Modeling Instance-Based Rule Learning in an Adaptive Retrieval Practice Task

Thomas Wilschut (t.j.wilschut@rug.nl)[1], Florian Sense[2], Myrthe Braam [3], Hedderik van Rijn[1]
[1]Department of Experimental Psychology, University of Groningen; [2]Infinite Tactics, LLC;
[2]MemoryLab BV

Abstract

Model-based adaptive learning systems have successfully improved the efficiency of fact learning in educational practice. Typically, such systems work by keeping track of a learner's memory processes by measuring behavior during learning, and using this information to tailor the learning process towards the needs of individual learners. Where many adaptive learning systems applied today focus on learning paired associates, we here focus on learning grammar rules based on instances of these general rules. We show that participants' (N = 42) behavioral responses on instance questions for a rule can be used to infer general performance on other questions associated to that rule, and that we can capture this rule performance in a single model-based *speed of forgetting* parameter. These findings could be used to develop and optimize adaptive learning systems that can be used to study general rules from instances.

Keywords: ACT-R; Adaptive Learning; Knowledge Tracing; Instance-Based Learning; Grammar

Introduction

Adaptive learning systems have successfully improved the process of memorizing factual information, such as vocabulary or glossary items, by tailoring learning schedules to the needs of individual learners. Typically, such systems aim to predict learner performance from behavioral measures that are recorded during learning, and use these predictions to tailor item repetition schedules towards the needs of individual learners (e.g., presenting fewer or easier items when predicted performance is low; and presenting more or more difficult items when predicted performance is high). This approach has proven to increase learning efficiency compared to traditional, less adaptive approaches in a wide range of materials, both in laboratory and classroom settings (e.g., see Lindsey, Shroyer, Pashler, & Mozer, 2014; Papousek, Pelánek, & Stanislav, 2014; Wozniak & Gorzelanczyk, 1994; Van Rijn, Van Maanen, & Van Woudenberg, 2009).

Existing adaptive learning systems are typically used to learn paired associates, such as vocabulary items or glossary items. For these materials, there is extensive evidence supporting the idea that it is possible to use behavioral proxies, recorded during the learning session, to infer the extent to which a learner has successfully memorised a specific paired-associate item. Most model-based adaptive learning systems present the learner with retrieval practice questions, and use response accuracy as a behavioral proxy of the extent to which an item is stored in memory (e.g., see Pavlik & Anderson, 2008; Van Rijn et al., 2009). As using accuracy scores only does not allow for meaningful discrimination within correct responses, (and as a consequence, accurate performance predictions require many incorrect responses,) many systems use response times in addition to accuracy scores to predict performance (Byrne & Anderson, 1998; Sense, Behrens, Meijer, & van Rijn, 2016, see). Finally, in recent implementations, information carried in the speech signal during spoken retrieval attempts has been used to infer the extent to which a learner has successfully memorised a specific paired-associate item (Wilschut, Sense, & van Rijn, 2024). Overall, for learning paired-associate items, there is extensive support for the idea that behavioral responses to retrieval practice questions can be used to infer model parameters that map on to latent memory processes.

A popular framework used in model-based adaptive learning systems is the ACT-R model of human declarative memory (Anderson et al., 2004). In ACT-R, learners' memory representations for individual facts are stored as *chunks* in declarative memory. Chunks are schematic units of information that possess an activation value: More active chunks are more likely to be retrieved during a search of declarative memory. Arguably, a limitation of this model is that it treats individual facts as independent units of information. As such, it is not straightforward to model a learner's memory for facts that are clustered or related to other facts that have been encountered in the learning session (although accounts of *spreading activation*, in which activation spreads through a semantic network, could account for such context effects (e.g., see Anderson, 1983; Thomson, Bennati, & Lebiere, 2014)).

Although model-based adaptive learning systems have proven to be successful in improving the efficiency of learning paired associates, it is unclear to what extent these findings generalise to situations where facts are not independent from each other (i.e., where the clustering of items plays an important rule). In this research project, we aim to extend existing adaptive learning models that keep track of memory performance for simple paired associates by modeling a learner's mastery of general/underlying rules from instances of that rule (i.e. instance-based learning, see Lejarraga, Dutt, & Gonzalez, 2012). If adaptive learning models are able to keep track of a learner's mastery of a common rule based on responses to instance questions, this would widen the scope of such systems and their possible application in a wide range of educational settings. For example, current teaching methods for learning language grammar rules, mathematics, physics or chemistry all heavily rely on teaching students to pick up regularities or general rules from instances.

There have been several successful attempts at modeling the process of learning common rules or patterns from a set of instances. For example, Stevens et al., 2018 showed that it is possible to model others' decisions from instances in a negotiation task. Instance-based rule learning models have also been made for learning the English past tense (Taatgen & Anderson, 2002), the German plural (Taatgen, 2001), as well as for other domains, such as the balanced-scale task (Van Rijn, Van Someren, & Van der Maas, 2003). Finally, within the context of ACT-R, studies have focused on using instance-based learning to explain human decision making (e.g., see Gonzalez, Dutt, & Lebiere, 2013). Yet, the above approaches all aim to model *inductive* rule learning from instances. In other words, the rule is never explicitly given to the learners. In the current work, we intend to explicitly provide feedback explaining the rule after each instance question, with the intention that the learners remember the rule, and recognize future instance questions that are associated to the same rule. To our knowledge, we are the first to model instance based rule learning in this exact setup.

In this project, we aim to explore if we can model instance-based rule learning in an adaptive retrieval practice task, where participants study Dutch grammar rules from specific instances. We specifically aim to track a learner's mastery of underlying rules, and therefore model these rules, and not the instances, as chunks in the memory model. We first examine the extent to which performance on instance questions for specific questions is associated to (a) other instance questions for the same grammar rule during learning and (b) new instance questions presented on a test following the learning session. Second, we will examine if we can use model-based estimates of speed of forgetting during learning to predict performance on the test. Finally, we aim to show that using a fully adaptive, model-based item scheduling algorithm—that takes both a learner's accuracy scores and response times into account to determine the most optimal item repetition schedule for each individual learner—can be used to successfully improve learning efficiency.

Methods

Participants

In total, 42 participants completed the experiment via the online participant pool *Prolific*. Participants were included if they had at least completed secondary education. Most participants had completed education at a university of applied science ('HBO'). In addition, they were required to speak Dutch fluently. Finally, participants were included only if they had completed at least 10 other Prolific studies prior to the current experiment. The mean age of the participants was 35 years, 18 participants identified as female and 24 participants as male.

Design and Procedure

The study consisted of two learning blocks and a test block, which were completed by all participants in a single session. All participants started with the learning blocks, which consisted of one rt-adaptive learning block and one stack-based learning block. Half of the participants (n = 21) started with the rt-adaptive learning block, and completed the stack-based block second. For the other half of the participants (n = 21), this order was reversed. After the learning blocks, a test followed.

Figure 1: Screenshots of the learning application. **a** shows a trial: the learner is asked to type the correct word **b** shows the feedback when an incorrect answer is given: the correct answer is provided along with the explanation of the underlying grammar rule.

During the learning blocks, participants studied Dutch grammar rules based on instance questions (see Materials). For each grammar rule, there were six instances/instance questions, that were randomly chosen to be presented to the learner (with the only exception that a specific instance question would not be repeated twice in a row). Participants were prompted with a request (e.g., '*write the plural form of the word between brackets*') and a context sentence in which the target word occurred. Participants were asked to type the item in correct spelling. If the answer was correct, a short feedback screen appeared prompting the participants that the answer was correct, after which the next item was presented. If the answer was incorrect, the correct answer, as well as an explanation of the associated grammar rule, were presented to the participants. The feedback screen after incorrect responses was self-paced: Participants were able to click 'next' to continue at their own pace. Response times were defined as the time elapsed between the presentation of the cue and the first keypress, unless the participants used the backspace button to erase the first keypress (in which case the response time would not be used). Both the stack-based learning block and the rt-adaptive learning block took 12 minutes in total. In the rt-adaptive scheduling block, rule repetition schedules were personalised based on the accuracy and response times that were recorded during the learning session (see Rule scheduling); in the stack-based learning block, rules were scheduled based on accuracy only (see Rule scheduling).

On the test, one new instance question was presented to the participants for each grammar rule, where one rule was presented at the time. During the test, response times were not recorded.

Materials

The materials for this study were generated in collaboration with *Hogeschooltaal* (see `https://www.hogeschooltaal.nl/?lang=en`, a Dutch institution facilitating the process of language proficiency development in Dutch applied university students. The total material set consisted of 18 grammar rules, for each of which there were seven instance questions. Six instance questions were used in the learning session, one instance question was used for the test. All participants saw the same questions on the test. The list of 18 grammar rules was split in two sets of nine rules, which were then randomly assigned to a specific scheduling block (rt adaptive or stack-based) for each participant.

Rule scheduling

In the rt-adaptive scheduling block, we used an adaptive algorithm to schedule rule repetitions in a way that is optimally tailored towards the individual learner. This adaptive algorithm is based on an ACT-R model of declarative memory (Anderson et al., 2004), and is described in more detail in Sense et al. (2016). In the current application, individual grammar rules—not individual instances—are stored as chunks in the declarative memory model. The algorithm aims to model the memory strength or activation of each to-be-learned grammar rule over time, and presents rules to the learner for retrieval practice whenever their activation decays to a threshold value. Activation values are continually updated using the learner's response times and accuracy scores.

In practice, this means that instances for which a learner gives slow and/or incorrect answers, activation values are adjusted downwards and rules are repeated more frequently, whereas if the learner gives quick and correct answers to a retrieval practice question, the activation will be adjusted upwards, and presented for practice less frequently. In addition to personalising the rule repetition schedule, the algorithm captures individual differences in ability through a learner- and item-specific *speed of forgetting* parameter (α), which it estimates from the learner's responses. Poorer learners will have a higher speed of forgetting value, which causes activation to decay faster, leading to more frequent repetition.

In the stack-based learning block, the rule repetition schedule was determined by a Leitner-inspired stack-based system (Mubarak & Smith, 2008), which groups words into virtual boxes: All words start in Box 1 and move to the next box if answered correctly. If a word is answered incorrectly, it moves back to the previous box. Words in Box 1 are presented first, followed by words in Box 2, followed by words in Box 3. If all rules are in Box 3 (and if they are all answered correctly) the rules are repeated in the order of first presentation until the learning time is over. This stack-based system allows for difficult rules to be rehearsed more often than easy rules and is a frequently used and effective study strategy (Bryson, 2012).

Analyses

Analyses were conducted in R 3.4.1 (R Core Team, 2020), with the mixed-effects modelling package lme4 1.1-28 (Bates, Mächler, Bolker, & Walker, 2015). The mixed effects models reported in this study include rule repetition, scheduling algorithm (contrast coded: rt-adaptive learning = 0; stack based learning = 1) and speed of forgetting. In all models, participant- and rule id were added as random intercepts (Baayen, Davidson, & Bates, 2008). The data was visualised using ggplot2 (Wickham, 2016).

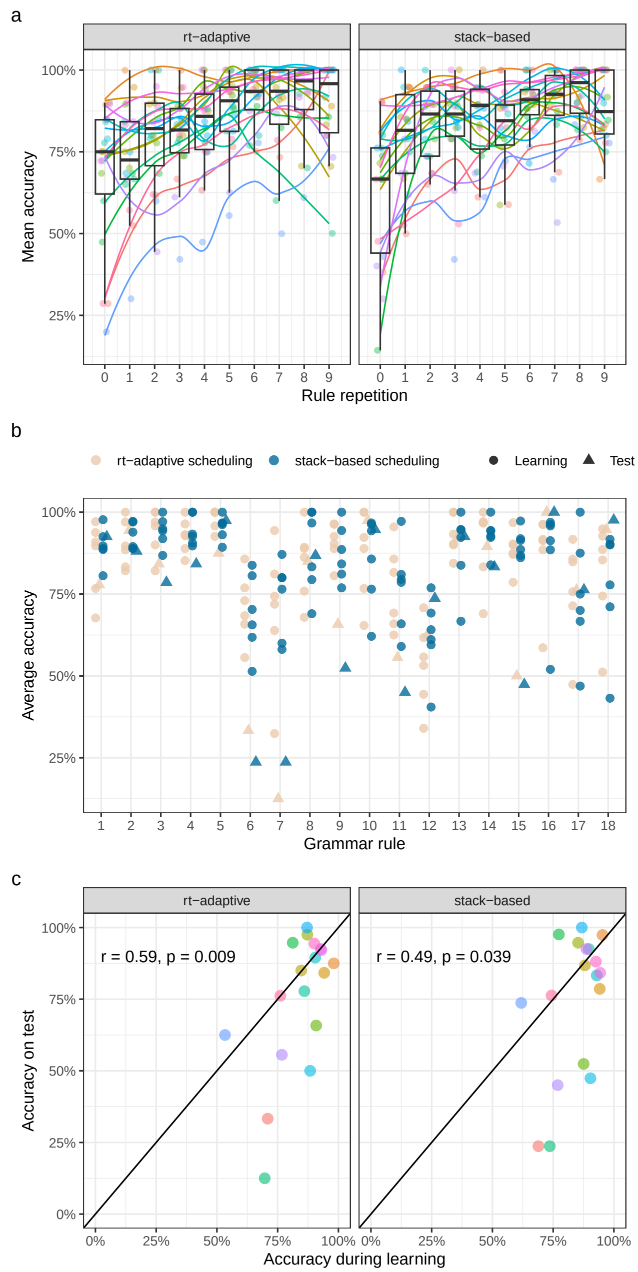

Figure 2: Inferring grammar rule mastery from instance-based learning behavior. Different colors represent unique grammar rules. **a** shows the mean accuracy over repetitions of a grammar rule, split by scheduling algorithm. Dots represent aggregate performance over randomly introduced instance questions for each rule. The graph shows that rule difficulty can be relatively reliably inferred from average scores on specific instances of that rule. For simplicity, only the first six of 18 grammar rules are shown here. **b** shows the main accuracy for each individual instance question during learning (dots) and associated test performance (triangles). **c** shows the association between mean accuracy during learning on instance questions associated to specific rules and accuracy during the test on new instance questions for the same rules.

Results[1]

Inferring grammar rule performance from instance-based learning behavior

The first aim of this research project was to examine the extent to which it is possible to use behavioral responses to instance questions to infer a learner's mastery of an underlying grammar rule. Figure 2a shows the mean accuracy on grammar rule questions over repetitions, split by scheduling algorithm. Colored lines represent individual grammar rules, and dots show average scores on repetitions of each rule, which are based on aggregates over instance questions. The figure clearly shows that it is possible to distinguish trends of rule difficulty: some grammar rules, aggregate accuracy scores over instance questions are lower than for other grammar rules. Correspondingly, for some grammar rules, initial performance is very low (close to 0), whereas for other rules, initial performance is quite high. Finally, the plot shows a trend of learning over repetitions (i.e., on average, accuracy increases over repetitions).

Figure 2b shows performance on individual instance questions for grammar rules. There is considerable variation between instance questions, but it seems reasonable to determine overall rule difficulty from a few observations of individual instances. Figure 2c shows that there is a strong positive association between average accuracy for grammar rules during learning and accuracy on new instance questions for the same grammar rules on the following test, both in the rt-adaptive scheduling block ($r = 0.59$, $p = 0.009$) and in the stack-based scheduling block ($r = 0.49$, $p = 0.039$).

Mixed effects models M1 and M2 (see Table 1 describe the effects of repetition, scheduling system, and their interaction on learning accuracy and response times, respectively. We found only significant main effects of repetition: participants became more accurate and responded faster over repetitions of a rule, regardless of the rule scheduling algorithm and despite the fact that rule repetitions consisted of randomly chosen instance questions. The effects of rule scheduling algorithm were not statistically significant. Overall, behavioral responses on instance questions for grammar rules seem to be indicative of performance on other instance questions that are associated to the same rule, both during the learning sessions and on the test that follows learning.

Model-based estimations of test performance

The second aim of this project was to capture rule mastery in a model-based *speed of forgetting* parameter. Figure 3a shows the mean estimated speed of forgetting over repetitions of grammar rules, based on instance questions for each rule. With each rule repetition, the estimated speed of forgetting was updated based on the accuracy and response time of the learner's answer (see Rule scheduling).

[1] Analysis code, data, and materials are available from https://osf.io/grdmw/.

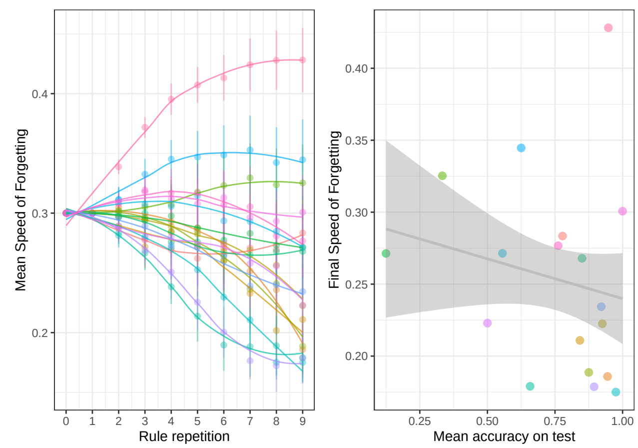

Figure 3: Estimating test performance based on model-inferred speed of forgetting for grammar rules. **a** shows the mean estimated speed of forgetting for each individual grammar rule, based on accuracy scores and response times for instance questions. Error bars represent (+/-) 1 standard error of the mean. **b** shows the mean test accuracy as a function of final speed of forgetting.

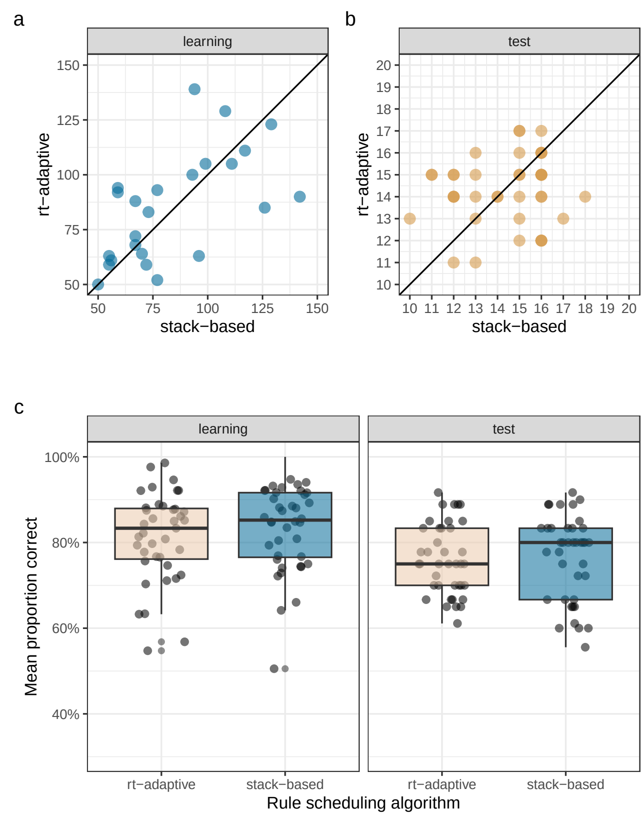

Figure 4: Performance during learning and test. **a** and **b** show the number of rules correctly recalled during learning and on test, respectively. Dots represent average scores for individual participants. **c** shows the proportion of correct responses during learning and on test for both rule scheduling algorithms.

Mixed effects models explaining performance during learning and on test from rule repetition and scheduling algorithm
*** = $p < 0.001$; ** = $p < 0.01$; * = $p < 0.05$.

M1. Accuracy during learning	*Estimate*	*SE*	*z*	*p*
Intercept	1.37	0.23	5.87	<0.001***
Rule repetition	0.08	0.01	8.85	<0.001***
Scheduling algorithm (stack-based = 1)	−0.06	0.01	−0.56	0.576
Rule repetition * Scheduling algorithm	0.018	0.01	1.28	0.212
M2. Reaction times during learning	*Estimate*	*SE*	*t*	*p*
Intercept	7711.91	1473.33	5.23	<0.001***
Rule repetition	−195.40	34.22	−5.71	<0.001***
Scheduling algorithm	−562.45	362.16	−1.55	0.121
Rule repetition * Scheduling algorithm	74.72	44.99	1.66	0.097
M3. Accuracy on test	*Estimate*	*SE*	*z*	*p*
Intercept	1.14	0.42	2.69	0.007
N. repetitions during learning	0.01	0.00	2.57	0.010*
Scheduling algorithm	−0.13	0.26	−0.50	0.616
Rule repetition * Scheduling algorithm	0.00	0.01	0.52	0.601

Estimating learning and test performance from rule repetitions and model-based speed of forgetting
*** = $p < 0.001$; ** = $p < 0.01$; * = $p < 0.05$.

M4. Accuracy during learning	*Estimate*	*SE*	*z*	*p*
Intercept	1.46	0.23	6.43	<0.001***
Speed of Forgetting	−0.84	0.21	−4.09	<0.001***
Rule Repetition	0.10	0.01	12.91	<0.001***
M5. Mean accuracy on test	*Estimate*	*SE*	*t*	*p*
Intercept	0.96	0.11	8.51	<0.001***
Final Speed of Forgetting	−0.76	0.31	−2.45	0.01*
N. repetitions during learning	0.00	0.00	0.71	0.47

Figure 3b shows the association between the final estimated speed of forgetting for a grammar rule during the learning session, and the mean accuracy during test for new instances of the same rules. We found that, overall, grammar rules for which a high speed of forgetting was estimated during learning, new instance questions were answered with lower accuracy on the test, indicating that the adaptive learning model could track rule performance during the learning session to estimate later test performance. Mixed effects models M4 and M5 (see Table 2) support these interpretations, as they show that the speed of forgetting for a grammar rule, estimated during learning based on responses to instance questions, can be used to estimate accuracy during learning, and on test, respectively.

Model-based optimization of learning

The final aim of this project was to explore the possibility of using the instance-based estimations of speed of forgetting for grammar rules to optimize repetition schedules, ultimately leading to a higher learning efficiency of grammar rules. To that end, we compared the learning efficiency with a fully adaptive scheduling algorithm, that uses both accuracy scores and response times to predict rule performance, to a stack-based rule scheduling algorithm that is based on the accuracy of rule instances only (see Rule scheduling). Figure 4a and 4b show the number of correct responses during learning and on test, respectively. Figure 4c shows the proportion of correct answers during learning and test with both scheduling algorithms. As is also supported by the mixed effects models M1 and M3 (see Table 1), we found no significant difference between using the stack-based scheduling system and the rt-adaptive scheduling system.

Discussion

In this study, we aimed to extend existing adaptive learning models that can keep track of memory performance for simple paired associate stimuli to estimating mastery of grammar rules, based on responses to randomly introduced instance questions. The results can be summarized in three main points. First, we examined the possibility of inferring rule performance from behavioral responses on randomly chosen instance questions. Our results suggest that, despite the fact that we found considerable variation in performance on individual instance questions within a rule, it seems sensible to keep track of a learner's mastery of a grammar rule using the behavioral responses on randomly chosen instance questions. More specifically, accuracy scores on instance questions for grammar rules were indicative of performance on other instance questions that are associated to the same rule, both during the learning sessions and on the test that follows learning.

Second, we showed that the adaptive learning model, that was originally developed to keep track of memory performance of individual paired associate rules, could capture the extent to which learners have mastered a grammar rule in a single speed of forgetting parameter. A higher speed of forgetting during learning for a specific grammar rule was associated with poorer memory performance for new instances of the same grammar rule.

Finally, we attempted to use the model-based estimations of a learner's performance of a certain rule by optimizing the repetition schedule for individual learners. We found that, despite the fact that our learning model could capture differences between grammar rules, model-based optimization of the rule repetition schedule did not lead to better learning performance compared to a more simple, stack-based adaptive learning system.

There are several possible reasons for the lack of a benefit of using the model-based rt-adaptive scheduling algorithm compared to using a stack-based accuracy adaptive system. First, in the current system, response times were defined as the time elapsed between the first presentation of an instance question and the first keypress of the response. The underlying rationale is that this response time mainly reflects *retrieval time*, and can therefore be used as a proxy of the memory strength for a specific rule (Byrne & Anderson, 1998). This way of measuring response latencies has proved to be effective for paired associate learning, but it is possible that response times should be decomposed more carefully when it comes to grammar rule learning. For instance, future research should examine whether the non-retrieval time (i.e., the time needed to process a question before retrieval takes place, or the time needed to prepare a response after retrieval has taken place) can be subtracted from the response times before being taken into account to determine scheduling for more complex materials such as grammar rule learning. Second, it is possible that the current experimental setup was not sensitive enough to statistically detect differences in learning efficiency between the accuracy-adaptive stack-based and the model-based, rt-adaptive scheduling system. Future studies should further examine this issue, in particular over multiple learning sessions and including longer-term retention tests.

Another possible direction for future studies is taking a data-driven approach of clustering items, rather than defining the common grammar rules beforehand. A post-hoc *k*-means clustering analysis of the current dataset suggests that only 5–7 clusters is enough to accurately describe the variability of performance on instance questions. In other words, learners performed very similar on some of the grammar rules, which makes the usefulness of treating them as separate knowledge chunks questionable. As in some situations it might be difficult to establish the most optimal common rule clustering upfront, it may be worthwhile exploring methods to use a data-driven approach to group items for individual learners in real time, and then track a learner's progress on each group of items.

Another important point that has received little attention in the current work concerns the explanatory feedback about the grammar rules that was shown to the learner after each incorrect response. Future work should examine the consequences of providing explanations of grammar rules after each response, and how the time taken to study these rules during the feedback moments impacts learning efficiency.

Despite these open questions, we show that it is possible to model learners' mastery general rules from answers to instance questions, and that we can use this information to optimize rule repetition schedules. These results demonstrate that—in the context of learning Dutch grammar rules—it is sensible to use performance on instance questions to infer a learner's mastery of the underlying rule. Despite the fact that our current attempts at using this information to personalise the repetition schedule did not result in increased learning efficiency, our results indicate that it is sensible to track rule performance from responses on corresponding instance questions. These findings underline the need to further investigate possible ways of using this information to improve repetition schedules for these rules. Ultimately, this could lead to learning systems that allow for instance-based rule learning, adapted to the needs and prior knowledge of individual learners.

Conclusion

In this project, we asked participants to study Dutch grammar and spelling rules through exposure to specific instances of each rule. We show that it is possible to use the learner's answers to instance questions to estimate their performance on new instances of the same rules. Using a cognitive model of memory retrieval, we show that we can estimate how well learners have memorized the rules. Although future research should explore how these estimations of a learner's rule performance can be exploited to increase learning efficiency, these results pave the way for the development of adaptive learning applications that allow for rule learning based on instances.

References

Anderson, J. R. (1983). A spreading activation theory of memory. *Journal of verbal learning and verbal behavior*, *22*(3), 261–295.

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological review*, *111*(4), 1036.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of memory and language*, *59*(4), 390–412.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Bryson, D. (2012). Using flashcards to support your learning. *Journal of visual communication in medicine*, *35*(1), 25–29.

Byrne, M. D., & Anderson, J. R. (1998). Perception and action. *The atomic components of thought*, *16*, 23–28.

Gonzalez, C., Dutt, V., & Lebiere, C. (2013). Validating instance-based learning mechanisms outside of act-r. *Journal of Computational Science*, *4*(4), 262–268.

Lejarraga, T., Dutt, V., & Gonzalez, C. (2012). Instance-based learning: A general model of repeated binary choice. *Journal of Behavioral Decision Making*, *25*(2), 143–153.

Lindsey, R. V., Shroyer, J. D., Pashler, H., & Mozer, M. C. (2014). Improving students' long-term knowledge retention through personalized review. *Psychological science*, *25*(3), 639–647.

Mubarak, R., & Smith, D. C. (2008). Spacing effect and mnemonic strategies: A theory-based approach to e-learning. In *e-learning* (pp. 269–272).

Papousek, J., Pelánek, R., & Stanislav, V. (2014). Adaptive practice of facts in domains with varied prior knowledge. In *Educational data mining 2014* (pp. 6–13).

Pavlik, P. I., & Anderson, J. R. (2008). Using a model to compute the optimal schedule of practice. *Journal of Experimental Psychology: Applied*, *14*(2), 101.

R Core Team. (2020). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Sense, F., Behrens, F., Meijer, R. R., & van Rijn, H. (2016). An individual's rate of forgetting is stable over time but differs across materials. *Topics in cognitive science*, *8*(1), 305–321.

Stevens, C. A., Daamen, J., Gaudrain, E., Renkema, T., Top, J. D., Cnossen, F., & Taatgen, N. A. (2018). Using cognitive agents to train negotiation skills. *Frontiers in psychology*, *9*, 154.

Taatgen, N. A. (2001). Extending the past-tense debate: A model of the german plural. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 23).

Taatgen, N. A., & Anderson, J. R. (2002). Why do children learn to say "broke"? a model of learning the past tense without feedback. *Cognition*, *86*(2), 123–155.

Thomson, R., Bennati, S., & Lebiere, C. (2014). Extending the influence of contextual information in act-r using buffer decay. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 36).

Van Rijn, H., Van Maanen, L., & Van Woudenberg, M. (2009). Passing the test: Improving learning gains by balancing spacing and testing effects. In *Proceedings of the 9th international conference of cognitive modeling* (Vol. 2, pp. 7–6).

Van Rijn, H., Van Someren, M., & Van der Maas, H. (2003). Modeling developmental transitions on the balance scale task. *Cognitive Science*, *27*(2), 227–257.

Wickham, H. (2016). *ggplot2: elegant graphics for data analysis*. springer.

Wilschut, T., Sense, F., & van Rijn, H. (2024). Speaking to remember: model-based adaptive vocabulary learning using automatic speech recognition. *Computer Speech & Language*, *84*, 101578.

Wozniak, P. A., & Gorzelanczyk, E. J. (1994). Optimization of repetition spacing in the practice of learning. *Acta neurobiologiae experimentalis*, *54*, 59–59.