Personalizing BERT to predict L2 learner masked sentences

Anonymous NLP4CALL submission

Abstract

Understanding learners' knowledge and modeling the use of linguistic skills is crucial for intelligent tutoring systems and digital learning platforms in second language teaching and learning. Motivated by the capability of language models to encode language representation, we investigate the performance of a standard BERT model in predicting masked tokens in learners' written sentences in the EFCAMDAT dataset and the effect of fine-tuning BERT with personalized data. Fine-tuned models, especially those pre-trained on nationality data, show significant improvements in recall values over the base model, while the standard BERT model also performs well. Corroborating the fact that language models can encode ungrammatical language representations.

1 Introduction

Modeling language learning is at the core of personalized and adaptive learning in computer-assisted language learning systems. Many aspects of the learning process can be modeled, and the generated models can be combined together in those systems to offer a significant advantage in the learning experience by enabling the improvement of students' learning plans, scheduling, and customization, which creates an interest in developing models that are efficient but also that can be adapted for different learners and scenarios.

Different approaches have been explored to model different aspects of language learning, from modeling learner language mistakes (Settles et al., 2018) and spaced repetition models (Settles and Meeder, 2016) to efficient grammatical error correction (Omelianchuk et al., 2020). Even with the success of such diverse tasks in tackling their specific modeling objectives, they typically model a global behavior, which in the scenario of modeling learners' specific linguistic skills leads to a problem due to the dynamic nature of how students knowledge and behavior might change.

With that gap in mind, we aim at investigating an approach for modelling a second language learner linguistic skills that can be dynamically adapted and personalized over time.

Traditionally, modelling a learner linguistic skills was tackled via linguistic approaches. In the past decade however, computational and hybrid approaches have shown to be very beneficial for many Natural Language Processing (NLP) tasks focused on improving language learning demonstrated by (Knowles et al., 2016), (Renduchintala et al., 2016) and (Avdiu et al., 2019).

Automatic proficiency assessment also benefited from advances in NLP enabling such assessments at scale and allowed individuals to assess their proficiency more regularly (Portnoff et al., 2021). It also allowed institutions to assess and group students based on proficiency. Such processes have always been a bottleneck for being labour intensive and time consuming and are actively being replaced with automatic computational proficiency assessments (Gaillat et al., 2022). Another research area that tremendously benefited from modern NLP and the increasing amount of corpora and computational resources is grammar error correction. Works such as (Omelianchuk et al., 2020) enabled systems for real time corrective feedback in ungrammatical texts with Grammarly.

Similar to how these tasks benefited from the availability of large scale resources, an increasing availability of large learners corpora as well as grammatical and ungrammatical resources such as the EFCAMDAT and the C4_200m creates the possibility efficiently scaling learners' modelling tasks. Modelling a learner personalized linguistic skills may help with predicting what exercises a learner will answer wrong, which sentences learners are more likely to produce or estimate lexical

profiles. Such prior information can be utilized to improve the learning experience of L2 learners.

This paper focuses on the task of modelling language learners token outcomes in the target second language. Our methodology tries to solve two problems when we model the linguistic skills of language learner who are not proficient, 1) We expect a significant amount of produced tokens to be ungrammatical either by misspelling errors or by sentences structured differently from a native speaker. 2) Students of different backgrounds have different types of patterns. To solve those problems we propose a simple heuristic based on vocabulary frequency and edit distance to add potential misspelled tokens to the BERT vocabulary and make it possible to predict ungrammatical tokens. We pre-train different models for each student by following a top-down approach. We first pre-train on the largest common subsets of the data, 50 million sentences from 200 million synthetic ungrammatical sentences form the C4_200m, we then pre-train over writings corresponding to the learner's nationality, followed by writings corresponding to proficiency, and finally the learner's specific writings.

2 Related Work

Previous work in modeling language learning focused on mistakes. Areas such as Grammatical Error Correction Research pursued the goal of efficiently correcting ungrammatical texts to grammatical texts by adapting modern NLP models to the task and using a variety of generated resources (authentic and synthetic). A second area was detecting mistakes with the Second Language Acquisition Modeling (SLAM) task, a task in computational language learning research (Ferreira-Mello et al., 2019). The task was popularized by a Shared Task created by Duolingo (Settles et al., 2018) where several teams submitted papers on how to forecast language learners' mistakes based on their learning history and demographics.

In another direction, we find works interested in modeling the learner and its knowledge, such as (Whitehill and Movellan, 2017) that developed a new student model that accounts for how a learner infers and updates vocabulary knowledge on an ITS for foreign language learning. More related to our goal of modeling learners' knowledge and linguistic skills modeling user language learning with Macaronic texts (Renduchintala et al., 2016) train a model to simulate a human subject's comprehension of foreign words. Authors resort to engineered features based on cognate and context clues from a Macaronic text to encode learners linguistic skills. Our work differ from them in two major aspects. First, they utilize handcrafted linguistic features related to cognates, history and context which typically do not scale as well as deep learning approaches for very large datasets. Second, they emphasize on modelling the capability of a learner by translating from their native language to the target language. On the contrary, we focus on modelling the learner capability of producing L2 in a L2 context. As a result, we do not depend on collecting pairwise data between both languages but only writings from the target L2 language. Additionally, (Renduchintala et al., 2016) use graph based models which cannot benefit from pre-training strategies that modern deep learning/transformers models can. Inspired by their work, we design a similar task over language learners corpora and adopt similar evaluation metrics. Several recent works also tackled modelling linguistic skills of language learners such as (Avdiu et al., 2019) and (Zylich and Lan, 2021). They show promising results indicating the utility of recent machine learning models for encoding language learners patterns. (Palenzuela et al., 2022) explored the benefits of transfer learning for modeling Second Language Acquisition, reinforcing the motivation for using pre-trained language models.

3 Data

In this section we will introduce the EFCAMDAT dataset (Geertzen et al., 2013) and the C4200m dataset (Stahlberg and Kumar, 2021) as the main resources for the proposed task. They are two of the largest available resources as learner's corpora and synthetic ungrammatical data in English.

3.1 EFCAMDAT learner corpus

The EFCAMDAT dataset is a Second Language Acquisition corpus collected and made available by the online school of EF (Education First) containing 1180309 writings with 5897793 sentences written by 174743 learners. The corpus is composed of writings produced by learners in English and manually graded. Every writing has metadata such as grade, topic, and proficiency level. Every learner has their nationality annotated. As the tar-

get of our work is modelling the production of specific learners, we are interested in learners with a reasonable amount of essays. We selected 34 students with more than 100 essays from 10 different nationalities shown in Table 1.

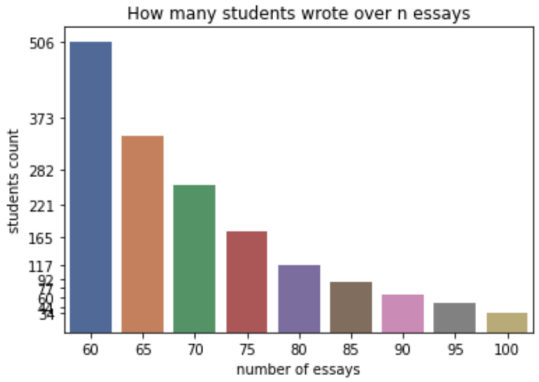

Figure 1: Count of students per number of essays on the EFCAMDAT dataset

As we achieve better performance results with our proposed models, this modelling approach could easily be extended to more learners as shown in Figure 1.

3.2 C4_200m dataset

Performance in Neural Grammatical Error Correction have been improving with new proposed neural architectures and with synthetic data generation. The C4_200m compares several models that can produce an ungrammatical sentence given a clean sentence and an error type tag following automatic annotations tools such as (Bryant et al., 2017). The generated ungrammatical sentences follow the distribution of error tags of the BEA-dev dataset. They validate the generated ungrammatical data by using it for fine-tuning and outperforming genuine parallel data on the CONLL-2014 and JFLEG-test. The dataset is generated by applying the proposed method to 200M randomly sampled sentences from the Colossal Clean Crawled Corpus (Raffel et al., 2020).

4 Methodology

4.1 Masked Language Modelling

Masked Language Modeling (MLM) is a task where, given a text with masked parts, a model is prompted to predict the missing tokens through their surrounding context. As a self-supervised task, it can be trained on massive quantities of text which are typically easily available in high resource languages. As a consequence, models have shown to learn meaningful representations of the language. Different from typical language modelling tasks that predict the next following token given an input text, the MLM task trains a bidirectional representation given relevance to both directions of the sentence. With a wide variety of tasks in NLP research benefiting from token representations learned during MLM, a natural interest is if such task would generate meaningful representations of ungrammatical structures (e.g, misspellings). In this work we identify that the MLM objective directly corresponds to the task of filling in the gaps in EFCAMDAT texts and use different ungrammatical resources in pre-training for encoding possible representations of ungrammatical structures . Typically the MLM task simply masks randomly some percentage of the tokens in the sentence and the trained model outputs a probability vector over the vocabulary. For our scenario, we only mask one token per sentence as predicting one single token is the simplest cloze task to understand the effects of the changes we propose in the models. With a good perfomance in this simpler task other more elaborated MLM strategies that are known to yield better languages representations could be explored.

4.2 User Modeling in Language Learning

User modelling generically deals with modelling how a user would respond in a task given a context clues. An instance of this in language learning is L1-L2 mixed text used to model the capability of incidental learning where using the L1 words as context, the learner can guess L2 translations (Renduchintala et al., 2016). Our work explores modelling the task in two different aspects. First guessing a ”L2 words” using ”L2 words” as context. Second, the ”L2 words” include ungrammatical L2 tokens generated by the learner instead of only correct native L2 vocabulary. In this way, we expect to implicitly encode language production patterns of a learner in the second language.

4.3 Model Architecture

We introduce BERT-based models aiming to solve the task of predicting masked tokens in the EFCAMDAT texts. BERT fits well for our scenario since its masked language modelling task is very similar to our downstream task. We use BERT as described in (Devlin et al., 2018) and made avail-

nationality	total writings	selected students	proficiency level	total writings
Brazil	476817	20	1	352106
China	165162	1	4	169291
Mexico	87259	4	2	164130
Germany	54597	2	3	109749
United States	33442	1	7	97882
Taiwan	29569	1	5	85746
United Kingdom	9000	2	6	52959
United Arab Emirates	3562	1	8	41926
Bulgaria	820	1	10	36485
Puerto Rico	645	1	9	28553

Table 1: Number of writings per nationality and proficiency levels used for pre-training

able through huggingface[1].

4.3.1 Embedding layer

The first layer in the BERT architecture is an embedding layer which maps tokens recognized by the BERT tokenizer to a vector representation. Apart from word tokens, BERT tokenizer uses WordPiece embeddings which in the case of unknown words tries to split it into smaller pieces and compound an embedding based on them. In our scenario, we expect to have many ungrammatical tokens that are not considered in BERT. For this we need to use strategies to represent those ungrammatical tokens. As manually curating which tokens should be added to the word embeddings is not feasible due to the massive amount of data. Thus, we try to automatically add tokens that are not in the BERT vocabulary or in a list of frequent English tokens but are reasonably close to at least one token in those vocabularies. A simple approach for this is to sort the tokens on our ungrammatical text based on the minimum edit distance to a token in the vocabularies and choose a threshold to filter the new vocabulary size. Similarly, the embedding representation of those new tokens will be the embedding of the closest token in the vocabulary.

4.3.2 Output layer

The Masked Language Model Output layer of BERT is a softmax over the language modelling head which outputs prediction scores for each vocabulary token. Again we have a similar problem of predicting out of vocabulary tokens since they will simply not have a prediction score. To overcome this problem, we use a similar approach by extending the dimensions of the prediction scores vector to the size of the new vocabulary with the new tokens from our ungrammatical datasets and assigning an initial weight to those new tokens based on the closest token in the vocabulary based on edit distance.

4.4 Vocabulary Modification

With the goal of mitigating negative effects of out of vocabulary tokens, we searched for tokens in the EFCAMDAT and C4_200m that were not present in the BERT vocabulary and not present in the google trillion word corpus. We show in table 2 some examples of out of the vocabulary tokens that were found. They correspond to approximately 3% (2 millions out of 72 millions) of the tokens on the EFCAMDAT dataset, which is not negligible but would have a marginal effect in our task. For datasets with more spelling errors this modification would be more critical. Given a smaller priority to this issue, we perform a simple solution by ordering the missing tokens by frequency and with an edit distance to a word in the

token	**frequency (EFCAMDAT)**
beatiful	5321
programing	1540
diferent	1456
recomend	1454
shoping	1306
peaple	1284
contry	1246
planing	1161
responsable	912

Table 2: Examples of frequent misspellings

[1] https://huggingface.co/

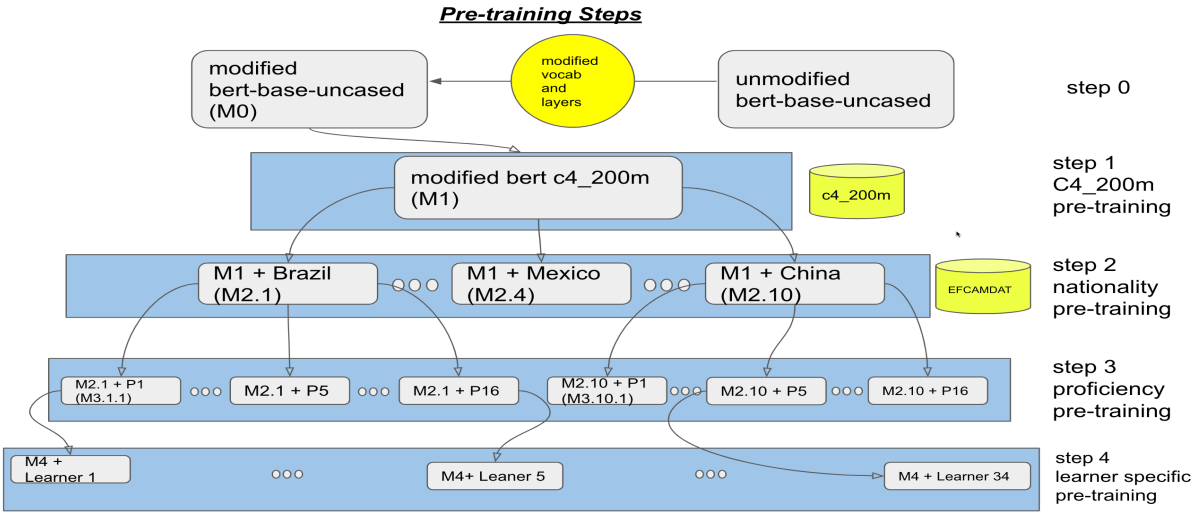

Figure 2: sequential pre-training steps described in section 3.5

vocabulary of at most 2. The intuition is to automatically add to the vocabulary misspellings that are relatively close to the correct language. After doing this selection, we add those tokens to the embeddings and output layer by simply copying the embeddings and output layer weight from any of the closest tokens.

4.5 Training steps

The main stages of developing our proposed models is injecting new vocabulary to the model and pre-training over each source of data we want to investigate. An important aspect of the training steps is that we generate more than one model that is specific to each selected learner as depicted in Figure 2. In the first step, we trained our modified vocabulary BERT over the C4_200m dataset, as it is an ungrammatical unpersonalized dataset it outputs only one single model used for all learners. The second training step creates one nationality specific model for each of the 10 nationalities of the 34 selected students. Doing the same in the third step for the proficiency of the same students. Lastly, we pre-train using specific learner data outputting a learner specific model.

5 Evaluation

Our experiment aims to evaluate two things: 1) How much the different data sources improve the precision of the models and 2) How injecting the vocabulary with potential ungrammatical tokens improve the capability of such models of recalling the masked token as candidate prediction

We use mean reciprocal ranking and average recall at k metrics, with the goal to have intuitions close to precision and recall in a classification task. We use the metrics as defined below:

$$MRR = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{rank(\hat{y}_i)}$$

$$AvgRecallatK = \frac{1}{N}\sum_{i=1}^{N}1[rank(\hat{y}_i) \leq k]$$

In both equations, the average is calculated over each sentence used for prediction in the test set. The rank function corresponds to the index where the correct label is positioned in the prediction.

6 Results

As a step to calculate isolated results, we create a test set by randomly selecting 90% of the writings of the selected learners for training the learner specific model and 10% for testing all pre-trained models. This becomes roughly 90 writings for training and 10 writings for testing per learner. So in total our test set are 340 writings from which we have 1380 sentences. For each sentence we mask a random token that contains only alpha characters, to avoid numbers and punctuations. One inconvenient aspect of the dataset is having entities such as people's, cities and other names which were relatively hard to filter automatically. Our improvements in relation to the initial base BERT model were very promising in terms of improvement in recall. One small component is related to

Model	MRR	recall at k 1	5	10	25	50	100
unmodified bert(baseline)	0.564	0.466	0.677	0.743	0.814	0.851	0.881
+ c4200m	0.552	0.460	0.666	0.712	0.777	0.803	0.830
+ nationality	0.667	0.575	0.780	0.822	0.871	0.893	0.908
+ proficiency	6	8	6	1	2	3	4
+ learner	8	7	3	1	2	3	4

Table 3: Results of each group of pre-trained models on the EFCAMDAT test set

including tokens that were not in the base BERT vocabulary but the significant improvement in all ranges of k for recall values gives evidence that by pre-training on the ungrammatical sources we're adapting well for every specific data source. Specially for the nationality data, we notice very positive results, showing how having a large scale human generated resource can bring drastic improvements. A non-intuitive result is that although most of pre-trained models achieve superior recall results for $k = 1, 5, 10$ the base BERT models have better results for higher values of k and this is reflected with a higher mrr. This shows the tail effect of such metrics.

One other important aspect is that the ordering of the pre-training can influence the results of the later models such as the learners being affected to the lower performance of the pre-training on the proficiency data source. This result gives evidence that training in different combinations might yield better results.

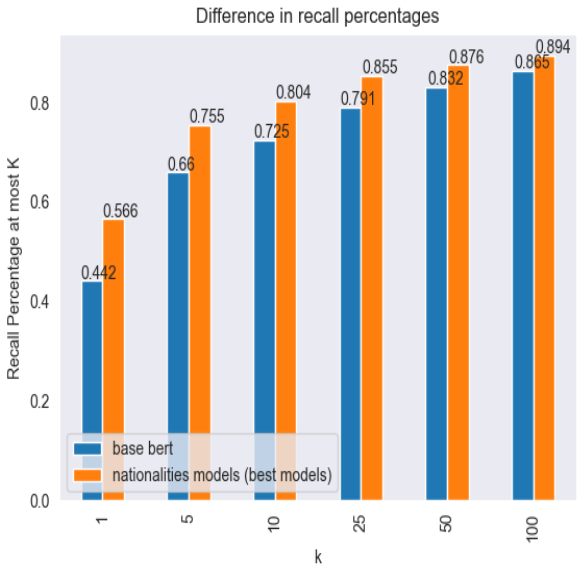

Figure 3: Difference in recall between base model and the average of the nationality models

7 Conclusion

Our work explored combining resources from learner corpora research and grammatical error correction to evaluate the feasibility of a BERT model being personalized to specific language learners in a masked token prediction task. To investigate the amount of misspelled tokens, we analyzed the frequency of misspellings in the EFCAMDAT Dataset. We concluded that out of vocabulary tokens were way below the expected in a ungrammatical dataset so we considered it as a marginal issue for the EFCAMADT dataset since only about 3% of tokens were not in the BERT vocabulary. We augment with model with a simple heuristic by adding the most frequent and close ungrammatical tokens to some token in the vocabulary to the BERT embeddings and output layer. To investigate the capability of personalization, we pre-train different models, each according to one source of data. Overall, pre-training a BERT model in those specific ungrammatical sources yielded positive results specially to the nationality data which had the biggest impact in terms of increase in all top k recall. Our intuition is that because almost half of the writings are from Brazilians, it has bias towards typical mistakes made by such students but at the same time shows evidence that BERT was able to encode those linguistics patterns. We see many directions for future work, two evident ones are direct exploring different transformers models such as MBERT and exBERT, different training tuning and strategies and investigating with error annotated sentences which types of ungrammatical structures our models performs well and which types it can have poorer performance.

References

Drilon Avdiu, Vanessa Bui, and Klára Ptačinová Klimčíková. 2019. Predicting learner knowledge of individual words using machine learning. In *Proceedings of the 8th Workshop on NLP for Computer Assisted Language Learning*, pages 1–9, Turku, Finland. LiU Electronic Press.

Christopher Bryant, Mariano Felice, and Ted Briscoe. 2017. Automatic annotation and evaluation of error types for grammatical error correction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 793–805, Vancouver, Canada. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Rafael Ferreira-Mello, Máverick André, Anderson Pinheiro, Evandro Costa, and Cristobal Romero. 2019. Text mining in education. *WIREs Data Mining and Knowledge Discovery*, 9(6):e1332.

Thomas Gaillat, Andrew Simpkin, Nicolas Ballier, Bernardo Stearns, Annanda Sousa, Manon Bouyé, and Manel Zarrouk. 2022. Predicting cefr levels in learners of english: The use of microsystem criterial features in a machine learning approach. *ReCALL*, 34(2):130–146.

Jeroen Geertzen, Theodora Alexopoulou, Anna Korhonen, et al. 2013. Automatic linguistic annotation of large scale l2 databases: The ef-cambridge open language database (efcamdat). In *Proceedings of the 31st Second Language Research Forum. Somerville, MA: Cascadilla Proceedings Project*, pages 240–254. Citeseer.

Rebecca Knowles, Adithya Renduchintala, Philipp Koehn, and Jason Eisner. 2016. Analyzing learner understanding of novel l2 vocabulary. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 126–135.

Koren Lazar, Benny Saret, Asaf Yehudai, Wayne Horowitz, Nathan Wasserman, and Gabriel Stanovsky. 2021. Filling the gaps in Ancient Akkadian texts: A masked language modelling approach. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4682–4691, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kostiantyn Omelianchuk, Vitaliy Atrasevych, Artem Chernodub, and Oleksandr Skurzhanskyi. 2020. GECToR – grammatical error correction: Tag, not rewrite. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 163–170, Seattle, WA, USA → Online. Association for Computational Linguistics.

Álvaro J Jiménez Palenzuela, Flavius Frasincar, and Maria Mihaela Truşcă. 2022. Modeling second language acquisition with pre-trained neural language models. *Expert Systems with Applications*, 207:117871.

Lucy Portnoff, Erin Gustafson, Joseph Rollinson, and Klinton Bicknell. 2021. Methods for language learning assessment at scale: Duolingo case study. *International Educational Data Mining Society*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

Adithya Renduchintala, Rebecca Knowles, Philipp Koehn, and Jason Eisner. 2016. User modeling in language learning with macaronic texts. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1859–1869, Berlin, Germany. Association for Computational Linguistics.

Burr Settles, Chris Brust, Erin Gustafson, Masato Hagiwara, and Nitin Madnani. 2018. Second language acquisition modeling. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 56–65, New Orleans, Louisiana. Association for Computational Linguistics.

Burr Settles and Brendan Meeder. 2016. A trainable spaced repetition model for language learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1848–1858, Berlin, Germany. Association for Computational Linguistics.

Felix Stahlberg and Shankar Kumar. 2021. Synthetic data generation for grammatical error correction with tagged corruption models. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 37–47, Online. Association for Computational Linguistics.

Jacob Whitehill and Javier Movellan. 2017. Approximately optimal teaching of approximately optimal learners. *IEEE Transactions on Learning Technologies*, 11(2):152–164.

Brian Zylich and Andrew Lan. 2021. Linguistic skill modeling for second language acquisition. In *LAK21: 11th International Learning Analytics and Knowledge Conference*, LAK21, page 141–150, New York, NY, USA. Association for Computing Machinery.