A Comparative Analysis of Automatic Speech Recognition Errors in Small Group Classroom Discourse

Jie Cao
University of Colorado Boulder
Boulder, Colorado, USA
jie.cao@colorado.edu

Ananya Ganesh
University of Colorado Boulder
Boulder, Colorado, USA
ananya.ganesh@colorado.edu

Jon Cai
University of Colorado Boulder
Boulder, Colorado, USA
jon.z.cai@colorado.edu

Rosy Southwell
University of Colorado Boulder
Boulder, Colorado, USA
roso8920@colorado.edu

E. Margaret Perkoff
University of Colorado Boulder
Boulder, Colorado, USA
margaret.perkoff@colorado.edu

Michael Regan
University of Colorado Boulder
Boulder, Colorado, USA
michael.regan@colorado.edu

Katharina Kann
University of Colorado Boulder
Boulder, Colorado, USA
katharina.kann@colorado.edu

James H. Martin
University of Colorado Boulder
Boulder, Colorado, USA
james.martin@colorado.edu

Martha Palmer
University of Colorado Boulder
Boulder, Colorado, USA
martha.palmer@colorado.edu

Sidney D'Mello
University of Colorado Boulder
Boulder, Colorado, USA
sidney.dmello@colorado.edu

ABSTRACT

In collaborative learning environments, effective intelligent learning systems need to accurately analyze and understand the collaborative discourse between learners (i.e., group modeling) to provide adaptive support. We investigate how automatic speech recognition (ASR) errors influence discourse models of small group collaboration in noisy real-world classrooms. Our dataset consisted of 30 students recorded by consumer off-the-shelf microphones (Yeti Blue) while engaging in dyadic- and triadic- collaborative learning in a multi-day STEM curriculum unit. We found that two state-of-the-art ASR systems (Google Speech and OpenAI Whisper) yielded very high word error rates (0.822, 0.847) but very different profiles of error with Google being more conservative, rejecting 38% of utterances instead of 12% for Whisper. Next, we examined how these ASR errors influenced down-stream small group modeling based on pre-trained large language models for three tasks: Abstract Meaning Representation parsing (AMRPARSING), on-task/off-task detection (ONTASK), and Accountable Productive Talk prediction (TALKMOVE). As expected, models trained on clean human transcripts yielded degraded performance on all three tasks, measured by the transfer ratio (TR). However, the TR of the specific sentence-level AMRPARSING task (.39 - .62) was much lower than that of the abstract discourse-level ONTASK (.63- .94) and TALKMOVE tasks (.64-.72). Furthermore, different training strategies that incorporated ASR transcripts alone or as augmentations of human transcripts increased accuracy for the discourse-level tasks (ONTASK and TALKMOVE) but not AMRPARSING. Simulation experiments suggested that the models were tolerant of missing utterances in the dialog context, and that jointly improving ASR accuracy on important word classes (e.g., verbs and nouns) can improve performance across all tasks. Overall, our results provide insights into how different types of NLP-based tasks might be tolerant of ASR errors under extremely noisy conditions and provide suggestions for how to improve accuracy in small group modeling settings for a more equitable, engaging, and adaptive collaborative learning environment.

CCS CONCEPTS

• **Human-centered computing** → **Collaborative and social computing**; • **Applied computing** → **Education**.

KEYWORDS

Group Discourse Analysis, Automatic Speech Recognition, Text Tagging, Collaborative Learning

ACM Reference Format:
Jie Cao, Ananya Ganesh, Jon Cai, Rosy Southwell, E. Margaret Perkoff, Michael Regan, Katharina Kann, James H. Martin, Martha Palmer, and Sidney D'Mello. 2023. A Comparative Analysis of Automatic Speech Recognition Errors in Small Group Classroom Discourse. In *UMAP '23: Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization (UMAP '23), June 26–29, 2023, Limassol, Cyprus.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3565472.3595606

UMAP '23, June 26–29, 2023, Limassol, Cyprus

ACM ISBN 978-1-4503-9932-6/23/06.
https://doi.org/10.1145/3565472.3595606

1 INTRODUCTION

Collaborative learning (CL) is a widely used educational approach involving joint intellectual effort among students in coordination with teachers [15, 74, 95]. It represents a significant shift away from traditional teacher-centered or lecture-centered classrooms. In CL, students usually work together in small groups of two or more towards a common goal, e.g., mutually developing consensus on an issue, sharing knowledge independently acquired, or coming up with solutions to problems [19, 37, 40, 43]. Through these constructive processes and interactive discourse, students are not simply taking in information, but also creating new ideas and having their voices included in the classroom discourse. Thus CL is increasingly used in today's classrooms, showing various benefits such as efficient knowledge acquisition [93], critical thinking [33], creative thinking [48], and developing collaborative problem solving skills [36]. Teachers play a key role in orchestrating effective CL in classrooms [73]. For example, teachers can provide guidance and content support when students get stuck and might need to intervene when students go off-task for extended periods of time. They may also need to help students be respectful of the other group members, encourage them to maintain classroom norms [73], and encourage high-quality disciplinary discourse [62]. At the same time, teachers can benefit from knowledge about insights and ideas that excite youth in their small group discussions.

A key factor limiting the effectiveness of CL is that it is difficult for teachers to monitor and coordinate CL activities across multiple groups at the same time for the simple reason that they cannot be omnipresent and omniscient. Thus, to increase teachers' awareness and assist them in orchestrating effective CL in classrooms, an exciting opportunity is to build adaptive AI-enabled learning environments via small-group discourse modeling (or group modeling). Because discourse is a key product of CL, a promising approach is to use natural language processing (NLP) to model students' collaborative discourse [27, 38, 58, 83]. Furthermore, group diversity [16, 39] has been widely studied in CL, impacting on both individual [13, 86] and group learning [96]. Hence, building effective learning environments requires adapting to different groups and individuals, which needs accurate discourse analysis to describe the behaviors of different groups, and the personality of users in various dimensions [22, 87]. Recent advances in **L**arge **L**anguage **M**odels (LLMs) [18, 57] offered novel opportunities to model a range of constructs such as the use of particular talk moves [91], collaborative problem-solving skills [12, 68], disruptive talk detection [65], and detection of off-task behaviors [8], etc.

However, a key obstacle to deploying accurate language-based systems in real-world is obtaining accurate transcripts from the multiparty student discourse. In most CL implementations, multiple small groups sit close to each other in a classroom resulting in a myriad of challenges including noisy audio, multiparty chatter, and other ambient noise [21]. Noisy environments such as these are a challenge for transcription by automatic speech recognition (ASR) and diarisation systems (i.e., attributing utterances to individual students) [68, 79]. While there are massive improvements in state-of-the-art ASR systems for recognizing adult speech [70, 81], there is the added complication of recognizing child speech (even in ideal, noise-free environments), where articulations are less clear than those of adults [53]. Many differences exist in the acoustic and linguistic features between child speech and adult speech [31, 53]. Most training data used in developing ASR models comes from adult speech, therefore underperformance on child speech is not unexpected due to this domain mismatch [72]. As a result, most existing attempts to leverage NLP to analyze and support collaborative discourse rely either on typed transcripts from chats or on human-transcribed speech [23, 26, 38, 58, 83].

Because perfect ASR is unlikely to be achieved in this setting, there is the question of how accurate ASR needs to be for functional modeling of small group collaborative learning discourse. We argue that imperfect ASR is not necessarily a death knell in applications where whole utterances or dialogues, rather than single words, are the unit of analysis. For instance, [27] simulated ASR errors input to a model of teamwork in adult teams in a military scenario and found that even with a WER of .57, their teamwork classifier only performed 20% worse. Similarly, [20] and [55] demonstrate effective human-computer tutorial dialog despite substantial ASR errors.

However, given that modern NLP approaches use LLM-based models, it is a critical question as to how different types and profiles of ASR errors affect various discourse models. We investigate this question by examining the accuracy of two state-of-the-art ASR systems (Google [35] and Whisper [70]) and the impact of ASR errors on three typical downstream discourse tasks (§2.3), which helps adaptive CL by knowledge grounding on student needs, identifying off-task talk, and understanding student discourse actions, respectively.

1.1 Background and Related Work

We review relevant organized as (1) domain-agnostic analysis, (2) domain-specific analysis for classroom conversations.

1.1.1 Domain-agnostic Analysis. ASR systems are generally evaluated based on word error rate (WER). This entails aligning the ASR (hypothesis) transcript to a human transcribed (reference) transcript, then counting the number of words missed (deletion), altered (substitution) or inserted relative to the reference. WER is domain-agnostic, in the sense that the quality of transcription is assessed purely by straightforward word-level matching between ground truth and hypothesis. Yet, this may not capture the utility of ASR for downstream applications when perfect transcripts are not needed [25, 76]. Further, WER reported alongside published models may give overly optimistic assessments of real-world performance, as real-world conversations have different characteristics from the isolated, single-party utterances used in these idealized experimental evaluations [92]. As analyzed by [34], there are a host of factors that can impact ASR errors, from acoustic features to lexicon frequency, word length, and part of speech. Regardless of the application domain, further breakdown of ASR errors by such factors can be informative for the design of downstream models. For instance, retaining phonemic features in ASR and passing these to an NLP model can be useful for robustness to pronunciation differences [64]. Domain-specific rescoring of the n-best hypotheses from a domain-general ASR can also yield improvements [56]. In addition to domain-agnostic metrics, many researchers have explored the use of ASR evaluation methods that are more relevant to a particular research domain, for instance by considering the

error rate for domain keywords separately from the remaining vocabulary [60] or by designing metrics that assess how well an ASR output captures higher-order meaning, for example by using word embeddings [52].

1.1.2 Domain-specific Analysis of Classroom Conversations. Child speech is generally associated with higher WER than adult speech due to differences in vocal parameters (fundamental and formant frequency), speech patterns [53], and increased variability relative to adults [14, 31, 66, 81]. The level of background noise typical in a classroom [94, 98] is also an issue, with multiple concurrent speakers resulting in multiparty chatter [69, 80]. There is also a tension between optimal recording conditions (such as individual wired headset microphones on each student) and minimizing invasiveness to the student (favouring distant placement of microphones, multiple speakers per recording channel, and allowing movement of the speakers relative to the recording device) - the latter approach will degrade the signal with documented impacts on ASR [29, 46]. Other factors to consider include the cost and scalability of the microphones and whether this disadvantages certain students in underfunded schools [21].

One efficient domain-specific evaluation method for ASR is to test it on downstream NLP tasks such as speech-based intelligent tutoring systems [20, 51], NLP-based learner modeling [24, 84], and modeling teacher and student talk moves [91]. For example, a recent study compared ASR performance on modeling collaborative problem-solving skills in different recording environments either in a school classroom or laboratory space [68]. Using IBM Watson's cloud-based ASR API, the WER was higher (78%) in the classroom setting than in the quieter lab space (54%). Crucially, the ASR transcripts were used as input to NLP-based models of collaborative problem-solving skills, reporting a 12% drop in performance, though the models still considerably outperformed chance guessing.

Most relevant to the present work is a previous study on ASR in the classroom with the end-goal of modeling CL discourse [85]. As a step toward minimally-invasive tracking of classroom collaboration, Southwell et al. [85] used low-cost tabletop-placed USB microphones to record small-group collaboration, and did transcription with several commercial ASR services. The results indicated very high WER (around 90%, particularly dominated by deletions (i.e., where words were missed by the ASR) and different patterns of errors for different ASRs. The study also found that most of the variance in performance occurred at the utterance level rather than the speaker or recording context. However, this study did not quantify the performance of ASR errors on downstream LLM models, for example, in the above [68] study.

1.2 Current Study and Contributions

We use a dataset of authentic small-group interactions recorded in middle school STEM classrooms using inexpensive, commercially-available equipment. We analyze the ASR errors of two state-of-the-art models (Google Speech and Whisper ASR) with both domain-agnostic and domain-specific approaches. Our contributions are as follows:

First, for domain-agnostic measurements, we conduct a detailed analysis of different ASR errors produced by the Google and Whisper systems. Whereas [85] provided a similar analysis on multiple cloud-based ASRs including Google, to our knowledge this is the first systematic analysis of Whisper on student conversation data in classrooms. We also break down WER according to specific word classes (part of speech or POS-WER), a novel analysis that will be inform understanding of how each word class affects downstream NLP tasks.

Second, for domain-specific modeling, we investigate the impact of ASR errors on three NLP-based tasks for small group discourse modeling in adaptive collaborative learning environments. First we chose an Abstract Meaning Representation parsing task (AMRParsing) [1], which helps representing course materials and group dialog in a unified graph-based representation: enabling explainable and robust personalized content support [4, 44]. Second, we identify Lesson-Focused and Classroom-Focused utterances (OnTask), which can be used to monitor and provide real-time support on unfolding CL discourse [30, 49]. Third, we analyze student utterances with the Accountable Productive Theory (TalkMove) framework focusing on the accountability to content knowledge, rigorous thinking, and the learningcommunity [62, 89–91]. These three tasks also vary with respect to specificity of language (with AMRParsing being most specific). We examine degradation in performance of models trained on human transcripts and tested on ASR data, and importantly, address the critical question of whether there are advantages to training models directly on ASR data and on combining human- and ASR-data, which is also novel in this context. Whereas the previous study [85] investigated ASR errors on downstream NLP, it used semantic similarity and a model of collaborative problem-solving models trained on a different data set, which did not afford systematic quantification of the influence of ASR errors as done here on three different tasks.

Finally, because a large number of utterances are simply not transcribed by the ASRs in this noisy setting (*empty utterances*), we conduct an analysis of how tolerant our LLM-based models are of empty utterances. More importantly, we also simulate ASR corrections based on different word classes based on the POS-WER analysis (see above), and investigate whether this improves performance on downstream user modeling. These results provide insights for future research on how to improve the modeling of CL discourse.

Taken together, to build and deploy adaptive AI-enabled learning environments in real-world noisy classrooms, the present study provides a comprehensive analysis of ASR errors on three NLP-based small group modeling tasks of child speech in authentic classroom environments. It also provides evidence of the feasibility of ASR in this noisy in-the-wild environment, and guidance on developing more robust systems for adaptive collaborative learning environments.

2 METHODS

2.1 Dataset

All procedures were approved by designated Institutional Research Boards and data were only collected from students who provided both personal assent and their parent's signed consent forms.

2.1.1 Data Collection. The data were collected from a US public middle school STEM classroom taught by one teacher in four class periods totaling 30 students in grades 5-8. The subject material was a STEM curriculum unit ("Sensor Immersion") that encourages students to work collaboratively in small groups to program and wire various types of environmental sensors and collect streams of data on their surroundings [11]. For the CL activities, each consenting student group sat around a table with a single Yeti Blue microphone connected to an iPad to record audio and video. The choice of this microphone was influenced by audio quality, cost, power source, form factor, and ease of use. As individual demographic information was not available on the specific students, we report the demographics of the school district as a whole. Ethnicity: 62% White, 30% Hispanic, 3% Asian, 3% two or more races, 1% Black, 0.3% American Indian or Alaska Native, and 0.1% Hawaiian/Pacific Islander. Sex: 49% female, 51% male.

2.1.2 Human Transcription & Automatic Speech Recognition. We selected 31 five-minute samples for analysis containing 30 unique speakers; a speaker could be in multiple samples. Samples were transcribed manually ("human" transcript) by a team of three transcribers. There were 2518 utterances (2179 from students, remainder from teachers). In situations where speaker's identity was clear, but the speech was too indistinct to transcribe, some or all of the utterance content was coded as "[inaudible]". Removing such utterances resulted in 1936 student utterances, which are the subset we use in our domain-agnostic analyses below. Different preprocessing in each downstream task resulted in slightly different numbers of utterances for the domain-specific analysis. All three downstream tasks are based on the same 31 sessions, which are randomly split at the session level into training (17), development (6), and testing (8) (all utterances from a given session are in one of the sets). Audio segments for utterances were extracted using the manually-annotated utterance timestamps and transcribed using two ASR models: (1) a cloud-based commercial provider (Google, video-optimized model configuration) and (2) a pretrained, open-access model trained on 680,000 hours of speech (Whisper, medium size model) [70]. Both the human and ASR transcripts were normalized to facilitate comparison. Specifically, non-word indicators used by the transcribers such as "[inaudible]" and "[shouting]" were stripped out, Numbers were spelled out, punctuation was stripped, words were transposed to lowercase, and hyphens were replaced by spaces.

2.2 Domain Agnostic Analysis

2.2.1 Word Error Rate and Empty Rate. By standard procedures [92], for each utterance, we used the Levenshtein algorithm at the word level to find the minimum edit distance between the reference (human transcript) and the hypothesis (ASR transcript) by optimizing on the minimum **edit operations**: substitution (S), insertion (I) and deletion (D) to align the reference to the hypothesis. Then the WER is computed by WER = $\frac{S+D+I}{N_{\text{ref}}}$, where the N_{ref} is the number of words in the reference human transcript. Finally, we compute the average WER over all utterances. We also analyzed the ASR **empty rate**, that is the the proportion of utterances where the ASR failed to detect any words.

2.2.2 POS-based Word Error Rate. Parts-of-speech (POS) fall into two categories: closed class and open class. Closed classes are those with relatively fixed memberships, such as pronouns since new pronouns are rarely coined. By contrast, nouns and verbs are open classes — new nouns and verbs like *Twitter* or *Google* are continually created. Breaking down the word errors according to POS will help us understand the ASR errors jointly with linguistic characteristics in our dataset. We denote this approach **POS-WER** as it is computed in a similar way to the standard WER. However, there are three differences: (1) We used Spacy [41] to produce the same tokenization for human and ASR transcripts. Critically, it expands contractions such as 'It's' into 'it' and ''s', thus with two separate POS tags for pronoun and auxiliary verbs. (2) Instead of reporting the WER as the mean over the WERs for each utterance, for POS-WER we first group the word-level errors (substitution, deletion, and insertion) by POS across all utterances, then compute the overall error rate for each POS. (3) For each substitution and deletion error, we use the POS of the corresponding aligned word in the reference for grouping, while for each insertion error (where there is no specific reference word to which it is aligned), we use the POS of the inserted word itself. For example, we compute POS-WER for VERB as $\text{WER}_{\text{VERB}} = \frac{S_{\text{ref-VERB}}+D_{\text{ref-VERB}}+I_{\text{hyp-VERB}}}{N_{\text{ref-VERB}}}$, $N_{\text{ref-VERB}}$ is the number of VERBS in the whole corpus of human transcripts.

2.3 Domain-specific Analysis

We focused on three group discourse modeling tasks: AMRParsing, OnTask and TalkMove. AMRParsing offers an utterance-level graph-based semantic analysis, which highlights student needs by extracting suggestions and claims in students' conversations. Matching the student's utterance with the curriculum in a unified AMR graph representation, enables explainable, robust, and personalized content support [44]. OnTask helps detect whether students' discussions lesson-focused or classroom-focused, thus providing timely and adaptive interventions and encouragement as needed. By analyzing student discourse actions with learning-community-focused talk moves, TalkMove can assist teachers and AI agents in helping students have more effective discussions. For the above tasks, we first examine how the models trained on a source setting (e.g., human transcripts) performed when evaluated on a target setting (e.g., ASR transcripts). We used the transfer ratio (TR) [67] to measure the difference in model performance between the source and target setting after adjusting a baseline for each setting ($\text{source}_{\text{base}}$, $\text{target}_{\text{base}}$). We then investigate whether **ASR-augmented training** using either ASR-data alone or combined human-ASR data perform relative to models trained on human transcripts alone. The details of three user modeling tasks as noted below.

$$\text{TR(source, target)} = \frac{target - target_{\text{base}}}{source - source_{\text{base}}} \quad (1)$$

2.3.1 AMRParsing: Knowledge Grounding on Student Needs. AMR represents the core semantics of a sentence using a graph structure,

capturing who did what to whom, when, where, and how [1]. An AMR graph is a directed, acyclic graph with two sets of nodes: concepts and predicates (representing the actions or predicative connections between concepts). Edges connect the nodes to indicate the types of relationships between them. In the first AMR graph shown in Figure 1, the predicate nodes are "confirm-01" and "go-06" and the other nodes are concept nodes. The "ARG0" and "ARG1" labels on the outgoing edges of "go-06" representing the prototypical agent and patient role of the action of "go" are "you" and "basic" respectively. For the ASR transcribed sentence "and then go base", the AMR (middle) will partially dropped. Furthermore, curriculum text can also be parsed into an AMR graph as knowledge base (right-most) and help grounding the understanding of "go basic" to the "tutorial" subgraph (rectangle) of "MakeCode" programming. By systematically constructing a unified graph-based meaning representation from both curriculum and dialog, AMR offers knowledge-grounded support such as answering curriculum-related questions [44] and detecting false claims [100].

AMRParsing refers to a task that automatically produces the AMR graph structure given a sentence (focusing on student utterance parsing in noisy environments). The state-of-the-art systems for AMR parsing are all data-driven and mostly sequence-to-sequence based (seq2seq-based) neural network models. We use the state-of-the-art SPRING AMR parser [3] to parse student utterances, which leverages the power of BART (an LLM for a sequence-to-sequence model) [54]. SPRING takes the tokenized sentence as input and is trained to generate a sequence of AMR tokens. Then a postprocessing module is applied to reconstruct the AMR graph structure. For our use case, we use the off-the-shelf version of SPRING as our baseline, which we finetune with our domain-specific data. Our evaluation is based on Smatch [6, 17], which is the F1 score for the best-effort matching of triples between gold-standard and parse generated AMR graphs, thus measures the similarity between two graphs.

2.3.2 OnTask: Lesson and Classroom Focused Discourse. Students' collaborative discourse often includes interactions that wander from the task at hand, such as chit-chat, singing, or jokes. Such speech, categorized in prior work as "off-task" is of considerable interest to educators and learning scientists [9, 32, 49, 50, 77]. The capability of an automated system [8, 30] to *identify* off-task talk is thus crucial for for understanding and supporting collaborative discourse. Towards this end, we incorporate "off-task utterance classification" [30] in our suite of downstream tasks. Our annotation scheme examines two facets of task-related collaborative dialog, Lesson-focused (LF) speech focuses on discussions about the specific problem or lesson that the students are required to work on, such as discussing ideas and solutions. Classroom-focused (CF) speech is when students discuss any relevant classroom activity, including peripheral tasks such as team management and other procedural information (see Table 1 for examples).

We annotate each utterance as *focused*, *unfocused*, or *unsure*, where a decision cannot be made. Annotations are done separately for the lesson-focused and classroom-focused categories (see Table 1). After developing and refining the guidelines, each transcript is annotated by two annotators from a pool of five annotators with experience with linguistic annotation tasks. The annotators are provided access to the lesson plans and descriptions of the curriculum unit to understand the context of student discussions. We instruct the annotators to evaluate every utterance in the context of the entire transcript, by looking at the past and the future utterances. Statistics of agreement between both annotators were 64.7% and 71.3% respectively for the LF and CF facets. We resolve disagreements as follows: if either of the annotators was unsure, while the other was decisive, we assign the decisive value (focused or unfocused) as the adjudicated label. In cases where one annotator chose focused whereas the other annotator chose unfocused, we manually adjudicate using a third trained annotator.

We use an LLM model RoBERTa [57] by adopting the base implementation from the Huggingface Transformers library [97] with the default hyperparameters of RoBERTa-base (i.e., an embedding size of 512 and a hidden layer size of 768). We use a dropout probability of 0.1 on the attention layers and the hidden layers. We train for 50 epochs, with early stopping based on the F1 score on the development set. We use the AdamW optimizer [47] and a learning rate of 1e-5. We focus on the F1 score pertaining to detecting the LF and CF classes as our outcome metric. Our best models on human data use 5 previous utterances as the dialogue context to predict the label for the current utterance.

2.3.3 TalkMove: Understanding Student Discourse Actions. We also analyze students' discourse based on Academic Productive Talk theory (TalkMove) [61]. This discourse framework emphasizes high-impact "talk moves" of either a student or teacher in relation to knowledge-building discourse [62]. Student talk moves are discourse actions such as making claims, using reasoning, relating to other students' actions, and asking questions. Understanding how the students are actively and equitably engaged in challenging academic work can orchestrate personalized [90] and equitable learning [89] environment. The task TalkMove is to automatically classify each student utterance into a talk move label as shown in Table 2. Prior works on predicting teacher and student talk moves [42, 78, 89], have yet to evaluate the impact of ASR errors on these models.

We mainly focus on the four student talk moves as shown in Table 2. There is a 'None' label for those utterances that do not correspond to any talk moves and a 'Not Enough Context' label for those utterances that human annotators deemed not classifiable given limited context. The talk moves are consensus coded by two trained coders. [1] Given the immense class imbalance as shown in Table 2, we merged the talk moves into three categories (Learning Community [LC], None, or Other) and the focus is on detecting LC moves.

To make a fair cross-task comparison on them, we use the same model architecture as our OnTask model and the same input preprocessing, dialog context modelling and training setup. Here, our best model setting is with window size 6 (instead of window size 5 in OnTask). We report the macro F1 score over all three labels.

[1] For utterances that contain multiple sentences, we annotate the talk moves for each sentence first, and use the majority code as the label for the whole utterance, when there are multiple sentences in the utterance.

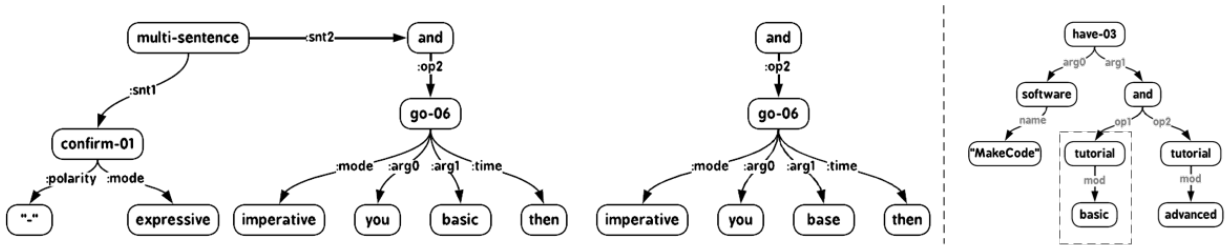


Figure 1: AMR for human transcribed utterance "No, no, no, no, no. And then go basic" (left-most), AMR for ASR transcribed utterance "and then go base" (middle), and AMR for curriculum text "MakeCode has basic and advanced tutorials." (right-most).

Table 1: Examples of utterances and label distribution for the Lesson-focused facet and the Classroom-focused facet in training data.

Category	Description	Label	Counts	Example
Lesson Focused	Whether related to the specific class room **topic** that the students are working on, including discussions about concepts, or about the artifact they are working with.	✓	723	This is actually a different button.
		✗	198	What are you doing over there [partner]?
		unsure	25	We should [inaudible].
Classroom Focused	Any relevant classroom activity, including peripheral tasks like introductions, preparations to work on an exercise, team management	✓	823	Teacher, their computer shut off.
		✗	198	Do your dino singing.
		unsure	25	"Here","Do you mean [inaudible]"

Table 2: Student Talk Moves Included in the training data.

Label	Category	Original TalkMove	Description	Counts	Example
NONE	None	None	Not a Talk Move	299	"OK","Alright","Let's do the next step."
LC	Learning Community	Relating to another student	Using, commenting on, or asking questions about a classmate's ideas	512	"My bad", "Press the button","You need to code that"
	Learning Community	Asking for more info	Student requests more info,says they are confused or need help	3	"I don't understand number four."
OTHER	Content Knowledge	Making a claim	Student makes a math claim, factual statement,or lists a step in their answer	41	"We should place the wire on P2.","We could do a winky face next."
	Rigorous Thinking	Providing evidence or reasoning	Student explains their thinking,provides evidence,or talks about their reasoning	1	"Because that's how loud our class usually is."
	N/A	Not Enough Context	The context is not enough to categorize the talk move	139	"Here","Do you mean [inaudible]"

2.4 Simulation-based Analysis

2.4.1 Simulating Empty Utterances. AMRParsing is an utterance-level parsing task, where each utterance u_i is parsed without considering the dialogue context. In contrast, OnTask and TalkMove classify each utterance u_i given a sequence of previous dialogue histories $\{u_{i-w}, u_{i-1}\}$ with a window size $w = 5$ or 6 respectively. Empty utterances may exist in the dialogue history, but it is unclear how they impact the classification of non-empty utterances. We therefore design comparative studies for two questions: **(1) How does context help each task?** Based on full human transcripts, we compare our best model for each task with a model using no context to predict each utterance u_i. **(2) How does the lack of context affect the models' accuracy?** We simulate 5 test sets with different empty utterance rates [0.04, 0.08,0.10, 0.30, 0.50] on human transcripts, then we compare the evaluation performance on the simulated datasets with the main results on the original human transcripts (where the empty rate is 0).

In total, we evaluate our best models trained on human transcripts over all 6 empty rate settings, [0, 0.04, 0.08,0.10, 0.30, 0.50]. For each empty rate, we generate 3 datasets by randomly masking out the empty utterances in the human transcripts, then we report the average performance over the 3 runs, and summarize the performance changes compared to their corresponding baselines.

2.4.2 Simulating Error-fixes based on POS tags. Rather than simulating ASR errors from clean human text for data-augmented training [20, 27, 28, 82, 88], we mainly investigate what kinds of improvements on ASR can help our LLM-based models. Hence, we simulate how to improve the ASR by incrementally fixing different kinds of word errors (with respect to POS) until utterances are fully restored to their human transcription source.

When calculating the WER between ASR and Human transcripts (§2.2.1), the edit distance algorithm produces word alignment and the edit operations needed to *edit the human utterance into the ASR utterance*. In order to simulate various improvements to ASR, we reverse those edit operations to *edit the ASR utterance back to the human utterance*. For the pair of human and ASR utterances in Figure 1, there are 5 *deletion* errors of missing "No"(with a POS as INTJ), and one *substitution* error of replacing "basic"(ADVERB) with "base"(NOUN). First, when fixing the word errors related to POS ADVERB, the ASR utterance will become "and then go basic" by *substituting* the wrong word "base' with"basic". By fixing the word errors related to both ADVERB and INTJ, the ASR utterance will be recovered to the original human utterance by *inserting* five "No's" in front.

According to Table 4, PRON, VERB, NOUN, INTJ, AUX are the top 5 most frequent word classes in our dataset. To simplify the analysis, we merged AUX-VERB with VERB together as "VERB"

for the simulation studies. Starting with Google or Whisper ASR, we simulate the error fixes for each word class one-by-one until we have recovered the corresponding human transcript. For the produced test set after each fix, we calculate its WER, and also evaluate it with our best model trained on human transcripts.

3 RESULTS

3.1 Domain Agnostic Analysis - Speech Recognition Errors

3.1.1 Word Error Rate and Empty Rate. Overall, word error rates are comparable across ASR engines and very high: 82.2% for Google and 84.5% for Whisper (Table 3). However, the two ASRs had rather different patterns of errors with Google being more conservative (Deletions > Substitutions > Insertions) compared to Whisper (Substitutions > Insertions > Deletions). As a result, Google produces an empty ASR rate of 37.7% whereas Whisper has an empty rate of just 12.1%. On the subset of utterances with non-empty ASR, Whisper has a higher WER (82.5%, n=1701) than Google (71.4%, n=1206), again suggestive of a tradeoff between the two.

3.1.2 Word Error Breakdown by Part Of Speech. Table 4 breaks down word errors by each POS. We noted the following patterns in the data. First, POS-WERs vary more widely for open classes (range 0.65 to 1.04 on Google, and 0.52 to 1.27 on Whisper) compared to closed classes (range 0.56 to 0.77 on Google, and 0.52 to 0.68 on Whisper)). Second, PRON (pronouns), VERB, AUX (auxiliary verbs), NOUN, and INTJ (interjections) are the five most frequent word classes in the recordings. Among them, Google ASR has the highest POS-WER (0.82) on INTJ, while Whisper's POS-WER is 0.64 for INTJ. The two are similar for AUX, whereas Google has higher POS-WERs for PRON, VERB, and NOUN. Third, we observe the same finding as in standard WER, across all POS tags, namely that Google has a relatively higher deletion rate, while Whisper has a higher insertion rate. Note that POS-WER for Whisper is on average lower than for Google, in contrast to the overall WER; this is because POS-WER is calculated using different text normalization and aggregated at the corpus-level instead of averaged over utterances.

3.2 Results on Downstream Tasks

3.2.1 Training on Human Data. The transfer ratio as shown in Equation 1 compares the relative change as shown below. Thus, a TR of 1 would indicate perfect transfer whereas a TR of 0.5 would indicate a 50% reduction in performance across settings. It is also possible for TRs to exceed 1 in cases where performance is higher in the target compared to the source setting. Table 5 shows the performance (Smatch) for AMRParsing and F1 scores for the other two tasks along with the transfer ratios. The top panel presents results for baselines for each task. Baselines for OnTask and TalkMove are obtained by averaging the performance on 10 random runs (by uniformly predicting a random label), while the baselines for AMRParsing use the off-the-shelf SPRING parser [3] without finetuning on our dataset. Note that baselines vary due to the removal of empty utterances in the test set when computing scores. The bottom panel shows model results along with the transfer ratio (TR) in parentheses. Human indicates human transcripts. For Google, we use Google ASR, but keep the original ground truth labels and include all utterances (i.e., even empty utterances) for fair comparisons. However, in a real-world application, we wouldn't have the human transcripts to align the ASR results to, so the empty utterances would not be included. Hence, we add a real-world setting called $\text{Google}_{\text{FILTER}}$, which filtered out those empty utterances. We use the same setting for Whisper: Whisper and $\text{Whisper}_{\text{FILTER}}$. After this part, we will only consider the real-world $\text{Google}_{\text{FILTER}}$ and $\text{Whisper}_{\text{FILTER}}$ for the remaining experiments.

We found that for AMRParsing, the models finetuned on human data degraded to the performance similar to off-the-shelf baseline when evaluated on Google ASR. Overall transfer ratios were low (0.39-0.46 for Google, and 0.62 on Whisper). Results were more positive for the other two downstream tasks with above-chance performance in all cases and transfer ratios as high as 0.94. Overall, as could be expected, TRs were higher for cases where empty utterances were filtered out. There was also considerable variance by task and ASR. Whereas, the ASRs were quite similar for Lesson-Focused OnTask predictions, Whisper yielded similar transfer ratios (.59 to .72) than Google (.64-.72) for TalkMove, whereas Google had much higher transfer ratios (.88 to .94) for Classroom-Focused OnTask predictions.

3.2.2 Results on ASR-Augmented Training. Here, we focus on the relative performance change of ASR-augmented training compared to training on human transcriptions reported above. As the four bars for each task in Figure 2, the blue and red ones mean the same model trained on clean human transcripts but evaluated on human and ASR transcripts respectively (where performance dropped drastically); the yellow and green means two models trained on ASR only and combined human-ASR respectively, while both evaluated in the real-world ASR setting. For AMR parsing, both ASR-only and combined human-ASR training under-perform models trained on human data for both Google and Whisper. This result likely occurred because using ASR utterances with human data for training confuses the seq2seq-based models with respect to the token alignments between the source and target sequences. Take Figure 1 as an example, if the model is trained to forcibly map the source sequence (ASR utterance) "and then go base" into the ground-truth graph on the left, then the left branch of that AMR graph will have no mapping in the source.

Findings for the discourse level tasks (OnTask and TalkMove) were more promising. There was always a setting (yellow or green) that outperformed training on human transcripts alone (red). Furthermore, the models trained only on Whisper (yellow on Figure 2b) consistently improve performance on both OnTask (4.65%,10.81%) and TalkMove (3.26%), while training on Google was largely unsuccessful. On the other hand, combined human-ASR training(green) also sometimes outperforms the human transcripts along models (red), but is not consistent. Overall, models trained only on Whisper (yellow on Figure 2b) will lead to robust performance in our cases.

3.3 Results on Simulation-based Analysis

3.3.1 Impact from Empty Utterances in the Context. As shown in Table 6, all reported performances are based on our best models trained on human transcripts. Each column denotes a setting of different

Table 3: Metrics for ASR errors: mean values computed over student utterances (n=1936). We report two rows of word error rate results for All utterances(the first row) and Non-Empty utterances(the second row).

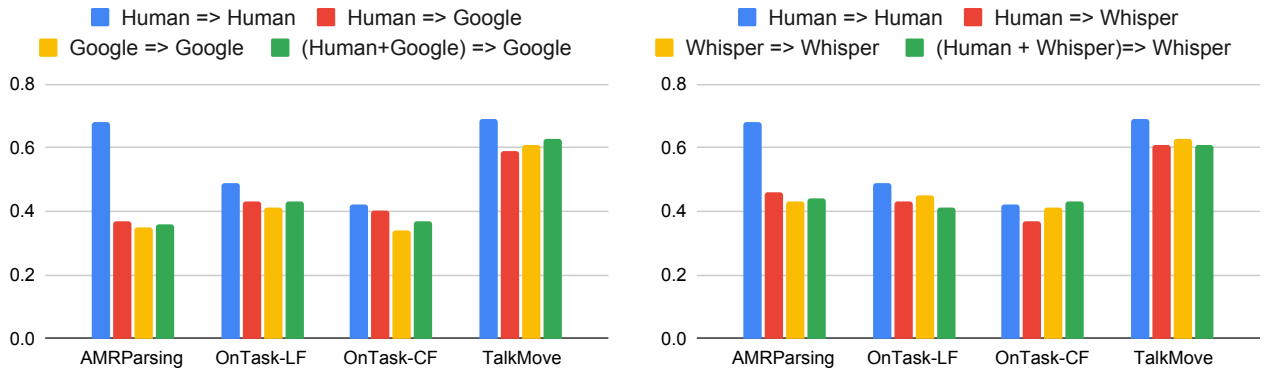
ASR	Empty rate		Substitution	Deletion	Insertion	WER
Google	0.377	All	0.209	0.552	0.062	0.822
		Non-Empty	0.335	0.280	0.099	0.714
Whisper	0.121	All	0.349	0.216	0.282	0.846
		Non-Empty	0.397	0.108	0.321	0.825

Table 4: Word errors broken down by POS for Google and Whisper ASR. Bold numbers show the Top five POS in our dataset.

POS	Counts	Google				Whisper			
		POS-WER	Substitution	Deletion	Insertion	POS-WER	Substitution	Deletion	Insertion
Open Classes									
ADJ	402	0.65	0.26	0.37	0.02	0.62	0.31	0.14	0.17
ADV	845	0.67	0.18	0.47	0.03	0.52	0.22	0.19	0.10
INTJ	**973**	**0.82**	0.12	0.68	0.01	**0.64**	0.30	0.27	0.08
NOUN	**1111**	**0.74**	0.30	0.41	0.02	**0.71**	0.38	0.15	0.17
PROPN	76	1.04	0.36	0.46	0.22	1.27	0.54	0.24	0.50
VERB	**1719**	**0.68**	0.21	0.43	0.04	**0.61**	0.27	0.17	0.17
Closed Classes									
ADP	520	0.635	0.22	0.38	0.04	0.56	0.24	0.15	0.17
AUX	**1014**	**0.64**	0.16	0.47	0.04	**0.64**	0.21	0.18	0.25
CCONJ	247	0.717	0.17	0.50	0.05	0.53	0.20	0.19	0.13
DET	490	0.561	0.16	0.38	0.03	0.52	0.21	0.13	0.18
NUM	267	0.772	0.35	0.42	0.01	0.68	0.36	0.23	0.08
PART	420	0.657	0.22	0.39	0.05	0.56	0.20	0.13	0.24
PRON	**2074**	**0.65**	0.15	0.46	0.04	**0.58**	0.20	0.19	0.20
SCONJ	231	0.68	0.17	0.47	0.05	0.57	0.26	0.18	0.13

Table 5: Results of Comparative Study on ASR Evaluation for AMRPARSING, ONTASK and TALKMOVE tasks.

Testing	AMRPARSING	ONTASK		TALKMOVE
	Smatch (TR)	Lesson-Focused F1 (TR)	Classroom-Focused F1 (TR)	Learning Community F1 (TR)
Base HUMAN	0.55	0.27	0.26	0.32
Base GOOGLE$_{FILTER}$	0.31	0.27	0.26	0.32
Base GOOGLE	0.26	0.27	0.26	0.31
Base WHISPER$_{FILTER}$	0.38	0.27	0.27	0.32
Base WHISPER	0.36	0.27	0.27	0.31
HUMAN	0.68 (1.00)	0.49 (1.00)	0.42 (1.00)	0.69 (1.00)
GOOGLE$_{FILTER}$	0.37 (0.46)	0.43 (0.73)	0.40 (0.88)	0.59 (0.72)
GOOGLE	0.31 (0.39)	0.42 (0.68)	0.41 (0.94)	0.54 (0.64)
WHISPER$_{FILTER}$	0.46 (0.62)	0.43 (0.72)	0.37 (0.63)	0.59 (0.72)
WHISPER	0.44 (0.62)	0.41 (0.64)	0.37 (0.63)	0.53 (0.59)



(a) ASR-Augmented Training with Google ASR

(b) ASR-Augmented Training with Whisper ASR

Figure 2: Main Results for ASR-Augmented Training on AMRPARSING, ONTASK (Lesson-Focused, Classroom-Focused) and TALKMOVE. In the legend, "A=>B" means the model is trained on A, and evaluated on B.

empty rates, the performances are averaged over 3 random runs. Our answers to the two questions (§2.4.1) are (1) compared to the models without using any dialog context (Human^0), our best models using context improve 4.55% on TALKMOVE, but don't consistently help on ONTASK. (2) empty utterances in the dialog context

Table 6: Analysis of the impact of empty utterances in the dialogue context. The numbers in parentheses are the relative performance changes compared to their corresponding baselines.

Empty rate		0	0.04	0.08	0.1	0.3	0.5
Baseline (%)		Human^0	Human				
OnTask	LF F1	0.49 (2.08)	0.48 (-1.84)	0.48 (-2.04)	0.48 (-2.04)	0.46 (-6.12)	0.44 (-10.20)
	CF F1	0.42 (-2.32)	0.41 (-2.38)	0.41 (-2.38)	0.44 (4.76)	0.41 (-2.38)	0.39 (-7.14)
TalkMove	F1	0.69 (4.55)	0.69 (0.00)	0.69 (0.00)	0.68 (-1.45)	0.65 (-5.80)	0.67 (-2.90)

only slightly hurt the performance when empty rates are less than 0.3 (6.12%, 2.38% and 5.80% drop on OnTask Lesson-Focused F1, OnTask Classroom-Focused F1, and TalkMove F1 respectively). Hence, we conclude that on both discourse-level tasks, the dialog context contributes much less than the current utterance for our classification, and TalkMove is more context-sensitive than OnTask. This simulation also indicates that the main performance drop when transferring from human to ASR is due to the word errors in the current utterances rather than the empty utterances in the dialog context.

3.3.2 Error Tolerance on Different Word Classes. In Table 7, the first row shows the source ASR datasets (Google and Whisper) where the simulation started from. The last row is the human test set with all errors fixed ideally. The second last row is the resultant test dataset after we fixed errors according to all POS (Ideally, reversing the edit options calculated in edit distance should exactly recover to zero WER (the same as the original human transcripts). However, due to the text normalization issue, e.g., one word 'it's' or 'gotta' may be split into two tokens for POS. Then after full fixing, some utterances still have around 1 or 2 word errors). Vertically, Table 7 gradually shows how WER and the model performance will change as the fixes continue. The WERs of Google and Whisper almost go down. Gradually fixing the word errors generally helps on our three tasks, and also suggests the following findings: (1) comparing two ASRs, we noticed that simulated error fixes help more on Google than Whisper. On Whisper, only when jointly fix VERB, PRON, NOUN, and INTJ will help all tasks. Before that, the error fixes by individual POS didn't help TalkMove. (2) comparing across tasks, continual improving WER will largely help on our lexically sensitive AMRParsing tasks, while not always helping on discourse tasks. The three tasks also benefit differently in an order (AMRParsing > OnTask > TalkMove). For example, fixing the word errors from 0.77 to 0.31 on Google, and 0.75 to 0.31 on Whisper (consider the row of jointly fixing VERB, NOUN, INTJ and PRON), this will help most on AMRParsing performance on Google and Whisper by (74.19%, 31.83%) respectively, while helping relatively less on OnTask and TalkMove. (3) comparing across different POS sets, jointly fixing VERB and NOUN robustly improves model performance for all tasks. However, only fixing individual POS may not help on some tasks. For example, although frequent, INTJ (e.g., Ouch! Hi! Oh!) contributes less than VERB in all three tasks. While PROUN helps on AMRParsing and OnTask, but not on TalkMove. However, when fixing PRON and INTJ jointly with VERB and NOUN, the performance boosts more on all tasks.

4 DISCUSSION

We investigated the impact of the noisy acoustic environment of collaborative group work in a K-12 classroom on downstream NLP-based group modeling. Using two state-of-the-art ASR systems (Google and Whisper) to transcribe the audio, we conducted three extensive comparative studies: (1) domain-agnostic analysis, (2) domain-specific analysis for classroom conversations, and (3) simulation-based analysis.

Main Findings: We found that both Google and Whisper suffer from high WER (0.822, 0.847) for student speech in the classroom, which replicates similar findings in previous work [67, 85]. We also found that Google is more conservative than Whisper in transcribing speech, thus returning more empty transcripts. Under this noisy classroom environment, we built models for three group modeling tasks (AMRParsing, OnTask and TalkMove) based on LLMs. Although we expect that models trained on human transcripts will result in a performance drop when evaluated on noisy ASR transcripts, the pertinent issue is to quantify the amount of performance reduction. Here, when evaluating on ASR transcripts, we found that the models were more robust for predicting the abstract discourse-level analysis (i.e., OnTask and TalkMove), but performance on AMRParsing dropped drastically to the performance of the off-the-shelf model. The transfer rate of AMRParsing (.39 - .62) is much lower than that of the abstract discourse-level OnTask (.63- .94) and TalkMove tasks (.64-.72). Since the real-world deployment of group modeling in classrooms would necessarily only have ASR input, we then asked if augmenting the training data with ASR transcripts would improve generalization. We found that simply augmenting the training data with ASR transcripts did not improve performance on the lexically-sensitive AMRParsing, but did for the discourse level tasks. Overall, the results do suggest potential advantages to training models directly on ASR data and a combination of human-ASR data. Finally, simulation-based analysis on empty utterances shows that our LLMs are robust to the missing context in OnTask and TalkMove tasks, while the main performance drops when transferring from human to ASR are due to the word errors. Fixing important word classes (e.g., verbs and nouns) can robustly help over all the tasks. The simulated error fixes help more on Google than on Whisper, and more on lexically sensitive AMRParsing task than abstract discourse-level tasks.

Limitations and Future Work A limitation of this work is that we only compared two off-the-shelf state-of-art ASRs: Google and Whisper. Including other existing ASR models (such as Watson, REV) will increase the diversity and coverage of our analysis. Another limitation lies with the LLMs we used for our three applications. We mainly rely on Roberta(base) for OnTask and TalkMove, however, it is known that various larger LLMs with emergent abilities may lead to different comparison results [45, 63]. Besides that, another limitation is the ASR-augmented training strategies. When considering the joint strategy of using both human and ASR, we always use the full human and ASR datasets. Better strategies for mixing human and ASR datasets still require further investigation.

Table 7: Simulating error-fixes based on parts-of-speech. Bold numbers highlight the error fixes that lead to performance boosts.

	Google					Whisper				
	WER	AMRParsing	OnTask		TalkMove	WER	AMRParsing	OnTask		TalkMove
		Smatch	LF F1	CF F1	F1		Smatch	LF F1	CF F1	F1
ASR	0.77	0.31	0.42	0.41	0.54	0.75	0.44	0.41	0.37	0.60
+ PRON	0.68	0.34	0.49	0.43	0.58	0.68	0.45	0.44	0.42	0.56
+ VERB	0.65	0.41	0.45	0.41	0.57	0.62	0.48	0.46	0.44	0.58
+ NOUN	0.66	0.34	0.46	0.42	**0.61**	0.68	0.45	0.49	0.42	0.59
+ INTJ	0.64	0.37	0.44	0.42	0.55	0.68	0.45	0.43	0.39	0.56
+ PRON + INTJ	0.54	0.40	**0.49**	**0.44**	0.57	0.59	**0.48**	0.45	0.40	0.56
+ VERB + NOUN	0.54	**0.44**	**0.46**	**0.43**	**0.61**	0.53	**0.51**	**0.49**	**0.47**	**0.62**
+ VERB + NOUN + INTJ	0.41	0.49	0.45	0.46	0.62	0.44	0.54	0.50	0.43	0.63
+ VERB + NOUN + PRON	0.45	0.50	0.47	0.47	0.63	0.40	0.55	0.49	0.47	0.63
+ VERB + NOUN + INTJ + PRON	0.31	**0.54**	**0.52**	**0.45**	**0.64**	0.31	**0.58**	**0.52**	**0.43**	**0.64**
+ All OPEN	0.33	0.52	0.48	**0.48**	0.66	0.39	0.56	0.53	0.46	0.64
+ ALL CLOSE	0.56	0.40	0.48	0.45	0.57	0.52	0.48	0.46	0.40	0.57
+ ALL	0.11	**0.65**	0.49	0.42	0.65	0.10	0.66	0.49	0.46	0.66
Human	0.00	0.68	0.49	0.42	0.69	0.00	0.68	0.49	0.42	0.69

Potential Applications: This work supports the provision of adaptive collaborative learning in real-world noisy classrooms, enabling teachers and AI agents to facilitate more effective collaborative learning experiences. Our empirical analysis suggests that despite considerable ASR errors in noisy classroom discourse, LLMs for coarse-grained discourse-level tasks (e.g., OnTask and TalkMove) that do not require perfect transcriptions, can be used for group modeling, and there might be benefits to adopting training strategies that incorporate noisy ASR data. However, further improvements are still needed for fine-grained knowledge supporting tasks (AMRParsing) under such a noisy environment. Finally, beyond the standalone analysis on our three NLP-based group modeling tasks, combining them with well-designed adaptive interventions [2], encouragements [71], content support [99], and creative brain-storming [10] can help facilitate more equitable, engaging, and effective collaborative learning experiences [5, 7, 59, 75].

5 CONCLUSION

Effective learning environments requires adapting to different groups and individuals, which needs accurate discourse analysis to describe diverse behaviors. A key obstacle of accurate discourse analysis is ASR errors in the real-world. We conducted three extensive comparative studies of ASR errors: (1) a domain-agnostic analysis of transcription errors, (2) domain-specific analysis on downstream NLP tasks (AMRParsing , OnTask and TalkMove) and (3) a simulation-based analysis to demonstrate the impact of empty utterances and what kinds of improvements on ASR will help our user modeling in the small-group classroom. Our results illustrated detailed ASR errors in this classroom setting with respect to empty rate, WER, and a novel POS-WER, which characterizes how each word class affects downstream NLP tasks. Then we show thorough examinations on deploying LLM-based models trained on clean human transcripts in noisy ASR setting. Our models yielded degraded performance on all three tasks, while varied on diverse transfer rates. Further, we found promising results using ASR-augmented training data for improving the performance of NLP models on ASR transcripts. Finally, we investigated how our LLM-based models tolerated to simulation-based errors. Overall, our paper demonstrates the characteristics and challenges of group discourse modeling for collaborative learning environments, especially in real-world noisy classrooms, and also provide practical guidance of using imperfect ASR on different tasks.

ACKNOWLEDGMENTS

This research was supported by the NSF National AI Institute for Student-AI Teaming (iSAT) under grant DRL 2019805. The opinions expressed are those of the authors and do not represent views of the NSF. We thank members of iSAT and Institute of Cognitive Science, especially Charis Harty and Brandon Booth, for their valuable insights and suggestions; and reviewers for helpful comments and corrections.

REFERENCES

[1] Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract Meaning Representation for Sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*. Association for Computational Linguistics, Sofia, Bulgaria, 178–186. https://aclanthology.org/W13-2322

[2] Alayne Benson. 2023. The Future of AI in Education: AI Classroom Partners. *XRDS* 29, 3 (apr 2023), 30–35. https://doi.org/10.1145/3589646

[3] Michele Bevilacqua, Rexhina Blloshmi, and Roberto Navigli. 2021. One SPRING to Rule Them Both: Symmetric AMR Semantic Parsing and Generation without a Complex Pipeline. *Proceedings of the AAAI Conference on Artificial Intelligence* 35, 14 (May 2021), 12564–12573. https://ojs.aaai.org/index.php/AAAI/article/view/17489

[4] Claire N Bonial, Lucia Donatelli, Jessica Ervin, and Clare R Voss. 2019. Abstract meaning representation for human-robot dialogue. *Proceedings of the Society for Computation in Linguistics* 2, 1 (2019), 236–246.

[5] Mariah Bradford, Ibrahim Khebour, Nathaniel Blanchard, and Nikhil Krishnaswamy. 2023. Automatic Detection of Collaborative States in Small Groups Using Multimodal Features. *In Proceedings of the 24th International Conference on Artificial Intelligence in Education* (2023).

[6] Shu Cai and Kevin Knight. 2013. Smatch: an evaluation metric for semantic feature structures. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 748–752.

[7] Jie Cao, Rachel Dickler, Marie Grace, Jeffrey B Bush, Alessandro Roncone, Leanne M Hirshfield, Marilyn A Walker, and Martha S Palmer. 2023. Designing an AI Partner for Jigsaw classrooms. *Workshop on Language-Based AI Agent Interaction with Children (AIAIC'2023)*.

[8] Dan Carpenter, Andrew Emerson, Bradford W Mott, Asmalina Saleh, Krista D Glazewski, Cindy E Hmelo-Silver, and James C Lester. 2020. Detecting off-task behavior from student dialogue in game-based collaborative learning. In *International Conference on Artificial Intelligence in Education*. Springer, 55–66.

[9] Suleyman Cetintas, Luo Si, Yan Ping Ping Xin, and Casey Hord. 2009. Automatic detection of off-task behaviors in intelligent tutoring systems with machine learning techniques. *IEEE Transactions on Learning Technologies* 3, 3 (2009), 228–236.

[10] Natawee Chaijum. 2020. Using Brainstorming through Social Media to Promote Engineering Students' Teamwork Skills. *European Journal of Science and Mathematics Education* 8, 4 (2020), 170–176.

[11] Alexandra Gendreau Chakarov, Quentin Biddy, Colin Hennessy Elliott, and Mimi Recker. 2021. The Data Sensor Hub (DaSH): A Physical Computing System to Support Middle School Inquiry Science Instruction. *Sensors (Basel, Switzerland)* 21, 18 (2021), 6243. https://doi.org/10.3390/s21186243

[12] Zhuang Chen and Tieyun Qian. 2020. Relation-aware collaborative learning for unified aspect-based sentiment analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 3685–3694.

[13] Elizabeth G Cohen. 1994. Restructuring the classroom: Conditions for productive small groups. *Review of educational research* 64, 1 (1994), 1–35.

[14] Rosanna Costaguta, Daniela Missio, and Pablo Santana-Mansilla. 2019. A Preliminary Analysis of Gender and Team Roles in Forum Interactions. In *Proceedings of the XX International Conference on Human Computer Interaction*. 1–2.

[15] National Research Council et al. 2000. *How people learn: Brain, mind, experience, and school: Expanded edition*. National Academies Press.

[16] Petru L Curşeu and Helen Pluut. 2013. Student groups as learning entities: The effect of group diversity and teamwork quality on groups' cognitive complexity. *Studies in Higher Education* 38, 1 (2013), 87–103.

[17] Marco Damonte, Shay B. Cohen, and Giorgio Satta. 2017. An Incremental Parser for Abstract Meaning Representation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. Association for Computational Linguistics, Valencia, Spain, 536–546. https://aclanthology.org/E17-1051

[18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 4171–4186.

[19] Pierre Dillenbourg. 1999. What do you mean by collaborative learning?

[20] Sidney K. D'Mello, Art Graesser, and Brandon King. 2010. Toward Spoken Human–Computer Tutorial Dialogues. *Human–Computer Interaction* 25, 4 (Nov. 2010), 289–323. https://doi.org/10.1080/07370024.2010.499850 Publisher: Taylor & Francis _eprint: https://www.tandfonline.com/doi/pdf/10.1080/07370024.2010.499850.

[21] Sidney K. D'Mello, Andrew M. Olney, Nathan Blanchard, Borhan Samei, Xiaoyi Sun, Brooke Ward, and Sean Kelly. 2015. Multimodal Capture of Teacher-Student Interactions for Automated Dialogic Analysis in Live Classrooms. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction* (Seattle, Washington, USA) *(ICMI '15)*. Association for Computing Machinery, New York, NY, USA, 557–566. https://doi.org/10.1145/2818346.2830602

[22] Nia M Dowell, Whitney L Cade, Yla Tausczik, James Pennebaker, and Arthur C Graesser. 2014. What works: Creating adaptive and intelligent systems for collaborative learning support. In *Intelligent Tutoring Systems: 12th International Conference, ITS 2014, Honolulu, HI, USA, June 5-9, 2014. Proceedings 12*. Springer, 124–133.

[23] Gregory Dyke, Iris Howley, David Adamson, Rohit Kumar, and Carolyn Penstein Rosé. [n. d.]. Towards academically productive talk supported by conversational agents. In *Proceedings of the International Conference on Intelligent Tutoring Systems*. Springer, 459–476.

[24] Michael Eagle, Albert Corbett, John Stamper, Bruce M McLaren, Ryan Baker, Angela Wagner, Benjamin MacLaren, and Aaron Mitchell. 2016. Predicting individual differences for learner modeling in intelligent tutors from previous learner activities. In *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization*. 55–63.

[25] Benoit Favre, Kyla Cheung, Siavash Kazemian, Adam Lee, Yang Liu, Cosmin Munteanu, Ani Nenkova, Dennis Ochei, Gerald Penn, Stephen Tratz, et al. 2013. Automatic human utility evaluation of ASR systems: Does WER really predict performance?. In *INTERSPEECH*. 3463–3467.

[26] Michael Flor, Su-Youn Yoon, Jiangang Hao, Lei Liu, and Alina von Davier. 2016. Automated classification of collaborative problem solving interactions in simulated science tasks. In *Proceedings of the 11th workshop on innovative use of NLP for building educational applications*. 31–41.

[27] Peter W Foltz, Darrell Laham, and Marcia Derr. 2003. Automated speech recognition for modeling team performance. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 47. SAGE Publications Sage CA: Los Angeles, CA, 673–677.

[28] Eric Fosler-Lussier, Ingunn Amdal, and Hong-Kwang Jeff Kuo. 2002. On the road to improved lexical confusability metrics. In *ISCA Tutorial and Research Workshop (ITRW) on Pronunciation Modeling and Lexicon Adaptation for Spoken Language Technology*.

[29] Hannes Gamper, Dimitra Emmanouilidou, Sebastian Braun, and Ivan J Tashev. 2020. Predicting word error rate for reverberant speech. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 491–495.

[30] Ananya Ganesh, Michael Change, Rachel Dickler, Michael Regan, Jon Cai, Kristin Wright-Bettner, James Pustejovsky, James Martin, Jeff Flanigan, Martha Palmer, and Katharina Kann. 2023. Navigating Wanderland: Highlighting Off-Task Discussions in Classrooms. *In Proceedings of the 24th International Conference on Artificial Intelligence in Education* (2023).

[31] Matteo Gerosa, Diego Giuliani, Shrikanth Narayanan, and Alexandros Potamianos. 2009. A review of ASR technologies for children's speech. In *Proceedings of the 2nd Workshop on Child, Computer and Interaction*. 1–8.

[32] Karrie E Godwin, Ma V Almeda, Howard Seltman, Shimin Kai, Mandi D Skerbetz, Ryan S Baker, and Anna V Fisher. 2016. Off-task behavior in elementary school children. *Learning and Instruction* 44 (2016), 128–143.

[33] Anuradha Gokhale. 1995. Collaborative learning enhances critical thinking. *Journal of Technology education* 7, 1 (1995).

[34] Sharon Goldwater, Dan Jurafsky, and Christopher D Manning. 2010. Which words are hard to recognize? Prosodic, lexical, and disfluency factors that increase speech recognition error rates. *Speech Communication* 52, 3 (2010), 181–200.

[35] Google. 2023. Google Speech-to-Text. https://cloud.google.com/speech-to-text/. [Online; accessed 20-Jan-2022].

[36] Arthur C Graesser, Samuel Greiff, Matthias Stadler, and Keith T Shubeck. 2020. Collaboration in the 21st century:: The theory, assessment, and teaching of collaborative problem solving.

[37] Arthur C Graesser, Natalie K Person, and Joseph P Magliano. 1995. Collaborative dialogue patterns in naturalistic one-to-one tutoring. *Applied cognitive psychology* 9, 6 (1995), 495–522.

[38] Jiangang Hao, Lei Chen, Michael Flor, Lei Liu, and Alina A von Davier. 2017. CPS-Rater: Automated sequential annotation for conversations in collaborative problem-solving activities. *ETS Research Report Series* 2017, 1 (2017), 1–9.

[39] David A Harrison and Katherine J Klein. 2007. What's the difference? Diversity constructs as separation, variety, or disparity in organizations. *Academy of management review* 32, 4 (2007), 1199–1228.

[40] Cindy E Hmelo-Silver and Howard S Barrows. 2008. Facilitating collaborative knowledge building. *Cognition and instruction* 26, 1 (2008), 48–94.

[41] Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. (2017). To appear.

[42] Jennifer Jacobs, Karla Scornavacco, Charis Harty, Abhijit Suresh, Vivian Lai, and Tamara Sumner. 2022. Promoting rich discussions in mathematics classrooms: Using personalized, automated feedback to support reflection and instructional change. *Teaching and Teacher Education* 112 (2022), 103631.

[43] Heisawn Jeong and Cindy E Hmelo-Silver. 2016. Seven affordances of computer-supported collaborative learning: How to support collaborative learning? How can technologies help? *Educational Psychologist* 51, 2 (2016), 247–265.

[44] Pavan Kapanipathi, Ibrahim Abdelaziz, Srinivas Ravishankar, Salim Roukos, Alexander Gray, Ramón Fernandez Astudillo, Maria Chang, Cristina Cornelio, Saswati Dana, Achille Fokoue-Nkoutche, et al. 2021. Leveraging Abstract Meaning Representation for Knowledge Base Question Answering. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. 3884–3894.

[45] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361* (2020).

[46] James Kennedy, Séverin Lemaignan, Caroline Montassier, Pauline Lavalade, Bahar Irfan, Fotios Papadopoulos, Emmanuel Senft, and Tony Belpaeme. 2017. Child speech recognition in human-robot interaction: evaluations and recommendations. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*. 82–90.

[47] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). http://arxiv.org/abs/1412.6980

[48] R Kusumawati, AF Hadi, et al. 2019. Implementation of integrated inquiry collaborative learning based on the lesson study for learning community to improve students' creative thinking skill. In *Journal of Physics: Conference Series*, Vol. 1211. IOP Publishing, 012097.

[49] Jennifer Langer-Osuna, Emma Gargroetzi, Rosa Chavez, and Jen Munson. 2018. Rethinking loafers: Understanding the productive functions of off-task talk during collaborative mathematics problem-solving. International Society of the Learning Sciences, Inc.[ISLS].

[50] Jennifer M Langer-Osuna. 2018. Productive disruptions: Rethinking the role of off-task interactions in collaborative mathematics learning. *Education Sciences* 8, 2 (2018), 87.

[51] Annabel Latham. 2022. Conversational Intelligent Tutoring Systems: The State of the Art. In *Women in Computational Intelligence: Key Advances and Perspectives on Emerging Topics*, Alice E. Smith (Ed.). Springer International Publishing, Cham, 77–101. https://doi.org/10.1007/978-3-030-79092-9_4

[52] Ngoc-Tien Le, Christophe Servan, Benjamin Lecouteux, and Laurent Besacier. 2016. Better Evaluation of ASR in Speech Translation Context Using Word Embeddings. In *Proc. Interspeech 2016*. 2538–2542. https://doi.org/10.21437/Interspeech.2016-464

[53] Sungbok Lee, Alexandros Potamianos, and Shrikanth Narayanan. 1999. Acoustics of children's speech: Developmental changes of temporal and spectral parameters. *The Journal of the Acoustical Society of America* 105, 3 (1999), 1455–1468.

[54] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 7871–7880.

[55] D. Litman, C. Rose, K. Forbes-Riley, K. VanLehn, D. Bhembe, and S. Silliman. 2006. Spoken versus typed human and computer dialogue tutoring. *International Journal of Artificial Intelligence In Education* 16, 2 (2006), 145–170.

[56] Linda Liu, Yile Gu, Aditya Gourav, Ankur Gandhe, Shashank Kalmane, Denis Filimonov, Ariya Rastrow, and Ivan Bulyko. 2021. Domain-aware neural language models for speech recognition. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 7373–7377.

[57] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[58] Luca Lugini, Christopher Olshefski, Ravneet Singh, Diane Litman, and Amanda Godley. 2020. Discussion Tracker: Supporting Teacher Learning about Students' Collaborative Argumentation in High School Classrooms. In *Conference Proceedings of the 28th International Conference on Computational Linguistics*.

[59] Ioannis Magnisalis, Stavros Demetriadis, and Anastasios Karakostas. 2011. Adaptive and intelligent systems for collaborative learning support: A review of the field. *IEEE transactions on Learning Technologies* 4, 1 (2011), 5–20.

[60] Salima Mdhaffar, Yannick Estève, Nicolas Hernandez, Antoine Laurent, Richard Dufour, and Solen Quiniou. 2019. *Qualitative Evaluation of ASR Adaptation in a Lecture Context: Application to the PASTEL Corpus*. https://doi.org/10.21437/Interspeech.2019-2661 Pages: 573.

[61] Sarah Michaels, Megan Williams Hall, and Lauren B Resnick. 2013. *Accountable talk sourcebook: For classroom conversation that works*. University of Pittsburgh Pittsburgh, PA.

[62] Sarah Michaels and Catherine O'Connor. 2015. Conceptualizing talk moves as tools: Professional development approaches for academically productive discussion. *Socializing intelligence through talk and dialogue* 347 (2015), 362.

[63] Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heinz, and Dan Roth. 2021. Recent advances in natural language processing via large pre-trained language models: A survey. *arXiv preprint arXiv:2111.01243* (2021).

[64] Yaroslav Nechaev, Weitong Ruan, and Imre Kiss. 2021. Towards NLU model robustness to ASR errors at scale. In *KDD 2021 Workshop on Data-Efficient Machine Learning*. https://www.amazon.science/publications/towards-nlu-model-robustness-to-asr-errors-at-scale

[65] Kyungjin Park, Hyunwoo Sohn, Wookhee Min, Bradford Mott, Krista Glazewski, Cindy E Hmelo-Silver, and James Lester. 2022. Disruptive Talk Detection in Multi-Party Dialogue within Collaborative Learning Environments with a Regularized User-Aware Network. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*. 490–499.

[66] Alexandros Potamianos and Shrikanth Narayanan. 2003. Robust recognition of children's speech. *IEEE Transactions on speech and audio processing* 11, 6 (2003), 603–616.

[67] Samuel L Pugh, Arjun Rao, Angela EB Stewart, and Sidney K D'Mello. 2022. Do Speech-Based Collaboration Analytics Generalize Across Task Contexts?. In *LAK22: 12th International Learning Analytics and Knowledge Conference*. 208–218.

[68] Samuel L. Pugh, Shree Krishna Subburaj, Arjun Ramesh Rao, Angela E.B. Stewart, Jessica Andrews-Todd, and Sidney K. D'Mello. 2021. Say What? Automatic Modeling of Collaborative Problem Solving Skills from Student Speech in the Wild. *Proceedings of The 14th International Conference on Educational Data Mining (EDM 2021* (3 2021).

[69] Yan-min Qian, Chao Weng, Xuan-kai Chang, Shuai Wang, and Dong Yu. 2018. Past review, current progress, and challenges ahead on the cocktail party problem. *Frontiers of Information Technology & Electronic Engineering* 19, 1 (2018), 40–63.

[70] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356* (2022).

[71] Parastoo Baghaei Ravari, Ken Jen Lee, Edith Law, and Dana Kulić. 2021. Effects of an adaptive robot encouraging teamwork on students' learning. In *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*. IEEE, 250–257.

[72] Ana Rodrigues, Rita Santos, Jorge Abreu, Pedro Beça, Pedro Almeida, and Sílvia Fernandes. 2019. Analyzing the performance of ASR systems: The effects of noise, distance to the device, age and gender. In *Proceedings of the XX International Conference on Human Computer Interaction (Interacción '19)*. Association for Computing Machinery, New York, NY, USA, 1–8. https://doi.org/10.1145/3335595.3335635

[73] Jeremy Roschelle, Yannis Dimitriadis, and Ulrich Hoppe. 2013. Classroom orchestration: Synthesis. *Computers & Education* 69 (2013), 523–526. https://doi.org/10.1016/j.compedu.2013.04.010

[74] Jeremy Roschelle and Stephanie D Teasley. 1995. The construction of shared knowledge in collaborative problem solving. In *Computer supported collaborative learning*. Springer, 69–97.

[75] Carolyn P Rosé and Sanna Järvelä. 2020. Building community together: towards equitable CSCL practices and processes. *International Journal of Computer-Supported Collaborative Learning* 15 (2020), 249–255.

[76] Somnath Roy. 2021. Semantic-WER: A Unified Metric for the Evaluation of ASR Transcript for End Usability. *arXiv preprint arXiv:2106.02016* (2021).

[77] Jennifer Sabourin, Jonathan P Rowe, Bradford W Mott, and James C Lester. 2011. When off-task is on-task: The affective role of off-task behavior in narrative-centered learning environments. In *Artificial Intelligence in Education: 15th International Conference, AIED 2011, Auckland, New Zealand, June 28–July 2011 15*. Springer, 534–536.

[78] Karla Scornavacco, Jennifer Jacobs, and C Clevenger. 2022. Automated feedback on discourse moves: Teachers' perceived utility of a big data tool.. In *Annual meeting of the American Educational Research Association*.

[79] Gregory Sell, David Snyder, Alan McCree, Daniel Garcia-Romero, Jesús Villalba, Matthew Maciejewski, Vimal Manohar, Najim Dehak, Daniel Povey, Shinji Watanabe, and Sanjeev Khudanpur. 2018. Diarization is Hard: Some Experiences and Lessons Learned for the JHU Team in the Inaugural DIHARD Challenge. In *Proc. Interspeech 2018*. 2808–2812. https://doi.org/10.21437/Interspeech.2018-1893

[80] Shane Settle, Jonathan Le Roux, Takaaki Hori, Shinji Watanabe, and John R Hershey. 2018. End-to-end multi-speaker speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 4819–4823.

[81] Prashanth Gurunath Shivakumar and Shrikanth Narayanan. 2022. End-to-end neural systems for automatic children speech recognition: An empirical study. *Computer Speech & Language* 72 (2022), 101289.

[82] Edwin Simonnet, Sahar Ghannay, Nathalie Camelin, and Yannick Estève. 2018. Simulating ASR errors for training SLU systems. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA), Miyazaki, Japan. https://aclanthology.org/L18-1499

[83] Yu Song, Shunwei Lei, Tianyong Hao, Zixin Lan, and Ying Ding. 2021. Automatic classification of semantic content of classroom dialogue. *Journal of Educational Computing Research* 59, 3 (2021), 496–521.

[84] Robert A Sottilare, Arthur Graesser, Xiangen Hu, and Heather Holden. 2013. *Design recommendations for intelligent tutoring systems: Volume 1-learner modeling*. Vol. 1. US Army Research Laboratory.

[85] Rosy Southwell, Samuel Pugh, E Margaret Perkoff, Charis Clevenger, Jeffrey B Bush, Rachel Lieber, Wayne Ward, Peter Foltz, and Sidney D'Mello. 2022. Challenges and Feasibility of Automatic Speech Recognition for Modeling Student Collaborative Discourse in Classrooms. *International Educational Data Mining Society* (2022).

[86] Leonard Springer, Mary Elizabeth Stanne, and Samuel S Donovan. 1999. Effects of small-group learning on undergraduates in science, mathematics, engineering, and technology: A meta-analysis. *Review of educational research* 69, 1 (1999), 21–51.

[87] Gerry Stahl. 2005. Group cognition in computer-assisted collaborative learning. *Journal of Computer Assisted Learning* 21, 2 (2005), 79–90.

[88] Matthew Stuttle, Jason Williams, and Steve Young. 2004. A framework for dialog systems data collection using a simulated asr channel. In *ICSLP 2004*.

[89] Abhijit Suresh, Jennifer Jacobs, Charis Clevenger, Vivian Lai, Chenhao Tan, James H Martin, and Tamara Sumner. 2021. Using ai to promote equitable classroom discussions: The talkmoves application. In *Artificial Intelligence in Education: 22nd International Conference, AIED 2021, Utrecht, The Netherlands, June 14–18, 2021, Proceedings, Part II*. Springer, 344–348.

[90] A Suresh, J Jacobs, V Lai, C Tan, W Ward, JH Martin, and T Sumner. 2021. Using transformers to provide teachers with personalized feedback on their classroom discourse: The TalkMoves application. *Association for the Advancement of Artificial Intelligence* (2021).

[91] Abhijit Suresh, Jennifer Jacobs, Margaret Perkoff, James H Martin, and Tamara Sumner. 2022. Fine-tuning Transformers with Additional Context to Classify Discursive Moves in Mathematics Classrooms. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*. 71–81.

[92] Piotr Szymański, Piotr Żelasko, Mikolaj Morzy, Adrian Szymczak, Marzena Żyła-Hoppe, Joanna Banaszczak, Lukasz Augustyniak, Jan Mizgajski, and Yishay Carmiel. 2020. WER we are and WER we think we are. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, Online, 3290–3295. https://doi.org/10.18653/v1/2020.findings-emnlp.295

[93] Patrick T Terenzini, Alberto F Cabrera, Carol L Colbeck, John M Parente, and Stefani A Bjorklund. 2001. Collaborative learning vs. lecture/discussion: Students' reported learning gains. *Journal of Engineering Education* 90, 1 (2001), 123–130.

[94] Faith LeAn Turner. 2022. *A Phenomonological Study of Teacher Induction in an Urban Middle School.* Ph. D. Dissertation. Tarleton State University.
[95] Lev Vygotsky. 1978. Interaction between learning and development. *Readings on the development of children* 23, 3 (1978), 34–41.
[96] Jeanne M Wilson, Paul S Goodman, and Matthew A Cronin. 2007. Group learning. *Academy of management review* 32, 4 (2007), 1041–1059.
[97] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*. 38–45.
[98] Ifat Yasin, Fangqi Liu, Vit Drga, Andreas Demosthenous, and Ray Meddis. 2018. Effect of auditory efferent time-constant duration on speech recognition in noise. *The Journal of the Acoustical Society of America* 143, 2 (2018), EL112–EL115.
[99] Charles YC Yeh, Hercy NH Cheng, Zhi-Hong Chen, Calvin CY Liao, and Tak-Wai Chan. 2019. Enhancing achievement and interest in mathematics learning through Math-Island. *Research and Practice in Technology Enhanced Learning* 14, 1 (2019), 1–19.
[100] Wanjun Zhong, Duyu Tang, Zhangyin Feng, Nan Duan, Ming Zhou, Ming Gong, Linjun Shou, Daxin Jiang, Jiahai Wang, and Jian Yin. 2020. LogicalFactChecker: Leveraging Logical Operations for Fact Checking with Graph Module Network. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 6053–6065.