CIMA: A Large Open Access Dialogue Dataset for Tutoring

Katherine Stasaski
UC Berkeley
katie_stasaski@berkeley.edu

Kimberly Kao
Facebook*
kimkao957@gmail.com

Marti A. Hearst
UC Berkeley
hearst@berkeley.edu

Abstract

One-to-one tutoring is often an effective means to help students learn, and recent experiments with neural conversation systems are promising. However, large open datasets of tutoring conversations are lacking. To remedy this, we propose a novel asynchronous method for collecting tutoring dialogue via crowdworkers that is both amenable to the needs of deep learning algorithms and reflective of pedagogical concerns. In this approach, extended conversations are obtained between crowdworkers role-playing as both students and tutors. The CIMA collection, which we make publicly available, is novel in that students are exposed to overlapping grounded concepts between exercises and multiple relevant tutoring responses are collected for the same input.

CIMA contains several compelling properties from an educational perspective: student role-players complete exercises in fewer turns during the course of the conversation and tutor players adopt strategies that conform with some educational conversational norms, such as providing hints versus asking questions in appropriate contexts. The dataset enables a model to be trained to generate the next tutoring utterance in a conversation, conditioned on a provided action strategy.

1 Introduction

There is a pressing societal need to help students of all ages learn new subjects. One-on-one tutoring is one of the most effective techniques for producing learning gains, and many studies support the efficacy of conversational tutors as educational aids (VanLehn et al., 2007; Nye et al., 2014; Graesser, 2015; Ruan et al., 2019).

Tutoring dialogues should exhibit a number of important properties that are not present in existing open datasets. The conversation should be grounded around common concepts that both the student and the tutor recognize are the topics to be learned (Graesser et al., 2009). The conversation should be *extended*, that is, long enough for the student to be exposed to new concepts, givin students the opportunity to recall them in future interactions. The collection should contain varied responses, to reflect the fact that there is more than one valid way for a tutor to respond to a student at any given point in the conversation. And lastly, the dialogue should not contain personally identifiable information so it can be available as open access data.

We propose a novel method for creating a tutoring dialogue collection that exhibits many of the properties needed for training a conversational tutor. In this approach, extended conversations are obtained between crowdworkers role-playing as both students and tutors. Students work through an exercise, which involves translating a phrase from English to Italian. The workers do not converse directly, but rather are served utterances from prior rounds of interaction asynchronously in order to obtain multiple tutoring responses for the same conversational input. Special aspects of the approach are:

- Each exercise is grounded with both an image and a concept representation.
- The exercises are linked by subsets of shared concepts, thus allowing the student to potentially transfer what they learn from one exercise to the next.
- Each student conversational turn is assigned three responses from distinct tutors.
- The exercises are organized into two datasets, one more complex (Prepositional Phrase) than the other (Shape).
- Each line of dialogue is manually labeled with a set of action types.

We report on an analysis of the Conversa-

*Research performed while at UC Berkeley.

tional Instruction with Multi-responses and Actions (CIMA) dataset,[1] including the difference in language observed among the two datasets, how many turns a student requires to complete an exercise, actions tutors choose to take in response to students, and agreement among the three tutors on which actions to take. We also report results of a neural dialogue model trained on the resulting data, measuring both quality of the model responses and whether the model can reliably generate text conditioned on a desired set of tutoring actions.

2 Prior Work

2.1 Tutoring Dialogue Corpus Creation

Past work in creation of large publicly-available datasets of human-to-human tutoring interactions has been limited. Relevant past work which utilizes tutoring dialogue datasets draws from proprietary data collections (Chen et al., 2019; Rus et al., 2015) or dialogues gathered from a student's interactions with an automated tutor (Niraula et al., 2014; Forbes-Riley and Litman, 2013).

Open-access human-to-human tutoring data has been released in limited contexts. In particular, we draw inspiration from the BURCHAK work (Yu et al., 2017b), which is a corpus of humans tutoring each other with the names of colored shapes in a made-up foreign language. In each session, an image is given to help scaffold the dialogue. The corpus contains 177 conversations with 2454 turns in total. This corpus has been utilized to ground deep learning model representations of visual attributes (colors and shapes) in dialogue via interacting with a simulated tutor (Ling and Fidler, 2017; Yu et al., 2017b). Follow-up work has used this data to model a student learning names and colors of shapes using a reinforcement learning framework (Yu et al., 2016, 2017a).

Our approach differs from that of Yu et al. (2017b) in several ways, including that we tie the colored shape tutoring interactions to the more complex domain of prepositional phrases. Additionally, by using a real foreign language (Italian) we are able to leverage words with similar morphological properties in addition to well-defined grammar rules.

[1] *Cima* is Italian for "top" and a target word in the dataset. The collection is available at: https://github.com/kstats/CIMA

2.2 Learning Tutoring Dialogue Systems

Modern work in dialogue falls into two categories: chit-chat models and goal-oriented models. Chit-chat models aim to creating interesting, diversely-worded utterances which further a conversation and keep users engaged. These models have the advantage of leveraging large indirectly-collected datasets, such as the Cornell Movie Script Dataset which includes 300,000 utterances (Danescu-Niculescu-Mizil and Lee, 2011).

By contrast, goal oriented dialogue systems have a specific task to complete, such as restaurant (Wen et al., 2017) and movie (Yu et al., 2017c) recommendations as well as restaurant reservations (Bordes et al., 2017).

Neural goal-oriented dialogue systems require large amounts of data to train. Bordes et al. (2017) include 6 restaurant reservation tasks, with 1,000 training dialogues in each dataset. Multi-domain datasets such as MultiWOZ include 10k dialogues spanning multiple tasks (Budzianowski et al., 2018). For longer-term interactions, a dataset involving medical diagnosis has approximately 200 conversations per disease (Wei et al., 2018).

By contrast, prior work in the field of intelligent tutoring dialogues has widely relied on large rule-based systems injected with human-crafted domain knowledge (Anderson et al., 1995; Aleven et al., 2001; Graesser et al., 2001; **?**; **?**). Many of these systems involve students answering multiple choice or fill-in-the-blank questions and being presented with a hint or explanation when they answer incorrectly. However, curating this domain knowledge is time-expensive, rule-based systems can be rigid, and the typical system does not include multiple rephrasings of the same concept or response.

Some recent work has brought modern techniques into dialogue-based intelligent tutoring, but has relied on hand-crafted rules to both map a student's dialogue utterance onto a template and generate the dialogue utterance to reply to the student (Dzikovska et al., 2014). A limitation of this is the assumption that there is a single "correct" response to show a student in a situation.

2.3 Crowdwork Dialogue Role-Playing

Prior work has shown that crowdworkers are effective at role-playing. Self-dialogue, where a single crowdworker role-plays both sides of a conversation, has been used to collect chit-chat data (Krause et al., 2017). Crowdworkers have been effective

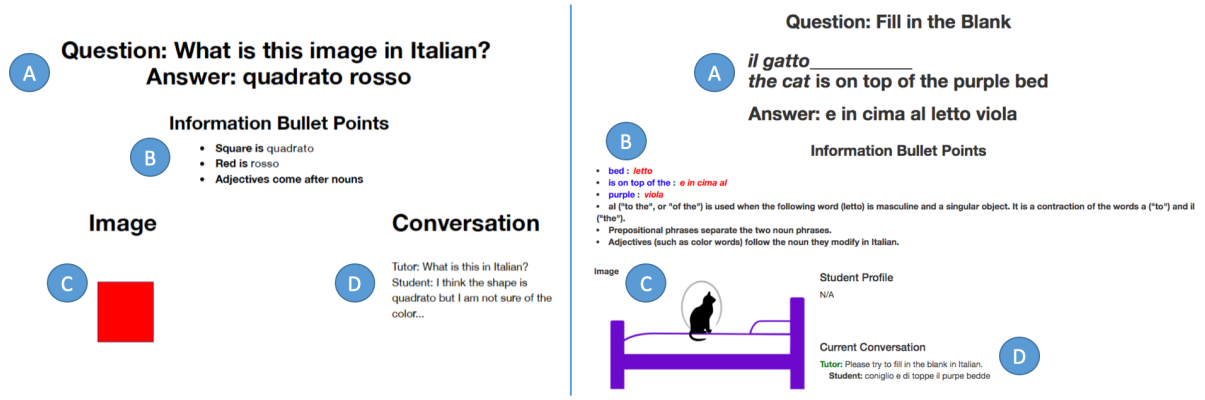

Figure 1: Example exercises as seen by a tutor (Left: Shape task, Right: Prepositional Phrase task). Shown are (A) the exercise with the correct answer that the student must produce, (B) knowledge in the form of information bullet points, (C) the image stimulus, and (D) the conversation so far. The student view is similar but does not include the information bullet points or the correct answer.

participants in peer learning studies (Coetzee et al., 2015); multiple crowdworkers can confirm lexical information within a dialogue (Ono et al., 2017).

3 Tutoring Dataset Creation

We create two dialogue datasets within CIMA: Shapes and Prepositional Phrases with colored objects.

3.1 Stimuli

We constructed stimuli for the two tasks at different levels of complexity. The Shape task follows the BURCHAK (Yu et al., 2017b) format of learning the words for adjective-noun modifiers when viewing shapes of different colors (see Figure 1, left). The Prepositional Phrase stimuli involves pairs of objects in relation to one another, with the task of learning the words for the prepositional phrase and its object, where the object is a noun with a color modifier and a determiner (see Figure 1, right).

Each stimulus consists of an image, a set of information points, and a question and answer pair. Importantly, the stimuli across the two tasks are linked by shared color terms. Intentionally including a set of common vocabulary words across datasets can potentially aid with transfer learning experiments (both human and machine). Initial tests were all done with English speakers learning the words in Italian. However, other language pairs can easily be associated with the image stimuli.

Vocabulary for the Shape task includes six colors (red, blue, green, purple, pink, and yellow) and five shapes (square, triangle, circle, star, and heart). There is only one grammar rule associated with the questions: that adjectives follow nouns in Italian.

The Prepositional Phrase task includes 6 prepositional phrases (on top of, under, inside of, next to, behind, and in front of) with 10 objects (cat, dog, bunny, plant, tree, ball, table, box, bag, and bed). Additionally, the same six colors as the Shape dataset modify the objects. Students are not asked to produce the subjects or the verbs, only the prepositional phrases. The full list of grammar rules (e.g. "l"' ("the") is prepended to the following word when it begins with a vowel) appears in Appendix A, and the full distribution of prepositional phrases, objects, and colors is in Appendix B.

3.2 Dialogue Collection with Crowdworkers

We hired crowdworkers on Amazon Mechanical Turk to role-play both the the student and the tutor. (Throughout this paper we will refer to them as students and tutors; this should be read as people taking on these roles.) In order to collect multiple tutoring responses at each point in a student conversation in a controllable way, student and tutor responses are gathered asynchronously. A diagram of this process can be seen in Figure 2. We collect several student conversations from crowdworkers with a fixed collection of hand-crafted and crowdworker-generated tutor responses. Afterwards, we show those student conversations to tutors to collect multiple appropriate crowdworker-generated responses. We then feed the newly-collected responses into the fixed collection of tutor responses for the next round of student data collection.

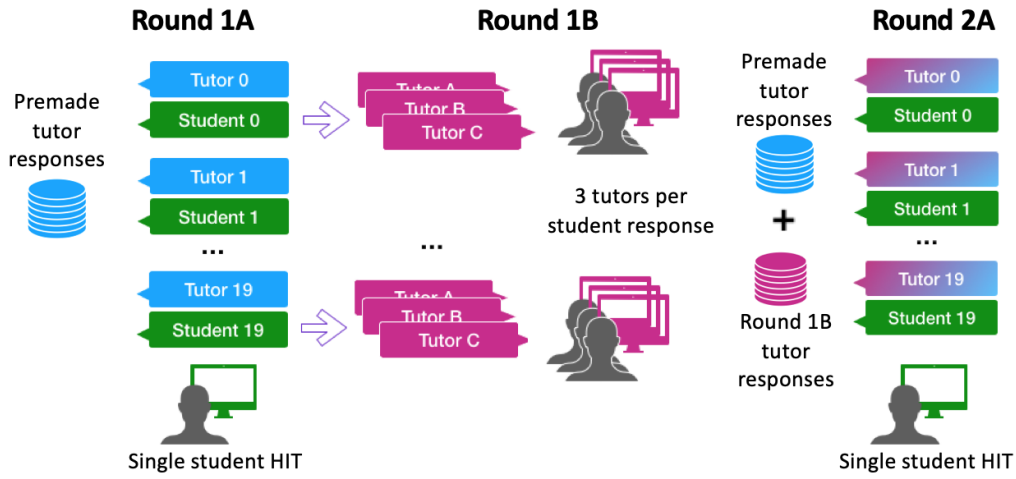

Figure 2: Progression of data collection process. In the first round of data gathering (1A), the student is exposed to 20 conversational responses from a hand-curated set of templates (shown in blue). After gathering data from 20-40 students, each student conversation is subsequently sent to three tutors to gather responses (1B). These responses (shown in pink) are placed into the pool of tutor responses for subsequent rounds (ex: 2A).

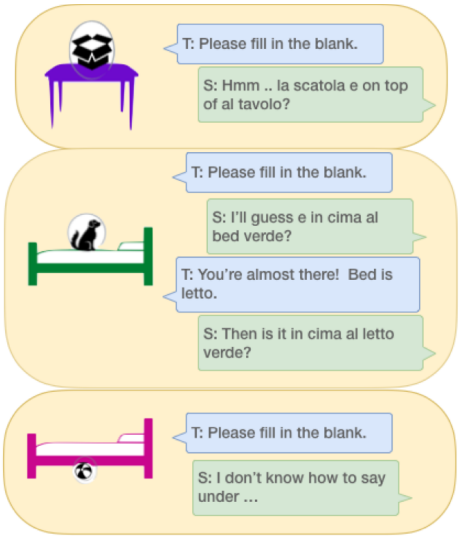

Figure 3: Example student exercise progression, showing shared features across stimuli. In this case, the student sees images for the words for bed and on top of twice within one session.

Tutors are asked to construct a *response* to the prior conversation with two outputs: the text of an *utterance* continuing the conversation and a discrete classification of the *action(s)* associated with the utterance. A summary of these actions for both the student and tutor can be seen in Table 1.

Similarly, students produce utterances which they label with actions as they work through *exercises* (defined as a question and corresponding answer, see (A) in Figure 1). Students complete as many exercises as possible in a *HIT*, defined as the crowdworking task consisting of a fixed number of turns. Each *turn* is defined as a pair consisting of the student's utterance and the most recent tutor response it is replying to. A *conversation* is defined as the set of utterances that comprise completion of an exercise.

Each participant can complete a maximum of 100 combined responses as a tutor or student for each task, to ensure diversity of responses. For the Shape task, students generate 5 responses per HIT. For the Prepositional Phrase task, however, we increase this to 20 responses per HIT due to the more complex domain.

To ensure response quality, crowdworkers were required to have 95% approval over at least 1,000 HITs. A subset of responses from each crowdworker were manually checked. We prohibited workers from copying from the prior conversation or writing a blank response. Crowdworkers were paid the equivalent of $8/hour and were required not to know Italian to participate as a student.

3.3 The Student Role

Figure 3 shows an example student interaction progression, in which students converse with the system to complete multiple exercises. Because the data collection process is asynchronous, when a student converses with the system, we serve a tutor response from a static collection to respond to them instantly. There are four rounds of data collection; in each phase, the pool of tutor responses is augmented with the student and tutor responses

Student Actions

Action Label	Description	Example
Guess	The student attempts to answer the question	"Is it 'il gatto e vicino alla scatola rosa'?"
Clarification Question	The student asks a question to the tutor, ranging from directly asking for a translated word to asking why their prior guess was incorrect.	"How would I say 'pink' in Italian?"
Affirmation	When the student affirms something previously said by the tutor.	"Oh, I understand now!"
Other	We allow students to define a category if they do not believe their utterance fits into the predefined categories.	"Which I just said."

Tutor Actions

Action Label	Description	Example
Hint	The tutor provides knowledge to the student via a hint.	"Here's a hint - "tree" is "l'albero" because l' ("the") is prepended to the following word when it begins with a vowel."
Open-Ended Question	The tutor asks a question of the student, which can attempt to determine a student's understanding or continue the conversation.	"Are you sure you have all the words in the right order?"
Correction	The tutor corrects a mistake or addresses a misconception a student has.	"Very close. Everything is correct, expect you flipped 'viola' and 'coniglio'."
Confirmation	The tutor confirms a student's answer or understanding is correct.	"Great! Now say the whole sentence, starting with the dog..."
Other	We allow tutors to define a category if they do not believe their response fits into the predefined categories.	"Correct! Although try to think of the complete word as 'la scatola.' I find that the easiest way to remember what gender everything is - I just think of the 'the' as part of the noun."

Table 1: Descriptions of Student and Tutor Actions that workers self-assign to their utterances.

from the prior round. For the Shape task, we gather responses from 20 students at each round; we increase this to 40 for Prepositional Phrase collection.

The conversation is always started by the tutor, with a pre-defined statement. For subsequent turns, we choose a tutor response conditioned on the student's most recent action, a keyword match of a student's most recent text response, and a log of what the student has been exposed to in the current conversation (details are in Appendix C). As tutor responses are gathered from crowdworkers in subsequent rounds, we add them to the collection.

3.3.1 Strategy for Student Exercise Selection

A student session is constrained to have 5 or 20 turns, depending on the task. At the start of the session, the system selects a list of stimuli for the student to work through that contains overlapping concepts (prepositions, colors, objects, shapes). From this list, one is chosen at random to show first to the student. After the student completes the exercise, if another exercise exists in the list which overlaps with at least one concept shown in the prior exercise, it is chosen next. If there is not a question with overlap, an exercise is selected at random. This process continues until the student reaches the required number of turns. An example of a resulting image chain can be seen in Figure 3.

3.3.2 Mitigating Effects of Potentially Erroneous Responses

We adopted two strategies to reduce the cost of potential errors that may arise from automatically selecting tutoring responses to show to students: (i) Student crowdworkers can explicitly indicate if the tutor response they were served does not make sense. (ii) Because there is more downside to a nonsensical answer to some kinds of student responses than others (e.g., in response to a student's question vs. to an affirmation), each student action type is assigned a probability of being served a templated vs crowdworker-collected response (details in Appendix D).

3.4 The Tutor Role

Tutors for both the Shape and Prepositional Phrase tasks complete five responses per HIT. Because the data collection is asynchronous, the tutor is responding not to five consecutive utterances from the same student, but rather to five different students' conversations.

To ensure good coverage, we inject three different tutors at each utterance within a student's conversation. This allows redundant generation of

Action	Shape	Prepositional Phrase
Student Actions		
Guess	448	1318
Question	313	840
Affirmation	289	406
Other	12	12
Tutor Actions		
Question	882	824
Hint	1002	1733
Correction	534	828
Confirmation	854	436
Other	37	59

Table 2: Distribution of student and tutor actions across the two datasets; multiple actions can be associated with each utterance.

tutor responses to the same student input. We show the tutor the entire conversation up to that point.[2]

To role-play a tutor, crowdworkers were not expected to have any proficiency in Italian. To simulate the knowledge a tutor would have, we show relevant domain information so the tutor could adequately respond to the student (see Figure 1(B)). This includes vocabulary and grammar information which are necessary to answer the question. This domain-specific information can also be used as input knowledge to inform a learning system. In the Prepositional Phrase task, we also showed summaries of prior student conversations, but do not describe this in detail due to space constraints.

4 Dataset Statistics

An analysis of the conversations found that the data contains several interesting properties from an educational perspective. This section summarizes overall statistics of the data collected; the subsequent two sections summarize phenomena associated with the student and tutor data.

4.1 Shape Dataset

A total of 182 crowdworkers participated in the Shape data collection process: 111 as tutors and 90 as students. 2,970 tutor responses were collected, responding to 350 student exercises. A student required an average of 3.09 (standard deviation: 0.85) turns to complete an exercise. The average student turn was 5.38 (3.12) words while the average tutor response length was 7.15 (4.53) words. 4.0% of tutor responses shown to students were explicitly flagged by the student as not making sense. Table 2 shows the distribution of action types.

4.2 Prepositional Phrase Dataset

A total of 255 crowdworkers participated in the creation of Prepositional Phrase data: 77 as students who completed a total of 391 exercises, and 209 as tutors who completed 2880 responses. The average number of turns a student requires before answering a question correctly is 3.65 (2.12). Of the tutor responses served to students, 4.2% were manually flagged as not making sense. The average student utterance is 6.82 words (2.90) while the average length of a tutor utterance is 9.99 words (6.99).

We analyze the proportion of tutoring responses which include the direct mention of an Italian color word, English translation of a color word, or "color," as this is the domain component which overlaps with the Shapes task. Of the set of tutor responses, 1,292 (40.0%) include a direct mention, indicating substantial overlap with the Shapes task.

5 Student Interactions

By examining a student's interactions over the course of a 20-turn HIT, we find that students take fewer turns on average to complete an exercise at the end than at the beginning of a HIT. We examine the number of turns students take before reaching the correct answer, as we hypothesize this will decrease as students have more exposure to domain concepts. We note this could be due to many factors, such as the students becoming more comfortable with the task or system or learning Italian phrases they were exposed to in prior questions.

We measure this with the prepositional phrase domain, because the students interacted with the system for 20 turns, compared to the 5-turn interactions with the Shape task. For a given HIT, we compare the number of student turns needed to produce their first correct answer with how many turns were needed for their final correct answer.[3]

For each student, we calculate the difference between the number of turns required between their first and final correct answers. The average difference is -0.723, indicating students required fewer turns to achieve their last correct answer than their first. Thus the data set might contain evidence of learning, although it could be as simple as student

[2] If a student conversation is longer than 10 turns, or if any point of the conversation has been marked as not making sense, the conversation is not shown to tutors.

[3] Note the final correct question might not be the final question the student attempted to answer, as the HIT is finished at 20-turns regardless of the state of a student's conversation.

workers learning how to more efficiently ask questions of the system.

6 Tutor Phenomena

We examine several characteristics about the tutor interactions: (i) properties about the language tutors use in their responses, (ii) how tutors respond to different student action types, and (iii) characterizing if and how tutors agree when presented with identical student input.

6.1 Tutoring Language

One feature of our dataset construction is the progression from the relatively simple Shape task to the linguistically richer Prepositional Phrase task. We analyze the resulting tutoring responses to see if more complex tutoring language emerges from the syntactically richer domain. We measure complexity in terms of number of non-vocabulary terms (where vocabulary refers to the words that are needed in the task, such as "rosa" for "pink").

We examine the set of tutoring responses from each domain. For each utterance, we remove Italian vocabulary, English translations, and stop words. We further restrict the utterance to words included in the English language[4] to remove typos and misspellings.

We find an average of 2.34 non-domain words per utterance (of average length 9.99 words) in the Prepositional Phrase dataset, compared to 0.40 per utterance (of average length 7.15 words) in the Shape dataset. While accounting for the average difference in length between the two datasets, the Prepositional Phrase dataset results in more non-domain English words than the Shape dataset.

This supports our hypothesis that the added domain complexity makes the Prepositional Phrase collection richer in terms of tutoring language than related work such as Yu et al. (2017b).

6.2 Tutor Response to Student Actions

We additionally examine the tutor action distributions conditioned on the student action taken immediately prior for the Prepositional Phrase dataset. We hypothesize if a student utterance is classified as a question, the tutor will be more likely to respond with the answer to the question (classified as a hint), conforming to conversational expectations. This is supported by the distributions, seen in Figure 4. For other student action type responses (e.g., guess, affirmation), we observe that the tutor actions are more evenly distributed.

[4]Stopwords and English vocabulary as defined by NLTK's stop words and English corpus, https://www.nltk.org/

Tutor Action(s)	Agreement	Individual
Hint	81.1%	39.0%
Question	5.7%	12.5%
Correction	5.2%	12.1%
Hint/Correction	2.8%	8.1%
Confirmation	2.8%	6.2%
Question/Hint	1.4%	7.5%
Correction/Confirmation	0.9%	2.1%
Total	**212**	**2880**

Table 3: Distribution of action sets agreed on by 3-tutor groups. Included are the proportion of individual tutor utterances labeled with each set of actions over the entire dataset for comparison.

6.3 Tutor Action Agreement

As there are three tutors responding to each student utterance, we analyze the conditions in which the tutors agree on a unified set of actions to take in response to a student (in the Prepositional Phrase task). In particular, when all three tutors take the same set of action types we measure (i) which action(s) are they agreeing on and (ii) which action(s) the student took in the prior turn.

In 212 out of 1174 tutor tasks, all 3 tutors agreed on the same set of actions to take. We show the distribution of these 212 cases over unified tutor action sets in Table 3. There is a particularly high proportion of agreement on giving hints compared to other action sets. While hint was the most common action taken by tutors compared to the next-highest action by 26.5%, tutor agreement on hint was the most common by 75.4% compared to the next-highest category, a 2.8 times larger difference.

Additionally, we examine how a student's most recent action might influence a group of tutor's potential for action agreement. We measure the proportion of tutor agreement on a unified action set per student action set (the analysis is restricted to student action sets with at least 10 examples). Results can be seen in Table 4.

We note the highest agreement occurs after a student has made a Question or Question/Affirmation. This is consistent with (i) the high likelihood of a tutor to give a hint in response to a question (Figure 4) and (ii) the high proportion of tutor agreement on hints (Table 3). On the contrary, there is relatively low agreement when a student makes a Guess, consistent with the more evenly-distributed tutor action distribution (Figure 4).

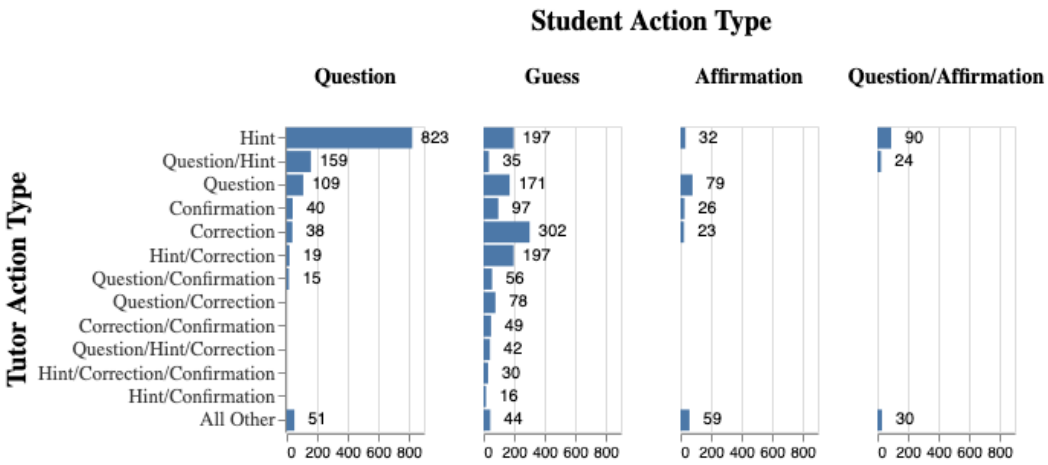

Figure 4: Distribution of tutor action classifications, grouped by the most recent set of student actions. The "All Other" category represents the combination of tutor action sets with fewer than 15 items.

Student Action(s)	Tutor Agreement
Question	36.8%
Question/Affirmation	37.5%
Affirmation	12.3%
Guess	6.4%
Guess/Affirmation	5.6%

Table 4: For each student action(s), percentage of tutor groups who agree on a unified action set in response.

7 Tutoring Model

We claim CIMA is useful to train neural models for tutoring tasks. To explore this, we train a Generation model (GM) aimed at producing a tutoring response conditioned on two past conversation utterances.[5] An example input would be:

> Hint, Correction, e di fronte al, giallo, coniglio, is in front of the, yellow, bunny, <EOC> Tutor: Well, "bunny" is "coniglio" Student: il gatto e di fronte al coniglio.

In this representation, domain information and an intended set of actions to take is separated with a special token <EOC> from two sentences of conversation. Model training details are in Appendix E. We split the data along conversations into 2296 train, 217 development, and 107 test utterances.

7.1 Generation Quality Results

One benefit of CIMA is the ability to compare generated text to multiple distinct reference sentences in order to measure quality. We apply two standard generation quality measures: BLEU (**?**) and BERT F1 Score (**?**), using the maximum score of the model's response compared to each of the three human-generated tutor responses for a turn in the conversation. We compare the quality of the GM's responses to Round 1A of the same rule-based system used to collect CIMA (see Appendix C).

Model	BLEU	BERT F1
Rule-Based Baseline	**0.34**	0.45
Generation Model	0.31	**0.53**

Table 5: Generation quality results comparing a rule-based baseline to the neural Generation model.

Results can be seen in Table 5. We note the rule-based baseline (which is guaranteed to be grammatical) performs slightly better than GM on BLEU score (which incentivizes exact word overlap) but that GM performs higher on BERT F1 Score (which incentivizes semantic word overlap). Given the comparable BLEU score and the gain on BERT F1 Score, we conclude that using CIMA to train a neural model can produce tutoring utterances of reasonable quality.

7.2 Action Evaluation Results

In addition to quality, we examine whether the Generation model is able to generate utterances consistent with the set of actions it is conditioned on. We train a separate Action Classifier (AC) to predict a set of actions from a tutoring utterance. For example, for the input *Tutor: Well, "bunny" is "coniglio." Do you know the word for yellow?* the classifier would output `Hint Question.` Training details appear in Appendix E.

To examine the classifier's reliability, we measure the F1 for the test set, both overall and for each

[5] As we did not see a gain in quality when including the full conversation, we simplify the task to responding to the most recent tutor and student utterance.

Overall	Question	Hint	Corr.	Conf.
0.72	0.85	0.93	0.67	0.61

Table 6: Action Classification model F1 scores for the test set, where the Overall metric is weighted by class.

of the four top action categories (excluding Other due to the low number of utterances). Results can be seen in Table 6. While the Overall, Hint, and Question F1 are relatively high, we note the lower Correction and Confirmation scores.

Using the classifier, we measure the GM's ability to generate utterances consistent with the set of actions it is conditioned on. For each item in the test set, we sample one of the three tutor's responses, identify the action(s) that tutor chose to make, and use GM to generate an utterance conditioned on that action type. To determine if the generated utterance is of the correct action type, we apply the classifier model. The average accuracy over the test set is 89.8%, indicating GM's ability to generate utterances consistent with an action strategy.

8 Discussion

Our analysis finds that tutors are more unified in an action strategy when a student asks a question than other actions. This is consistent with the findings that (i) when tutors agree, they are more likely to agree on a hint and (ii) the most likely action in response to a student question is a hint. Overall tutor agreement was low among the dataset (18.1%), indicating the potential capture of divergent tutoring strategies. Future work can leverage this disagreement to explore the multiple potential actions to take when responding to a student.

Our preliminary experiments show CIMA can be used to train a model that can generate text conditioned on a desired actions. Future work should explore more complex models utilizing CIMA, as well as exploring the other unique qualities of the collection, such as the shared image representation, multiple tutoring utterances for each conversation, and link between the two domains.

Tutoring responses marked as not making sense should be explored, to both improve the process of serving student responses as well as correcting a model when a generated response veers the conversation off track. A benefit to having this explicitly logged is that the collection contains labeled negative examples of tutoring responses, which can be leveraged in training models.

9 Limitations

While past work utilized crowdworkers to collect tutoring utterances (Yu et al., 2017b) and for peer learning studies (Coetzee et al., 2015), future work should examine the similarities and differences between the language and actions taken by crowdworkers and actual tutors and students engaged in the learning process.

Because we were working with untrained crowdworkers, we were constrained in the complexity of language learning concepts we could include in CIMA. It is possible that the resulting dataset only transfers to novice language learners. Future work should examine how well this generalizes to a real language learning setting and how general tutoring language and strategies that emerge from our domain transfer to more complex ones.

The dataset currently does not distinguish the type of hint or correction tutors make. Examples include providing direct corrections versus indirect feedback which states the error and allows the student to self-correct (**?**). Future work on CIMA can examine the prevalence of these different types of feedback and potential benefits or shortcomings.

10 Conclusion

We present CIMA: a data collection method and resulting collection of tutoring dialogues which captures student interactions and multiple accompanying tutoring responses. Two datasets of differing complexity have direct applicability to building an automatic tutor to assist foreign language learning, as we examine with a preliminary model. CIMA has the potential to train personalized dialogue agents which incorporate longer-term information, have a well-defined goal to have a student learn and recall concepts, and can explore different correct utterances and actions at given times.

Acknowledgements

This work was supported by an AWS Machine Learning Research Award, an NVIDIA Corporation GPU grant, a UC Berkeley Chancellor's Fellowship, and a National Science Foundation (NSF) Graduate Research Fellowship (DGE 1752814).We thank Kevin Lu, Kiran Girish, and Avni Prasad for their engineering efforts and the three anonymous reviewers for their helpful comments.

References

Vincent Aleven, Octav Popescu, and Kenneth R Koedinger. 2001. A tutorial dialogue system with knowledge-based understanding and classification of student explanations. In *Working Notes of 2nd IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems*.

John R Anderson, Albert T Corbett, Kenneth R Koedinger, and Ray Pelletier. 1995. Cognitive tutors: Lessons learned. *The journal of the learning sciences*, 4(2):167–207.

Antoine Bordes, Y-Lan Boureau, and Jason Weston. 2017. Learning end-to-end goal-oriented dialog. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gasic. 2018. Multiwoz - a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026. Association for Computational Linguistics.

Guanliang Chen, David Lang, Rafael Ferreira, and Dragan Gasevic. 2019. Predictors of student satisfaction: A large-scale study of human-human online tutorial dialogues. In *Proceedings of the 12th International Conference on Educational Data Mining, EDM 2019, Montréal, Canada, July 2-5, 2019*. International Educational Data Mining Society (IEDMS).

D. Coetzee, Seongtaek Lim, Armando Fox, Bjorn Hartmann, and Marti A. Hearst. 2015. Structuring interactions for large-scale synchronous peer learning. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '15, page 1139–1152, New York, NY, USA. Association for Computing Machinery.

Cristian Danescu-Niculescu-Mizil and Lillian Lee. 2011. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, pages 76–87, Portland, Oregon, USA. Association for Computational Linguistics.

Myroslava O. Dzikovska, Natalie B. Steinhauser, Elaine Farrow, Johanna D. Moore, and Gwendolyn E. Campbell. 2014. Beetle ii: Deep natural language understanding and automatic feedback generation for intelligent tutoring in basic electricity and electronics. *International Journal of Artificial Intelligence in Education*, 24:284–332.

Kate Forbes-Riley and Diane Litman. 2013. When does disengagement correlate with performance in spoken dialog computer tutoring? *Int. J. Artif. Intell. Ed.*, 22(1–2):39–58.

Arthur C. Graesser. 2015. Conversations with autotutor help students learn. *International Journal of Artificial Intelligence in Education*, 26:124–132.

Arthur C Graesser, Sidney D'Mello, and Natalie Person. 2009. Meta-knowledge in tutoring. *The educational psychology series. Handbook of metacognition in education*, pages 361–382.

Arthur C. Graesser, Kurt VanLehn, Carolyn Penstein Rosé, Pamela W. Jordan, and Derek Harter. 2001. Intelligent tutoring systems with conversational dialogue. *AI Magazine*, 22:39–52.

Ben Krause, Marco Damonte, Mihai Dobre, Daniel Duma, Joachim Fainberg, Federico Fancellu, Emmanuel Kahembwe, Jianpeng Cheng, and Bonnie Webber. 2017. Edina: Building an open domain socialbot with self-dialogues. *Alexa Prize Proceedings*.

Huan Ling and Sanja Fidler. 2017. Teaching machines to describe images via natural language feedback. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 5075–5085, Red Hook, NY, USA. Curran Associates Inc.

Nobal Niraula, Vasile Rus, Rajendra Banjade, Dan Stefanescu, William Baggett, and Brent Morgan. 2014. The DARE corpus: A resource for anaphora resolution in dialogue based intelligent tutoring systems. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3199–3203, Reykjavik, Iceland. European Language Resources Association (ELRA).

Benjamin D Nye, Arthur C Graesser, and Xiangen Hu. 2014. Autotutor and family: A review of 17 years of natural language tutoring. *International Journal of Artificial Intelligence in Education*, 24(4):427–469.

Kohei Ono, Ryu Takeda, Eric Nichols, Mikio Nakano, and Kazunori Komatani. 2017. Lexical acquisition through implicit confirmations over multiple dialogues. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 50–59, Saarbrücken, Germany. Association for Computational Linguistics.

Sherry Ruan, Liwei Jiang, Justin Xu, Bryce Joe-Kun Tham, Zhengneng Qiu, Yeshuang Zhu, Elizabeth L. Murnane, Emma Brunskill, and James A. Landay. 2019. Quizbot: A dialogue-based adaptive learning system for factual knowledge. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, New York, NY, USA. Association for Computing Machinery.

Vasile Rus, Nabin Maharjan, and Rajendra Banjade. 2015. Unsupervised discovery of tutorial dialogue modes in human-to-human tutorial data. In *Proceedings of the Third Annual GIFT Users Symposium*, pages 63–80.

Kurt VanLehn, Arthur C Graesser, G Tanner Jackson, Pamela Jordan, Andrew Olney, and Carolyn P Rosé. 2007. When are tutorial dialogues more effective than reading? *Cognitive science*, 31(1):3–62.

Zhongyu Wei, Qianlong Liu, Baolin Peng, Huaixiao Tou, Ting Chen, Xuanjing Huang, Kam-Fai Wong, and Xiangying Dai. 2018. Task-oriented dialogue system for automatic diagnosis. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 201–207. Association for Computational Linguistics.

Tsung-Hsien Wen, David Vandyke, Nikola Mrkšić, Milica Gašić, Lina M. Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2017. A network-based end-to-end trainable task-oriented dialogue system. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 438–449, Valencia, Spain. Association for Computational Linguistics.

Yanchao Yu, Arash Eshghi, and Oliver Lemon. 2016. Training an adaptive dialogue policy for interactive learning of visually grounded word meanings. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 339–349, Los Angeles. Association for Computational Linguistics.

Yanchao Yu, Arash Eshghi, and Oliver Lemon. 2017a. Learning how to learn: An adaptive dialogue agent for incrementally learning visually grounded word meanings. In *Proceedings of the First Workshop on Language Grounding for Robotics*, pages 10–19, Vancouver, Canada. Association for Computational Linguistics.

Yanchao Yu, Arash Eshghi, Gregory Mills, and Oliver Lemon. 2017b. The BURCHAK corpus: a challenge data set for interactive learning of visually grounded word meanings. In *Proceedings of the Sixth Workshop on Vision and Language*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.

Zhou Yu, Alan W. Black, and Alexander I. Rudnicky. 2017c. Learning conversational systems that interleave task and non-task content. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, page 4214–4220. AAAI Press.

A Prepositional Phrase Collection Grammar Rules

Listed below is the complete collection of Prepositional Phrase grammar rules:

- "il" ("the") is used for when the following word is masculine.
- "alla" ("to the") is used when the following word is feminine and a singular object. It is a contraction of a ("to") and la ("the").
- "al" ("to the") is used when the following word is masculine and a singular object. It is a contraction of the words a ("to") and il ("the").
- "l'" ("the") is prepended to the following word when it begins with a vowel.
- "all" ("to the" is prepended to the following word when it begins with a vowel. This is a contraction of al ("to") and l' ("the").
- "rossa" is the feminine form of red because the noun it modifies is feminine
- Adjectives (such as color words) follow the noun they modify in Italian.
- Prepositional phrases separate the two noun phrases.

B Phrase Breakdown

Table 7 shows the coverage of Prepositional Phrase exercises over the potential objects, prepositional phrase, and colors.

Category	Number of Questions
First Object (students don't translate these)	
'the dog'	134
'the cat'	161
'the plant'	49
'the bunny'	129
'the ball'	47
'the bag'	28
'the box'	17
Prepositional phrases	
'is in front of the'	126
'is next to the'	106
'is inside of the'	74
'is under the'	73
'is behind the'	127
'is on top of the'	59
Colors	
'green'	88
'pink'	77
'blue'	100
'yellow'	91
'red'	101
'purple'	108
Second Object	
'tree'	128
'box'	160
'plant'	14
'cat'	13
'bunny'	49
'dog'	23
'bed'	69
'table'	89
'bag'	20

Table 7: Phrase Breakdown of Student conversations

C Algorithm

Algorithmic specifications for data collection can be viewed in Figure 5. In order to serve a tutor-crafted response to a student, we match the current student utterance to a prior-collected student utterance which has been responded to by a tutor. The most similar student utterance is determined by maximizing word overlap of the student's most recent utterance to the prior-collected student utterances, excluding domain vocabulary words. The English and Italian words are replaced with the information relevant to the current exercise in the associated tutor utterance before showing this to the student.

D Hand-Crafted Response Probabilities

Throughout different rounds of data collection, we balance the probability of a student receiving a pre-made tutor response with a crowdworker-generated response from a prior round of data collection. As we collect more tutoring responses in subsequent rounds, the probabilities shift from pre-made, safe choices to the crowd-worker generated responses, because with more data, the choices should be more likely to more closely match a student utterance. The probabilities were manually set and can be seen in Table 8.

	G	Q	A	O
Shape	1.0	1.0	0.0	0.5
PP R1	1.0	1.0	1.0	1.0
PP R2	0.75	0.75	0.5	0.5
PP R3	0.75	0.75	0.5	0.5
PP R4	0.65	0.65	0.4	0.4

Table 8: Probabilities for serving hand crafted responses instead of tutor-provided responses for the shape and the prepositional phrase task, rounds 1 - 4. for Guess, Question, Action, and Other student question types.

E Model Training Details

For the Generation Model, we use OpenNMT (**?**) to train a 4-layer LSTM of size 1000 with global attention. We use the Adam optimizer (**?**) with a learning rate of 0.001. We allow the model to have a copy mechanism to copy relevant words (such as translation information) from the input (**?**). We use 300-dimensional pre-trained GloVe embeddings (**?**), which are allowed to be updated throughout training. At decode time, we replace unknown words with the input word with the highest attention.

We train the Action Classification model using OpenNMT (**?**). The model is a 4-layer bidirectional LSTM with 1024 hidden state size, general attention, a learning rate of 0.001, and batch size of 16. It utilizes pre-trained 300-dimensional GloVe embeddings (**?**) which can be updated. This model is trained on the same training set as the generation model, taking in the human-created utterances and predicting the corresponding classifications.

```
if Guess and Correct then
    Move on to next question
else if Guess and Incorrect then
    Flip a coin with probability = G
    if Heads then
        Compile a list of pre-defined responses containing
        vocabulary missing from the student response
        Randomly select from this list
    else
        Return the most similar past-tutor response from the
        set of responses of type G.
    end if
end if


if Question then
    Flip a coin with probability = Q
    if Heads then
        Attempt to find words from a set list of pre-defined
        hints associated with each vocabulary word in the
        question.
        if Match is Found then
            Serve that hint
        else
            Choose a random hint that the student has not
            seen and serve that.
        end if
    else
        Return the most similar past-tutor response from the
        set of responses of type Q.
    end if
end if


if Affirmation or Other then
    Flip a coin with probability = A / O
    if Heads then
        Flip a coin with probability = 0.5
        if Heads then
            Ask the student for an attempt at a guess.
        else
            Give a pre-defined hint for a vocabulary or gram-
            mar concept that the student has not yet seen.
        end if
    else
        Return the most similar past-tutor response from the
        set of responses of type A/O.
    end if
end if
```

Figure 5: Algorithm for serving tutoring responses.