Words Matter: Automatic Detection of Teacher Questions in Live Classroom Discourse using Linguistics, Acoustics, and Context

Patrick J. Donnelly
University of Notre Dame
Notre Dame, IN, 46556, USA
pdonnel4@nd.edu

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

Andrew M. Olney
University of Memphis
Memphis, TN 38152
aolney@memphis.edu

Sean Kelly
University of Pittsburgh
Pittsburgh, PA 15260
spkelly@pitt.edu

Martin Nystrand
University of Wisconsin, Madison
Madison, WI 53715
nystrand@wisc.edu

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

ABSTRACT
We investigate automatic detection of teacher questions from audio recordings collected in live classrooms with the goal of providing automated feedback to teachers. Using a dataset of audio recordings from 11 teachers across 37 class sessions, we automatically segment the audio into individual teacher utterances and code each as containing a question or not. We train supervised machine learning models to detect the human-coded questions using high-level linguistic features extracted from automatic speech recognition (ASR) transcripts, acoustic and prosodic features from the audio recordings, as well as context features, such as timing and turn-taking dynamics. Models are trained and validated independently of the teacher to ensure generalization to new teachers. We are able to distinguish questions and non-questions with a weighted F_1 score of 0.69. A comparison of the three feature sets indicates that a model using linguistic features outperforms those using acoustic-prosodic and context features for question detection, but the combination of features yields a 5% improvement in overall accuracy compared to linguistic features alone. We discuss applications for pedagogical research, teacher formative assessment, and teacher professional development.

CCS Concepts
• Social and professional topics~K-12 education • Computing methodologies~Discourse, dialogue and pragmatics • Computing methodologies~Supervised learning by classification • Information systems~Information extraction

Keywords
Automatic Speech Recognition; Natural Language Processing; Classroom Analytics; Question Detection

LAK '17, March 13-17, 2017, Vancouver, BC, Canada
 ISBN 978-1-4503-4870-6/17/03…$15.00
DOI: http://dx.doi.org/10.1145/3027385.3027417

1. INTRODUCTION
Teachers employ a variety of pedagogical practices in their classrooms. Instructional activities may include lectures, asking questions and evaluating student responses, or assigning students individualized seatwork or group work. A growing body of research indicates that certain activities, such as asking particular types of questions or engaging in classroom-wide discussion, predicts increased levels of student engagement and achievement gains net of socio-demographics [4, 20, 42]. Furthermore, providing teachers with training [18] and data-driven analysis [23] about their use of such instructional strategies can have positive downstream effects on student achievement.

But just how do we generate this feedback to share with teachers? Currently, efforts to assess classroom practices rely on observations by trained human judges [19]. For example, the Nystrand and Gamoran coding scheme [14, 30] provides a general template for documenting teachers' activities and can be used to analyze their instructional strategies. Unfortunately, this is an expensive and time-consuming practice that cannot be deployed uniformly at scale.

In order to facilitate wide-scale analysis of teachers' practices, computational methods that can automatically analyze classroom instruction are needed. We take a step in this direction by automatically detecting teacher questions in live classrooms. We focus on questions because they are a central component of dialogic instruction, often serving as a catalyst for in-depth classroom discussions and so called 'dialogic spells', characterized by student questions that spawn reflection, debate, and deviation from pre-scripted lesson plans [31]. We acknowledge that all questions are not created alike. The so called authentic questions (questions without prescripted answers) and questions with uptake (follow-up on the respondent's answer) are much more highly predictive of achievement compared to test questions, where the answers are known apriori [30, 31]. Nevertheless, we focus on detecting all teacher questions in this early stage of work with an eye for categorizing among different types of questions in future work (as we have begun to do from text transcripts [36]).

Classrooms provide a unique set of challenges for automatic detection of questions. Questions in the classrooms often vary from traditional information-seeking questions in other contexts, such as office meetings or conversational speech. For instance, teachers ask different types of questions, such as managerial questions (e.g., "*who hasn't finished the assignment yet?*"), rhetorical questions

(e.g., "*that is indeed relevant, isn't?*"), closed questions ("*what is the capital of Pakistan?*"), and open questions ("*why do you think the author makes this argument?*") [7]. Some of these questions, such as attendance-taking or teacher test questions, often do not exhibit the prosodic rises in inflections typically associated with question-asking and may instead sound like declarative statements when taken out of context.

Furthermore, a system designed to automatically analyze teachers' practices must fulfill a number of practical constraints [10]. For one, it cannot be disruptive to either teacher or students. Secondly, the approach must be cost-effective and easy to setup in order to enable widespread adoption. Additionally, for privacy concerns, video recordings are not possible unless students can be de-identified.

We attempt to overcome these challenges by designing a system that includes a low cost, wireless headset microphone to record teachers as they freely move about the classroom. Our system accommodates various seating arrangements, classroom sizes, and room layouts, and attempts to mitigate complications due to ambient classroom noise, muffled speech, or classroom interruptions [5], factors that reflect the reality of real-world environments.

2. RELATED WORK

There has been a large-body of work on detecting questions from text-based corpora spanning several decades [1]. However, there has been comparably little research on question detection from audio recordings. A few key studies are discussed below.

Boakye et al. [8] examined the ability of machine learning models to automatically detect questions from the ICSI Meeting Recorder Dialog Act (MRDA) corpus [40], a dataset of 75 hour-long meetings recorded with headset and lapel microphones. Using the SRI CTS automatic speech recognition (ASR) engine, they achieved a word error rate (WER), a measure of edit distance comparing the ASR hypothesis to the original transcript, of 0.38 on the corpus. They then trained an AdaBoost classifier to detect questions from the corpus using word, natural language, and parse tree features derived from the ASR transcriptions, achieving F_1 scores of 0.52, 0.35, and 0.50, respectively. Adding contextual and acoustic features improved the F_1 score to 54.0. This modest improvement suggests that linguistic information may be more important for question detection compared to contextual or acoustic information.

Stolcke et al. [41] presented a statistical approach for modeling dialogue acts in conversational telephone speech. The authors used a hidden Markov model on the Switchboard corpus [15] to identify speech acts, such as questions, statements, or apologies, achieving an accuracy of 65% on ASR transcriptions (WER 0.41) and 71% based on human transcriptions (chance level 35%; human agreement 84%).

The authors also attempted to distinguish questions from statements, two dialogic components often confused by their model. The authors improved their accuracy to 75% on a subset of their dataset containing equal proportions of questions and statements. This result, based on a balanced dataset limited to a small number of pre-specified topics and segmented by human observers, achieved only modest accuracy, highlighting the difficulty of question detection in real world environments.

More recently, Orosanu and Jouvet [33] used ASR transcriptions to train models to differentiate between statements and questions using three French language corpora consisting of 7,005 statements and 831 questions. They classified 72.6% of questions and 77.7% of statements correctly using linguistic and prosodic features derived from human transcriptions. However, when they substituted ASR transcriptions in two corpora (WERs of 0.22 and 0.28), they found a 3% reduction in classification accuracy. Additionally, the F_1 score was only 0.40 for questions compared to a weighted F_1 score of 0.63, demonstrating greater difficulty in identifying questions compared to statements. The authors found that models trained on linguistic features significantly outperformed those trained with prosodic features, and the combination of both provided only trivial improvements. Furthermore, the models trained on ASR transcriptions of unscripted, spontaneous speech from the third corpus was slightly less accurate (accuracy of 70%) compared to a corpus of scripted dialogue from radio interviews (73%) (chance=50%). This underscores the potential added difficulty of automatically detecting questions based on spontaneous speech, as in our study.

Additionally, Orosanu and Jouvet examined classification using imperfect sentence boundaries. The authors manipulated the boundaries of the human segmented utterances based on the longest silence preceding or following an utterance. Following this perturbation of the manual segmentation, they observed a 3% drop in accuracy. This reveals an additional difficulty in question detection based on imperfectly segmented utterances, a challenge we also confront in this work.

Other studies have attempted to identify questions using prosodic information, often in tonal languages such as Chinese [45] or Vietnamese [34, 44]. Another study used prosodic features and a decision tree to detect questions from a dataset of Arabic language audio lectures containing an equal number of questions and statements [21]. Although the authors reported 76% accuracy (chance = 50%), the dataset consisted of only three speakers and the results were not validated independent of the speaker. Nevertheless, the authors found that the fundamental frequency and the energy level were the most useful features, features we also consider in this work.

One English language study combined acoustic, lexical, and syntactic features to identify questions from Wikipedia talk pages [24]. The authors noted that models using lexical and prosodic features (AU-ROC 0.92) only slightly outperformed models using lexical features alone (AU-ROC 0.91). This study benefited from perfect transcripts and their corpus did not contain properties of spontaneous speech (e.g., backchannels, disfluencies, or interruptions), as in our study.

In preliminary work, we explored question detection from automatically-segmented utterances derived from live-recordings of classroom audio. Using leave-one-speaker-out cross-validation, we achieved an overall weighted F_1 score of 0.69 using only lexical and syntactic features [28] demonstrating that question detection was possible from noisy classroom audio. Presently, we expand upon this work to explore the potential of acoustic and contextual features in conjunction with linguistic features.

3. CONTRIBUTIONS AND NOVELTY

We describe an approach to automatically identify teacher questions solely from an audio recording of the teacher in a real-world classroom. Given the noisy environment, we face a number of technical challenges. Classroom speech is particularly noisy as there are disruptions, accidental microphone contact, sounds of students shuffling papers or moving desks, alarms, background media, and so on. Classroom speech is also more informal and conversational compared to more formal settings such as meetings. Furthermore, we must automatically segment utterances from the

audio stream, an analysis itself that is prone to error. Lastly, ASR transcription is imperfect, thereby adding additional noise.

We make several contributions while addressing these challenges. First, we examine a dataset of full length recordings of real world class sessions, drawn from multiple teachers and schools. Second, we only use teacher audio because it is the most practical option given privacy and scalability concerns. Third, we automatically segment audio recordings into individual teacher utterances in a fully automated pipeline. Fourth, we combine multiple ASR engines at the feature level to ameliorate errors. Fifth, we consider high-level linguistic features, acoustic and prosodic features, and contextual features, all derived from the audio stream. Finally, we design our models to generalize across teachers rather than optimizing to individual teachers.

4. METHODS

We begin by reviewing our data collection procedure (Section 4.1), followed by discussion of our approach to automatically segment the audio streams (Section 4.2). Next, we describe the process used to manually label the detected speech as containing a question or not (Section 4.3) and transcribe the audio using ASR (Section 4.4). Finally, we discuss our feature sets (Section 4.5) and machine learning approach (Section 4.6).

4.1 Data Collection

A dataset of audio recording was collected at six rural Wisconsin middle schools during literature, language arts, and civics classes taught by 11 different teachers (three male; eight female). A total of 37 class sessions were recorded over 17 separate days in the course of a year. The length of each class session varied depending on the school, lasting between 30 and 90 minutes. The dataset contains a total of 32 hours and five minutes of audio.

In order to capture unbiased samples of discourse in real-world classrooms, each teacher was asked to carry out their normal lesson plan. The teachers were recorded using a wireless microphone, which allowed them to move about the classroom freely. A Samson 77 Airline wireless microphone (Figure 1) was chosen based on previous work [10] and for its noise-canceling abilities and relatively low-cost ($300 in 2014). The recording of the teacher's speech was saved as a 16 kHz, 16-bit single channel audio file (Figure 2).

Figure 1. Samson 77 Airline Microphone

Each class session was live coded by observers trained in the Nystrand and Gamoran coding scheme (see below), using specialized software developed for this task (Figure 3) [25]. Coders were also trained on proper recording techniques to ensure consistent levels of quality in the recordings. Following the class session, the coded annotations were reviewed and refined by the original coder until reaching full agreement with a second coder.

The Nystrand and Gamoran coding scheme [14, 30] tracks classroom activity across the following (in order of increasingly fine granularity) three parallel levels: (1) episodes, which denote

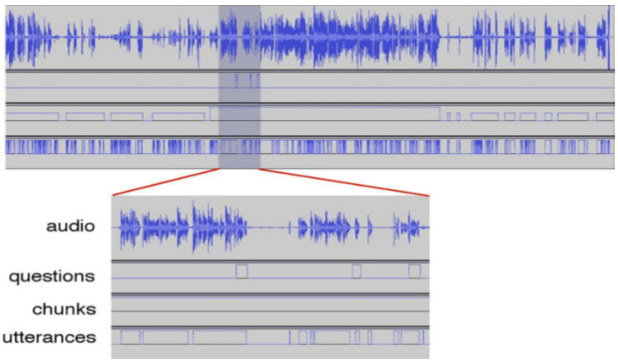

Figure 2. Excerpt of classroom recording marked with utterances, instructional segments, and dialogic questions

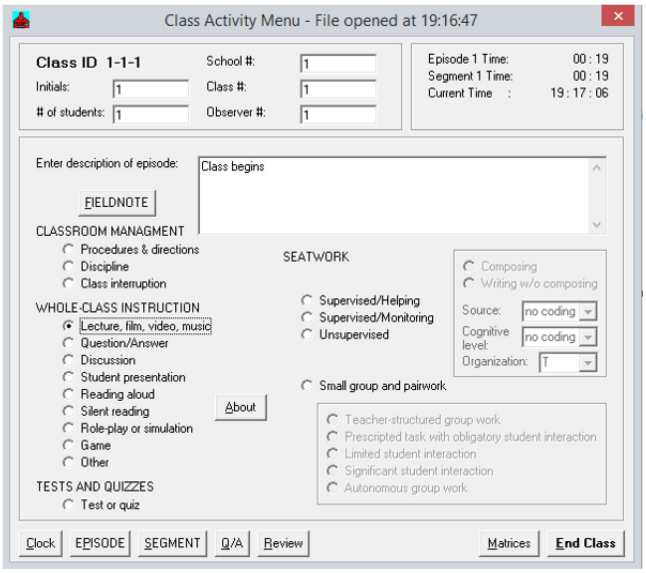

Figure 3. Screenshot of the Class 4 program

the current activity/topic; (2) instructional segments, 17 categories that represent possible classroom activities (e.g., Lecture, Group Work, Discussion) used to implement an episode; and (3) certain questions (e.g., non-procedural, non-rhetorical) asked by teachers and students [30]. We focus on questions asked by the teacher here.

4.2 Teacher Speech Extraction

Recordings of the teacher's speech was segmented into utterances using a voice activity detection (VAD) technique **[5]**. First, a low-pass filter was applied to the recordings. Next, the amplitude envelope of the signal was examined in non-overlapping 20-millisecond windows. Whenever the amplitude of the signal exceeded a preset threshold, it was assumed that the teacher was speaking, otherwise, silence was assumed. Consecutive windows of teacher speech were considered part of a potential teacher utterance until silence was detected continuously for 1 second.

We set the VAD threshold low enough to prioritize capturing all instances of the teachers' speech. However, this caused a high rate of false alarms in the form of non-speech utterances, such as classroom noise, body movements, coughing, or heavy breathing. In order to filter out these false alarms, we processed the potential utterances with the Bing ASR system [27]. We considered any potential utterances rejected by the ASR as non-speech.

Additionally, utterances shorter than 125 milliseconds were removed as they were unlikely to contain meaningful speech.

We evaluated the effectiveness of this utterance detection approach in prior work [10]. Briefly, we extracted a random subset of 1,000 potential utterances and manually coded them as containing speech or not speech. We observed high levels of both precision (96.3%) and recall (98.6%) and an F_1 score of 0.97, which we deemed sufficient for our purpose. We extracted a total of 10,080 utterances from the 37 classroom recordings. The average utterance length was 5.01 seconds with a standard deviation of 7.64 seconds.

4.3 Question Coding

Many of the previous studies in automatic question detection used datasets that contained questions and statements that had been segmented manually by humans. However, in our study we rely on automatic, and thus imperfect, segmentation. We also used fixed amplitude envelope and silence thresholds while segmenting utterances as opposed to learning teacher-specific thresholds in order to increase generalizability to new teachers. A side-effect of this procedure is that each utterance may contain multiple questions, or conversely, a question may be spread across multiple utterances [22].

To address this concern, we manually coded the 10,080 extracted utterances as either "containing a question" or "not containing a question" rather than "question" or "statement." This distinction, although subtle, addressed the cases in which a question phrase was embedded within a longer utterance, coding these as "containing a question." Similarly, we also handled cases in which a question phrase spans one or more consecutive utterances, also coding these utterances as "containing a question."

We define a "question" following a coding scheme that was specifically designed to analyze questions in classroom discourse [31]. Here, questions are defined as utterances in which the teacher solicits information from a student either procedurally (e.g., "*Is everyone ready?*"), rhetorically (e.g., "*Oh good idea James, why don't we just have recess instead of class today*"), or for knowledge assessment/information solicitation purposes (e.g., "*What is the capital of Indiana, Michael?*"). Likewise, the teacher calling on a different student to answer the same question (e.g., "*Nope. Shelby?*") would also be considered a question. In some coding schemes, the previous example would be classified as "turn eliciting" questions [3]. Cases in which the teacher reads from a novel in which a character asked a question or calls on a student for other reasons (e.g., such as to discipline them) would not be considered questions.

The coders were seven research assistants and researchers whose native language was English and who had no experience with the Nystrand and Gamoran [25] coding scheme. The coders first engaged in a training task by labeling a common evaluation set of 100 utterances. These 100 utterances were manually selected to exemplify questions that were difficult to identify. Once coding of the evaluation set was completed, the expert coder who initially selected and coded the example utterances reviewed the codes for any discrepancies. Coders were required to achieve a minimal level of agreement with the expert coder (Cohen's kappa, $\kappa = 0.80$). If the agreement was lower than 0.80, mistakes were identified and explained to the coders.

After this training task was completed, each coder coded a subset of utterances from the complete dataset across multiple sessions. Coders listened to the utterances in temporal order and assigned a label to each, based on the words spoken by the teacher, the teachers' tone (e.g., prosody, inflection), and the context of the previous utterance. Coders could also flag an utterance for review by the expert, although this occurred only rarely.

Of the 10,080 utterances, 3,584 were labeled as "containing a question" (36%) and 6,496 as "not containing a question" (64%), across teachers. To ensure reliability, a random subset of 117 utterances from the full dataset were selected and coded by the expert coder, which resulted in high agreement (Kappa $\kappa = 0.85$).

4.4 Automatic Speech Recognition

In previous work [6, 28] we examined three different automatic speech recognition (ASR) systems for this task: Bing [27], AT&T Watson [16] and the Azure [26]. Both the Bing and AT&T ASR systems transcribed individual utterances segmented as described in Section 4.2. The Azure system, however, processed full-length classroom recording to produce a set of time-stamped words, from which we reconstructed the individual utterances.

We evaluated performance of the three ASR systems on a subset of 1,000 utterances chosen randomly without replacement, considering two metrics commonly used in speech recognition: word error rate (WER) and simple word overlap (SWO) [28]. WER accounts for word order between ASR and human transcripts and was computed by summing the number of substitutions, deletions, and insertions required to transform the human transcript into the ASR transcript, divided by the total number of words in the human transcript. SWO, however, does not account for word order and was computed by dividing the number of words that appear in both the human and ASR transcripts by the total number of words in the human transcript.

In Table 1 we show the WER and SWO for the three different ASR systems. We note that all three systems achieved moderate accuracy, despite the complexity of the task of automatically transcribing noisy conversational speech recorded in a real-world environment.

Table 1. ASR word error rate (WER) and simple word overlap (SWO) averaged by teacher for 1,000 utterances, with standard deviations shown in parentheses

ASR	WER	SWO
Bing Speech	0.45 (0.10)	0.55 (0.06)
AT&T Watson	0.63 (0.11)	0.42 (0.11)
Azure Speech	0.49 (0.07)	0.64 (0.16)

The Azure ASR engine failed to return transcriptions for 201 of these 1,000 utterances. This resulted in a WER of 1.0 and SWO of 0.0 for those utterances. Discarding those instances improves WER for Azure to 0.37 (SD = 0.07) and SWO to 0.68 (0.06). Thus Azure, when able to transcribe an utterance, is the best performing ASR. However, this failure to return a transcription for 20% of utterances requires that we consider it only in conjunction with another ASR engine.

We also compared ASR transcriptions with each other across our full dataset of 10,080 utterances. The results in terms of SWO were as follows: Bing versus AT&T (0.43), Bing versus Azure (0.51), and Azure versus AT&T (0.48). On average the ASR engines only agree on half of the words in each transcript. Therefore, the combined use of multiple ASR systems could provide additional information, as the strengths and weaknesses of individual ASRs may vary across teachers, class sessions, and instruction types. We combined them at the feature-level as discussed in the next section.

4.5 Features

We extracted 218 linguistic, acoustic, and contextual features to train our classification models.

Linguistic Features. We considered a set of natural language features generated from ASR transcripts for each utterance obtained from Bing Speech, AT&T Watson, and Azure Speech engines. The majority of these features (n=34) were obtained by processing each utterance with the Brill Tagger [9] and analyzing each token with a question type classifier [32]. The question type classifier, developed for speech act classification in educational systems, used a cascaded finite state transducer to tag utterances according to a taxonomy of potential question types, as well as part of speech tags. Features included the presence of particular question words (e.g., *what, why, how*), simple disambiguation rules (e.g., the presence of words that start with *wh-*), part of speech tags (e.g., presence of nouns, presence of adjectives), and hypothesized categories based on simple keywords (e.g., definition, comparison, procedural). We used these features in prior work to detect domain-independent question properties from human-transcribed questions [36] and for automatic classification of teacher questions [6, 28]. We included three additional features: proper nouns (e.g., student names), pronouns associated with uptake (teacher questions that incorporate student responses), and pronouns not associated with uptake, as recommended by a domain expert on teacher questions.

In all, we extracted 37 binary linguistic (NLP) features for each ASR's transcription. Next, we computed a combined NLP feature set by taking the mean of the individual ASR binary features. For example, if the *what* feature was detected by Bing and AT&T, but not Azure, then it would take on a value of 0.67. All subsequent analyses focus on this combined feature set as it has been shown to be superior to the individual feature sets in our prior work [28]

Acoustic Features. We extracted prosodic, spectral, and voice quality features with the OpenSmile audio feature extraction tool [13] using the feature set from the Interspeech Emotion Challenge [38]. The low-level audio descriptors were: zero-crossing-rate (ZCR) from the time signal, root mean square (RMS) frame energy, Mel-frequency cepstral coefficients (MFCC) 1-12, fundamental frequency computed from the cepstrum (normalized to 500 Hz), and a voicing probability computed from the autocorrelation of the power spectrum. For each feature, 12 statistical functionals were computed: mean, standard deviation, kurtosis, skewness, minimum and maximum value, relative position of the minimum and maximum values (in frames), range, two linear regression coefficients (slope, offset), as well as the linear regression mean square error (MSE). Additionally, for each feature, we considered the smoothed moving averaged window (length 3) and the 1st order delta coefficient (differential) of the smoothed low-level descriptor. This results in $16 \cdot 2 \cdot 12 = 384$ features for each utterance. Given the large number of acoustic features compared to our other feature sets, we used tolerance analysis to eliminate features with high multicollinearity (variance inflation factor > 5) [2], after which, 168 acoustic features remained.

Context Features. We considered a number of context features for each utterance. These included: the length of utterance, lengths of the previous and subsequent utterances, the duration of the pause preceding and following the utterance, the position of the utterance within the class session normalized to (0,1), verbosity (the number of words from the Bing ASR transcript), and the rate of speech (number of words in the Bing ASR transcript divided by the length of the utterance in seconds).

We also considered the likelihood that each utterance occurred during one of the instructional segments from the Nystrand and Gamoran coding scheme [30]. We considered the five most commonly occurring instructional segments in the dataset: Question & Answer, Procedures & Directions, Group Work, Supervised Seatwork, and Lecture. In prior work, we predicted the occurrence of the segments by training supervised machine learning models using timing, acoustic, and linguistic features, resulting in F_1 scores ranging from 0.64 to 0.78 [12]. For each utterance, the probability it is contained within each of the instructional segments was considered, resulting in five additional context features. In total, we extracted 13 context features.

4.6 Classification Models

We trained and tested supervised classification models to predict if an utterance contained a partial or complete question (question), or did not contain a question at all (non-question). The model building process involved the following steps.

Feature Standardization. We z-scored standardized all 218 features, resulting in a mean of 0.0 and a standard deviation of 1.0 for each feature. Features were standardized within each teacher.

Classification Models. We explored a number of common machine learning classifiers using implementations from the WEKA toolkit [43]: Naïve Bayes, logistic regression, random forest, J48 decision tree, J48 with Bagging, Bayesian network, k-nearest neighbor (k = 7, 9, and 11). We also combined the classifiers with MetaCost [11], which penalized misclassifications of the minority class (weights of 2 and 4).

Validation. We validated the classification models with leave-one-teacher-out cross-validation. Each model was built on data from 10 teachers (the training set) and validated on the held-out teacher (the testing set). The process was repeated for 11 folds so that each teacher appeared in the testing set once and the results were calculated from a confusion matrix aggregated across teachers. This cross-validation technique tests the potential of our models to generalize to new teachers despite variations in speaking patterns, word-choice, and rate of question-asking.

5. RESULTS

5.1 Classification Accuracy

The best performing model was the J48 decision tree with bagging and MetaCost (miss weight of 2) and used all features. This model achieved the highest F_1 score for the classification of utterances containing questions and was consistent with respect to precision (0.69) and recall (0.70). We selected the best performing model based on the F_1 score for questions to prioritize the model's ability to detect the class of interest, which was always the minority, rather than prioritizing the dominant class label (i.e. non-questions). Figure 4 shows the results of the J48 decision tree with bagging and MetaCost for each set of features: linguistic (NLP; n = 37), acoustic (n = 168), and contextual (n = 13), as well as the combination of all features (n = 218). The remainder of the results given in this work use this classifier.

With respect to individual feature sets, we found that linguistic features outperformed acoustic and context features in the classification of both questions and non-questions, a result that is consistent with previous studies in question detection from automatic transcriptions from audio [8, 24, 33]. However, we note that while acoustic-prosodic features were not particularly successful in identifying questions (0.36), they were significantly more successful in identifying non-questions (0.67). This suggests that acoustic-prosodic features may be useful in identifying other types of statements compared to questions themselves, albeit less useful than the linguistic features alone.

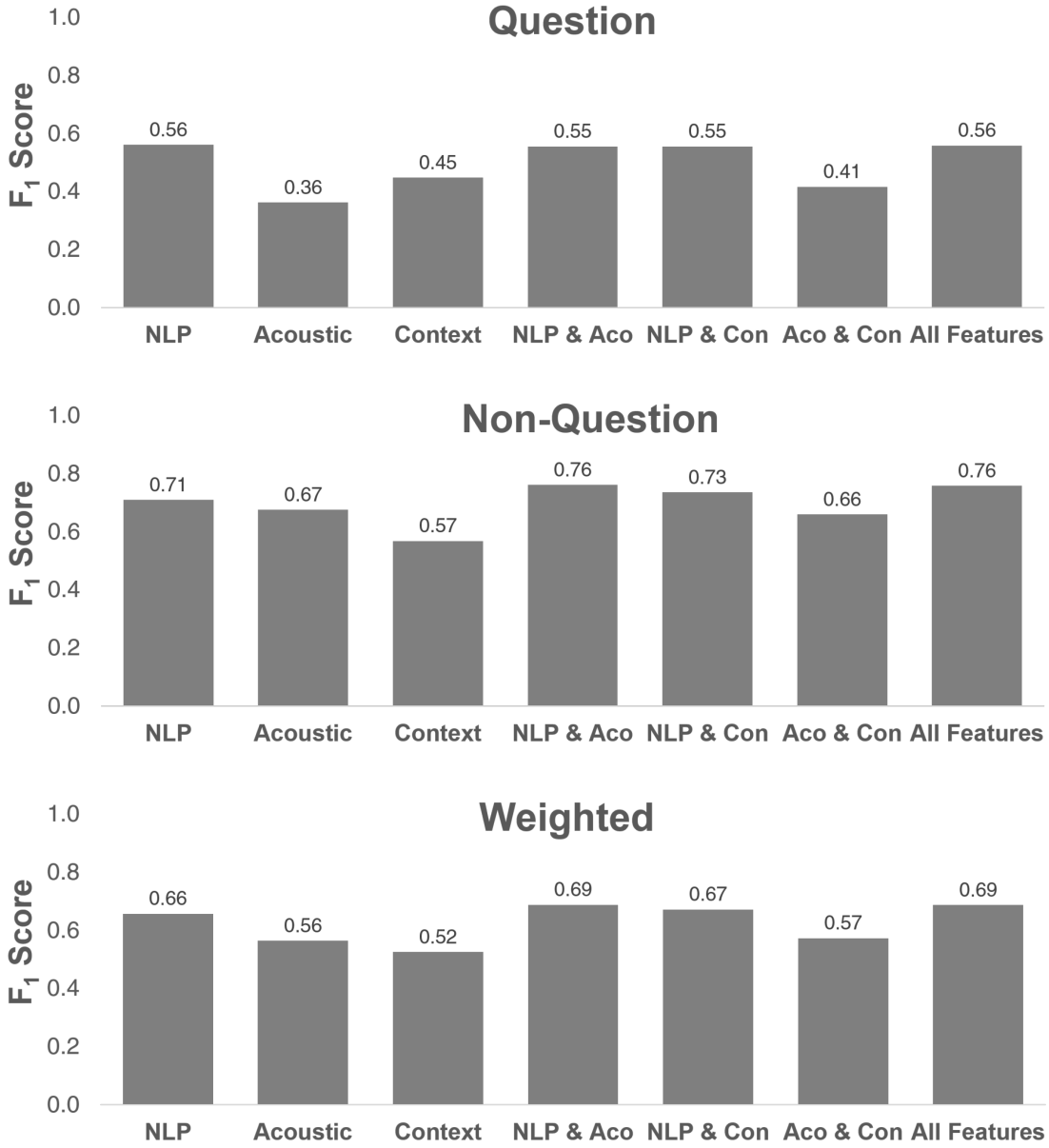

Figure 4. Results for question detection comparing linguistic (NLP), acoustic (Aco), and context (Con) features, pairwise combinations of feature sets, and the combination of all features

Table 2. Confusion matrix showing agreement and disagreement between the three feature sets: linguistic (NLP), acoustic (Aco), and context (Con) in classifying questions (Q) and non-questions (NQ)

	Predicted							
	Agreement		**Disagreement**					
	NLP Q	NLP NQ	NLP Q	NLP Q	NLP Q	NLP NQ	NLP NQ	NLP NQ
	Aco Q	Aco NQ	Aco Q	Aco NQ	Aco NQ	Aco NQ	Aco Q	Aco Q
Actual	Con Q	Con NQ	Con NQ	Con Q	Con NQ	Con Q	Con NQ	Con Q
Q	0.64	0.36	0.37	0.35	0.40	0.32	0.35	0.35
NQ	0.36	0.64	0.63	0.65	0.60	0.68	0.65	0.65

For the question class, the combined model demonstrated no improvement compared to using only linguistic features (both F_1 = 0.56). However, we found that the additional features did result in improvement for the classification of non-questions (0.76 vs. 0.71), a 7% improvement. This resulted in a 5% improvement to the overall weighted F_1 score (0.69 vs. 0.66). Table 2 provides a confusion matrix that shows agreement and disagreement between the three feature sets. In general, there was no clear pattern and the results were similar to the base rates in the dataset (see Section 4.2).

Our best performing classification model (J48 with bagging and MetaCost) returns a confidence rate with each prediction, allowing comparison of the confidence of predictions between feature sets (Pearson's r): acoustic vs. linguistic (0.15), context vs. linguistic (0.22), and acoustic vs. context (0.30), indicating a lack of consensus between the individual feature set models. Unsurprisingly, given our findings on the importance of the linguistic features, the prediction confidence of the linguistic model and the combined model was strongly correlated (r =.80) compared to acoustic (r = .25) and context (r = .27).

Additionally, we compared pairwise combinations of the three feature sets, shown in Figure 4. We found that combining linguistic features with acoustic (0.55) or context (0.55) features resulted in no improvement over linguistic features alone (0.56) for the detection of questions. However, the combination of acoustic (0.76) or context (0.73) features with linguistic features did yield a 7% and 3% improvement, respectively, over linguistic features alone (0.71) for the detection of non-questions. When taken together, these results indicate that a combination of acoustic and linguistic features may be sufficient for this task, with the context features not contributing much more.

5.2 Feature Analysis

Motivated by the observation that linguistic features were the most useful in identifying questions, we analyzed the diagnosticity of each feature individually. We hypothesized that linguistic features documenting the use of certain question words (e.g., *what, why, how*) will be the most useful to discriminate between questions and non-questions. We re-ran our models by considering only a single feature, one at a time, using a J48 decision tree with bagging and Meta-Cost. In Figure 5 we show the F_1 scores for each individual feature ranked by the overall weighted F_1 score. For completeness, we focused on all 218 features rather than the NLP features alone.

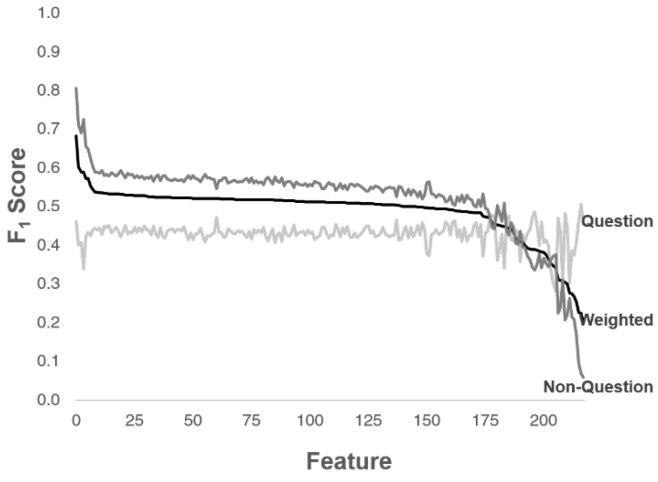

Figure 5. Analysis of individual features ranked by overall weighted F_1 score

Of the top 100 features, 91 were acoustic features, eight were linguistic features, and only one was a context feature (the length of the utterance; ranked 13th). The dominance of acoustic features is unsurprising as they constitute 79% of the features. However, the top seven features were all linguistic, perhaps explaining the success of models trained with linguistic features alone compared to models using acoustic or context features. We show the top 10 features sorted by their overall weighted F_1 scores in Table 3.

The top ranked feature was the presence or absence of the question word "*what*" in the ASR transcript. We note that this feature alone achieved an overall F_1 score of 0.68, rivaling the model's performance using all features (0.69). However, this feature only achieved a F_1 of 0.46 for the question class compared to 0.56 using all features. An analysis of the ASR transcriptions showed that the "*what"* feature appeared in 5.9% of non-questions and 12.8% questions, indicating this one word alone was insufficient to distinguish questions from non-questions. Contrary to our expectations, we found that no other question-word features were as successful: *how* (ranked 198th), *why* (ranked 203rd), *wh-* (ranked 207th), *should* (ranked 218th). We also note that no single feature model exceeded an F_1 of 0.50 for question detection, implying that the combination of features was needed.

Table 3. Top ten features ranked by overall F_1 score

Type	Feature	F_1
NLP	Presence of word "*what*"	0.68
NLP	Presence of phrase "*do…have*"	0.60
NLP	Presence of a pronoun	0.59
NLP	Presence of word "*be*"	0.59
NLP	Presence of a proper noun	0.57
NLP	Pronouns associated with uptake	0.57
NLP	Presence of a verb	0.55
Acoustic	MFCC [1] - maximum value	0.54
Acoustic	Voice probability - maximum position	0.54
Acoustic	MFCC [11] - linear regression MSE	0.54

5.3 Analysis by Class Session

The models were trained using leave-one-teacher-out cross-validation, but we performed additional post-hoc analyses exploring the model's accuracy across the 37 individual class sessions. These analyses allow an investigation of the stability of our models for individual class sessions, which will be essential for generalizability to future class sessions with different topics.

In Figure 6 we show histograms of F_1 scores for questions, non-questions, and the overall weighted average by individual class session. We note that model accuracy was distributed across class sessions, rather than a bimodal distribution of successes and failures. The model yielded an interquartile range 0.44 to 0.65 for the question class. In the classification of questions, we observed that the model had greater difficulty with some class sessions, and that this was most often associated with classes that contained relatively fewer questions. For example, the best performing class session (0.76) contained 57% questions while the poorest-performing class session (0.19) contained only 29% questions. Over the 37 class sessions, the rate of questions and the F_1 scores for questions was strongly correlated (Pearson's r = 0.76). This demonstrates that our model had greater difficulty in identifying questions for class sessions that contained a low proportion of

questions to non-questions. detection of questions might have important consequences for both research on effective instructional strategies and on teacher professional development. Thus, our current work centers on a fully-automated process for predicting teacher questions in a noisy real-world classroom environment, using only a full-length audio recording of teacher speech.

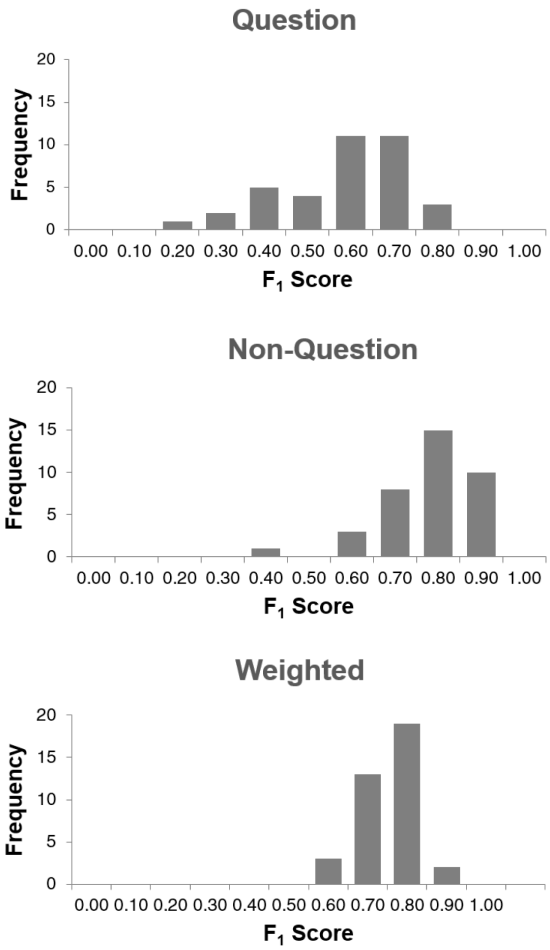

Figure 6. Histogram of F_1 scores by class session

6. DISCUSSION

The importance of teacher questions in classroom discourse is widely acknowledged in both policy (e.g., [39]) and research [4, 29, 31]. Teacher questions play a central role in promoting student engagement and achievement, suggesting that automating the

6.1 Summary of Contributions

We present encouraging results with our automated processes, consisting of voice activity detection to automatically segment teacher speech, combining three different ASR transcriptions, three different features sets (linguistic, acoustic-prosodic, and context), and machine learning with teacher-independent validation.

A key contribution of our work over previous research is that our models were trained and tested on automatically, and thus imperfectly, segmented utterances. This builds upon the work of Orosanu and Jouvet [33] which artificially explored perturbations of a subset of utterance boundaries using automatic detection of silence within human-segmented spoken sentences. We note that despite automatic segmentation, which may split individual questions across utterances, we outperformed the previous work (weighted F_1 scores 0.69 vs. 0.63). To our knowledge, our work is the first to detect spoken questions using a fully automated process.

Our best performing model using all features achieved an F_1 score of 0.56 for the question class and an overall weighted F_1 score of 0.69. Furthermore, we demonstrated that models built using only linguistic features outperformed those built using either acoustic or context features, consistent with similar findings in the literature [8, 24, 33]. Although models built using a combination of features did not improve the identification of questions, they were more successful at detecting non-questions. Here, also linguistics and paralinguistics seemed to suffice; contextual features had little more to add – at least for the small set investigated here.

Additionally, we investigated the utility of each feature. We found that the top seven individual features were all linguistic features, underscoring the importance of the specific content of the spoken utterances compared to paralinguistic or contextual clues for identifying questions.

We validated our models using leave-one-teacher-out cross-validation, demonstrating generalizability of our approach across teachers in this dataset. Furthermore, we analyzed model performance by class session, finding that our model was consistent across class sessions, an encouraging result supporting our goals of class session-independent question detection.

6.2 Limitations and Future Work

This study is not without limitations. We were also unable to record individual students for practical reasons and extraction of student speech was not feasible from the teacher's microphone. This precluded us a potentially key feature that signals a question – the student response. Fortunately, additional data collection includes a second microphone that captures general classroom activity. This second channel of audio when combined with the recording of the teacher, will afford modeling patterns of teacher-student interactions, potentially revealing question-response patterns between teachers and students.

We designed our approach to avoid overfitting to specific classes, teachers, or schools. However, all of our recordings were collected in Wisconsin, a state that has adopted the common core standard [39]. It is possible that the common core may impose aspects of a particular style of teaching that our models may overfit to. Similarly, although we used speaker-independent ASR and teacher-independent validation techniques to improve generalizability to new teachers, our sample of teachers were from a single region with traditional Midwestern accents and dialects. Therefore, broader generalizability across the U.S. and beyond remains to be seen [17]. Finally, our models are likely English language specific. However, because we limited the linguistic features to high-level part of speech features, it may be possible to adapt these features to other languages that share similar linguistic structures. Our finding that the most useful features are linguistic features is encouraging as these features could be readily tagged in many languages.

We acknowledge that our method for teacher utterance segmentation may potentially be improved using proposed techniques in related works. For example, Komatani et al. [22] explored detecting and merging utterances segmented mid-sentence, allowing analysis to take place on a full sentence, rather than a fragment, which may improve detection of questions split across utterances. An alternative approach would be to automatically detect sentence boundaries within utterances, and extract features from each detected sentence. Furthermore, Raghu et al. [35] explored using context to identify non-sentential utterances (NSUs), defined as utterances that are not full sentences

but convey complete meaning in context. Identification of NSUs may improve our model's ability to differentiate between difficult cases (e.g., calling on students, saying a student's name to discipline them).

In this work, we compared the performance of three different features sets (linguistic, acoustic-prosodic, and context). While this approach allowed us to compare the utility of individual feature sets, the ideal set of features may derive from a subset drawn from different feature types. More sophisticated fusion methods in lieu of the simple feature level fusion explored here might also be needed. In future work, we will examine empirical feature selection as well as explore decision- and model-based fusion techniques that combine the three feature sets. We will also explore temporal models, such as hidden Markov models and conditional random fields, that might better capture questions in the larger context of the classroom dialogue. Such a temporal analysis may help find sequences of consecutive questions, such as those present in question and answer sessions or in classroom discussions.

Lastly, informed by our observation of the utility of linguistic features, and more specifically those that capture certain question words, we will explore additional linguistic feature to identify additional words important to the detection of questions. Because the topics varied between individual class sessions, a traditional bag-of-word analysis may not be useful since the course material is not likely to repeat between teachers and sessions. However, this approach may yield insight into additional key words which could be aggregated into high-level features that may be useful to detect questions, similar to our binary question word features.

Finally, as noted in the Introduction, it is not merely the amount of questions asked but the types of questions that correlate with achievement. Thus, future work will focus on classifying question properties defined by Nystrand and Gameron [25], such as authenticity, uptake, and cognitive level. We have explored these properties in previous work [36, 37] using perfectly segmented and human transcribed text. We will continue this work using our approach that employs automatic segmentation, ASR transcriptions, and question detection.

6.3 Applications

The ability to identify questions asked by teachers in the classroom is necessary in order to generate personalized formative feedback for the teacher about their use of class time. Our approach would permit automating such analysis, enabling a cost-effective scalable deployment that would be accessible to many schools and teachers. Using our system, teachers could record their class and receive automated feedback following the class session. Such feedback will afford teachers reflection on their teaching style and better enable collaboration with professional development personnel towards improvement of their pedagogy with the ultimate goal of increasing student engagement and achievement. It will also facilitate research into effective pedagogy by providing educational researchers with an automated approach to collect and code classroom discourse.

6.4 Concluding Remarks

We took steps towards fully-automated detection of teacher questions in noisy real-world classroom environments using linguistic, acoustic-prosodic, and context features. The present contribution is one component of a broader effort to automate the collection and coding of classroom discourse.

7. ACKNOWLEDGMENTS

We thank Marci Glaus, Xiaoyi Sun, and Brooke Ward of the University of Wisconsin-Madison for their help with the collection and annotation of the classroom data. 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.

8. REFERENCES

[1] Aichroth, P., Björklund, J., Stegmaier, F., Kurz, T. and Miller, G. 2015. State of the art in cross-media analysis, metadata publishing, querying and recommendations. *Media in Context (MICO)*. 1, (2015).

[2] Allison, P.D. 1999. *Multiple regression: A primer*. Pine Forge Press.

[3] Allwood, J., Cerrato, L., Jokinen, K., Navarretta, C. and Paggio, P. 2007. The MUMIN coding scheme for the annotation of feedback, turn management and sequencing phenomena. *Language Resources and Evaluation*. 41, 3-4 (2007), 273–287.

[4] Applebee, A.N., Langer, J.A., Nystrand, M. and Gamoran, A. 2003. 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.

[5] Blanchard, N., Brady, M., Olney, A.M., Glaus, M., Sun, X., Nystrand, M., Samei, B., Kelly, S. and D'Mello, S. 2015. A study of automatic speech recognition in noisy classroom environments for automated dialog analysis. *Artificial Intelligence in Education* (2015), 23–33.

[6] Blanchard, N., Donnelly, Patrick J, Olney, A.M., Samei, B., Ward, B., Sun, X., Kelly, S., Nystrand, M. and D'Mello, S.K. 2016. Automatic detection of teacher questions from audio in live classrooms. *Proceedings of the 9th International Conference on Educational Data Mining (EDM 2016), International Educational Data Mining Society* (2016).

[7] Blosser, P.E. 1975. *How to ask the right questions*. National Science Teachers Association Press.

[8] Boakye, K., Favre, B. and Hakkani-Tür, D. 2009. Any questions? Automatic question detection in meetings. *Automatic Speech Recognition & Understanding, 2009. ASRU 2009. IEEE Workshop on* (2009), 485–489.

[9] Brill, E. 1992. A simple rule-based part of speech tagger. *Proceedings of the workshop on Speech and Natural Language* (1992), 112–116.

[10] D'Mello, S.K., Olney, A.M., Blanchard, N., Samei, B., Sun, X., Ward, B. and Kelly, S. 2015. Multimodal capture of teacher-student interactions for automated dialogic analysis in live classrooms. *Proceedings of the ACM on International Conference on Multimodal Interaction* (2015), 557–566.

[11] Domingos, P. 1999. Metacost: A general method for making classifiers cost-sensitive. *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining* (1999), 155–164.

[12] Donnelly, P.J., Blanchard, N., Samei, B., Olney, A.M., Sun, X., Ward, B., Kelly, S., Nystrand, M. and D'Mello, S.K. 2016. Automatic teacher modeling from live classroom audio. *Proceedings of the 24th Conference on User Modeling, Adaptation and Personalization,* 45–53.

[13] Eyben, F., Wöllmer, M. and Schuller, B. 2010. OpenSmile: the Munich versatile and fast open-source audio feature extractor. *Proceedings of the 18th ACM international conference on Multimedia* (2010), 1459–1462.

[14] Gamoran, A. and Kelly, S. 2003. Tracking, instruction, and unequal literacy in secondary school English. *Stability and*

change in American education: Structure, process, and outcomes. (2003), 109–126.

[15] Godfrey, J.J., Holliman, E.C. and McDaniel, J. 1992. SWITCHBOARD: Telephone speech corpus for research and development. *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on* (1992), 517–520.

[16] Goffin, V., Allauzen, C., Bocchieri, E., Hakkani-Tür, D., Ljolje, A., Parthasarathy, S., Rahim, M.G., Riccardi, G. and Saraclar, M. 2005. The AT&T WATSON Speech Recognizer. *ICASSP (1)* (2005), 1033–1036.

[17] Hall, J.K. 2008. Language education and culture. *Encyclopedia of language and education*. Springer. 45–55.

[18] Juzwik, M.M., Borsheim-Black, C., Caughlan, S. and Heintz, A. 2013. *Inspiring dialogue: Talking to learn in the English classroom*. Teachers College Press.

[19] Kane, T. and Staiger, D. 2012. Gathering feedback for teachers: Combining high-quality observations with student surveys and achievement gains. Research Paper. MET Project. Bill & Melinda Gates Foundation.

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

[21] Khan, O., Al-Khatib, W.G. and Lahouari, C. 2007. Detection of questions in Arabic audio monologues using prosodic features. *Multimedia, 2007. ISM 2007. Ninth IEEE International Symposium on* (2007), 29–36.

[22] Komatani, K., Hotta, N., Sato, S. and Nakano, M. 2015. User adaptive restoration for incorrectly segmented utterances in spoken dialogue systems. *16th Annual Meeting of the Special Interest Group on Discourse and Dialogue* (2015), 393.

[23] Lai, M.K. and McNaughton, S. 2013. Analysis and discussion of classroom and achievement data to raise student achievement. *Data-based decision making in education*. Springer. 23–47.

[24] Margolis, A. and Ostendorf, M. 2011. Question detection in spoken conversations using textual conversations. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2* (2011), 118–124.

[25] Martin Nystrand 2016. *Class 4.5 user's manual: A window's laptop-computer system for the in-class analysis of classroom discourse.* The Wisconsin Center for Education Research, University of Wisconsin-Madison.

[26] Microsoft 2016. *Azure Speech API*.

[27] Microsoft 2014. *The Bing Speech Recognition Control*.

[28] Blanchard, N., Donnelly, P.J., Olney, A.M., Samei, B., Sun, X., Ward, B., Kelly, S., Nystrand, M., and D'Mello, S.K. 2016. Identifying teacher questions using automatic speech recognition in live classrooms. *Proceedings of the Special Interest Group on Discourse and Dialogue (SIGDIAL)* (2016), 191–201.

[29] Nystrand, M. and Gamoran, A. 1991. Instructional discourse, student engagement, and literature achievement. *Research in the Teaching of English*. (1991), 261–290.

[30] Nystrand, M., Gamoran, A., Kachur, R. and Prendergast, C. 1997. *Opening dialogue: Understanding the dynamics of language and learning in the English classroom. Language and Literacy Series*. Teachers College Press.

[31] Nystrand, M., Wu, L.L., Gamoran, A., Zeiser, S. and Long, D.A. 2003. Questions in time: Investigating the structure and dynamics of unfolding classroom discourse. *Discourse Processes*. 35, 2 (2003), 135–198.

[32] Olney, A., Louwerse, M., Matthews, E., Marineau, J., Hite-Mitchell, H. and Graesser, A. 2003. Utterance classification in AutoTutor. *Proceedings of the HLT-NAACL 03 workshop on building educational applications using natural language processing-Volume 2* (2003), 1–8.

[33] Orosanu, L. and Jouvet, D. 2015. Detection of sentence modality on French automatic speech-to-text transcriptions. *Proceedings ICNLSP'2015, International Conference on Natural Language and Speech Processing* (2015).

[34] Quang, V.M., Besacier, L. and Castelli, E. 2007. Automatic question detection: prosodic-lexical features and cross-lingual experiments. *Proc. Interspeech* (2007), 2257–2260.

[35] Raghu, D., Indurthi, S., Ajmera, J. and Joshi, S. 2015. A statistical approach for non-sentential utterance resolution for interactive QA system. *Special Interest Group on Discourse and Dialogue* (SIGDIAL) (2015), 335–343.

[36] Samei, B., Olney, A., Kelly, S., Nystrand, M., D'Mello, S., Blanchard, N., Sun, X., Glaus, M. and Graesser, A. 2014. Domain independent assessment of dialogic properties of classroom discourse. *Proceedings of the 7th International Conference on Educational Data Mining. International Educational Data Mining Society*. (2014), 223–236.

[37] Samei, B., Olney, A.M., Kelly, S., Nystrand, M., D'Mello, S., Blanchard, N. and Graesser, A. 2015. Modeling classroom discourse: Do models that predict dialogic instruction properties generalize across populations? *Proceedings of the 8th International Conference on Educational Data Mining, International Educational Data Mining Society.* (2015), 444–447.

[38] Schuller, B., Steidl, S., Batliner, A. and others 2009. The INTERSPEECH 2009 emotion challenge. *INTERSPEECH* (2009), 312–315.

[39] Schutz, D. 2011. The Common Core State standards for English language arts & literacy in history/social studies, science, and technical subjects: An analysis and an alternative. *Social Studies, Science, and Technical Subjects: An Analysis and an Alternative*. (2011), 1–9.

[40] Shriberg, E., Dhillon, R., Bhagat, S., Ang, J. and Carvey, H. 2004. The ICSI meeting recorder dialog act (MRDA) corpus. *In Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue*, 97–100.

[41] Stolcke, A., Coccaro, N., Bates, R., Taylor, P., Van Ess-Dykema, C., Ries, K., Shriberg, E., Jurafsky, D., Martin, R. and Meteer, M. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*. 26, 3 (2000), 339–373.

[42] Sweigart, W. 1991. Classroom talk, knowledge development, and writing. *Research in the Teaching of English*. (1991), 469–496.

[43] Witten, I.H., Frank, E., Trigg, L.E., Hall, M.A., Holmes, G. and Cunningham, S.J. 1999. Weka: Practical machine learning tools and techniques with Java implementations. Workshop on emerging Engineering and Connectionnist-based Information Systems. (1999), 192–196.

[44] Xiong, S., Guo, W. and Liu, D. 2014. The Vietnamese speech recognition based on rectified linear units deep neural network and spoken term detection system combination. *9th International Symposium on Chinese Spoken Language Processing (ISCSLP)* (2014), 183–186.

[45] Yuan, J. and Jurafsky, D. 2005. Detection of questions in Chinese conversational speech. *IEEE Workshop on Automatic Speech Recognition and Understanding* (2005), 47–52.