THE QUARTERLY JOURNAL OF EXPERIMENTAL PSYCHOLOGY, 2016
Vol. 69, No. 6, 1049–1055, http://dx.doi.org/10.1080/17470218.2015.1130068

Rapid communication

Applying an exemplar model to an implicit rule-learning task: Implicit learning of semantic structure

Chrissy M. Chubala[1], Brendan T. Johns[2], Randall K. Jamieson[1], and D. J. K. Mewhort[3]

[1]Department of Psychology, University of Manitoba, Winnipeg, MB, Canada
[2]Department of Communicative Disorders and Sciences, State University of New York at Buffalo, Buffalo, NY, USA
[3]Department of Psychology, Queen's University at Kingston, Kingston, ON, Canada

(Received 11 September 2015; accepted 15 October 2015; first published online 16 February 2016)

Studies of implicit learning often examine peoples' sensitivity to sequential structure. Computational accounts have evolved to reflect this bias. An experiment conducted by Neil and Higham [Neil, G. J., & Higham, P. A.(2012). Implicit learning of conjunctive rule sets: An alternative to artificial grammars. *Consciousness and Cognition, 21*, 1393–1400] points to limitations in the sequential approach. In the experiment, participants studied words selected according to a conjunctive rule. At test, participants discriminated rule-consistent from rule-violating words but could not verbalize the rule. Although the data elude explanation by sequential models, an exemplar model of implicit learning can explain them. To make the case, we simulate the full pattern of results by incorporating vector representations for the words used in the experiment, derived from the large-scale semantic space models LSA and BEAGLE, into an exemplar model of memory, MINERVA 2. We show that basic memory processes in a classic model of memory capture implicit learning of non-sequential rules, provided that stimuli are appropriately represented.

Keywords: Implicit learning; Instance theory; MINERVA 2; LSA; BEAGLE.

People are sensitive to regularities in their environment. In some cases, people can respond to those regularities without deliberate effort, a phenomenon typically referred to as *implicit learning*. Implicit learning has been studied in the laboratory using two primary tasks. One is a classification task: Participants first study letter strings created using a finite-state grammar. After studying grammatical exemplars, participants can sort novel test exemplars into grammatical versus ungrammatical sets with an accuracy of about 60%. The second is a performance task: Participants identify a target's position by pressing an associated response key. If the sequence of targets is repeated, participants learn to respond more quickly. In both tasks, the participants cannot articulate the basis of their performance.

Structure is defined in both tasks by sequential rules. Not surprisingly, standard models of implicit

Correspondence should be addressed to Chrissy M. Chubala, Department of Psychology, University of Manitoba, Winnipeg, MB, Canada R3T 2N2. E-mail: umchubal@myumanitoba.ca

We thank Greg Neil and Phil Higham for their cooperation and discussion.

This research was supported by a Natural Sciences and Engineering Research Council of Canada (NSERC) Alexander Graham Bell Canada Graduate Scholarship (doctoral program) held by C.M.C. and NSERC Discovery Grants held by R.K.J. and D.J.K.M.

learning, including the simple recurrent network (Dienes, Altmann, & Gao, 1999) and the competitive chunking model (Servan-Schreiber & Anderson, 1990), are focused on how people learn sequential structure. Implicit learning, however, is not restricted to cases involving sequential structure. For example, people can learn non-sequential functions in a video game (Berry & Broadbent, 1984), can learn to discriminate rule-consistent from rule-violating line drawings (Pothos & Bailey, 2000), and can appreciate semantic structure (Higham & Brooks, 1997; Neil & Higham, 2012).

In Neil and Higham's (2012) experiment, participants rated the meaning of 80 English words and subsequently learned that the words had been selected according to a rule. In one condition (labelled RC-CA), each study word was either rare and concrete (e.g., manganese) or common and abstract (e.g., indicate). In a second condition (labelled RA-CC), each study word was either rare and abstract (e.g., beguiled) or common and concrete (e.g., hotel). In a test phase, participants tried to discriminate rule-consistent from rule-violating words in a two-alternative forced-choice (2AFC) procedure. Whereas participants were correct on 57.3% of the trials, they could not articulate the rule.

Neil and Higham's (2012) results parallel those from the standard artificial grammar classification task (rule discrimination without rule knowledge). However, in Neil and Higham's example, the rule was not sequential. Clearly, accounts of implicit learning based on learning sequential structure are incomplete. How, then, should implicit learning be explained?

Brooks (1978; Vokey & Brooks, 1992) argued that implicit learning in the artificial grammar task reflected exemplar-based generalization: Test probes that remind the learner of the studied items are judged to be grammatical, and those that do not are rejected as ungrammatical. Jamieson and Mewhort (2009, 2010) implemented Brooks's idea in a computational account of memory and showed that it accommodates performance in the judgement of grammaticality, string completion, and serial reaction-time tasks. Their account was adapted from Hintzman's (1986) MINERVA 2, a multiple-trace model of memory designed to model recognition and cued recall in non-sequential stimulus domains. Given that the model has been used to explain learning in sequential and non-sequential domains, it might provide an account of Neil and Higham's (2012) data.

THE MODEL

Previous work has expanded the explanatory breadth of MINERVA 2 by borrowing the architectural and retrieval assumptions from Hintzman (1986), but altering the way items in memory are represented. In MINERVA 2, each item is represented by a random vector of binary feature values (+1 and −1). To accommodate representation of nonsense letter strings, Chubala and Jamieson (2013) altered the random representation assumption to encode strings as collections of letter subunits (e.g., bigrams; see also Jamieson & Mewhort, 2011). In order to model Neil and Higham's (2012) task with MINERVA 2, we extend this theme and represent items using semantic vectors derived with two current models of semantic representation: LSA (Landauer & Dumais, 1997) and BEAGLE (Jones & Mewhort, 2007).

Architecture and retrieval

Memory is an $m \times n$ matrix $\mathbf{M}$, where each row represents an event in the model's history, m is the number of independent traces stored in memory, and n is the number of elements in each trace. To model that fact that people's encoding is imperfect, we set a proportion of elements in $\mathbf{M}$ to zero. The proportion of elements reset to zero is controlled by a learning rate parameter L that specifies the probability of storing each feature correctly. Each element in $\mathbf{M}$ has a probability $1 - L$ of reverting to zero.

Retrieval follows a resonance metaphor. Presenting a probe to memory activates all traces in memory in parallel. The degree of activation is

a function of the similarity of each item to the probe. Activation in memory is collected in a vector called the *echo*, which is equal to a weighted sum of the activated traces in memory:

$$c = \sum_{i=1}^{m} \left(\frac{\sum_{j=1}^{n} p_j \times M_{ij}}{\sqrt{\sum_{j=1}^{n} p_j^2} \sqrt{\sum_{j=1}^{n} M_{ij}^2}} \right)^3 \times M_i \quad (1)$$

where **c** is the echo, m is the number of traces in memory, n is the dimensionality of words in memory, p_j is feature j in the probe, M_{ij} is feature j in row i in memory, and M_i is trace i in memory.

Evidence that a probe is grammatical is indexed by its echo intensity, I, which is computed as,

$$I = \frac{\sum_{j=1}^{n} p_j \times c_j}{\sqrt{\sum_{j=1}^{n} p_j^2} \sqrt{\sum_{j=1}^{n} c_j^2}} \quad (2)$$

where n is the dimensionality of a representation, p_j is feature j in the probe, and c_j is feature j in the echo. To model performance in a 2AFC test, the echo intensity is computed for the two response alternatives, and the alternative with the higher echo intensity is selected.

Simulations

In the simulations, we used LSA and BEAGLE to derive vectors that corresponded to the words used in Neil and Higham's (2012) study. In a first simulation, we used LSA (Landauer & Dumais, 1997) vectors (of dimensionality 300) derived by Günther, Dudschig, and Kaup (2015) using the TASA text corpus (Touchstone Applied Science Associates, Inc.).[1] In a second simulation, we used vectors (also of dimensionality 300) derived using the BEAGLE algorithm described by Jones and Mewhort (2007) and by Recchia, Sahlgren, Kanerva, and Jones (2015).

Because we assumed imperfect encoding (L = .7), results differed from one simulation to the next. Thus, we conducted 2000 independent simulations in total: 1000 using Neil and Higham's RA-CC training list and 1000 using their RC-CA training list. In all simulations, the test was conducted using the same test pairs (see Neil & Higham, 2012).[2]

Except for the difference in training lists, the simulations were the same. We stored the relevant training list to memory, forced information loss (L = .7), computed the echo for each word in the test list, and recorded the percentage of test trials in which the echo intensity for the rule-consistent alternative was greater than the echo intensity for its rule-violating counterpart.

RESULTS

The simulation results are shown in Figure 1. Results based on the LSA vectors are shown in the left column; results based on the BEAGLE vectors are shown in the right column.

As shown in the top row in Figure 1, the model predicted a mean percentage correct score of 58.4% with the LSA vectors and 55.2% with the BEAGLE vectors. In both cases, the mean percentage correct was better than chance, $p < .01$, and in both cases the simulated mean percentage score was close to the observed empirical mean percentage correct score of 57.3% (SD = 5.1%) in Neil and Higham's (2012) study.

The middle and bottom rows in Figure 1 show the distributions of percentage correct scores for simulations with the RA-CC and RC-CA training lists, respectively. As shown, the model predicted better discrimination of rule-consistent from rule-violating test items after studying the RA-CC than after studying the RC-CA training list. The prediction held for simulations using either the LSA (M = 60.2% versus 56.6%) or BEAGLE (M = 56.4% versus 54.0%) vectors. The difference

[1]The matrix of word vectors is publicly available as an .rda file at http://www.lingexp.uni-tuebingen.de/z2/LSAspaces/

[2]Eight (5%) of the 160 words in the training lists and 15 (4.8%) of the 320 words in the test list were not included in the LSA and BEAGLE vectors. To solve the problem we selected category-appropriate semantically similar words as replacements.

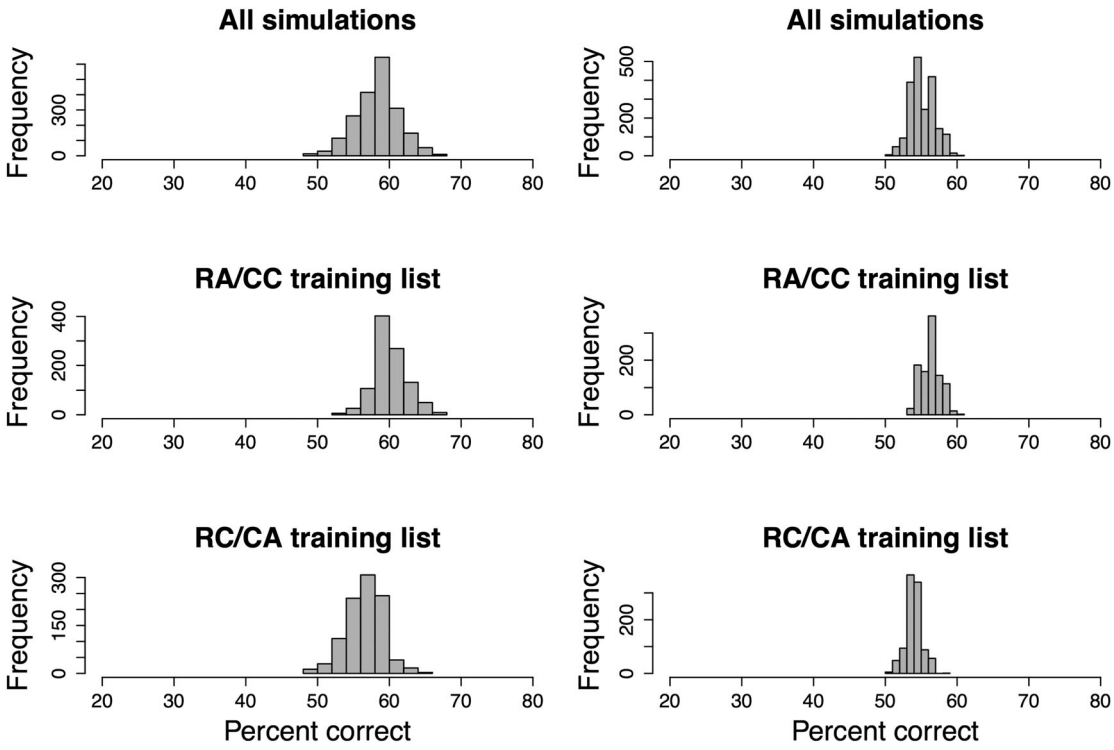

Figure 1. *Distributions of simulated percentage correct classification scores as a function of training list. Results using LSA vectors are shown on the left. Results using BEAGLE vectors are shown on the right.*

matches results in Neil and Higham's (2012) experiment, where participants who studied the RA-CC list ($M = 59.8\%$, $SD = 5.4\%$) outperformed participants who studied the RC-CA list ($M = 55.6\%$, $SD = 4.2\%$), $t(31) = 2.52$, $p = .017$.[3] Generally speaking, increasing L betters performance on this task, and decreasing L worsens it. However, decreasing L does not reduce discrimination below the level of chance, except at extremely small values ($L < .25$), and does not eliminate the advantage in test item discrimination after studying the RA-CC training list.

In summary, a standard exemplar-based model of memory discriminates rule-conforming from rule-violating test words without knowledge of the rule. The result holds whether words are represented using the LSA or the BEAGLE method.

GENERAL DISCUSSION

After reading words chosen according to a rule, participants can discriminate novel rule-consistent from rule-violating test words but cannot articulate the rule. The results are consistent with the idea that participants implicitly learned the rule. Indeed, this is the conclusion that Neil and Higham (2012) drew from their data. However, an exemplar model of memory that does not know the rule can reproduce the result. In short, the semantic structure of materials used in the experiment was correlated with the rule. Thus, when the model responded to semantics, it behaved *as if* it were responding to the rule. The computational demonstration shows that the data do not force the conclusion of implicit rule knowledge.

[3]Neil and Higham (2012) reported the number of participants in each study condition as approximately equal (p. 1396); therefore, our estimates are approximate assuming 17 participants in the RA-CC condition and 16 participants in the RC-CA condition.

We are by no means the first to use exemplar models to understand performance in implicit categorization tasks. Brooks (1978) promoted the principles of an exemplar model as an explanation for categorization in the judgement of grammaticality task, but never developed a formal model (see also Vokey & Brooks, 1992). Nosofsky's generalized context model (Nosofsky, 1987) has been applied to numerous results from implicit categorization tasks (e.g., Nosofsky & Johansen, 2000; Nosofsky & Zaki, 1998; Pothos & Bailey, 2000). To our knowledge, however, we are the first to combine exemplar-based processes with large-scale semantic representations in an account of performance in an implicit semantic categorization task.

Why does our model work? A word's properties, including concreteness and word frequency, are encompassed in the information captured by LSA and BEAGLE. Accordingly, vectors for training words selected using Neil and Higham's (2012) rule will, on average, be more similar to vectors for test words selected using the same rule. As a result, MINERVA 2's familiarity-based decisions perform the discrimination between rule-consistent and rule-discriminating words without knowledge of the rule.

In light of the model's success in reproducing Neil and Higham's (2012) results, to argue that participants implicitly learned the rule implies that LSA and BEAGLE tacitly and proactively sorted words into separate regions in semantic space so that a rule-based discrimination process could reflect the relevant properties. There is no mechanism in either LSA or BEAGLE to support this conclusion. Even assuming there were, MINERVA 2 has no mechanism with which to apply a conjunctive rule. The model computes familiarity and selects the alternative with the greater familiarity. Because words that share concreteness and frequency are more likely to be semantically related than words that do not, the model discriminates the test stimuli without knowing the conjunctive rule.

Neil and Higham (2012) designed their experiment to challenge current theories of implicit learning. As they pointed out, theories of implicit learning have focused on sequential structure and, therefore, are ill equipped to explain learning of non-sequential structure. They argued that the limitation follows naturally from a stereotyped reliance on the artificial grammar task and the design of models to accommodate data from that task. We have demonstrated that a classic global-similarity memory model—a model known to capture judgement of grammaticality in the artificial grammar task (e.g., Chubala & Jamieson, 2013; Jamieson & Mewhort, 2011)—also discriminates words in Neil and Higham's analogous rule-learning task.

Importantly, however, our simulations allow us to draw more forceful conclusions than Neil and Higham (2012) could draw from their data alone. Whereas they argued against sequential models for the process of implicit learning, we argue against implicit learning as a process altogether. For instance, Neil and Higham took self-report measures of the features that participants used to classify test words (see their Table 4). A majority of participants reported using word familiarity (62%) and word meaning (52%), followed closely by semantic category and word associations (48% each). Whereas these choices do not reflect the components of the conjunctive rule, they do reflect the semantic features captured by LSA and BEAGLE vectors, as well as the familiarity-based process by which MINERVA 2 accomplishes retrieval. Participants' behaviour in the experiment could be explained by either non-verbalizable learning of a conjunctive rule, or by a more general appreciation of semantic relationships. The same behaviour observed in our model supports only the latter explanation.

Although the MINERVA 2 approach explains both the artificial-grammar and semantic-rule tasks, it uses different representations for the two kinds of task. The fact that different representations are used follows from the task demands. As Jamieson and Mewhort (2011) argue, faced with nonsense letter strings, participants encode materials as best they can by parsing the strings into subunits and remember the strings as aggregations of those units. Faced with words, by contrast, participants encode meaning rather than subgroups

of letters. In both cases, a common set of retrieval processes accommodates performance within a given stimulus environment. Whether other approaches, such as auto-associative neural networks (e.g., Vokey & Higham, 2004), can also accommodate the results is an open and decidedly worthwhile question.

Dienes (1992) suggested that a competent model ought to capture not only participants' average performance on a task, but also the pattern of performance on individual test items (see also Jamieson & Mewhort, 2010). In the context of the semantic implicit learning task, a competent model ought to predict any systematic patterns of bias toward certain test words or word categories. That we limited our simulations to average performance should not be taken as a stand against Dienes's dictum. An item-level analysis on a much larger set of data presents a rational next step to more fully evaluate our model's performance in this task.

As it stands, we take the model's success as an echo of advice from Newell (1973). He argued that an understanding of human behaviour demands more than a list of phenomena paired with local explanations. Rather, it demands formal theory to approach a wide range of phenomena from a more global perspective, in which common principles are shown to accommodate diverse datasets (see also Simon, 1956; Surprenant & Neath, 2009). The current paper contributes to this overarching goal and suggests a fruitful direction for future contributions: the merging of semantic representations with basic processes of memory.

REFERENCES

Berry, D. C., & Broadbent, D. E. (1984). On the relationship between task performance and associated verbalizable knowledge. *The Quarterly Journal of Experimental Psychology Section A*, *36*, 209–231.

Brooks, L. R. (1978). Nonanalytic concept formation and memory for instances. In E. Rosch & B. Lloyd (Eds.), *Cognition and Categorization* (pp. 3–170). Hillside, NJ: Lawrence Elbaum Associates.

Chubala, C. M., & Jamieson, R. K. (2013). Recoding and representation in artificial grammar learning. *Behavior Research Methods*, *45*, 470–479.

Dienes, Z. (1992). Connectionist and memory-array models of artificial grammar learning. *Cognitive Science*, *16*, 41–79.

Dienes, Z., Altmann, G. T. M., & Gao, S. (1999). Mapping across domains without feedback: A neural network model of transfer of implicit knowledge. *Cognitive Science*, *23*, 53–82.

Günther, F., Dudschig, C., & Kaup, B. (2015). LSAfun - An R package for computations based on Latent Semantic Analysis. *Behavior Research Methods*, *47*, 930–944.

Higham, P. A., & Brooks, L. R. (1997). Learning the experimenter's design: Tacit sensitivity to the structure of memory lists. *The Quarterly Journal of Experimental Psychology*, *50*, 199–215.

Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review*, 93, 411–428.

Jamieson, R. K., & Mewhort, D. J. K. (2009). Applying an exemplar model to the serial reaction-time task: Anticipating from experience. *The Quarterly Journal of Experimental Psychology*, *62*, 1757–1783.

Jamieson, R. K., & Mewhort, D. J. K. (2010). Applying an exemplar model to the artificial-grammar task: String completion and performance on individual items. *The Quarterly Journal of Experimental Psychology*, *63*, 1014–1039.

Jamieson, R. K., & Mewhort, D. J. K. (2011). Grammaticality is inferred from global similarity: A reply to Kinder (2010). *The Quarterly Journal of Experimental Psychology*, *64*, 209–216.

Jones, M. N., & Mewhort, D. J. K. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review*, *114*, 1–37.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*, 211–240.

Neil, G. J., & Higham, P. A. (2012). Implicit learning of conjunctive rule sets: An alternative to artificial grammars. *Consciousness and Cognition*, *21*, 1393–1400.

Newell, A. (1973). You can't play 20 questions with nature and win: Projective comments on the papers of this symposium. In W. G. Chase (Ed.), *Visual*

information processing (pp. 283–308). New York, NY: Academic Press.

Nosofsky, R. M. (1987). Attention and learning processes in the identification and categorization of integral stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*, 87–108.

Nosofsky, R. M., & Johansen, M. K. (2000). Exemplar-based accounts of "multiple-system" phenomena in perceptual categorization. *Psychonomic Bulletin and Review*, *7*, 375–402.

Nosofsky, R. M., & Zaki, S. R. (1998). Dissociations between categorization and recognition in amnesic and normal individuals: An exemplar-based interpretation. *Psychological Science*, *9*, 247–255.

Pothos, E. M., & Bailey, T. M. (2000). The role of similarity in artificial grammar learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26, 847–862.

Recchia, G., Sahlgren, M., Kanerva, P., & Jones, M. N. (2015). Encoding sequential information in semantic space models: Comparing holographic reduced representation and random permutation. *Computational Intelligence and Neuroscience*, *2015*, Article 986574, 1–18.

Servan-Schreiber, E., & Anderson, J. R. (1990). Learning artificial grammars with competitive chunking. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*, 592–608.

Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review*, *63*, 129–138.

Surprenant, A. M., & Neath, I. (2009). *Principles of Memory*. New York, NY: Psychology Press.

Vokey, J. R., & Brooks, L. R. (1992). Salience of item knowledge in learning artificial grammars. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*, 328–344.

Vokey, J. R., & Higham, P. A. (2004). Opposition logic and neural network models in artificial grammar learning. *Consciousness and Cognition*, *13*, 565–578.