An Open Vocabulary Approach for Estimating Teacher Use of Authentic Questions in Classroom Discourse

Connor Cook
Institute of Cognitive Science
University of Colorado Boulder
Boulder, CO 80309
connor.cook@colorado.edu

Andrew M. Olney
Institute for Intelligent Systems
University of Memphis
Memphis, TN 38152
aolney@memphis.edu

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

Sidney K. D'Mello
Institute of Cognitive Science
University of Colorado Boulder
Boulder, CO 80309
sidney.dmello@colorado.edu

ABSTRACT

Automatic assessment of the quality of classroom discourse can have a transformative effect on research and practice on improving teaching effectiveness. We improve on a previous automated method to measure teacher authentic questions – open-ended questions without pre-scripted responses that predict student achievement growth – using classroom audio and expert question codes from two sources: (1) a large archival database of text transcripts of 428 class-sessions from 116 classrooms, and (2) a newly collected sample of 132 high-quality audio recordings with automatic speech recognition transcripts from 27 classrooms. Whereas previous work utilized a "closed vocabulary" approach, consisting of 732 pre-defined word, sentence, and discourse level features, the present "open vocabulary" approach exclusively utilized word and phrase counts from the transcripts themselves. The two approaches yielded substantial, but statistically equivalent, correlations with gold-standard human codes of authenticity (Pearson *r*'s of 0.396 vs. 0.424 and 0.602 vs. 0.613 for datasets 1 and 2, respectively). Importantly, averaging estimates from the two approaches resulted in statistically significant improvements over either approach (*r's* of 0.492 and 0.686 for datasets 1 and 2, respectively). We discuss implications of our findings for automated analysis of classroom discourse.

Keywords

Open vocabulary, authentic questions, classroom discourse

1. INTRODUCTION

(Example 1)

Teacher*: "How does a person become a noble?"*

Student: "*They're born into it."*

Teacher: *"They're born into it, right? It's by family. It gets passed down so if you're a noble, your child would be a noble, their child would be…it's a tradition, right?"*

(Example 2)

Teacher: *"How did that make you guys feel, I mean what was your gut reaction to all that?"*

Student*: "Ashamed."*

Teacher: "*Ashamed in what way?"*

Consider these discourse exchanges between a teacher and his/her students from an actual classroom. The first follows the oft-used, but ineffective, Initiate-Response-Evaluate (IRE) [40] mode of questioning. Now contrast this with the second case, where the teacher asks an open-ended question or a question without a pre-scripted response. Although it only elicited a one-word answer from the student, the teacher withheld evaluation, instead building on the student's response, thereby "opening up" the conversation.

Such questions – called *authentic questions* — whose answers are not presupposed by the teacher (e.g. "Do you think Abigail is going to tell the truth?" [33]) are a core dimension of dialogic instruction related to student engagement and achievement growth [24, 25, 42], and are central to many conceptual models of effective discourse practices [39, 50, 63]. Prior research utilized expert human coders to identify discourse practices at the level of individual questions and thus provided exceptionally precise measures of instructional practice. Our goal is to precisely estimate the prevalence rate of teacher authentic questions using fully-automated methods.

Why bother in the first place? It is because teacher observation has become increasingly central to educational research and school improvement efforts [2, 26, 28, 35, 58]. Observations of classroom practice are valuable because they identify specific domains of practice for improvement [36] and can target dimensions of schooling not captured by test scores, such as socialization processes in elementary school [32]. Classroom observations also enhance school principals' role in managing teachers' work [30]. Yet current in-person observational methods are logistically complex, require observer training, are an expensive allocation of administrators' time [4], and simply do not scale.

Can computers help? We think so, and report the results of ongoing research efforts to automate the analysis of teacher question-asking behavior, a common component across various well-known observation protocols (e.g., Domain 3 of Danielson's Framework for Teaching [16]; PLATO's Classroom Discourse Element [27]). Our specific emphasis on authentic questions is motivated by the strong research base linking them to engagement and achievement as cited above.

1.1 Related Work

There has been considerable work on detecting questions from text [1], with fewer studies focusing on audio [8, 45, 61]. These studies also largely focus on general question detection from meetings and other interactions, which is quite different from the present goal of detecting authentic questions from real-world classrooms. Blanchard et al. [6] and Donnelly et al. [20] investigated question detection from classroom audio, but again, their emphasis was on discriminating questions from other utterances, which is a related but distinct problem from authenticity detection. There has also been research on automated analysis of teacher and student discourse [18, 19, 62], but these studies emphasize modeling of general instructional activities (e.g., distinguishing between lecture vs. group work vs. discussion) rather than authentic questions.

To our knowledge, there have only been three studies germane to our goal of detecting authentic questions from classroom discourse. Samei et al. [53] focused on identifying authenticity from human-transcribed questions from the Partnership for Literacy Study, a large sample of over 20,000 questions and associated "gold-standard" human codes (see section 2.1). The authors repurposed features (e.g., part of speech tags) from an existing speech act classifier [44] to train a J48 classifier to detect authenticity of individual questions. They achieved a Cohen's kappa of 0.34 and accuracy of 67%, which they deemed promising but in need of improvement.

In a follow-up study, Samei et al. [54] focused on testing the generalizability of this model. They split the data based on whether it was collected in an urban or non-urban area and whether the teacher had been trained in dialogic practices (including the use of authentic questions and other effective teacher talk strategies). They found that classifiers trained on a subset (e.g. urban) and tested on the dual subset (e.g. non-urban) were fairly close in accuracy to one another, but that some subpopulations were more representative of the data than others, making them better for classifier training.

Of utmost relevance to the present study is work by Olney et al. [43] on detecting authentic questions from the aforementioned Partnership dataset as well as a newly collected CLASS 5 dataset with automatic speech recognition (ASR) transcriptions (see Section 2.1). Their main goal was to address heavily imbalanced classes, which occur because of the relatively infrequent proportion of authentic questions (about 3%) compared to all teacher utterances. The class imbalance problem was so severe that they forewent identification of individual authentic questions, instead focusing on predicting the proportion of all utterances in a class session that were authentic questions. In other words, an utterance-level binary prediction problem (i.e., labeling an utterance as an authentic question or not) was recast as the problem of predicting the proportion of authentic questions at the class level.

Using a combination of 242 pre-defined features, extracted at the word, sentence, and discourse level, they first attempted aggregating utterance-level predictions of authentic questions, obtained with SMOTEBoost [11], to the class level. This yielded correlations of 0.27 and 0.44 between the predicted and actual (human-coded) authenticity proportions on the Class 5 and Partnership datasets, respectively. The difference in correlations was attributed to the differences in the degree of class imbalance across the two datasets because the Partnership data only contained instructional questions whereas the Class 5 data contained all teacher utterances. Next, they aggregated their utterance-level features to the class level (by taking their mean, sum, and standard deviation to yield 726 features) and then trained a M5P regression tree [23] on the resulting class-level features. The resulting correlation increased from 0.27 to 0.50 for the Class 5 dataset (with the most severe imbalance) but remained similar (0.42 vs. 0.44) for the Partnership dataset (with minor imbalance). Further refinements by Kelly et al. [37], including adding 6 new class-level features, resulted in correlations of 0.61 and 0.42 on the Class 5 and Partnership datasets, respectively.

We attempt to improve on these results using an open vocabulary approach for class-level authenticity prediction. In an open vocabulary approach, the features used to train a classifier are determined from the data itself and are not pre-determined. To illustrate, albeit in a different domain, Schwartz et al. [56] used an open vocabulary approach to predict gender, age, and personality traits based on social media posts. They computed counts of words and phrases (i.e., n-grams) per participant, and then filtered phrases based on pointwise mutual information (PMI) [13, 38], which ensured that they only kept phrases with high informational value. They then normalized the word and phrase counts by the total number of words for each participant and applied the Anscombe transformation [3] to the normalized values to stabilize their variances. They also generated topics using Latent Dirichlet Allocation (LDA) [7, 59]. Using words, phrases, and topics as features, the authors were able to predict gender, age, and personality traits more accurately than a closed vocabulary approach using features from Linguistic Inquiry and Word Count (LIWC) [48, 49]. We apply a variant of this basic approach in the present study.

1.2 Novelty and Contributions

We expand on and improve upon previous work [43] on automatically estimating the proportion of authenticity in classroom discourse using the same datasets. We call this previous approach a closed vocabulary approach since the features are predefined and are independent of the dataset. An advantage of the closed vocabulary approach is that it is less likely to overfit to the dataset at hand because it does not directly encode (as features) specific words from the corpus. This might be particularly important in the case of classroom discourse because generalizable models should encode language that correlates with authentic questions vs. being specific to the particular topic being discussed in class (e.g., The American Civil War).

In contrast, an open vocabulary approach uses counts of words and phrases found in the corpus. The vocabulary is "open" in that the features change depending on the corpus. A potential disadvantage of this approach is that it is more likely to overfit to the training dataset. However, we think this problem can be alleviated by careful selection of words and phrases for use as features. The advantage of this approach is that it ostensibly allows for the detection of a wider variety of instructional constructs due to a lack of pre-determined features. It also yields more interpretable models in that one can examine the specific words, phrases, and utterances that signal authenticity compared to some of the pre-defined features used in the closed vocabulary approach.

Previous research [56] has indicated that an open vocabulary approach outperforms the closed vocabulary approach on a different task of gender, age, and personality prediction from social media. How might it fare for the present task of authenticity prediction and what are the words and phrases that signal authenticity? Is there an advantage to combining both approaches? These are the questions that motivated the present study.

2. METHOD

2.1 Datasets

CLASS 5 (new) data. CLASS 5 data were collected between January 2014 and May 2016 from 132 classes taught by 14 different teachers at seven schools in rural Wisconsin. The data consisted of in-class observations in the form of live coding of authenticity by trained researchers and subsequent offline refinement of the coding from recorded audio. Both teacher and school identifiers were preserved with the data.

Given the logistical constraints of using individual microphones for each student, the recording instrumentation instead focused on high-quality teacher audio suitable for ASR (see [15] for a description of the setup). Classroom audio, which included both teacher and student speech, was recorded from a stationary boundary microphone, and is not of sufficient quality to be used for ASR; it is useful for marking when students speak but is not analyzed further here. Thus this dataset differs from the archival data (see below) in that the audio is automatically segmented into utterances, which are converted into transcripts using Bing Speech ASR with accompanying errors. Further, only teacher speech is transcribed, and the transcripts contain all utterances rather than just questions.

Partnership (archival) data. The archival data was collected in the Partnership for Literacy Study (Partnership), a study of professional development, instruction, and literacy outcomes in middle school English and language arts classrooms. The study collected data from 7th- and 8th- grade English and language arts teachers in Wisconsin and New York State from 2001 to 2003. Over that two-year period, 119 classrooms in 21 schools were observed twice in the fall and twice in the spring. Three of the classrooms had missing question data and could not be used for this study, leaving us with 116 classrooms. Classroom observations for Partnership were conducted using a near-real-time computer-based annotation system [41]. The primary focus of the system was to annotate the dialogic properties of questions asked by both teachers and students. During this process, the instructional questions were transcribed by humans, and the transcriptions were mostly accurate, but not verbatim. Reliability studies indicate that raters agree on question properties approximately 80% of the time, with observation-level inter-rater correlations averaging approximately .95 [42].

Table 1 shows a comparison of both datasets. Note that the same rubric was used to code authentic questions in both datasets.

2.2 Natural Language Processing

Closed vocabulary approach. The closed vocabulary approach used 732 specific features to predict the proportion of authentic questions in class sessions. This feature set includes specific words (like “Why” and “What”), part-of-speech tags, named entity type categorizations (such as PERSON, LOCATION, and DATE), syntactic dependencies (like subject, direct object, and indirect object), and discourse-level features (such as contrast and elaboration discourse relations, and joint, nucleus, and satellite elementary discourse units). There were 242 utterance-level features, which were aggregated at the class level by taking their mean, sum, and standard deviation [43]. Two more features were later added at the utterance level, leading to six more features at the class level, for a total of 732 class-level features [37].

Open vocabulary approach. The open vocabulary approach used a variable number of features depending on the dataset. This method was adapted from the open vocabulary language model developed by Park et al. [46]. To start, counts of words, two-word phrases, and three-word phrases were computed from the corpus. See Table 1 for a comparison of n-gram counts prior to filtering (see below).

We used a stop word list from Pedregosa et al. [47] to filter out the most common English words (such as “the” and “and”), and so these words and phrases including them were filtered out. We also required each word or phrase to occur in at least some percentage of documents, which we call the *cutoff* (we investigated multiple cutoffs, with results shown in Section 3).

We then calculated the pointwise mutual information (PMI) of each phrase, defined as:

$$pmi(phrase) = \log(\frac{p(phrase)}{\Pi\, p(word)})$$

where $p(phrase)$ is the probability of a phrase based on its relative frequency in the training data and $\Pi\, p(word)$ is the product of the probabilities of each word in the phrase in the training data. We filtered out phrases where the PMI was less than three times the number of words in the phrase [13, 38]. This helped ensure that we only used meaningful phrases (such as “language arts”), rather than phrases that were just the result of frequent words occurring next to one another (such as “next we will”). We experimented with PMI thresholds ranging from zero to four times the number of words in the phrase, but no difference in performance was observed. Cutoff and PMI filtering were based only on data in the training folds, ensuring that the test was not affected (see Section 2.3).

Combined approach. We simply averaged predictions from the closed and open vocabulary approaches.

Table 1. Summary of the two datasets

Item	Class 5	Partnership
# Utterances	45,044	Unknown
# Instructional Questions	4,377	25,711
# Authentic Questions	1,510	12,862
% Authentic Utterances	3%	Unknown
% Authentic Questions	34%	50%
Unigrams	17,520	8,358
Bigrams	152,023	61,460
Trigrams	319,545	117,049

Note. % Authentic Utterances refers to teacher utterances aligned with authentic questions. % Authentic Questions refers to instructional questions that were also authentic. N-gram counts are prior to filtering.

2.3 Model Training

We used M5P model trees, which are decision trees that have regression functions at each leaf node [23]. Starting at the root of the tree, decisions to follow a left or right branch are based on the value of a particular feature until a leaf with the appropriate regression model is reached. We chose the M5P model to enable comparisons with previous work [43].

All models used cross-validation, with selection of words and phrases to use as features for the open vocabulary approach based only on the training folds; we did not peek into the testing folds. For generalizability to new teachers, it was important that a teacher would not appear in both the training and testing folds. For the CLASS 5 data, this was achieved using leave-one-teacher-out cross-validation. For the archival Partnership data, the mapping between teachers and data files was incomplete, and so the mapping between schools and data files was used instead. This leave-one-school-out cross-validation assumes that a teacher did not transfer between schools during the study (a likely assumption), and in a sense is even more conservative than leave-one-teacher-out because it controls for similarities shared by teachers at the same school.

It should be noted that the unit of analysis is always a class-session. That is, counts for the language model, feature aggregation, and authenticity aggregation are all done at the level of an individual class-session.

2.4 Method Pseudocode

Below is pseudocode outlining our method for teacher-level cross-validation.

```
Aggregate utterance-level transcripts to the class session level
For each cutoff percentage:
    For each teacher:
        Split data into training set (class sessions from other teachers) and
                        test set (class sessions from this teacher)
        Get counts of n-grams (words, bigrams, and trigrams) for each class session in training set
        Remove n-grams that contain words from stop word list
        Remove n-grams that appear less than once in cutoff percentage of class sessions
        Filter phrases (bigrams and trigrams) using pointwise mutual information
        Get counts of kept n-grams for each class session in test set
        Train M5P model on n-gram counts from training set class sessions
        Use M5P model to predict authenticity on test set class sessions
    Pool class session authenticity predictions across teachers
    Compute correlation between predicted and actual authenticities for cutoff percentage
```

3. RESULTS

Our outcome measure is the Pearson correlation between the computer- and human-coded estimates of proportion authenticity per class session. We recomputed the previous results [37] obtained with the closed vocabulary approach and replicated the previous findings.

3.1 Cutoff Percentage (Open Vocabulary Approach)

As mentioned in Section 2.2, we tested various cutoff percentages for the open vocabulary approach. As can be seen in Figure 1, the correlation starts out low as the model is overwhelmed by the sheer number of features (Figure 2). However, as the cutoff becomes more stringent and the number of features decreases, the results improve, until the correlations peaks at 0.602, achieved with 52 features at an 82% cutoff. Beyond this point, the correlation steeply drops as too few features remain.

We observed a different pattern for the Partnership data as noted in Figure 3 and Figure 4. Here, the results were less dependent on the number of features, though the best correlation of 0.396 was obtained at the 61% cutoff with only 6 features retained. It should be noted that we only considered up to a 70% cutoff for this dataset because there were only three remaining features beyond this point. This is unsurprising because the Partnership data, though more diverse, only contains questions compared to the full transcripts in the CLASS 5 dataset, and consequently contains far fewer unique n-grams (see Section 2.2).

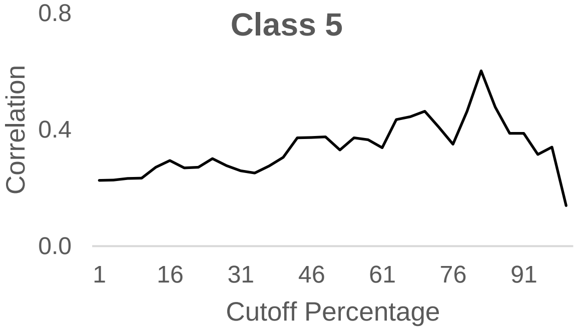

Figure 1. Correlation by cutoff % for Class 5 dataset

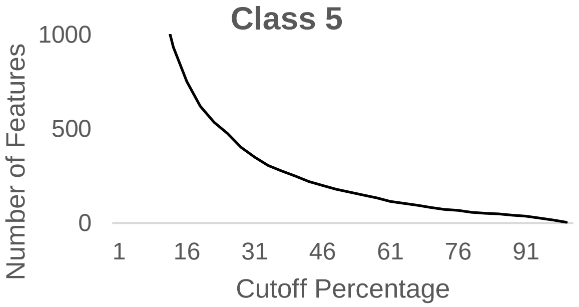

Figure 2. # of features by cutoff % for Class 5 dataset

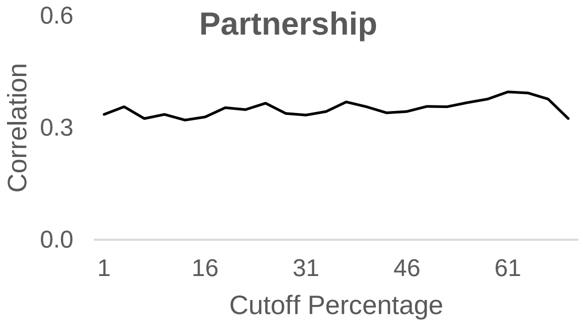

Figure 3. Correlation by cutoff % for the Partnership dataset

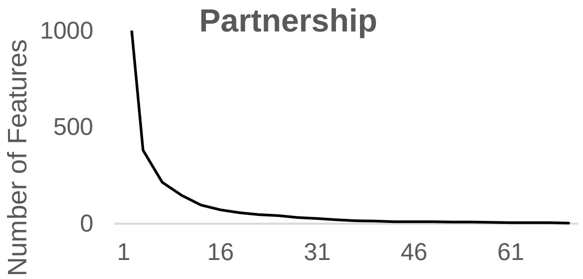

Figure 4. # of features by cutoff % for the Partnership dataset

3.2 Comparison with Closed Vocabulary Results

For the Class 5 data, the best correlation of 0.602 obtained via the open vocabulary approach was significant ($p < .001$) and similar to the significant 0.613 ($p < .001$) correlation obtained from the closed vocabulary approach. Zou's [66] test of the difference between two overlapping dependent correlations with one common variable (i.e., the gold-standard authenticity codes) indicated that the two correlation coefficients were statistically equivalent at the $p < .05$ level. A similar pattern of results was obtained for the Partnership data in that the significant 0.396 ($p < .001$) correlation from the open vocabulary approach was statistically equivalent to the 0.421 significant ($p < .001$) correlation from the closed vocabulary approach at the $p < .05$ level. Subsequent results focus on these two "best" models.

3.3 Combined Models

The analyses thus far indicate that the closed and open vocabulary approaches were equally predictive of authenticity across both datasets. Authenticity estimates from both methods correlated at .559 ($p < .001$) and .371 ($p < .001$) for the Class 5 and Partnership datasets, respectively, suggesting some, but not substantial, redundancy. This raises the question of whether a combination of the two approaches might improve predictive power.

We addressed this question by averaging the predictions of the two best models (we also attempted feature-level fusion, but this resulted in lower performance; results not shown here). For Class 5, the combined model predicted authenticity with a significant correlation of .686 ($p < .091$), which was quantitatively and statistically higher ($p < .05$) than the 0.602 and 0.613 correlations obtained from the open and closed vocabulary approaches, respectively (see Figure 5).

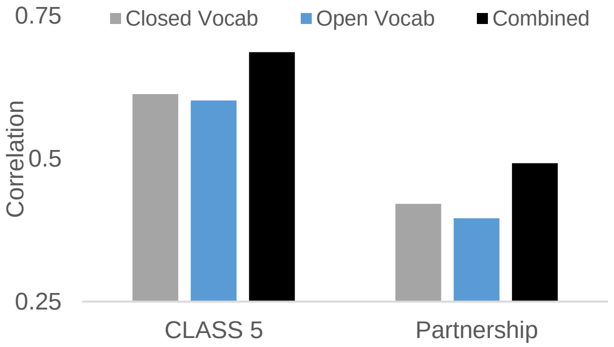

Figure 5. Comparison of closed, open, and combined models

These results can be visualized as a density plot (see left of Figure 6). The plot illustrates smoothed histograms of class-level computer- and human-provided proportional authenticity estimates. We note the combined model tends to slightly overestimate the mean compared to the human-coded data. Its predictions are also less positively skewed, ostensibly because it underpredicts some cases with considerable human-coded authenticity (also see right of Figure 6).

A similar pattern of results was obtained for the Partnership data. Specifically, the combined model's correlation of .492 was significant ($p < .001$) and also significantly higher ($p < .05$) than the 0.396 and 0.421 correlations obtained from the open and closed vocabulary approaches, respectively (see Figure 5). As noted in the density plot in Figure 7, the combined model is "peakier" with a reduced range in either direction compared to the human-coded data. The model has difficulty with cases associated with very low and very high human-coded authenticity (see scatterplot in Figure 7).

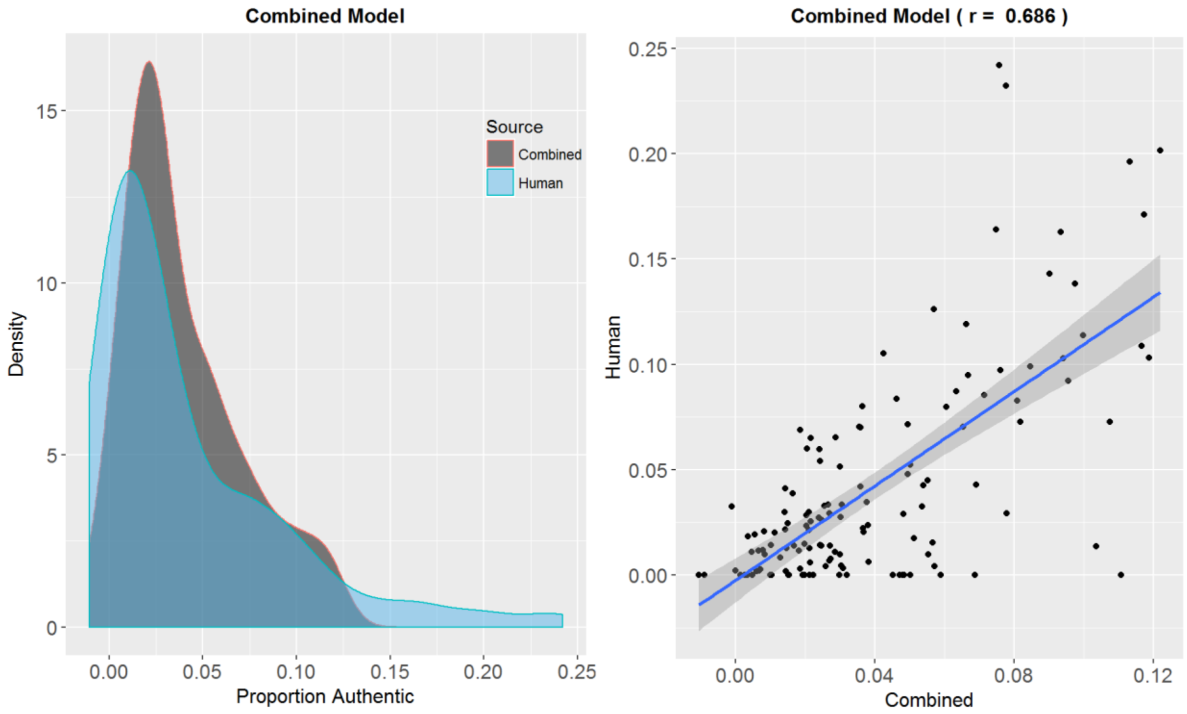

Figure 6. Density plot and scatter plot showing the resulting predictions from combining both the open and closed vocabulary models on the Class 5 dataset compared to human codes.

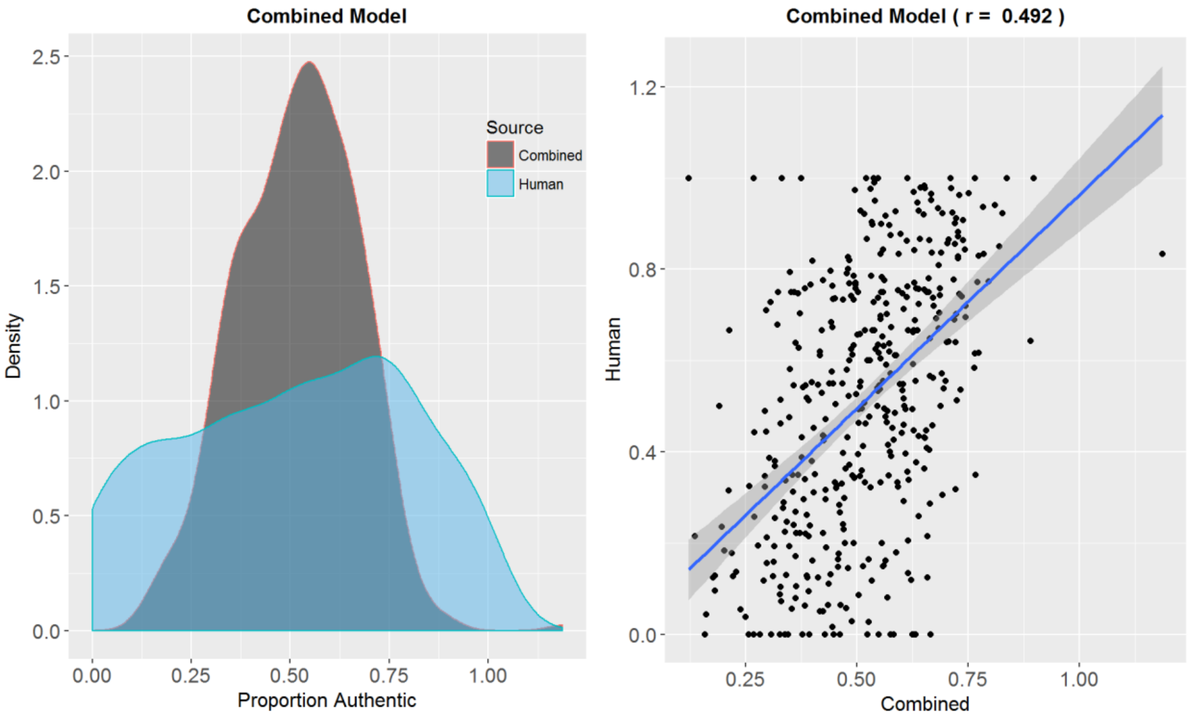

Figure 7. Density plot and scatter plot showing the resulting predictions from combining both the open and closed vocabulary models on the Partnership dataset compared to human codes.

3.4 Feature Analysis

We investigated the features (words and phrases) from the best open vocabulary model in the form of word clouds[1] scaled using correlations of individual features with authenticity rather than by absolute frequency in the corpus. Figure 8 shows words that positively correlate with authenticity for the Class 5 dataset. The words "Question," "Maybe," and "Ok" correlated most strongly with authenticity (correlation values of .254, .229, and .219 respectively). These words are used to ask questions, indicate uncertainty, or to accept another's response. This might suggest the teacher is setting the stage for open dialogue, which is precisely what authentic questioning signals.

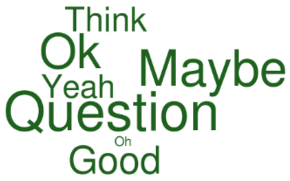

Figure 8. Words that are positively correlated with authenticity in the Class 5 dataset.

Alternatively, the words "Need," "Work," and "Doing" were most negatively correlated with authenticity (correlation values of -.383, -.330, and -.302 respectively) – see Figure 9 for the full word cloud. These words might be more likely to occur during non-dialogic activities, such as lecture or individual work.

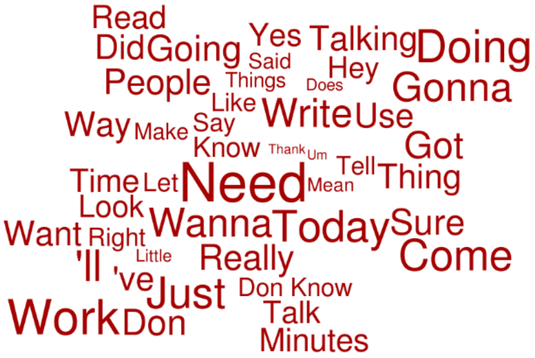

Figure 9. Words and phrases that are negatively correlated with authenticity for the Class 5 dataset.

For the Partnership dataset, only "Like," "Think," and "Say" were positively correlated with authenticity (correlation values of .177, .158, and .055 respectively). It is plausible that these terms accompany more open-ended authentic questions (e.g., "Why do you *like* the last story?" or "What do you *think* about that?" or "Why did you *say* that?") compared to their non-authentic counterparts that solicit specific responses (e.g., "What do we *know* about the beginning?" – these are all hypothetical examples).

There were also only three words that negatively correlated with authenticity. "Does" was more strongly correlated than "Know" and "Did" (correlation values of -.246, -.062, and -.032 respectively). "Does" might be more likely to accompany information-seeking questions, such as "What *does* mandible mean?" or "How *does* Jim know he is in danger?" compared to more authentic questions. Of course, these are only speculative suggestions that need to be verified by more systematic analyses.

4. DISCUSSION

We addressed the task of automated prediction of the proportion of authentic questions in a class session from real-world classroom discourse. We compared a previous closed vocabulary approach to an open vocabulary approach, combined the two, and tested them on two datasets. In the remainder of this section, we discuss our main findings, possible applications of this work, as well as limitations and directions for future work.

4.1 Main Findings

We found that the open and closed vocabulary approaches yielded equitable performance on both datasets, but a simple combination of the two resulted in statistically better results. This suggests that knowledge of the domain, as reflected in some of the closed vocabulary features (the question specific ones), is very important, but missed patterns can be captured using the open vocabulary approach. Thus, the combined approach capitalized on the strengths while mitigating the weaknesses of each individual approach.

The fact that the result replicated across two rather different datasets increases our confidence in the findings. This is particularly important because the datasets differ in a number of substantial ways – for example, one contained ASR transcripts of entire class sessions while the other contained human transcriptions of question text; one was much more variable, larger in size, and was validated at the school-level compared to the smaller, more homogenous dataset that was validated at the teacher level.

The open vocabulary approach provided key insights into the specific words used to guide its predictions. Of particular interest was the fact that the word "think" was positively correlated with authenticity in both datasets, but the word "like" was negatively correlated with authenticity in one and positively in another. This suggests the importance of examining the broader context in which these words appear.

4.2 Applications

Like anyone, teachers need feedback to improve. But in contrast to an expert musician or athlete who receives continual feedback across the countless hours spent in practice for the occasional performance, a teacher delivers approximately 1,000 "performances" a year with almost no feedback [22, 60]. Given the pivotal role of feedback to learning [5, 14, 21, 57], the lack of immediate and objective feedback is a critical barrier that needs to be cracked if we are truly going to innovate teaching.

Accordingly, one key application of our work is in an automated teacher feedback system with the goal of improving teaching effectiveness and consequently student learning. Such a system needs to be able to detect different measures of teaching effectiveness beyond authentic questions (e.g., goal clarity, disciplinary concepts, strategy use, elaborated feedback), and the open vocabulary approach is particularly suited for this task.

Ultimately, we envision technology that will autonomously analyze teachers' behaviors as they go about their daily activities, both within and beyond the classroom. The technology would provide formative feedback (i.e., feedback aimed at improvement rather

[1] Word clouds were generated via https://worditout.com

than evaluation [57]), which the teacher can use as a form of DIY (do it yourself) professional development or share with support staff. The feedback can enable reflective practice, defined as thoughtfully considering one's own actions and experiences to refine one's skill in a selected discipline [55]. Due to its emphasis on contextualized analysis and metacognition, reflective practice holds great promise in improving teaching effectiveness [9, 10], which should result in positive downstream influences on student achievement given the robust relationship between the two [12, 17, 29, 34, 51, 52, 65].

Such a technology can also be used to streamline research into teaching effectiveness, which currently relies on cumbersome human observation (see the introduction). Going beyond question authenticity, at a broader level, such a technology could be used to advance basic research on student-teacher discourse, essentially opening up the methods of "big data" science to real-world classrooms.

4.3 Limitations & Future Work

One limitation of this study is the amount and variety of classroom transcriptions with corresponding authenticity labels. The Class 5 dataset was collected in a very limited geographical location. The Partnership dataset, although much more variable in terms of the sample, only included transcriptions of questions rather than transcriptions of all teacher utterances.

Our models also detect authenticity at the level of an entire class session, rather than at the individual utterance level. Finer grain size is needed to provide actionable feedback to teachers, at least with respect to the vision articulated above. We also did not correlate our results with more objective measures, particularly achievement growth, due to a lack of available data.

In addition to addressing the aforementioned limitations, future work should include using the open vocabulary approach to predict measures beyond authenticity. We are taking a step in this direction by re-coding current CLASS 5 audio as well as collecting new audio files and coding them for the following broader dimensions of discourse linked, or hypothesized to be linked, to student achievement growth: goal clarity, disciplinary concepts, and strategy use for teacher-led discourse, and challenge, connection, and elaborated feedback for transactional discourse.

We are also streamlining the data collection process, essentially providing usable tools for teachers to collect their own data, and have collected over 65 hours of audio (in about two months) using this approach. When coupled with existing data from CLASS 5, we estimate that the combined datasets will be sufficiently large to experiment with deep natural language processing methods, such as long short-term recurrent neural networks [31] and hierarchical attention networks [64].

4.4 Concluding Remarks

We applied an open vocabulary approach to the task of predicting authentic questions in classroom discourse and compared it to a previous closed vocabulary approach applied to the same problem. We found that the two approaches yielded equivalent performance, but a combination led to higher accuracies than either method alone. We achieved a correlation of close to 0.70 on real-world audio, which suggests that fully-automated methods might complement or even replace humans on the difficult task of determining the level of dialogism in classroom discourse.

5. ACKNOWLEDGMENTS

We acknowledge the CLASS 5 team for their contributions to the study. This research was supported by the National Science Foundation (NSF IIS 1735785) and Institute of Education Sciences (IES R305A130030). Any opinions, findings, conclusions, or recommendations expressed in this paper are those of the author and do not represent the views of the funding agencies.

6. 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*. Technical Report. Media in Context (MICO).

[2] American Institutes for Research. 2013. *Databases on state teacher and principal evaluation policies*. Retrieved from http://resource.tqsource.org/stateevaldb/Compare50States.aspx.

[3] Anscombe, F. J. 1948. The transformation of poisson, binomial and negative binomial data. *Biometrika*. 35, 3/4 (Dec. 1948), 246-254. DOI= https://doi.org/10.1093/biomet/35.3-4.246.

[4] Archer, J., Cantrell, S., Holtzman, S. L., Joe, J. N., Tocci, C. M., and Wood, J. 2016. *Better Feedback for Better Teaching: A Practical Guide to Improving Classroom Observations*. Jossey-Bass, San Francisco, CA.

[5] Azevedo, R. and Bernard, R. M. 1995. A meta-analysis of the effects of feedback in computer-based instruction. *Journal of Educational Computing Research*. 13, 2 (Sep. 1995), 111-127. DOI= https://doi.org/10.2190/9LMD-3U28-3A0G-FTQT.

[6] Blanchard, N. et al. 2016. Identifying Teacher Questions Using Automatic Speech Recognition in Classrooms. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue* (Los Angeles, CA, USA, September 13 - 15, 2016). 191-201.

[7] Blei, D. M. D., Ng, A. Y., and Jordan, M. I. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*. 3, 1 (Jan. 2003), 993-1022. DOI= https://doi.org/10.1162/jmlr.2003.3.4-5.993.

[8] Boakye, K., Favre, B., and Hakkani-Tür, D. 2009. Any questions? Automatic question detection in meetings. In *Proceedings of the 11th Biannual IEEE Workshop on Automatic Speech Recognition and Understanding (*Merano, Italy, December 13 - 17, 2009). ASRU '09. IEEE, Piscataway, NJ, 485-489.

[9] Camburn, E. M. 2010. Embedded Teacher Learning Opportunities as a Site for Reflective Practice: An Exploratory Study. *American Journal of Education*. 116, 4 (Jun. 2010), 463-489. DOI= https://doi.org/10.1086/653624.

[10] Camburn, E. M. and Han, S. W. 2015. Infrastructure for teacher reflection and instructional change: An exploratory study. *Journal of Educational Change*. 16, 4 (Nov. 2015), 511-533. DOI= https://doi.org/10.1007/s10833-015-9252-6.

[11] Chawla, N. V. 2005. Data Mining for Imbalanced Datasets: An Overview. In *Data Mining and Knowledge Discovery Handbook*, O. Maimon and L. Rokach, Eds. Springer, Boston, MA, 875-886. DOI= https://doi.org/10.1007/978-0-387-09823-4_45.

[12] Chetty, R., Friedman, J. N., and Rockoff, J. E. 2014.

Measuring the Impacts of Teachers I: Evaluating bias in teacher value-added estimates. *American Economic Review*. 104, 9 (Sep. 2014), 2593-2632. DOI= https://doi.org/10.3386/w19423.

[13] Church, K. W. and Hanks, P. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*. 16, 1 (Mar. 1990), 22-29.

[14] D’Mello, S. K., Lehman, B., and Person, N. K. 2010. Expert tutors feedback is immediate, direct, and discriminating. In *Proceedings of the 23rd Florida Artificial Intelligence Research Society Conference* (Daytona Beach, Florida, USA, May 19 - 21, 2010). AAAI, Palo Alto, CA, 504-509.

[15] D’Mello, S. K. et al. 2015. Multimodal Capture of Teacher-Student Interactions for Automated Dialogic Analysis in Live Classrooms. In *Proceedings of the 2015 ACM International Conference on Multimodal Interaction* (London, United Kingdom, January 19 - 20, 2015). ICMI '15. ACM, New York, NY, 557-566. DOI= https://doi.org/10.1145/2818346.2830602.

[16] Danielson, C. 2007. *Enhancing Professional Practice: A Framework for Teaching*. Association for Supervision and Curriculum Development, Alexandria, VA.

[17] Darling-Hammond, L. 2000. Teacher Quality and Student Achievement. *Education policy analysis archives*. 8, 1 (Jan. 2000), 1-44. DOI= https://doi.org/10.14507/epaa.v8n1.2000.

[18] Donnelly, P. J. et al. 2016. Automatic Teacher Modeling from Live Classroom Audio. In *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization* (Halifax, Canada, July 13 - 16, 2016). UMAP ’16. ACM, New York, NY, 45-53. DOI= https://doi.org/10.1145/2930238.2930250

[19] Donnelly, P. J. et al. 2016. Multi-sensor modeling of teacher instructional segments in live classrooms. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction* (Tokyo, Japan, November 12 - 16, 2016). ICMI '16. ACM, New York, NY, 177-184. DOI= https://doi.org/10.1145/2993148.2993158.

[20] Donnelly, P. J., Blanchard, N., Olney, A. M., Kelly, S., Nystrand, M., and D'Mello, S. K. 2017. Words Matter: Automatic Detection of Teacher Questions in Live Classroom Discourse using Linguistics, Acoustics, and Context. In *Proceedings of the Seventh International Learning Analytics and Knowledge Conference* (Vancouver, BC, Canada, March 13 - 17, 2017). LAK '17. ACM, New York, NY, 218-227. DOI= https://doi.org/10.1145/3027385.3027417.

[21] Ericsson, K. A., Krampe, R. T., and Tesch-Römer, C. 1993. The role of deliberate practice in the acquisition of expert performance. *Psychological Review*. 100, 3 (Jul. 1993), 363. DOI= https://doi.org/10.1037/0033-295X.100.3.363.

[22] Fadde, P. J. and Klein, G. A. 2010. Deliberate performance: Accelerating expertise in natural settings. *Performance Improvement*. 49, 9 (Oct. 2010), 5-14. DOI= https://doi.org/10.1002/pfi.20175.

[23] Frank, E., Wang, Y., Inglis, S., Holmes, G., and Witten, I. H. 1998. Using model trees for classification. *Machine Learning*. 32, 1 (Jul. 1998), 63-76.

[24] Gamoran, A. and Kelly, S. 2003. Tracking, instruction, and unequal literacy in secondary school English. In *Stability and change in American education: Structure, process, and outcomes*, M. T. Hallinan et al., Eds. Eliot Werner Publications Incorporated, Clinton Corners, NY, 109-126.

[25] Gamoran, A. and Nystrand, M. 1992. Taking students seriously. In *Student Engagement and Achievement in American Schools*, F. M. Newman, Ed. Teachers College Press, New York, NY, 40-61.

[26] Goe, L., Biggers, K., and Croft, A. 2012. *Linking Teacher Evaluation to Professional Development: Focusing on Improving Teaching and Learning.* Research & Policy Brief. National Comprehensive Center for Teacher Quality.

[27] Grossman, P., Greenberg, S., Hammerness, K., Cohen, J., Alston, C., and Brown, M. 2009. Development of the protocol for language arts teaching observation (PLATO). In *annual meeting of the American Educational Research Association* (San Diego, California, USA, 2009).

[28] Hamilton, L. 2012. Measuring teaching quality using student achievement tests: Lessons from educators’ response to No Child Left Behind. In *Assessing teacher quality: Understanding teacher effects on instruction and achievement*, S. Kelly, Ed. Teachers College Press, New York, NY, 49-76.

[29] Hanushek, E. A. and Rivkin, S. G. 2006. Teacher quality. In *Handbook of the Economics of Education*, E. A. Hanushek and F. Welsh, Eds. North-Holland, Amsterdam, The Netherlands, 1051-1078.

[30] Harris, D. N., Ingle, W. K., and Rutledge, S. A. 2014. How Teacher Evaluation Methods Matter for Accountability: A Comparative Analysis of Teacher Effectiveness Ratings by Principals and Teacher Value-Added Measures. *American Educational Research Journal*. 51, 1 (Feb. 2014), 73-112. DOI= https://doi.org/10.3102/0002831213517130.

[31] Hochreiter, S. and Schmidhuber, J. 1997. Long Short-Term Memory. *Neural Computation*. 9, 8 (Nov. 1997), 1735-1780. DOI= https://doi.org/10.1162/neco.1997.9.8.1735.

[32] Jennings, J. L. and Corcoran, S. P. 2012. Beyond high-stakes tests: Teacher effects on other educational outcomes. In *Assessing teacher quality: Understanding teacher effects on instruction and achievement*, S. Kelly, Ed. Teachers College Press, New York, NY, 77-95.

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

[34] Kane, T., Kerr, K., and Pianta, R. 2014. *Designing teacher evaluation systems: New guidance from the measures of effective teaching project*. Jossey-Bass, San Francisco, CA.

[35] Kane, T. J., McCaffrey, D. F., Miller, T. and Staiger, D. O. 2013. *Have We Identified Effective Teachers? Validating Measures of Effective Teaching Using Random Assignment*. Bill & Melinda Gates Foundation, Seattle, WA.

[36] Kane, T. J. and Staiger, D. O. 2012. *Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains*. Bill & Melinda Gates Foundation, Seattle, WA.

[37] Kelly, S., Olney, A. M., Donnelly, P. J., Nystrand, M., and D'Mello, S. K. Automatically Measuring Question Authenticity in Real-World Classrooms. *In Review*.

[38] Lin, D. 1998. Extracting collocations from text corpora. In

First Workshop on Computational Terminology (Montreal, Canada, August 15, 1998). 57-63.

[39] McKeown, M. G. and Beck, I. L. 2015. Effective classroom talk is reading comprehension instruction. In *Socializing intelligence through academic talk and dialogue*, L. B. Resnik et al., Eds. American Educational Research Association, Washington, D.C., 51-62.

[40] Mehan, H. 1979. *Learning Lessons: Social Organization in the Classroom*. Harvard University Press, Cambridge, MA.

[41] Nystrand, M. 1988. *CLASS (Classroom language assessment system) 2.0: A Windows laptop computer system for the in-class analysis of classroom discourse*. Wisconsin Center for Education Research, Madison, WI.

[42] Nystrand, M. and Gamoran, A. 1997. The big picture: Language and learning in hundreds of English lessons. In *Opening dialogue: Understanding the dynamics of language and learning in the English classroom*. M. Nystrand, Ed. Teachers College Press, New York, NY, 30-74.

[43] Olney, A. M., Samei, B., Donnelly, P. J., and D'Mello, S. K. 2017. Assessing the Dialogic Properties of Classroom Discourse: Proportion Models for Imbalanced Classes. In *Proceedings of the 10th International Conference on Educational Data Mining* (Wuhan, China, June 25 - 28, 2017). EDM '17. 162-167.

[44] Olney, A. M., Louwerse, M., Matthews, E., Marineau, J., Hite-Mitchell, H., and Graesser, A. 2003. Utterance Classification in AutoTutor. In *Proceedings of the Human Language Technology - North American Chapter of the Association for Computational Linguistics 03 Workshop on Building Education Applications Using Natural Language Processing* (Philadelphia, PA, May 31, 2003). Association for Computational Linguistics, Stroudsburg, PA, 1-8.

[45] Orosanu, L. and Jouvet, D. 2015. Detection of sentence modality on French automatic speech-to-text transcriptions. In *Proceedings of International Conference on Natural Language and Speech Processing* (Algiers, Algeria, October 18 - 19, 2015). IEEE.

[46] Park, G. et al. 2015. Automatic personality assessment through social media language. *Journal of Personality and Social Psychology*. 108, 6 (Jun. 2015), 934-952. DOI= https://doi.org/10.1037/pspp0000020.

[47] Pedregosa, F. et al. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*. 12, 1, (Oct. 2011), 2825-2830.

[48] Pennebaker, J. W., Mehl, M. R., and Niederhoffer, K. G. 2003. Psychological Aspects of Natural Language Use: Our Words, Our Selves. *Annual Review of Psychology*. 54, 1 (Feb. 2003), 547-577. DOI= https://doi.org/10.1146/annurev.psych.54.101601.145041.

[49] Pennebaker, J. W., Chung, C. K., Ireland, M., Gonzales, A., and Booth, R. J. 2007. *The Development and Psychometric Properties of LIWC2007*. LIWC.net, Austin, TX. DOI= https://doi.org/10.1068/d010163.

[50] Resnick, L., Michaels, S., and O'Connor, C. 2010. How (well structured) talk builds the mind. In *Innovations in educational psychology, Perspectives on learning, teaching, and human development*, D. Preiss and R. J. Sternberg, Eds. Springer, Boston, MA, 163-194.

[51] Rivkin, S. G., Hanushek, E. A., and Kain, J. F. 2005. Teachers, schools and academic achievement. *Econometrica*. 73, 2 (Mar. 2005), 417-458.

[52] Rockoff, J. E. 2004. The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data. *The American Economic Review*. 94, 2 (May. 2004), 247-252. DOI= https://doi.org/10.1257/0002828041302244.

[53] Samei, B. et al. 2014. Domain Independent Assessment of Dialogic Properties of Classroom Discourse. In *Proceedings of the 7th International Conference on Educational Data Mining* (London, United Kingdom, July 04 - 07, 2014). EDM '14. 233-236.

[54] Samei, B. et al. 2015. Modeling Classroom Discourse: Do Models that Predict Dialogic Instruction Properties Generalize across Populations? In *Proceedings of the 8th International Conference on Educational Data Mining* (Madrid, Spain, June 26 - 29, 2015). EDM '15. 444-447.

[55] Schon, D. A. 1987. *Educating the Reflective Practitioner: Toward a New Design for Teaching and Learning in the Professions*. Jossey-Bass, San Francisco, CA.

[56] Schwartz, H. A. et al. 2013. Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. *PLoS ONE*. 8, 9 (Sep. 2013), e73791. DOI= https://doi.org/10.1371/journal.pone.0073791.

[57] Shute, V. J. 2008. Focus on Formative Feedback. *Review of Educational Research*. 78, 1 (Mar. 2008), 153-189. DOI= https://doi.org/10.3102/0034654307313795.

[58] Stein, M. K. and Matsumura, L. C. 2009. Measuring instruction for teacher learning. In *Measurement issues and assessment for teacher quality*, D.H. Gitomer, Ed. Sage Publications, Los Angeles, CA, 179-205.

[59] Steyvers, M. and Griffiths, T. 2007. Probabilistic Topic Models. In *Handbook of latent semantic analysis*, T. K. Landauer et al., Eds. Lawrence Erlbaum Associates, Inc., Mahwah, NJ, 424-440.

[60] Stigler, J. and Miller, K. 2006. Expertise and Expert Performance in Teaching. In *The Cambridge Handbook of Expertise and Expert Performance*, K. A. Ericsson et al, Eds. Cambridge University Press, Cambridge, United Kingdom, 431-452.

[61] Stolcke, A. et al. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*. 26, 3 (Sep. 2000), 339-373.

[62] Wang, Z., Miller, K., and Cortina, K. 2013. Using the LENA in teacher training: Promoting student involvement through automated feedback. *Unterrichtswissenschaft*. 4, 4 (Nov. 2013) 290-305.

[63] Wilkinson, I. A. G., Soter, A. O., and Murphy, P. K. 2010. Developing a model of Quality Talk about literary text. In *Bringing reading research to life*, M. G. McKeown and L. Kucan, Eds. Guilford, New York, NY, 142-169.

[64] Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., and Hovy, E. 2016. Hierarchical Attention Networks for Document Classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (San Diego, California, June 12 - 17, 2016). Association for Computational Linguistics, Stroudsburg, PA, 1480-1489.

[65] Yoon, K. S., Duncan, T., Lee, S. W. Y., Scarloss, B., and

Shapley, K. L. 2007. *Reviewing the evidence on how teacher professional development affects student achievement*. REL 2007-No. 033. Regional Educational Laboratory Southwest (NJ1).

[66] Zou, G. Y. 2007. Toward Using Confidence Intervals to Compare Correlations. *Psychological Methods*. 12, 4 (Dec. 2007), 399-413. DOI= https://doi.org/10.1037/1082-989X.12.4.399.