A Memory-Sensitive Classification Model of Errors in Early Second Language Learning

Brendan Tomoschuk, Jarrett T. Lovelett

[1]University of California, San Diego

Contact: btomoschuk@ucsd.edu, jlovelet@ucsd.edu

INTRODUCTION

Acquiring a second language (L2) as an adult is notoriously difficult.
By understanding where individual learners make mistakes, we can improve efficiency and durability of L2 learning

- Linguistic factors:
  - E.g. Cognates, concrete words are easier (de Groot & Keizer, 2000) while interlingual homographs are harder (Dijkstra, Timmermans & Schriefers, 2000)
- Memory factors:
  - Since language is learned, it must be stored in memory.
    - What improves memory in general should also improve memory for language
  - Spaced repetition: words (and other items) are remembered better when they are encountered repeatedly, with temporal gaps in between (vs. repeated all at once).
    - Longer gaps are better (e.g. Cepeda et al. 2006)
    - Robust over seconds, minutes, days, weeks, years (e.g. Cepeda et al. 2008)
    - Applies to a wide variety of materials (e.g. Donovan & Radosevich, 1999)
    - Including language (e.g. Ullman & Lovelett, 2018)
  - Retrieval Practice: Recalling information from memory makes that information easier to recall in the future
    - Duolingo frequently prompts users to retrieve from memory
    - Retrieval practice enhances the efficacy of spaced repetition

By better understanding the factors that influence learning and retention of L2, systems like Duolingo can:

- Devote more resources to the most difficult aspects of the L2 (for each learner)
- Schedule review of learned material when it is of most benefit to the learner
- Leverage their own users' data to improve understanding of the learning process, and in turn improve learning outcomes

DATA Settles et al. 2018

POPULATIONS

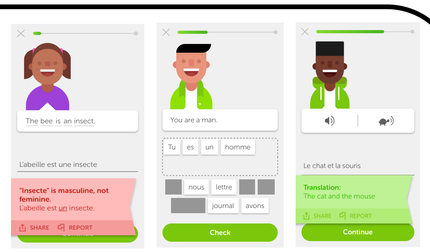
Three groups were analyzed separately:

- English-speaking learners of Spanish
- English-speaking learners of French
- Spanish-speaking learners of English

THREE SETS

The first 30 days of each users learning broken are broken down into:

- Training: each user's first 80% of sessions
- Development: the next 10% of each user's data
- Test: the final 10% of exercises for each user



Reverse Translate | Reverse Tap | Listen

Figure 1. *Examples of Duolingo exercises and error markings present in the data*

MODEL

Random forest classifier

- Each decision tree branched a number of times equal to the square root of the total number of features
- An ensemble of 1000 trees was created for each of the three language datasets
- Each tree branched until leaves were pure (contained only a single label: "error" or "no error")
- Out-of-bag error was used to estimate prediction error of the classifier
- The classifier was trained in Python 3, using sklearn.ensemble.RandomForestClassifier()

ENGINEERED FEATURES

Linguistic		Memory		Categorical		Interactions
orthoLength	Word length in characters	**nthOccurance**	Number of times a token has been seen	**pos**	Part of speech	**stemLag1 x stemLag2**
phonLength	Word length in phonemes	**userTrial**	Number of trials a user has seen	**format**	Trial format (see Figure 1)	**stemLag1 x stemLag2 x lagTr1Tr2**
orthoNei	Number of orthographic word neighbors	**tokenLag1**	Amount of time since token last seen	**prevFormat**	Previous trial format	**lagTr1Tr2 x morphoComplexity**
phonNei	Number of phonological word neighbors	**tokenLag2**	Amount of time between last time a word has been seen and the time before that	**client**	User's client (collapsed to mobile or web)	**lagTr1Tr2 x morphoLag1**
logWordFreq	log-transformed word frequency	**stemLag1**	Amount of time since stemmed token has been seen	**userMeanError**	Average of a user's accuracy across trials	**Format x prevFormat**
logOrthoNeiFreq	Average log-transformed word frequency of orthographic neighbors	**stemLag2**	Amount of time between last time a stemmed token has been seen and the time before that	**userVarError**	Variance in a user's accuracy across trials	**orthoNei x format**
logPhonNeiFreq	Average log-transformed word frequency of phonological neighbors	**morphoLag1**	Amount of time since morphological features have last been seen			**phonNei x format**
Edit Distance	Levenshtein distance between translations of word	**lagTr1Tr2**	Amount of time between first and second trials containing that token			**format x client**
Interlingual homograph	Whether a given translation was identical to a different word in the source language					**morphoComplexity x pos**
morphoCompexity	Number of morphological features					
Concreteness	Subject ratings of how perceptible an entity is					

Table 1. *Names and descriptions of the engineered features*

RESULTS & DISCUSSION

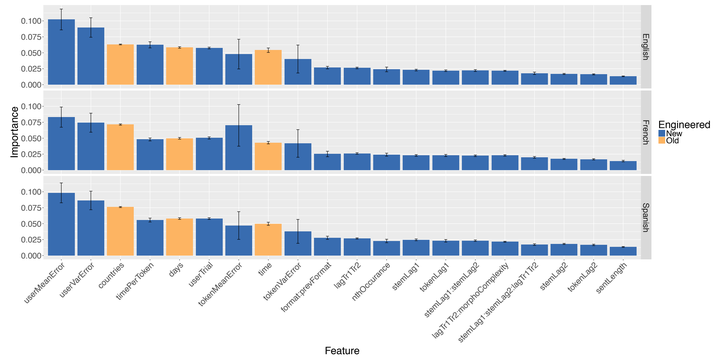


Figure 2. *Importance measures for each of the top 20 features.*

- Most important features: u*serMeanError*; *userVarError*:
  - mean and variance of each user's error rate (under each combination of levels of a small set of features)
  - Computational savings over fitting a more comprehensive random effect structure (i.e. random effects for all users, all tokens, and all user-token combinations, at minimum)
- The more time users spend per token (on average) within an exercise (*timePerToken*) the more likely they are to make errors in that exercise
- Users make more errors on average the longer they've spent using the app (*Days, userTrial*). Perhaps because item difficulty also increases with experience.
- Words that repeat more often (*nthOccurrence*) are remembered better.
- The more time that passed since the previous occurrence of a word, the higher the error rate (*tokenLag1*, *tokenLag2)*
  - Contra spacing effect: perhaps more consideration of full item history is needed (or gaps too long; see Cepeda et al. 2008)
- There seems to be a cost to switching formats: error rates are higher when the current task type is different from the previous (*format:prevFormat*)
- Future models will include ablation experiments and word embeddings

	AUROC	F1	Log-loss
SLAM English	.7730	.1899	.3580
English	.8286	.4242	.3191
SLAM French	.7707	.2814	.3952
French	.8228	.4416	.3561
SLAM Spanish	.7456	.1753	.3862
Spanish	.8027	.4353	.3571

Table 2. *Model outcomes compared to SLAM baselines.*

REFERENCES

Ben Ambridge, Anna L. Theakston, Elena V.m. Lieven, and Michael Tomasello. 2006. The distributed learning effect for children's acquisition of an abstract syntactic construction. *Cognitive Development*, *21*(2), 174–193. **Harry P. Bahrick** and Elizabeth Phelphs. 1987. Retention of Spanish vocabulary over 8 years. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*(2), 344–349. **David A. Balota**, Janet M. Duchek, and Ronda Paullin. 1989. Age-related differences in the impact of spacing, lag, and retention interval. *Psychology and Aging*, *4*(1), 3–9. **Marc Brysbaert**, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior research methods*, *46*(3), 904-911. **Shana K. Carpenter**. 2009. Cue strength as a moderator of the testing effect: The benefits of elaborative retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(6), 1563–1569. **Shana K. Carpenter** and Edward L. DeLosh. 2006. Impoverished cue support enhances subsequent retention: Support for the elaborative retrieval explanation of the testing effect. *Memory and Cognition*, 268-276. **Nicholas J. Cepeda**, Harold Pashler, Edward Vul, John T Wixted, and Doug Rohrer. 2006. Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin, 132*(3), 354-380. **Nichlaos J. Cepeda**, Edward Vul, Doug Rohrer, John T. Wixted, and Harold Pashler. 2008. Spacing effects in learning: A temporal ridgeline of optimal retention. *Psychological Science, 19*(11), 1095-1102. **William L. Cull**. 2000. Untangling the benefits of multiple study opportunities and repeated testing for cued recall. *Applied Cognitive Psychology, 14*(3), 215-235. **Annette De Groot** and Rineke Keijzer. 2000. What is hard to learn is easy to forget: The roles of word concreteness, cognate status, and word frequency in foreign-language vocabulary learning and forgetting. *Language Learning*, *50*(1), 1-56. **Ton Dijkstra**, Mark Timmermans, and Herbert Schriefers. 2000. On being blinded by your other language: Effects of task demands on interlingual homograph recognition. *Journal of Memory and Language*, *42*(4), 445-464. **John J. Donovan** and David J. Radosevich. 1999. A meta-analytic review of the distribution of practice effect: Now you see it, now you don't. *Journal of Applied Psychology, 84*(5), 795-805. **Hermann Ebbinghaus**. 1964. Memory: A contribution to experimental psychology (H.A. Ruger, C.E. Bussenius, & E. R. Hilgard, Trans.). New York, NY: Dover. (Original work published in 1885). **Jeffrey D. Karpicke** and Henry L. Roediger. 2007. Expanding retrieval practice promotes short-term retention, but equally spaced retrieval enhances long-term retention. *Journal of Experimental Psychology-Learning Memory and Cognition, 33*(4), 704-719. **Thomas K. Landauer** and Robert A. Bjork. 1978. Optimum rehearsal patterns and name learning In M. M. Gruneberg, P. E. Morris, & R. N. Sykes (Eds.), *Practical aspects of memory* (pp. 625-632). London: Academic Press. **Viorica Marian**, James Bartolotti, Sarah Chabal, and Anthony Shook. 2012. CLEARPOND: Cross-linguistic easy-access resource for phonological and orthographic neighborhood densities. *PloS one*, *7*(8), e43230. **Cornelius P. Rea** and Vito Modigliani. 1985. The effect of expanded versus massed practice on the retention of multiplication facts and spelling lists. *Human Learning: Journal of Practical Research and Applications, 4*(1), 11-18. **B. Settles**, C. Brust, E. Gustafson, M. Hagiwara, and N. Madnani. 2018. Second Language Acquisition Modeling. In *Proceedings of the NAACL-HLT Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*. ACL. **Michael T. Ullman** and Jarrett T. Lovelett. 2016. Implications of the declarative/procedural model for improving second language learning: The role of memory enhancement techniques. *Second Language Research*, *39*(1), 39-65. **Eleanor Vander Linde**, Barbara A. Morrongiello, and Carolyn Rovee-Collier. 1985. Determinants of retention in 8-week-old infants. *Developmental Psychology*, *21*(4), 601–61.