A Model of Information State in Situated Multimodal Dialogue

Kenneth Lai[1], **Lucia Donatelli**[2], **Richard Brutti**[1], **James Pustejovsky**[1]
[1]Brandeis University, Waltham, MA, USA
[2]Vrije Universiteit Amsterdam, Amsterdam, Netherlands
{klai12, brutti, jamesp}@brandeis.edu, l.e.donatelli@vu.nl

Abstract

In a successful dialogue, participants come to a mutual understanding of the content being communicated through a process called *conversational grounding*. This can occur through language, and also via other communicative modalities like gesture. Other kinds of actions also give information as to what has been understood from the dialogue. Moreover, achieving common ground not only involves establishing agreement on a set of facts about discourse referents, but also agreeing on what those entities refer to in the outside world, i.e., *situated grounding*. We use examples from a corpus of multimodal interaction in a task-based setting, annotated with Abstract Meaning Representation (AMR), to explore how speech, gesture, and action contribute to the construction of common ground. Using a simple model of information state, we discuss ways in which existing annotation schemes facilitate this analysis, as well as information that current annotations do not yet capture. Our research sheds light on the interplay between language, gesture, and action in multimodal communication.

1 Introduction

In dialogue, the concept of *common ground* refers to the set of presuppositions held by the participants, propositions that they agree to treat as true (Stalnaker, 1978). The process by which common ground is constructed over a dialogue is known as *(conversational) grounding* (Clark and Brennan, 1991). Formal models of dialogue have been developed to track how common ground (and more generally, information state) evolves over the course of an interaction (Poesio and Traum, 1997; Cooper and Larsson, 1999; Ginzburg, 2012).

Much work examining the role of non-linguistic modalities in communication focuses on gesture (Kendon, 2004; McNeill, 2008; Lascarides and Stone, 2009). This includes analyses of the semantic contents of gestures (Ebert and Ebert, 2014; Schlenker, 2018), and proposals for integrating gesture into models of dialogue (Lücking and Ginzburg, 2020).

Figure 1: Example of multimodal communication in a task-based setting (Wang et al., 2017). On the left, the signaler describes part of the structure to be built: he says, “It starts in the top left; there’s a block”, and makes a deictic gesture with his left hand. On the right, the actor puts a block in the top left corner of the table (note that both videos are mirrored).

More general types of actions can also affect dialogue context, especially in real-world or embodied settings (Tam et al., 2023). Within these settings, *referential grounding* is the process by which interlocutors anchor linguistic expressions to actual entities, relations, or events in the shared environment. When considering the perception and embodiment of participants, *situated grounding* is used (Kordjamshidi et al., 2025). In other words, while conversational grounding focuses on “what was said”, referential grounding ensures everyone agrees on “what is being talked about”.

In this paper, we present a simple information state model of dialogue that integrates both propositional updates (conversational grounding) and referential anchoring (situated grounding). We walk through a dialogue fragment from a corpus of task-based multimodal interaction (Lai et al. (2024); Wang et al. (2017); an example is shown in Fig-

Proceedings of the 16th International Conference on Computational Semantics, pages 292–298
September 22-23, 2025,

ure 1), annotated with AMR (Banarescu et al., 2013) for speech and gesture (Brutti et al., 2022; Donatelli et al., 2022), illustrating how speech, gesture, and object-directed actions co-construct and update the common ground. We assess the strengths and limitations of current annotations for capturing multimodal grounding phenomena, and argue for the importance of situational information in dialogue interpretation.

2 Related Work

Information state theories of dialogue are based on the idea that dialogue acts change the context available to participants (Fernández, 2022). At the most basic level, this includes the common ground, or shared assumptions of the participants (Stalnaker, 1978). Over time, the scope of the information state has expanded to handle different types of utterances beyond assertion; interrogatives are commonly handled via a set or stack of questions under discussion (Roberts, 2012), while there are various theories for the meaning of imperatives (Kaufmann, 2012; Portner, 2004; Barker, 2012). Formal information state theories include Poesio and Traum (1997); Cooper and Larsson (1999); and Ginzburg (2012).

Several dialogue corpora analyze the conversational grounding process and the impact of situated grounding or information about the shared environment. Among them, Mohapatra et al. (2024) annotate two corpora with (conversational) grounding acts and grounding units (Traum, 1995). The STAC corpus contains multi-party Settlers of Catan chats annotated with discourse structure and dialogue acts (Asher et al., 2016); Martinenghi et al. (2024) experiment with using large language models to predict the dialogue acts. Zhu et al. (2023) present the FIREBALL dataset of Dungeons & Dragons games, showing that adding game state information to the dialogue history can improve narration generation. Kruijt et al. (2024) develop the SPOTTER framework to investigate linguistic convention formation in a task referentially grounded in vision. The SCOUT corpus of situated human-robot dialogues (Lukin et al., 2024) is annotated with Dialogue-AMR (Bonial et al., 2020) and relations between utterances (Carletta et al., 1996; Traum et al., 2018). The Weights Task Dataset of situated interaction is annotated with several modalities including speech and gesture (Khebour et al., 2024a); Khebour et al. (2024b) perform common ground tracking, focusing on the emergence of facts.

3 Analyzing Multimodal Interaction

3.1 Setting

We draw examples in this paper from the EGGNOG corpus of task-based multimodal communication (Wang et al., 2017). Two participants are located in separate rooms, connected through video and audio. One person, the signaler, has an image of a block structure, and instructs the other person, the actor, on how to build the structure. For part of the corpus, Lai et al. (2024) annotated the signaler's speech and gesture with AMR (Banarescu et al., 2013; Brutti et al., 2022; Donatelli et al., 2022). While they did not annotate the actor's actions, our examples use another AMR extension, Action AMR (Tam et al., 2023), to describe them.

3.2 Information State

We use a simple model of information state, inspired by Ginzburg (2012)'s dialogue gameboard. Our model $M = (C, Q, T_s, T_a, E, g)$ contains the common ground C, which we assume to have a similar structure to a file card (Heim, 1982) or Discourse Representation Structure (Kamp, 2002), namely, that it stores a set of discourse referents and facts or shared beliefs about them. It also contains a set of questions under discussion Q. We take imperatives to denote actions; while Barker (2012) does not prescribe any specific data structure for these, we adopt Portner (2004)'s concept of a To-Do List T (one each for the signaler s and actor a) to handle actions. To describe the environment in which the participants are situated, we use a list E containing the objects in the environment (including the agents themselves), and the previous actions performed, both communicative and not; this is similar to the "common ground structure" in Pustejovsky and Krishnaswamy (2021) and Lai et al. (2021). Finally, to represent the situated grounding of objects and actions to the environment, we use an embedding or grounding function g. This is similar to the notion of an embedding in Discourse Representation Theory (Kamp, 2002), a function mapping discourse referents to elements in a model; here, the "model" comprises the environment E in which the agents are situated. For simplicity, we assume that the information state is an objective structure (i.e., not relative to any particular agent), and that all of its components are public; while each agent is assigned their own To-Do List, they also have access to the other participant's list.

Figure 2: Initial state for our example.

3.3 Example

We illustrate the dynamics of our information state using an example from the corpus. We note that because of the task-based nature of the interaction, the state does not begin empty. Both participants have prior information about the task, given by the experimenter or from previous trials; the common ground begins with these task-based presuppositions (see additional discussion in Section 4). Similarly, “what is the shape of the structure?” can be seen as an overarching question that begins in Q, the signaler has the task of communicating how to build the structure in T_s, and the actor has the task of actually building the structure in T_a. The environment contains the participants, the actor’s table[1], and the blocks, at least[2]. Finally, our example begins with the signaler already having given one instruction and the actor having put a block on the table, as shown in Figure 2.

In the corpus, signalers generally communicate their instructions through a combination of direct commands, and/or describing some aspect of the eventual structure. Here, the signaler does the former, issuing the imperative “Take another block; put it next to it” and gesturing towards a location on his table, as shown in Figure 3 (an example of the latter follows in Section 4). The signaler’s communicative act adds discourse referents to the common ground and actions to the actor’s To-Do List; the communicative act is itself recorded in

(1) “Take another block; put it next to it.”

```
(a / and
  :op1 (t / take-01 :mode imperative
          :ARG0 (y / you)
          :ARG1 (b / block
                  :mod (a2 / another)))
  :op2 (p / put-01 :mode imperative
          :ARG0 y
          :ARG1 b
          :ARG2 (n / next-to
                  :op1 (i / it))))
```

(2) Gesture for “put here”.

```
(g / gesture-unit
  :op1 (d / deixis-GA
          :ARG0 (s / signaler)
          :ARG1 (l / location)
          :ARG2 (a / actor))
  :op2 (i / icon-GA
          :ARG0 s
          :ARG1 (p / put-01)
          :ARG2 a))
```

Figure 3: The signaler gives the actor an instruction using speech (1) and gesture (2). Colors denote coreference relations between the AMRs.

the environment. These discourse referents and actions come from the speech and gesture AMRs, also shown in Figure 3. In this case, the signaler references a new `block` `b` to be placed at a new `location` `l`, and places `take` (`t`) and `put` (`p`) actions into T_a.

The actor shows her understanding of the signaler’s instructions by performing the referenced actions. The action and its corresponding AMR are shown in Figure 4. In the action AMR, note that the action and its arguments are not discourse objects, but rather objects in the world, that is, they are elements of E; for clarity, we use capital letters in the action AMR to mark this distinction. In performing the action, the actor *identifies* entities in the discourse with entities in the world, and *proposes* this identification to the signaler. That is, she is suggesting that $g(\texttt{b}) = \texttt{B2}$, $g(\texttt{l}) = \texttt{L2}$, $g(\texttt{p}) = \texttt{P2}$, and (given a suitably subevent structure for `put`, such as in Krishnaswamy and Pustejovsky (2021)), $g(\texttt{t})$ is a subevent of `P2`.

Note that the actor’s action does not automati-

[1]The signaler and actor being in different rooms complicates things somewhat. The signaler and actor both have tables in their rooms, and the signaler often uses locations on their table to refer to locations on the actor’s table, raising interesting questions of perspective and frame of reference. Ultimately, the actor’s table and the locations on it are the ones relevant to the completion of the task.

[2]One could argue that the environment should also include the locations in space available to the participants. Assuming a continuous space, enumerating every possible location would not be possible, so we allow for actions to dynamically generate locations as needed, a strategy employed by Krishnaswamy and Pustejovsky (2021).

(3) Actor puts another block next to the first block.

```
(P2 / put-01
  :ARG0 (A / actor)
  :ARG1 (B2 / block)
  :ARG2 (L2 / location))
```

Figure 4: The actor carries out the signaler's instruction. Proposed situated grounding between the action AMR and the communicative act is shown with the same colors as above (a subevent of the actor's put action corresponding to the signaler's take instruction).

cally update the situated grounding function g; it is now up to the signaler to accept or reject the actor's proposals. Mirroring Ginzburg (2012)'s treatment of statements yet to be accepted, the actor's suggestions become questions under discussion, $(g(\mathtt{b}) = \mathtt{B2})$?, and so on. If the signaler is satisfied with the actor's action, they can either give explicit positive acknowledgment, or implicitly accept by moving on to the next instruction; either way it is the signaler's acceptance that updates g. Otherwise, if there is something wrong, the signaler can either say or gesture so, and/or provide additional instruction to correct the misunderstanding.

In this example, the speaker's next communicative act is the utterance "Spread them apart a little bit but not as wide as a full block", with a corresponding "spread apart" gesture. While the actor's choice of block may have been appropriate, and $(g(\mathtt{b})$ is thus set to B2), the signaler intended there to be a gap between the blocks, and the actor's proposal of $(g(\mathtt{l}) = \mathtt{L2})$? is *not* accepted. The actor responds by moving both blocks to new locations a suitable distance apart; this represents a proposal not only to set the location of the second block, but also to update the location of the first block. The new proposals are eventually accepted by the signaler, and the dialogue continues.

4 Discussion and Conclusion

Within the corpus, some signalers use what we can call the *result present tense*, describing the configuration resulting from an action in the present, rather than giving an imperative. In fact, excluding one-word utterances, declarative sentences outnumber imperatives by almost two to one (191 to 97). In one example, the signaler says "Starting from the top, moving to your left, down four diagonally a row with the corners touching." The analysis of such utterances can be formalized in a number of ways. One approach, suggestive of Ross (1970)'s performative analysis, is to treat them like implicit imperatives: one could imagine each statement beginning with a covert "Make it true that...". These instructions would then be added to the actor's To-Do List, in the same way as explicit imperatives[3]. Another approach is to treat them as standard declaratives, with the actor's subsequent actions determined by pragmatic effects. Following Ginzburg (2012), declarative statements are offered as questions under discussion, which the actor can either accept or reject. Without an imperative, there is no direct update to the actor's To-Do List; however, assuming that they accept the statement, and the initial overarching task of building the structure remains in T_a, they will change the state of the world (i.e., move blocks around) to make the signaler's description true.

The challenge of ambiguous statements that require context for correct interpretation are well-established in dialogue literature (Grice, 1975). In sampling our corpus, we encounter two distinct kinds of ambiguity that require situated information to arrive at the correct interpretation. First, we notice several instances of presuppositions that are connected to the setup of the block-building task. These presuppositions are triggered with canonical utterances such as "again", "the same", or "also" (Frege, 1892; Strawson, 1950; Stalnaker, 1975). In one interaction, the signaler begins with the statement, "so you will begin with a grid structure **again**", referencing a previous interaction that required a grid-like spatial understanding of the block orientation on the table. We notice this throughout interactions: both signalers and actors approach the task with an implicit and often shared understanding of constraints on block structures and their orientation in the physical space.

In the same interaction, the signaler instructs the actor to create "**the same** pattern" with blocks in a new area of the table. Here, we encounter a second, partially overlapping challenge of multimodal ambiguity: multimodal coreference. In the case of the block pattern, the instruction and subsequent action

[3]We thank an anonymous reviewer for this suggestion.

are potential instances of the so-called *sloppy identity* effect (Ross, 1967), in which the same phrase can be interpreted with different arguments, i.e., blocks (Partee, 1975; Webber, 1978; Carnie, 2021). Such multimodal coreference can also be understood as *coreference under transformation* (Rim et al., 2023), a category easier to annotate and helpful in understanding sequences of events. Here, while the concept of a block pattern is stable in identity, the concept is applied to a new instance that requires situated knowledge to enact correctly.

Using AMR for both speech and gesture allows multimodal coreference relations throughout the dialogue and between the modalities to be marked using Multi-sentence AMR (O'Gorman et al., 2018). Meanwhile, using AMR for action facilitates alignment and binding from the communicative modalities to the local environment, allowing for easier identification of situated grounding. However, as the Lai et al. (2024) corpus annotates only communication from the signaler, there are certain aspects of conversational grounding, such as the signaler's understanding of the actor's communicative acts, that the annotations do not capture yet. A complete analysis of bidirectional grounding processes will require the rest of the corpus to be annotated with the actor's actions, in addition to their speech and gesture. Our model, focusing on describing *what* identifications are made between discourse entities and objects in the real world, sidesteps the question of *how* agents make these identifications. Kennington and Schlangen (2015) describe a "words as classifiers" approach to situated grounding of words and phrases in perceptual scenes. Furthermore, our findings are limited to a single corpus, and applying this approach to other dialogue types will reveal new insights. For example, in the block structure-building task, the signaler knows what structure is to be built, and the actor knows this, and therefore accepts the signaler as an authoritative source of information. Additionally, the task-specific presuppositions that define the initial dialogue state require knowledge of each new context. These factors point to clear next steps for extending multimodal semantic annotation for the analysis of situated dialogue.

Acknowledgments

We would like to thank Derrick Kim and Yifan Zhu for their assistance with this research. We would also like to thank the three anonymous reviewers for their detailed comments and suggestions.

This research was supported by the NSF National AI Institute for Student-AI Teaming (iSAT) under grant DRL 2019805 to James Pustejovsky. The opinions expressed are those of the authors and do not represent views of the NSF.

References

Nicholas Asher, Julie Hunter, Mathieu Morey, Benamara Farah, and Stergos Afantenos. 2016. Discourse structure and dialogue acts in multiparty dialogue: the STAC corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2721–2727, Portorož, Slovenia. European Language Resources Association (ELRA).

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.

Chris Barker. 2012. Imperatives denote actions. In *Proceedings of Sinn und Bedeutung*, volume 16, pages 57–70.

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.

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.

Jean Carletta, Amy Isard, Stephen Isard, Jacqueline Kowtko, Gwyneth Doherty-Sneddon, and Anne Anderson. 1996. *HCRC dialogue structure coding manual*. Human Communication Research Centre.

Andrew Carnie. 2021. *Syntax: A generative introduction*. John Wiley & Sons.

Herbert H. Clark and Susan E. Brennan. 1991. Grounding in communication. In Lauren B. Resnick, John M. Levine, and Stephanie D. Teasley, editors, *Perspectives on Socially Shared Cognition*, pages 127–149. American Psychological Association.

Robin Cooper and Staffan Larsson. 1999. Dialogue moves and information states. In *Proceedings of* *the Third International Workshop on Computational Semantics*.

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*, pages 293–312, Cham. Springer International Publishing.

Cornelia Ebert and Christian Ebert. 2014. Gestures, demonstratives, and the attributive/referential distinction. *Handout of a talk given at Semantics and Philosophy in Europe (SPE 7), Berlin*, 28.

Raquel Fernández. 2022. Dialogue. In *The Oxford Handbook of Computational Linguistics*. Oxford University Press.

Gottlob Frege. 1892. On sense and reference.

Jonathan Ginzburg. 2012. *The Interactive Stance*. Oxford University Press.

H. Paul Grice. 1975. Logic and conversation. In Donald Davidson, editor, *The logic of grammar*, pages 64–75. Dickenson Pub. Co.

Irene Heim. 1982. *The Semantics of Definite and Indefinite Noun Phrases*. Ph.D. thesis, University of Massachusetts Amherst.

Hans Kamp. 2002. A theory of truth and semantic representation. In Paul H. Portner and Barbara H. Partee, editors, *Formal Semantics - the Essential Readings*, pages 189–222. Blackwell.

Magdalena Kaufmann. 2012. *Interpreting Imperatives*. Springer Netherlands, Dordrecht.

Adam Kendon. 2004. *Gesture: Visible Action as Utterance*. Cambridge University Press.

Casey Kennington and David Schlangen. 2015. Simple learning and compositional application of perceptually grounded word meanings for incremental reference resolution. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 292–301, Beijing, China. Association for Computational Linguistics.

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 Hirshfield, Sadhana Puntambekar, Nathaniel Blanchard, James Pustejovsky, and Nikhil Krishnaswamy. 2024a. When text and speech are not enough: A multimodal dataset of collaboration in a situated task. *Journal of Open Humanities Data*.

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, Torino, Italia. ELRA and ICCL.

Parisa Kordjamshidi, Marie-Francine Moens, and James Pustejovsky. 2025. *Spatial Language Understanding: Representation, Reasoning, and Grounding*. Springer - Synthesis Lectures on Human Language Technologies, Switzerland.

Nikhil Krishnaswamy and James Pustejovsky. 2021. The role of embodiment and simulation in evaluating hci: Experiments and evaluation. In *International Conference on Human-Computer Interaction*, pages 220–232. Springer.

Jaap Kruijt, Peggy van Minkelen, Lucia Donatelli, Piek T.J.M. Vossen, Elly Konijn, and Thomas Baier. 2024. SPOTTER: A framework for investigating convention formation in a visually grounded human-robot reference task. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15202–15215, Torino, Italia. ELRA and ICCL.

Kenneth Lai, Richard Brutti, Lucia Donatelli, and James Pustejovsky. 2021. Situated umr for multimodal interactions. In *Proceedings of the 25th Workshop on the Semantics and Pragmatics of Dialogue - Poster Abstracts*, Potsdam, Germany. SEMDIAL.

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.

Alex Lascarides and Matthew Stone. 2009. A Formal Semantic Analysis of Gesture. *Journal of Semantics*, 26(4):393–449.

Andy Lücking and Jonathan Ginzburg. 2020. Towards the score of communication. In *Proceedings of the 24th Workshop on the Semantics and Pragmatics of Dialogue - Full Papers*, Virtually at Brandeis, Waltham, Massachusetts, USA. SEMDIAL.

Stephanie M. Lukin, Claire Bonial, Matthew Marge, Taylor A. Hudson, Cory J. Hayes, Kimberly Pollard, Anthony Baker, Ashley N. Foots, Ron Artstein, Felix Gervits, Mitchell Abrams, Cassidy Henry, Lucia Donatelli, Anton Leuski, Susan G. Hill, David Traum, and Clare Voss. 2024. SCOUT: A situated and multi-modal human-robot dialogue corpus. In *Proceedings* *of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14445–14458, Torino, Italia. ELRA and ICCL.

Andrea Martinenghi, Gregor Donabauer, Simona Amenta, Sathya Bursic, Mathyas Giudici, Udo Kruschwitz, Franca Garzotto, and Dimitri Ognibene. 2024. LLMs of catan: Exploring pragmatic capabilities of generative chatbots through prediction and classification of dialogue acts in boardgames' multi-party dialogues. In *Proceedings of the 10th Workshop on Games and Natural Language Processing @ LREC-COLING 2024*, pages 107–118, Torino, Italia. ELRA and ICCL.

David McNeill. 2008. *Gesture and Thought*. University of Chicago Press.

Biswesh Mohapatra, Seemab Hassan, Laurent Romary, and Justine Cassell. 2024. Conversational grounding: Annotation and analysis of grounding acts and grounding units. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3967–3977, Torino, Italia. ELRA and ICCL.

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.

Barbara Partee. 1975. Montague grammar and transformational grammar. *Linguistic Inquiry*, 6(2):203–300.

Massimo Poesio and David R. Traum. 1997. Conversational actions and discourse situations. *Computational Intelligence*, 13(3):309–347.

Paul Portner. 2004. The semantics of imperatives within a theory of clause types. In *Semantics and Linguistic Theory*, pages 235–252.

James Pustejovsky and Nikhil Krishnaswamy. 2021. Embodied human computer interaction. *KI - Künstliche Intelligenz*, 35(3):307–327.

Kyeongmin Rim, Jingxuan Tu, Bingyang Ye, Marc Verhagen, Eben Holderness, and James Pustejovsky. 2023. The coreference under transformation labeling dataset: Entity tracking in procedural texts using event models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12448–12460, Toronto, Canada. Association for Computational Linguistics.

Craige Roberts. 2012. Information structure in discourse: Towards an integrated formal theory of pragmatics. *Semantics and Pragmatics*, 5(6):1–69.

John Robert Ross. 1967. *Constraints on Variables in Syntax*. Ph.D. thesis, MIT.

John Robert Ross. 1970. On declarative sentences. In Roderick A. Jacobs and Peter S. Rosenbaum, editors, *Readings in English Transformational Grammar*, pages 222–272. Georgetown University Press, Washington, DC, USA.

Philippe Schlenker. 2018. Gesture projection and cosuppositions. *Linguistics and Philosophy*, 41(3):295–365.

Robert Stalnaker. 1975. Presuppositions. In *Contemporary Research in Philosophical Logic and Linguistic Semantics: Proceedings of a Conference Held at the University of Western Ontario, London, Canada*, pages 31–41. Springer.

Robert Stalnaker. 1978. Assertion. *Syntax and Semantics*, 9:315–332.

Peter F Strawson. 1950. On referring. *Mind*, 59(235):320–344.

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.

David Traum, Cassidy Henry, Stephanie Lukin, Ron Artstein, Felix Gervits, Kimberly Pollard, Claire Bonial, Su Lei, Clare Voss, Matthew Marge, Cory Hayes, and Susan Hill. 2018. Dialogue structure annotation for multi-floor interaction. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

David Rood Traum. 1995. *A computational theory of grounding in natural language conversation*. Ph.D. thesis, University of Rochester, USA. UMI Order No. GAX95-23171.

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, multi-modal 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.

Bonnie Lynn Webber. 1978. *A Formal Approach to Discourse Anaphora*. Routledge.

Andrew Zhu, Karmanya Aggarwal, Alexander Feng, Lara J. Martin, and Chris Callison-Burch. 2023. FIREBALL: A dataset of dungeons and dragons actual-play with structured game state information. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4171–4193, Toronto, Canada. Association for Computational Linguistics.