Predictors of Student Satisfaction: A Large-scale Study of Human-Human Online Tutorial Dialogues

Guanliang Chen
Monash University
guanliang.chen@monash.edu

David Lang
Stanford University
dnlang86@stanford.edu

Rafael Ferreira
Universidade Federal Rural de Pernambuco
rafael.ferreira@ed.ac.uk

Dragan Gasevic
Monash University
dragan.gasevic@monash.edu

ABSTRACT

For the development of successful human-agent dialogue-based tutoring systems, it is essential to understand what makes a human-human tutorial dialogue successful. While there has been much research on dialogue-based intelligent tutoring systems, there have been comparatively fewer studies on analyzing large-scale datasets of human-human online tutoring dialogues. A critical indicator of success of a tutoring dialogue can be student satisfaction, which is the focus of the study reported in the paper. Specifically, we used a large-scale dataset, which consisted of over 15,000 tutorial dialogues generated by human tutors and students in a mobile app-based tutoring service. An extensive analysis of the dataset was performed to identify factors relevant to student satisfaction in online tutoring systems. The study also engineered a set of 325 features as input to a Gradient Tree Boosting model to predict tutoring success. Experimental results revealed that (i) in a tutorial dialogue, factors such as efforts spent by both tutors and students, utterance informativeness and tutor responsiveness were positively correlated with student satisfaction; and (ii) Gradient Tree Boosting model could effectively predict tutoring success, especially with utterances from the later period of a dialogue, but more research effort is needed to improve the prediction performance.

Keywords

Intelligent Tutoring Systems, Student Satisfaction, Educational Dialogue Analysis, Gradient Tree Boosting

1. INTRODUCTION

Intelligent tutoring systems (ITS) are computer systems that are designed to act as human tutors and provide personalized instruction or feedback to students in online learning environments [3, 41]. Ultimately, ITS aim at replicating the benefits of one-to-one tutoring in contexts where students cannot receive such tutoring during the learning process [49]. In the past decades, numerous researchers have been actively involved in the investigation and development of various types of ITS, among which representative examples include AutoTutor [19], BEETLE [16], ASSISTments [25], and Cognitive Tutor [36]. More importantly, these systems have been applied in different educational contexts for hundreds of thousands of students to use and have facilitated student learning. With the aid of ITS, students with an internet connection can receive guidance tailored to their needs and enhance their learning anytime and anywhere. At the same time, instructors and educational institutions can improve their teaching quality and educational programs by analyzing the fine-grained data collected by ITS [1, 3].

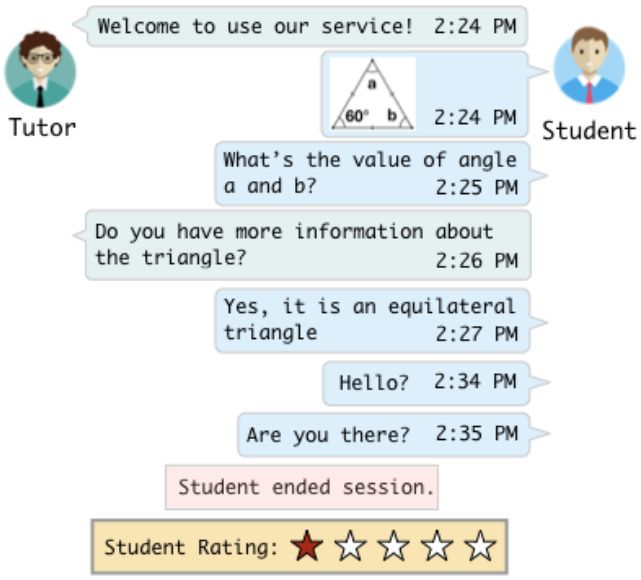

Figure 1: A tutoring dialogue example.

A special class of ITS is dialogue-based intelligent tutoring systems such as AutoTutor and BEATTLE, which emphasize the use of human-agent dialogue in one-to-one tutoring. Such systems have been built on advances in psycho-/sociolinguistics, computational linguistics, and natural language processing [14, 35] to create productive learning experiences in human-agent dialogue tutoring.

In line with [47], we argue that the future development of

dialogue-based systems can benefit greatly from the analysis of massive datasets collected in online tutoring. Online tutoring has been promoted by a growing number of applications, e.g., Chegg[1], Skooli[2] and Wyzant[3], that offer human-human online tutoring at scale and the learning subjects covered by these applications include math, science, language learning, humanities, etc. Specifically, we are interested in understanding what constitutes successful human-human online tutoring. In this paper, we report on the findings of a study that looked at factors that predict student satisfaction with human-human online tutoring.

The analysis of human-human online tutoring requires consideration of factors that shape the entire tutorial process. Online tutoring, especially helping students to solve problems, is a complex process, in which a student is expected to clearly explain the problems to be solved to a tutor and the tutor is expected to use her knowledge as well as appropriate tutoring strategies to guide the student to solve the problems. The success of such a process depends on many factors, e.g., whether the student clearly explains the problem, whether the tutor asks appropriate questions to guide the student, and whether the tutor provides sufficient emotional support. As shown in the example in Figure 1, the student ended the tutoring session because the tutor did not respond to the student in a timely manner and the student only rated the tutoring session as 1 on the scale of $(1, 5)$, which indicated that the student was not satisfied with the tutoring service at all and thus the tutorial dialogue was unsuccessful.

To our knowledge, few studies have attempted to identify the crucial factors that are correlated with the success of a tutoring session. Thus, our work aimed at (i) identifying factors that are correlated with the success of a dialogue-based tutoring session, and further (ii) utilizing the identified features as input to a state-of-the-art machine learning model to predict the tutoring success. Formally, our work was guided by the following research question:

RQ: What factors are related to student satisfaction with online tutoring service?

By investigating the RQ, we expected to (i) help tutors in existing online tutoring systems to better direct their efforts in guiding students, and (ii) inform the design of future dialogue-based ITS.

To this end, we first formulated a set of hypotheses about potential factors that were correlated with the success of a dialogue-based tutoring, which were grounded in previous research findings on online tutoring or relevant educational topics. Then, we conducted an extensive analysis of a large-scale dataset provided by a company offering online tutoring services to students, which contained transcripts of over 15,000 dialogue-based tutoring sessions generated by more than 5,000 students, to test the formulated hypotheses. Based on the analysis, we designed a set of 325 features and used these features as input to a state-of-the-art machine learning model (i.e., Gradient Tree Boosting) to predict whether a tutorial dialogue would be successful or not.

Experimental results showed that the success of a dialogue-based online tutoring session was associated with factors such as the efforts made by both tutors and students, the informativeness of the utterances, and the sentiment polarity conveyed through the utterances. We further showed that Gradient Tree Boosting was an effective method in predicting the success of tutoring sessions. In particular, we observed that the utterances from the later period of a tutoring session (e.g., the last 20% utterances in a dialogue) could deliver prediction accuracy comparable to that using the whole dialogue as input, and more research effort can be invested to further boost the prediction performance.

2. RELATED WORK

Our work is mainly related to research on educational dialogue analysis [39]. One common theme that has been investigated for years is the development and refinement of a coding scheme for educational dialogue acts [30, 38]. For instance, by building upon the language-as-action theory, [21] proposed a coding scheme that attempts to map utterances to their inherent functions in a dialogue and validated the effectiveness of the proposed scheme in two different learning contexts (one from primary school and the other from secondary school).

Another common strand of work in the field is the investigation of the relationship between tutorial dialogues and student performance [29]. For example, [47] adopted correlation analysis to capture the effects of dialogues on student performance. In particular, the dialogue acts of tutors (e.g., those related to providing explanations) were found to be significantly predictive of students' learning gain. Similarly, [5] found that the choice of corrective tutorial acts adopted by tutors, which serves as an approach to deal with incorrect problem-solving actions, has a significant influence on students' learning gain. In a different vein, [32] measured the quality of a tutorial dialogue with the Classroom Assessment Scoring System-Secondary observational instrument and demonstrated that the quality of educational dialogues was positively associated with student performance. Other relevant works include [13, 17, 24, 31].

Compared to the related works described above, our work distinguished itself in several aspects. Firstly, our work aimed at discovering factors that are related to student satisfaction instead of student performance, though both of them can be regarded as indicative predictors for the success of a tutorial dialogue. Secondly, our work analyzed various types of factors associated with student satisfaction (which are described in Section 3), while prior works have mainly analyzed one or two specific types of factors, e.g., dialogue acts [5, 47] and dialogue quality [32]. Thirdly, the tutorial dialogue dataset used in our work consists of dialogues collected from over 15,000 tutoring sessions initialized by more than 5,000 students, while previous work often used datasets containing a few hundred tutorial dialogues generated by dozens of students.

[1]https://www.chegg.com/
[2]https://www.skooli.com/
[3]https://www.wyzant.com/

3. APPROACH

In this section, we first describe the dataset we used for analysis. Then, we outline and justify the hypotheses upon which we grounded our work to explore factors that are associated with student satisfaction, and then introduce the method we used to test these hypotheses. Lastly, we describe the machine learning model for predicting student satisfaction.

3.1 Dataset

The dataset used in our work was prepared by an educational technology company that provides on-demand tutoring services via a mobile application and covers topics including mathematics, chemistry, and physics. With the mobile application, a student can take a picture of the problem she encounters or directly write down the problem and select the category to which the problem belongs to. Then, the student will be connected to a professional tutor who can guide the student to solve the problem by leveraging texts and pictures to communicate. Originally, the dataset consisted of dialogues of 18,203 tutoring sessions, which accounted for over 7,000 tutoring hours. To ensure the validity and generalizability of the experimental results, we filtered out dialogues with less than 10 utterances or of duration less than 60 seconds. This was carried out because tutors were unlikely to deliver meaningful tutoring in those sessions.

Table 1: Dataset statistics.

Category	Row ID	Metric	Value
Basic statistics	1	# Sessions	15,756
	2	# Utterances	1,250,270
	3	# Tutors	116
	4	# Students	5,468
	5	Avg. ratings	4.22
Dialogue length	6	Avg. session duration (mins)	28.75
	7	Avg. # utterances / session	79.35
	8	Avg. # words / session	610.87
	9	Avg. # unique words / session	183.80
Activeness	10	Avg. % utterances sent by tutors	57.92
	11	Avg. % words sent by tutors	78.02
	12	Avg. % new words sent by tutors	74.92
Platform experience	13	Avg. # sessions guided by tutors	135.83
	14	Avg. # sessions owned by students	2.88

After filtering, the dataset contained a total of 15,756 dialogues generated by 116 tutors and 5,468 students together, as described in Table 1. It is noteworthy that more than 79% of the dialogues received a rating of 4 or 5 (out of a scale of $(1, 5)$), as shown in Figure 2, and only about 16% of the dialogues were of rating 1 or 2. This indicates that most of the students were satisfied with the help they received from tutors and those tutoring sessions were successful.

To enable a better understanding of the characteristics of the tutor/student behavior in online tutoring, we further analyzed the dataset from the following perspectives: the length of dialogues, how active tutors/students were in dialogues, and the experiences of tutors/students in using the tutoring platform, and the results are given in Table 1 (Rows 6-14). Firstly, the average duration of all tutorial sessions is about 29 minutes, and we observed that about 50% of the sessions

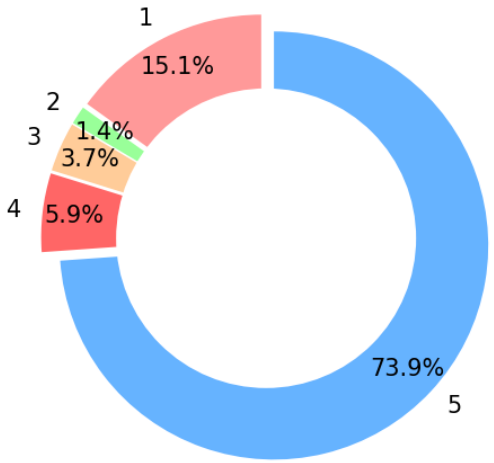

Figure 2: The distribution of student ratings for tutoring sessions.

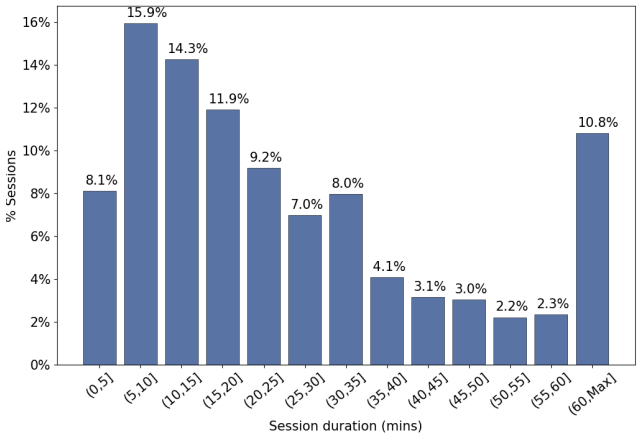

Figure 3: The distribution of the duration of tutoring sessions.

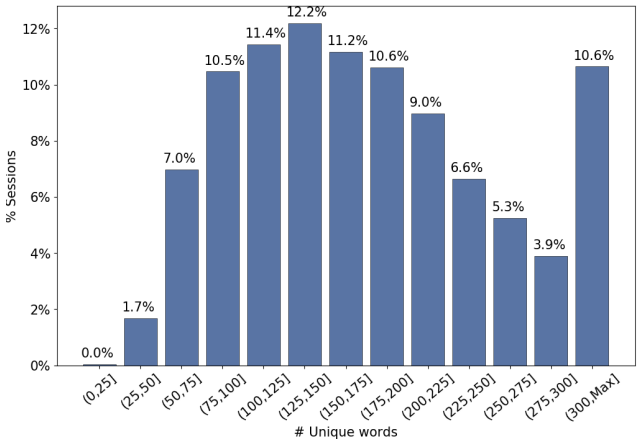

Figure 4: The distribution of the number of unique words in dialogues.

were less than 20 minutes, as shown in Figure 3. On aver-

age, about 80 utterances and 610 words were contained in a dialogue. However, we found only 184 unique words were contained in a dialogue, which was only about 30% of the average number of words used in a dialogue. In fact, most of the dialogues (over 70%) only contained 50∼250 unique words (Figure 4). This is in line with previous research finding [6], i.e., people tend to use a relatively small number of words during the conversational process. Also, we observed that tutors were more active than students, i.e., on average, 58% of the utterances in a dialogue were sent by tutors (as shown in Figure 5, tutors sent 50%∼70% of the utterances in almost 80% of the dialogues). We had a similar observation when analyzing the average fraction of words sent by tutors, i.e., in over 70% of the dialogues, tutors sent 70%∼90% of the words. In particular, 75% of the new words (i.e., words which never appeared in previous utterances) were from tutors. In fact, tutors were in charge of introducing 60%∼90% of the new words in 88% of the dialogues (as depicted in Figure 6). This implies that, most of the dialogues were led by the tutors, e.g., tutors were responsible for introducing new concepts to help students solve the problem and guiding students by providing detailed explanations. In terms of platform experience, on average, tutors guided more than 135 sessions, while students only had less than 3 sessions. A detailed analysis revealed that only 43% of the students used the tutoring service for more than once.

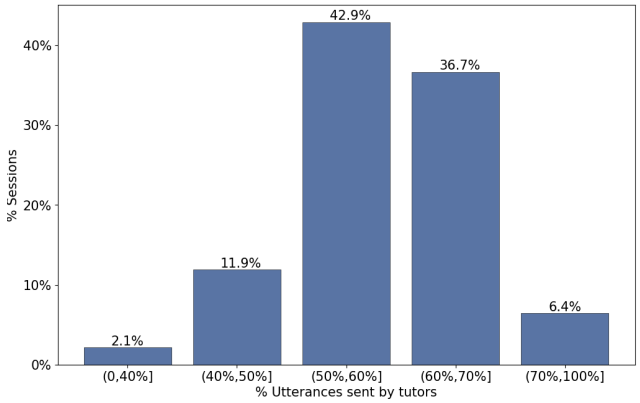

Figure 5: The distribution of the fraction of utterances sent by tutors in dialogues.

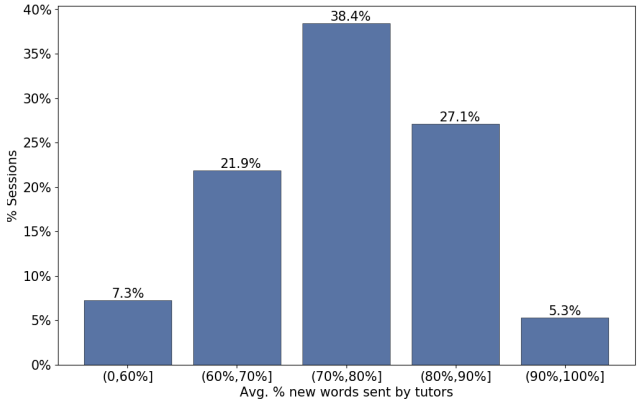

Figure 6: The distribution of the fraction of new words sent by tutors in dialogues.

3.2 Research Hypotheses

Based on prior work, we can make the following hypotheses related to our RQ.

H1 *The more efforts a student/tutor spends in a tutorial dialogue, the more likely the dialogue will be successful..*
The efforts spent by students in learning (e.g., the engagement with course materials) have long been regarded as predictive indicators of their performance [8, 15, 37]. Similarly, we hypothesized that the amount of effort spent by tutors, which directly determines how much help students can receive, also affect students' performance and tutoring success.

H2 *The more informative the utterances sent by a student/tutor are, the more likely a tutorial dialogue will be successful.*
Generally, informative tutoring feedback provided by tutors to students plays a positive role in assisting students in most learning contexts [33]. Here we argue that the informativeness of student utterances also contributes to the success of a tutoring session because it helps tutors quickly understand the difficulties faced by students and correspondingly come up with effective tutoring strategies to help the students.

H3 *The less time a student spends waiting to receive a response from a tutor, the more likely a tutorial dialogue will be successful.*
Previous research on investigating the design and delivery of feedback in online learning environments showed that not only the feedback itself but also the timing of feedback provision impacted student learning [27]. As suggested in [42], effective feedback should be timely so that students can recall the steps of addressing a learning task. Given the fact that a tutoring session is initialized by a student seeking help to solve a problem, this, to a certain extent, implies that the student lacks necessary knowledge but is eager to receive responses from a tutor and solve the problem.

H4 *The higher the lexical entrainment of a tutorial dialogue is, the more likely the dialogue will be successful.*
[6] pointed out that people involved in a conversation tend to coordinate with each other in terms of the words they use (so-called *lexical entrainment*), e.g., both the tutor and the student mentioned the word *triangle* in the dialogue in Figure 1. [34] argued that lexical entrainment is key to facilitate both production and comprehension in dialogues, and more importantly, correlated with task success.

H5 *The less complex the utterances sent by a student/tutor are, the more likely the dialogue will be successful.*
[40] suggested that, in the setting of classroom-based education, tutors should intend to gradually increase the complexity level of their verbal communication with students so as to foster students' level of competency. However, in the setting of online tutoring, in which a tutoring session usually lasts no more than half an hour (as shown in Section 3.1) and the learning task is relatively simple (e.g., solving a math problem), we hypothesized that the complexity level of tutor/student

utterances would be negatively correlated with the tutoring success as complex utterances usually take more time to understand and respond.

H6 *The more questions a tutor/student asks, the more likely a tutorial dialogue will be successful.*
Previous research demonstrated that questioning is an essential method for tutors to help students build up their understanding and promote effective learning [20, 44, 50]. Likewise, questions asked by a student represent the student's activeness in learning what is unknown to her at the moment and are generally viewed as positively related to her learning performance [43, 45, 46].

H7 *The more positive sentiment conveyed through the utterances sent by a student/tutor, the more likely a tutorial dialogue will be successful.*
[9, 51] suggested that students' sentiment expressed via forum posts in a MOOC is correlated with the retention rate of the course. This led us to postulate that the success of a tutoring session could be revealed by students' sentiment conveyed in the dialogue utterances. Also, the positive sentiment contained in tutors' utterances, e.g., those used to encourage students, can be indicative of the success of a tutoring session.

H8 *The more prior tutorial dialogues a student/tutor has, the more likely the current dialogue will be successful.*
On the one hand, the number of prior tutorial dialogues that tutors have can be used to estimate their prior tutoring experience, which is generally believed to have a positive effect on students' learning outcome [48]. On the other hand, if a student has multiple sessions before the current session, this may imply that (i) the student is familiar in using the tutoring platform; and (ii) the tutoring platform has gained the trust of the student by providing satisfactory learning experiences; thus the student repeatedly returns to the platform and uses the service.

3.3 Hypotheses Testing

To test the formulated hypotheses, we first classified tutorial dialogues receiving ratings of 4 or 5 from students as the *Success* group and those of ratings of 1 or 2 as the *Failure* group. Then, we defined a set of metrics to describe the factors investigated in each hypothesis and compared the two groups with Mann-Whitney test on the relevant metrics to test our hypotheses.

For **H1**, we quantified the efforts of tutors/students made in a tutoring session from three perspectives:

M1 **Session duration**: the duration of a tutoring session;

M2 **# Utterances**: the number of utterances made by a tutor/student;

M3 **# Words**: the number of words contained in the utterances made by a tutor/student;

To investigate **H2**, we considered four metrics to measure utterance informativeness:

M4 **# Unique words**: the number of unique words contained in utterances sent by a tutor/student;

M5 **# Unique concepts**: the number of concepts contained in utterances introduced by a tutor/student;

M6 **% New words**: the fraction of unique words sent by a tutor/student for the first time (so-called *new words*);

M7 **% New concepts**: the fraction of unique concepts introduced by a tutor/student for the first time (so-called *new concepts*);

Counting the number of unique words (**M4**) was one indicator to measure the informativeness of utterances. Besides, given that both tutors and students often use concepts during the conversational problem-solving process (as shown in Figure 1 where *triangle* was mentioned by both the tutor and the student), and such concepts often bring new information, we also calculated the number of unique concepts (**M5**) to measure the utterance informativeness. As concepts typically appear as nouns, we extracted the nouns contained in an utterance and used them as proxies to capture the mentioned concepts. For this, we use NLTK[4] to extract nouns from utterances. In addition, we also defined **M6** and **M7** to measure the extent to which the new words/concepts were spoken by tutors/students, as indicators to distinguish the main contributor in bringing new information in a dialogue.

We tested **H3** from two angles:

M8 **Wait time**: the amount of time between a student initialized a request for help and the student was connected to a tutor;

M9 **Avg. response time**: the average amount of time that a student needed to wait before receiving a reply from a tutor after the student sent an utterance;

To investigate **H4**, we defined the following metric:

M10 **Entrainment**: the score describing the level of entrainment between the tutor utterances and the student utterances;

Inspired by [4], we calculated **M10** as the similarity between the distribution of respective words used by tutors and students. Specifically, we first needed to decide the set of considered words used to calculate entrainment score. [34] suggested that function words (i.e., frequent words like *is*, *do*, *can*) and punctuation marks are important for measuring the degree to which people align with each other in successful dialogues. Therefore, we took all of the words appearing in dialogues into account for calculating **M10** (denoted as *Entrainment (All)*). In addition, as indicated before, both tutors and students often use concepts during the tutoring process and we hypothesized that the entrainment between such concepts was of particular importance to indicate whether a dialogue would be successful or not. Therefore, we also calculated **M10** by only considering concepts (denoted as *Entrainment (Concepts)*). Similar to **H2**, we extracted nouns in utterances and regarded them as the concepts mentioned by tutors and students. With the set of considered words defined, we counted the occurrence of

[4]https://www.nltk.org

each word for the utterances made by a tutor/student in a dialogue, respectively, which were represented as two vectors (one for the tutor and the other for the student). Then, we measured the similarity of the two vectors by computing their cosine similarity [22], which served to describe how close the tutor and the student were in terms of the vocabulary they used in a dialogue and thus to indicate to what extent they coordinated the words to each other.

To test **H5-8**, we respectively defined the following metrics:

M11 Complexity: the average complexity of utterances sent by a tutor/student;

M12 # questions: the number of questions asked by a tutor/student;

M13 Sentiment: the overall sentiment polarity scores of utterances sent by a tutor/student;

M14 Platform experience: the number of tutorial dialogues that a tutor/student has prior to the current one;

Specifically, we measured the complexity (**M11**) of an utterance by calculating its Flesch readability score [12], which specify to what extent a piece of text is readable to people by returning a score between $[0, 100]$. A piece of text with a higher Flesch readability score indicates it is easier to understand. Given that questions often ended with a question mark, we, therefore, computed **M12** as the number of sentences ending with a question mark in the utterances made by a tutor/student. For **M13**, we again use *NLTK* to determine the sentiment polarity score for each utterance. The returned score was of range $(-1, 1)$ with -1 being very negative and 1 being very positive. Then, the values of all utterances sent by a tutor/student were summed up as the overall score of the tutor/student in a dialogue, and the scores of all dialogues in a group were averaged as the final score for the group.

3.4 Tutoring Success Prediction

We aimed to predict whether a tutoring session would be successful or not based on the transcript of the tutorial dialogue, which could be regarded as a binary classification problem. Previous research indicated that there are various techniques that can be used for binary classification problems, such as logistic regression, decision trees, random forests, support vector machines, and neural networks. Gradient Tree Boosting (GTB) [11, 18] is a machine learning technique which can be used for both regression and classification problems. Similar to random forests, GTB is based on the belief that multiple predictors aiming to predict the same target variable will do a better job than any single predictor alone. Therefore, GTB constructs a set of predictors (i.e., decision trees), which are typically trained with a random sub-sample of the data (thus each predictor is slightly different from the others) and the predictions of all predictors are taken into account to give a final prediction. In random forests, the predictors are built independently and the predictions are combined by using techniques like weighted average and majority vote. However, in GTB, the predictors are built sequentially in which the later predictors can learn from mistakes committed by previous predictors and thus reduce prediction errors. This usually takes less time to reach close to actual predictions. Previous research has demonstrated that GTB is one of the most robust machine learning approaches and can deal with various types of feature data and has reliable predictive power when dealing with unbalanced data (as in our case) [10]. Therefore, we select GTB over other approaches for our prediction task.

We used all of the metrics described in Section 3.3, i.e., M1-14, as features for the Gradient Tree Boosting model. Note that we calculated M2-7, M11-14 for both tutors and students, respectively. M10 was calculated by taking all of the words as well as only the concepts into consideration. In addition, as a common practice in solving text classification problems, we extracted *N-grams* features, i.e., unigrams, bigrams, and trigrams, from the dialogue transcripts. Prior to the N-grams extraction, we preprocessed the dialogue transcripts by removing stopwords (e.g., *can*, *a*, *be*, *is*, *are*), which are of high frequency but seldom carry useful information for classification purposes. To avoid overfitting, we only took the top 100 most frequent unigrams, bigrams, and trigrams into consideration. In total, we designed 325 features.

To set up the experiment, we randomly sampled 80% of the data as the *training* dataset, and the remaining 20% as the *validation* and *testing* datasets (10% for each). To demonstrate the effectiveness of GTB in predicting tutoring success, we selected random forests as the baseline method for comparison. We implemented random forests as well as GTB by using the machine learning library `scikit-learn`[5] for Python. The parameters for both random forests and GTB were optimized through grid search on the validation dataset, and then we evaluated the models' performance on the testing dataset. In line with previous works on classification problems, especially those dealing with imbalanced data, we adopted three representative metrics for measuring the competency of the two methods, i.e., Area Under the Curve (AUC), F1 score, and Cohen's kappa coefficient (Cohen's κ) [23]

In particular, the design of our experiments was guided by the following three questions:

Q1 How does GTB perform in predicting the success of a tutorial dialogue?

Q2 How much data is needed to successfully predict tutoring success?

Q3 Which of the designed features are of particular importance for the prediction performance?

To our knowledge, there have been few works attempting to predict the success of a tutorial dialogue with a large-scale dataset and our work has contributed to this by enabling a better understanding of this problem. By investigating **Q1**, we expected to examine the capability of GTB, which is regarded as a state-of-the-art machine learning technique, in solving this particular prediction task. Previous works demonstrated that the earlier a student is identified as being at risk, the more help a tutor can offer to help the student

[5]https://scikit-learn.org/

continue to learn. Similarly, the earlier an unsuccessful tutoring session can be identified, the more effective intervention can be provided to the student by a tutor. Therefore, by investigating **Q2**, we expected to identify how early an unsuccessful tutoring session can be identified and to shed light on the practicability of using GTB to assist tutors during their interaction with students. Lastly, by answering **Q3**, we expected to examine the contributions made by each type of features for the prediction performance.

4. RESULTS

In this section, we describe the experimental results on hypotheses testing as well as tutoring success prediction.

4.1 Results on Hypotheses Testing

For **H1**, we calculated the mean values of M1-3 for all the dialogues contained in the *Failure* and *Success* groups, which are given in Table 2 (so as the other results for H2-8). Based on the results, we observed that the *Success* dialogues were 50% longer than the *Failure* dialogues (31.20 vs. 19.27). In addition, compared to the *Failure* dialogues, both tutors and students had more utterances in the *Success* dialogues. We had similar observations when comparing the number of words sent by tutors/students in the two groups. Therefore, we conclude that H1 was supported.

To validate **H2**, we computed M4-7 over all the utterances sent by a tutor(student) in a dialogue and summed up the values to measure how informative the tutor(student) was in the dialogue. Then, the metric values of all dialogues contained in a group were averaged as the final value. We found that both tutors and students used a higher number of unique words as well as unique concepts (M4-5) in the *Success* group than those in the *Failure* group. In particular, students of *Success* group used about 50% more unique words and concepts than their peers in the *Failure* group. These results suggest that, in order to solve problems, both tutors and students in the *Success* group introduced a greater variety of words during tutoring process and thus were more informative. Interestingly, when inspecting the results of M6-7, we found that the *Success* students introduced a larger fraction of new words as well as new concepts compared to the *Failure* students, and correspondingly the *Success* tutors were less active than the *Failure* tutors in bringing new words and concepts to their dialogues. This motivates us to design further experiments to investigate, during the tutoring process, whether tutors should intentionally encourage students to use more new words and concepts to explain problems as well as their thoughts so as to help students solve the problems. To summarize, the observed results indicate that H2 was supported by the analysis of our dataset.

For **H3**, we only observed a significant difference between the two groups in terms of M9. Compared to *Failure* students, *Success* students spent less time (about 5 seconds) in waiting for responses from tutors. Thus, there was some support for H3.

From the reported results of M10, we concluded that **H4** was supported, i.e., the tutors and students were more likely to align with each other in terms of the words they used in the *Success* group than in the *Failure* group. In particular, when only taking concepts into account, we had a slightly higher entrainment score, which implies a higher degree to which tutors and students coordinated concepts than other words in the tutorial dialogues.

Table 2: Results on validating the formulated hypotheses. T represents tutors and S represents students. Significant differences (according to Mann-Whitney test) between *Failure* group and *Success* group are marked with ** ($p < 0.001$).

Hypotheses	Metrics	Failure	Success
H1	Session length (mins) **	19.27	31.20
	# Utterances (T) **	28.79	51.46
	# Utterances (S) **	21.32	35.36
	# Words (T) **	315.95	518.21
	# Words (S) **	82.33	146.44
H2	# Unique words (T) **	117.1	157.12
	# Unique words (S) **	47.43	74.52
	# Unique concepts (T) **	102.84	138.43
	# Unique concepts (S) **	41.73	64.34
	# New words (T) **	76.75	74.38
	# New words (S) **	23.25	25.62
	# New concepts (T) **	76.39	74.66
	# New concepts (S) **	23.61	25.34
H3	Wait time	24.09	24.37
	Avg. response time **	32.93	27.89
H4	Alignment (All) **	0.83	0.86
	Alignment (Concepts) **	0.87	0.89
H5	Complexity (T) **	83.93	85.11
	Complexity (S)	100.71	101.26
H6	# Questions (T) **	10.34	17.14
	# Questions (S) **	1.85	4.05
H7	Sentiment (T) **	4.58	9.38
	Sentiment (S) **	1.54	3.32
H8	Experience (T)	160.66	162.56
	Experience (S) **	9.11	12.67

By inspecting the results of M11, we discovered that the utterances made by tutors in the *Success* group were slightly less complex than those in the *Failure* group (85.11 vs. 83.93). However, we did not observe a significant difference between the utterances made by students in the two groups. Therefore, **H5** was only supported for tutors.

For **H6**, the results of M12 were in line with our assumption: both tutors and students asked more questions in successful tutoring sessions than those in unsuccessful ones. Particularly, the *Success* students asked more than two times of questions than *Failure* students. Also, it is important to note that, in both groups, tutors asked many more questions than students (about 4∼5 times). This is aligned with our previous findings related to the testing of **H1**: tutors tended to make more efforts than students in tutoring sessions.

For **H7**, we noted that the tutors as well as students in the *Success* group displayed a higher level of positive sentiment

than those in the *Failure* group. Also, the tutors were more likely to use words of positive sentiment than students in both groups. Therefore, we concluded H7 was supported.

Lastly, we observed a significant difference on M14 for students, i.e., the platform experience of the *Success* students was significantly higher than that of the *Failure* students. Thus, we concluded that **H8** was only supported for students.

4.2 Results on Tutoring Success Prediction

For **Q1** described in 3.4, i.e., whether GTB is capable of predicting the success of a tutoring session, we reported the results of GTB as well as the baseline method (random forests) in Rows 1-2 in Table 3. The results indicated that GTB outperformed random forests on all of three evaluation metrics. This demonstrated the effectiveness of GTB for predicting whether a tutoring session will be successful or not. Particularly, GTB attained an improvement of 21% over random forests in terms of Cohen's κ, though the value was only 0.4323, which implied that the constructed prediction model achieved a *moderate* performance level [28]. This calls for further research effort in developing more effective prediction models for this particular task.

To answer **Q2**, we trained GTB by using different portions of the dialogue utterances, i.e., the first 20%/40%/60%/80%. The results are reported in Rows 3-6 in Table 3. To our surprise, even using the first 80% of the data to train GTB, the achieved performance was still much inferior to that of using the whole dataset. For instance, the AUC of using the first 80% data was 0.7368, which was 10% lower than that of using the whole dataset. When it comes to Cohen's κ, the difference became even larger (28% lower). This may imply that the utterances made by tutors and students in the later stage of a tutoring dialogue (especially the last 20%) possibly contained more information for predicting the success of a tutoring session. This motivated us to train the model with the last 20%/40%/60%/80% of the dialogue utterances and reported their performance in Rows 7-10 in Table 3. The results aligned with our assumption. Specifically, solely using the last 20% data already achieved performance that was comparable to that of using the whole dataset. For AUC, it even achieved slightly better performance. This could be explained by the fact that, at the end of successful tutoring sessions, tutors tended to praise students and acknowledge their achievements and students were likely to express their gratitude to the tutors. As a sanity check, we randomly selected 100 successful and unsuccessful dialogues and checked the last 20% utterances in these dialogues. We found that, most of the successful dialogues contained N-grams like *(appreciate)*, *(thanks)*, *(well, done)*, and *(good, job)*, which were seldom observed in unsuccessful dialogues. Undoubtedly, these linguistic features served as good indicators for GTB to determine a dialogue's success.

Lastly, we conducted an ablation study to answer **Q3**. An ablation study is a frequently-used method to determine to what extent a feature contributes to the performance of a model. Typically, the contribution of a feature is determined by comparing the performance of a model including the feature with that without the feature. The more the performance decreases after removing a feature, the more contribution the feature makes to the model. Instead of identifying the contributions made by each feature we engineered, we were more interested in determining the contributions made by each type of features, i.e., the eight types of feature investigated in 3.3 (*efforts*, *informativeness*, *responsiveness*, *entrainment*, *complexity*, *questions*, *sentiment* and *platform experience*) and the linguistic features (*unigrams*, *bigrams*, *trigrams*). Therefore, we removed each type of feature at a time and reported the model performance in Row 11-20 in Table 3.

We observed that, the top 3 types of feature that made the most contributions to the prediction performance were *unigrams*, *bigrams*, and *efforts*. This was in line with the observation we had when answering **Q2**, i.e., the linguistic features were predictive in terms of distinguishing successful dialogues from unsuccessful ones.

5. DISCUSSION AND CONCLUSION

Implications for Online Tutors. Through the extensive analysis presented in Section 3.3, we demonstrated that student satisfaction is correlated with a set of dialogue features, which include (i) the efforts invested by tutors/students; (ii) the informativeness of tutor/student utterances; (iii) the readability level of tutor utterances; (iv) tutor responsiveness; (v) the number of questions asked by tutors/students; (vi) the entrainment level of a tutorial dialogue; (vii) the positive sentiment level of tutor/student utterances; and (viii) students' experience in using the tutoring service. This may shed some light on how to better direct online tutors' efforts in guiding students. For example, tutors may consider to provide prompt responses, use more words of positive sentiment and suitable readability level, and ask a suitable number of questions to assist students to solve problems. However, it should be noted that the identified dialogue features (as well as the corresponding tutoring implications) may be correlated with each other, e.g., the increased number of utterances might introduce a higher number of questions asked by tutors/students. Further experiments, e.g., online A/B testing, are needed to verify which factors are actually affecting student satisfaction in this context. Also, it is necessary to further investigate whether there are any other factors contributing to the observed correlation. For example, though we observed that students' experience (measured by the number of tutoring sessions they had before) is associated with their satisfaction, it is still unclear whether this is because of students' familiarity in using the platform, which enables them to quickly find a tutor and solve a problem, or because of their established loyalty in using the tutoring service. For the former case, it would be beneficial to develop guidelines to help novice students quickly learn how to use the tutoring service. For the latter case, it would be necessary to scrutinize the tutoring sessions that students had before so as to better investigate the elements contributing to students' established loyalty for the tutoring platform.

Improvement space for satisfaction prediction. Our study demonstrated that Gradient Tree Boosting model is effective in predicting tutoring success with all of the utterances or the utterances from the later period of a tutorial dialogue as input. However, this might be of little value to improve online tutoring service in the real-world setting, i.e.,

Row ID	Method	Data usage	Features	Compared row	AUC	F1	Cohen's κ
1	Random Forests	All	All	-	0.7913	0.8885	0.3571
2	GTB			1	0.8225 (↑ 3.95%)	0.9018 (↑ 1.49%)	0.4323 (↑21.05%)
3	GTB	First 20%	All	2	0.6847 (↓16.75%)	0.8612 (↓ 4.50%)	0.1584 (↓63.37%)
4		First 40%			0.7088 (↓13.83%)	0.8705 (↓ 3.47%)	0.2098 (↓51.47%)
5		First 60%			0.7086 (↓13.84%)	0.8783 (↓ 2.60%)	0.2473 (↓42.79%)
6		First 80%			0.7368 (↓10.42%)	0.8880 (↓ 1.52%)	0.3115 (↓27.95%)
7	GTB	Last 20%	All	2	0.8275 (↑ 0.61%)	0.8735 (↓ 3.13%)	0.3901 (↓ 9.76%)
8		Last 40%			0.8388 (↑ 1.98%)	0.8808 (↓ 2.32%)	0.4038 (↓ 6.58%)
9		Last 60%			0.8349 (↑ 1.51%)	0.8924 (↓ 1.04%)	0.4239 (↓ 1.93%)
10		Last 80%			0.8271 (↑ 0.56%)	0.8882 (↓ 1.51%)	0.3754 (↓13.17%)
11	GTB	All	w/o Efforts	2	0.8217 (↓ 0.09%)	0.8887 **(↓ 1.45%)**	0.3749 **(↓13.29%)**
12			w/o Infomativeness		0.8145 **(↓ 0.97%)**	0.8965 (↓ 0.58%)	0.4032 (↓ 6.73%)
12			w/o Complexity		0.8205 (↓ 0.24%)	0.8961 (↓ 0.63%)	0.4018 (↓ 7.06%)
13			w/o Responsiveness		0.8170 (↓ 0.66%)	0.8974 (↓ 0.49%)	0.4180 (↓ 3.31%)
14			w/o Questions		0.8196 (↓ 0.35%)	0.9002 (↓ 0.17%)	0.4213 (↓ 2.54%)
15			w/o Entrainment		0.8230 (↑ 0.06%)	0.8988 (↓ 0.33%)	0.4205 (↓ 2.73%)
16			w/o Sentiment		0.8204 (↓ 0.26%)	0.8985 (↓ 0.36%)	0.4156 (↓ 3.86%)
17			w/o Experience		0.8178 (↓ 0.57%)	0.8974 (↓ 0.49%)	0.4180 (↓ 3.31%)
18			w/o Unigrams		0.8045 **(↓ 2.18%)**	0.8818 **(↓ 2.22%)**	0.3692 **(↓14.59%)**
19			w/o Bigrams		0.8168 **(↓ 0.69%)**	0.8954 **(↓ 0.70%)**	0.3829 **(↓11.44%)**
20			w/o Trigrams		0.8233 (↑ 0.10%)	0.8993 (↓ 0.27%)	0.4302 (↓ 0.49%)

Table 3: Experimental results on tutoring success prediction. The percentage value within brackets indicates the increased/decreased (denoted by ↑/↓, respectively) performance of evaluation metrics, which were computed by taking the results of the compared row as a comparison. The results in bold represent the top 3 decreased performance among Rows 10-20.

if an unsuccessful dialogue can only be identified (almost) until the end, there is not much a tutor can do to change the situation. Therefore, more research is needed to build effective satisfaction prediction models, especially with only the utterances close to the beginning of a dialogue as input. Since we only engineered relatively shallow linguistic features (i.e., unigrams, bigrams, trigrams) as input for the prediction model, which made much larger contributions to the prediction performance compared to other types of feature, it is worthwhile to explore more in-depth linguistic features (e.g., word/phrase/sentence embedding [2]) to boost the prediction performance. Also, noteworthy is that all the features we designed as input for Gradient Tree Boosting model is derived from dialogue utterances without considering the sequential nature between them. In the future, it would be useful to explore the suitability of time series models to capture the underlying time-aware interaction patterns between tutors and students for this prediction task. In addition, it is recognized that data imbalance (as in our case) can have a big impact on the classification performance [26]. We posit that techniques used to reduce impacts of data imbalance like SMOTE [7] would probably help in future research on this problem.

6. ACKNOWLEDGEMENTS

The authors would like to acknowledge (i) the Institute for Educational Sciences and Grant Number R305B14009, and (ii) the funding of the Faculty of Education at Monash University through the Education Futures initiative.

7. REFERENCES

[1] A. Alkhatlan and J. Kalita. Intelligent tutoring systems: A comprehensive historical survey with recent developments. *arXiv preprint arXiv:1812.09628*, 2018.

[2] S. Arora, Y. Liang, and T. Ma. A simple but tough-to-beat baseline for sentence embeddings. In *ICLR 2017*, 2017.

[3] R. S. Baker. Stupid tutoring systems, intelligent humans. *IJAIED*, 26(2):600–614, 2016.

[4] L. Benotti, J. Bhaskaran, and D. Lang. Modeling student response times: Towards efficient one-on-one tutoring dialogues. In *NUT@EMNLP*, 2018.

[5] K. E. Boyer, R. Phillips, M. D. Wallis, M. A. Vouk, and J. C. Lester. Learner characteristics and feedback in tutorial dialogue. In *Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications*, pages 53–61. Association for Computational Linguistics, 2008.

[6] S. E. Brennan. Lexical entrainment in spontaneous dialog. *Proceedings of ISSD*, 96:41–44, 1996.

[7] C. Bunkhumpornpat, K. Sinapiromsaran, and C. Lursinsap. Safe-level-smote: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem. In T. Theeramunkong, B. Kijsirikul, N. Cercone, and T.-B. Ho, editors, *Advances in Knowledge Discovery and Data Mining*, pages 475–482, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg.

[8] R. M. Carini, G. D. Kuh, and S. P. Klein. Student engagement and student learning: Testing the linkages*. *Research in Higher Education*, 47(1):1–32, Feb 2006.

[9] D. S. Chaplot, E. Rhim, and J. Kim. Predicting student attrition in moocs using sentiment analysis and neural networks. In *AIED Workshops*, 2015.

[10] N. V. Chawla. Data mining for imbalanced datasets: An overview. In *Data mining and knowledge discovery handbook*, pages 875–886. Springer, 2009.

[11] T. Chen and C. Guestrin. 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, 2016.

[12] K. Collins-Thompson. Computational assessment of text readability: A survey of current and future research. *ITL-International Journal of Applied Linguistics*,

165(2):97–135, 2014.

[13] M. G. Core, J. D. Moore, and C. Zinn. The role of initiative in tutorial dialogue. In *EACL*, 2003.

[14] M. W. Crocker. *Computational psycholinguistics: An interdisciplinary approach to the study of language*, volume 20. Springer Science & Business Media, 2012.

[15] J. Davies and M. Graff. Performance in e-learning: online participation and student grades. *British Journal of Educational Technology*, 36(4):657–663, 2005.

[16] M. Dzikovska, N. Steinhauser, E. Farrow, J. Moore, and G. Campbell. Beetle ii: Deep natural language understanding and automatic feedback generation for intelligent tutoring in basic electricity and electronics. *IJAIED*, 24(3):284–332, 2014.

[17] K. Forbes-Riley, D. Litman, A. Huettner, and A. Ward. Dialogue-learning correlations in spoken dialogue tutoring. In *Proceedings of the 2005 Conference on Artificial Intelligence in Education: Supporting Learning Through Intelligent and Socially Informed Technology*, pages 225–232, Amsterdam, The Netherlands, The Netherlands, 2005. IOS Press.

[18] J. H. Friedman. Stochastic gradient boosting. *Computational statistics & data analysis*, 38(4):367–378, 2002.

[19] A. C. Graesser, K. Wiemer-Hastings, P. Wiemer-Hastings, R. Kreuz, T. R. Group, et al. Autotutor: A simulation of a human tutor. *Cognitive Systems Research*, 1(1):35–51, 1999.

[20] J. Hattie. *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. routledge, 2008.

[21] S. Hennessy, S. Rojas-Drummond, R. Higham, A. M. MÃąrquez, F. Maine, R. M. RÃ■os, R. GarcÃ■a-CarriÃşn, O. Torreblanca, and M. J. Barrera. Developing a coding scheme for analysing classroom dialogue across educational contexts. *Learning, Culture and Social Interaction*, 9:16 – 44, 2016.

[22] A. Huang. Similarity measures for text document clustering. In *Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand*, volume 4, pages 9–56, 2008.

[23] L. A. Jeni, J. F. Cohn, and F. De La Torre. Facing imbalanced data–recommendations for the use of performance metrics. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, pages 245–251. IEEE, 2013.

[24] S. Katz, G. O'Donnell, and H. Kay. An Approach to Analyzing the Role and Structure of Reflective Dialogue. *IJAIED*, 11:320–343, 2000. Part I of the Special Issue on Analysing Educational Dialogue Interaction (editor: Rachel Pilkington).

[25] K. R. Koedinger, E. A. McLaughlin, and N. T. Heffernan. A quasi-experimental evaluation of an on-line formative assessment and tutoring system. *Journal of Educational Computing Research*, 43(4):489–510, 2010.

[26] B. Krawczyk. Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, 5(4):221–232, Nov 2016.

[27] J. A. Kulik and C.-L. C. Kulik. Timing of feedback and verbal learning. *Review of Educational Research*, 58(1):79–97, 1988.

[28] J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174, 1977.

[29] N. Maharjan, V. Rus, and D. Gautam. Discovering effective tutorial strategies in human tutorial sessions. In *The Thirty-First International Flairs Conference*, 2018.

[30] J. C. Marineau, P. M. Wiemer-Hastings, D. Harter, B. A. Olde, P. Chipman, A. Karnavat, V. Pomeroy, S. Rajan, and A. Graesser. Classification of speech acts in tutorial dialog. 2000.

[31] D. E. Meltzer. Relation between students âĂŹ problem-solving performance and representational format. 2005.

[32] H. Muhonen, E. Pakarinen, A.-M. Poikkeus, M.-K. Lerkkanen, and H. Rasku-Puttonen. Quality of educational dialogue and association with studentsâĂŹ academic performance. *Learning and Instruction*, 55:67 – 79, 2018.

[33] S. Narciss and K. Huth. Fostering achievement and motivation with bug-related tutoring feedback in a computer-based training for written subtraction. *Learning and Instruction*, 16(4):310 – 322, 2006.

[34] A. Nenkova, A. Gravano, and J. Hirschberg. High frequency word entrainment in spoken dialogue. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, HLT-Short '08, pages 169–172, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics.

[35] D. Nguyen, A. S. Doğruöz, C. P. Rosé, and F. de Jong. Computational sociolinguistics: A survey. *Computational linguistics*, 42(3):537–593, 2016.

[36] J. F. Pane, B. A. Griffin, D. F. McCaffrey, and R. Karam. Effectiveness of cognitive tutor algebra i at scale. *Educational Evaluation and Policy Analysis*, 36(2):127–144, 2014.

[37] T. Phan, S. G. McNeil, and B. R. Robin. Students âĂŹ patterns of engagement and course performance in a massive open online course. *Computers & Education*, 95:36–44, 2016.

[38] R. Pilkington. *Analysing educational discourse: The DISCOUNT scheme*. University of Leeds, Computer Based Learning Unit, 1999.

[39] R. Pilkington. Analysing educational dialogue interaction: Towards models that support learning. *IJAIED*, 12:1–7, 2001.

[40] S. Podschuweit, S. Bernholt, and M. Brückmann. Classroom learning and achievement: how the complexity of classroom interaction impacts studentsâĂŹ learning. *Research in Science & Technological Education*, 34(2):142–163, 2016.

[41] J. Psotka, L. D. Massey, and S. A. Mutter. *Intelligent tutoring systems: Lessons learned*. Psychology Press, 1988.

[42] J. R. Anderson, A. T. Corbett, K. Koedinger, and R. Pelletier. Cognitive tutors: Lessons learned. *Journal of the Learning Sciences*, 4:167–207, 04 1995.

[43] L. B. Resnick and L. E. Klopfer. *Toward the Thinking Curriculum: Current Cognitive Research. 1989 ASCD Yearbook*. ERIC, 1989.

[44] C. P. Rosé, D. Bhembe, S. Siler, R. K. Srivastava, and K. VanLehn. The role of why questions in effective human tutoring. 2003.

[45] A. Taboada and J. T. Guthrie. Growth of cognitive strategies for reading comprehension. *Motivating reading comprehension: Concept-oriented reading instruction*, pages 273–306, 2004.

[46] A. Taboada and J. T. Guthrie. Contributions of student questioning and prior knowledge to construction of knowledge from reading information text. *Journal of literacy research*, 38(1):1–35, 2006.

[47] A. K. Vail and K. E. Boyer. Identifying effective moves in tutoring: On the refinement of dialogue act annotation schemes. In *Intelligent Tutoring Systems*, 2014.

[48] H. J. Van Berkel and D. H. Dolmans. The influence of tutoring competencies on problems, group functioning and student achievement in problem-based learning. *Medical Education*, 40(8):730–736, 2006.

[49] K. VanLehn. The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. 2011.

[50] J. A. Walsh and B. D. Sattes. *Quality questioning: Research-based practice to engage every learner*. Corwin Press, 2016.

[51] M. Wen, D. Yang, and C. P. Rosé. Sentiment analysis in mooc discussion forums: What does it tell us? In *EDM*,

2014.