Measuring Conversational Uptake: A Case Study on Student-Teacher Interactions

Dorottya Demszky[1] **Jing Liu**[2] **Zid Mancenido**[3] **Julie Cohen**[4]
Heather Hill[3] **Dan Jurafsky**[1] **Tatsunori Hashimoto**[1]
[1]Stanford University [2]University of Maryland [3]Harvard University [4]University of Virginia
{ddemszky, thashim}@stanford.edu

Abstract

In conversation, *uptake* happens when a speaker builds on the contribution of their interlocutor by, for example, acknowledging, repeating or reformulating what they have said. In education, teachers' uptake of student contributions has been linked to higher student achievement. Yet measuring and improving teachers' uptake at scale is challenging, as existing methods require expensive annotation by experts. We propose a framework for computationally measuring uptake, by (1) releasing a dataset of student-teacher exchanges extracted from math classroom transcripts annotated for uptake by experts; (2) formalizing uptake as pointwise Jensen-Shannon Divergence (PJSD), estimated via next utterance classification; (3) conducting a linguistically-motivated comparison of different unsupervised measures and (4) correlating these measures with educational outcomes. We find that although repetition captures a significant part of uptake, PJSD outperforms repetition-based baselines, as it is capable of identifying a wider range of uptake phenomena like question answering and reformulation. We apply our uptake measure to three different educational datasets with outcome indicators. Unlike baseline measures, PJSD correlates significantly with instruction quality in all three, providing evidence for its generalizability and for its potential to serve as an automated professional development tool for teachers.

1 Introduction

Building on the interlocutor's contribution via, for example, acknowledgment, repetition or elaboration (Figure 1), is known as uptake and is key to a successful conversation. Uptake makes an interlocutor feel heard and fosters a collaborative interaction (Collins, 1982; Clark and Schaefer, 1989), which is especially important in contexts like education. Teachers' uptake of student ideas promotes

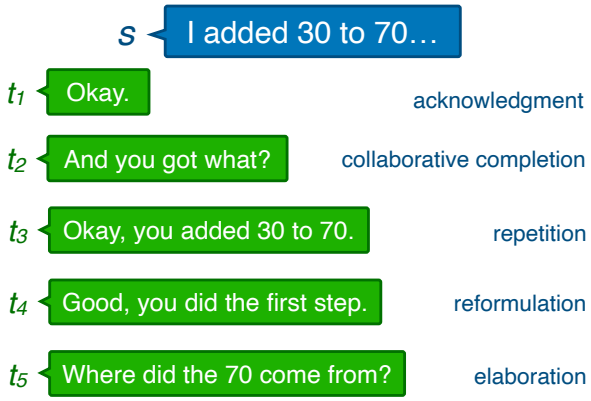

Figure 1: Example student utterance s and possible teacher replies t, illustrating different uptake strategies.

dialogic instruction by amplifying student voices and giving them agency in the learning process, unlike monologic instruction where teachers lecture at students (Bakhtin, 1981; Wells, 1999; Nystrand et al., 1997). Despite extensive research showing the positive impact of uptake on student learning and achievement (Brophy, 1984; O'Connor and Michaels, 1993; Nystrand et al., 2003), measuring and improving teachers' uptake at scale is challenging as existing methods require manual annotation by experts and are prohibitively resource-intensive.

We introduce a framework for computationally measuring uptake. First, we create and release a **dataset** of 2246 student-teacher exchanges extracted from elementary math classroom transcripts, each annotated by three domain experts for teachers' uptake of student contributions.

We take an **unsupervised approach** to measure uptake in order to encourage domain-transferability and account for the fact that large amounts of labeled data are not possible in many contexts due to data privacy reasons and/or limited resources. We conduct a careful analysis of the role of **repetition** in uptake by measuring utterance overlap and

similarity. We find that the proportion of student words repeated by the teacher (%-IN-T) captures a large part of uptake, and that surprisingly, word-level similarity measures consistently outperform sentence-level similarity measures, including ones involving sophisticated neural models.

To capture uptake phenomena beyond repetition and in particular those relevant to teaching (e.g. question answering), we **formalize uptake** as a measure of the reply's dependence on the source utterance. We quantify dependence via **pointwise Jensen-Shannon divergence** (PJSD), which captures how easily someone (e.g., a student) can distinguish the true reply from randomly sampled replies. We show that PJSD can be estimated via cross-entropy loss obtained from next utterance classification (NUC).

We **train a model** by fine-tuning BERT-base (Devlin et al., 2019) via NUC on a large, combined dataset of student-teacher interactions and Switchboard (Godfrey and Holliman, 1997). We show that scores obtained from this model significantly outperform our baseline measures. Using dialog act annotations on Switchboard, we demonstrate that PJSD is indeed better at capturing phenomena such as reformulation, question answering and collaborative completion than %-IN-T, our best-performing baseline. Our manual analysis also shows qualitative differences between the models: the examples where PJSD outperforms %-IN-T are enriched by teacher prompts for elaboration, an exemplar for dialogic instruction (Nystrand et al., 1997).

Finally, we find that our PJSD measure shows a **significant linear correlation with outcomes** such as student satisfaction and instruction quality across three different datasets of student-teacher interactions: the NCTE dataset (Kane et al., 2015), a one-on-one online tutoring dataset, and the SimTeacher dataset (Cohen et al., 2020). These results provide evidence for the generalizability of our PJSD measure and for its potential to serve as an automated tool to give feedback to teachers.

2 Background on Uptake

Uptake has several linguistic and social functions. (1) It creates *coherence* between two utterances, helping structure the discourse (Halliday and Hasan, 1976; Grosz et al., 1977; Hobbs, 1979). (2) It is a mechanism for *grounding*, i.e. demonstrating understanding of the interlocutor's contribution by accepting it as part of the common ground (shared set of beliefs among interlocutors) (Clark and Schaefer, 1989). (3) It promotes *collaboration* with the interlocutor by sharing the floor with them and indicating what they have said is important (Bakhtin, 1981; Nystrand et al., 1997).

There are multiple linguistic strategies for uptake, such as acknowledgment, collaborative completion, repetition, and question answering — see Figure 1 for a non-exhaustive list. A speaker can use multiple strategies at the same time, for example, t_3 in Figure 1 includes both acknowledgment and repetition. Different strategies can represent lower or higher uptake depending on how effectively they achieve the aforementioned functions of uptake. For example, Tannen (1987) argues that repetition is a highly pervasive and effective strategy for ratifying listenership and building a coherent discourse. In education, high uptake has been defined as cases where the teacher follows up on the student's contribution via a question or elaboration (Collins, 1982; Nystrand et al., 1997).

Even though there are strategies that tend to represent higher degrees of uptake than others, the best-fitting uptake strategy often depends on the discourse context (Clark and Schaefer, 1989). For example, if the interlocutor asks a question, an answer is usually the most appropriate, or if they are struggling to find words, collaborative completion may be a very effective uptake strategy.

3 A New Educational Uptake Dataset

Despite the substantial literature on the functions of uptake, we are not aware of a publicly available dataset labeled for this phenomenon. To address this, we recruit domain experts (math teachers and raters trained in classroom observation) to annotate a dataset of exchanges between students and teachers. The exchanges are sampled from transcripts of 45-60 minute long 4th and 5th grade elementary math classroom observations collected by the National Center for Teacher Effectiveness (NCTE) between 2010-2013 (Kane et al., 2015). The transcripts represent data from 317 teachers across 4 school districts in New England that serve largely low-income, historically marginalized students. Transcripts are fully anonymized: student and teacher names are replaced with terms like "Student", "Teacher" or "Mrs. H".[1]

[1] Parents and teachers gave consent for the study (Harvard IRB #17768), and for de-identified data to be retained and used in future research. The transcripts were anonymized at the time they were created.

Preparing utterance pairs. We prepare a dataset of utterance pairs (S, T), where S is a student utterance and T is a subsequent teacher utterance. The concept of uptake presupposes that there is something to be taken up; in our case that the student utterance has substance. For example, short student utterances like "yes" or "one-third" do not present many opportunities for uptake. Based on our pilot annotations, these utterances are difficult for even expert annotators to label. Therefore, we only keep utterance pairs where S contains at least 5 tokens, excluding punctuation. We also remove all utterance pairs where the utterances contain an [Inaudible] marker, indicating low audio quality. Out of the remaining 55k (S, T) pairs, we sample 2246 for annotation.[2]

Annotation. Our annotation framework for uptake is designed by experts in math quality instruction, including our collaborators, former and current math teachers and former and current raters for the Mathematical Quality Instruction (MQI) coding instrument, used to assess math instruction (Teaching Project, 2011). In the annotation interface, raters can see (1) the utterance pair (S, T), (2) the lesson topic, which is manually labeled as part of the original dataset, and (3) two utterances immediately preceding (S, T) for context. Annotators are asked to first check whether (S, T) relates to math – e.g. "Can I go to the bathroom?" is unrelated to math. If both S and T relate to math, raters are asked to select among three labels: "low", "mid" and "high", indicating the degree to which a teacher demonstrates that they are following what the student is saying or trying to say. The annotation framework is included in Appendix A. We recruit 13 expert annotators for the task, excluding our coauthors, to eliminate bias. We randomly assign each example to three raters.

Table 1 includes a sample of our annotated data. Inter-rater agreement for uptake is ρ = .47, measured by (1) excluding examples where at least one rater indicated that the utterance pair does not relate to math[3]; (2) converting rater's scores into numbers ("low": 0, "mid": 1, "high": 2); (3) z-scoring each rater's scores; (4) computing a leave-out Spearman ρ for each rater by correlating their judgments with the average judgments of the other two raters[4]; and (5) taking the average of the leave-out correlations across raters. The inter-rater agreement we obtain is similar to inter-rater agreement in classroom observation protocols (Kelly et al., 2020). We obtain a single label for each example by averaging the z-scored judgments across raters.

Example	Uptake
S: 'Cause you took away 10 and 70 minus 10 is 60. T: Why did we take away 10?	high
S: There's not enough seeds. T: There's not enough seeds. How do you know right away that 128 or 132 or whatever it was you got doesn't make sense?	high
S: Teacher L, can you change your dimensions like 3-D and stuff for your bars? T: You can do 2-D or 3-D, yes. I already said that.	mid
S: The higher the number, the smaller it is. T: You got it. That's a good thought.	mid
S: An obtuse angle is more than 90 degrees. T: Why don't we put our pencils down and just do some brainstorming, and then we'll go back through it?	low
S: Because the base of it is a hexagon. T: Student K?	low

Table 1: Examples from our annotated data.

4 Uptake as Overlap & Similarity

As we see in Table 1, examples labeled for high uptake tend to have overlap between S and T; this is expected, since incorporating the previous utterance in some form is known to be an important aspect of uptake (Section 2). Therefore, we begin by carefully analyzing repetition and defer discussion of more complex uptake phenomena to Section 5. To accurately quantify repetition-based uptake, we evaluate a range of metrics and surprisingly find that *word overlap* based measures correlate significantly better with uptake annotations than more sophisticated, utterance-level similarity measures.

4.1 Methods

We use several algorithms to better understand if word- or utterance-level similarity is a better measure of uptake. For each token-based algorithm,

[2] To enable potential analyses on the temporal dynamics of uptake, we randomly sampled 15 transcripts where we annotate all (S, T) pairs (constituting 29% of our annotations). The rest of the pairs are sampled from the remaining data.

[3] This step is motivated by widely used education observation protocols such as MQI (Hill et al., 2008), which also clearly separate on- vs off-task instruction.

[4] Kappa values (Fleiss κ=.286) lead to similar conclusions for interrater agreement. We use correlations instead because Kappa has undesirable properties (see Delgado and Tibau, 2019) and correlations are more interpretable and directly comparable to our models' results (see later sections).

we experimented with several different choices for pre-processing and we include symbols for the set of choices yielding best performance: removing punctuation ♠, removing stopwords using NLTK (Bird, 2006) ⊕, and stemming via NLTK's SnowballStemmer †.

String- and token-overlap.

LCS: Longest Common Subsequence.

%-IN-T: Fraction of tokens from S that are also in T (Miller and Beebe-Center, 1956). [♠ ⊕ †]

%-IN-S: Fraction of tokens from T that are also in S. [♠⊕]

JACCARD: Jaccard similarity. [♠⊕]

BLEU: BLEU score (Papineni et al., 2002) for up to 4-grams. We use S as the reference and T as the hypothesis.[♠ ⊕ †]

Embedding-based similarity. For the word vector-based metrics, we use 300-dimensional GloVe vectors pretrained on 6B tokens from Wikipedia and Gigaword (Pennington et al., 2014)

GLOVE [ALIGNED]: Average pairwise cosine similarity of word embeddings between tokens from S and its most similar token in T. [♠]

GLOVE [UTT]: Cosine similarity of utterance vectors representing S and T. Utterance vectors are obtained by averaging word vectors from S and from T. [♠⊕]

SENTENCE-BERT: Cosine similarity of utterance vectors representing S and T, obtained using a pre-trained Sentence-BERT model for English (Reimers and Gurevych, 2019).[5]

UNIVERSAL SENTENCE ENCODER: Inner product of utterance vectors representing S and T, obtained using a pre-trained Universal Sentence Encoder for English (Cer et al., 2018).

4.2 Results

We compute correlations between model scores and human labels via Spearman rank order correlation ρ. We perform bootstrap sampling (for 1000 iterations) to compute 95% confidence intervals.

The results are shown in Table 2. Overall, we find that token-based measures outperform utterance-based measures, with %-IN-T (ρ = .523), GLOVE [ALIGNED] (ρ = .518) (a soft word overlap measure) and BLEU (ρ = .510) performing the best. In fact, all three of them correlate better with averaged human judgments than the humans among themselves (ρ = .474). Even embedding-based algorithms that are computed at the utterance-level do not outperform %-IN-T, a simple word overlap baseline. It is noteworthy that all measures have a significant correlation with human judgments.

Model	ρ	95% CI
LCS	.283	[.240, .329]
%-IN-T	**.523*****	[.488, .559]
%-IN-S	.440	[.399, .480]
JACCARD	.450	[.413, .487]
BLEU	.510	[.472, .543]
GLOVE [ALIGNED]	.518	[.483, .550]
GLOVE [UTT]	.424	[.378, .465]
SENTENCE-BERT	.390	[.350, .432]
UNIVERSAL SENTENCE ENCODER	.448	[.408, .486]
Human (leave-out)	.474	[.456, .494]

Table 2: Results on our labeled data from our baseline measures. Asterisks indicate that %-IN-T significantly outperforms GLOVE [ALIGNED] ($p < 0.001$), measured by a paired bootstrap test, comparing the difference between the ρ obtained by %-IN-T and the one by GLOVE [ALIGNED] across 1000 iterations, then using a t-test.

The surprisingly strong performance of %-IN-T, GLOVE [ALIGNED] and BLEU provide further evidence that the extent to which T repeats words from S is important for uptake (Tannen, 1987), especially in the context of teaching. The fact that removing stopwords helps these measures suggests that the repetition of function words is less important for uptake; an interesting contrast to linguistic style coordination in which function words play a key role (Danescu-Niculescu-Mizil and Lee, 2011). Moreover, the amount of words T adds in addition to words from S also seems relatively irrelevant based on the lower performance of the measures that penalize T containing words that are not in S.

5 Uptake as Dependence

Now we introduce our main uptake measure, used to capture a broader range of uptake phenomena beyond repetition including, e.g., acknowledgment and question answering (Section 2). We formalize uptake as dependence of T on S, captured by the Jensen-Shannon Divergence, which quantifies the extent to which we can tell whether T is a response

[5] https://github.com/UKPLab/sentence-transformers

to S or is it a random response (T'). If we cannot tell the difference between T and T', we argue that there can be no uptake, as T fails all three functions of coherence, grounding and collaboration (Section 2).

We can formally define the dependence for a single teacher-student utterance pair (s, t) in terms of a pointwise variant of JSD (PJSD) as

$$pJSD(t,s) := -\frac{1}{2}\Big(\log P(Z=1|M=t,s) + \mathbb{E}\log(1-P(Z=1|M=T',s))\Big) + \log(2) \quad (1)$$

where (S, T) is a teacher-student utterance pair, T' is a randomly sampled teacher utterance that is independent of S, and $M := ZT + (1-Z)T'$ is a mixture of the two with a binary indicator variable $Z \sim \text{Bern}(p=0.5)$.

This pointwise measure relates to the standard JSD for $T|S=s$ and T' by taking expectations over the teacher utterance via $\mathbb{E}[pJSD(T,s)|S=s] = JSD(T|S=s||T')$. We consider the pointwise variant for the rest of the section, as we are interested in a measure of dependence between a specific (t, s) rather than one that is averaged over multiple teacher utterances.

5.1 Next Utterance Classification

The definition of PJSD naturally suggests an estimator based on the *next utterance classification* task — a task previously used in neighboring NLP areas like dialogue generation and discourse coherence. We fine-tune a pre-trained BERT-base model (Devlin et al., 2019) on a dataset of (S, T) pairs to predict if a specific (s, t) is a true pair or not (i.e., whether t came from T or T'). The objective function is cross-entropy loss, computed over the output of the final classification layer that takes in the last hidden state of t. Let Z be a binary indicator variable representing the model's prediction. Then, the cross entropy loss for identifying z is

$$L(t,s) = -\log f_\theta(t,s) - \mathbb{E}\log(1-f_\theta(T',s)) \quad (2)$$

Which can be used directly as an estimator for the log-probability terms in Equation 1,

$$\widehat{pJSD}(t,s) := \frac{1}{2}L(t,s) + \log 2. \quad (3)$$

Standard variational arguments (Nowozin et al., 2016) show that any classifier f_θ forms a lower bound on the JSD,

$$JSD(T|S=s||T') \geq \mathbb{E}[\widehat{pJSD}(T,s)|S=s].$$

Thus, our overall procedure is to fit $f_\theta(t, s)$ by maximizing $\mathbb{E}[\widehat{pJSD}(t, s)]$ over our dataset and then use $\widehat{pJSD}(t, s)$ as our pointwise measure of dependence.

Model	ρ	95% CI
%-IN-T	.523	[.488, .559]
PJSD	**.540*****	[.505, .574]
Human (leave-out)	.474	[.456, .494]

Table 3: Results on our labeled data, obtained by the PJSD model. The asterisks, calculated analogically to Table 2, indicate that the difference between the two models' performance is significant.

Training data. We use (S, T) pairs from three sources to form our training data: the NCTE dataset (Kane et al., 2015) (see Section 3), Switchboard (Godfrey and Holliman, 1997) and an one-on-one online tutoring dataset (see Section 6) — we use a combination of datasets instead of one dataset in order to support the generalizability of the model. Filtering out examples with $S < 5$ tokens or [Inaudible] markers (see Section 3), our resulting dataset consists of 259k (S, T) pairs. For each (s, t) pair, we randomly select 3 negative (s, t') pairs from the same source dataset, yielding 777k examples.[6]

Parameter Settings. We fine-tune our model for 1 epoch to avoid overfitting with a batch size of 32×2 gradient accumulation steps, max length of 120 tokens for S and T each (the rest is truncated), learning rate of 6.24e-5 with linear decay and the AdamW optimizer (Loshchilov and Hutter, 2017). Training took about 13hrs on a single TitanX GPU.

5.2 Results & Analysis

Table 3 shows that the PJSD model ($\rho = .540$) significantly outperforms %-IN-T. Now we turn to our motivating goals for proposing PJSD and quantitatively analyze its ability to capture more sophisticated forms for uptake.

Fine-grained analysis. To understand if there is a pattern explaining PJSD's better performance, we

[6] We do not split the data into training and validation sets, as we found that using predictions on the training data vs those on the test data as our uptake measure yield similar results, so we opted for maximizing training data size.

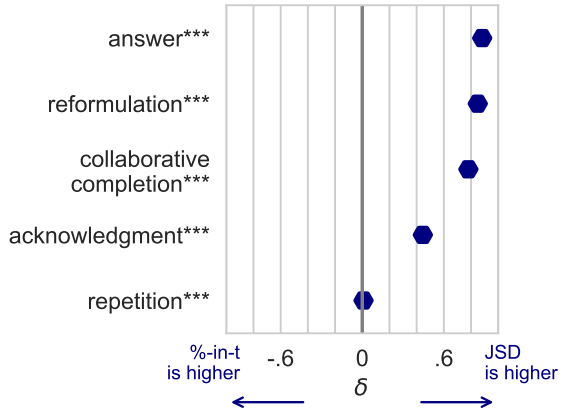

Figure 2: The difference (δ) between the scores from %-IN-T and PJSD for five uptake phenomena labeled in Switchboard. Asterisks indicate significance (***: $p < 0.001$), estimated via a median test.

quantify the occurence of different linguistic phenomena for examples where PJSD outperforms %-IN-T. Concretely, we compute the residuals for each model, regressing the human labels on their predictions. Then, we take those examples where the difference between the two models' residuals is 1.5 standard deviations above the mean difference between their residuals. We label teacher utterances in these high uptake examples for four linguistic phenomena associated with uptake and good teaching (elaboration prompt, reformulation, collaborative completion, and answer to question), allowing multiple labels (e.g. elaboration prompt and completion often co-occur). As Table 4 shows, elaboration prompts, which are exemplars of high uptake in teaching (Nystrand et al., 1997) are significantly more likely to occur in this set — suggesting that there is a qualitative difference between what these models capture that is relevant for teaching. We do not find a significant difference in the occurrence of reformulations, collaborative completions and answers between the two sets, possibly due to the small sample size (n=67). To see whether these differences are significant on a larger dataset, we now turn to the Switchboard dialogue corpus.

Switchboard dialog acts. We take advantage of dialog act annotations on Switchboard (Jurafsky et al., 1997), to compare uptake phenomena captured by %-IN-T and PJSD at a large scale. We identify five uptake phenomena labeled in Switchboard and map them to SWBD-DAMSL tags: acknowledgment, answer, collaborative completion, reformulation and repetition (see details in Appendix B).

Label	Examples
elaboration prompt (4.25*)	S: so it means that the whole equation is only the same. T: what does it mean? i still don't understand what is it?
reformulation (2.6)	S: multiplication is like, say, for instance, nine times twenty. you just take - nine just nine times and add it up. T: okay, so repeated addition.
answer (2.67)	S: do we look at the d or the m first? T: the m. what's this called, that i'm writing?
collaborative completion (0)	S: we had to add twenty-four plus twenty-four. T: because there are how many triangles?

Table 4: Examples for linguistic phenomena, manually labeled in the dataset where PJSD and %-IN-T make significantly different predictions. Parenthetical numbers after the labels represent the odds ratio of examples occurring in the set where PJSD performs better over the set where %-IN-T performs better (p-values are computed using a Fisher exact test (*: $p < 0.05$)).

We estimate scores for %-IN-T and PJSD for all utterance pairs (S, T) in Switchboard, filtering out ones where $S < 5$ tokens. We apply our PJSD model from Section 5.1, which was partially fine-tuned on Switchboard. Since both measures are bounded, we quantile-transform the distribution of each measure to have a uniform distribution. For each uptake phenomenon, we compute the difference (δ) between the median score from PJSD and the median score from %-IN-T for all (S, T) pairs where T is labeled for that phenomenon.

The results (Figure 2) show that PJSD predicts significantly higher scores than %-IN-T for all phenomena, especially for answers, reformulations, collaborative completions and acknowledgments. For repetition, δ is quite small, but still significant due to the large sample size. These findings corroborate our hypothesis that %-IN-T and PJSD capture repetition similarly, but PJSD is able to better capture other uptake phenomena.

6 Downstream Application

To test the generalizability of our uptake measures and their link to instruction quality, we correlate PJSD and %-IN-T with educational outcomes on three different datasets of student-teacher interactions (Table 5).

NCTE dataset. We use all transcripts from the NCTE dataset (Kane et al., 2015) (see Section 3) with associated classroom observation scores based on the MQI coding instrument (Teaching Project,

Dataset	Size	Genre	Topic	Class size	Outcome	PJSD (β)	%-IN-T (β)
NCTE	1.6k conv. 55k (S,T)	in-person spoken	math	whole class	use of student contributions math instruction quality	.101*** .091***	.113*** .121***
SimTeacher	338 conv. 2.7k (S,T)	virtual spoken	literature	small group	quality of feedback	.127*	.123*
Tutoring	4.6k conv. 85k (S,T)	virtual written	math, science	one-on-one	student satisfaction external reviewer rating	.069*** .063***	.008 .021

Table 5: The correlation of uptake scores from PJSD and %-IN-T and outcomes for three educational datasets. The β values represent z-scored coefficients, each obtained from an ordinary least squares regression, controlling for the number of (S,T) pairs we have uptake scores for in each conversation. Asterisks represent p-values.

2011). We select two items from MQI relevant to uptake as outcomes: (1) use of student math contributions and (2) overall quality of math instruction. Since these items are coded at a 7-minute segment-level, we take the average ratings across raters and segments for each transcript.

Tutoring dataset. We use data from an educational technology company (same as in Chen et al., 2019), which provides on-demand text-based tutoring for math and science. With a mobile application, a student can take a picture of a problem or write it down, and is then connected to a professional tutor who guides the student to solve the problem. Similarly to Chen et al. (2019), we filter out short sessions where the tutors are unlikely to deliver meaningful tutoring. Specifically, we create a list of (S,T) pairs for all sessions, keeping pairs where $S \geq 5$ tokens, and then remove sessions with fewer than ten (S,T) pairs. This results in 4604 sessions, representing 108 tutors and 1821 students. Each session is associated with two outcome measures: (1) student satisfaction scores (1-5 scale) and (2) a rating by the tutor manager based on an evaluation rubric (0-1 scale).

SimTeacher dataset. We use a dataset collected by Cohen et al. (2020), via a mixed reality simulation platform in which novice teachers get to practice key classroom skills in a virtual classroom interface populated by student avatars. The avatars are controlled remotely by a trained actor; hence the term "mixed" reality. All pre-service teachers from a large public university complete a five-minute simulation session at multiple timepoints in their teacher preparation program, and are coached on how to better elicit students' thinking about a text. We use data from Fall 2019, with 338 sessions representing 117 teachers. Since all sessions are based on the *same scenario* (discussed text, leading questions, avatar scripts), this dataset uniquely allows us to answer the question: controlling for student avatar scripts, does a greater teacher uptake lead to better outcomes? For the outcome variable, we use their holistic "quality of feedback" measure (1-10 scale), annotated at the transcript-level by the original research team.[7]

6.1 Results & Analysis

As outcomes are linked to conversations, we first mean-aggregate uptake scores to the conversation-level. We then compute the correlation of uptake scores and outcomes using an ordinary least squares regression, controlling for the number of (S,T) pairs in each conversation.

The results (Table 5) indicate that PJSD correlates with all of the outcome measures significantly. %-IN-T also shows significant correlations for NCTE and for SimTeacher, but not for the tutoring dataset. We provide more details below.

For NCTE and SimTeacher, we find that two measures show similar positive correlations with outcomes. These results provide further insight into our earlier findings from Section 5.2. They suggest that the teacher's repetition of student words, also known as "revoicing" in math education (Forman et al., 1997; O'Connor and Michaels, 1993), may be an especially important mediator of instruction quality in classroom contexts and other aspects of uptake are relatively less important. The significant correlation of PJSD with the outcome in case of SimTeacher is especially noteworthy because PJSD was *not* fine-tuned in this dataset (Section 5.1); this provides evidence for the adaptability of a pre-trained model to other (similar) datasets.

The gap between the two measures in case of the tutoring dataset is an interesting finding, possi-

[7] This overall quality scale accounts for the extent to which teachers actively work to support student avatars' development of text-based responses, highlighting the importance of probing student responses (e.g. "Where in the text did you see that?"; "What made you think this about the character?").

high student feedback (%-IN-T < PJSD)	low student feedback (PJSD < %-IN-T)
S: if they're the same length i think T: that's right! all we need is the length, and that's enough. S: the energy from the one pendulum moving will transfer the same frequency to the second pendulum once they touch? T: they don't even need to touch! we can swing them so they swing side by side, like two swings on a swingset. S: pendulum one will start to absorb energy from pendulum two? T: exactly! and eventually, the whole process will reverse until pendulum one is moving full speed again.	S: when you are saying mixture are you talking about nitrogen? T: thanks for your question. S: no i don't think so T: great answer! S: i don't know , just made an educated guess T: great try! S: i want further explanation about volume and number moles when using nitrogen T: sure. no worries!

Table 6: Examples from the tutoring dataset — for both examples, the predictions by PJSD are more accurate than the ones by %-IN-T that predicts too low and too high values, respectively, when compared to student ratings.

bly explained by the conversational setting: repetition may be an effective uptake strategy in multi-participant & spoken settings, ensuring that everyone has heard what the student said and is on the same page; whereas, in a written 1:1 teaching setting, repetition may not be necessary or effective as both participants are likely to assume that that their interlocutor has read their words. Our qualitative analysis suggests PJSD might be outperforming %-IN-T because it is better able to pick up on cues related to teacher responsiveness (we include two examples in Table 6). To test this, we detect coarse-grained estimates of teacher uptake: teacher question marks (estimate of follow-up question) and teacher exclamation marks (estimate of approval). We then follow the same procedure as in Section 5.2 and find that dialogs where PJSD outperforms %-IN-T, in terms of predicting student ratings, have a higher ratio of exchanges with teacher questions ($p < 0.05$, obtained from two-sample t-test) and teacher exclamation marks ($p < 0.01$).

To put these effect sizes from Table 5 (where significant) in the context of education interventions that are designed to increase student outcomes (typically test scores), the coefficients we report here are considered average for an effective educational intervention (Kraft, 2020). Further, existing guidelines for educational interventions would classify uptake as a promising potential intervention, as it is highly scalable and easily quantified.

7 Related Work

Prior computational work on classroom discourse has employed supervised, feature-based classifiers to detect teachers' discourse moves relevant to student learning, such as authentic questions, elaborated feedback and uptake, treating these moves as binary variables (Samei et al., 2014; Donnelly et al., 2017; Kelly et al., 2018; Stone et al., 2019; Jensen et al., 2020). Our labeled dataset, unsupervised approach (involving a state-of-the art pre-trained model), and careful analysis across domains are novel contributions that will enable a fine-grained and domain-adaptable measure of uptake that can support researchers and teachers.

Our work aligns closely with research on the computational study of conversations. For example, measures have been developed to study constructiveness (Niculae and Danescu-Niculescu-Mizil, 2016), politeness (Danescu-Niculescu-Mizil et al., 2013) and persuasion (Tan et al., 2016) in conversations. Perhaps most similar to our work, Zhang and Danescu-Niculescu-Mizil (2020) develop an unsupervised method to identify therapists' backward- and forward-looking utterances, with which they guide their conversations.

We also draw on work measuring discourse coherence via embedding cosines (Xu et al., 2018; Ko et al., 2019), or via utterance classification (Xu et al., 2019; Iter et al., 2020), the latter of which is used also for building and evaluating dialog systems (Lowe et al., 2016; Wolf et al., 2019). Our work extends these two families of methods to human conversation and highlights the different linguistic phenomena they capture. Finally, our work shows the key role of coherence in the socially important task of studying uptake.

8 Conclusion

We propose a framework for measuring uptake, a core conversational phenomenon with particularly high relevance in teaching contexts. We release an annotated dataset and develop and compare unsupervised measures of uptake, demonstrating significant correlation with educational outcomes across three datasets. This lays the groundwork (1) for scaling up teachers' professional development on uptake thereby enabling improvements to educa-

tion, (2) for conducting analyses on uptake across domains and languages where labeled data does not exist and (3) for studying the effect of uptake on a wider range of socially relevant outcomes.

9 Ethical Considerations

Our objective in building a dataset and a framework for measuring uptake is (1) to aid researchers studying conversations and teaching and (2) to (ultimately) support the professional development of educators by providing them with a scalable measure of a phenomenon that supports student learning. Our second objective is especially important, since existing forms of professional development aimed at improving uptake are highly resource intensive (involving classroom observations and manual evaluation). This costliness has meant that teachers working in under-resourced school systems have thus far had limited access to quality professional development in this area.

The dataset we release is sampled from transcripts collected by the National Center for Teacher Effectiveness (NCTE) (Kane et al., 2015) (Harvard IRB #17768). These transcripts represent data from 317 teachers across 4 school districts in New England that serve largely low-income, historically marginalized students. The data was collected as part of a carefully designed study on teacher effectiveness, spanning three years between 2010 and 2013 and it was de-identified by the original research team, meaning that in the transcripts, student names are replaced with "Student" and teacher names are replaced with "Teacher". Both parents and teachers gave consent for the de-identified data to be retained and used in future research. The collection process and representativeness of the data are all described in great detail in (Kane et al., 2015). Given that the dataset was collected a decade ago, there may be limitations to its use and ongoing relevance. That said, research in education reform has long attested to the fact that teaching practices have remained relatively constant over the past century (Cuban, 1993; Cohen and Mehta, 2017) and that there are strong socio-cultural pressures that maintain this (Cohen, 1988).

The data was annotated by 13 raters, whose demographics are largely representative of teacher demographics in the US[8]. All raters have domain expertise, in that they are former or current math teachers and former or current raters for the Mathematical Quality Instruction (Teaching Project, 2011). The raters were trained for at least an hour each on the coding instrument and spent 8 hours on average on the annotation (over the course of several weeks) and were compensated $16.5 / hr.

In Section 6, we apply our data to to two educational datasets besides NCTE. We do not release either of these datasets. The SimTeacher dataset was collected by Cohen et al. (2020) (University of Virginia IRB #2918), for research and program improvement purposes. The participants in the study are mostly white (82%), female (90%), and middle class (71%), mirroring the broader teaching profession. As for the tutoring dataset, the data belongs to a private company; the students and tutors have given consent for their data to be used for research, with the goal of improving the company's services. The company works with a large number of tutors and students; we use data that represents 108 tutors and 1821 students. 70% of tutors in the data are male, complementing the other datasets where the majority of teachers are female. The company does not share other demographic information about tutors and students.

Similarly to other data-driven approaches, it is important to think carefully about the source of the training data when considering downstream use cases of our measure. Our unsupervised approach helps address this issue as it allows for training the model on data that is representative of the population that it is meant to serve.

[8] https://nces.ed.gov/fastfacts/display.asp?id=28

References

M. M. Bakhtin. 1981. *The dialogic imagination: four essays*. University of Texas Press.

Steven Bird. 2006. NLTK: The Natural Language Toolkit. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pages 69–72, Sydney, Australia. Association for Computational Linguistics.

Jere E Brophy. 1984. *Teacher behavior and student achievement*. 73. Institute for Research on Teaching, Michigan State University.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder for english. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 169–174.

Guanliang Chen, Rafael Ferreira, David Lang, and Dragan Gasevic. 2019. Predictors of student satisfaction: A large-scale study of human-human online tutorial dialogues. *International Educational Data Mining Society*.

Herbert H Clark and Edward F Schaefer. 1989. Contributing to discourse. *Cognitive science*, 13(2):259–294.

David K Cohen. 1988. Teaching practice: Plus ca change... issue paper 88-3.

David K Cohen and Jal D Mehta. 2017. Why reform sometimes succeeds: Understanding the conditions that produce reforms that last. *American Educational Research Journal*, 54(4):644–690.

Julie Cohen, Vivian Wong, Anandita Krishnamachari, and Rebekah Berlin. 2020. Teacher coaching in a simulated environment. *Educational Evaluation and Policy Analysis*, 42(2):208–231.

James Collins. 1982. Discourse style, classroom interaction and differential treatment. *Journal of Reading Behavior*, 14(4):429–437.

Larry Cuban. 1993. *How teachers taught: Constancy and change in American classrooms, 1890-1990*. Teachers College Press.

Cristian Danescu-Niculescu-Mizil and Lillian Lee. 2011. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. *ACL HLT 2011*, page 76.

Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A computational approach to politeness with application to social factors. In *51st Annual Meeting of the Association for Computational Linguistics*, pages 250–259. ACL.

Rosario Delgado and Xavier-Andoni Tibau. 2019. Why cohen's kappa should be avoided as performance measure in classification. *PloS one*, 14(9):e0222916.

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)*, pages 4171–4186.

Patrick J Donnelly, Nathaniel Blanchard, Andrew M Olney, Sean Kelly, Martin Nystrand, and Sidney K D'Mello. 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 & Knowledge Conference*, pages 218–227.

Ellice A Forman, Dawn E McCormick, and Richard Donato. 1997. Learning what counts as a mathematical explanation. *Linguistics and Education*, 9(4):313–339.

John J Godfrey and Edward Holliman. 1997. Switchboard-1 release 2. *Linguistic Data Consortium, Philadelphia*, 926:927.

Barbara J Grosz et al. 1977. The representation and use of focus in a system for understanding dialogs. In *IJCAI*, volume 67, page 76. Citeseer.

Michael Alexander Kirkwood Halliday and Ruqaiya Hasan. 1976. *Cohesion in English*. London: Longmans.

Heather C Hill, Merrie L Blunk, Charalambos Y Charalambous, Jennifer M Lewis, Geoffrey C Phelps, Laurie Sleep, and Deborah Loewenberg Ball. 2008. Mathematical knowledge for teaching and the mathematical quality of instruction: An exploratory study. *Cognition and instruction*, 26(4):430–511.

Jerry R Hobbs. 1979. Coherence and coreference. *Cognitive science*, 3(1):67–90.

Dan Iter, Kelvin Guu, Larry Lansing, and Dan Jurafsky. 2020. Pretraining with contrastive sentence objectives improves discourse performance of language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4859–4870.

Emily Jensen, Meghan Dale, Patrick J Donnelly, Cathlyn Stone, Sean Kelly, Amanda Godley, and Sidney K D'Mello. 2020. Toward automated feedback on teacher discourse to enhance teacher learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13.

Daniel Jurafsky, Elizabeth Shriberg, and Debra Biasca. 1997. Switchboard SWBD-DAMSL Labeling Project Coder's Manual, Draft 13. Technical Report 97-02, University of Colorado Institute of Cognitive Science.

T Kane, H Hill, and D Staiger. 2015. National center for teacher effectiveness main study. icpsr36095-v2.

Sean Kelly, Robert Bringe, Esteban Aucejo, and Jane Cooley Fruehwirth. 2020. Using global observation protocols to inform research on teaching effectiveness and school improvement: Strengths and emerging limitations. *Education Policy Analysis Archives*, 28:62.

Sean Kelly, Andrew M Olney, Patrick Donnelly, Martin Nystrand, and Sidney K D'Mello. 2018. Automatically measuring question authenticity in real-world classrooms. *Educational Researcher*, 47(7):451–464.

Wei-Jen Ko, Greg Durrett, and Junyi Jessy Li. 2019. Linguistically-informed specificity and semantic plausibility for dialogue generation. In *Proceedings of NAACL 2019*, pages 3456–3466.

Matthew A Kraft. 2020. Interpreting effect sizes of education interventions. *Educational Researcher*, 49(4):241–253.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Ryan Lowe, Iulian Vlad Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. On the evaluation of dialogue systems with next utterance classification. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 264–269.

George A. Miller and J. G. Beebe-Center. 1956. Some psychological methods for evaluating the quality of translations. *Mechanical Translation*, 3:73–80.

Vlad Niculae and Cristian Danescu-Niculescu-Mizil. 2016. Conversational markers of constructive discussions. In *Proceedings of NAACL-HLT*, pages 568–578.

Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. 2016. f-gan: training generative neural samplers using variational divergence minimization. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 271–279.

Martin Nystrand, Adam Gamoran, Robert Kachur, and Catherine Prendergast. 1997. *Opening dialogue*. New York: Teachers College Press.

Martin Nystrand, Lawrence L Wu, Adam Gamoran, Susie Zeiser, and Daniel A Long. 2003. Questions in time: Investigating the structure and dynamics of unfolding classroom discourse. *Discourse processes*, 35(2):135–198.

Mary C O'Connor and Sarah Michaels. 1993. Aligning academic task and participation status through revoicing: Analysis of a classroom discourse strategy. *Anthropology & Education Quarterly*, 24(4):318–335.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3973–3983.

Borhan Samei, Andrew M Olney, Sean Kelly, Martin Nystrand, Sidney D'Mello, Nathan Blanchard, Xiaoyi Sun, Marcy Glaus, and Art Graesser. 2014. Domain independent assessment of dialogic properties of classroom discourse. *Grantee Submission*.

Cathlyn Stone, Patrick J Donnelly, Meghan Dale, Sarah Capello, Sean Kelly, Amanda Godley, and Sidney K D'Mello. 2019. Utterance-level modeling of indicators of engaging classroom discourse. *International Educational Data Mining Society*.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th international conference on world wide web*, pages 613–624.

Deborah Tannen. 1987. Repetition in conversation: Toward a poetics of talk. *Language*, pages 574–605.

Learning Mathematics for Teaching Project. 2011. Measuring the mathematical quality of instruction. *Journal of Mathematics Teacher Education*, 14:25–47.

Gordon Wells. 1999. *Dialogic inquiry: Towards a socio-cultural practice and theory of education*. Cambridge University Press.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. Transfertransfo: A transfer learning approach for neural network based conversational agents. *CoRR*, abs/1901.08149.

Peng Xu, Hamidreza Saghir, Jin Sung Kang, Teng Long, Avishek Joey Bose, Yanshuai Cao, and Jackie Chi Kit Cheung. 2019. A cross-domain transferable neural coherence model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 678–687.

Xinnuo Xu, Ondřej Dušek, Ioannis Konstas, and Verena Rieser. 2018. Better conversations by modeling, filtering, and optimizing for coherence and diversity. In *Proceedings of EMNLP 2018*, pages 3981–3991.

Justine Zhang and Cristian Danescu-Niculescu-Mizil. 2020. Balancing objectives in counseling conversations: Advancing forwards or looking backwards. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5276–5289.

A Annotation Framework

Figure 3 shows a screenshot of our annotation interface. In the annotation framework, we used the term "active listening" to refer to uptake, since we found that active listening is more interpretable to raters, while uptake is too technical. However, the difference in terminology should not affect the annotations, since the two constructs are synonymous and we designed the annotation instructions entirely based on the linguistics and education literature on uptake. For example, the title of the instruction manual is "Annotating Teachers' Uptake of Student Ideas", and we define different levels of uptake with phrasings such as "the teacher provides evidence for following what the student is saying or trying to say", linking our definition to Clark and Schaefer (1989)'s theory on grounding. We include annotation instructions with the dataset.

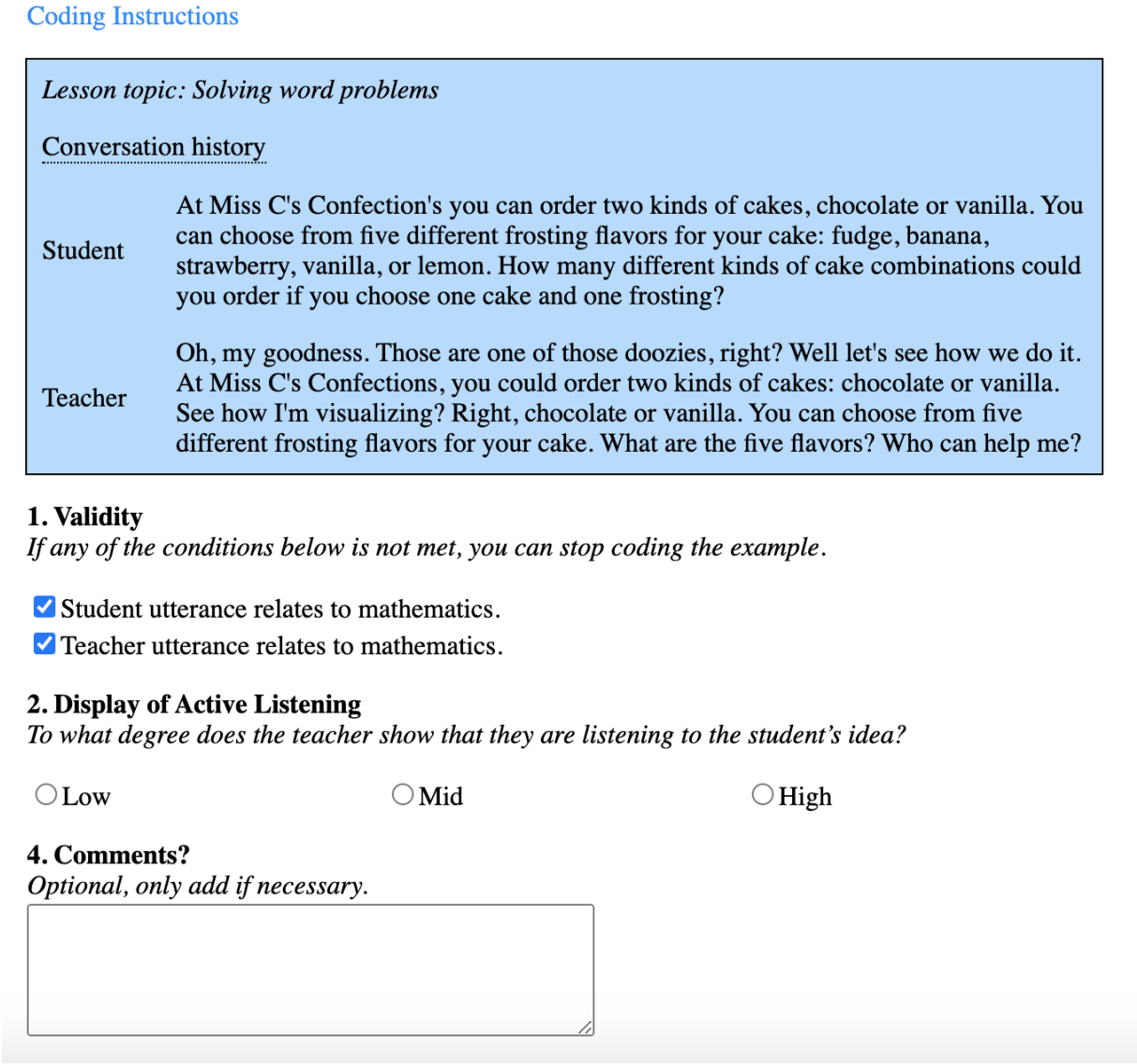

Figure 3: Screenshot of the annotation interface.

B Mapping the SWBD-DAMSL Tagset to Uptake Phenomena

We map tags from SWBD-DAMSL (Jurafsky et al., 1997) to five salient uptake phenomena: acknowledgment, answer, reformulation, collaborative completion and repetition. Table 7 summarizes our mapping. Since acknowledgment is highly frequent and it can co-occur with several other dialog acts, we consider those examples to be acknowledgments that are labeled exclusively for this phenomenon (using either the tag *b*, *bh* or *bk*).

Uptake phenomenon	**DAMSL Tags**	**% of Examples**
acknowledgment	b, bh, bk	81%
answer	tags containing "n"	13%
reformulation	bf	2%
collaborative completion	^2	2%
repetition	^m	2%

Table 7: Mapping between uptake phenomena and tags from SWBD-DAMSL (Jurafsky et al., 1997).