Building Bridges Between Student and Curricular Language: Creating a Corpus of Abstract Meaning Representations for the Classroom

Kristin Wright-Bettner[1], **Jon Z. Cai**[1], **Zekun Zhao**[2], **James H. Martin**[1], **Jeffrey Flanigan**[2], **Martha Palmer**[1]
[1] University of Colorado Boulder, [2] University of California Santa Cruz
first.last@colorado.edu

Abstract

The potential of AI conversational agents to foster student learning and reduce teacher strain in classroom settings has made the development of pedagogical agents a prime research target. An effective AI agent in particular must be able to understand both student language and the content they are learning and, furthermore, map between them. Curricular terminology and student speech, though topically and semantically related, differ significantly in surface-form expression. We present the JIA-AMRs Collection, a new resource for exploring whether Abstract Meaning Representations (AMRs) can optimize interventions by a conversational AI agent in a middle-school classroom by providing structured semantic representations of classroom language. This resource also provides an avenue by which we can verify interventions by the agent. We discuss the challenges of creating a corpus of meaning representations that map across highly-dissimilar classroom data (multimedia curriculum, student spoken language, and student written language) and our promising results of a nearly 30-point gain in trained-parser performance over the off-the-shelf model.

Keywords: Corpus Creation and Annotation, Meaning Representation, Semantics, Parsing, Training and Domain Adaptation

1. Introduction

Conversational AI agents have been proposed as one possible solution to the heavy demands on teachers in classrooms (Labadze et al., 2023). For example, an interactive AI agent could enhance learning by simulating teacher behaviors, such as intervening when students are confused, or reminding them of key curricular concepts.

The work in this paper was initiated as part of our NSF Institute for AI-Student Teaming, iSAT,[1] which is aimed at supporting more effective collaboration in classroom environments (D'Mello et al., 2024). For our conversational AI partner, we focused on *jigsaw* activities (Liao et al., 2018). *Jigsaws* are collaborative learning techniques in which each student in a group first learns a different part of the learning materials and then the students gather together to discuss tasks or projects that require them to share their knowledge (Aronson et al., 1978). These are typically done in break-out groups, as many as 8 or 10 per classroom, making it difficult for a single teacher to supervise all of the groups simultaneously. We were interested in exploring whether AI conversational agents could assist the teacher in supporting *jigsaw* groups, hence JIA (Jigsaw Interactive Agent), and chose the Schoolwide Labs Sensor Immersion curriculum unit (discussed in Section 3).

One of our first challenges involved the highly-contextualized semantic content of classroom discourse that a successful AI agent must be able to parse correctly. As an example, consider the following middle-school dialogue fragment:

1. Student 1: *In button A*
2. Student 2: *I- I- on like in on button A*
3. Student 1: *On button a?*
4. Student 2: *The in like the block*
5. Student 1: *What the? In or on?*

Without background knowledge of the course material, this exchange appears opaque, even nonsensical. However, the conversation is meaningful and on-topic; the students are participating in a task in which they are asked to describe blocks of code in MakeCode, a blocks-based programming platform that is part of their curriculum. Here, Student 2 is attempting to describe the code block in Figure 1, but Student 1 is confused, creating a prime opportunity for an agent to engage. To intervene effectively,

Figure 1: MakeCode block

an AI agent itself needs to follow the conversation. It should be able to correctly interpret utterance (2),

[1] https://www.colorado.edu/research/ai-institute/

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 11960–11971
11–16 May 2026.

for example, as follows:

- The first few words are disfluency
- "in" is part of meaningful predicate-argument structure – something is located in something else (the code block in Figure 1)
- "on button A" is not a prepositional phrase but a named entity (NE), a reference to the code block in Figure 1. Identifying this means a system must have knowledge of the curriculum the students are learning.

It is useful to explore how the Abstract Meaning Representation (AMR) (Banarescu et al., 2013) represents this utterance. AMRs are rooted, directed, acyclic graphs that can capture all of this information in a structured way: disfluency is dropped; "in" is captured by a locative frame' and "on button A" is accurately represented:

```
(b / be-located-at-91
   :ARG2 (s / software
      :name (n / name :op1 "on" :op2 "button"
         :op3 "A")))
```
[2]

We claim that AMRs are desirable for such data for at least three reasons. First, AMRs have been shown to improve large language model (LLM) performance in other dialogue-understanding tasks such as dialogue evaluation (Yang et al., 2024) and summarization (Hua et al., 2023), so we hypothesize they could also optimize understanding of classroom dialogue semantics by an LLM-powered AI agent. Second, because AMRs are traceable directly to the source text, they provide an avenue for verifying and explaining an agent's responses. It is well known that LLMs are prone to hallucinations (Li et al., 2024), making evaluative tools crucial (Zhu et al., 2024), such as the way AMRs have been used to validate medical summaries (Landes et al., 2025). This validation is particularly vital in the real-world application of youth education – an AI agent that propagates inaccuracies causes more harm than benefit. Finally, AMRs lay the groundwork for integration with visual processing and situational grounding to support multimodal functionality (Tam et al., 2023, e.g.), which is the explicit long-term goal of our student discourse analysis. The predicate-argument structure represented by an AMR parse plays a central role in neurosymbolic approaches such as Common Ground Tracking for Situated Multimodal Dialogues (Lai et al., 2025).

We therefore manually created the JIA-AMRs Collection, a new gold-standard database of 4,711 English-language AMRs to support further research and evaluation in these areas.[3] The Collection comprises three diverse but related datasets: (1) a middle-school science curriculum; (2) dialogues of the students collaborating to answer questions about the curriculum content; and (3) the students' finalized (typed) answers to those questions. It also includes document-level coreference relations for a subset of each of these corpora. The datasets include classroom conversations and lab study conversations collected under IRB approval by members of iSAT (Doherty et al., 2025).

In this paper, we discuss the challenges of creating consistent and comprehensive meaning representations across such specialized and diverse source data. In part, our solution included adopting the two-pass annotation strategy for specialized domains presented in Cai et al. (2024). This proved effective, yielding a nearly 30-point gain in parser performance over the off-the-shelf model.

2. Background

2.1. AMRs and Related Work

AMRs are graphical representations of single-sentence semantic phenomena, including but not limited to semantic roles, coreference, and named entities (NEs). In general, events are represented by PropBank frames (Palmer et al., 2005), and semantic relations are identified by a frame's predefined numbered arguments or one of the roles from AMR's inventory[4]. Figure 2 demonstrates several of these features. *Mary* is identified as the *learner* by filling the ARG0 slot of learn-01, while "programming her robot" is the thing she is learning (ARG1). Mary is also the *programmer* and the *possessor* of the robot. AMR captures this coreference by repeating the variable associated with Mary, (*p*).

AMR can also capture coreference relations at the document level (O'Gorman et al., 2018; Naseem et al., 2022), including with arguments that are implicit at the sentence level. For example, if a later sentence mentions Mary's teacher, that explicit concept can be linked to the implicit ARG2 slot of learn-01 in the Figure 2 graph.

The AMR schema has been adapted for a variety of domains and applications, including search-and-rescue navigation tasks (Bonial et al., 2024), clinical-text understanding (Cai et al., 2024), grounded representation of spatial language (Bonn et al., 2020), and even non-linguistic semantics such as gestures (Brutti et al., 2022; Donatelli et al., 2022; Lai et al., 2024). There has also been an exploratory study that uses German AMRs to evaluate Language learner responses (Dellert, 2020). However, to our knowledge, this is the first time AMRs

[2] This is the PENMAN-style representation of the graph, which is used for human annotation (Banarescu et al., 2013).

[3] The de-identified, annotated transcripts and typed responses are available at https://github.com/NSF-iSAT/JIA-AMR-for-classroom

[4] https://umr4nlp.github.io/web/amr/lib/roles.html

((l / learn-01 :ARG0 (p / person :name (n / name :op1 "Mary")) :ARG1 (p2 / program-01 :ARG0 p :ARG1 (r / robot) :poss p)))	*learn-01* learn, absorbing information ARG0-PAG: student ARG1-PPT: subject ARG2-DIR: teacher

Figure 2: AMR for the sentence *Mary is learning to program her robot* (left) and example PropBank frame (right).

have been applied to multi-party student conversations in a classroom setting for a specific curriculum unit. Furthermore, we believe the dialogue subset of our corpus pioneers AMR annotation of human-human spoken dialogue with typical turn-taking. While there have been several ventures into dialogue AMRs, these have either been human-robot (Bonial et al., 2020, 2021, 2024), typed dialogue (Bonn et al., 2020), or dialogue that is in fact primarily one-way; several projects have utilized the EGGNOG corpus (Wang et al., 2017). It includes speech data from a two-party setting, consisting mainly of one person (the signaler) telling the second person (the actor) how to build a structure out of blocks, such that Brutti et al. (2022) says the speech is "largely one-way communication" and "does not follow typical conversational turn-taking patterns." Lai et al. (2024) only annotate the signaler's speech for this reason. Interestingly, all three of these other AMR-dialogue collections are highly command-based: with a commander and a follower. Our dialogue corpus, by contrast, is sourced from multi-party, highly-collaborative settings with no designated leader or follower. This early effort therefore breaks the ground for developing AMR to better support representation of more natural human conversational semantics. This paper should be read in conjunction with the paper discussing the dialogue annotation, (Cai et al., 2025a).

2.2. Connecting the Dots: the Varied Landscape of Classroom Language

Semantic annotation is often done on datasets consisting of homogeneous data types and linguistic styles. AMR was originally developed on data like Wall Street Journal articles, which typically contain complete sentences, are self-contained, and are written for a general audience in a narrative style. Little or no background knowledge is required for annotators to interpret the semantic content, and such texts can simply be extracted from a source file and sentence-segmented for annotation.

The data required for an AI interactive agent in a classroom are more complex, specialized, and varied. Transcripts of classroom dialogue are largely incomprehensible (to human annotators or AI agents) without supplemental knowledge about the dialogue subject matter and physical environment. Furthermore, like many curricula today, our data is multimedia in nature: Key content is scattered across PowerPoints, Word documents, and online tutorials, as well as embedded in images and tables, non-linguistic modes of communication. These phenomena must be taken into account and mapped to create a coherent semantic framework.

For example, our curriculum[5] teaches students how to wire scientific sensors to computers. A successful AI agent must first interpret the specialized, elliptical language in the following dialogue[6] in which Student-4 reports that the red wire connects the power to the 3V3 electrical port:

1. Student-4: *We only had two set wires Red and black*
2. Student-1: *And what did the red connect?*
3. Student-1: *What what port?*
4. Student-4: *Three, the red was the power 3V3*

The agent could then either confirm Student-4's information or nudge them in a different direction if it is incorrect[7], which means being able to map to a knowledge base (KB) about the subject matter.

However, for a multimedia source curriculum, much of the necessary content for that KB is lost with a simple text extraction process. Consider Figure 3, which contains the knowledge needed to confirm or reject Student-4's understanding. This knowledge is compiled from several modes of communication:

- **linguistic**: *This is where the power gets connected* (but this text itself is embedded in the red, upper right speech-bubble image).
- **graphic**: The speech bubble arrow (along with the deictic pronoun *This*) refers us to the *3V3* port identifier, itself embedded in a separate image.

[5]The Sensor Immersion (SI) Curriculum by Schoolwide Labs. Discussed in Section 3.

[6]Truncated for space. This example is from the classroom Worksheet Collection (see Section 4).

[7]Whether or not it should in fact intervene here is an important but separate question for dialogue policy teams.

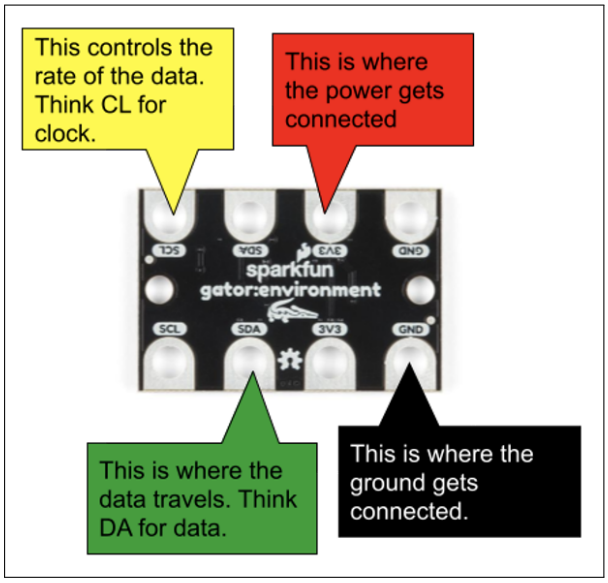

Figure 3: Screenshot of wiring diagram from the SI curriculum, illustrating the meaning embedded interdependently in linguistic, graphic, and iconic modalities.

- **iconic**: The red coloring of the speech bubble signifies the red wire.

Even extracting the text from the speech bubble in Figure 3 would not capture *where* the power gets connected; without some kind of reference resolution, deictic pronouns like *This* are of little use in a knowledge base.

While the semantic content in this example is unique to our data, the types of phenomena encountered are not: specialized terminology; deixis; image-based content; fragmented utterances; and so forth, are common in group educational settings. How to bridge between these features in order to provide the semantic scaffolding for an interactive classroom agent is a crucial and generalizable issue. We discuss our exploratory methods below.

3. Data and Annotation Process

We annotated three distinct but related corpora, with amounts and IAA reported in Table 1:

1. **Sensor Immersion (SI) Curriculum by SchoolWide Labs** (the "Curriculum Corpus"). This middle-school curriculum (Chakarov et al., 2021) integrates science and programming. As mentioned, the students learn about scientific sensors and how to program them using MakeCode, an introductory blocks-based coding platform.
2. **transcripts of small-group dialogues about the SI Curriculum** (the "Dialogue Corpus"). These dialogues occurred in a lab study setting of 2-3 students each, in which they discuss what they learned and brainstorm ideas of how to use the sensors in the real world.
3. **students' typed responses to questions about the curriculum** (the "Typed-Response Corpus"). The students' conversation during the lab study was guided by four preformulated questions. They used a single shared keyboard to type their answers to the questions.

corpus	AMRs	AMRs for IAA	IAA
all	4,711	2,531	0.91
Curric	1,631	1602	0.93
Typed-Resp	157	70	0.89
Dialogue	2,923	859	0.84

Table 1: JIA-AMR Collection: single-sentence AMRs for the curriculum; student typed responses; and student dialogues datasets. All AMRs underwent annotation followed by a quality-control pass or were single-annotated by an expert annotator. Inter-annotator agreement (IAA) represents F1 Smatch scores, calculated on a subset of the AMRs, as shown. IAA is discussed in section 3.4.

This paper focuses on our results with the typed responses, which were always intended to be our initial application of AMR parsing. However, our long-term goal is to process the spoken student utterances with equivalent levels of accuracy, hence the inclusion of the student dialogues. We also had very few instances of typed responses, 157, so including the conversations from the same lesson gave us more training data, albeit with differences. As indicated by Table 3, training on both datasets gave us the best test results on each dataset.

Process To account for the diverse phenomena discussed in section 2.2, we applied the following three overarching strategies to all three datasets:

Domain-specific training Annotators were trained in curriculum materials, including hardware and software components (the SI kit and MakeCode program). This strategy supported accurate interpretation of contextualized semantics, exemplified by the two dialogue fragments above.

Terminology dictionary definition Following Cai et al. (2024), we implemented a preprocessing pass in which specialized terminology were identified, analyzed, and added to a searchable resource (the "Jargon Dictionary") by our lead linguist in close consultation with the computer scientists and education experts on our team. For example, in this pass it was determined that *gator:bit* (a piece of the SI kit) should be annotated as shown in Figure 4. In the following AMR-building pass, upon encountering terminology, annotators searched the Dictionary for the appropriate entry and followed the established annotation. This strategy enabled uni-

	curric	typed-resp	dial.	all
files	2	8	11	21
clusters	66	331	588	985

Table 2: JIA-AMR Collection - Multi-sentence AMRs by the numbers. "Clusters" refers to the number of document-level coreference clusters (identity relations).

fied representations of terminology across datasets and annotators, leading to a 29.6 percent increase in parser performance (details in section 4). The preprocessing pass also included various corpus-unique components, discussed below.

Multi-sentence coreference annotation We conducted multi-sentence coreference annotation (O'Gorman et al., 2018) on a subset of each corpus (Table 2). This strategy facilitated cross-utterance reference resolution. We discuss this more in the Dialogue Corpus section (3.3) below since it is especially key for dialogue interpretation.

Dataset-specific phenomena are discussed below.

3.1. Annotating Multimedia Source Data: the Curriculum Corpus

The main challenge of annotating the SI curriculum was ensuring we captured all necessary information – linguistic and non-linguistic – for an agent, while avoiding wasting time on unnecessary information, like certain metadata. We found a manual data-selection and paraphrase pass accomplished both goals while optimizing the pipeline. These tasks were wrapped into the Jargon Dictionary creation pass, dispensing with an additional pass.

In this pass, we paraphrased non-linguistic information into natural language[8]. The content of Figure 3 was readily reformulated: *The 3V3 port is where the power is connected via the red wire*. In addition to resolving deixis and image-based content, paraphrasing allowed us to reassemble non-narrative textual information from data structures like lists and tables. The table in Figure 5 contains key information for a KB, but the meaning is lost in a standard text extraction process:

Environmental Sensor
Sound Sensor
Soil Moisture Sensor
What data can the sensor collect?
Temperature (C & F)
...

A paraphrase easily re-captures the meaning, e.g.: *The environmental sensor collects data about temperature (C and F)...*

Of course, there are other ways to tackle this problem. For example, document-level coreference on the fragments can retrieve some of the predicate-argument structure, and we did use this strategy to some degree. Ultimately, we found that paraphrasing was needed either way since human-targeted tabular categories and content are unpredictable; a later row (not shown) in the Figure 5 table included information about all three sensors in the environmental sensor column, for example. Furthermore, "front-loading" single-sentence AMRs like this allowed us to capitalize on both the more-evolved capabilities of graph-level automatic parsers (compared to document-level parsers) and the much broader range of semantic relations available from PropBank frames and AMR's inventory.

We also identified and filtered out unnecessary information. In many cases a single curriculum file included both relevant and irrelevant knowledge for an AI agent. The manual pass allowed us to preserve useful information about hardware set-up (*The lights are controlled on P12*), for example, while dropping supra-curricular information about teaching standards (*Building Toward Target NGSS PE*) from the same file. This avoided spending valuable time and effort on purposeless AMRs.

Finally, we tagged certain information that may be useful for downstream agent applications, including:

1. Perspective tags. The curriculum freely mixes student-targeted and teacher-targeted content, so we labeled each sentence for its intended audience. This information is not included in the AMRs themselves but could be added later automatically if desired.

2. Pointers to images. While we translated much image-based content into natural language, we also set up skeletal structures to support downstream multimodal exchanges between the meaning representations and curriculum images. To do this, we automatically generated bounding boxes with numerical identifiers on images (Figure 6) and introduced a new AMR concept, *image-entity*, which takes a single role, *:value*. For cases where we wanted to link the textual and visual content, we included the keyword *[IMAGE]* in the paraphrases: (*This [IMAGE] is a micro:bit*). This cued annotators in the AMR-building stage to use *image-entity* and reference the bounding box for the appropriate identifier value:

```
(h / hardware :name (n / name :op1 "micro:bit")
   :domain (i2 / image-entity :value 9.1
      :mod (t2 / this)))
```

[8] Early experiments with prompting an LLM to paraphrase the curriculum showed inadequate results. Since our goal was to create an evaluative dataset, we wanted to be sure of the faithfulness and comprehensiveness of our annotations. Furthermore, the manual nature of the pass was also motivated by the Jargon Dictionary and other steps described in this section.

concept	Phrase	NE type	AMR
gator:bit	gator:bit, gator bit, gator, Gator bit	hardware	(h / hardware :name (n / name :op1 "gator:bit"))

Figure 4: Example entry in the Jargon Dictionary. Annotators search the Dictionary for terminology upon encountering it in the data and find the standardized annotation for that term under the "AMR" column.

	Environmental Sensor	Sound Sensor	Soil Moisture Sensor
What data can the sensor collect?	Temperature (C & F) Humidity Pressure Carbon Dioxide (CO2) Total Volatile Organic Compounds (TVOC)	Sound Intensity or Sound level	Soil Moisture level as a percentage (0-100)

Figure 5: Screenshot of a worksheet key from the SI curriculum. Preserving the semantic relations embedded in table structures like this one requires more configuring in preprocessing than a simple text extraction process.

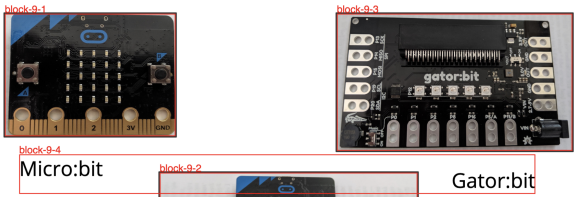

Figure 6: Screenshot of slide from the SI curriculum showing two key components of the hardware kit.

Thus, the scaffolding is in place for an agent to potentially access and link a picture for a student struggling to remember, for example, what the *micro:bit* is. In some cases, for image-based content that was not easily reformulated into natural language, we simply used *image-entity* instead of a fully-resolved paraphrase.

3.2. Annotating Student Written Language: the Typed-Response Corpus

During the lab study, the students' dialogue was guided by four questions that appeared one-by-one on a computer screen. Students were required to type their answers in a text box on the same screen. As with the Curriculum Corpus, we conducted a preprocessing paraphrase pass. For this corpus, the paraphrase task involved turning answer fragments into complete sentences, incorporating predicates and other concepts from the questions when possible, and resolving anaphora. Student typos and grammar were otherwise preserved. For example:

- Question: *How did each sensor system display the data it collected (music, lights, numbers, letters)?*
- Original student typed response: *the LEDS, And numbers* [sic]
- Paraphrased typed response: *The sensor systems displayed the data they collected by using the LEDS, And numbers.*

Paraphrasing added more information at the graph level which will support algorithmic matching with the AMRs in the curriculum KB and again capitalizes on sentence-level parser functionality. Furthermore, we aim to use the paraphrased data as an evaluative tool to see whether an LLM could paraphrase previously-unseen student answers in the same way, provided the questions as context; this would facilitate creation of a much larger dataset for parser training. Finally, the paraphrase task was easily constrained since each typed answer comprises at most a few sentences that are clearly contextualized by the question.

3.3. Annotating Student Spoken Language: the Dialogue Corpus

We annotated eight deidentified human-transcribed dialogues from eight separate lab studies, in which middle school students (2-3 per study) collaborated to answer questions about the curriculum. Following a Jigsaw educational approach (Kaplan and Dillenbourg, 2010), each student first individually learned about a different sensor (an environmental sensor, a sound sensor, or a soil moisture sensor) and how to program it to collect data. They then came together as a group to share their individual knowledge and brainstorm real-world applications of their sensors. The dialogues we annotated were of this culminating group stage.

As noted above, we believe this is the first attempt to create an AMR corpus for highly-collaborative

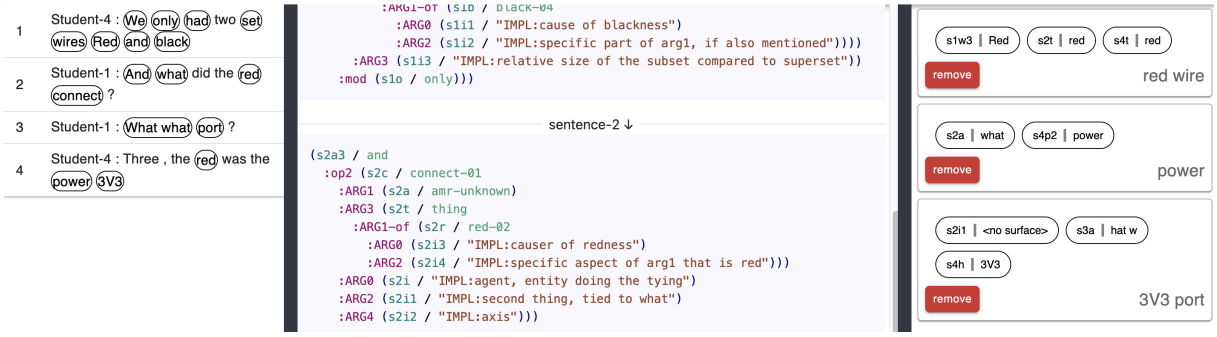

Figure 7: LiDARR's annotation interface for document-level coreference. The text is in the left-hand panel; AMRs in the middle; and the clusters in the right-hand panel represent coreference relations. The names in the lower right-hand corner of each cluster are assigned by the annotator.

human-human spoken dialogue. This linguistic setting introduces many phenomena as potential targets for AMR expansion. We focused our research on the following question in alignment with our overarching goals: *How can we recover meaning from the fragmented predicate-argument structure, ellipsis, and non-standard vocabulary that characterizes student dialogues such that an agent could map this meaning to the curriculum KB to identify gaps or mistakes in the students' knowledge?*

To address the first two features (fragmentation and ellipsis), we conducted multi-sentence coreference annotation, using the LiDARR annotation tool (Cai et al., 2025b)[9]. We did not attempt the paraphrase strategy for the Dialogue Corpus because dialogues are vastly more complex and unconstrained than the controlled typed-response context. The students jump back and forth between discussing the question content, metalanguage (for example, discussing how to phrase an answer or spell a word), the physical and virtual environment (discussing microphones, where to click on the screen, etc.), and off-topic subjects. Delimiting a paraphrase task for such phenomena would require grounding metalanguage and references in the environment, among other things, which was outside the scope of the current task.

Therefore, we turned to document-level coreference. Consider again the dialogue fragment from section 2.2, re-displayed in Figure 7. In order to check the student's knowledge, an agent must be able to resolve the ellipsis in utterances #2 and #4 (What does *red* refer to? What does *3V3* refer to?) and interpret *power* and *3V3* in #4 as arguments of the *connect* predicate in #2.

Multi-sentence coreference on the graphs recovers much of this information. For example, we are able to cluster the explicit mentions of *power* and *3V3* in #4 with ARG1 (*first thing being tied*) and ARG2 (*second thing, tied to what*) of *connect-01* in #2, respectively. Note that doing coreference on the graph level is what allows us to say that *3V3* is the second thing being connected, by adding the implicit ARG2 variable (s2i1) from the connect-01 frame in #2 to the *3V3 cluster*. The surface text does not make this connection explicit.

To address the most pervasive non-standard terminology in our data – references to the MakeCode blocks – we introduced a new domain-specific AMR role, *:isat-wiki* which facilitated normalization. For example, the block in Figure 1 is referred to by 34 different "names" in the annotated data (including the SI curriculum and typed responses corpora as well), including *input A*, *on B press*, and *when button*. Patterning after the pre-existing AMR *:wiki* role, we created standardized names for these elements and attached them using *:isat-wiki* to support mapping diverse surface-form realizations to the same concept in the KB, such that each of the 34 names are tagged with the unifying label *:isat-wiki "on-button-pressed-block"*. This was done on all three datasets but is mentioned here since students in particular refer to these items in creative ways.

3.4. Inter-Annotator Agreement

Table 1 shows inter-annotator agreement scores, calculated using the SMATCH (Semantic Match) metric (Cai and Knight, 2013). In SMATCH, F1-scores are micro-averaged over all triples in the AMR graphs, which means common or "easy" relations (e.g., frequent roles like :ARG0) contribute more to the score than rare, difficult ones. Scores range from 0.0 to 1.0, where 1.0 indicates perfect agreement and 0.0 indicates no agreement. All scores are above average for IAA SMATCH scores on AMRs, which are typically between 0.7 and 0.8 (Bonial et al., 2020). While still above average, the dialogue corpus shows the lowest agreement, as anticipated. Two unique features of the lab study setting contributed to a high degree of semantic ambiguity which had a negative impact on annotator agreement. First, the fact that the students

[9]For single-sentence AMR creation, we used UMR-Writer 2.0 (Zhao et al., 2021; Ge et al., 2023)

were typing while talking and, second, the fact that they were often discussing code. For example, a single word like *temperature* sometimes referred to the actual property of temperature, sometimes to a MakeCode block used to measure temperature, and sometimes to the word itself (as in *Can you fix temperature?*, as in, *Can the spelling of the word "temperature" be fixed?* These concepts are annotated differently in AMR to reflect their different semantics, but often the students' intended meaning was difficult to disambiguate, leading to divergent analyses by annotators. However, we have confidence that our thorough adjudication followed by careful Quality Control mitigated the impact of these disagreements. The high parsing results discussed in the following section validates our confidence.

3.5. Annotation time

Developing the Jargon Dictionary as a preprocessing step represents a time cost in addition to the actual AMR annotation time. We concluded however, that, compared to our previous single pass approach, our two-pass process overall took less annotation time and improved parser performance. Once the Dictionary is available for consultation, the AMR sentence annotation proceeds quite similarly to standard general news annotation. This is in contrast with the single pass annotation approach on this type of jargon heavy domain-specific data which is quite slow.

As mentioned above, our two-pass approach mirrored the terminology dictionary creation approach that was first utilized in our project for processing clinical text, i.e., electronic patient records, (Cai et al., 2024). This project found a similar substantial parsing performance gain based on the more consistent data annotation. Surprisingly, it also found that a very similar performance gain could be achieved with as few as 2000 new AMR graphs.

This is welcome news, since if much less high-quality consistently annotated data can achieve the same parser gains, then the additional annotation time for developing the dictionary can be offset by annotating less data. This makes the two-pass approach actually more efficient than the single pass approach while resulting in the highest parser performance. To illustrate, a parser trained on an inherited, preliminary round of single-pass annotation (predating our two-pass methodology) comprising over double the number of AMRs only showed a 1.4-point performance gain over the off-the-shelf parser (see Section 4. In this early round, in which annotators determined terminological analyses essentially on their own, *gator:bit* was represented five different ways. The performance gain differential demonstrates the importance of consistency over quantity of annotated data for parser training. The two-pass methodology not only promoted consistency, it decreased annotator time. The end result is less annotation time for better parser performance.

4. Parser Domain Adaptation and Early Experiments

We investigated the effect of domain shift on Abstract Meaning Representation (AMR) parsing using a model whose performance is near the current state of the art on the general LDC2020 AMR 3.0 release. Our parser is based on the SPRING AMR parser (Bevilacqua et al., 2021), which adopts BART (Lewis et al., 2020) as its underlying model. BART is a sequence-to-sequence model built on the transformer architecture (Vaswani et al., 2017). We fine-tuned the base model of the SPRING parser using 5,530 AMRs ("all" in Table 3) from different subsets of our whole dataset, segmented by genre: written and spoken. The written genre comprises curriculum documents and student typed responses, using 1,572 sentence-AMR pairs from the JIA-AMRs Collection (Table 1) for parser development. The spoken genre consists of 2,327 lab study transcript AMRs from the Collection[10], plus an additional 1,631 classroom-transcript AMRs[11] from the preliminary round of annotation (Section 3) for a total of 3,958 sentence-AMR pairs. For both genres, we used the same split ratio of 85% for training, 5% for development, and 10% for testing. The parser was continuously trained with each training set for 30 epochs, with initial learning rate to be 5×10^{-5} on 2 NVIDIA Titan RTX GPUs.

We observed that training on one genre leads to reduced performance when testing on the other. This indicates a substantial divergence in textual characteristics between the two genres. However, combining the written and spoken subsets for training yields improved parsing performance on both genres. This suggests that texts from similar domains, despite differences in linguistic expression, contribute complementary information that enhances domain-specific understanding. Table 3 shows the training and evaluation performance across genre-specific splits. These results include parser SMATCH scores when trained and tested within and across genres.

In addition, as introduced in Section 3, we conducted a preliminary analysis of the impact of our two pass annotation methodology: "all" in Table 3 represents the AMRs collected primarily through the two-pass strategy (excepting only the 1,631

[10]Annotation was ongoing at the time of parser training, which is why the full amount of lab study transcript AMRs from Table 1 was not used.

[11]This data was collected in a classroom setting before the lab study protocol was defined and will be released separately as the Worksheet Collection.

train \ test	written	spoken	all
written	70.53%	53.09%	59.71%
spoken	67.65%	82.85%	76.87%
all	**82.34**%	**83.97**%	**83.33**%
preliminary	57.09%	53.64%	55.12%
off the shelf	55.39%	52.64%	53.72%

Table 3: Parser performance across genre-specific test sets. Columns represent evaluation results on test sets from different genres, while rows indicate the training conditions. "All" refers to the combined spoken and written 5,530 AMRs. "Preliminary" indicates the 12,002 AMRs collected prior to introducing the two-pass methodology. "Off the shelf" denotes the baseline SPRING parser without domain-specific fine-tuning.

classroom graphs). When compared with the “preliminary” AMRs, which were built in a single pass prior to the introduction of the Jargon Dictionary, we see that the model trained on the two-pass annotations achieved significantly higher performance. This finding underscores the benefit of conventionalizing analyses of terminology for improving AMR quality and downstream parser training efficacy in specialized domains. This aligns well with p

Finally, we have conducted very preliminary experiments that suggest AMRs may improve LLM performance in identifying key missing concepts from student written answers.

5. Conclusion and Future Work

Assembling meaning representations for training an AI agent requires bridging heterogeneous source data. A preprocessing step was critical for creating a faithful and comprehensive database. In particular, we found that putting the bulk of specialized knowledge in this pass (and therefore removing it from annotators’ shoulders) supported consistency and speed. Rather than annotators having to reanalyze the same term each time they encountered it in the data – inevitably in different ways – difficult terminological analyses were determined once in preprocessing and standardized in the searchable Jargon Dictionary. This methodology is generalizable to other genres. While at face value a two-pass approach appears to be more time-consuming than a single pass, the fact that significantly fewer two-pass AMRs were required for a much greater performance boost suggests that it is in fact the more scalable strategy for specialized domains.

Preliminary experiments indicate that inclusion of AMR features improves JIA performance on the task of identifying gaps in student’s knowledge. We built two models, one using text data alone and one with text and AMR features, in which we asked the model to identify key missing concepts from the students’ typed responses to one of the lab study questions by comparing them with the correct answer. The model that incorporated AMR performed better in our 17-sample set, achieving a macro-F1 score of nearly 0.90 versus the macro-F1 score of 0.725 as the baseline method. In future work, we want to test on a larger, diversified dataset, investigate class imbalance issues, and improve the interpretability and deployment of the model in the field.

When we started this project with the very daunting goal of dynamically achieving accurate analysis of student conversations, no one was convinced that AMRs would provide the needed structure and scaffolding, or that the AMR parsing results would be acceptable. We therefore focused on a single curriculum unit with the goal of providing proof of concept. The results reported here encourage us to now turn our attention to the question of scalability. Is the adoption of this approach for multiple curriculum units feasible?

The first step is to determine the minimum number of AMRs for adequate parser performance on this type of data, similarly to the approach taken for the Spring THYME parser and the clinical domain (Cai et al., 2024). The next step is to experiment with using LLMs to assist in building the Jargon Dictionary for a new curriculum unit, and in using it to annotate the resulting AMR graphs on newly acquired student conversations as automatically as possible. The more we can compress the annotation time per curriculum unit without sacrificing quality, the more feasible our portability task becomes. We make our data and results available with the hope of persuading others to join us in this endeavor.

6. Acknowledgments

This research was made possible through support from the NSF National AI Institute for Student- AI Teaming (iSAT), Grant DRL 2019805. Any opinions, findings, or conclusions expressed are solely those of the authors and do not necessarily reflect the views of the NSF. We sincerely thank Ahmed Elsayed for his diligent annotation and our anonymous reviewers for their very constructive comments.

7. Limitations

Currently, the JIA-AMR multi-sentence corpus only includes full-identity coreference annotations at the document level, not partial-coreference relations such as set-member and whole-part. Additionally,

as yet, no structures have been put in place to integrate the dialogues with the students' physical and virtual environment. This means that currently there are "floating" deictic references in the annotated data – for example, if a student says *Put it there* while pointing to something on the screen, *there* is unanchored to its referent. In future work we hope to integrate our annotations with GestureAMR (GAMR) (Donatelli et al., 2022) to retrieve some of this meaning. Finally, we did not extend the schema to support evolution of reference over time during the dialogues. For example, one of the lab study questions asks the students to choose their best idea about how to use the sensors after they finish brainstorming. In one case the students first choose one idea and then later change their minds. In this and in similar cases, only the final answer was treating as a coreference. In the future we wish to expand coverage to include these semantic features as such steps will support an even more comprehensive semantic framework for a classroom agent.

8. Ethical Considerations

The authors of this paper collaborated with the data managers on the project to ensure all student-related information was carefully and thoroughly de-identified. This included replacing student names with untraceable labels like *Student 1*.

Human annotators were assigned from our research staff based on skillset and experience and were paid based on level of experience within the pay range set by our institution.

While efforts were made in the data collection stage to include students from a range of economic and racial/ethnic backgrounds, several groups were underrepresented, meaning their language is not well represented in the lab study AMRs. Therefore, one possible risk of the collection could be decreased mapping capability of their knowledge to the curriculum knowledge base.

9. Bibliographical References

Elliot Aronson, Nancy Blaney, Cook Stephan, Jev Sikes, and Matthew Snapp. 1978. *The Jigsaw Classroom*. Sage Publications, Beverly Hills, CA.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract Meaning Representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria. Association for Computational Linguistics.

Michele Bevilacqua, Rexhina Blloshmi, and Roberto Navigli. 2021. One SPRING to rule them both: Symmetric AMR semantic parsing and generation without a complex pipeline. In *Proceedings of AAAI*.

Claire Bonial, Mitchell Abrams, David Traum, and Clare Voss. 2021. Builder, we have done it: Evaluating & extending dialogue-AMR NLU pipeline for two collaborative domains. In *Proceedings of the 14th International Conference on Computational Semantics (IWCS)*, pages 173–183, Groningen, The Netherlands (online). Association for Computational Linguistics.

Claire Bonial, Lucia Donatelli, Mitchell Abrams, Stephanie M. Lukin, Stephen Tratz, Matthew Marge, Ron Artstein, David Traum, and Clare Voss. 2020. Dialogue-AMR: Abstract Meaning Representation for dialogue. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 684–695, Marseille, France. European Language Resources Association.

Claire Bonial, Stephanie M. Lukin, Mitchell Abrams, Anthony Baker, Lucia Donatelli, Ashley Foots, Cory J. Hayes, Cassidy Henry, Taylor Hudson, Matthew Marge, Kimberly A. Pollard, Ron Artstein, David Traum, and Clare R. Voss. 2024. Human–robot dialogue annotation for multi-modal common ground. *Language Resources and Evaluation*, 59:1525–1575.

Julia Bonn, Martha Palmer, Zheng Cai, and Kristin Wright-Bettner. 2020. Spatial AMR: Expanded spatial annotation in the context of a grounded Minecraft corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4883–4892, Marseille, France. European Language Resources Association.

Richard Brutti, Lucia Donatelli, Kenneth Lai, and James Pustejovsky. 2022. Abstract Meaning Representation for gesture. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1576–1583, Marseille, France. European Language Resources Association.

Jon Cai, Brendan King, Peyton Cameron, Susan Windisch Brown, Miriam Eckert, Dananjay Srinivas, George Arthur Baker, V Kate Everson, Martha Palmer, James Martin, and Jeffrey Flanigan. 2025a. In search of the lost arch in dialogue: A dependency dialogue acts corpus for multi-party dialogues. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 20135–20149, Vienna, Austria. Association for Computational Linguistics.

Jon Cai, Kristin Wright-Bettner, Martha Palmer, Guergana Savova, and James Martin. 2024. Adapting Abstract Meaning Representation parsing to the clinical narrative – the SPRING THYME parser. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 271–282, Mexico City, Mexico. Association for Computational Linguistics.

Jon Cai, Kristin Wright-Bettner, Zekun Zhao, Shafiuddin Rehan Ahmed, Abijith Trichur Ramachandran, Jeffrey Flanigan, Martha Palmer, and James Martin. 2025b. LiDARR: Linking document AMRs with referents resolvers. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 426–435, Vienna, Austria. Association for Computational Linguistics.

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

Alexandra Gendreau Chakarov, Quentin Biddy, Colin Hennessy Elliott, and Mimi Recker. 2021. The data sensor hub (dash): A physical computing system to support middle school inquiry science instruction. *Sensors*, 21(18).

Johannes Dellert. 2020. Exploring probabilistic soft logic as a framework for integrating top-down and bottom-up processing of language in a task context. *arXiv preprint arXiv:2004.07000*.

Sidney K. D'Mello, Quentin Biddy, Thomas Breideband, Jeffrey Bush, Michael Chang, Arturo Cortez, Jeffrey Flanigan, Peter W. Foltz, Jamie C. Gorman, Leanne Hirshfield, Mon-Lin Monica Ko, Nikhil Krishnaswamy, Rachel Lieber, James Martin, Martha Palmer, William R. Penuel, Thomas Philip, Sadhana Puntambekar, James Pustejovsky, Jason G. Reitman, Tamara Sumner, Michael Tissenbaum, Lyn Walker, and Jacob Whitehill. 2024. From learning optimization to learner flourishing: Reimagining ai in education at the institute for student-ai teaming (isat). *AI Magazine*, 45(1):61–68.

Emily Doherty, E. Margaret Perkoff, Sean von Bayern, Rui Zhang, Indrani Dey, Michal Bodzianowski, Sadhana Puntambekar, and Leanne Hirshfield. 2025. Piecing together teamwork: A responsible approach to an llm-based educational jigsaw agent. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–17.

Lucia Donatelli, Kenneth Lai, Richard Brutti, and James Pustejovsky. 2022. Towards situated amr: Creating a corpus of gesture amr. In *Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management. Health, Operations Management, and Design*, page 293–312.

Sijia Ge, Jin Zhao, Kristin Wright-bettner, Skatje Myers, Nianwen Xue, and Martha Palmer. 2023. UMR-writer 2.0: Incorporating a new keyboard interface and workflow into UMR-writer. In *Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII)*, pages 211–219, Toronto, Canada. Association for Computational Linguistics.

Yilun Hua, Zhaoyuan Deng, and Kathleen McKeown. 2023. Improving long dialogue summarization with semantic graph representation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13851–13883, Toronto, Canada. Association for Computational Linguistics.

Frederic Kaplan and Pierre Dillenbourg. 2010. Scriptable classrooms. *Classroom of the Future*, pages 141–160.

Lasha Labadze, Maya Grigolia, and Lela Machaidze. 2023. Role of ai chatbots in education: systematic literature review. *International Journal of Educational Technology in Higher Education*, 20(56).

Kenneth Lai, Richard Brutti, Lucia Donatelli, and James Pustejovsky. 2024. Encoding gesture in multimodal dialogue: Creating a corpus of multimodal AMR. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5806–5818, Torino, Italia. ELRA and ICCL.

Kenneth Lai, Lucia Donatelli, Richard Brutti, and James Pustejovsky. 2025. A model of information state in situated multimodal dialogue. In *Proceedings of the 16th International Conference on Computational Semantics*, pages 292–298, Dusseldorf, Germany. Association for Computational Linguistics.

Paul Landes, Sitara Rao, Aaron J. Chaise, and Barbara Di Eugenio. 2025. Abstract meaning representation for hospital discharge summarization. arXiv:2506.14101. Version 1.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence

pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Junyi Li, Jie Chen, Ruiyang Ren, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2024. The dawn after the dark: An empirical study on factuality hallucination in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10879–10899, Bangkok, Thailand. Association for Computational Linguistics.

Soohyun Nam Liao, William G. Griswold, and Leo Porter. 2018. Classroom experience report on jigsaw learning. ITiCSE 2018, page 302–307, New York, NY, USA. Association for Computing Machinery.

Tahira Naseem, Austin Blodgett, Sadhana Kumaravel, Tim O'Gorman, Young-Suk Lee, Jeffrey Flanigan, Ramón Astudillo, Radu Florian, Salim Roukos, and Nathan Schneider. 2022. DocAMR: Multi-sentence AMR representation and evaluation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3496–3505, Seattle, United States. Association for Computational Linguistics.

Tim O'Gorman, Michael Regan, Kira Griffitt, Ulf Hermjakob, Kevin Knight, and Martha Palmer. 2018. AMR beyond the sentence: the multi-sentence AMR corpus. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3693–3702, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

Christopher Tam, Richard Brutti, Kenneth Lai, and James Pustejovsky. 2023. Annotating situated actions in dialogue. In *Proceedings of the Fourth International Workshop on Designing Meaning Representations*, pages 45–51, Nancy, France. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Isaac Wang, Mohtadi Ben Fraj, Pradyumna Narayana, Dhruva Patil, Gururaj Mulay, Rahul Bangar, J. Ross Beveridge, Bruce A. Draper, and Jaime Ruiz. 2017. Eggnog: A continuous, multimodal data set of naturally occurring gestures with ground truth labels. In *2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017)*, pages 414–421.

Bohao Yang, Kun Zhao, Liang Zhan, and Chenghua Lin. 2024. Emphasising structured information: Integrating abstract meaning representation into llms for enhanced open-domain dialogue evaluation. arXiv:2404.01129. Version 1.

Jin Zhao, Nianwen Xue, Jens Van Gysel, and Jinho D. Choi. 2021. UMR-writer: A web application for annotating uniform meaning representations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 160–167, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yilun Zhu, Joel Ruben Antony Moniz, Shruti Bhargava, Jiarui Lu, Dhivya Piraviperumal, Site Li, Yuan Zhang, Hong Yu, and Bo-Hsiang Tseng. 2024. Can large language models understand context? In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 2004–2018, St. Julian's, Malta. Association for Computational Linguistics.