Feature Engineering for Second Language Acquisition Modeling

Guanliang Chen, Claudia Hauff, Geert-Jan Houben
Delft University of Technology
Delft, The Netherlands
{guanliang.chen, c.hauff, g.j.p.m.houben}@tudelft.nl

Abstract

Knowledge tracing serves as a keystone in delivering personalized education. However, few works attempted to model students' knowledge state in the setting of Second Language Acquisition. The Duolingo Shared Task on Second Language Acquisition Modeling (Settles et al., 2018) provides students' trace data that we extensively analyze and engineer features from for the task of predicting whether a student will correctly solve a vocabulary exercise. Our analyses of students' learning traces reveal that factors like exercise format and engagement impact their exercise performance to a large extent. Overall, we extracted 23 different features as input to a Gradient Tree Boosting framework, which resulted in an AUC score of between 0.80 and 0.82 on the official test set.

1 Introduction

Knowledge Tracing plays a crucial role in providing adaptive learning to students (Pelánek, 2017): by estimating a student's current knowledge state and predicting her performance in future interactions, students can receive personalized learning materials (e.g. on the topics the student is estimated to know the least about).

Over the years, various knowledge tracing techniques have been proposed and studied, including Bayesian Knowledge Tracing (Corbett and Anderson, 1994), Performance Factor Analysis (Pavlik Jr et al., 2009), Learning Factors Analysis (Cen et al., 2006) and Deep Knowledge Tracing (Piech et al., 2015). Notable is that most of the existing works focus on learning performance within mathematics in elementary school and high school due to the availability of sufficiently large datasets in this domain, e.g. ASSISTment and OLI (Piech et al., 2015; Xiong et al., 2016; Zhang et al., 2017; Khajah et al., 2016). The generalization to other learning scenarios and domains remains under-explored.

Particularly, there are few studies attempted to explore knowledge tracing in the setting of Second Language Acquisition (SLA) (Bialystok, 1978). Recent studies showed that SLA is becoming increasingly important in people's daily lives and should gain more research attention to facilitate their learning process (Larsen-Freeman and Long, 2014). It remains an open question whether the existing knowledge tracing techniques can be directly applied to SLA modeling—the release of the Duolingo challenge datasets now enables us to investigate this very question.

Thus, our work is guided by the following research question: **What *factors* impact students' language learning performance?**

To answer the question, we first formulate six research hypotheses which are built on previous studies in SLA. We perform extensive analyses on the three SLA Duolingo datasets (Settles et al., 2018) to determine to what extent they hold. Subsequently, we engineer a set of 23 features informed by the analyses and use them as input for a state-of-the-art machine learning model, *Gradient Tree Boosting* (Ye et al., 2009; Chen and Guestrin, 2016), to estimate the likelihood of whether a student will correctly solve an exercise.

We contribute the following major findings: (i) students who are heavily engaged with the learning platform are more likely to solve words correctly; (ii) contextual factors like the device being used and learning format impact students' performance considerably; (iii) repetitive practice is a necessary step for students towards mastery; (iv) Gradient Tree Boosting are demonstrated to be an effective method for predicting students' future performance in SLA.

2 Data Analysis

Before describing the six hypotheses we ground our work in as well as their empirical validation, we first introduce the Duolingo datasets.

2.1 Data Description

To advance knowledge modeling in SLA, Duolingo released three datasets[1], collected from students of English who already speak Spanish (EN-ES), students of Spanish who already speak English (ES-EN), and students of French who already speak English (FR-EN), respectively, over their first 30 days of language learning on the Duolingo platform (Settles et al., 2018). The task is to predict what mistakes a student will make in the future. Table 1 shows basic statistics about each dataset. Interesting are in particular the last two rows of the table which indicate the unbalanced nature of the data: across all languages correctly solving an exercise is far more likely than incorrectly solving it. Note that the datasets contain rich information not only on students, words and exercises[2] but also on students' learning process, e.g., the amount of time a student required to solve an exercise, the device being used to access the learning platform and the countries from which a student accessed the Duolingo platform.

Table 1: Statistics of the datasets.

	FR-EN	ES-EN	EN-ES
#Unique students	1,213	2,643	2,593
#Unique words	2,178	2,915	2,226
#Exercises	326,792	731,896	824,012
#Words in all exercises	926,657	1,973,558	2,622,958
#Avg. words / exercise	2.84	2.7	3.18
%Correctly solved words	84%	86%	87%
%Incorrectly solved words	16%	14%	13%

In our work, we use *learning session* to denote the period from a student's login to the platform until the time she leaves the platform. We use *learning type* to refer to the "session" information in the original released datasets, whose value can be *lesson*, *practice* or *test*.

2.2 Research Hypotheses

Grounded in prior works we explore the following hypotheses:

H1 A student's *living community* affects her language acquisition performance.
Previous works, e.g., (Dixon et al., 2012) demonstrated that the surrounding living community is a non-negligible factor in SLA. For instance, a student learning English whilst living in an English-speaking country is more likely to practice more often and thus more likely to achieve a higher learning gain than a student not living in one.

H2 The more *engaged* a student is, the more words she can master.
Educational studies, e.g., (Carini et al., 2006), have shown that a student's engagement can be regarded as a useful indicator to predict her learning gain, which is the number of mastered words in our case.

H3 The *more time* a student spends on *solving an exercise*, the more likely she will get it wrong.

H4 *Contextual factors* such as the device being used (e.g. iOS or Android), learning type (lesson, practice or test) and exercise format (such as transcribing an utterance from scratch or formulating an answer by selecting from a set of candidate words) will impact a student's mastery of a word.
We hypothesize that, under specific contexts, a student can achieve a higher learning gain due to the different difficulty level of exercises. For instance, compared to transcribing an utterance from scratch, a student is likely to solve more exercises correctly when being provided with a small set of candidate words.

H5 *Repetition* is useful and necessary for a student to master a word (Young-Davy, 2014; Gu and Johnson, 1996; Lawson and Hogben, 1996).

H6 Students with a high-spacing learning routine are more likely to learn more words than those with a low-spacing learning routine.
Here, high-spacing refers to a larger number of discrete learning sessions. Correspondingly, low-spacing refers to relatively few learning sessions, which usually last a relatively long time. In other words, students with a low-spacing routine tend to acquire words in a "cramming" manner (Miyamoto et al., 2015; Donovan and Radosevich, 1999; Bjork, 1994).

[1] http://sharedtask.duolingo.com/#task-definition-data

[2] An exercise usually contains multiple words.

2.3 Performance Metrics

We now define four metrics we use to measure a student's exercise performance.

Student-level Accuracy (Stud-Acc) measures the overall accuracy of a student across all completed exercises. It is calculated as the ratio between the number of words correctly solved by a student and the total number of words she attempted.

Exercise-level Accuracy (Exer-Acc) measures to what extent a student answers a particular exercise correctly. It is computed as the number of correctly solved words divided by the total number of words in the exercise.

Word-level Accuracy (Word-Acc) measures the percentage of times of a word being answered correctly by students. For a word, it is calculated as the number of times students provided correct answers divided by the total number of attempts.

Mastered Words (Mast-Word) measures how many words have been mastered by a student. As suggested in (Young-Davy, 2014), it takes about 17 exposures for a student to learn a new word. Thus, we define a word being mastered by a student only if (i) it has been exposed to the student at least 17 times and (ii) the student answered the word accurately in the remaining exposures.

Table 2: Avg. student-level accuracy (%) and the number of mastered words of students living in different locations (approximated by the countries from which students have finished the exercises). Significant differences (compared to *Avg.*, according to Mann-Whitney) are marked with $*$ ($p < 0.001$).

Datasets	Locations	Stud-Acc	Mast-Word
FR-EN	Avg.	83.57	3.37
	CA	84.12	3.13
	US	83.01	3.40
	GB	83.66	3.46
	AU	85.69	3.70
ES-EN	Avg.	85.91	2.74
	CA	84.89	3.26
	US	86.22	2.58
	AU	85.82	3.50
	GB	83.94 *	3.30
	NL	87.15	2.86
EN-ES	Avg.	87.62	4.39
	CO	87.49	4.14
	US	87.98	5.02
	ES	87.85	5.66 *
	MX	86.92 *	3.71 *
	CL	88.95	4.42
	DO	87.26	4.40
	AR	89.58	4.75
	VE	89.47 *	4.99
	PE	88.83	4.37

2.4 From Hypotheses To Validation

To verify **H1**, we use the location (country) from where a student accessed the Duolingo platform as an indicator of the student' living community. We first bin students into groups according to their locations. Next, we calculate the average student-level accuracy and the number of mastered words of students in each group. We report the results in Table 2. Here we only consider locations with more than 50 students. If a student accessed the platform from more than one location, the student would be assigned to all of the identified location groups. In contrast to our hypothesis, we do not observe the anticipated relationship between living community and language learning (e.g. Spanish-speaking English-students living in the US do not perform better than other students).

For **H2** (student engagement), we consider three ways to measure engagement with the platform: (i) number of attempted exercises, (ii) number of attempted words and (iii) amount of time spent learning. To quantify the relationship between students' engagement and their learning gain, we report the Pearson correlation coefficient between the three engagement metrics and Stud-Acc as well as Mast-Word (Table 3). We note a consistent negative correlation between accuracy and our engagement metrics. This is not surprising, as more engagement also means more exposure to novel vocabulary items. When examining the number of mastered words, we can conclude that—as stated in **H2**—higher engagement does indeed lead to a higher learning gain. This motivates us to design engagement related features for knowledge tracing models.

To determine the validity of **H3**, in Table 4 we report the Pearson correlation coefficient between the amount of time spent in solving each exercise and the corresponding exercise-level accuracy. The moderate negative correlation values indicate that the hypothesis holds to some extent.

For **H4**, we investigate three types of contextual factors: (i) device used (i.e., Web, iOS, Android); (ii) learning type (i.e., Lesson, Practice, Test) and (iii) exercise format (i.e., Reverse Translate, Listen, Reverse Tap). To verify whether these contextual factors impact students' exercise performance, we partition exercises into different groups

Table 3: Pearson Correlation between student engagement (measured by # attempted exercises/words and the amount of time spent in learning) and student-level accuracy as well as # mastered words. Significant differences are marked with $*$ ($p < 0.001$).

	Stud-Acc			Mast-Word		
	FR-EN	**ES-EN**	**EN-ES**	**FR-EN**	**ES-EN**	**EN-ES**
# Exercises Attempted	-0.05 *	-0.09 *	-0.08 *	0.85 *	0.87 *	0.79 *
# Words Attempted	-0.06 *	-0.08 *	-0.08 *	0.85 *	0.86 *	0.80 *
Time Spent	-0.13 *	-0.14 *	-0.22 *	0.73 *	0.79 *	0.61 *

Table 4: Pearson Correlation between the amount of time spent in solving each exercise and exercise-level accuracy. Significant differences are marked with $*$ ($p < 0.001$).

	FR-EN	**ES-EN**	**EN-ES**
Correlation	-0.16 *	-0.18 *	-0.18 *

Table 5: Average exercise-level accuracy (%) in different contextual conditions. Significant differences (compared to *Avg.*, according to Mann-Whitney) are marked with $*(p < 0.001)$.

	FR-EN	**ES-EN**	**EN-ES**
Avg.	84.29	86.31	87.96
Client			
Web	80.64 *	85.44 *	85.68 *
iOS	86.45 *	87.90 *	88.10 *
Android	83.92 *	84.88 *	88.92 *
Session			
Lesson	85.43 *	87.23 *	88.76 *
Practice	80.94 *	83.92 *	84.19 *
Test	82.19 *	84.34 *	84.66 *
Format			
Reverse Translate	77.92 *	85.88 *	85.42 *
Listen	78.30 *	77.01	82.78 *
Reverse Tap	92.51 *	94.84 *	95.48 *

according to the contextual condition in which they were completed and calculate the average of their exercise-level accuracy within each group. Table 5 shows the results. Interestingly, students with *iOS* devices perform better than those using *Web* or *Android*. Students' learning accuracy is highest in the *Lesson* type. Learning formats also have an impact: *Reverse Tap* achieves the highest accuracy followed by *Reverse Translate* and then *Listen*. This result is not surprising as active recall of words is more difficult than recognition. Finally, we note for English students who speak Spanish (EN-ES) and Spanish students who speak English (ES-EN), the accuracy of *Reverse Translate* is considerably higher than *Listen*, which is not the case in FR-EN (where both are comparable). These results suggest that contextual factors should be taken into account in SLA modeling.

Table 6: Avg. word-level accuracy (%) of words with different number of exposures.

	# Words	**Word-Acc**	**Correlation**
FR-EN			
≥ 1	2,178	72.30	-0.08 *
≥ 10	1,007	75.01	0.13 *
≥ 20	756	75.78	0.15 *
≥ 50	756	76.41	0.19 *
≥ 100	580	77.47	0.25 *
ES-EN			
≥ 1	2,915	75.33	-0.10 *
≥ 10	1,798	77.10	0.12 *
≥ 20	1,511	77.29	0.19 *
≥ 50	1,163	77.92	0.25 *
≥ 100	900	78.67	0.31 *
EN-ES			
≥ 1	2,226	75.58	0.00
≥ 10	1,587	77.12	0.25 *
≥ 20	1,401	77.88	0.28 *
≥ 50	1,171	78.90	0.28 *
≥ 100	963	79.57	0.34 *

Table 7: Pearson Correlation between student performance and the number of previous attempts and the amount of time elapsed since the last attempt for a word.

	FR-EN	**ES-EN**	**EN-ES**
# Previous attempts	-0.05 *	-0.04 *	-0.07 *
Time elapsed	0.05 *	0.06 *	0.07 *

We investigate **H5** from two angles. Firstly, we investigate whether words with very different exposure amounts will differ from each other in terms of word-level accuracy as they are practiced by students to different degrees. For this purpose, we only retain words with more than n exposures (with n being ≥ 1, ≥ 10, ≥ 20, ≥ 50, ≥ 100)

and calculate Pearson correlation coefficient between the word-level accuracy and their number of exposures (Table 6). As expected, the more low-exposure words we filter out, the higher the average word-level accuracy and the stronger the correlation scores (albeit at best these are moderate correlations).

Secondly, we believe that whether a student will solve a word correctly (0 mean solving correctly and 1 incorrectly) is affected by two factors related to word repetition. One factor is the number of previous attempts that a student has for a word, and the other is the amount of time elapsed since her last attempt at the word. Therefore, we compute Pearson correlation coefficient between students' performance on exercises and the two repetition related factors (Table 7). The resulting correlations are even weaker than in our preceding analysis, though they do point towards a (very) weak relationship: if a student gets more exposed to a word or practices the word more frequently, she is more likely to get it correct. Clearly, the results indicate that other factors at play here too.

Lastly, to study **H6**, we partition all students into low-spacing and high-spacing groups according to (Miyamoto et al., 2015). Initially, all students are sorted in ascending order according to their total time spent in learning words. Subsequently, these students are binned into ten equally-sized groups labeled from 0 (spending the least amount of time) to 9 (spending the most amount of time). Therefore, we can regard students from the same group as learning roughly the same amount of time. Next, within each group, the students are sorted based on their number of distinct learning sessions[3], and we further divide them into two equally-sized subgroups: students with few sessions (low-spacing) and students with many sessions (high-spacing). In this way, students spending similar total amounts of time can be compared with each other. We plot the average student-level accuracy as well as the number of mastered words within each low-spacing and high-spacing subgroup in Figure 1. We do not observe consistent differences between low-spacing and high-spacing groups. Therefore, we conclude **H6** to not hold.

[3]Here we consider all learning activities occurring within 60 minutes as belonging to the same learning session.

3 Knowledge Tracing Model

We now describe the machine learning model we adopt for knowledge tracing and then introduce our features.

3.1 Gradient Tree Boosting

Various approaches have been proposed for modeling student learning. Two representatives are Bayesian Knowledge Tracing (Corbett and Anderson, 1994) and Performance Factor Analysis (Pavlik Jr et al., 2009), both of which have been studied for years. Inspired by the recent wave of deep learning research in different domains, deep neural nets were also recently applied to track the knowledge state of students (Piech et al., 2015; Xiong et al., 2016; Zhang et al., 2017). In principal, all of these methods can be adapted to predict students' performance in SLA. As our major goal is to investigate the usefulness of the designed features, we selected a robust model that is able to take various types of features as input and works well with skewed data. Gradient Tree Boosting (GTB) is a machine learning technique which can be used for both regression and classification problems (Ye et al., 2009). It is currently one of the most robust machine learning approaches that is employed for a wide range of problems (Chen and Guestrin, 2016). It can deal with various types of feature data and has reliable predictive power when dealing with unbalanced data (as in our case). We selected it over a deep learning approach as we aim to built an interpretable model.

3.2 Feature Engineering

Based on the results in §2.4, we designed 23 features. The features are categorized into two groups: features directly available in the datasets (7 *given features*) and features derived from the datasets (16 *derived features*). Note that the features differ in their granularity—they are computed per student, or per word, per exercise or a combination of them, as summarized in Table 8.

Given features:

- *Student ID*: the 8-digit, anonymized, unique string for each student;
- *Word*: the word to be learnt by a student;
- *Countries*: a vector of dimension N (N denotes the total number of countries) with

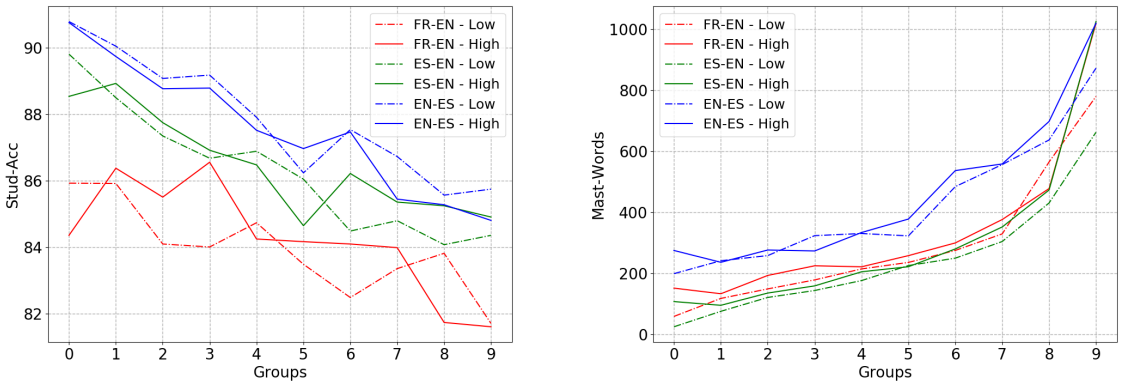

Figure 1: The average student-level accuracy, i.e., Stud-Acc (Left), and the average number of mastered words, i.e., Mast-Word (Right), of students in high-spacing and low-spacing groups.

Table 8: Granularity levels on which each feature is retrieved or computed. Features marked with *b* are used as input in the baseline provided by the benchmark organizers.

Features	**Granularity Level**		
	User	**Word**	**Exercise**
Student ID[b]	√		
Word[b]		√	
Countries	√		
Format[b]			√
Type			√
Device			√
Time spent (exercise)			√
# Exercises attempted	√		
# Words attempted	√		
# Unique words attempted	√		
# sessions	√		
Time spent (learning)	√		
# Previous attempts	√	√	
# Correct times	√	√	
# Incorrect times	√	√	
Time elapsed	√	√	
Word-Acc	√	√	
Std. timestamps (exercise)	√		√
Std. timestamps (word)	√	√	
Std. timestamps (session)	√		
Std. timestamps (word-session)	√	√	
Std. timestamps (word-correct)	√	√	
Std. timestamps (word-incorrect)	√	√	

binary values indicating whether a student complete an exercise in one or multiple countries;

- *Format*: the exercise format in which a student completed an exercise, i.e., Reverse Translate, Reverse Tap and Listen;
- *Type*: the learning type in which a student completed an exercise, i.e., Lesson, Practice and Test;
- *Device*: the device platform which is used by a student to complete an exercise, i.e., iOS, Web and Android;
- *Time spent (exercise)*: the amount of time a student spent in solving an exercise, measured in seconds;

Derived features:

- *# Exercises attempted*: the number of exercises that a student has attempted in the past;
- *# Words attempted*: the number of words that a student has attempted in the past;
- *# Unique Words attempted*: the number of unique words a student has attempted in the past;
- *# Sessions*: the number of learning sessions a student completed;
- *Time spent (learning)*: the total amount of time a student spent learning, measured in minutes;
- *# Previous attempts*: a student's number of previous attempts at a specific word;
- *# Correct times*: the number of times that a student correctly solved a word;
- *# Incorrect times*: the number of times that a student incorrectly solved a word;
- *Time elapsed*: the amount of time that elapsed since the last exposure of a word to a student;

- *Word-Acc*: the word-level accuracy that a student gained for a word in the training dataset;
- *Std. timestamps (exercise)*: the standard deviation of the timestamps that a student solved exercises;
- *Std. timestamps (word)*: the standard deviation of the timestamps that a student solved a word;
- *Std. timestamps (session)*: the standard deviation of timestamps that a student logged in to start a learning session;
- *Std. timestamps (word-session)*: the standard deviation of session starting timestamps that a student solved a word;
- *Std. timestamps (word-correct)*: the standard deviation of timestamps that a student answered a word correctly;
- *Std. timestamps (word-incorrect)*: the standard deviation of timestamps that a student answered a word incorrectly.

Finally, we note that none of the features in our feature set make use of external data sources. We leave the inclusion of additional data sources to future work.

4 Experiments

In this section, we first describe our experimental setup and then present the results.

4.1 Experimental Setup

Each of the three Duolingo datasets consists of three parts: TRAIN and DEV sets for offline experimentations and one TEST set for the final evaluation. We use the TRAIN and DEV sets to explore features that are useful in predicting a student's exercise performance and then combine TRAIN and DEV sets to train the GTB model; we report the model's performance on the TEST set.

We trained the GTB model using XGBoost, a scalable machine learning system for tree boosting (Chen and Guestrin, 2016). All model parameters[4] were optimized through grid search and are reported in Table 9.

[4]For a detailed explanation of the parameters, please refer to https://github.com/dmlc/xgboost/blob/v0.71/doc/parameter.md.

We also report the official baseline provided by the benchmark organizers as comparison. The baseline is a logistic regression model which takes six features as input, which include student ID, word, format and three morpho-syntactic features of the word (e.g., Part of Speech). As suggested by the benchmark organizers, we use the AUC and F1 scores as our evaluation metrics.

Table 9: Model parameters of the GTB model; determined by using grid search per dataset.

	FR-EN	**ES-EN**	**EN-ES**
learning_rate	0.4	0.5	0.6
n_estimatorss	800	1100	1550
max_depth	6	6	5
min_child_weight	7	8	13
gamma	0.0	0.0	0.1
subsample	1.0	1.0	1.0
colsample_bytree	0.7	0.7	0.85
reg_alpha	4	6	5

4.2 Results

In order to evaluate the impact of the features described in §3.2, we report in Table 10 different versions of GTB training, starting with three features (Student ID, Word, Format) and adding additional features one at a time. We incrementally added features according to the order presented in Section 3.2 and only kept features that boost the prediction performance (i.e. the AUC score improves on the DEV set). Among all 23 evaluated features, seven are thus useful for SLA modeling. Here, we only report the results in the ES-EN dataset; we make similar observations in the other two datasets. In contrast to our expectations, a large number of the designed features did not boost the prediction accuracy. This implies that further analyses of the data and further feature engineering efforts are necessary. The extraction of features from external data sources (which may provide insights in the difficulty of words, the relationship between language families and so on) is also left for future work.

In our final prediction for the TEST set, we combine the TRAIN and DEV data to train the GTB model with the nine features listed in Table 10 and student ID as well as the word as input. The results are shown in Table 11. Compared to the logistic regression baseline, GTB is more effective with a 6% improvement in AUC and 83% improvement in F1 on average.

Table 10: Experimental results reported in AUC on ES-EN. Each row indicates a feature added to the GBT feature space; the model of row 1 has three features.

	TRAIN	DEV
Student ID & Word & Format	0.8095	0.7758
Mode	0.8111	0.7780
Client	0.8137	0.7790
Time spent (exercise)	0.8270	0.7828
# Previous attempts	0.8323	0.7835
# Wrong times	0.8348	0.7871
Std. time (word-session)	0.8348	0.7871

Table 11: Final prediction results on the TEST data. Significant differences (compared to Baseline, according to paired t-test) are marked with $*$ ($p < 0.001$).

	Methods	AUC	F1
FR-EN	Baseline	0.7707	0.2814
	GTB	0.8153 *	0.4145 *
ES-EN	Baseline	0.7456	0.1753
	GTB	0.8013 *	0.3436 *
EN-ES	Baseline	0.7737	0.1899
	GTB	0.8210 *	0.3889 *

5 Conclusion

Knowledge tracing is a vital element in personalized and adaptive educational systems. In order to investigate the peculiarities of SLA and explore the applicability of existing knowledge tracing techniques for SLA modeling, we conducted extensive data analyses on three newly released Duolingo datasets. We identified a number of factors affecting students' learning performance in SLA. We extracted a set of 23 features from student trace data and used them as input for the GTB model to predict students' knowledge state. Our experimental results showed that (i) a student's engagement plays an important role in achieving good exercise performance; (ii) contextual factors like the device being used and learning format should be taken into account for SLA modeling; (iii) repetitive practice of words and exercises affect students performance considerably; (iv) GTB can effectively use some of the designed features for SLA modeling and there is a need for further investigation on feature engineering. Apart from the future work already outlined in previous sections, we also plan to investigate deep knowledge tracing approaches and the inclusion of some of our rich features into deep models, inspired by (Zhang et al., 2017). Also, instead of developing a one-size-fits-all prediction model, it will be interesting to explore subsets of students that behave similarly and develop customized models for different student groups.

References

Ellen Bialystok. 1978. A theoretical model of second language learning. *Language learning*, 28(1):69–83.

Robert A. Bjork. 1994. Memory and metamemory considerations in the training of human beings. *Metacognition: Knowing about knowing*, pages 185–205.

Robert M. Carini, George D. Kuh, and Stephen P. Klein. 2006. Student engagement and student learning: Testing the linkages*. *Research in Higher Education*, 47(1):1–32.

Hao Cen, Kenneth Koedinger, and Brian Junker. 2006. Learning factors analysis–a general method for cognitive model evaluation and improvement. In *International Conference on Intelligent Tutoring Systems*, pages 164–175. Springer.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794. ACM.

Albert T Corbett and John R Anderson. 1994. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User modeling and user-adapted interaction*, 4(4):253–278.

L Quentin Dixon, Jing Zhao, Blanca G Quiroz, and Jee-Young Shin. 2012. Home and community factors influencing bilingual childrens ethnic language vocabulary development. *International Journal of Bilingualism*, 16(4):541–565.

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.

Yongqi Gu and Robert Keith Johnson. 1996. Vocabulary learning strategies and language learning outcomes. *Language learning*, 46(4):643–679.

Mohammad Khajah, Robert V. Lindsey, and Michael C. Mozer. 2016. How deep is knowledge tracing? *CoRR*, abs/1604.02416.

Diane Larsen-Freeman and Michael H Long. 2014. *An introduction to second language acquisition research*. Routledge.

Michael J Lawson and Donald Hogben. 1996. The vocabulary-learning strategies of foreign-language students. *Language learning*, 46(1):101–135.

Yohsuke R. Miyamoto, Cody A. Coleman, Joseph J. Williams, Jacob Whitehill, Sergiy O. Nesterko, and Justin Reich. 2015. Beyond time-on-task: The relationship between spaced study and certification in moocs. *SSRN 2547799*.

Philip I Pavlik Jr, Hao Cen, and Kenneth R Koedinger. 2009. Performance factors analysis–a new alternative to knowledge tracing. *Online Submission*.

Radek Pelánek. 2017. Bayesian knowledge tracing, logistic models, and beyond: an overview of learner modeling techniques. *User Modeling and User-Adapted Interaction*, 27(3):313–350.

Chris Piech, Jonathan Bassen, Jonathan Huang, Surya Ganguli, Mehran Sahami, Leonidas J Guibas, and Jascha Sohl-Dickstein. 2015. Deep knowledge tracing. In *Advances in Neural Information Processing Systems*, pages 505–513.

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.

Xiaolu Xiong, Siyuan Zhao, Eric Van Inwegen, and Joseph Beck. 2016. Going deeper with deep knowledge tracing. In *EDM*, pages 545–550.

Jerry Ye, Jyh-Herng Chow, Jiang Chen, and Zhaohui Zheng. 2009. Stochastic gradient boosted distributed decision trees. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, pages 2061–2064, New York, NY, USA. ACM.

Belinda Young-Davy. 2014. Explicit vocabulary instruction. *ORTESOL Journal*, 31:26.

Liang Zhang, Xiaolu Xiong, Siyuan Zhao, Anthony Botelho, and Neil T. Heffernan. 2017. Incorporating rich features into deep knowledge tracing. In *Proceedings of the Fourth (2017) ACM Conference on Learning @ Scale*, L@S '17, pages 169–172, New York, NY, USA. ACM.