Toward Automated Feedback on Teacher Discourse to Enhance Teacher Learning

Emily Jensen[1], Meghan Dale[2], Patrick J. Donnelly[3], Cathlyn Stone[1], Sean Kelly[2], Amanda Godley[2], Sidney K. D'Mello[1]

[1] University of Colorado Boulder
Boulder, CO, USA
emily.jensen@colorado.edu

[2] University of Pittsburgh
Pittsburgh, PA, USA

[3] Oregon State University
Bend, OR, USA

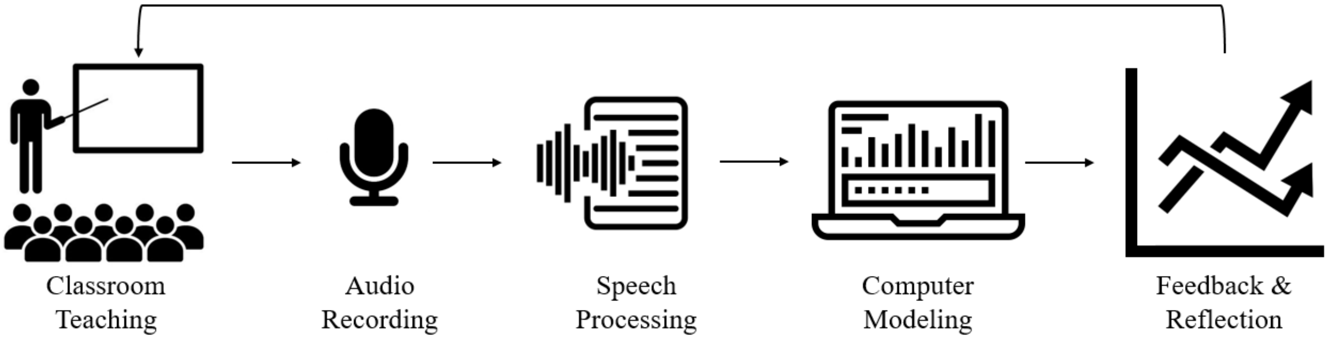

Figure 1. Overview of automated teacher feedback approach

ABSTRACT

Like anyone, teachers need feedback to improve. Due to the high cost of human classroom observation, teachers receive infrequent feedback which is often more focused on evaluating performance than on improving practice. To address this critical barrier to teacher learning, we aim to provide teachers with detailed and actionable automated feedback. Towards this end, we developed an approach that enables teachers to easily record high-quality audio from their classes. Using this approach, teachers recorded 142 classroom sessions, of which 127 (89%) were usable. Next, we used speech recognition and machine learning to develop teacher-generalizable computer-scored estimates of key dimensions of teacher discourse. We found that automated models were moderately accurate when compared to human coders and that speech recognition errors did not influence performance. We conclude that authentic teacher discourse can be recorded and analyzed for automatic feedback. Our next step is to incorporate the automatic models into an interactive visualization tool that will provide teachers with objective feedback on the quality of their discourse.

Author Keywords

automatic speech recognition; audio recording; classroom discourse; dialogic instruction; natural language processing

CHI '20, April 25–30, 2020, Honolulu, HI, USA
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-6708-0/20/04...$15.00.
DOI: https://doi.org/10.1145/3313831.3376418

CSS Concepts

- Human-centered computing~Usability testing
- Computing methodologies~Machine learning

INTRODUCTION

A good teacher must first be a good learner. Designing lesson plans, pursuing lesson goals, and analyzing post-lesson success and growth requires reflexivity and learning. Due to the contextualized nature of teaching, the cycle of planning, enacting, reflecting, and adjusting are at the core of good teaching. Beyond this type of daily practice which is important for proficiency, achieving expertise requires deliberate practice, usually under the guidance of a coach, with carefully designed training tasks at an appropriate level of difficulty such that progress should be achievable in the short-term with effort, feedback, and guidance [19, 21, 22].

Unfortunately, current teacher learning opportunities are a far cry from this type of idealized practice. Few teachers receive intensive sustained professional development and many schools report needing assistance to improve the professional development they can offer [6]. A recent multi-district analysis [70] of professional development effectiveness yielded sobering results. Despite spending about 10% of out-of-class time in a typical work year on professional development at a cost of $18,000 per teacher, a mere 30% teachers showed any performance improvement over 2-3 years; performance of the vast majority either stayed the same (50%) or even declined (20%). This is likely because the most common forms of professional development (including stand-alone conferences and workshops) have been criticized for being intellectually superficial [5]. These findings confirm the results of two federally-funded studies showing that when decontextualized from the classroom, traditional professional development has failed to make an impact on teacher performance or student outcomes [8, 27, 28].

Despite this discouraging news, the [70] report offers some promising recommendations. In lieu of simply "tinkering with the types and amount of professional development teachers receive," it argues for a dramatic change in the nature of teacher professional development. Realizing that teacher learning is a highly contextualized process, their chief recommendation is to "give teachers a clear, deep understanding of their own performance and progress" (p. 3). Simply put, teachers need feedback to improve.

Unfortunately, due to the high cost and complexity of human classroom observation [2], teachers receive infrequent feedback. The little feedback they do receive is often provided by a superior and the focus is on evaluating performance rather than improving practice [24, 65]. Given the pivotal role of feedback to learning [4, 14, 22, 63], the lack of immediate and objective feedback is a critical barrier that needs to be overcome if we are truly going to innovate teacher learning. Accordingly, we develop an approach to provide timely, accurate, and objective feedback to teachers to help them improve their practice.

In our approach (summarized in Figure 1), teachers begin with a typical classroom session (Classroom Teaching). During the lesson, teachers record high-quality audio of their own classroom talk (Audio Recording). This audio is then uploaded to the cloud, where an automated process transcribes it and extracts speech and language information from the transcripts (Speech Processing). Using these extracted features, the automated system then identifies the presence of key discourse variables based on pre-trained machine learning models (Computer Modeling). Finally, the results of the automated analysis are presented to teachers along with their long-term trends. Teachers will be able to use this feedback to adjust their classroom discourse and monitor their progress over time (Feedback & Reflection).

A key aspect of our approach is the emphasis on high-quality audio recordings of teacher discourse. Focusing on teacher feedback allows us to provide precise and robust feedback to support effective classroom instruction [39, 47]. We focus on using teacher audio rather than video or student audio for several reasons. First, there are severe privacy concerns with using video recordings, whereas recording of audio is exempted as it constitutes observations of normal practice. We also only record teachers as it is infeasible to scale up audio recording to individually mic each student in a classroom; recording students with a single omnidirectional microphone produced noisy audio [15]. However, since our intended focus is on teacher discourse, teacher audio should be sufficient for our purposes as we elaborate below.

Related Work

We review work on recording classroom audio, modeling classroom discourse, and automating teacher feedback.

Classroom Discourse Recording

There is a long tradition of recording live class sessions in order to study instructional practices [1, 11, 29]. In a particularly ambitious study [64], researchers recorded 8th grade mathematics classes in Germany, Japan, and the United States in order to analyze the differences in teaching strategies between cultures. However, these recording methods were not designed to produce data of high enough fidelity to accommodate automatic analyses.

Recently, researchers such as Wang and colleagues [72] have pioneered approaches to classroom recording that enable some forms of automated analysis. In this study, the LENA [25] wearable recording system was used to record classroom audio in 1st and 3rd grade classrooms. Researchers were able to distinguish between teacher and student speech, from which they could infer turn-taking dynamics. However, they did not analyze any utterance content. Further, the high cost ($10,000) of this recording approach prohibits usage at scale.

More recent work has used classroom video recordings with the specific intent of developing automated analysis systems. In [52], Ramakrishnan et al. used classroom video and audio with deep learning techniques to automatically model classroom climate. Although they were able to differentiate positive and negative climate using both modalities, the logistic and privacy concerns about using video raise challenges to widespread usage.

Turning back to audio, D'Mello et al. [15], proposed a comprehensive set of design constraints of a teacher recording system aimed for automatic analysis. In particular: (1) the system must be easy to use by teachers and other non-experts, (2) the cost must be affordable for schools, (3) the system must be easy to set up in a few minutes, (4) recording should not interfere with classroom activities, and (5) the recorded audio must be of high enough fidelity to allow automatic speech recognition (ASR). In [15], they tested several designs in middle school English Language Arts (ELA) classrooms and found moderate ASR performance (word accuracy of 66%, simple word overlap of 69%) using a simple teacher headset microphone.

Discourse Modeling

Previous work has used audio recordings to model classroom discourse at several different levels. The coarsest level is activity classification, which attempts to classify classroom recordings according to broad categories such as discussion or individual seatwork. Using the Language Environment Analysis (LENA) [25] system, Wang et al. [72] used an analysis of turn taking dynamics to identify general classroom activities from audio. Similarly, Donnelly et al. [17] segmented audio into 60-second windows and labeled each with the dominant classroom activity. They then trained models to identify general classroom activities using utterance timing, language, and acoustic features. Although these methods provide valuable information on the structure of class time, they do not provide fine-grained feedback on the content and quality of teacher discourse.

Recent work focuses on analysis of classroom discourse at the individual utterance level. For example, Donnelly et al. [18] expands the work in [7] to specifically identify teacher questions. In this work, they transcribed classroom speech using ASR and predicted teacher questions using acoustic, linguistic, and context features. Stone et al. [66] used similar features as well as specific words and phrases (n-grams) to improve utterance-level modeling of several discourse variables such as content-specific questions and instructional statements. Suresh and colleagues have also successfully used deep learning methods to detect specific dialogic strategies in middle school mathematics classrooms [67, 68]; however, this work relied on human-transcribed utterances.

Other work has focused on analysis of classroom discourse at the class session level rather than individual utterances. This approach has been promising when predicting the prevalence of infrequent discourse strategies such as open-ended questions [38]. For example, Olney et al. [48] used word, part of speech, syntactic, and other discourse structure to directly predict the proportion of open-ended questions for a class session. They also showed that this model outperformed models that aggregated predictions at the utterance level. Building off of the work in [48], Cook et al. [13] found that models trained on words and phrases performed similarly to those trained on predefined part of speech and discourse structure [12] features for predicting open-ended questions. Importantly, combining the predictions from these models yielded improved performance over the individual models.

Automated Teacher Feedback

Using automated models to provide teacher feedback is an area still in the beginning stages of research. Dashboards are a popular method of giving student feedback; however, there is little work analyzing potential benefits to teachers and how this may improve student learning [36, 42, 43, 55, 73].

There are a few notable examples of automated teacher feedback systems. Holstein and colleagues have developed real-time systems that can inform and guide teachers during live class sessions. In [36], they introduce Lumilo, which pairs smart glasses with an Intelligent Tutoring System; this system alerts teachers when students need help that the tutoring system cannot provide. Additionally, Poskin et al. developed a smartphone application TeachFX (teachfx.com) which models the proportion of teacher talk using classroom audio recordings. Finally, Aslan et al. developed a real-time system to alert teachers of student disengagement [3]. However, none of these systems provide teachers with automatic discourse feedback as in the present work.

Research Questions

We aim to automatically provide feedback for teachers on the quality of their discourse from audio alone. We address three specific research questions.

RQ1: To what extent can teachers easily record high-quality audio of their own classes to enable automatic feedback? An important goal of our work is to allow teachers to record their classroom talk independently and with few technological issues. To do this, we: (1) used high-quality but relatively inexpensive recording equipment, (2) created user-support materials, (3) conducted one in-person training with each teacher. We asked 16 teachers to record their classroom audio over a two-month period. We analyzed the quality of the recorded audio data and collected usability feedback.

RQ2: To what extent can we use the recorded audio to automate the analysis of teacher discourse? We used real data collected from RQ1 and machine learning methods to train models that predict seven teacher discourse variables. We show that our automated analysis is able to model the session-level prevalence of these discourse variables with moderate accuracy compared to human annotations, which is the current gold-standard.

RQ3: How robust is our approach to differences in speech recognition quality? Real-world classrooms are noisy and teachers are not experts in audio recording. Thus, perfect speech recognition is currently implausible, nor is it the goal. There is then the question of how speech recognition quality affects our results. To address this, we analyze the quality of ASR transcriptions using word error rate (WER) and simple word overlap (SWO) metrics and investigate the relationship of these metrics to the accuracy of our modeling approach (RQ2). We show the quality of audio transcription is not strongly related to the accuracy of our automated models.

Novelty and Contribution

This research is novel in three areas. First, we developed a system for teachers to easily record their own audio data. This is an improvement over previous studies [1, 11, 29, 64, 72] in that teachers no longer require external researchers to assist with set-up or recording. This approach also allows teachers to obtain frequent feedback since they are able to record any class session at any time. Data collected from a user study shows the self-recorded teacher audio data is of good quality despite using commercial off-the-shelf devices.

Our work is also novel in that we are able to automatically model the presence of specific key discourse variables using teacher-recorded audio data. This extends previous work [13, 18, 38, 48, 66], which provided automated time usage summaries (e.g., percent of time on discussions vs. lecture) or individual discourse variables (percent of questions) on researcher-collected audio. We extend this work by successfully modeling the presence of seven discourse variables from teacher-recorded audio. Our results highlight the potential to provide teachers with frequent and actionable feedback on discourse variables related to student learning.

Lastly, it is known that classrooms are noisy environments, rendering fully automatic speech and language processing a major challenge. Here, we show that with carefully designed recording methods and computational modeling, it is possible to obtain important insights into teacher audio. Further, we demonstrate that our models are largely robust to errors with automatic speech transcription. In sum, our results point to the feasibility of fully automated teacher feedback mechanisms as outlined in Figure 1.

TEACHER TALK RECORDING SYSTEM

Goals and Rationale

This section addresses the Audio Recording segment of Figure 1. Our design builds off several design requirements, such as the ones discussed in [15]. Teachers should be able to use this recording system regularly in a natural classroom setting. To do this, we created a streamlined teacher talk recording system using microphones and laptop computers with accompanying user-support materials. Through surveys, we evaluated the usability of the teacher talk recording system after minimal training. In addition to usability, the recording system aided in the collection of a large data set of high-fidelity classroom recordings for automatic analysis.

Recording Setup and Teacher Training

Teachers were provided with a Samson 77 Airline microphone system (AH1 Headset Transmitter with CR77 Wireless Receiver; retails for $230). The only other equipment teachers used was a laptop with the necessary recording programs installed (see Figure 2 for equipment). We selected this headset based on the analysis in [15] which reported good noise cancellation properties for classroom background noise. The microphone is a unidirectional microphone with a cardioid pickup pattern which results in it being more accurate than lapel or earsets available at the time (based on pilot testing with multiple microphones).

To begin set up, the receiver and laptop were both connected to a power source. Teachers then logged in to the laptop, verified it was connected to the internet, and connected the receiver to the laptop. Next, teachers turned on the wireless headset and ensured it was positioned correctly.

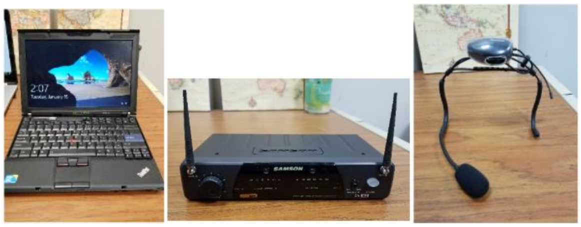

Figure 2. Recording equipment (from left to right): laptop, audio receiver, and wireless headset

Once the hardware was set up, teachers created a new project in the recording program Audacity. This then allows teachers to record test audio to check the volume level of the microphone. Teachers then recorded their normal classroom session while wearing the headset. After the recording was completed, teachers saved the Audacity recording project, which was automatically backed up to a Dropbox folder. Teachers were instructed how to do these steps in an in-person training session. During this training, members of the research team walked teachers through each step of the recording process and answered any questions. Teachers could later refer back to these steps in user-support materials developed by the research team. These training materials include images and step-by-step directions to set up the equipment as well as troubleshooting instructions. In addition, the training materials include several examples of how to verify that the microphone is set at an appropriate volume level, an important factor for automatic analyses.

Data Collection with the Recording Setup

To test the recording system, we recruited 16 secondary ELA teachers in the spring of 2018. Teachers were recruited from three suburban school districts in Western Pennsylvania—including two high schools and one middle school. Our sample included 11 female and 5 male teachers. Teachers in our sample averaged 14.6 years of experience, though our sample also included three early career teachers (i.e., teachers in their first five years of teaching). All teachers in our sample: (1) were ELA teachers, (2) identified as white, (3) held a master's degree or higher, (4) received their state certification, and (5) taught grades 7-12. We note that the lack of demographic diversity in our sample was due to the fact that data was collected from suburban districts with a heavily white population. Teachers were compensated for their participation with a $600 pre-paid card.

After initial training, teachers were expected to be able to independently record their own classroom talk. Each teacher was asked to record at least four sessions of two classes (eight recordings in all) that they identified as differing in some way (e.g., by grade level, by academic level, by class size, or by pace of the class). All teacher talk recordings were stored using anonymized identifiers for teachers and schools.

After each classroom recording, we asked teachers to complete a post-observation survey using Qualtrics. The Qualtrics link was located on the desktop of the recording laptop and asked teachers to include the name of their most recently saved audio recording. The survey included four questions relating to the usability of the teacher talk recording system as well as an open-ended response for any additional comments. All questions were required except for the open-ended comments question. The usability questions and response options can be found in Table 1.

Results of the Teacher Talk Recording System

Evaluating Audio Quality

We obtained the desired eight recordings per teacher for a total of 142 attempted recordings across 39 unique classes. Of those, 15 recordings were omitted from the sample due to technical difficulty (discussed below), leaving 127 (or 89% of total recordings) usable for ASR and further analysis. The average length of the teacher talk recordings was 43 minutes.

We evaluated audio quality by using an A, B, C, or F rating. An "A" indicated excellent recording quality, "B" indicated

Table 1. User survey questions and responses

Question	Possible Responses	Average Response (SD)
At the start of the observation, did you check the audio levels and adjust the recording volume in Audacity as needed?	1 = Yes 0 = No	0.83 (0.327)
Thinking about the *ease of set-up*, how difficult or easy was it to get your recording equipment set-up and running?	1 = very easy 2 = somewhat easy 3 = somewhat difficult 4 = very difficult	1.54 (0.510)
Thinking about *wearing the headset*, how comfortable or uncomfortable was the equipment during the class session?	1 = basically comfortable 2 = somewhat uncomfortable 3 = very uncomfortable	1.60 (0.451)
Thinking about being recorded during class, did conducting a recording make you *more focused* on classroom communication?	1 = I was focused the usual amount on classroom communication 2 = Somewhat more focused on classroom communication 3 = I was much more focused on classroom communication	1.45 (0.390)

acceptable quality with minimal volume or background noise problems, "C" indicated recordings containing flawed segments, and an "F" indicated audio files that were lost or had irreparable technical error. The majority (65%) of total recordings were rated as "A" quality. We then combined ratings of "B" and "C" (additional 24%) to form a general category indicating acceptable audio quality (see Figure 3).

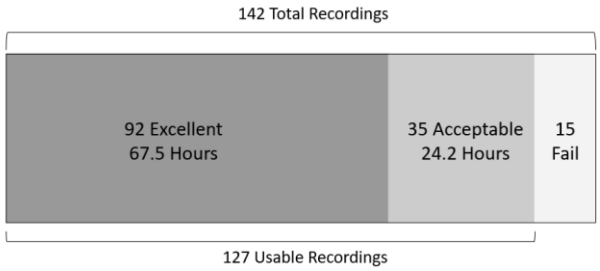

Figure 3. Breakdown of recorded audio quality

Results of User Questionnaire

Teachers completed surveys for 114 of these 127 usable class sessions for an average teacher response rate of 91%.

Most teachers (12 of 16) reported checking the audio quality for all of their recording sessions. On average, teachers report checking the audio quality 83% of the time. The comments show that teachers were competent troubleshooters; they knew what good audio levels should look like in the recording software and made appropriate volume adjustments to produce good audio quality.

Overall, teachers had very little difficulty with the recording process. When considering ease of set-up, teachers reported an average rating of 1.54. In comments, teachers reported feeling "*a bit uncertain with the equipment*" for early recordings but expected "*to get better with the process for future recordings*." Some teachers also reported making several recordings because of recording difficulty at the beginning of class. The most common reported problem was forgetting to turn on the microphone.

Teachers reported minimal discomfort with the headset. When reporting comfort, teachers reported an average rating of 1.60. Of the 114 completed surveys, only 3 sessions were reported as "*very uncomfortable*". In comments, one teacher reported feeling the microphone might be "*intrusive*" or make the lesson feel "*staged*". In addition, one teacher commented, "*students were very curious about the headset.*"

This survey also provides preliminary evidence that the recording process encourages teachers to focus more on their classroom instruction. The average teacher response was 1.45, which is between the "usual amount" of focus and "somewhat more" focus. More focus is a desirable outcome for the feedback approach as it aims to elicit self-reflection.

Discussion of Recording Errors

Most recording errors were caused by data saving problems. Of the recording sessions labeled as "F", five sessions were missing an Audacity project file. Four other sessions did not sync data files to the corresponding Dropbox folder. Two class sessions contained duplicate incomplete project files. These problems may have been caused by internet connectivity issues or failure of teachers to follow the set-up instructions. An additional three sessions were rated as "F" due to extremely short recording time or recordings largely consisting of silence. These errors were likely caused by improper set-up or calibration of the wireless headset. Overall, we conclude that the present approach is a viable method to obtain good quality teacher-recorded audio.

AUTOMATIC DISCOURSE MODELING

Goals and Rationale

This section aims to answer RQ2, which addresses whether the teacher-recorded audio data can be used to automatically model features of teacher discourse aimed at providing automated feedback. To measure the accuracy of our models, we compared our predictions to human-coded labels, which are the current gold-standard.

We framed this problem as a regression problem (i.e., predicting the proportion of utterances that contain a discourse variable) because this is the level of granularity at which we will provide feedback to teachers. This level of feedback is standard in the teacher professional development literature [46] and more intuitive for teachers to understand (e.g., focus on increasing questions from 15% to 20%).

Selection of Discourse Variables

Our chosen teacher discourse variables draw from a set of practices that are well-established in the literature on teaching effectiveness, student engagement, and achievement that are exhibited through classroom discourse. In observational studies of ELA classroom discourse, much research has focused on interactive forms of talk encompassing genuine discussions, deliberations, etc., which are often collectively termed *dialogic*. A key feature of dialogic teacher discourse is that it takes students' ideas seriously [26] and elicits both increased cognition and engagement [39, 46, 54]. Thus, the theoretical framework employed in the present study builds very closely on this program of research on dialogically organized instruction, emphasizing *questions*, *authentic questions*, and *uptake* amongst other dialogic discourse variables (see also [37, 44]).

In order to move beyond the study of interactive discourse occurring in question and answer sessions—a guiding empirical component of prior work—our theoretical framework incorporates insights from research by Shernoff and colleagues [61, 62] and Grossman et al. [31]. We study several new discourse variables including *goal specificity* [45, 61], *use of ELA terms* [20, 33], *cognitive level* [9, 30, 53, 69], and *elaborated feedback*. In addition, we include a variable to distinguish *instructional talk* from non-instructional talk.

The selected set of variables has been shown to challenge and support students [60] as well as promote discussion and engagement among students. The final set of variables and proportional occurrence in our data are listed in Table 2. It is important to note that these categories are not mutually exclusive; for instance, each of the other discourse variables are considered instances of instructional talk and authentic questions are a subset of the question category.

Transcribing Audio & Coding Discourse Variables

Once a teacher has recorded an audio file of their class session (see Teacher Talk Recording System section), the next step is to process and transcribe the audio and code the utterances for the discourse variables. We used an ASR for this since we intended to develop an automated approach.

Audio Segmentation and Transcription

Each audio file was transcribed using the IBM Watson speech recognizer [58], which segments and transcribes the audio into utterances. For each utterance, the transcription includes the start and end time of the utterance, the transcript of spoken words, and the confidence of the speech recognizer. We found that the speech recognizer sometimes erroneously splits an utterance into two. To address this, we merged consecutive utterances if there was less than one second pause between transcribed utterances; this threshold was selected after extensively testing several others. There were a total of 35,142 utterances after merging.

Utterance Coding

Resource limitations precluded coding the entire set of 35,142 utterances. Therefore, we sampled 200 randomly selected, sequential teacher utterances per transcript so that our "gold standard" human coding dataset could include as many teachers as possible. Overall, we coded 16,977 teacher utterances (48% of total utterances).

The coding system included a Microsoft Excel template with pre-programmed macros for skip patterns and legitimate values. Coding was done by raters trained and supervised by ELA content experts. Coders reached a reliability threshold of 80% (using Gwet's AC metric [32]) across all discourse variables prior to independent coding. The coders evaluated the Excel template interface favorably because it allowed them to easily reference surrounding utterances for context. On average, each 200-row file took approximately 96 minutes to code.

Automatic Modeling of Teacher Discourse

We adopted a supervised learning approach for modeling teacher discourse. The goal was to automatically estimate the proportional occurrences of each discourse variable as indicated in Table 2 in a manner that is generalizable to new teachers with similar characteristics as our sample.

Feature Generation. We used four types of features. First, we extracted acoustic features using OpenSmile [23], using a standard feature set from the 2009 Interspeech Emotion challenge [59]. We used statistical functionals of 16 low-level features: zero-crossing-rate from the time signal, root mean square frame energy, Mel-frequency cepstral coefficients 1-12, fundamental frequency computed from the cepstrum (normalized to 500 Hz) and voicing probability computed from the autocorrelation of the power spectrum.

Second, we computed 13 context features that provide insights into turn-taking dynamics. These include utterance length, normalized position in the session, speech rate, length of surrounding pauses, and probability an utterance occurred in one of the instructional segments discussed in [46].

Third, we used 37 binary linguistic features. Of these features, 34 come from part of speech [10] and question type

Table 2. Description of key teacher discourse variables

Discourse Variable	**Definition**	**Prevalence**	**Positive Example**
Instructional Talk	Focuses on the lesson and learning goals rather than on other topics, such as classroom management or procedural talk.	81%	Let's think about the tone of this poem.
Questions	Requests for information.	31%	Do you have a pencil?
Authentic Questions	Open-ended question for which the teacher does not have a pre-scripted answer.	5%	What was your reaction to the end of the story?
Elaborated Evaluation	Expression of judgment or correctness of a student's utterance with explicit guidance for student learning and thinking.	6%	That's right. You're dying with each breath, and this is what the poet tries to bring to the consciousness of the beloved.
High Cognitive Level	Emphasizes analysis (e.g., compare, interpret, synthesize, etc.) rather than reports or recitation of facts (e.g., define, recall, identify)	4%	How were their reactions to the accident different?
Uptake	Incorporation of ideas from a student utterance into a subsequent statement or question.	2%	You think he can't get help, can you expand on that?
Goal Specificity	Extent to which the teacher explains the process and end goals of a particular activity.	9%	Your writing partner should give you three overall comments, before editing supporting details.
ELA Terms	The use of disciplinary terms in teacher talk.	9%	Ensure that you include a topic sentence in each one of your paragraphs.

[49] taggers and include the presence of question words, parts of speech, and categories such as definition or comparison. These have been previously linked to some of our discourse variables [56, 57]. We included three other features: proper nouns (e.g., student names), pronouns associated with uptake, and pronouns not associated with uptake as recommended by a domain expert.

Finally, we used a bag of n-grams approach to represent common words and phrases derived from automatically transcribed utterances and filtered so only n-grams that frequently occurred in the corpus were included. Some n-grams correlated with discourse variables include *does, did, think, say, make*.

Model Building & Validation. We modeled the occurrence of each discourse variable using three approaches. The first two approaches used supervised learning models from the scikit-learn library [50]. The first approach models the data at the utterance level (i.e., identifying the presence or absence of each discourse variable per utterance) and then averages the per-utterance predictions over the class session. One can think of this as a *local context model*. The second approach is a *global session model* that merges all the utterances and directly predicts the proportional occurrence of each discourse variable. The third model simply averages the predictions of the local and global models.

For all models, we implemented nested 5-fold cross-validation at the teacher level, which means that all utterances from the same teacher are kept within the same train/validation/test fold. This approach ensures the generalizability of our models to new teachers since it maintains independence between the testing and training sets. An overview of the model-building process can be found in Figure 4. We did not include *uptake* in our final analyses due to its infrequency in the data.

For the utterance-level model, we trained a Random Forest classifier with 100 trees to predict the presence of a discourse variable in each utterance using the acoustic, linguistic, context, and n-gram features described above. Since the prevalence of the discourse variables are imbalanced (Table 2), we used the imblearn library [40] to oversample minority class utterances during training; the class distributions in the testing sets were unchanged. In cross-validation, we also experimented with removing stopwords and using n-grams with a minimum frequency. Averaged across discourse variables, this model achieved an AUROC of 0.77 (compared to a 0.5 baseline) and an average accuracy of 0.71. After predicting the presence or absence of the discourse variable for each utterance, we averaged these predictions at the class session level to get a proportional occurrence of utterances containing that discourse variable per class session. Finally, we normalized these predictions to match the range of values generated by human labelling.

For the second (global) model, we trained a Random Forest regressor with 100 trees to directly predict the proportion of a discourse variable in each class session. Since we are not predicting on individual utterances, we only used n-grams with a minimum frequency and did not use oversampling. These models predicted occurrence of each discourse variable per class session, normalized as above.

For the combined model, we averaged the normalized predictions of the above two models per class session and then re-normalized them again.

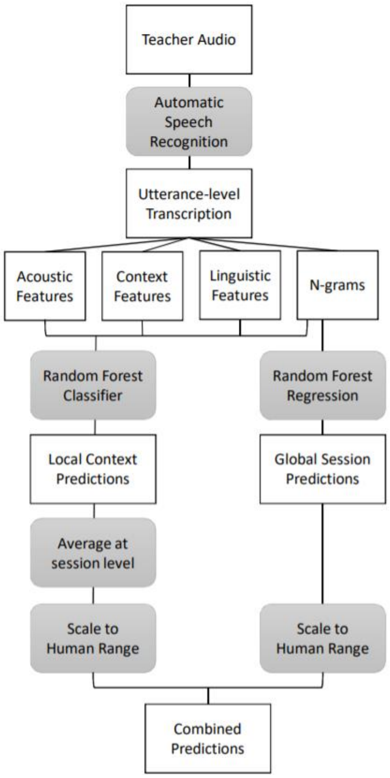

Figure 4. Audio processing and model-building pipeline

Results

We evaluated the performance of these three models using Spearman rank-order correlation (since the distributions were non-normal and continuous) and Mean Absolute Error (MAE), calculated as the absolute error between predicted proportion and actual proportion. We report the results for the best-performing model of each discourse variable. Correlations ranged from 0.305 to 0.565 with an average of 0.437 (Table 3). MAE scores ranged from 0.052 to 0.127 with an average of 0.086. Some model predictions more closely matched human-labeled means (i.e., 0.306 vs. 0.270 for questions) than others (0.135 vs. 0.035 for high cognitive level). The best performing models were for questions and authentic questions. This is unsurprising since questions have unique prosodic characteristics (e.g., raised pitch at the end of an utterance) that can be captured in our audio processing. The worst performing model was for high cognitive level. This is also unsurprising because this variable has fewer obvious identifying characteristics and can be subjective to code.

We also calculated the MAE per teacher, which ranged from 0.064 to 0.100, indicating mostly equitable performance across teachers (correlations could not be computed since there are a maximum of eight sessions per teacher).

In order to analyze the potential effectiveness of our models for providing feedback, we compared the distribution of predictions compared to human-labeled data. Some model predictions closely matched the shape and spread of the human-labeled data; however, our computer models struggled to capture zero- or one-inflated distributions (see Figure 5 for representative examples). Additionally, our computer predictions tended to have wider spread around the center of the distribution. This indicates that computer-generated feedback may be more variable; in the future it may be necessary to simplify automated feedback (e.g., to high, medium, or low instead of a specific proportion) to provide consistent feedback to teachers.

Table 3. Best model results for each discourse variable

	Distribution Mean			
Discourse Variable	**Human**	**Computer**	**Spearman *r***	**MAE**
Instructional Talk	0.809	0.716	0.349	0.127
Questions	0.306	0.270	0.564	0.088
Authentic Questions	0.051	0.094	0.565	0.061
Elaborated Evaluation	0.064	0.099	0.351	0.052
High Cognitive Level	0.035	0.135	0.305	0.108
ELA Terms	0.089	0.168	0.469	0.104
Goal Specificity	0.087	0.199	0.456	0.063

Analysis and Modeling of Speech Recognition Quality

Goals and Rationale

This section addresses RQ3 regarding the robustness of our models to transcription quality differences. Automatic speech recognizers are often optimized for certain types of speech, such as meetings, phone conversations, or video captioning. Classroom discourse is different than other types of speech in that it is inherently noisy, contains multiple speakers, features frequent occurrences of cross-talk, and uses domain-specific vocabulary. Given the challenges of transcribing live classroom discourse, we aimed to quantify the relationship between the transcription quality and the accuracy of our automatic modeling of discourse.

Methods

To evaluate the quality of the transcriptions, we manually transcribed a random sample of 20 utterances per usable class session and compared these transcriptions to those generated by the IBM Watson ASR. This yielded 2,540 utterances with an average duration of 7.9 seconds, length of 22.4 words, and average ASR confidence score of 0.88. We use the standard metric of word error rate (WER) for speech to text accuracy, computed as the percent of insertion, deletion, and substitution errors incurred by the ASR. Since WER is sensitive to word order, we also compute simple word overlap (SWO) as the proportion of words that appear in both transcripts. We obtained a WER of 0.28 and an average SWO of 0.72 on our data.

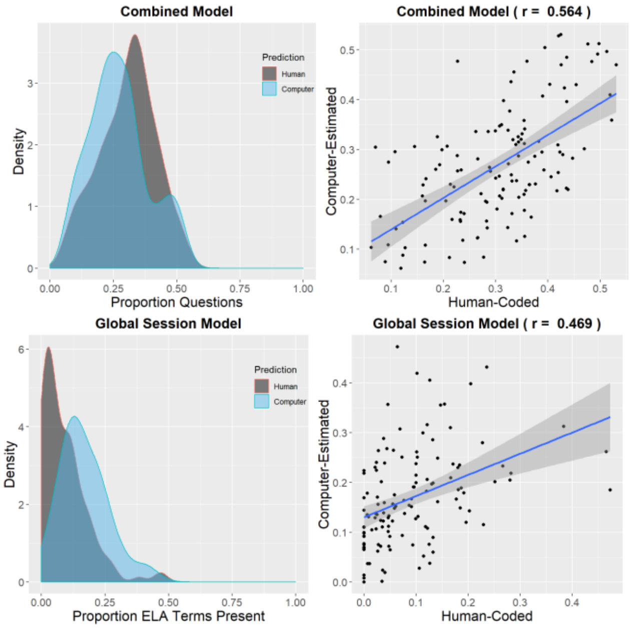

Figure 5. Distribution of human and computer-predicted labels for Questions (top) and ELA Terms (bottom)

Results

To evaluate if the quality of speech transcription had an effect on the accuracy of the automatic discourse modeling, we compared the MAE from our automated models with measures of ASR quality (Table 4). Correlations were generally small, indicating that the models were quite robust to ASR errors. These low correlations also indicate that our errors in modeling discourse variables are not primarily driven by the inherent noisiness of automatic transcriptions.

DISCUSSION

We addressed the problem of automatically providing feedback to teachers to improve their classroom practice. Our goal was to make it easy for teachers to record their classroom audio so they can get quick, automated feedback on important features of their classroom discourse. This should enable them to reflect on each class period and incorporate insights into their next class period. The hope is that this iterative cycle will increase the frequency that teachers reflect on their practice compared to traditional in-class evaluations which occur a few times per year, if at all.

We addressed three main steps towards this goal. First, we developed an audio recording system which allows teachers to record data from their classrooms independently and with commercial-off-the-shelf equipment (RQ1). Second, we used the audio collected from teachers to train automated computer models that can accurately and consistently provide feedback in a manner that generalizes to new teachers (RQ2). Lastly, we investigated the relationship between ASR transcription quality and automated feedback quality (RQ3). We discuss our main findings with respect to a set of usability themes.

Table 4. Session-level Spearman correlation between MAE and WER/SWO

Discourse Variable	WER	SWO
Instructional Talk	0.19	-0.06
Questions	-0.02	-0.01
Authentic Questions	-0.12	0.02
Elaborated Evaluation	-0.03	0.01
High Cognitive Level	-0.01	-0.12
ELA Terms	-0.05	-0.03
Goal Specificity	-0.10	0.15

Usability Themes (Main Findings)

Ease of Use. The ability to easily record high-quality audio lies at the heart of the automated effort. Compared to humans, computers still require relatively high-fidelity audio. Thus, the low-to-medium quality audio traditionally obtained in classroom recordings would not suffice for automated analyses; this was empirically confirmed in [15]. Using our system, teachers were able to successfully record audio from their own classrooms without external help beyond a short training session. Results of usability surveys also indicated that teachers found the system easy to use, generally comfortable, and suggest that it might increase focus on classroom communication.

Accuracy. We also found that teacher–recorded audio was sufficient for automated analyses of several discourse variables. In particular, correlations between our automated estimates and gold-standard human coding ranged from 0.305 to 0.565 with an average of 0.437. These results are within the range of what is obtained from human observations of classroom practice. Consider the Measures of Effective Teaching (MET) project, which collected data from more than 1,500 teachers in six urban districts. This project trained human coders to score classroom videos using two state-of-the-art protocols including Framework For Teaching (FFT [16]) and Protocol for Language Arts Teaching Observation (PLATO [31]). Correlations between the two human judges ranged from a low of (0.21 FFT, 0.33 PLATO) to a maximum of only (0.34 FFT, 0.56 PLATO) [34]. Our models perform well within the higher end of this range. Further, these existing protocols categorize instruction discretely (e.g., rating as high, medium, or low) rather than a proportion. We similarly discretized our data (middle 80% is rated medium) and found that rates of exact agreement range from 60-80%, compared to the 70-76% agreement in the MET study. This again suggests that our results are comparable to the state of the art human coding.

Robustness. The use of automated models likely provides more consistency over traditional in-class observations, where the observer may change between sessions or have different subjective interpretations over time. We found that despite moderate ASR transcription errors, the transcripts were of good enough quality to train useful discourse models. Importantly, our models were quite robust to ASR errors as this was negligibly correlated with model accuracy.

Potential Risks for Teachers

One potential concern is that our proposed system might be used to evaluate teachers or enforce adherence to particular teaching practices. The features of teacher talk that our project centers on target teaching quality. However, the application that we are developing does not provide a rating system or measures for teaching quality that could be used to assess or enforce specific practices. We intend for teachers to use this system as an opportunity for self-initiated reflection to improve their future practice.

Limitations & Improvement Plans

We discovered several areas for improvement of our approach. Regarding audio recording, a researcher had to be present to initially show teachers how to use the equipment and software, which might limit scalability to large school districts. The process of automatically transferring data from the laptop to the cloud-based servers was sometimes error prone when teachers shut off the computer and when bandwidth was low. We lost 11% of the recordings due to these and other factors (incorrect setting of audio levels).

We aim to address this limitation by streamlining the recording process. We will develop an integrated application that will allow teachers to directly record audio from the device microphone. At the beginning of each recording session, the app will allow the teacher to enter meta-data for the lesson and connect and adjust the microphone levels. After this 1-minute setup, the teacher can begin teaching and the app will automatically adjust the audio levels as necessary. Teachers can end the recording with another button click. The app will upload audio to cloud-based servers and notify teachers and researchers of any problems.

Another area of improvement pertains to the accuracy of the automated models. Teacher discourse is inherently sequential in that the current utterance is strongly influenced by the previous set of utterances. Thus, one method of improvement is to consider deep sequence learning models, such as long short-term memory [35], that are able to learn long temporal dependencies in the data. We will also consider representational learning techniques by using word embeddings (e.g., Word2vec [41] and GloVe [51]), obtained by encoding words as vectors that capture the similarity with other words, thereby providing semantics. Finally, we will consider incorporating an attention mechanism in the networks to allow it to focus on specific parts of the input, which has led to performance gains in other domains [71].

The third limitation relates to our relatively homogeneous sample, since dialect and speaker variation are expected to degrade model accuracy. To address this, we aim to collect additional data using the streamlined audio recognition approach from a more diverse sample of teachers. That said, dialect and speaker differences are complex and pose a challenge to any automated language analysis technology. Hence, it might be that our approach will be initially limited for teachers beyond the dialects and demographic backgrounds represented in the data.

Future Work

We have demonstrated that teachers can record high-quality audio on their own and that we can automatically estimate the prevalence of several important features of classroom discourse. The next step is to incorporate the models in a visualization app to provide teachers with automated feedback. We are currently completing development on a smartphone application that will display automated feedback to teachers. This application will serve as the central component of our approach (Feedback & Reflection step of Figure 1) as it will collect teacher audio data, send it to the cloud for processing, and finally display feedback. Using this application, teachers can track their performance over time and compare different lessons and classes. Once implemented and user tested, we will be able to evaluate how feedback changes teacher behaviors over time and whether this improves student learning.

CONCLUSION

In this work, we develop an approach which generates feedback on key teacher discourse variables using teacher self-recorded audio data from real classrooms. Using a set of design requirements, we develop a relatively easy-to-use audio recording system that is affordable and employs minimal use of specialized equipment. We show that teachers would record high-quality audio data with this approach (89% success rate). We further show that the recorded audio is of good enough quality to automate the measurement of several discourse variables known to promote student engagement and learning. Finally, we show that these models are largely robust to ASR errors, suggesting that speech recognition should not be a limiting factor for these and similar systems. Our next step is to incorporate the automated feedback into an interactive visualization application so that teachers can received detail feedback as frequently as they choose, get improvement tips, and track progress towards goals that they set. Our hypothesis is that frequent, immediate, and automated feedback on core dimensions of effective discourse will enhance the quality of teacher reflection, leading to improvements in practice, ultimately increasing student engagement and achievement.

ACKNOWLEDGEMENTS

This research was supported by the National Science Foundation (NSF IIS 1735785). Any opinions, findings, conclusions, or recommendations expressed in this paper are those of the author and do not represent the views of the funding agencies.

REFERENCES

[1] Alibali, M.W. et al. 2014. How teachers link ideas

in mathematics instruction using speech and gesture: a corpus analysis. *Cognition and Instruction*. 32, 1 (2014), 65–100. DOI:https://doi.org/10.1080/07370008.2013.858161.

[2] Archer, J. et al. 2016. *Better Feedback for Better Teaching: A Practical Guide to Improving Classroom Observations*.

[3] Aslan, S. et al. 2019. Investigating the impact of a real-time, multimodal student engagement analytics technology in authentic classrooms. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19* (New York, New York, USA, 2019), 1–12. DOI:https://doi.org/10.1145/3290605.3300534.

[4] Azevedo, R. and Bernard, R.M. 1995. A meta-analysis of the effects of feedback in computer-based instruction. *Journal of Educational Computing Research*. 13, 2 (1995), 111–127. DOI:https://doi.org/10.2190/9lmd-3u28-3a0g-ftqt.

[5] Ball, D.L. and Cohen, D.K. 1999. Developing practice, developing practitioners: toward a practice-based theory of professional education. *Teaching as the learning profession: Handbook of policy and practice*. G. Sykes and L. Darling-Hammond, eds. Jossey-Bass Inc. 3–32.

[6] Birman, B.F. et al. 2007. State and local implementation of the no child left behind act. *Volume II - Teacher Quality Under NCLB: Interim Report*. U.S. Department of Education.

[7] Blanchard, N. et al. 2016. Automatic detection of teacher questions from audio in live classrooms. *Proceedings of the 9th International Conference on Educational Data Mining (EDM 2016)* (2016).

[8] Borko, H. 2004. Professional development and teacher learning: mapping the terrain. *Educational Researcher*. 33, 8 (2004), 3–15. DOI:https://doi.org/10.3102/0013189X033008003.

[9] Bransford, J.D. et al. eds. 2000. *How people learn: brain, mind, experience, and school*. National Academy Press.

[10] Brill, E. 1992. A simple rule-based part of speech tagger. *Proceedings of the Workshop on Speech and Natural Language* (Harriman, NY, 1992), 112–116. DOI:https://doi.org/10.3115/1075527.1075553.

[11] Cantrell, S. and Kane, T.J. 2013. *Ensuring Fair and Reliable Measures of Effective Teaching: Culminating Findings from the MET Project's Three-Year Study*.

[12] Church, K.W. and Hanks, P. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*. 16, 1 (1990), 76–83. DOI:https://doi.org/10.3115/981623.981633.

[13] Cook, C. et al. 2018. An open vocabulary approach for estimating teacher use of authentic questions in classroom discourse. *Proceedings of the 11th International Conference on Educational Data Mining* (2018).

[14] D'Mello, S.K. et al. 2010. Expert tutors' feedback is immediate, direct, and discriminating. *Proceedings of the 23rd International Florida Artificial Intelligence Research Society Conference, FLAIRS-23* (Menlo Park, CA, 2010), 504–509.

[15] D'Mello, S.K. et al. 2015. Multimodal capture of teacher-student interactions for automated dialogic analysis in live classrooms. *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction - ICMI '15* (New York, New York, USA, 2015), 557–566. DOI:https://doi.org/10.1145/2818346.2830602.

[16] Danielson, C. 2012. Observing classroom practice. *Educational Leadership*. 70, 2 (2012), 32–37.

[17] Donnelly, P.J. et al. 2016. Automatic teacher modeling from live classroom audio. *Proceedings of the 24th Conference on User Modeling, Adaptation, and Personalization (UMAP 2016)* (2016).

[18] Donnelly, P.J. et al. 2017. Words matter: automatic detection of questions in classroom discourse using linguistics, paralinguistics, and context. *LAK '17: Proceedings of the Seventh International Conference on Learning Analytics and Knowledge* (2017).

[19] Duckworth, A.L. et al. 2010. Deliberate practice spells success: why grittier competitors triumph at the national spelling bee. *Social Psychological and Personality Science*. 2, 2 (2010), 174–181. DOI:https://doi.org/10.1177/1948550610385872.

[20] Duke, N. et al. 2012. *Reading and writing genre with purpose in a k-8 classroom*. Heinemann.

[21] Ericsson, K.A. 2006. The influence of experience and deliberate practice on the development of superior expert performance. *The Cambridge Handbook of Expertise and Expert Performance*. K.A. Ericsson et al., eds. Cambridge University Press. 685–706.

[22] Ericsson, K.A. et al. 1993. The role of deliberate practice in the acquisition of expert performance. *Psychological Review*. 100, 3 (1993), 363–406. DOI:https://doi.org/10.1037/0033-295x.100.3.363.

[23] Eyben, F. et al. 2013. Recent developments in opensmile, the munich open-source multimedia feature extractor. *Proceedings of the 21st ACM international conference on Multimedia - MM '13* (New York, New York, USA, 2013), 835–838. DOI:https://doi.org/10.1145/2502081.2502224.

[24] Fadde, P.J. and Klein, G.A. 2010. Deliberate performance: accelerating expertise in natural settings. *Performance Improvement*. 49, 9 (2010), 5–14. DOI:https://doi.org/10.1002/pfi.

[25] Ford, M. et al. 2008. *The LENA Language Environment Analysis System*.

[26] Gamoran, A. and Nystrand, M. 1992. Taking

students seriously. *Student Engagement and Achievement in American Secondary Schools*. F.M. Newmann, ed. Teachers College Press. 40–61.

[27] Garet, M.S. et al. 2011. *Middle School Mathematics Professional Development Impact Study: Findings After the Second Year of Implementation*.

[28] Garet, M.S. et al. 2008. *The Impact of Two Professional Development Interventions on Early Reading Instruction and Achievement*.

[29] Goldman, R. et al. eds. *Video research in the learning sciences*. Erlbaum.

[30] Graesser, A.C. et al. 2009. What is a good question? *Threads of coherence in research on the development of reading ability*. M.G. McKeown and L. Kucan, eds. Guilford Press. 112–141.

[31] Grossman, P. et al. 2013. Measure for measure: the relationship between measures of instructional practice in middle school english language arts and teachers' value-added scores. *American Journal of Education*. 119, 3 (2013), 445–470. DOI:https://doi.org/10.1086/669901.

[32] Gwet, K.L. 2008. Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*. 61, (2008), 29–48.

[33] Hill, H.C. et al. 2008. Mathematical knowledge for teaching and the mathematical quality of instruction: an exploratory study. *Cognition and Instruction*. 26, 4 (2008), 430–511.

[34] Ho, A.D. and Kane, T.J. 2013. *The Reliability of Classroom Observations by School Personnel*.

[35] Hochreiter, S. and Schmidhuber, J. 1997. Long short-term memory. *Neural Computation*. 9, 8 (Nov. 1997), 1735–1780. DOI:https://doi.org/10.1162/neco.1997.9.8.1735.

[36] Holstein, K. et al. 2018. Student learning benefits of a mixed-reality teacher awareness tool in ai-enhanced classrooms. *Artificial Intelligence in Education* (2018), 154–168. DOI:https://doi.org/10.1007/978-3-319-93843-1_12.

[37] Juzwik, M.M. et al. 2013. *Inspiring dialogue: talking to learn in the english classroom*. Teachers College Press.

[38] Kelly, S. et al. 2018. Automatically measuring question authenticity in real-world classrooms. *Educational Researcher*. 47, 7 (2018), 451–464.

[39] Kelly, S. 2007. Classroom discourse and the distribution of student engagement. *Social Psychology of Education*. 10, 3 (2007), 331–352.

[40] Lemaître, G. et al. 2017. Imbalanced-learn: a python toolbox to tackle the curse of imbalanced datasets in machine learning. *The Journal of Machine Learning Research*. 18, (2017), 559–563. DOI:https://doi.org/http://www.jmlr.org/papers/volume18/16-365/16-365.pdf.

[41] Mikolov, T. et al. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems* (2013), 3111–3119.

[42] Molenaar, I. and Knoop-van Campen, C. 2017. Teacher dashboards in practice: usage and impact. *12th European Conference on Technology Enhanced Learning, EC-TEL 2017* (Tallinn, Estonia, 2017), 125–138. DOI:https://doi.org/10.1007/978-3-319-66610-5_10.

[43] Molenaar, I. and Knoop-van Campen, C.A.N. 2019. How teachers make dashboard information actionable. *IEEE Transactions on Learning Technologies*. 12, 3 (Jul. 2019), 347–355. DOI:https://doi.org/10.1109/TLT.2018.2851585.

[44] Murphy, P.K. et al. 2009. Examining the effects of classroom discussion on students' comprehension of text: a meta-analysis. *Journal of Educational Psychology*. 101, 3 (2009), 740–764. DOI:https://doi.org/10.1037/a0015576.

[45] Newman, F.M. et al. 1992. The significance and sources of student engagement. *Student Engagement and Achievement in American Secondary Schools*. F.M. Newman, ed. Teachers College Press. 11–39.

[46] Nystrand, M. et al. 1997. *Opening dialogue: understanding the dynamics of language and learning in the english classroom*. Teachers College Press.

[47] Nystrand, M. et al. 2003. Questions in time: investigating the structure and dynamics of unfolding classroom discourse. *Discourse Processes*. 35, 2 (Mar. 2003), 135–198. DOI:https://doi.org/10.1207/S15326950DP3502_3.

[48] Olney, A.M. et al. 2017. Assessing the dialogic properties of classroom discourse: proportion models for imbalanced classes. *Proceedings of the 10th International Conference on Educational Data Mining* (2017), 162–167.

[49] Olney, A.M. et al. 2003. Utterance classification in autotutor. *Proceedings of the HLT-NAACL 03 workshop on Building educational applications using natural language processing* (2003), 1–8. DOI:https://doi.org/10.3115/1118894.1118895.

[50] Pedregosa, F. et al. 2011. Scikit-learn: machine learning in python. *Journal of Machine Learning Research*. 12, (2011), 2825–2830. DOI:https://doi.org/10.1007/s13398-014-0173-7.2.

[51] Pennington, J. et al. 2014. Glove: global vectors for word representation. *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (2014), 1532–1543.

[52] Ramakrishnan, A. et al. 2019. Toward automated classroom observation : predicting positive and negative climate. *IEEE Conference on Automatic Face and Gesture Recognition* (2019).

[53] Raudenbush, S.W. et al. 1993. Higher order

instructional goals in secondary schools: class, teacher, and school influences. *American Educational Research Journal*. 30, 3 (Jan. 1993), 523–553. DOI:https://doi.org/10.3102/00028312030003523.

[54] Resnick, L.B. and Schantz, F. 2015. Re-thinking intelligence: schools that build the mind. *European Journal of Education*. 50, 3 (2015), 340–349. DOI:https://doi.org/10.1111/ejed.12139.

[55] Rodriguez-Triana, M.J. et al. 2017. Monitoring, awareness and reflection in blended technology enhanced learning: a systematic review. *IJTEL*. 9, 2–3 (2017), 126–150.

[56] Samei, B. et al. 2014. Domain independent assessment of dialogic properties of classroom discourse. *Proceedings of the 7th International Conference on Educational Data Mining (EDM)*. (2014), 233–236.

[57] Samei, B. et al. 2015. Modeling classroom discourse: do models that predict dialogic instruction properties generalize across populations? *Proceedings of the 8th International Conference on Educational Data Mining* (2015), 444–447.

[58] Saon, G. et al. 2015. The ibm 2015 english conversational telephone speech recognition system. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH* (Dresden, Germany, 2015), 3140–3144. DOI:https://doi.org/10.21437/Interspeech.2016-1460.

[59] Schuller, B. et al. 2009. The interspeech 2009 emotion challenge. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*. (2009), 312–315.

[60] Shernoff, D.J. 2013. *Optimal learning environments to promote student engagement*. Springer New York.

[61] Shernoff, D.J. et al. 2016. Student engagement as a function of environmental complexity in high school classrooms. *Learning and Instruction*. 43, (2016), 52–60. DOI:https://doi.org/10.1016/j.learninstruc.2015.12.003.

[62] Shernoff, D.J. et al. 2003. Student engagement in high school classrooms from the perspective of flow theory. *School Psychology Quarterly*. 18, 2 (2003), 158.

[63] Shute, V.J. 2008. Focus on formative feedback. *Review of Educational Research*. 78, 1 (2008), 153–189. DOI:https://doi.org/10.3102/0034654307313795.

[64] Stigler, J.W. et al. 1999. *The TIMSS Videotape Classroom Study: Methods and Findings from an Exploratory Research Project on Eighth-Grade Mathematics Instruction in Germany, Japan, and the United States. A Research and Development Report.*

[65] Stigler, J.W. and Miller, K.F. 2018. Expertise and expert performance in teaching. *The Cambridge Handbook of Expertise and Expert Performance*. K.A. Ericsson et al., eds. Cambridge University Press. 431–452.

[66] Stone, C. et al. 2019. Utterance-level modeling of indicators of engaging classroom discourse. *The 12th International Conference on Educational Data Mining* (Montreal, Canada, 2019), 420–425.

[67] Suresh, A. et al. 2019. Automating analysis and feedback to improve mathematics' teachers' classroom discourse. *Proceedings of the Ninth Symposium on Educational Advances in Artificial Intelligence (EAAI)* (2019).

[68] Suresh, A. et al. 2018. Using deep learning to automatically detect talk moves in teachers'mathematics lessons. *Proceedings - 2018 IEEE International Conference on Big Data, Big Data 2018*. (2018), 5445–5447. DOI:https://doi.org/10.1109/BigData.2018.8621901.

[69] Taylor, B.M. et al. 2003. The influence of teacher practices that encourage cognitive engagement in literacy learning. *The Elementary School Journal*. 104, 1 (2003), 3–28.

[70] TNTP 2015. *The Mirage: Confronting the Hard Truth About Our Quest for Teacher Development*.

[71] Vaswani, A. et al. 2017. Attention is all you need. *Advances in neural information processing systems* (2017), 5998–6008.

[72] Wang, Z. et al. 2013. Using the lena in teacher training: promoting student involvement through automated feedback. *Unterrichtswissenshaft*. 4, (2013), 290–305.

[73] Xhakaj, F. et al. 2017. Effects of a teacher dashboard for an intelligent tutoring system on teacher knowledge, lesson planning, lessons and student learning. *EC-TEL*. 315–329.