Automatic Classification of Question & Answer Discourse Segments from Teacher's Speech in Classrooms

Nathaniel Blanchard
University of Notre Dame
384 Fitzpatrick Hall
Notre Dame, IN 46556, USA
nblancha@nd.edu

Sidney D'Mello
University of Notre Dame
384 Fitzpatrick Hall
Notre Dame, IN 46556, USA
sdmello@nd.edu

Andrew M. Olney
University of Memphis
365 Innovation Drive
Memphis, TN 38152, USA
aolney@memphis.edu

Martin Nystrand
University of Wisconsin-Madison
685 Education Sciences
Madison WI, 53706-1475
mnystrand@ssc.wisc.edu

ABSTRACT

Question-answer (Q&A) is fundamental for dialogic instruction, an important pedagogical technique based on the free exchange of ideas and open-ended discussion. Automatically detecting Q&A is key to providing teachers with feedback on appropriate use of dialogic instructional strategies. In line with this, this paper studies the possibility of automatically detecting segments of Q&A in live classrooms based solely on audio recordings of teacher speech. The proposed approach has two steps. First, teacher utterances were automatically detected from the audio stream via an amplitude envelope thresholding-based approach. Second, supervised classifiers were trained on speech-silence patterns derived from the teacher utterances. The best models were able to detect Q&A segments in windows of 90 seconds with an AUC (Area Under the Receiver Operating Characteristic Curve) of 0.78 in a manner that generalizes to new classes. Implications of the findings for automatic coding of classroom discourse are discussed.

Keywords

Dialogic instruction, teacher feedback, professional development, live classrooms, speech, learning

1. INTRODUCTION

Dialogic instruction, a form of classroom discourse based around the free exchange of ideas and open-ended discussion, is considered to be an important pedagogical approach to increase student engagement [11] and improve student achievement [24]. However, the quality of implementation of dialogic instruction in classrooms varies widely. Recent research has demonstrated the importance of formative assessment of teacher use of dialogic instruction in classrooms [10]. Providing formative feedback based on what actually occurs in classrooms allows teachers to focus their efforts on improving the quality of dialogic instruction over time. Providing formative feedback efficiently, accurately, and automatically on a day-to-day basis will ensure that teachers receive the feedback they need to better incorporate dialogic instructional practices into their classrooms. However, large-scale efforts to assess the quality of classroom discourse have relied on manual, labor-intensive, and expensive excursions into classrooms. The automation of classroom discourse analysis to inform personalized formative assessment and training programs has the potential to transform teachers' use of dialogic instruction and thereby improve student outcomes. This is the overarching goal of the current project, called CLASS 5.

The CLASS 5 project is focused on automatically analyzing classroom discourse as a means of providing feedback to teachers. CLASS 5 is intended to be a modern adaptation of the traditional model of requiring trained observers to manually code classroom discourse, an unsustainable task for providing day-to-day feedback for professional development. The automated analyses are grounded in the coding scheme of Nystrand and Gamoran [6,19], who observed thousands of students across hundreds of middle and high school English Language Arts classes. They found that the overall dialogic quality of classroom discourse through teacher's use of authentic questions (questions without prescripted responses), uptake (integration of previous speaker's ideas into future questions), and classroom discussion had positive effects on student achievement. The Nystrand and Gamoran coding scheme has been validated in multiple studies across a multitude of classrooms [2,7,17,18], hence, we are optimistic that by automating this coding scheme, we will replicate the well substantiated results of finding positive effects of dialogic instruction on student achievement. In the remainder of this section, we provide a brief overview of the Nystrand and Gamoran coding scheme, review prior work on automated classroom discourse analysis, and provide a brief overview of the present study, which is focused on automatically detecting question-answer (Q&A) segments via audio recordings of teachers during normal classroom instruction.

1.1 Coding Classroom Discourse

The Nystrand and Gamoran [6,19] coding scheme can be subdivided into three key '*tracks*,' of increasingly fine granularity: 1) episodes, which refer to the activity/topic being addressed by the teacher; 2) segments, seventeen categories that represent possible techniques used to implement the episode; and 3) questions asked by teachers or students embedded within segments [19]. Each track can be further understood by its own nuance and properties. For example, many classes typically begin

and end with procedural episodes (i.e., "getting started"; "preparing to leave") with one or more instructional episodes permeating the core of the class. All episodes consist of one or more segments, which can be broadly subdivided into four categories: classroom management activities, direct instruction, seatwork, and tests and quizzes. Questions are coded along dimensions of authenticity, uptake, and cognitive level as elaborated in [19].

Our current focus is on classifying key *segments* in classroom discourse. Of the seventeen segment categories the most frequent segments are lecture (including film, music, or video), Q&A, reading aloud, supervision/helping, and small group work [19]. Lecture incorporates instances where a teacher speaks for at least 30 seconds on a topic unrelated to the procedural aspects of running a class (discussing assignment instructions, for example, would not be considered lecture). Q&A segments include a question or series of questions which are non-rhetorical, non-procedural, and non-discourse management questions. Reading aloud segments consist of students reading aloud. Supervised/helping segments occur when teachers help students complete individual work. Small group work segments occurs when a group of students participates in some activity.

Discussions constitute an important, but rare, segment of particular relevance to dialogic instruction. According to the coding scheme, discussion segments consist of a free exchange among three or more participants that lasts longer than 30 seconds. Discussions typically include relatively few questions. Questions that are asked tend to focus on clarification of ideas. Discussions are typically initialized when a student makes an observation, rather than asking a question, and another student or a teacher asks for clarification on that observation. In contrast, Q&A segments usually consist of three parts – an initiation, a response, and an evaluation (IRE). The most common example of these parts begins with a teacher question, followed by a student answer, and then a teacher response to the student's answer. The teacher's response is often perfunctory (e.g. 'right' or 'wrong') – and sometimes non-vocalized (i.e., a nod) [16,18].

Q&A and discussion segments have traditionally positively correlated with achievement, and it is recommended that teachers should attempt to maximize use of these segments [19]. As mentioned above, discussion segments are rare in classrooms. In Nystrand's observations there was on average less than one minute of discussion per class [19]. Traditionally Q&A segments have dominated between 30% - 42% of class time [19]. In fact, when discussion does occur it tends to do so in the midst of Q&A segments. Therefore, the present study focuses on the automated detection of Q&A segments as an initial approach to automating the coding of classroom discourse.

1.2 Related Work

The closest work in this area stems from research by Wang and colleagues. In particular, Wang et. al. [26] used teacher and student speech features obtained by the Language Environment Analysis system (LENA) [5] to analyze discourse profiles from 1st to 4th grade math classes. LENA is a wearable system which records and measures the quality of language produced by and directed at young children. Wang et. al. had two trained coders listen to 30-second audio windows and classify if the window represented discussion, lecture, or group work. Coders also provided their confidence in their annotation on a scale of 1 to 3 (1 indicating a lack of confidence and 3 indicating very confident).

LENA was adapted to assess when teachers were speaking, students were speaking, speech was overlapping, or there was silence. Wang et al. [25] previously found that LENA coded many student utterances as teacher utterances and modified LENA to improve its voice detection accuracy by changing the categorization algorithm to account for volume as an indicator of the distance between the speaker and the microphone. Their precision for teacher speech detection ranged from 0.95 – 0.99 and their precision for student speech detection ranged 0.70 to 0.86.

They then trained a random-forest classifier to classify the 30-second windows based on the results of speech segmentation. They used one coder's confidence labels of 3 for training data. This constituted 62% of the windows. They validated their model on all of the windows (including the training windows), but with the annotations provided by a different coder. The coders agreed on 83% (Kappa 0.72) of the annotations, so there was considerable overlap between training and testing data. Their model achieved an accuracy of 83% (Kappa of 0.73) in discriminating between lecturing, discussion, and group work.

Although Wang et. al. [26] reported success at classifying classroom discourse at course-grained levels, their audio solution was focused on what occurred in the context of individual windows, rather than using the broader classroom context to code segments. Further, according to Wang's coding, discussion occurred approximately 33% of the time, indicating their definition of discussion was much more inclusive than the Nystrand & Gamoran coding scheme [6,19]. Their definition of discussion, which involved students and teachers having conversations about the learning content on the whole class level (the conversation should be accessible to the majority of students in class), is not incorrect, but more closely aligns with our definition of Q&A segments. In addition, their validation method did not include an independent class-level hold-out set, thus evidence for generalizability to new classes is unclear.

1.3 Current Study

The present study takes inspiration from Wang et al.'s pioneering work, but also differs from it in significant ways. The LENA system is a research-grade solution and is thereby cost prohibitive and might not be scalable. This raises the question of whether classroom discourse can be automatically analyzed using more cost effective consumer-grade sensors. Of particular interest is addressing which signals are needed for accurate automatic classification of classroom discourse. Teachers lead dialogic instruction and one possibility is the only signals needed to capture classroom activity are signals that capture teacher activity. Since teachers may be anywhere in a classroom, data needs to be collected from a device that accompanies their movements with high fidelity. One attractive candidate for such a sensor is a microphone to record teacher speech, which is the approach adopted here.

Recording teacher speech is not a difficult task, but distilling the signal into appropriate features for classification of Q&A segments is more complicated. Thus, we first focused our efforts on teacher utterance detection in an attempt to find the onsets and offsets of teacher speech. Features extracted from these onsets and offsets, signaling periods of speech and rest, were then used to train classifiers to discriminate Q&A segments from all other

segments combined (i.e., Q&A vs. "other" discriminations). Note that all classification is done by analyzing these utterance onsets and offsets in an attempt to establish the accuracy of Q&A segment classification using a minimalistic approach.

The key differences between the present approach and Wang's previous work include: (a) our use of a consumer-grade microphone rather than the LENA system; (b) segments are coded during live classrooms, so that the overarching classroom context can be incorporated in the coding; (c) we study Q&A segment classification by exclusively focusing on the teacher speech signal; and (d) our models are validated across class sessions, thereby ensuring generalizability to new classes.

The remainder of the paper is organized as follows. First, we discuss our data collection, which involved coders trained in Nystrand's coding scheme collecting data from three teachers in 21 class sessions over the course of a semester (Section 2). We recorded teacher speech using a headset microphone and the audio signal was temporally synchronized with the human codings. Next, we developed an amplitude envelop-based utterance detection approach to segment the teacher audio into periods of speech and rest (Section 3). Then, supervised classifiers were used to detect Q&A segments from features extracted by the utterance detection algorithm (Section 4). Implications of our findings towards the broader goal of automating the analysis of classroom discourse at multiple-levels are discussed (Section 5).

2. Data Collection

Audio recordings were collected at a rural Wisconsin middle school during literature, language arts, and civics classes. The recordings were of three different teachers: two males – Speaker 1 and Speaker 2 – and one female – Speaker 3. The recordings spanned classes of about 45 minutes each on 9 separate days over a period of 3-4 months. Due to the occasional missed session, classroom change, or technical problem, a total of 21 classroom recordings were available for analyses. During each class session, teachers wore a Samson AirLine 77 'True Diversity' UHF wireless headset microphone that recorded their speech, with the headset hardware gain adjusted to maximum. This microphone was chosen for its high noise-cancelling ability and is not cost-prohibitive ($300 per unit). Audio files were saved in 16 kHz, 16-bit mono .wav format. Teachers were recorded naturalistically while they taught their class as usual.

Two observers trained in Nystrand et. al.'s dialogic coding technique [19,20] were present in the classroom during recordings. Observers used a specialized coding software developed by Nystrand [15] to mark episodes, segments, and teacher's dialogic questions with the appropriate labels, as well as start and stop times as the class progressed. Later, these same observers reviewed the recordings to ensure labels were accurate and engaged in discussion until all discrepancies were resolved.

Table 1 lists the proportion of time spent on each of the segments. We note that Q&A segments were the most frequent, while discussions were highly infrequent. Other somewhat frequent segments include small group work, supervised/helping, and lecture/film/video/music. The subsequent analyses focus on detecting the 28.6% Q&A segments from all other segments combined.

Table 1. Proportion of class time on each segment

Segment	Proportion
Question/answer	**0.286**
Small Group Work	0.160
Supervised/helping	0.158
Lecture/film/video/music	0.150
Reading Aloud	0.093
Procedures and directions	0.091
Supervised/monitoring	0.019
Silent Reading	0.017
Other	0.012
Unsupervised seatwork	0.006
Class interruption	0.003
Game	0.002
Discussion	0.001

3. TEACHER UTTERANCE DETECTION

Our overall objective was to use teacher speech to detect instances of question-and-answer using recorded audio from classrooms. Before this could be done, recorded audio needed to be distilled into instances of teacher speech vs. rest (silence or no speech). Thus, we developed and validated an utterance detection method as discussed below.

3.1 Method

Our first assumption was that all sound was voice because teacher speech was recorded from a high-quality noise-canceling headset microphone, all sound was voice and that no advanced voice activity detection (VAD) techniques were required[1]. Thus, a simple binary procedure was used for utterance detection. The amplitude envelope of the teacher's low-pass filtered speech was passed through a threshold function in 20 millisecond increments. Where the amplitude envelope was above threshold, the teacher was considered to be speaking. Where the amplitude envelope was below threshold, the teacher was assumed to not be speaking. Any time speech was detected, that speech was considered part of an utterance, meaning there was no minimum threshold for how short an utterance could be. Utterances were marked as complete when speech stopped for 1000 milliseconds (1 second). A typical result of this automatic utterance labeling method is depicted in Figure 1.

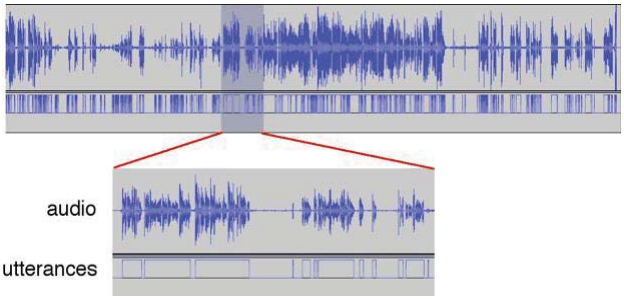

Figure 1: A 45-minute class recording (top) is depicted, while a small portion of the recording is enlarged for a detailed view (bottom). The upper track visualizes the .wav form of the audio. The lower track visualizes detected utterances.

[1] We also experimented with off-the-shelf voice activity detection algorithms [22], with comparable, if not slightly inferior, results.

The speech delimiter and threshold were both low to ensure all speech was detected, resulting in no known cases of missed speech. This process resulted in 8662 utterances, which we call *potential speech utterances*. An examination of a subset of these potential speech utterances indicated that there were a large number of false alarms. These were mainly attributed to instances of background noise permeating the audio. Common examples of background noise that the microphone picked up included voices of students who were being exceptionally loud, sounds from a film or audio clip being played in the classroom, and sounds of the teacher's breathing.

A two-step filtering approach was taken to eliminate the false alarms. First, potential utterances less than 125 milliseconds in length (12% in all) were deemed to be too short to contain meaningful speech and were eliminated. Second, the remaining potential speech utterances were submitted through an automatic speech recognizer (Bing Speech) in an effort to identify the false alarms. Bing Speech [13] is a freely available, cloud-based automatic speech recognition service which supports seven languages. Bing returns a recognition result and a confidence score for that speech. Instances where Bing rejected the speech or where it returned no transcribed text were considered to be false alarms. After eliminating the false alarms, we were left with a total of 5502 utterance (64% of the 8662 potential utterances).

3.2 Validation

A small study was conducted to evaluate the aforementioned utterance detection method. A random sample of 500 potential utterances was selected and manually annotated for speech/non-speech. Speech was defined to include all articulations (i.e., "um", "hm", "sh", etc) in addition to normal spoken segments. Potential speech utterances that included noise (i.e., loud students) in addition to teacher speech, the utterance was deemed as being a spoken utterance since it contained teacher speech. In total, 63% of potential utterances contained teacher speech and 37% did not. Thus, the effective false alarm rate prior to discarding utterances less than 125 milliseconds in length and accepted by Bing Speech was 37%.

Table 2 presents the confusion matrix obtained when using the 125 millisecond utterance duration threshold and Bing Speech to eliminate false alarms in the sample of 500 potential utterances. The filtering approach was highly successful, resulting in a kappa of 0.93 (agreement between computer-detected teacher utterances and human-detected teacher utterances). We note a substantially high hit and correct rejection rates and very low false alarms and miss rates. This was deemed to be sufficiently accurate for the present goal of detecting Q&A segments from teacher speech.

Table 2. Descriptive Statistics of Utterances

	Predicted	
Actual	*Speech*	*Non-Speech*
Speech	0.96 (hit)	0.04 (false alarm)
Non-Speech	0.03 (miss)	0.97 (correct rejection)

4. CLASSIFYING Q&A SEGMENTS

Segments were coded in the classrooms of three teachers in 21 classes by trained coders over the course of a semester. Our goal was to differentiate Q&A segments, which are key for dialogic instruction, from all other types of segments (a binary Q&A segment vs. "other" classification task). Features for Q&A segment classification were obtained from the automated teacher speech utterance detection approach discussed above.

4.1 Method

4.1.1 Creating and labeling instances

Audio was sectioned into non-overlapping windows of 30, 45, 60, 75, and 90 seconds in length. Each window was assigned a label of "Q&A" or "other" based on the annotations by the trained coders (see Section 2). In some cases, there was overlap, defined as a window with multiple segment labels (e.g., first 20 seconds are Q&A and the last 10 seconds are lecture). For windows with overlap, the label of "Q&A" or "other" was assigned based on the label of the majority segment (e.g., Q&A in the example above).

Table 3 presents the number of windows and the proportion of windows that contain overlap for each window size. As expected, the proportion of windows with overlap increases as window size is increased.

Table 3. Number of instances and proportion of instances with overlap

Window	**N**	**N (with overlap)**	**Proportion with overlap**
30 seconds	1886	163	0.09
45 seconds	1253	145	0.12
60 seconds	937	126	0.13
75 seconds	748	126	0.17
90 seconds	620	112	0.18

Note: N = Total number of windows in a dataset

4.1.2 Feature Engineering

Features were based on teacher utterance detection as discussed in Section 3. The features attempt to capture the temporal speech patterns that teachers use in Q&A segments as defined by the initiation (speech), response (rest), and evaluation (speech) pattern of Q&A discussed in Section 1.1. They include: 1) number of utterances, 2) mean utterance duration, 3) standard deviation of utterance duration 4) minimum utterance duration 5) maximum utterance duration, 6) number of rests, 7) mean rest duration (rests were the intervals of silence between utterances), 8) standard deviation of rest duration, 9) minimum rest duration, 10) maximum rest duration, and 11) window number, the number of windows into a class session.

4.1.3 Model Building

Supervised classifiers were built using the Waikato Environment for Knowledge Analysis (WEKA) [9] an open source data mining tool. Models were cross validated on the class level to ensure generalizability across class sessions. In each fold, a random 67% of the classes were used for training and the remaining 33% were used for testing. This process was repeated for 25 iterations and the classification accuracy metrics was averaged across these iterations. A large number (N = 43) of standard classifiers were tested because of a lack of knowledge regarding what classifier works best for this type of data.

Various data treatments were applied in order to determine which combination resulted in the best model. First, tolerance analysis was used to eliminate features that exhibited multicollinearity [1]. Second, four feature selection algorithms: 1) Information Gain Ratio (Info-Gain) [14], 2) RELIEF-F [12], 3) Gain-Ratio [21], and 4) Correlation-based Feature Selection (CFS) [8] were used

(on training data only) to select either 25%, 50%, or 75% of the top features (the specific percentage of features was another parameter). Third, the data was Winsorized by setting outliers greater than 3 standard deviations from the mean to the corresponding value 3 standard deviations from the mean. Finally, synthetic minority oversampling technique (SMOTE) [4] was applied to the training data by creating synthetic instances of the minority Q&A class until the classes were balanced. Testing data was not sampled.

4.2 Results

4.2.1 Best Models

Classification accuracy was evaluated with area under the receiver operating characteristic curve (AUC), a metric bounded on [0, 1] with 1 indicating perfect classification and 0.5 indicating chance level classification. Table 4 presents an overview of the AUCs associated with the best models for each window size. The mean AUC across all windows was 0.73 (SD = 0.05). Classification accuracy was greater for longer window sizes with the best results obtained for the 90 second window. This model used a logistic regression classifier and had 5 features (discussed below). Table 5 presents the confusion matrix for this 90 second window model. The main source of errors appear to be misses rather than false alarms.

Table 4. AUC for best models at each window size

Window Size	AUC
30 secs	0.67 (0.04)
45 secs	0.69 (0.05)
60 secs	0.75 (0.04)
75 secs	0.75 (0.04)
90 secs	0.78 (0.05)

Note: Standard Deviation in parenthesis

Table 5. Confusion matrix for best model using class-level cross-validation

	Predicted		
Actual	*Q&A*	*Other*	**Priors**
Q&A	0.78 (hit)	0.22 (false alarm)	0.26
Other	0.36 (miss)	0.64 (correct rejection)	0.74

4.2.2 Robustness to Overlap

One concern was whether classification accuracy was degraded due to instances where Q&A segments overlapped other segments within a window. As presented in Section 4.1, the larger the window size, the greater proportion of instances that contain overlap. To study the effect of overlap, we built another set of models with overlapping segments removed.

Performance of models without overlapping windows was consistent compared to models with overlapping windows (see Table 4). Mean AUC for the models built without overlap was 0.74 (SD = 0.04) compared with mean AUC from Section 4.2.1: 0.73 (SD = 0.05). Thus, our best models were robust to instances where Q&A segments overlapped with other segments within a window.

4.2.3 Feature Analysis

We analyzed the five features used in the best model (90 second window). These features were 1) number of utterances, 2) mean utterance duration, 3) maximum utterance duration, 4) mean rest duration, 5) maximum rest duration. Table 6 presents the mean and standard deviation for these top features across the four most frequent segments (see Table 1). All non-Q&A segments included a fewer number of utterances, shorter utterance durations, and fewer silences (rest). For lecture/media this was likely a result of the all-inclusiveness nature of lecture/media which could include instances of only speech, a traditional lecture, or instances of no speech (e.g., when a film is played). For group work, this was likely because speech consisted of clarifying instructions or addressing individual group concerns. Supervised/helping was likely similar to group work, but rather than group concerns, individual concerns were addressed.

Table 6. Mean and standard deviation for features across most frequent segments

Feature	Q&A	Lecture/ Media	Small Group Work	Supervised/ Helping
Number of utterances	10.45 (4.82)	4.86 (5.16)	8.90 (4.32)	7.38 (4.46)
Mean utterance duration	5.19 (4.15)	3.23 (4.37)	2.76 (1.83)	2.80 (1.92)
Maximum utterance duration	14.62 (9.85)	7.77 (9.44)	7.80 (5.69)	8.14 (7.02)
Mean rest duration	5.40 (4.67)	38.71 (37.26)	12.22 (19.23)	17.57 (24.77)
Max rest duration	15.92 (11.71)	50.42 (33.53)	27.91 (22.31)	35.51 (25.60)

Note: Standard Deviation in parenthesis

5. General Discussion

Dialogic instruction is considered to be an important pedagogical approach for promoting learning and engagement in classrooms. However, analyzing the effective use of dialogic instruction in classrooms has traditionally required the presence of trained live coders and is inherently non-scalable. In the present paper, we considered the possibility of automating the coding of classroom discourse. As an initial step, we focused on automatically detecting question-and-answer (Q&A) segments, an important component of dialogic instruction, using teacher speech. We were able to detect instances of Q&A from teacher speech with moderate success in live classrooms. In this section, we compare our results to previous work in this area, discuss major findings, limitations of the present study, and consider next steps with this research.

5.1 Comparing with Previous Work

Our goal was to compare our approach, which only uses features from teacher speech, with models from Wang et al. [26], which were based on teacher speech, student speech, overlapping speech, and silence. A perfect comparison is complicated due to many differences across approaches, most importantly with

respect to how classroom activities were coded and how the models were validated. In particular, coders in the Wang et al. study annotated their data using 30-second intervals and specified a confidence level for each annotation. This allowed them to train their models on only the high-confidence labels. In comparison, we used a variety of different window sizes and our labels did not include a confidence level.

Our best model, which used a logistic regression classifier, had a kappa of 0.32, which is much lower than Wang et al.'s kappa of 0.77. To equate models, we also experimented with using a random forest model [3], used by Wang et al. Using a random forest model and validating at the class-level resulted in an AUC of 0.71 (SD = 0.04) and a lower kappa of 0.25 (SD = 0.07). However, we noted that Wang et al. validated their data using both training and testing data, while our models were validated on held-out class sessions. In other words, 62% of their testing data contained training instances. We attempted to replicate their validation approach by randomly selecting 62% of training instances for inclusion in the testing data. This drastically increased the AUC to 0.87, with a Kappa of 0.57.

In conclusion, although our model's performance is lower than Wang et al's, there are many possible reasons for this difference. For example, differences in our definitions of Q&A, their coding of each window devoid of context (which could lead to misinterpreting a window due to lack overall of context), different recording setups (LENA vs. microphone), different class structures (elementary mathematics vs. middle-school literature, language arts, and civics classes), and so on. Future work needs to equate these differences so the two approaches can be compared more equitably.

5.2 Major Findings

We were moderately successful in detecting Q&A segments despite considerable challenges associated with automatically recording classroom discourse using only teacher speech recorded via a headset microphone. Our major contribution is the use of consumer grade equipment to filter teacher utterances from non-teacher utterances in a noisy classroom environment. We found that we could use those utterances to develop and validate Q&A segment detectors in classrooms using only teacher speech.

Our approach consisted of two steps. Step 1 involved segmenting teacher utterances and Step 2 involved analyzing speech-silence dynamics from this segmentation to train classifiers suitable for discriminating Q&A segments from all other coded segments. For utterance detection, we used an amplitude enveloping approach to identify a large subset of potential teacher utterances and filtered them based on both duration and by submitting them to a web-based automatic speech recognizer (Bing Speech). We validated the utterance detection approach using a sample of 500 potential speech utterances randomly sampled from three teachers and 21 class sessions. We reliably and accurately discriminated speech from non-speech (kappa of 0.93) and this was accomplished despite the complexities of teacher utterance detection in noisy classrooms such as loud student speech, classroom disruptions, the use of media (i.e., video, music), and non-articulations of the teacher (such as breathing).

For Step 2, we built models to classify instances of Q&A from other instructional activities using speech-silence dynamics from the utterance segmentation. The best model was a logistic regression classifier trained on speech and silence features in 90 second windows which yielded an AUC of 0.78 when validated at the class-level. We also built models without overlap in order to determine their effect. The models without overlap were equitable to models with overlap, indicating our models were robust to this issue. Finally, we analyzed the top features from our best model and the main finding was that Q&A segments were associated with more teacher speech and fewer rests compared to the other segments.

5.3 Limitations and Future Work

This study was not without its limitations. First, data was collected from three teachers who taught different subjects. However, this is a small number of teachers and all taught at the same school, so replication with a larger and more diverse sample is warranted. Second, discussion is a key indicator of dialogic discourse in classrooms [19], but our data set had only one instance of discussion, which lasted 77 seconds. Thus models could not be built for this key activity. Finally, our method focuses on a coarse-grained measure of classroom discourse. Future research is needed before a fine-grained analysis of the types of questions being asked in Q&A segments can be done (see Samei et al. [23]). When we use Bing to filter speech, it returns recognition results which could potentially be used for these fine-grained analysis. This is an important item for future work.

In general, future data collection should include more teachers, schools, social environments, and class diversity. Future work should also consider ways to capture student speech in an equally cost effective way. One possibility would be to record the entire room with a boundary microphone. However, it should be noted that every additional sensor increases the complexity of data collection and raises the threshold of adaptation in terms of cost and complexity of use. For example, if using a boundary microphone to capture student speech, a teacher needs to learn where best to position the microphone. However, a headset microphone only requires a teacher to turn it on and wear it. Nevertheless, we anticipate much improved results in Q&A detection when student speech is available.

5.4 Concluding Remarks

The overall purpose of this research was to automate the coding of classroom discourse and the present paper made some advances in this direction. As Nystrand et al. found [19], professional development activities focused on increasing the quality of dialogic instruction can have measurable effects on student achievement. The automated classroom discourse analysis techniques developed here can contribute to this goal by providing daily feedback to teachers for their professional development. Although this feedback alone may allow teachers to better reflect on their classroom instruction, it remains to be seen whether this increases their use of appropriate techniques for dialogic instruction. If not, tracking key components of dialogic instruction allows for interventions to increase dialogic instruction in classrooms. The research presented here represents an important initial step toward these goals, the next step involving an analysis of individual question-events at a more fine-grained level.

6. ACKNOWLEDGMENTS

We would like to thank Dr. Michael Brady for the amplitude envelope processing method.

This research was supported by the Institute of Education Sciences (IES) (R305A130030). Any opinions, findings and

conclusions, or recommendations expressed in this paper are those of the author and do not represent the views of the IES.

7. REFERENCES

1. Allison, P.D. *Multiple regression: A primer.* Pine Forge Press, 1999.
2. Applebee, A.N., Langer, J.A., Nystrand, M., and Gamoran, A. Discussion-Based Approaches to Developing Understanding: Classroom Instruction and Student Performance in Middle and High School English. *American Educational Research Journal 40*, 3 (2003), 685–730.
3. Breiman, L. Random forests. *Machine learning 45*, 1 (2001), 5–32.
4. Chawla, N.V., Bowyer, K.W., Hall, L.O., and Kegelmeyer, W.P. SMOTE: synthetic minority over-sampling technique. *arXiv preprint arXiv:1106.1813*, (2011).
5. Ford, M., Baer, C.T., Xu, D., Yapanel, U., and Gray, S. *The LENA Language Environment Analysis System.* Technical Report LTR-03-2. Boulder, CO: LENA Foundation, 2008.
6. Gamoran, A. and Kelly, S. Tracking, instruction, and unequal literacy in secondary school English. *Stability and change in American education: Structure, process, and outcomes*, (2003), 109–126.
7. Gamoran, A. and Nystrand, M. Background and instructional effects on achievement in eighth-grade English and social studies. *Journal of Research on Adolescence 1*, 3 (1991), 277–300.
8. Hall, M.A. *Correlation-based Feature Selection for Machine Learning.* 1999.
9. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I.H. The WEKA Data Mining Software: An Update. *SIGKDD Explor. Newsl. 11*, 1 (2009), 10–18.
10. Juzwik, M.M., Borsheim-Black, C., Caughlan, S., and Heintz, A. *Inspiring Dialogue: Talking to Learn in the English Classroom.* Teachers College Press, 2013.
11. Kelly, S. Classroom discourse and the distribution of student engagement. *Social Psychology of Education 10*, 3 (2007), 331–352.
12. Kononenko, I. Estimating attributes: analysis and extensions of RELIEF. *Machine Learning: ECML-94*, Springer (1994), 171–182.
13. Microsoft. *The Bing Speech Recognition Control.* 2014. http://www.bing.com/dev/en-us/speech. Accessed 14 Jan 2015
14. Mitchell, T.M. Machine learning. 1997. *Burr Ridge, IL: McGraw Hill 45*, (1997).
15. Nystrand, M. *CLASS (Classroom language assessment system) 2.0: A Windows laptop computer system for the inclass analysis of classroom discourse.* Wisconsin Center for Education Research, Madison, 1988.
16. Nystrand, M. CLASS 4.0 user's manual. *The National Research Center on*, (2004).
17. Nystrand, M. Research on the Role of Classroom Discourse as It Affects Reading Comprehension. *Research in the Teaching of English 40*, 4 (2006), 392–412.
18. Nystrand, M. and Gamoran, A. Instructional discourse, student engagement, and literature achievement. *Research in the Teaching of English*, (1991), 261–290.
19. Nystrand, M., Gamoran, A., Kachur, R., and Prendergast, C. Opening dialogue. *Teachers College, Columbia University, New York and London*, (1997).
20. Nystrand, M., Wu, L.L., Gamoran, A., Zeiser, S., and Long, D.A. Questions in time: Investigating the structure and dynamics of unfolding classroom discourse. *Discourse processes 35*, 2 (2003), 135–198.
21. Quinlan, J.R. Induction of decision trees. *Machine learning 1*, 1 (1986), 81–106.
22. Rouvier, M., Dupuy, G., Gay, P., Khoury, E., Merlin, T., and Meignier, S. *An open-source state-of-the-art toolbox for broadcast news diarization.* Idiap, 2013.
23. Samei, B, Olney, A, Kelly, S, et al. Domain independent assessment of dialogic properties of classroom discourse. *Proceedings of the 7th International Conference on Educational Data Mining*, Stamper, J., Pardos, Z., Mavrikis, M., McLaren, B.M. (2014), 233–236.
24. Sweigart, W. Classroom Talk, Knowledge Development, and Writing. *Research in the Teaching of English 25*, 4 (1991), 469–496.
25. Wang, Z., Miller, K., and Cortina, K. *Using the LENA in Teacher Training: Promoting Student Involement through automated feedback.* na, 2013.
26. Wang, Z., Pan, X., Miller, K.F., and Cortina, K.S. Automatic classification of activities in classroom discourse. *Computers & Education 78*, (2014), 115–123.