Modeling One-on-one Online Tutoring Discourse using an Accountable Talk Framework

Renu Balyan
SUNY College at Old Westbury,
New York, USA
balyanr@oldwestbury.edu

Tracy Arner
Arizona State University,
Arizona, USA
tarner@asu.edu

Karen Taylor
Arizona State University,
Arizona, USA
karnetaylor.sb@gmail.com

Jinnie Shin
University of Florida,
Florida, USA
jinnie.shin@coe.ufl.edu

Michelle Banawan
Arizona State University,
Arizona, USA
mbanawan@asu.edu

Walter L. Leite
University of Florida,
Florida, USA
walter.leite@coe.ufl.edu

Danielle S. McNamara
Arizona State University,
Arizona, USA
dsmcnamara1@gmail.com

ABSTRACT

The National Council of Teachers of Mathematics (NCTM) has been emphasizing the importance of teachers' pedagogical communication as part of mathematical teaching and learning for decades. Specifically, NCTM has provided guidance on how teachers can foster mathematical communication that positively impacts student learning. A teacher may have different academic goals towards what needs to be achieved in a classroom, which require a variety of discourse-based tools that allow students to engage fully in mathematical thinking and reasoning. Accountable or academically productive talk is one such approach for classroom discourse that may ensure that the discussions are coherent, purposeful and productive. This paper discusses the use of a transformer model for classifying classroom talk moves based on the accountable talk framework. We investigate the extent to which the classroom Accountable Talk framework can be successfully applied to one-on-one online mathematics tutoring environments. We further propose a framework adapted from Accountable Talk, but more specifically aligned to one-on-one online tutoring. The model performance for the proposed framework is evaluated and compared with a small sample of expert coding. The results obtained from the proposed framework for one-on-one tutoring are promising and improve classification performance of the talk moves for our dataset.

Keywords

accountable talk framework, classroom discourse, one-on-one online tutoring, transfer learning.

R. Balyan, T. Arner, K. Taylor, J. Shin, M. Banawan, W. Leite, and D. McNamara. Modeling one-on-one online tutoring discourse using an accountable talk framework. In A. Mitrovic and N. Bosch, editors, *Proceedings of the 15th International Conference on Educational Data Mining*, pages 477–483, Durham, United Kingdom, July 2022. International Educational Data Mining Society.

https://doi.org/10.5281/zenodo.6852936

1. INTRODUCTION

Productive classroom discourse is positively associated with student learning [6,7,13,15] across multiple content areas, including reading comprehension [18], academic vocabulary learning [9] development of collaborative reasoning [14], persuasive writing performance [1] historical reasoning [26], scientific argumentation [5], and mathematical reasoning [13]. Additionally, academically rigorous classroom discussions are explicitly promoted in mathematical pedagogy [3,19] and in the Common Core State Standards [20] that guide teachers' instructional practices.

The oft-cited, common pattern of classroom talk, Initiation-Response-Evaluation (IRE) [11] emerged in early research investigating the forms and functions of classroom talk [12]. This minimal unit of interactional exchange includes a teacher's initiation, then a student's response, followed by the teacher's evaluation of the response. The IRE pattern is most commonly noted in teacher-led lessons (i.e., direct instruction) which tends to be the default in many classrooms. The IRE pattern of classroom talk demonstrates a transmission style of instruction such that the teacher is in a position of authority and controls the content of the discourse as well as who engages in it [4]. Indeed, this pattern of monologic classroom discourse is still common in many classrooms despite research demonstrating the benefit of dialogic and reasoning-based discourse [6,15,17].

Dialogic discourse, on the other hand, aligns with Vygotsky's [25] sociocultural theory stipulating that learning is more likely to occur when thinking is socialized, particularly within the range of a student's ability when provided with appropriate guidance (i.e., Zone of Proximal Development; ZPD). Interactions between teachers and students serve as a scaffold for knowledge acquisition during instruction [21,24]. A hallmark of this type of scaffolded, dialogic classroom discourse is the *Accountable Talk framework* [15].

2. ACCOUNTABLE TALK FRAMEWORK

The Accountable Talk Framework divides teacher and student discourse into three broad categories (see Table 1 for definitions and respective examples): accountability to the community, accountability to accepted standards of reasoning, and accountability to knowledge [15]. Talk that is accountable to the community is characterized as cutting across disciplines and "attends seriously to and builds on the ideas of others" [15, p. 286]. Accountability to standards of reasoning involves making logical and reasonable connections, explanations, and conclusions. Accountability to knowledge emphasizes sensemaking based on facts and authoritative knowledge. The latter categories of discourse practices promoting accountability to standards of reasoning and knowledge are generally more discipline-specific.

Table 1. Accountable Talk framework teacher and student talk moves definitions and examples

Talk Moves Definition and Example(s)	
Teacher Talk Move	
Learning Community	
Keeping Together	Keeping everyone together; prompting students to be active listeners and orienting students to each other. For example, turn to the student sitting next to you; what do you think; What did X just say the equation was?
Relating Students	Getting students to relate to other students' idea; prompting students to respond to what a classmate said. For example, would someone like to add on to what X said; do you agree with Y that the answer is; does anyone understand how Z solved this problem?
Restating	Repeating all or in part of what a student said. For example, Student: the factors, Teacher: factors
Content Knowledge	
Accuracy	Pressing for accuracy; prompting students to make a mathematical contribution or use mathematical language. For example, can you give an example of X; What's another word for that?
Rigorous Thinking	
Revoicing	Repeating what a student said but changing the wording or adding to it; using at least a key mathematical word or idea from what the student said. For example, Student: it is x squared; Teacher: so, instead of a cube it will be squared.
Reasoning	Pressing for reasoning; prompting students to explain or provide evidence. For example, can you explain why you think this is the answer; why would you add two and not three?
Student Talk Move	
Learning Community	
Relate to another student	Comment on or ask questions about another student's idea or use another student's idea to form your own basis. For example, I got the same answer as X; I was about to say what Y just said.
Asking for more information	Ask for help in case confused or request more information about a math topic. For example, can you please explain this again; I did not understand this; Is this multiplication or division?
Content Knowledge	
Making a claim	Make a factual statement about a mathematical concept; list a step to arrive at an answer; make a mathematical claim. For example, I got this answer by dividing the two numbers; X is the profit.
Rigorous Thinking	
Providing evidence or explanation	Provide the evidence or explanation for the mathematical claim or explain their thinking. For example, you cannot divide the number by zero as that will result in an infinite number; four multiplied by five gives me twenty.

Prior research evaluating implementation and benefit of discourse using dialogic frameworks was conducted by transcribing and hand coding hours and hours of classroom audio or video recordings [22]. This method is labor intensive and time consuming thus making it difficult to provide teachers with feedback on their implementations of academic discourse. Suresh and colleagues [22] used automatic speech recognition and machine learning to evaluate teacher-student interactions using the Accountable Talk framework. The machine learning models trained as part of their TalkBack application showed promise as a medium to provide teachers with important feedback regarding their pedagogical practices. However, this work has only been applied to traditional classroom instruction. Therefore, less is known about the application of the Accountable Talk framework in discourse occurring outside of the traditional classroom. The recent shift to hybrid or online learning necessitated by the COVID-19 pandemic, provides a rich opportunity to apply machine learning models to teacher-student interaction in an online environment. The capability to evaluate academic discourse and provide feedback to teachers that may lead to improved student learning is particularly important given the concern over learning gaps being exacerbated by the pandemic [2].

The present study builds on this work using the machine learning models developed by Suresh and colleagues to detect evidence of the Accountable Talk framework [15] in teacher-student interactions in online, one-on-one tutoring. For the purpose of this study, we adopted the same talk moves outlined in Suresh [22]. We further refined our analysis by including moves that would be included in a one-on-one tutoring interaction and excluded those that could only occur in traditional, whole class settings. Specifically, talk moves related to accountability to the community were removed from data for the model to be re-trained for the proposed framework as they are focused on ensuring learners have a shared focus and on developing collegial interactions between students. Since there was only one student and one teacher in each tutoring session, no utterances exemplifying these two talk moves were present in our data.

3. METHOD

3.1 Participants

The student population was recruited for online tutoring in summer and fall of 2019 from four high schools in the Broward school district. Participants included 40 students who did not pass the Algebra 1 course or were not successful on the end of course (EOC) exam. Tutoring was provided during school hours on school grounds in the summer or fall semester. Sessions were planned for up to 10 hours, but the number of hours varied from 1 to 20 with a mean of 5 hours of tutoring. Tutoring was conducted by credentialed math teachers with at least two years of experience in the Florida system to ensure that they were familiar with the targeted curriculum standards. Tutoring sessions were conducted online using Study Edge's GoBoard video conferencing system which supports shared note-taking (pen-casts). During the online tutoring sessions, both the student's and the tutor's computer screens were recorded, which included an audio and video recording of their conversation. The audio recordings were later transcribed to obtain the student-teacher discourse during the tutoring session. This study was approved by the Institutional Review Board and all participants in this study consented to the use of their data in accordance with APA guidelines.

3.2 Data

Three different data sets were used in this study. The first set was the talk moves labeled data available from prior research. The second data set was a subset of the first set that was filtered for talk moves. The second data set was more aligned with the one-on-one tutoring discourse. The third data set consisted of the unlabeled data containing teacher and student utterances from one-on-one tutoring discourse. Examples of teacher and student talk moves from the data sets discussed below are shown in Table 2.

Table 2. Talk Moves examples (Teacher Talk Moves – 1: none; 2: Keeping everyone together, 3: getting students to relate, 5: revoicing, 6: Press for accuracy; **Student Talk Moves –** 1: none, 2: relating to another student, 3: asking for more information, 4: making a claim, 5: providing evidence/explaining reasoning**)**

Speaker	Sentence	Teacher Tag	Student Tag
Teacher	Who can help her notice where she went wrong	3	nan
Teacher	<<student1 name>>	2	nan
Student	She kept the six	nan	2
Teacher	<<student2 name>>	2	nan
Student	She put the eight so everything after the decimal number that you need changed	nan	5
Teacher	You underline you look at an arrow then what	6	nan
Student	Whats that one negative 51	nan	3
Teacher	Mixed numbers you were right	5	nan
Student	Its like one and one half or one and three thirds or something like that	nan	4
Teacher	All right I see some disagreements here	1	nan
Student	I was waiting	nan	1

3.2.1 Data Set 1

The Data Set 1 included annotated transcripts of classroom discourse in mathematics collected in public schools in the US and are, therefore, aligned with our Algebra tutoring data set. This talk moves data was obtained from [23], which included entire lessons and short excerpts from lessons. The data set consisted of 230,778 utterances (172,309 teacher utterances and 58,469 student utterances) from 559 lesson transcripts.

3.2.2 Data Set 2

The full data set (Data Set 1) was filtered for talk moves applicable to one-on-one tutoring discourse from the original data set [23]. The filtered dataset consisted of 199,123 utterances (147,145 teacher utterances and 51,978 student utterances) from 558 lesson transcripts.

3.2.3 Data Set 3

The tutoring data used to classify the talk moves consists of 87,100 utterances 62,370 teacher and 24,730 student utterances). The final dataset consisted of transcripts from 130 one-on-one tutoring sessions between 25 teachers and 39 students.

3.2.4 Talk Moves data set creation via expert coding

Expert coding of the data was conducted in two phases: training and data coding. Training was conducted using the Data Set 1 consisting of transcripts for individual student-teacher conversations. The transcripts in this original data were coded for six mutually exclusive teacher talk moves, and five student talk moves that were adapted from the Accountable Talk framework. The teacher or student utterances that did not contain a talk move were labeled as "None-1". Strings of utterances by the teachers that were aimed at direct instruction were considered "None-1" as they were not aimed at fulfilling one of the six accountable talk moves outlined in the framework. Four undergraduate research assistants were trained to code a sample of each teacher and student 'talk move'.

The first step in training the coders involved acquainting them with each of the types of talk moves explicated in the Accountable Talk framework [15]. Research assistants met with a researcher and discussed examples of each type of Talk Move for both teachers and students. Importantly, coders learned to determine which utterances were part of direct instruction or not consequential (e.g., classroom management utterances such as "get out your textbook") to eliciting student responses; these are coded as "1-None". The talk moves being classified in this framework consist of turns of an interaction between student(s) and teachers with the goal of knowledge construction. Following the initial group training, each coder was given one of two practice sets created from the large data set provided by [23]. Equal numbers of each type of move were included in the data set such that coders had 100 target sentences for each of the six mutually exclusive teacher talk moves and five of the student talk moves. The teacher and student talk moves were separated (i.e., one practice sheet for teachers and one for students) and randomized so that no patterns could be detected. The interrater reliability (weighted kappa) for the first pair of coders for teacher practice set A was 0.44 and 0.61 for student practice set A. The interrater reliability (weighted kappa) for the second pair of coders was 0.65 for both teacher and student practice set B. The moderately low interrater reliability suggested that either coders were not quite familiar enough with the codes or that the utterances were too difficult to code when taken out of the context of the discourse. Using the student utterance was necessary to accurately determine the context, particularly for the restating and revoicing teacher talk moves. Additionally, multiple utterances may be included in one sentence and there may be multiple sentences included within a turn. Therefore,

all four coders were given a second practice data set (revised version) that included utterances in context. For example, coders received spans of sentences that included both the teacher utterances and the student utterances. Prior to starting on the new practice data set, coders were provided with a short training refresher video. The interrater reliability for the student tags in the second practice set ranged from 0.42 to 0.74 with a weighted kappa of 0.60 for all raters. The interrater reliability for the teacher tags in the second practice set ranged from 0.52 to 0.71 with a weighted kappa of 0.60 for all raters. Weighted kappa ranging from 0.41 to 0.60 suggest moderate agreement [8]. Following the training sets, the two coders with the highest interrater reliability were provided with a subset of Data Set 3 to code, containing 570 teacher utterances and 183 student utterances. The interrater reliability for the experimental data was 0.90 weighted kappa for teacher utterances and 0.95 weighted kappa for student utterances. Weighted kappa over 0.81 suggest nearly perfect agreement [8].

3.2.5 Model architecture

Because our data (Data Set 3) was not labeled/coded for talk moves initially, we used the labeled data (Data Sets 1 and 2) available from prior research [23] and the filtered data set for initial model development. The labeled data was split into training (75%), validation (5%) and test (20%) sets to verify that the prior research results could be replicated. We used the RoBERTa-base transformer model [10], a model pretrained with Masked Language Modeling (MLM) using five data sets including BookCorpus - a dataset consisting of 11,038 unpublished books; English Wikipedia; CC-News - a dataset containing 63 million English news articles; OpenWebText - an open-source recreation of the WebText dataset used to train GPT-2; and Stories - a dataset containing a subset of CommonCrawl data, to train our model. RoBERTa allows the model to learn a bidirectional representation of a sentence.

The models were trained using sentences that were preprocessed into *turns*, where each turn constituted a student utterance followed immediately by a teacher utterance for the teacher talk moves model. Similarly, for student talk moves, each turn consisted of a teacher utterance immediately followed by a student utterance. A teacher or a student turn could also include multiple utterances. There were very few student-student pair utterances as compared to the original/prior data (classroom setting) because our data constituted one-on-one tutoring transcripts. Both the student and teacher utterances were required to set the context particularly for the *restate* and *revoice* teacher talk moves, and *make a claim* and *provide evidence or reasoning* student talk moves. The input sentences were also cleaned of any punctuation and converted to lowercase. The pretrained model from the Hugging Face library was used and the model was trained using a Tesla P100 GPU. Code was implemented using TensorFlow framework with Python version 3.7. The batch_encode_plus method was used for tokenizing and encoding the pair of sentence sequences and AdamW was the optimizer. The hyperparameters used for training and tuning the model included a learning rate of 2e-5, and number of epochs and batch size as 4. The talk moves were used as the dependent variables.

3.2.6 Analytic framework

The study was performed as a set of five experiments. First, a transformer model was trained and tested using Data Set 1 [23], the model trained using Data Set 1 was then used to classify our unlabeled tutoring data (Data Set 3) for both student and teacher talk moves. The second experiment was performed to compare the consistency between the distributions of the talk moves for Data Set 1 and the predicted talk moves for Data Set 3. In the third experiment, after determining the consistency between the talk moves distribution of the two data sets (Data Set 1 and Data Set 3), the prior data (i.e., Data Set 1) was filtered for talk moves that are more applicable to one-on-one online tutoring. The teacher talk moves *keeping everyone together* and *getting students to relate* were removed as these talk moves are more relevant to classroom teaching rather than one-on-one online tutoring. Similarly, the utterances labeled as student talk move *relating to another student* were also removed. Prior research data, filtered for one-on-one online tutoring aligned talk moves (Data Set 2), was used to re-train and re-test the model. The re-trained model was used to classify our unlabeled data (Data Set 3) again for both the teacher and student one-on-one tutoring related talk moves. The distribution of the talk moves for the filtered datasets was compared again to confirm whether the newly trained model classified the talk moves consistently with the prior research data that was labeled in the fourth experiment. The fifth experiment evaluated the performance of the model that was trained using talk moves applicable to one-on-one tutoring. The predicted talk moves for Data Set 3 were compared with the expert coded talk moves for a small sample of Data Set 3.

4. RESULTS

4.1 Model Performance for Data Set 1: Original Framework

The first set of experiments included training a transformer model using the Data Set 1 (training set) and classifying the teacher and student talk moves on the test set for the talk moves defined in the original accountable talk framework. The confusion matrix for the teacher and student talk moves for Data Set 1 (test set) are shown in Figure 1 (a and b), respectively. The micro F1 and macro F1 scores obtained for the teacher test set were 0.89 and 0.79, respectively. The Matthew correlation coefficient (MCC) was 0.79. The F1 scores for the student talk moves were lower than the teacher talk moves and were 0.80 (micro F1) and 0.76 (macro F1), and the MCC was 0.71. The precision, recall, and F scores of the talk moves for both teachers and students were also computed. The teacher talk moves *getting students to relate* and *revoicing* had the lowest F scores, 0.68 and 0.67 respectively. The teacher talk moves *press for reasoning*, *press for accuracy*, and *restating* had higher and comparable F scores (0.80-0.84). The student talk move *relating to another student* performed the worst with an F score of 0.58 and other student talk moves had higher and comparable F scores (0.76-0.79).

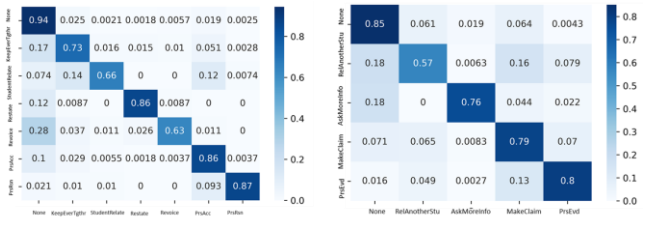

Figure 1. (a) Teacher Test set Talk Moves

Figure 1. (b) Student Test set Talk Moves

Figure 1. The teacher and student talk move accuracy and labels: None (*no talk move*), KeepEverTgthr (*keeping everyone together*), StudentRelate (*getting students to relate to other students' ideas*), Restate (*restating*), Revoice (*revoicing*), PrsAcc (*press for accuracy*), and PrsRsn (*press for reasoning*), RelAnotherStu (*relating to another student*), AskMoreInfo (*asking for more information*), MakeClaim (*making a claim*), and PrsEvd (*providing evidence or explaining reasoning*).

4.2 Talk Moves Distribution Comparison (Data Set 1 vs. Data Set 3): Original Framework

The second experiment included comparing the talk moves distribution of the original data (Data Set 1) and the predictions for our data (Data Set 3) obtained using the model trained in the first experiment. The distribution of the teacher and the student talk moves across the two datasets was found to be comparable and consistent with slight variations. Both datasets found 67.42-76.03% of utterances did not contain a teacher talk move. The remaining teacher talk moves in order of frequency were, *pressing for accuracy* (10.59-13.07%), *keeping everyone together* (9.44-12.97%), and *revoicing* (2.15-2.27%). *Getting students to relate to another student's idea* was the least classified (0.14%) teacher talk move in our dataset (Data Set 3). The remaining teacher talk moves shown in Figure 2 (a and b) had similar distributions.

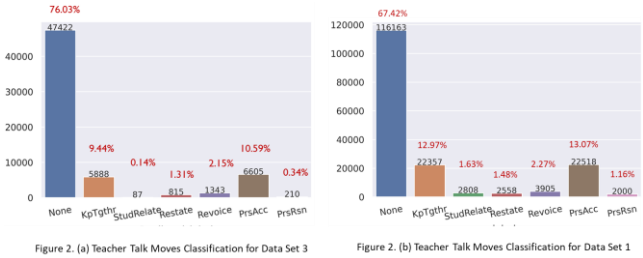

Figure 2. (a) Teacher Talk Moves Classification for Data Set 3

Figure 2. (b) Teacher Talk Moves Classification for Data Set 1

Figure 2. The teacher talk move distribution and labels for Data sets 3 and 1: None (*no talk move*), KpTgthr (*keeping everyone together*), StudRelate (*getting students to relate to other students' ideas*), Restate (*restating*), Revoice (*revoicing*), PrsAcc (*press for accuracy*), and PrsRsn (*press for reasoning*).

The predicted student talk moves also had consistent distribution with the prior labeled data (Data Set 1). Following *not a talk move*, the student talk move *making a claim* was the second most abundant with 28.25-30.66% utterances falling into that category. *Asking for more information* was the least observed talk move in both the data sets (see Figure 3 (a and b)).

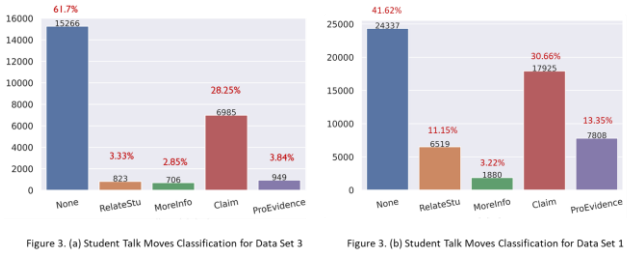

Figure 3. (a) Student Talk Moves Classification for Data Set 3

Figure 3. (b) Student Talk Moves Classification for Data Set 1

Figure 3. The student talk move distribution labels for Data sets 3 and 1 are: None (*no talk move*), RelateStu (*relating to another student*), MoreInfo (*asking for more information*), claim (*making a claim*), and ProEvidence (*providing evidence or explaining reasoning*).

The consistent distributions of talk moves across the two datasets combined with very few utterances classified in the talk moves categories of *getting students to relate to another student's idea* (teachers) and *relate to another student* (students) suggest that the classification model trained on Data Set 1 can be used for Data Set 3. The two minimally classified talk moves are not directly applicable or aligned to one-on-one tutoring thus increasing our confidence that the classification model can be used for our one-on-one tutoring dataset with minor updates to the model.

4.3 Model Performance for Data Set 2: Proposed Framework

Experiment 3 was similar to the first experiment except the transformer model was trained using prior data for talk moves that are more aligned and applicable to one-on-one tutoring (i.e., Data Set 2) rather than the original set (Data Set 1) of talk moves. Figure 4 (a and b) shows the confusion matrix for the teacher and student talk moves for the test data from Data Set 2, respectively. The micro F1 for the teacher test set improved from 0.89 to 0.94 and macro F1 score increased from 0.79 to 0.83, respectively. The Matthew correlation coefficient (MCC) was higher (0.79 vs. 0.83). The F1 scores for the student dataset were lower than the teacher talk moves but improved from 0.80 to 0.86 (micro) and 0.76 to 0.82 (macro), and the MCC is higher (0.71 vs 0.78). The precision, recall, and F scores for each talk move, for both teachers and students, are also computed. The teacher talk move *revoicing* had the lowest F score, 0.67, which aligns with the original accountable talk framework. The student talk move *asking for more information* performed the worst with an F score of 0.76.

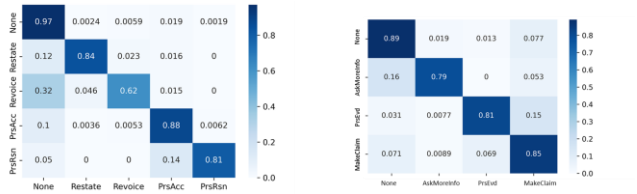

Figure 4.a. Teacher Test set Talk Moves – Proposed Framework

Figure 4.b. Student Test set Talk Moves – Proposed Framework

Figure 4. The teacher and student talk move labels for the proposed framework are: None (*no talk move*), Restate (*restating*), Revoice (*revoicing*), PrsAcc (*press for accuracy*), and PrsRsn (*press for reasoning*) and the student talk moves in Figure 4(b) are: None (*no talk move*), AskMoreInfo (*asking for more information*), PrsEvd (*providing evidence or explaining reasoning*) and MakeClaim (*making a claim*).

4.4 Talk Moves Distribution Comparison (Data Set 2 vs. Data Set 3): Proposed Framework

This experiment is comparable to the second experiment except that the distributions of talk moves are for the data aligned to one-on-one tutoring instead of the talk moves from the original accountable talk framework [22,23]. The teacher talk moves distributions for both Data Set 2 and the predicted talk moves for Data set 3 using the model trained in the third experiment were found to be consistent with *press for reasoning* having the lowest frequency (0.38% and 1.36%) and *press for accuracy* as the second highest frequency (11.84% and 15.3%) next to *no talk move*.

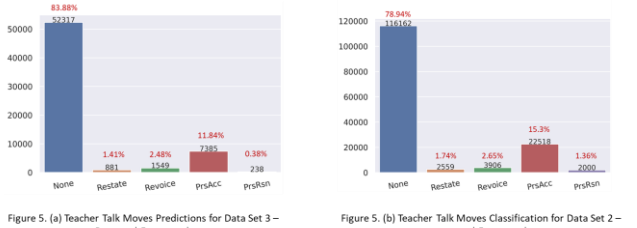

Figure 5. (a) Teacher Talk Moves Predictions for Data Set 3 – Proposed Framework

Figure 5. (b) Teacher Talk Moves Classification for Data Set 2 – proposed Framework

The distributions for both Data Set 2 and the predictions of student talk moves for Data Set 3 were similar except for the *providing evidence* talk move (see Figure 6 (a and b)). Very few (3.39% utterances) were classified as *providing evidence* talk moves for

Data Set 3, whereas the original labeled filtered dataset had a higher percentage (~15%) of talk moves belonging to this category.

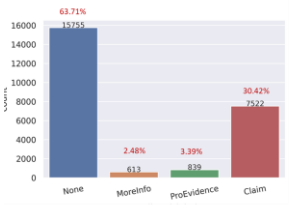

Figure 6. (a) Student Talk Moves Predictions for Data Set 3 – Proposed Framework

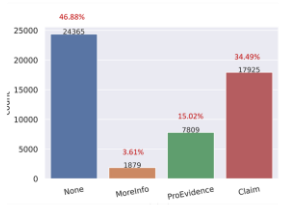

Figure 6. (b) Student Talk Moves Classification for Data Set 2 – proposed Framework

4.5 Model Performance for Data Set 3: Proposed Framework

The student and teacher talk move predictions using the proposed framework model for Data Set 3 when compared with the expert coding achieved 0.88 accuracy for teacher talk moves and 0.81 for student talk moves. The weighted average and micro precision, recall and F scores were in the range of 0.87-0.88 for teacher talk moves and 0.79-0.83 for the student talk moves. The teacher talk move *pressing for reasoning* had the worst performance and the student talk move *asking for more information* performed the worst of all the talk moves. The confusion matrix for both the talk moves are shown in Figure 7 (a and b). The teacher and student talk moves labels in the Figures 7 (a and b) are: *None*; *restating*; *revoicing*; *pressing for accuracy* and *pressing for reasoning, and None*; *asking for more information*; *providing evidence or explaining reasoning*; and *making a claim for (1-5) and (1-4) respectively*.

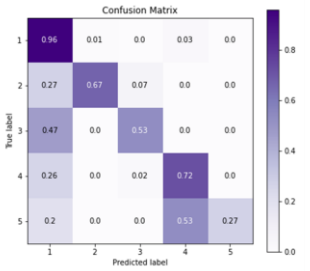

Figure 7. (a) Teacher Talk Moves Classification Predicted vs. Expert Coding – proposed Framework

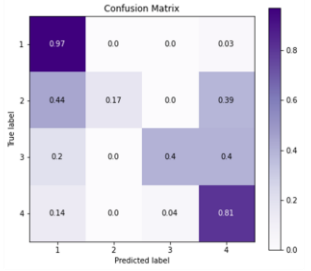

Figure 7. (b) Student Talk Moves Classification Predicted vs. Expert Coding – proposed Framework

5. DISCUSSION

The Accountable Talk framework has been used to classify teacher and student talk moves as a method of providing feedback and guidance to teachers on the discourse strategies they use in their classrooms. Until recently, this type of discourse analysis has been cumbersome and labor intensive. Furthermore, it has been restricted to face-to-face or in-person classrooms mostly. The present work extends the work of Suresh [22,23] in the application of deep learning to a transcribed data set of student and teacher utterances gathered from an online, one-on-one algebra tutoring system. The data set provided by Suresh [23] was used to train our model which was then applied to our unlabeled data set. The data available from prior research (Data Set 1) was used to train a transformer model (RoBERTa) and the test set achieved an accuracy higher than that mentioned in the literature on a similar data set for several deep learning models including Long Short Term Memory (LSTM) unit, Bi-LSTM, gated recurrent unit (GRU), and recurrent neural network (RNN). The Bi-LSTM had outperformed all the other models with an F1 measure of 65% [23]. However, our model RoBERTa achieved an accuracy of 0.89 and 0.94 for the original framework and the proposed framework respectively for the teacher talk moves. The student talk move model accuracy for the original framework and the proposed framework are 0.80 and 0.86 respectively. The talk move distributions for both teacher and student for our data set (Data Set 3) were also found to be consistent with the prior data (Data Set 1) and the filtered prior data (Data Set 2). In addition, high performance was observed for most of the talk moves for the proposed framework model when applied to our unlabeled data set (Data Set 3) and compared with the expert coding. The teacher talk move *press for reasoning* and the student talk move *asking for more information* performed the worst when compared with the expert coding. One probable reason for low performance could be very few instances belonging to these talk moves in the test set or a skewed distribution of the talk moves. Therefore, this needs to be explored further to determine the reason for low performance of these specific talk moves.

6. LIMITATIONS

The work described in this paper is novel and not without limitations. The first limitation is that we could not evaluate the classification model for Data Set 3 using Data Set 1, the model based on the full Accountable Talk framework from discourse in a traditional classroom. Our data was gathered from one-on-one online tutoring, which inherently prevents classification of talk moves that require additional community members as in the original dataset based on classroom discourse.

The second limitation to note is the limited amount of human-coded data to evaluate the classification of talk moves by the proposed model. We recognize that a larger percentage of human-coded validation data is desirable and plan to address this limitation in future work.

7. CONCLUSION

The academic discourse that occurs between teachers and students in the interest of knowledge creation has a rich history of research demonstrating the importance of each turn. Despite the importance of these classroom exchanges, providing feedback to teachers is nearly impossible due to the labor-intensive task of collecting and categorizing discourse according to an evidence-based framework such as the Accountable Talk framework [16].

Recent events have demonstrated that academic discourse is not restricted to traditional, whole group classroom instruction; therefore, it is equally important to evaluate the quality of discourse occurring in online, one-on-one tutoring sessions. The work described in this paper is both novel and promising as a means to reliably categorize teacher and student talk moves by applying machine learning models slightly modified from those validated on a complete framework. Future work is needed to further test and validate (with human coding) the machine learning model that was modified to accommodate one-on-one instruction (i.e., without talk moves for *accountability to the community)*. Overall, this work has the potential to introduce a beneficial and simplified mechanism to provide feedback for teachers, thus affording strong potential to improve instructional practice and student learning outcomes.

8. ACKNOWLEDGMENTS

The research reported here was supported by the Institute of Education Sciences (IES), U.S. Department of Education through Grant R305C160004, the National Science Foundation (NSF) through Award# 2131052, and the Office of Naval Research (ONR) through Grant N000142012623. The opinions expressed are those of the authors and do not represent views of the IES, the NSF, or the ONR.

9. REFERENCES

[1] Al-Adeimi, S., & O'Connor, C. (2021). Exploring the relationship between dialogic teacher talk and students' persuasive writing. *Learning and Instruction*, 71, 101388.

[2] Bailey, D. H., Duncan, G. J., Murnane, R. J., & Au Yeung, N. (2021). Achievement gaps in the wake of COVID-19. *Educational Researcher*, *50*(5), 266-275.

[3] Candela, A. G., Boston, M. D., & Dixon, J. K. (2020). Discourse actions to promote student access. *Mathematics Teacher: Learning & Teaching, 113*, 266-277.

[4] Cazden, C. B. (1988). *Classroom discourse: The language of teaching and learning* (2nd ed.). Portsmouth, NH: Heinemann.

[5] Chin, C., & Osborne, J. (2010). Supporting argumentation through students' questions: Case studies in science classrooms. *Journal of the Learning Sciences, 19*, 230-284.

[6] Correnti, R., Stein, M. K., Smith, M. S., Scherrer, J., McKeown, M., Greeno, J., & Ashley, K. (2015). Improving teaching at scale: Design for the scientific measurement and learning of discourse practice. In L. Resnick, C. Asterhan, & S. Clarke (Eds.), *Socializing intelligence through academic talk and dialogue* (pp. 303-320). Washington, DC: American Educational Research Association.

[7] Howe, C., Hennessy, S., Mercer, N., Vrikki, M., & Wheatley, L. (2019). Teacher-student dialogue during classroom teaching: Does it really impact on student outcomes? *Journal of the Learning Sciences, 28*, 462-512.

[8] Landis J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1),159–174.

[9] Lawrence, J. F., Crosson, A., Paré-Blagoev, E. J., & Snow, C. E. (2015). Word Generation randomized trial: Discussion mediates the impact of program treatment on academic word learning. *American Educational Research Journal, 52*, 750-786.

[10] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

[11] Mehan, H. (1979). *Learning lessons*. Cambridge, MA: Harvard University Press.

[12] Mercer, N., & Dawes, L. (2014). The study of talk between teachers and students, from the 1970s until the 2010s. *Oxford Review of Education, 40*, 430-335.

[13] Mercer, N., & Sams, C. (2006). Teaching children how to use language to solve maths problems. *Language and Education, 20*, 507-528.

[14] Mercer, N., Wegerif, R., & Dawes, L. (1999). Children's talk and the development of reasoning in the classroom. *British Educational Research Journal, 25*, 95-111.

[15] Michaels, S., O'Connor, C., & Resnick, L. B. (2008). Deliberative discourse idealized and realized: Accountable talk in the classroom and in civic life. *Studies in Philosophy and Education, 27*(4), 283-297.

[16] Michaels, S., O'Connor, C., Hall, M. W., & Resnick, L. B. (2010). Accountable talk® sourcebook. Pittsburgh, PA: University of Pittsburgh Institute for Learning.

[17] Michener, C. J., Proctor, C. P., & Silverman, R. D. (2018). Features of instructional talk predictive of reading comprehension. *Reading and Writing, 31,* 725-756.

[18] Murphy, P. K., Wilkinson, I. A. G., Soter, A. O., Hennessey, M. N., & Alexander, J. F. (2009). Examining the effects of classroom discussion on students' comprehension of text: A meta-analysis. *Journal of Educational Psychology, 101,* 740-764.

[19] National Council of Teachers of Mathematics (NCTM). (2014). Principles to action: Ensuring mathematical success for all. Reston, VA: National Council of Teachers of Mathematics.

[20] National Governors Association Center for Best Practices (NGA Center) & Council of Chief State School Officers (CCSSO). (2010). The common core state standards. Washington, DC: NGA Center, CCSSO.

[21] Puntambekar, S. (2022). Distributed scaffolding: Scaffolding students in classroom environments. *Educational Psychology Review, 34*, 451-472.

[22] Suresh, A., Sumner, T., Huang, I., Jacobs, J., Foland, B., & Ward, W. (2018). Using deep learning to automatically detect talk moves in teachers' mathematics lessons. In 2018 *IEEE International Conference on Big Data* (pp. 5445-5447). IEEE.

[23] Suresh, A., Sumner, T., Jacobs, J., Foland, B., & Ward, W. (2019, July). Automating analysis and feedback to improve mathematics teachers' classroom discourse. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 33, No. 01, pp. 9721-9728).

[24] van de Pol, J., Volman, M., & Beishuizen, J. (2010). Scaffolding in teacher-student interaction: A decade of research. *Educational Psychology Review, 22,* 271-296.

[25] Vygotsky, L. S. (1978). In M. Cole, V. John-Steiner, S. Scribner, & E. Souberman (Eds.), *Mind in society – The development of higher psychological processes.* Cambridge, MA: Harvard University Press.

[26] Wissinger, D. R., & De La Paz, S. (2015). Effects of critical discussions on middle school students' written historical arguments. *Journal of Educational Psychology, 108*, 43-59.

Columns on Last Page Should Be Made as Close As Possible to Equal Length