Multimodal Common Ground Annotation for Partial Information Collaborative Problem Solving

Yifan Zhu[1], **Changsoo Jung**[2], **Kenneth Lai**[1], **Videep Venkatesha**[2], **Mariah Bradford**[2]
Jack Fitzgerald[2], **Huma Jamil**[2], **Carine Graff**[2], **Sai Kiran Ganesh Kumar**[2]
Bruce Draper[2], **Nathaniel Blanchard**[2], **James Pustejovsky**[1], **Nikhil Krishnaswamy**[2]
[1]Brandeis University, Waltham, MA USA
[2]Colorado State University, Fort Collins, CO USA
{zhuyifan,jamesp}@brandeis.edu, nkrishna@colostate.edu

Abstract

This project note describes challenges and procedures undertaken in annotating an audio-visual dataset capturing a multimodal situated collaborative construction task. In the task, all participants begin with different partial information, and must collaborate using speech, gesture, and action to arrive a solution that satisfies all individual pieces of private information. This rich data poses a number of annotation challenges, from small objects in a close space, to the implicit and multimodal fashion in which participants express agreement, disagreement, and beliefs. We discuss the data collection procedure, annotation schemas and tools, and future use cases.

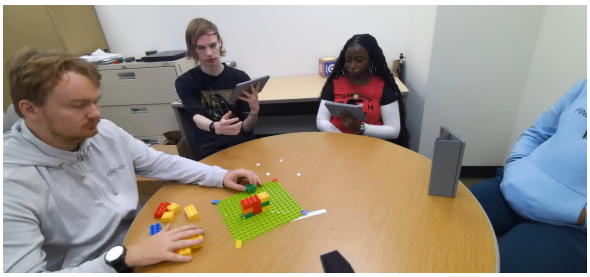

Figure 1: Left to right: a *builder* and 3 *directors* participating in the collaborative construction task with a partially-completed structure on the board. Director 1 (second from left) is indicating the position of a block using a combination of language and gesture with the accompanying utterance "Coming towards me then it's the red long block."

1 Introduction

In collaborative tasks, participants may convey their beliefs, desires, and intentions (BDI) through language, gesture, gaze, and action. These modalities communicate explicit beliefs, disambiguate references, and signal implicit attitudes, enabling participants with different backgrounds or knowledge to build a shared *common ground*—the set of task-relevant facts and evidence jointly accepted by the group. The Edinburgh Map Task (Anderson et al., 1991) is a well-known example of multimodal, conversational, collaborative task annotation and has long served as a benchmark for studying dialogue, spatial reference, and grounding. Our work builds on this tradition with a more complex, co-situated construction task with multiple *instruction givers* and integrates gesture, speech, and action while supporting the study of common ground under structural and spatial ambiguity.

In this project note, we briefly describe the collection and annotation of a novel collaborative problem solving dataset centered around this task. The data is being annotated with multiple modal channels, and the process implicates a number of interesting challenges toward creating semantic annotations that are interoperable across modalities.

The problem of *common ground tracking* (CGT) has been addressed in previous work such as Clark and Brennan (1991); Traum (1994); Ginzburg et al. (1996); Stalnaker (2002); Asher and Gillies (2003); Traum and Larsson (2003), and Hadley et al. (2022). Multimodal approaches to common ground tracking include Khebour et al. (2024b) and Vanderhoeven et al. (2025). However, the tasks addressed in these and similar approaches (Khebour et al., 2024a) suffer from a number of drawbacks, including problems with 1) **agreement/disagreement**: there are few opportunities for disagreement as the task is well-structured with clear solutions at each step; 2) **complexity**: cognitive and interpretive complexity is low as disagreements typically center questions of single-step procedures or computations; 3) **reusability**: once a group has completed the task, they know the answer and cannot organically perform the task again. Our task has been designed to mitigate these shortcomings to to enable the robust study of common ground tracking in multimodal dialogue.

Proceedings of the 21st Joint ACL - ISO Workshop on Interoperable Semantic Annotation (ISA-21), pages 85–91

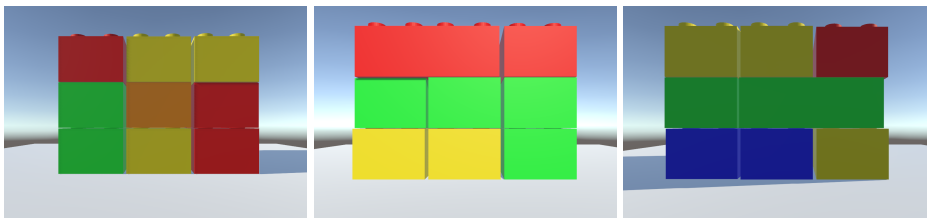

Figure 2: 3 individual side views of a complete structure, each given to a director.

2 Task Description

The task we focus on is a group collaborative construction task structured to satisfy the three conditions enumerated in Sec. 1, that we previously identified as being shortcomings in existing tasks used in the study of multimodal CGT. Namely, we designed the task to create meaningful disagreements within the group about the right course of action, be sufficiently complex such that there are multiple likely solutions toward the goal, and allow participants to do the task multiple times by creating a novel goal each session.

The task is designed for 4 people: 3 *directors* and 1 *builder* (see Fig. 1). Each builder receives a different side view of a 3D structure made of large blocks (see Fig. 2; the directors receive their images on a personal tablet). There are an assortment of blocks on the table before the group, but only the builder (identified in Fig. 1 as the only person without a tablet) is allowed to touch the blocks. The directors are not allowed to show their private images to each other or to the builder. The group must then collaborate to instruct the builder to build a single coherent structure that is consistent with the images given to all the directors. Dialogue is free form and there are no restrictions on what the participants may say, do, or ask each other, as long as the directors do not touch the blocks or show their private images to anyone else. Since there are four sides to the structure but only three images provided, there may be multiple valid solutions. A novel test pattern is generated at the beginning of each session. Thus this task satisfies the 3 desiderata listed above by distributing *partial* information throughout the group, and also simulates a scenario in which a group of people with different background knowledge, expertise, and skills must collaborate to solve a problem.

3 Data Collection

Data was collected at 2 sites, both universities in the United States. The task takes place on a tabletop and is captured using 3 Microsoft Kinect Azure cameras to capture different angles of the task space. Audio is recorded on a single conference-style tabletop microphone. The study was approved by university Institutional Review Boards (IRBs) and participants received USD 15.00 each.

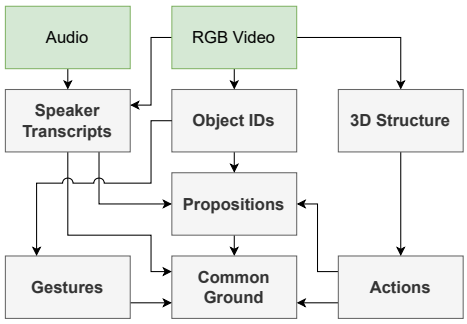

Figure 3: Dependencies between annotations of different modalities. Arrows represent required inputs to the target annotation. Boxes with bolded text are each described in subsections in Sec. 4.

Novel test structures were generated for each group, either manually by the researchers, or procedurally using a script written in the Unity game engine. Test structures consist of blocks arranged in a 3D grid configuration, and screenshots of 3 side views are taken and distributed to the 3 directors. Each session consists of two phases. In the first phase, the test structure contains strictly square or rectangular blocks arranged in a $3 \times 3 \times 3$ grid (wdh – see Fig. 2), with no gaps permitted in the structure. In phase 2, the footprint of the structure is expanded to $4 \times 4 \times 3$, the blocks involved may have curved or angled components, and gaps in the structure are permitted.

In total, after removing recordings with technical or procedural errors, 38 usable group recordings were retained. Most were 20-40 minutes in length.

4 Annotation Schemas

The technical challenges in annotating this data are manifold. Speech overlaps, the objects are close together, actions and gestures may have multiple physical manifestations and interpretations. Additionally, complete annotation of one modality frequently depends on information from another modality, creating dependencies in the annotation

pipeline (Fig. 3). Finally, the sheer amount of data makes purely manual annotation an infeasible task. Therefore, for most modalities, we adopt a semi-automated machine annotation with human validation and post-correction strategy. Specific challenges and methods for each individual modality are given in their respective sections below.

4.1 Speech Transcriptions

Spoken dialogue is transcribed via automatic transcription with the Whisper ASR model (Radford et al., 2023), combined with PyAnnote (Plaquet and Bredin, 2023) for speaker diarization. Annotators review the ASR transcriptions while watching the relevant video, and correct errors in segmentation, transcription, or speaker attribution. We save both the manually-corrected and automatic transcriptions, as research has shown that automated segmentation and transcription errors can have an impact on downstream task performance (Terpstra et al., 2023; Ibarra et al., 2025; VanderHoeven et al., 2025; Venkatesha et al., 2025a), but training models against noisy transcriptions can mitigate this effect (Nath et al., 2025). A zero-shot LLM (Llama-3.1-8B-Instruct) is then used to extract relations among blocks referenced in dialogue, which constitute the project's primary task-relevant signal. All outputs were subsequently manually reviewed and corrected by a human annotator, with omitted instances added, yielding a curated annotation set.

4.2 Gestures

Participant gestures are annotated using Gesture AMR (GAMR; Brutti et al. (2022)), an abstract meaning representation format designed to capture gesture semantics. Gestures may be deictic, iconic, or emblematic, indicating structural descriptions (e.g., *side by side*), block attributes (e.g., *square*, *curved*), or actions (e.g., *bring forward/backward*, *rotate*). These annotations are time-stamped against the video, enabling alignment with utterances and actions. Annotation is performed while watching the video, with access to object IDs so that gestures referring to specific objects can be concretely recorded. For example, a GAMR annotation:

```
(d / deixis-GA
  :ARG0 (d1 / director-1)
  :ARG1 (bs1 / blue-square-1)
  :ARG2 (g / group))
```

indicates that Director 1 (`ARG0`) is pointing not just at a generic blue square, but at a specific block (`ARG1`), with the intended recipient of this reference being the group (`ARG2`). In practice, we treat deictic references as pointing to objects or locations. More fine-grained distinctions—such as path or manner— are considerably harder to interpret. While GAMR could in principle encode these with roles like `:manner` or `:mode`, we do not capture them in our current annotations.

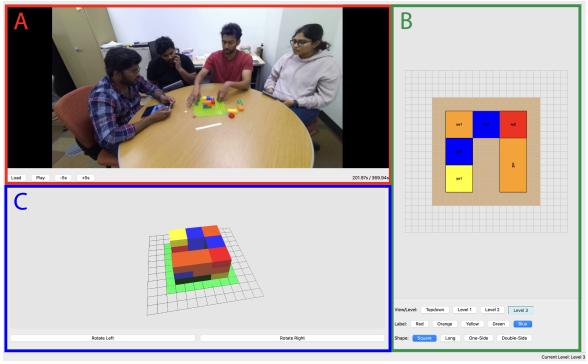

Figure 4: Structure Annotation Tool. **A**: Video Player, **B**: Interactive Area, **C**: 3D View.

Gesture signals are inherently context-dependent and do not provide direct evidence for belief states alone. To address this, we adopt a two-pass contextualization procedure. In the first pass, emblematic attitudinal gestures (e.g., nods, head-shakes) are identified, as these directly license belief updates from gesture evidence. In the second pass, gestures are aligned with the discourse and action layers to assess co-occurrence and temporal contingency (e.g., accompanying an utterance or preceding an action). When such alignment holds, we annotate the corresponding gesture-derived belief state.

4.3 3D Structure

3D structure annotations are intended to capture the state of the board after each time the builder places, moves, or removes a block. To capture these we created a Structure Annotation Tool (SAT), whose interface is shown in Fig. 4. The SAT interface consists of a Video Player (A), an Interactive Area (B), and a 3D View (C). The Video Player displays the video being annotated and the annotator can scrub back and forth as needed. The Interactive Area is a drag-and-drop tool where, if a block is placed on the board in the video, the annotator chooses the color (red, orange, yellow, green, blue) and shape (square, long, single curve, and double curve) of the block and places it on a grid representing the placement board. Blocks can be placed on the bottom level or on top of other blocks in the top-down view, and can then be selected, moved,

rotated, or deleted. The 3D view shows the current structure in 3D, which is rotatable for better visibility. Any actions taken in the Interactive Area are instantly reflected in the 3D View. The construction history is autosaved to JSON as timestamped data that contains all actions along with object IDs and coordinate information on the grid.

4.4 Actions

Annotator actions in the Structure Annotation Tool reflect block placement, movement, or removal actions in the task video. Therefore, actions are automatically extracted from the saved structure annotations. If a block appears at a location where there was none previously, a put action is registered. If a block disappears from a location, a $remove$ action is registered. $move$ can be considered a combination of $remove$ and put such that $move(b, \ell_1, \ell_2)$ can be reified to $remove(b, \ell_1)$ followed by $put(b, \ell_2)$. Locations may be absolute coordinates or relational predicates extracted over coordinates. Relational predicates are restricted to a fixed set such as on and $left$ to avoid creating ambiguous annotations, e.g., where $left(a, b)$ and $right(b, a)$ refer to the same configuration.

4.5 Object Identification

Since the small size of individual objects and the dense configurations of stuctures created during the task pose a tractability challenge for manual annotation, we use a semi-automated approach. However, since automated object trackers struggle to reidentify objects that have disappeared from view, some level of human annotation, validation, and correction is required. We adopt a pipeline as shown in Fig. 5 to maximize accuracy while minimizing human labor cost. The original video is split into 30 second segments and in the first frame of each segment, an annotator manually labels points on distinct objects in the frame. These labeled video segments are then fed into the Segment Anything 2 (SAM-2) model (Ravi et al., 2024) which makes an initial prediction of bounding boxes. The same segments are also fed to a fine-tuned instance of YOLOv11x (Khanam and Hussain, 2024) and the YOLO and SAM-2 bounding boxes are compared for validation. Where SAM-2 missed detections, the failed frames are extracted and returned to the manual keypoints annotation stage and the process repeats. We find that compared to strictly manual bounding box annotation of objects, this pipeline results in up to a 240× speed-up in processing time. Object detection was intended to automatically identify (a) the targets of deictic gestures, (b) the targets of actions, and (c) the positions of blocks within or outside of the structure. The broader goal was to use automatically detected objects to generate complete 3D structures. However, this proved challenging in practice, so to ensure usable data for downstream annotation and analysis, we provided teams with the 3D structure annotations directly.

4.6 Propositions

In Khebour et al. (2024b), common ground is computed in part by extracting expressed task-relevant propositions. Relatedly, Venkatesha et al. (2024, 2025b) develop propositional extraction methods in multiple tasks that realize task relevant propositions as relations between task items or between items and properties. Similarly, in this task, each proposition is indexed by participant ID, timestamp, and the relative relation among blocks, enabling participant-specific retrieval and temporal alignment in the general form "`<timestamps> <person> <block> <relation> <block>`". Propositions must also capture the perspective from which the annotated relation is seen, and the layer of the structure.

The structured propositions are extracted from three modalities—speech, action, and gesture—and integrate them into a unified representation format. Annotations from each modality are first collected independently and then merged into a single CSV file. The entries are sorted chronologically and converted into a standardized belief-annotation schema to support common-ground computation. Speech-based propositions are derived using off-the-shelf large language models (LLMs). For each target mention, the model receives a ten-utterance window (five preceding, five following) to capture discourse cues and resolve coreference. Generic block descriptors are then replaced with specific block identifiers from the action annotations, resulting in propositions such as $on(o1, y2, D3_{side}, layer_1)$. Action-based propositions are converted from absolute spatial coordinates to relative positions between blocks, while side information is supplemented using cues from the speech modality. As gesture annotations contain only the gesturer's identity and gesture content, we incorporate contextual information from both speech and action to complete the final set of gesture-related propositions.

4.7 Common Ground

We track participants' epistemic state updates and define common ground as propositions mutually

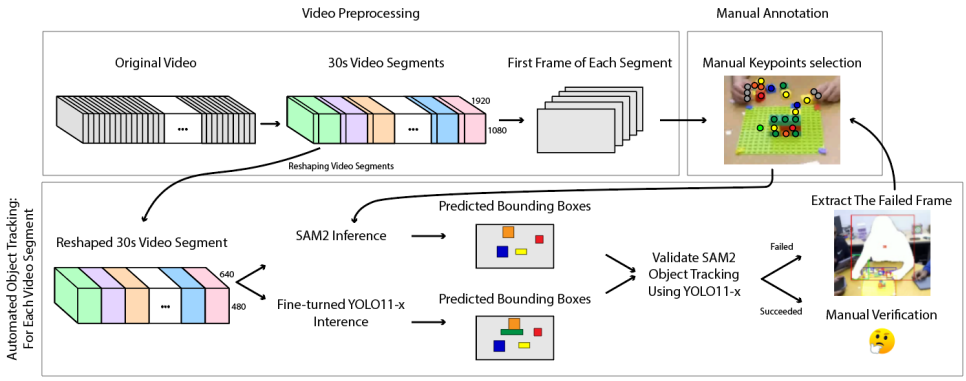

Figure 5: Semi-automated pipeline for capturing object bounding boxes and IDs.

agreed upon. Propositions from different modalities are normalized into a belief-annotation format, $B_x\phi$, where participant x believes (B) propositional content ϕ. Belief formation is licensed by three axioms—Seeing is Believing (Bolander, 2014), Saying is Believing, and Acting is Believing—aligning with the three modalities from which we extract evidence: speech, gesture, and action (gaze is treated as implicit). For example, if participant x asserts φ in dialogue at time t, we record $B_x\varphi$ under the axiom Saying is Believing.

Annotations across different modalities are merged into a single chronologically ordered file and converted into the standardized belief-annotation schema for time-indexed computation. A proposition is considered common ground when the same normalized content is attributed to two or more participants, represented as $CG_{a,b,\ldots}\phi$, where $a, b, \ldots$ denotes the set of participants jointly committed to proposition ϕ. In our dataset, the maximal common-ground set involves four participants. Because beliefs are dynamic, B_x may be revised over time as implicit intentions become explicit in speech or as actions provide evidence that licenses new belief updates.

To capture commitment to or rejection of others' propositions, our scheme further includes $ACCEPT$ and $DOUBT$ labels. When a participant accepts another's proposition, the corresponding common ground annotation is updated to reflect shared understanding; when a participant expresses doubt, a disagreement annotation is recorded to mark epistemic conflict.

5 Conclusion

We have outlined the desiderata, processes, and challenges involved in annotating common ground in a co-situated, multimodal, partially observable collaborative problem-solving task. This type of annotation requires integrating multiple communicative channels with converging dependencies and raises a range of technical, design, and interpretive challenges, for which we have described our approaches and techniques. More broadly, annotation of data of this kind presents challenges familiar to the annotation community, and we hope that our experiences can serve as a useful reference point. Although a gold-standard annotation set and corresponding inter-annotator agreement (IAA) analysis are not yet available, developing them remains a priority for future work. We plan to obtain human annotations, quantify agreement using standard measures (e.g., Cohen's κ, Krippendorff's α), and evaluate the computed annotations against this gold standard. The annotation remains ongoing and a fully-annotated dataset will be released at a future date.

The task's multi-party, partial information setting represents a novel contribution in the age of LLMs. The resulting corpus captures the conversational and information dynamics of a collaboration that is not fully transparent to any of the participants, including any AI system observing the interaction. Therefore, in the context of LLM-driven agents for problem-solving support or human-AI collaboration, our data captures how each participant expresses their implicit "theory of mind" of the other participants' beliefs and goals. The ability to infer such belief states has been shown to be challenging for modern LLMs (Ullman, 2023; Hu et al., 2025), and this challenge is amplified by how even granular task-relevant propositions may be expressed multimodally in this task, including through speech, gesture, and actions. Thus, to fine-tune or assess LLMs for this and similar tasks, or to provide a modern LLM or VLM with sufficient information to interpret participant behaviors in context, an interoperable annotation scheme that captures the semantic relations across modalities

and across time, is required. The efforts described represent a step toward a corpus that would be suitable for fine-tuning, constructing scenarios suitable for assessment of zero-shot prompting, or for benchmarking the recoverability of information in modalities of interest from other modalities.

Acknowledgments

This material is based in part upon work supported by Other Transaction award HR00112490377 from the U.S. Defense Advanced Research Projects Agency (DARPA) Friction for Accountability in Conversational Transactions (FACT) program, and by award W911NF-25-1-0096 from the U.S. Army Research Office (ARO). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government.

References

Anne H Anderson, Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, et al. 1991. The hcrc map task corpus. *Language and speech*, 34(4):351–366.

Nicholas Asher and Anthony Gillies. 2003. Common ground, corrections, and coordination. *Argumentation*, 17:481–512.

Thomas Bolander. 2014. Seeing is believing: Formalising false-belief tasks in dynamic epistemic logic. In *European conference on social intelligence (ECSI 2014)*, pages 87–107.

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.

Herbert H Clark and Susan E Brennan. 1991. Grounding in communication.

Jonathan Ginzburg et al. 1996. Dynamics and the semantics of dialogue. *Logic, language and computation*, 1:221–237.

Lauren V Hadley, Graham Naylor, and Antonia F de C Hamilton. 2022. A review of theories and methods in the science of face-to-face social interaction. *Nature Reviews Psychology*, 1(1):42–54.

Jennifer Hu, Felix Sosa, and Tomer Ullman. 2025. Re-evaluating theory of mind evaluation in large language models. *Philosophical Transactions B*, 380(1932):20230499.

Benjamin Ibarra, Brett Wisniewski, Corbyn Terpstra, Videep Venkatesha, Mariah Bradford, and Nathaniel Blanchard. 2025. Investigating automated transcriptions for multimodal cps detection in groupwork. In *International Conference on Human-Computer Interaction*, pages 214–224. Springer.

Rahima Khanam and Muhammad Hussain. 2024. Yolov11: An overview of the key architectural enhancements. *arXiv preprint arXiv:2410.17725*.

Ibrahim Khebour, Richard Brutti, Indrani Dey, Rachel Dickler, Kelsey Sikes, Kenneth Lai, Mariah Bradford, Brittany Cates, Paige Hansen, Changsoo Jung, Brett Wisniewski, Corbyn Terpstra, Leanne M Hirshfield, Sadhana Puntambekar, Nathaniel Blanchard, Nikhil Krishnasamy, and James Pustejovsky. 2024a. When text and speech are not enough: A multimodal dataset of collaboration in a situated task. *Journal of Open Humanities Data*, 10(1).

Ibrahim Khalil Khebour, Kenneth Lai, Mariah Bradford, Yifan Zhu, Richard A Brutti, Christopher Tam, Jingxuan Tu, Benjamin A Ibarra, Nathaniel Blanchard, Nikhil Krishnaswamy, and James Pustejovsky. 2024b. Common ground tracking in multimodal dialogue. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3587–3602.

Abhijnan Nath, Carine Graff, Andrei Bachinin, and Nikhil Krishnaswamy. 2025. Frictional Agent Alignment Framework: Slow Down and Don't Break Things. In *Annual Meeting of the Association for Computational Linguistics (ACL)*. ACL.

Alexis Plaquet and Hervé Bredin. 2023. Powerset multi-class cross entropy loss for neural speaker diarization. In *Proc. INTERSPEECH 2023*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. 2024. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*.

Robert Stalnaker. 2002. Common ground. *Linguistics and philosophy*, 25(5/6):701–721.

Corbyn Terpstra, Ibrahim Khebour, Mariah Bradford, Brett Wisniewski, Nikhil Krishnaswamy, and Nathaniel Blanchard. 2023. How good is automatic segmentation as a multimodal discourse annotation aid? In *Proceedings of the 19th Joint ACL-ISO Workshop on Interoperable Semantics (ISA-19)*, pages 75–81.

David Traum. 1994. A computational theory of grounding in natural language conversation.

David R Traum and Staffan Larsson. 2003. The information state approach to dialogue management. *Current and new directions in discourse and dialogue*, pages 325–353.

Tomer Ullman. 2023. Large language models fail on trivial alterations to theory-of-mind tasks. *arXiv preprint arXiv:2302.08399*.

Hannah VanderHoeven, Brady Bhalla, Ibrahim Khebour, Austin C Youngren, Videep Venkatesha, Mariah Bradford, Jack Fitzgerald, Carlos Mabrey, Jingxuan Tu, Yifan Zhu, and James Krishnaswamy, Nikhil ane Pustejovsky. 2025. Trace: Real-time multimodal common ground tracking in situated collaborative dialogues. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations)*, pages 40–50.

Videep Venkatesha, Mariah Bradford, and Nathaniel Blanchard. 2025a. Dude, where's my utterance? evaluating the effects of automatic segmentation and transcription on cps detection. In *International Conference on Artificial Intelligence in Education*, pages 144–151. Springer.

Videep Venkatesha, Abhijnan Nath, Ibrahim Khebour, Avyakta Chelle, Mariah Bradford, Jingxuan Tu, James Pustejovsky, Nathaniel Blanchard, and Nikhil Krishnaswamy. 2024. Propositional extraction from natural speech in small group collaborative tasks. In *Proceedings of the 17th International Conference on Educational Data Mining*, pages 169–180.

Videep Venkatesha, Abhijnan Nath, Ibrahim Khebour, Avyakta Chelle, Mariah Bradford, Jingxuan Tu, Hannah VanderHoeven, Brady Bhalla, Austin Youngren, James Pustejovsky, and Nikhil Krishnaswamy. 2025b. Propositional extraction from collaborative naturalistic dialogues. *Journal of educational data mining*, 17(1):183–216.