implicit training using the match task (see Mathews, Buss, et al., 1989, Exps. 3 & 4). Thus, people can acquire knowledge of the biconditional grammar implicitly. Moreover, when transferred to a new letter set, subjects continue to perform well without additional feedback trials (some subjects even do better when the letter set is changed). When asked how they performed the task, these subjects said things like "I picked the most symmetrical strings," or "I picked the strings with the best pattern." Such symmetry rules relate to similarities in patterns of repetition of letters across halves of valid strings (e.g., PTTT.CXXX or XXVV.TTSS), and they are partially valid rules for selecting valid strings.

There are several aspects of such implicitly acquired "symmetry" rules that are interesting. First, they cannot be built up from previously acquired bigram or trigram rules because no such rules are valid in this grammar. Second, positional information is inherent in such symmetry rules (e.g., repetition of a pair of letters in the first two positions will be matched by repetition of a pair of letters in Positions 5-6). Third, these rules are acquired without any attempt to consciously discover rules. As far as the subjects know, each trial in the match task involves memorizing a unique string for a test of short-term memory. Their goal is to remember the string presented on that trial long enough (a few seconds) to pick it out on a subsequent screen containing five choices. There is no incentive to organize or remember strings across sets of trials (that would only make a subject's task harder because of proactive interference). Yet subjects acquire abstract symmetry rules under such implicit training conditions. Finally, it is interesting that, on the basis of verbal reports taken during the string discrimination task, none of our implicitly trained subjects ever learned the specific letter association rules (e.g., that S goes with V). The symmetry rules they acquired seem to involve abstract features of the strings (e.g., pattern goodness) and they do not get more specific after additional implicit training trials. The forgetting algorithm is capable of generating abstract rules like the symmetry rules if (a) abstract features like the occurrence of repeated letters or runs are encoded and (b) spatial informa-

tion associated with the abstract features (e.g., positions of the runs in the strings) is also retained.

We believe the disagreements embodied in these commentaries are small relative to the agreements of researchers concerned with implicit learning and that a good theory of implicit learning is close at hand. Most researchers seem to agree that implicit learning is based on memories of prior experiences (e.g., Brooks, 1978, 1987; Dulany, Carlson, & Dewey, 1984; Estes, 1986; Vokey & Brooks, in press) or what my colleagues and I have called *memory-based processing* (Mathews, Buss, et al., 1989). At the present stage, there are disagreements about how complete memories of prior instances might be (e.g., bigrams, chunks, or whole exemplars) and how such memories are used to guide performance. Other disagreements seem to concern value judgments rather than matters of fact (e.g., is this type of knowledge really abstract).

We view THYOS as a system that integrates notions from both exemplar and prototype models. Once subjects are allowed to have less than perfect memory for exemplars, such fragmentary memories of exemplars correspond to THYOS's rules. After THYOS has had an opportunity to perform a task and adjust the relative strength of its rules, its highest strength set of rules identify most typical instances. However, lower strength rules remain available and they are capable of overriding stronger rules when (a) several weaker rules act in concert (through support) or when (b) the weaker rules are more specific to the current episode compared with the more general default rules (see Holland et al., 1986). Thus, THYOS is capable of demonstrating preferences for typical instances and it is also capable of learning to successfully respond to local regularities, like exemplar models. Also, this type of model is capable of being implemented across diverse task domains in which evidence of implicit learning has been found, ranging from controlling sugar production in an imaginary sugar factory (Berry & Broadbent, 1984, 1988; Stanley, Mathews, Buss, & Kotler-Cope, 1989) to visual search tasks (Lewicki, 1986; Lewicki, Czyzewska, & Hoffman, 1987; Stadler, 1989). Perhaps the most intriguing aspect of this theory is that it suggests that we may have to get used to the idea that forgetting is learning.

References

- Berry, D. C., & Broadbent, D. E. (1984). On the relationship between task performance and associated verbalizable knowledge. *Quarterly Journal of Experimental Psychology*, *36A*, 209-231.
- Berry, D. C., & Broadbent, D. E. (1988). Interactive tasks and the implicit-explicit distinction. *British Journal of Psychology*, *79*, 251-272.
- Brooks, L. R. (1978). Non-analytic concept formation and memory for instances. In E. Rosch & B. Lloyd (Eds.), *Cognition and Concepts*. Hillsdale, NJ: Erlbaum.
- Brooks, L. R. (1987). Decentralized control of categorization: The role of prior processing episodes. In U. Neisser (Eds.), *Concepts and conceptual development: Ecological and intellectual factors in categorization* (pp. 141-174). Cambridge, England: Cambridge University Press.
- Druhan, B. B., & Mathews, R. C. (1989). THYOS: A classifier system model of implicit knowledge of artificial grammars. *Proceedings of the Eleventh Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.
- Dulany, D. E., Carlson, R. A., & Dewey, G. I. (1984). A case of syntactical learning and judgment: How conscious and how abstract? *Journal of Experimental Psychology: General*, *113*, 541-555.
- Estes, W. K. (1986). Array models for category learning. *Cognitive Psychology*, *18*, 500-549.
- Goldberg, D. E. (1988). *Genetic algorithms in search, optimization & machine learning*. New York: Addison-Wesley.
- Holland, J. H., Holyoak, K. J., Nisbett, R. E., & Thagard, P. R. (1986). *Induction: Processes of inference, learning, and discovery*. Cambridge, MA: MIT Press.
- Lewicki, P. (1986). *Nonconscious social information processing*. New York: Academic Press.
- Lewicki, P., Czyzewska, M., & Hoffman, H. (1987). Unconscious acquisition of complex procedural knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*, 523-530.
- Mathews, R. C. (1990). Abstractness of implicit grammar knowledge: Comments on Perruchet and Pacteau's analysis of synthetic grammar learning. *Journal of Experimental Psychology: General*, *119*, 412-416.
- Mathews, R. C., Buss, R. R., Stanley, W. B., Blanchard-Fields, F., Cho, J., & Druhan, B. (1989). The role of implicit and explicit processes in learning from examples. A synergistic effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*, 1083-1100.
- Mathews, R. C., Druhan, B. B., & Roussel, L. G. (1989). *Forgetting is learning: Evaluation of three induction algorithms for learning artificial grammar*. Paper presented at the annual meeting of The Psychonomic Society, Boston, MA.
- Medin, D. L. (1990). Concepts and conceptual structure. *American Psychologist*, *44*, 1469-1481.
- Perruchet, P., & Pacteau, C. (1990). Synthetic grammar learning: Implicit rule abstraction or explicit fragmentary knowledge. *Journal of Experimental Psychology: General*, *119*, 264-275.
- Perruchet, P., & Pacteau, C. (1991). Implicit acquisition of abstract knowledge about artificial grammar: Some methodological and conceptual issues. *Journal of Experimental Psychology: General*, *120*, 112-116.
- Roussel, L. G., Mathews, R. C., & Druhan, B. B. (1990). Rule induction and interference in the absence of feedback: A classifier system model. *Proceedings of The Twelfth Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.
- Servan-Schreiber, E., & Anderson, J. R. (1990). Learning artificial grammars with competitive chunking. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*, 592-608.
- Stadler, M. A. (1989). On learning complex procedural knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*, 1061-1069.
- Stanley, W. B., Mathews, R. C., Buss, R. R., & Kotler-Cope, S. (1989). Insight without awareness: On the interaction of verbalization, instruction and practice in a simulated process control task. *Quarterly Journal of Experimental Psychology*, *41a*, 553-577.
- Vokey, J. R., & Brooks, L. R. (in press). The salience of item knowledge in learning artificial grammars. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Received September 10, 1990

Accepted September 11, 1990 ■