Revisiting Half-Life Regression for Explainable and Personalized Spaced Repetition in Duolingo

Pedro Ilídio* PEDRO.ILIDIO@KULEUVEN.BE
Itec, imec research group at KU Leuven, Belgium
Department of Public Health and Primary Care, KU Leuven, Belgium
Achilleas Ghinis* ACHILLEAS.GHINIS@KULEUVEN.BE
Itec, imec research group at KU Leuven, Belgium
Department of Public Health and Primary Care, KU Leuven, Belgium
Sameh Said Metwaly SAMEH.METWALY@KULEUVEN.BE
Itec, imec research group at KU Leuven, Belgium
Faculty of Psychology and Educational Sciences, KU Leuven, Belgium
Celine Vens CELINE.VENS@KULEUVEN.BE
Itec, imec research group at KU Leuven, Belgium
Department of Public Health and Primary Care, KU Leuven, Belgium
Frederik Cornillie FREDERIK.CORNILLIE@KULEUVEN.BE
Itec, imec research group at KU Leuven, Belgium
Linguistics Research Unit, Faculty of Arts, KU Leuven, Belgium
Alireza Gharahighehi ALIREZA.GHARAHIGHEHI@KULEUVEN.BE
Itec, imec research group at KU Leuven, Belgium
Department of Public Health and Primary Care, KU Leuven, Belgium

Abstract

Spaced repetition is a technique widely used in digital learning applications (notably for language practice) that determines the frequency with which concepts need to be reviewed by learners in order to boost long-term retention, scheduling reviews at increasing intervals. Research in the area of AI in education investigates how data from digital learning applications can be leveraged to develop predictive models that enable personalization of spaced repetition, and as such maximize learning efficiency. Previous research on Duolingo's digital platform for language practice developed a personalized model (the Half-Life Regression model) that predicts when lexemes (abstract units of mearning) are likely to be forgotten by individual learners. A limitation of previous research is its reliance on a small set of features, which restricts the performance of the original model. Moreover, attempts to improve this performance often rely on deep learning methods, which in turn reduce the model's explainability. In the current study, we propose an alternative approach (combining psychometrics with machine learning) that incorporates relevant learner, lexeme and contextual features (e.g., time interval since the lexeme was last shown) aimed at improving both the performance and explainability of earlier Half-Life Regression models. More specifically, we introduce a mixed-effects version of Half-Life Regression, and use the estimated random effects as input to a Random Forest. In comparison with the original model, our results show improvements in predictive performance across a range of measures. SHAP analysis of the best-performing method (the Random Forest) indicates the importance of contextual features and random-effects features that are likely to reflect aspects like lexeme difficulty and learner ability. By achieving both explainability and higher predictive performance than the earlier models, our work contributes to more impactful and responsible use of AI for personalized spaced repetition in language learning.

Keywords: spaced repetition, half-life regression, language learning, AI in education

* These two authors contributed equally.

1. Introduction

Systematic and deliberate practice of a second or foreign language (L2) can be instrumental for helping learners to develop automaticity of L2 skills. Automaticity is key for communicating fluently and effectively in an L2 (Suzuki, 2023), an essential competence for education, work and life. Digital applications for such practice (e.g., online platforms or mobile applications for computer-assisted language learning; CALL) can be a cornerstone of effective environments for L2 learning and teaching, when designed thoughtfully on the basis of established findings in cognitive psychology and applied linguistics (Cornillie et al., 2026), and when the affordances of AI are harnessed responsibly for personalizing L2 learning, which have been shown to benefit L2 learning outcomes both at the cognitive level (e.g., gains in vocabulary knowledge) and non-cognitive level (e.g., increased self-efficacy) (Chen et al., 2021). However, as learners typically practice L2 knowledge and skills through digital applications on their own (in the absence of teachers), developers of digital applications are seeking ways to keep learners engaged through nudging and gamification techniques, with a view to maximizing learning efficiency (Cornillie et al., 2026).

Duolingo is the most widely used digital application for L2 practice globally, enabling learners to autonomously study more than 40 languages through gamified exercises that target the development of both knowledge (e.g., vocabulary) and skills (e.g., speaking) in the L2. The application incorporates several personalization mechanisms. For example, it allows learners to set daily practice goals, and varies the frequency with which vocabulary items are reviewed to support long-term retention (Settles and Meeder, 2016). With a view to improving the learning effectiveness of its applications, the Duolingo company invests quite intensively in research and development on AI in education (AIEd), among other through shared tasks around its datasets of learner logs – this makes Duolingo the first company to deploy an Intelligent CALL (Schulze, 2008) application at scale and simultanously fuel research on student modeling.

An active research area in the field of AIEd that brings together the need for systematic and deliberate practice in an L2 with the affordances of AI for personalized learning concerns the concept of spaced repetition. Spaced repetition refers to a family of techniques in which concepts to be learned are reviewed at increasing intervals rather than through cramming, thereby boosting long-term retention. One of the earliest methods for spaced repetition was developed by Pimsleur (1967), which prescribes fixed (i.e. non-personalized), exponentially increasing review intervals. Leitner (1972) introduced another well-known method that is widely implemented in flashcard applications. Although Leitner introduces partial adaptivity, where intervals are adjusted based on learner performance (e.g. correctly recalling a vocabulary item), it is also not fully personalized, as scheduling rules remain largely fixed.

To achieve more personalized practice schedules, researchers at Duolingo proposed the Half-Life Regression (HLR) model (Settles and Meeder, 2016). In this framework, a lexeme (i.e., an abstract unit of meaning) is scheduled for review when the estimated probability of recalling it falls below a threshold, indicating that the learner is likely to forget it. HLR is a parametric logistic regression model grounded in the exponential relationship (Ebbinghaus, 1913) between the probability of correctly recalling an item, the time elapsed since it was last practiced, and the learner's long-term memory strength (half-life).

Several studies have attempted to improve HLR in various directions. Some studies have reformulated the HLR model using different neural architectures, for example, a feed-forward neural network with one hidden layer, referred to as Neural Half-Life Regression (Zaidi et al., 2020), a gated recurrent unit-based network (Ye et al., 2022), and a model incorporating attention mechanisms (Ma et al., 2023). Beyond changes in modeling approaches, some studies have focused on expanding the feature space, most notably by including item difficulty (Zaidi et al., 2020; Ye et al., 2022). Although these studies have advanced the HLR model from different perspectives, they rely on a limited set of features, which restricts adaptation across multiple criteria (Gharahighehi et al., 2024), and on deep learning approaches that reduce explainability. The lack of explainability is a significant limitation for both users (learners and teachers) and developers of digital applications that implement AI-driven spaced repetition, preventing them from knowing why certain predictions and recommendations are made. Such explanations would not only help learners to get a better insight into their learning process but also would provide an overall picture of the learners' behavior to be explored during application design.

Therefore, the current study aims to achieve both high performance and explainability of algorithms for personalized spaced repetition in L2 practice, by constructing a richer and more pedagogically-informed feature space, and subsequently training parametric and non-parametric models to predict recall of lexemes. Our method combines psychometrics with machine learning and is grounded in applied linguistics and cognitive psychology. We evaluate our approach on real data from Duolingo and compare it with the HLR model proposed by Settles and Meeder (2016), with the eventual goal of contributing to impactful and responsible use of AI in L2 learning.

2. Methodology

2.1. Data description

The data used in the paper come from two weeks of Duolingo log data (Settles and Meeder, 2016), containing 12.9 million learner-lexeme practice sessions. The dataset includes approximately 115k unique learners and 19k unique lexemes across six L2s. For each practice session, a recall rate is provided, representing the proportion of times the learner correctly recalled the lexeme; this value ranges between zero and one.

2.2. Feature extraction

To be able to interpret and explain the predictions for spaced repetition, the extracted features need to be explainable. We categorized these features into three groups: lexeme, learner and contextual features (see Table 1).

Regarding the lexemes, the extracted features represent semantic, formal and frequency-related properties of each lexeme. For *lexeme embeddings*, since the lexemes span six different languages, we use aligned FastText word embeddings (Joulin et al., 2018) to ensure that similar lexemes across languages receive comparable representations. *Part of speech* (e.g., noun, verb, adjective), *morphological features* (e.g., number, gender, case, tense, person) and *number of letters/vowels* reflect the syntactic role and other formal properties of the lexeme. *Lexeme frequency* reflects the normalized occurrence of words, derived from a large

Table 1: Feature Descriptions

Feature type	Feature Description
Lexeme Features	Lexeme Embedding Part of Speech Morphological Features Number of Letters in the Lexeme Number of Vowels in the Lexeme Lexeme Frequency Lexeme BLUP
Learner Features	Number of Languages that Learner is Learning Number of Historical Sessions Number of Unique Lexemes that Learner Has Practiced Number of Lexemes Learned by the Learner Time Elapsed from the First Session of the Learner Average Accuracy Across lexemes Variance in Accuracy Across Lexemes Learner BLUP
Contextual Features	Number of Exercises in the Session Number of Unique Lexemes in the Session Time Interval Since Lexeme was Last Shown (Δ) Session Time of the Day Is Session in Weekend? Accuracy of Responses on this Lexeme

multilingual corpus that includes Wikipedia, subtitles, news articles, books, web text, Twitter, and Reddit (Speer, 2022). The underlying assumption is that more frequent words are easier to recall correctly.

On the learner side, no explicit features are provided in the dataset apart from user IDs; therefore, learner features are extracted from their activity logs. The extracted features consist of frequency-, temporal-, and performance-based indicators. For the *Number of Lexemes Learned* feature, we use an empirically motivated heuristic: a lexeme is treated as learned if it is recalled correctly in four consecutive sessions (Schuetze, 2015).

Contextual features capture information related to the configuration of the practice session and its temporal attributes, as well as learner-lexeme interactions such as aggregated accuracy of the learner on the lexeme.

Incorporating a diverse set of features is important, as L2 learning outcomes arise from the interaction between learner characteristics (e.g., general ability, prior knowledge related to the lexeme that is being learned, working memory) and properties of items (in our case lexemes) (e.g., difficulty, frequency, and linguistic features) (Elgort and Warren, 2014). In spaced repetition contexts, the effectiveness of practice schedules varies between individuals and items, making models that ignore either source of variability inherently suboptimal (Eglington and Pavlik Jr, 2020). Research in L2 proficiency modeling shows that combining learner, item, and contextual features improves predictive accuracy and interpretability compared to single-source models (Awwad and Tavakoli, 2022; Cornillie et al., 2025). By capturing both learner-specific and item-specific influences, along with practice history and contextual dynamics, such models can disentangle their contributions, leading to more accurate, interpretable, and personalized predictions of learning outcomes.

In addition to the outlined learner and lexeme features in Table 1, we extracted the Best Linear Unbiased Predictors (BLUPs) (Robinson, 1991) for the random effects for each learner and lexeme (see Section 2.4 for more detail).

2.3. Half-Life Regression

The original HLR model (Settles and Meeder, 2016) directly parametrizes the half-life h as a log-linear function of inputs $\mathbf{x}$. h represents the time required for the probability of recall p to decay by half. p is then obtained from h according to:

$$\hat{h}_\Theta = 2^{\Theta \cdot \mathbf{x}} \qquad\qquad p = 2^{-\Delta/h}, \tag{1}$$

where Δ is the time interval between sessions and Θ is the model parameters, obtained by minimizing the loss function ℓ:

$$\ell(\mathbf{x}; \Delta; p; \Theta) = (p - \hat{p}_\Theta)^2 + \alpha(h - \hat{h}_\Theta)^2 + \lambda\|\Theta\|_2^2. \tag{2}$$

2.4. Proposed methods

We first extended the original HLR model (Settles and Meeder, 2016), which included previous number of correct and incorrect recalls for a lexeme as features, to a linear mixed model by adding in crossed random effects for learner and lexeme (HLR+LMM in Table 2). More concretely, if we re-write equation 1 as a linear regression on $\log_2(\hat{h}_{ui})$ where u, i denotes the learner-item pair then the linear mixed model can be expressed as:

$$\log_2(h_{ui}) = \Theta \cdot \mathbf{X} + b_u + c_i + \epsilon_{ui} \tag{3}$$

$$b_u \sim N(0, \sigma_b^2); \quad c_i \sim N(0, \sigma_c^2); \quad \epsilon_{ui} \sim N(0, \sigma_\epsilon^2) \tag{4}$$

where b_u is a learner level random effect and c_i is the lexeme level random effect. From this model we extracted the estimated random effects for each learner and lexeme in the training set using their Best Linear Unbiased Predictors (BLUPs) (Robinson, 1991). In this context, the random effects can be interpreted as reflecting learner ability and item difficulty, similar to parameters estimated in Item Response Theory (IRT) (Wilson and De Boeck, 2004). In other words, instead of modeling item difficulty (Tack et al., 2024) or learner ability, these random effects are estimated from the learner's logs on lexemes. We extended this model by expanding the feature set to include additional features shown in Table 1, with the exception of learner and item BLUPs, morphological features and lexeme embeddings, as additional fixed effects to fit a more complex fixed effects structure (HLR+LMM+EF in Table 2). In this model, we incorporated an L_2 regularization term with a weight of $\lambda = 10^{-5}$.

Finally, we analysed the contribution of each feature in the extended set. This feature set is highly heterogeneous, comprising continuous variables (e.g. lexeme embeddings), categorical variables (e.g. the target language), and ordinal variables (e.g. the number of lexemes in a session). Such heterogeneity poses challenges for strongly parameterized models like the original HLR, which assume relatively simple relationships both among features and between features and the outcome. To mitigate these assumptions, we trained a Random Forest (Breiman, 2001) on all the extracted features (outlined in Table 1), including

Table 2: Model performance comparison. The models in the bottom of the table were proposed in this study. EF indicates we used the extended feature set of Table 1.

Model	Acc.	MAE	AUC	Spearman (h)	Spearman (p)
Leitner	0.7055	0.2313	0.5100	-0.0156	0.0128
Pimsleur	0.5396	0.4383	0.5067	-0.0193	0.0084
HLR	0.8197	0.1153	0.5079	0.0295	0.0107
HLR+LMM	0.8248	0.1132	0.5584	0.2528	0.0766
HLR+LMM+EF	0.8232	0.1174	0.6030	0.2967	0.1347
RF+EF	**0.8357**	**0.1044**	**0.6312**	**0.6084**	**0.1788**

learner and item BLUPs, to predict the recall rate. We then extracted Shapley Additive Explanations (SHAP) (Lundberg and Lee, 2017) to investigate the contribution of each feature to the Random Forest's predictions.

2.5. Experimental setup

We held out as test samples the last 10% of the 12.9 million interactions in the original dataset, as described by (Settles and Meeder, 2016). The Random Forest model used 100 trees with maximum depth of 15 and mean squared error as the impurity function. As evaluation metrics, we report accuracy (Acc.), mean absolute error (MAE), area under the ROC curve (AUC), and Spearman correlation for estimated half-life (h) and recall probability (p).

3. Results and discussion

Table 2 summarizes the predictive performance of all evaluated models across multiple metrics. RF+EF outperforms all other competitors, moderately in Accuracy and MAE, and substantially in AUC and Spearman correlations. Among the HLR-based approaches, HLR+LMM+EF achieves the best AUC and Spearman correlations, while performing very similarly to the other HLR variants in Accuracy and MAE. HLR+LMM outperforms HLR across all metrics, indicating that incorporating mixed-effects modeling provides a meaningful improvement over the standard HLR formulation. Although MAE is used in the original HLR paper as the main evaluation measure, we believe it is less relevant for this application, where the goal is to determine which lexeme should be practiced first. In this setting, it is more important to correctly order the items to be practiced than obtain precise point-wise probability estimates, making metrics such as Spearman correlation and AUC more informative. In addition to the superior predictive performance, tree-based models such as Random Forests offer greater potential for explainability, as their structure enables inspection of decision rules, feature importance and decision paths.

The SHAP analysis in Figure 1 shows that the most influential features for the Random Forest predictions are the *time interval* since the lexeme was last shown (Δ), the *number of exercises in the session*, the *number of unique lexemes* that the learner has seen, and

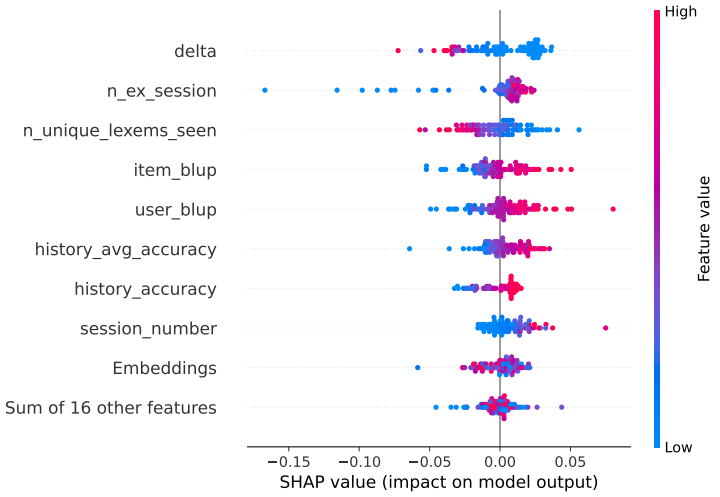
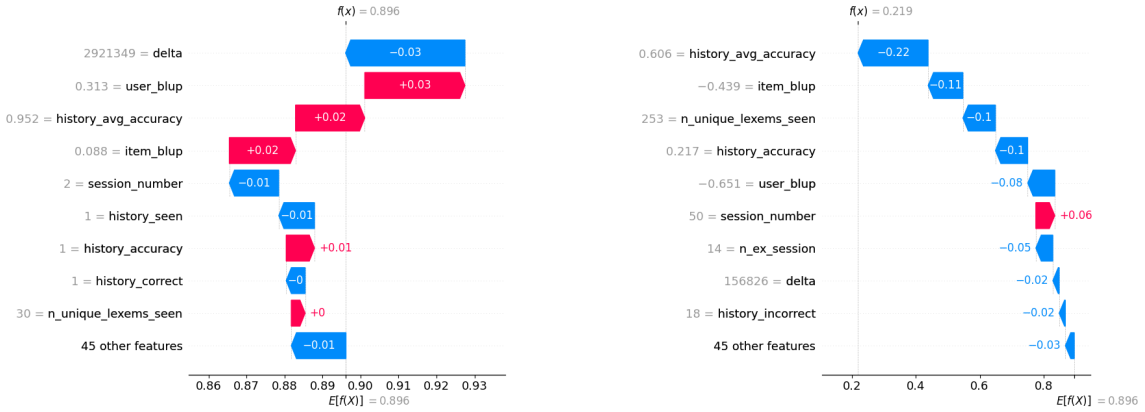


Figure 1: Shapley Additive Explanations (SHAP) for the Random Forest with random effects features.

the *learner* and *lexeme BLUPs*. The first two are contextual features that characterize the practice session for a given learner-lexeme pair. The time interval is the most evident factor and is typically considered the primary driver of forgetting. We hypothesize that the number of exercises in the session informs the model about the quantization of the recall probabilities. The SHAP values also highlight the importance of including random-effects features: lexeme and learner BLUPs are the other influential contributors, capturing lexeme difficulty and learner ability. This information could help learning scientists at Duolingo identify which features most strongly influence learner performance and adjust the content and instructional design of the practice sessions to improve the learning process accordingly. For example, as shown in Figure 1, having more exercises in a practice session appears to have a positive impact on learning outcomes. This discourages sessions with very few exercises in favor of longer sessions in which learners can practice the same lexemes repeatedly.

In addition to the global explanations that can be interpreted by learning scientists at Duolingo, local-level explanations on predictions could also be shown to individual learners. This would help them understand the factors influencing their predicted outcomes and potentially guide them toward more effective learning behaviors. Two examples of such local SHAP explanations are presented in Figure 2. In Figure 2(a), for this particular learner–lexeme pair, high *user_blup* and *historical accuracy* have a positive effect on the predicted outcome, whereas a long *time interval* (Δ) has a negative impact. This suggests that although the learner appears to have strong underlying ability, practicing lexemes at shorter intervals would substantially improve their performance. In the second example (Figure 2(b)), the learner seems to have lower overall ability (*user_blup*) and has practiced many unique lexemes, which increases the probability of rarer, more challenging lexemes being suggested.



(*a*) Local SHAP for a high predicted p_recall

(*b*) Local SHAP for a low predicted p_recall

Figure 2: Local SHAP diagrams

While we have successfully addressed the research question of this study, we acknowledge two main limitations. First, we applied the proposed approach only to the Duolingo dataset, which may affect the generalizability of the findings. Although it is ideal to validate the proposed approach on other relevant datasets, we were unable to find publicly available datasets with similar characteristics and volume as the Duolingo dataset. Furthermore, since we did not have access to the L2 learners on Duolingo, it was not possible to assess the impact of the new spaced repetition algorithm on the learning outcomes of L2 learners who would follow the recommendations from the proposed spaced repetition algorithm. Therefore, in this study, we conducted an offline evaluation to assess the performance of the proposed method on the unseen test set and to identify the most influential features in its predictions.

4. Conclusion and future work

In this study, we proposed an explainable model that enables personalized spaced repetition in L2 practice, and benchmarked our results on data and earlier models from Duolingo. We extracted learner, lexeme and contextual features, such as multilingual semantic embeddings and linear random effects for learners and lexemes. The best-performing method extracts random effects from an HLR-like model and uses them as additional features in a Random Forest. The importance of these features was confirmed by their SHAP values. The use of this diverse set of features enables meaningful explanations at both the global and local levels. These insights can help learning scientists at Duolingo improve the effectiveness of their digital application for L2 practice (global level), while also allowing learners themselves to understand and adapt their learning behaviors to ultimately learn more effectively (local level).

In future work, the potential of Random Forests can be further explored to enhance explainability. 1) One could analyse how different features interact constructively or destructively to produce the model outcomes. 2) One could explore techniques such as BEL-

LATREX (Dedja et al., 2023) to summarize large ensembles of trees and retrieve important decision rules from the tree structure. 3) Large language models could be used to convert explanations into textual representations that are accessible to end-users and to specialists from a wider range of domains. 4) There is still much to be explored regarding the alignment of the patterns discovered in Duolingo data with various pedagogical and psychological frameworks. Another important direction for future work is to validate the results of this study on other relevant datasets and, if possible, assess the impact of the proposed algorithm on the learning experience and outcomes of L2 learners both quantitatively and qualitatively. Finally, we aim to model spaced repetition using survival analysis, where the event of interest is forgetting a lexeme. The goal is therefore to predict learner-lexeme pairs with a high risk of forgetting and disengagement (Gharahighehi et al., 2023), and thus a higher importance of practicing.

Acknowledgments

The authors would like to thank the Research Fund Flanders (FWO, personal mandate 11A7U26N to PI), as well as imec's Smart Education research program and the Flanders AI Research program (both supported by the Flemish Government).

References

Anas Awwad and Parvaneh Tavakoli. Task complexity, language proficiency and working memory: Interaction effects on second language speech performance. *International Review of Applied Linguistics in Language Teaching*, 60(2):169–196, 2022. doi: 10.1515/iral-2018-0378.

Leo Breiman. Random forests. *Machine Learning*, 45:5–32, 2001. doi: 10.1023/A:1010933404324.

Xieling Chen, Di Zou, Haoran Xie, and Gary Cheng. Twenty Years of Personalized Language Learning: Topic Modeling and Knowledge Mapping. *Educational Technology & Society*, 24(1):205–222, 2021. ISSN 1176-3647.

Frederik Cornillie, Julie Gijpen, Sameh Said-Metwaly, Steffen Luypaert, and Wim Van Den Noortgate. Toward adaptive spoken dialogue systems for language learning: Predicting task completion from learning process data. *CALICO Journal*, 42(3):413–436, 2025. doi: 10.3138/calico-2025-0035.

Frederik Cornillie, Joan-Tomàs Pujolà, and Christine Appel. Gaming for focused language practice. In Fiona Farr and Liam Murray, editors, *Routledge Handbook of Language Learning and Technology*. Routledge, 2 edition, 2026. doi: 10.4324/9781003412212-37.

Klest Dedja, Felipe Kenji Nakano, Konstantinos Pliakos, and Celine Vens. BELLATREX: Building Explanations Through a LocaLly AccuraTe Rule EXtractor. *IEEE Access*, 11:41348–41367, 2023. ISSN 2169-3536. doi: 10.1109/ACCESS.2023.3268866.

Hermann Ebbinghaus. *Memory: A contribution to experimental psychology*. Teachers College Press, 1913. doi: 10.1037/10011-000.

Luke G Eglington and Philip I Pavlik Jr. Optimizing practice scheduling requires quantitative tracking of individual item performance. *npj Science of Learning*, 5(1):15, 2020. doi: 10.1038/s41539-020-00074-4.

Irina Elgort and Paul Warren. L2 vocabulary learning from reading: Explicit and tacit lexical knowledge and the role of learner and item variables. *Language Learning*, 64(2): 365–414, 2014. doi: https://doi.org/10.1111/lang.12052.

Alireza Gharahighehi, Michela Venturini, Achilleas Ghinis, Frederik Cornillie, and Celine Vens. Extending bayesian personalized ranking with survival analysis for mooc recommendation. In *Adjunct Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization*, pages 56–59, 2023.

Alireza Gharahighehi, Rani Van Schoors, Paraskevi Topali, and Jeroen Ooge. Adaptive lifelong learning (all). In *International Conference on Artificial Intelligence in Education*, pages 452–459. Springer, 2024.

Armand Joulin, Piotr Bojanowski, Tomáš Mikolov, Hervé Jégou, and Edouard Grave. Loss in translation: Learning bilingual word mapping with a retrieval criterion. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 2979–2984, 2018. doi: 10.48550/arxiv.1804.07745.

Sebastian Leitner. *So lernt man lernen. Angewandte Lernpsychologie - ein Weg zum Erfolg.* Herder, 1972.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Boxuan Ma, Gayan Prasad Hettiarachchi, Sora Fukui, and Yuji Ando. Each encounter counts: Modeling language learning and forgetting. In *LAK23: 13th international learning analytics and knowledge conference*, pages 79–88, 2023. doi: 10.1145/3576050.3576062.

Paul Pimsleur. A memory schedule. *The Modern Language Journal*, 51(2):73–75, 1967.

G. K. Robinson. That BLUP is a good thing: The estimation of random effects. *Statistical Science*, 6(1):15–32, 1991. ISSN 08834237, 21688745. doi: 10.1214/ss/1177011926.

Ulf Schuetze. Spacing techniques in second language vocabulary acquisition: Short-term gains vs. long-term memory. *Language Teaching Research*, 19(1):28–42, 2015. doi: 10.1177/1362168814541726.

Mathias Schulze. Modeling SLA processes using NLP. In Carol A Chapelle, Yoo-Ree Chung, and Jing Xu, editors, *Towards adaptive CALL: Natural language processing for diagnostic language assessment*, pages 149–166, Ames, IA, 2008. Iowa State University.

Burr Settles and Brendan Meeder. A trainable spaced repetition model for language learning. In Katrin Erk and Noah A. Smith, editors, *Proceedings of the 54th Annual Meeting of*

the Association for Computational Linguistics (Volume 1: Long Papers), pages 1848–1858, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1174.

Robyn Speer. wordfreq v3.0, September 2022. URL https://doi.org/10.5281/zenodo.7199437.

Yuichi Suzuki, editor. *Practice and Automatization in Second Language Research. Perspectives from Skill Acquisition Theory and Cognitive Psychology*. Second Language Acquisition Research Series. Routledge, 2023. doi: 10.4324/9781003414643.

Anaïs Tack, Siem Buseyne, Changsheng Chen, Robbe D'hondt, Michiel De Vrindt, Alireza Gharahighehi, Sameh Metwaly, Felipe Kenji Nakano, and Ann-Sophie Noreillie. Itec at bea 2024 shared task: Predicting difficulty and response time of medical exam questions with statistical, machine learning, and language models. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 512–521, 2024.

Mark Wilson and Paul De Boeck. Descriptive and explanatory item response models. In *Explanatory item response models: A generalized linear and nonlinear approach*, pages 43–74. Springer, 2004.

Junyao Ye, Jingyong Su, and Yilong Cao. A stochastic shortest path algorithm for optimizing spaced repetition scheduling. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pages 4381–4390, 2022. doi: 10.1145/3534678.3539081.

Ahmed Zaidi, Andrew Caines, Russell Moore, Paula Buttery, and Andrew Rice. Adaptive forgetting curves for spaced repetition language learning. In *International Conference on Artificial Intelligence in Education*, pages 358–363. Springer, 2020. doi: 10.1007/978-3-030-52240-7_65.