Context Is Key: Annotating Situated Dialogue Relations in Multi-floor Dialogue

Claire Bonial[1], Mitchell Abrams[2], Anthony L. Baker[1], Taylor Hudson[3], Stephanie M. Lukin[1], David Traum[4], and Clare R. Voss[1]

[1]U.S. Army Research Laboratory, Adelphi, MD 20783
[2]Institute for Human and Machine Cognition, Pensacola, FL 32502
[3]Oak Ridge Associated Universities, Oak Ridge, TN 37831
[4]USC Institute for Creative Technologies, Playa Vista, CA 90094
claire.n.bonial.civ@mail.mil

Abstract

In order to account for the features of *situated* dialogue, we extend a multi-party, multi-floor dialogue annotation schema so that it uniquely marks turns with language that must be grounded to the conversational or situational context. We then annotate a dataset of 168 human-robot dialogues using our extended, situated relation schema. Despite the addition of nuanced dialogue relations that reflect the kind of context referenced in the language, our inter-annotator agreement rates remain similar to those of the original annotation schema. Crucially, our updates separate data that can be used to train dialogue systems in essentially any context from those utterances in the data that are only appropriate in a particular situated environment.

1 Introduction

In order to account for the features of situated dialogue, we extend our multi-floor dialogue annotation schema described in Traum et al. (2018) to better capture the nuances that arise when a human and a robot collaborate on a search-and-navigation task. Using the same data collection procedure—a "Wizard-of-Oz" experimental design (Riek, 2012), in which participants directed what they believed to be an autonomous robot to complete search-and-navigation tasks (Marge et al., 2016, 2017)—we collect 168 human-robot dialogues and subsequently annotate them with a novel situated dialogue relation schema we present in this paper. While the original dialogue annotation schema is effective for multi-floor dialogue, it does not provide any indicator in the annotation schema demonstrating where the language requires grounding in the conversational or physical context.

In this paper, we address this problematic gap in the original annotation schema (described in §2) by introducing eight new annotation categories that uniquely mark where a particular interpretation/execution of the input natural language instruction relies upon some knowledge of the context—physical context, conversational context, or the robot's own physical form and abilities, and the interplay of these factors (updates described in §3). We annotate 168 additional dialogues with the augmented annotation schema and provide a corpus analysis (§4.1) and inter-annotator agreement (IAA) analysis (§4.2), which demonstrates that IAA remains high despite introducing new and somewhat nuanced annotation categories. We thus contribute an annotation schema and corpus that is better suited to serve as training data for situated dialogue systems by identifying precisely where the language must be grounded within the conversational or physical context in order to be interpreted and executed correctly.

2 Background

2.1 Human-Robot Dialogue Data

The experimental design of the human-robot dialogue data collection breaks up the planned autonomous robot capabilities into dialogue and navigation components, with one human experimenter, or "wizard," standing in for each component (depicted in Figure 1). A participant, acting as the "Commander," issues verbal instructions to their remotely located robot partner. Their instructions are heard and responded to by the "Dialogue Manager"

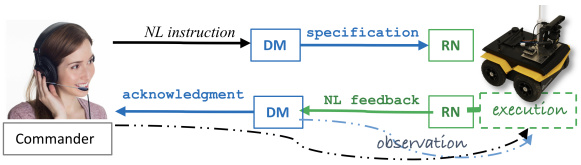

Figure 1: Natural Language (NL) and execution pairings for translate right (top) and acknowledgement (bottom). Dialogue turns for Commander in black, dialogue manager (DM) blue, robot navigator (RN) green.

(DM), whose role is to pass on a simplified version of the instructions, or "specification", via text message to the "Robot Navigator" (RN). The RN then joysticks the robot to execute the instruction, and this motion can be observed by the Commander and DM on a dynamically updating 2D LIDAR[1] map. The RN indicates completion or any problems in spoken natural language, and that status is acknowledged/described by the DM for the Commander. This data is therefore *multi-floor dialogue,* where communications between the Commander and DM is considered one conversational floor, while communications between the DM and the RN is considered another conversational floor. Note also that the information available to interlocutors within each floor is distinct—The Commander is unfamiliar with the remote environment and can *only* understand this environment based on the 2D LIDAR map that builds and updates as the robot enters a new space, while the DM and RN (i.e., the robot) are familiar with the space and are furnished with a map of the environment populated with unique names for each of the landmarks and spaces.

In previous work, we collected 60 human-robot dialogues following this protocol (Bonial et al., 2017; Marge et al., 2017). These dialogues were annotated using the schema presented in Traum et al. (2018), and used to train robot dialogue systems (Lukin et al., 2018; Gervits et al., 2021). In these efforts, input language from the Commander is associated with both feedback utterances from the DM to the Commander, as well as specifications for executing the input instructions in the form of the DM's "translation" of those instructions that are sent to the RN. Thus, on a high level, the dataset can be thought of as comprised of pairings of input natural language instructions with specifications for execution, and the complementary pairings of execution and generated natural language descriptions of what will be done, what is being done, or what has been done (see Figure 1).[2] Although effective for some natural language input, the annotation schema did not distinctly mark places where a particular pairing of natural language and execution was *only* valid in the conversational or physical context in which it arose. As a result, dialogue systems trained on this data could not handle, for example, input language that referenced a particular landmark in the current physical context, such as *the door ahead on the right*. Our modified annotation schema addresses this gap.

[1]Light Detecting and Ranging sensor

[2]Specifications are a controlled language, constrained to utterances included in the DM's wizard GUI described in Bonial et al. (2017), but are not a robotic planning language.

2.2 Dialogue Annotation

There are a variety of annotation schemas for dialogue available, including ISO standards for both dialogue acts (Bunt et al., 2012) and discourse relations (Prasad and Bunt, 2015). While these offer relations appropriate to dialogue, and perhaps even multi-party dialogue, they do not address the intricacies and challenges of *multi-floor dialogue*. Multi-floor dialogue is the focus of our original annotation schema, and is defined as "cases in which the high-level dialogue purposes are the same, and some content is shared, but other aspects of the information state, such as the participant structure and turn-taking expectations, are distinct" (Traum et al., 2018, p. 104).

This dialogue annotation schema was used to annotate the multi-floor, human-robot dialogue dataset described in Section 2.1. A dialogue excerpt from the dataset is given in Table 1. The annotation follows Grosz and Sidner (1986)'s *intentional structure* using the TRANSACTION UNIT (TU), which comprises an initial message from one speaker and all subsequent utterances across all floors that address the intention of that initial message. The internal structure of the TU is annotated using RELATIONS (rels) describing how a subsequent utterance relates to, or addresses, a previous utterance or its ANTECEDENT (ant). Relations are organized into a taxonomy of types where higher-order categories are distinguished based on whether they describe relations between utterances within or across conversational floors, and within or across speakers in a single floor. EXPANSIONS are relations between utterances of the same speaker and within the same conversational floor. RESPONSES are relations between utterances by different speakers within the same floor. TRANSLATIONS are relations between utterances in different conversational floors. Within each of these broad relation types, there is at least one but often two levels of relation subtypes. For example, TRANSLATIONS have two subtypes to characterize whether the information is being translated from the left floor to the right floor, TRANSLATION-**R**IGHT, or from the right floor to the left floor, TRANSLATION-**L**EFT (see Table 1). In contrast, RESPONSE has 17 subtypes, including ACKNOWLEDGMENT relations, which in turn has 8 subtypes, including **ACK**NOWLEDGMENT-DOING

	Left Floor		Right Floor		Annotations			
#	**Commander**	**DM→Commander**	**DM→RN**	**RN**	**TU**	**Ant**	**Rel**	**New Rel**
1	turn east ninety degrees				1			
2	and travel three feet				1	1	continue	
3		processing...			1	2*	processing	
4			turn left 90 degrees		1	1	translation-r	translation-r-situated
5			then...		1	4	link-next	
6			move forward 3 feet		1	2	translation-r	translation-r-default
7		turning...			1	1	ack-doing	
8		moving...			1	2	ack-doing	
9				done	1	6*	ack-done	
10		done			1	9	translation-l	

Table 1: Annotation exemplifying one TU with a situated translation (#4) and a default assumption that *travel* involves forward movement (#6), shown with original relations of Traum et al. (2018) and updated relations.

and ACKNOWLEDGMENT-DONE, which indicate that an instruction is being or has been carried out (see Table 1). For full details of all relation types, we refer the reader to Traum et al. (2018).

3 Situated Annotation Schema

We are addressing not only multi-party, multi-floor dialogue, but also *situated* dialogue that often draws upon the surrounding physical context, as well as the dialogue history and some assumptions relevant to the robot's own embodied form and capabilities. While our original annotation schema is uniquely suited to multi-floor dialogue, we have made several modifications to address the situated, contextual nature of the multi-floor dialogue found in the data.[3] These additions are summarized in Table 2, where we have also listed the directly relevant original annotation categories that we expanded upon. Our additions are made to two main relation types:

i. TRANSLATIONS from the left floor to the right floor, which allow us to pinpoint where and how certain translations draw upon the physical or conversational context (§3.1);

ii. ACKNOWLEDGMENTS of a preparatory action, not explicitly instructed, that is needed given the particular situated context or the particular capabilities and behaviors of the robot (§3.2).

[3]Note that although we see broad applicability of the annotation scheme, particularly the high-level types, with initial attempts to annotate other multi-floor dialogue corpora, such as (Martinovski et al., 2003), our approach has been to articulate only the low-level actions that appear in the analyzed data. Thus this research showcases the challenges of a specific domain, task, and robot (see Bonial et al. (2021) for related efforts extending dialogue annotations to a new task and domain).

3.1 Translation Across Floors

Translation across the conversational floors occurs when the speaker conveys the content from one conversational floor to an addressee in another conversational floor. In the data of interest here, TRANSLATIONS occur when the DM passes a message from the Commander in the left floor to the RN in the right floor, TRANSLATION-R (often instructions to be executed by the RN), or when the DM passes a message from the RN in the right floor to the Commander in the left floor, TRANSLATION-L (often feedback on the execution status of instructions). In our original annotation schema, TRANSLATION-L and TRANSLATION-R were the only two translation relations, along with a -PARTIAL flag that was used to indicate if the translation only addressed part of the original instruction. TRANSLATIONS from the left floor to the right are a critical aspect of training dialogue systems, as they provide the association between an unconstrained natural language instruction and a specification for execution by a robot (which has only a constrained behavior set). Essentially, TRANSLATIONS provide critical data for associating language and behavior.

However, we found that one cannot assume that a particular association is applicable in all physical and conversational contexts. In fact, a particular translation from the left floor to the right (TRANSLATION-R) is often valid only in the unique situational and conversational context where it was originally uttered. If such cases are not annotated distinctly from TRANSLATIONS that are valid in any context, this can lead to system responses that were learned in the training data but are not appro-

Relation	Definition
Translation-left	Provides the same content from speaker in right floor to addressee in left floor.
Translation-right	Provides the same content from speaker in left floor to addressee in right floor.
Translation-right-Direct	Uses the same or synonymous words, where the translation is applicable in any physical or conversational context.
Translation-right-Contextual	Draws upon situational or conversational context, but precisely what contextual information is being used is unclear, underspecified, or there are two or more factors.
Translation-right-Landmark	Refers to a unique landmark name known only to members of the right floor.
Translation-right-Situated	Relevant and/or synonymous to the original instruction in the current physical context but does not refer to a unique landmark.
Translation-right-History	All or part of the translation is only relevant given the dialogue history, in which it was established that a certain instruction should be interpreted in a particular way.
Translation-right-Default	Supplements information by relying on some default assumption related to a robot behavior or capability.
Translation-left-Partial	Only translates part of the command of an utterance or sequence.
Translation-right...Partial	Any of the above Translation-r subtypes that only translates part of the command of an utterance or sequence.
Acknowledgment-Will-comply	Acknowledgment of a command and a promise to do it in the future.
Acknowledgment-Doing	Acknowledgment that the speaker understands the command and that its execution is underway.
Acknowledgment-Done	Acknowledgment that a command or prior planned act has been completed successfully.
Acknowledgment-Will-comply Preparation	Acknowledgment of commitment to the preparation step consistent with compliance with the previous command, but not a promise/commitment of full compliance to the complete command (in contrast to will-comply)
Acknowledgment-Doing Preparation	Acknowledgment that the speaker understands the command and a preparation step required for compliance with the command is underway.

Table 2: Summary of added subcategories (shown in grey) and relevant categories from Traum et al. (2018).

priate in the test or use context. To address this issue, we introduce six new TRANSLATION-R subtypes to uniquely distinguish DIRECT TRANSLATIONS, which relate input language and execution specifications that can be consistently linked in any context, from those TRANSLATIONS that are only valid given particular aspects of the situational or conversational context.

3.1.1 Direct Translation

DIRECT TRANSLATIONS convey the content or intent of the speaker in one floor to the addressee in another floor, using the same or synonymous wording, without adding or subtracting content. These translations therefore relate specifications for execution with input language where this relation is consistently applicable—`Turn right 90 degrees` is *always* a valid expression of the specification for executing the instructions *Rotate right 90 degrees* or *Pivot 90 degrees to the right*, etc.[4] Accordingly, when used as training data, a strong association between input language and a particular execution is appropriate, regardless of context. Another example, translating *Take a photo* as `send image` is found in Table 3.

3.1.2 Landmark

In some cases, the instructions given refer to a particular object or landmark in the environment. Because of the nature of the experimental design where the DM and RN experimenters have complete information of the environment, all salient objects, rooms, hallways, and doorways were pre-coded with a unique identifier name. Thus, when the Commander mentions a particular landmark with a general reference (e.g., *Move to the doorway ahead on the right*), the TRANSLATION-R execution specification includes the specific name for that landmark (e.g., `Move through Kitchen-hall doorway`). Like other situational TRANSLATIONS, the reference used in the instructions cannot be consistently paired with a particular referent since *the doorway ahead on the right* will change depending upon the position of the robot. In Table 3, the translation `move into Conf Room` involves a LANDMARK TRANSLATION as the original destination reference of *through the doorway directly in front of you* is shifted to the named landmark `Conf Room`. Having an annotation category specifically for landmark mentions paves the way for experimentation incorporating a grounding system that will associate linguistic references to their referent in the environment.

3.1.3 Situated

In other cases, the instructions leverage spatial references to the environment as opposed to particular landmarks, where the execution specification

[4]Natural language instructions are italicized, while DM specifications are shown in Courier font.

	Left Floor		Right Floor		Annotations		
#	**Commander**	**DM→Commander**	**DM→RN**	**RN**	**TU**	**Ant**	**Rel**
1	go through the doorway directly in front of you				1		
2	and take a photo				1	1	continue
3		processing...			1	2*	processing
4			move into Conf Room		1	1	translation-r-landmark
5			then...		1	4	link-next
6			send image		1	2	translation-r-direct
7		moving...			1	1	ack-doing
8				uh done and sent	1	6*	ack-done
9		done, sent			1	8	translation-l

Table 3: Dialogue exchange exemplifying a LANDMARK TRANSLATION, referring to the unique identifier of the room that is referenced in the movement instruction (#1), and a DIRECT TRANSLATION of the second piece of the instructions (#2) that has distinct wording, but is applicable in any context.

for these instructions use a spatial reference that is only synonymous to the original in the current situated context. For example, in Table 1, the Commander instructs the robot *Turn east ninety degrees* where this is translated to `turn left 90 degrees`, which is only a valid specification for execution in that particular situated context—the robot's current heading is such that left and East are the same. Thus, again, the input language and execution specification cannot be consistently linked in all contexts. Although SITUATED TRANSLATIONS are conceptually a superset that includes LANDMARK TRANSLATIONS, we mark these distinctly as we expect grounding the references of SITUATED TRANSLATIONS to their referents will be more complex; they leverage abstract spatial language and regions as opposed to physical objects with clearer boundaries.

3.1.4 History

In some cases, an expectation for a certain behavior or certain manner of interpreting instructions may be set in the dialogue history, and then referenced later in the specification for execution. For example, several Commanders requested that the robot take a picture of what is in front of it after each movement instruction, to avoid repeating such requests as part of each instruction. As a result, a movement instruction such as *back up five feet* is linked with the translation `back up 5 feet...send image`, despite the fact that the direct antecedent instructions did not mention sending a picture. Instead, this portion of the specification for execution stems from the globally applicable request established in the dialogue history.[5] Other examples of HISTORY include anaphora (e.g., "take a picture of it"), deixis (g., e.g. "do that again"), or ellipsis (e.g., "two more feet"), where the translation includes the full content, part coming from previous TUs. Annotating these cases uniquely from other kinds of situational TRANSLATIONS again prevents the assumption that all cases of, for example, *back up five feet* be associated with an execution specification involving sending a picture, but also paves the way for incorporating higher-level instructions that apply throughout a dialogue by identifying where such instructions are deployed in the specification.

3.1.5 Default

DEFAULT TRANSLATION is applied when the input instruction does not make explicit some information which is instead inferred using a default assumption, generally regarding the robot's behaviors and capabilities. For example, in Table 1, the instruction *travel three feet* is linked with the translation `move forward three feet`, as it is assumed that the robot's default travel behavior would be a forward movement. Such assumptions are changeable based on the task, physical environment, and type of robot, thus the unique annotation category allows for identification of language/execution pairs that are only valid given a certain set of default assumptions.

[5] Instructions drawing upon utterances within the same TU are not annotated as HISTORY. E.g., an open-ended instruction *Move forward*, with a clarification—`How far?` *Three feet*—would have the DIRECT TRANSLATION `Move forward three feet`; the antecedents are the original instruction and clarification.

	Left Floor		Right Floor		Annotations		
#	**Commander**	**DM→Commander**	**DM→RN**	**RN**	**TU**	**Ant**	**Rel**
1	take a picture of the wall on your left				1		
2		processing...			1	1	processing
4			move to left wall		1	1	translation-r-contextual-partial
5			send image		1	4	continue
6		moving...			1	1	ack-doing-prep
7				done and sent	1	5*	ack-done
8		done, sent			1	7	translation-l

Table 4: Dialogue exchange where the translation of the instruction (#1) with the initial movement (#4) is motivated by underspecified and unknown aspects of the situated context, combined with default assumptions regarding where the robot needs to be to take an appropriate picture.

3.1.6 Contextual Translation, Underspecified

The back-off category CONTEXTUAL TRANSLATION is applied in cases where the kind of context used is underspecified such that it is not clear to the annotator what context, precisely, is being drawn upon, or more than one kind of contextual information is drawn upon within the same translation. In Table 4, the translation of the instruction *take a picture of the wall on your left* is translated with multiple steps, starting with the instruction to `move to left wall`, which is motivated by both the current position and orientation of the robot, as well as some default assumptions about the robot's abilities and requirements for taking a picture from a sensible vantage point. Since two types of context are used (the current situated context and default assumptions), CONTEXTUAL type is used.

3.2 Preparatory Actions

ACKNOWLEDGMENTS (Acks) and feedback to the participant are essential for establishing and maintaining common ground as well as trust and transparency in the system. As TRANSLATION-R relates pairs of input language to a specification of execution, ACKS relate the robot behavior/execution to a natural language description of that behavior that provides feedback and insight into what the robot intends to do, is doing, or has done. Reflecting the importance of this in dialogue, our original annotation schema included 8 subtypes of ACKS, capturing not only completion status, but also the level of confidence of the speaker in what was understood and the commitment to complete the instructed task (e.g., ACKNOWLEDGE-UNDERSTAND, ACKNOWLEDGE-UNSURE, ACKNOWLEDGE-TRY).

ACKS that do not match up with the Commander's intended instructions can be perceived as red flags that some miscommunication has taken place. This is beneficial when there is true misunderstanding, which can then be repaired. However, we found other cases where the robot acknowledged an action required to prepare for execution of the main instructed action. The preparatory action was not explicitly requested by the commander, and the commander might not understand the connection, so the DM utterance may be perceived as irrelevant by the Commander, and therefore signal misunderstanding. In fact, the robot has understood the instructions and is simply executing them in a way that reflects additional preparatory steps needed given its abilities. For example, an instruction *Take a picture of what's behind you* requires that the robot used for data collection first turn around 180 degrees before taking a picture, as its camera is a static, front-facing camera. ACKS that the robot will turn or is turning around, however, might not be perceived as appropriate acknowledgments of the original instruction, and therefore may actually undermine trust that the system has understood and is executing the instructions. To capture the fact that such ACKS are distinctly providing feedback on preparatory actions, we introduced two ACK relations described below. In the future, we will explore the potential value of making these acknowledgments explicitly mention the preparatory nature of the action (e.g., `I am turning in preparation to take a picture...`) to prevent the perception of a mismatch between the Commander's intention and the action being carried out.

3.2.1 Will Comply - Preparation

The first added subtype of acknowledgments, WILL COMPLY - PREP is used to mark acknowledg-

ments that reflect the speaker's commitment to do a preparatory step consistent with compliance with the previous command. While the relation type WILL COMPLY is acknowledgment of the speaker's commitment to comply with the command, this is not the case for WILL COMPLY - PREP, as there may be some intervening problem or clarification needed for full compliance, so the latter is a commitment restricted to the preparatory step. For example, the instruction *Go five feet north* requires the preparatory step of the robot turning to face North when its current heading is not in that direction. Note that just acknowledging the preparatory step `I will turn to face North` could be perceived as a mis-hearing or misunderstanding of the original instruction. Furthermore, this preparatory step is only needed in a physical context where the robot is not already facing North. Thus, like the situational TRANSLATIONS, this association of the natural language description of the execution of this instruction is only valid in a particular physical context. Marking these cases distinctly again allows us to pinpoint communications that rely upon context to be appropriate.

3.2.2 Doing - Preparation

The second added subtype of acknowledgments DOING - PREP is used to mark acknowledgments that a preparatory step consistent with compliance with the previous command is underway. For example, in Table 4, the instruction *Take a picture of the wall on your left* requires that the robot first move to the left wall (line 4) in order to take an appropriate picture. Like the acknowledgment WILL COMPLY - PREP, the feedback `moving` (line 6) could be perceived as evidence of a misunderstanding, since the original instruction does not mention any motion at all, and is instead focused on taking a picture. Again, this feedback is also only an appropriate natural language description of the execution specification given the specific physical context that requires the preparatory action.

4 Corpus & Annotation

We apply the updated dialogue annotation schema to a total of 168 dialogues collected from 56 Commanders, using the data collection procedure described in Section 2, however, using a virtual environment and robot.[6] Each Commander participates in three trials, corresponding to a different search-and-navigation task, which each lasts about 20 minutes. The spoken input of the Commander and spoken feedback of the RN were transcribed and time-aligned with the text chat messages of the DM. The aligned streams are compiled into a spreadsheet with rows and columns corresponding to the examples shown here in Tables 1, 3 and 4.

[6]This data can be released via a data-sharing agreement.

Relation	#	%
Translation-r	8556	
Direct	6017	70
Direct-partial	123	1
Contextual	163	2
Contextual-partial	47	<1
Landmark	766	9
Landmark-partial	67	<1
Situated	708	8
Situated-partial	201	2
History	200	2
History-partial	4	<1
Default	251	3
Default-partial	9	<1
Updated Ack Types	5573	
Will-comply	2092	38
Doing	3379	61
Will-comply-prep	27	<1
Doing-prep	75	1

Table 5: Frequencies and % of updated relations.

Annotations are added to the spreadsheet by one of a pool of undergraduate and graduate-level annotators with backgrounds in linguistics or computer science. All annotations are then validated by one of the senior project members.

4.1 Corpus Analysis

Across the entire corpus of 168 dialogues, there are 40,873 relations, and the most prevalent general relation types are ACKNOWLEDGEMENTS, making up 36.4% of corpus relations, and TRANSLATIONS, making up 36.5% of relations, with TRANSLATION-R comprising 20.9% of the corpus and TRANSLATION-L comprising 15.6%. Thus, the general relation types we update have a large impact on the corpus.

The frequencies of the extended relation types (and directly relevant original relations) are summarized in Table 5. DIRECT TRANSLATIONS, which do not draw upon any contextual information, are the majority (70%) of TRANSLATION-R. Thus, TRANSLATIONS that do draw on contextual information make up the remaining third of the corpus TRANSLATION-R relations, with LANDMARK and SITUATED TRANSLATIONS accounting for the largest percentages of 9% and 8%, respectively.

The updates to the ACK relations have a smaller impact on the corpus, as the new types make up only about 1.8% of the ACKS considered in this paper. However, we note that these complex cases

where the new PREP ACKS apply may still be prevalent enough to be problematic in training data for a dialogue system if they are not marked distinctly, and they can now be confidently separated out from potential noise or errors in the data where the wrong acknowledgment is mistakenly given, or the input instructions are genuinely misunderstood.

4.2 Inter-Annotator Agreement

Following the same procedure as Traum et al. (2018), we compute IAA on the three markables in the annotation schema: antecedents, relations, and transaction units (TUs). Three expert coders annotated a subset of 3 dialogues (a total of 896 utterances) using our extended schema. Results appear in Table 6, which also shows the reported IAA from the unmodified schema. Note that in the unmodified schema, two rounds of IAA were conducted, the first round on 3 dialogues of 482 utterances using 5 coders, and the second round on a single dialogue of 314 utterances using 6 coders. We compare this range of IAA from the four trials of the unmodified schema, to the range of IAA for the three trials annotated with the new schema.

Markable Type	Agreement		Distance Metric
	Unmodified Schema	Modified Schema	
Antecedents	0.72–0.82	0.79-**0.94**	Nominal[a]
Relation Types	0.77–0.89	0.83-**0.93**	Nominal[a]
Transaction Units	0.48–**0.93**	0.65-0.85	MASI[b]

[a]Krippendorff (1980) [b]Passonneau (2006)

Table 6: IAA of the original, unmodified schema of Traum et al. (2018) and our modified schema.

Our modified schema yields comparable or higher IAA than the original schema for antecedents (maximum 0.94) and relation types (maximum 0.93). Our TU IAA (maximum 0.85) is higher than the range of TU IAA reported for the first round of annotations with the unmodified schema (0.48-0.70), but the final round of TU annotation from in the unmodified schema achieves the highest agreement rate of 0.93. Note that our modified schema adapts the same coding for antecedents and TUs. Thus, although one might expect that adding annotation categories would lead to lower IAA, the addition of our new subtype relations did not produce significantly lower agreement scores, demonstrating that the new annotation categories are clearly identifiable.

5 Related Work

Speech and dialogue acts have been used as part of the meaning representation of task-oriented dialogue systems since the 1970s (Bruce, 1975; Cohen and Perrault, 1979; Allen and Perrault, 1980). For a summary of some of the earlier work in this area, see Traum (1999). Although the refinement and extension of Austin's (1962) hypothesized speech acts by Searle (1969) remains a canonical work on this topic, there have since been a number of widely used speech act taxonomies that differ from or augment this work, including an ISO standard (Bunt et al., 2012). Nevertheless, these taxonomies often have to be fine-tuned to the domain of interest to be fully useful.

With the aim of developing dialogue systems, Narayan-Chen et al. (2019) propose a dialogue act schema that is somewhat more limited than Traum et al. (2018), in order to annotate dialogue focused on a collaborative building task in the Minecraft gaming environment. Bonn et al. (2020) further annotate the Minecraft corpus with Abstract Meaning Representation (AMR) (Banarescu et al., 2013) that has been updated with more detailed spatial relations. Bonial et al. (2020) also propose an annotation schema that combines both illocutionary force and propositional content into an augmented version of AMR and use this to annotate a sample of the same human-robot dialogue dataset described in Traum et al. (2018). We plan to explore the contrasts and complementarity of these annotation schemas that have been used to annotate task-oriented dialogue.

6 Conclusions & Future Work

We extend the annotation schema presented in Traum et al. (2018) so that it now uniquely marks where the language requires grounding in the physical or conversational context. While much more work is needed to provide a schema capable of training a system on how it should relate language to context, our extensions take the first critical step towards such exploration, while also separating out the training data that is largely applicable in any context. We demonstrate that the new categories introduced, which all mark up distinct features of situated language, are clearly discernible to human annotators through IAA that remains high. We are optimistic that these extensions will improve performance of dialogue systems trained on this data, which we are currently implementing.

References

James F Allen and C Raymond Perrault. 1980. Analyzing intention in utterances. *Artificial Intelligence*, 15(3):143–178.

John Langshaw Austin. 1962. *How to Do Things with Words*. Harvard University Press and Oxford University Press.

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.

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 12th Language Resources and Evaluation Conference*, pages 684–695, Marseille, France. European Language Resources Association.

Claire Bonial, Matthew Marge, Ron Artstein, Ashley Foots, Felix Gervits, Cory J. Hayes, Cassidy Henry, Susan G. Hill, Anton Leuski, Stephanie M. Lukin, Pooja Moolchandani, Kimberly A. Pollard, David Traum, and Clare R. Voss. 2017. Laying Down the Yellow Brick Road: Development of a Wizard-of-Oz Interface for Collecting Human-Robot Dialogue. In *AAAI Fall Symposium*.

Claire N Bonial, Mitchell Abrams, David Traum, and Clare R Voss. 2021. Builder, we have done it: Evaluating & extending dialogue-AMR NLU pipeline for two collaborative domains. *Proceedings of the the 14th International Conference on Computational Semantics*.

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 12th Language Resources and Evaluation Conference*, pages 4883–4892, Marseille, France. European Language Resources Association.

Bertram C. Bruce. 1975. Generation as a social action. In *Theoretical Issues in Natural Language Processing*, pages 64–67.

Harry Bunt, Jan Alexandersson, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Volha Petukhova, Andrei Belis-Popescu, and David Traum. 2012. ISO 24617-2: A semantically-based standard for dialogue annotation. In *International Conference on Language Resources and Evaluation (LREC)*, Istanbul, Turkey.

Philip R Cohen and C Raymond Perrault. 1979. Elements of a plan-based theory of speech acts. *Cognitive science*, 3(3):177–212.

Felix Gervits, Anton Leuski, Claire Bonial, Carla Gordon, and David Traum. 2021. A classification-based approach to automating human-robot dialogue. In *Increasing Naturalness and Flexibility in Spoken Dialogue Interaction: 10th International Workshop on Spoken Dialogue Systems*, pages 115–127. Springer Singapore.

Barbara J. Grosz and Candace L. Sidner. 1986. Attention, intention, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

Klaus Krippendorff. 1980. *Content Analysis: An Introduction to Its Methodology*, chapter 12. Sage, Beverly Hills, CA.

Stephanie M. Lukin, Felix Gervits, Cory J. Hayes, Anton Leuski, Pooja Moolchandani, John G. Rogers, III, Carlos Sanchez Amaro, Matthew Marge, Clare R. Voss, and David Traum. 2018. ScoutBot: A Dialogue System for Collaborative Navigation. In *Proceedings of ACL 2018, System Demonstrations*, pages 93–98, Melbourne, Australia.

Matthew Marge, Claire Bonial, Brendan Byrne, Taylor Cassidy, A. William Evans, Susan G. Hill, and Clare Voss. 2016. Applying the Wizard-of-Oz technique to multimodal human-robot dialogue. In *RO-MAN 2016: IEEE International Symposium on Robot and Human Interactive Communication*.

Matthew Marge, Claire Bonial, Ashley Foots, Cory Hayes, Cassidy Henry, Kimberly Pollard, Ron Artstein, Clare Voss, and David Traum. 2017. Exploring variation of natural human commands to a robot in a collaborative navigation task. In *Proceedings of the First Workshop on Language Grounding for Robotics*, pages 58–66.

Bilyana Martinovski, David Traum, Susan Robinson, and Saurabh Garg. 2003. Functions and patterns of speaker and addressee identifications in distributed complex organizational tasks over radio. In *Diabruck: seventh workshop on semantics and pragmatics of dialogue*. Citeseer.

Anjali Narayan-Chen, Prashant Jayannavar, and Julia Hockenmaier. 2019. Collaborative dialogue in Minecraft. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5405–5415, Florence, Italy. Association for Computational Linguistics.

Rebecca Passonneau. 2006. Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. In *Proc. of LREC*.

Rashmi Prasad and Harry Bunt. 2015. Semantic relations in discourse: The current state of iso 24617-8. In *Proceedings 11th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-11)*, pages 80–92.

Laurel Riek. 2012. Wizard of Oz Studies in HRI: A Systematic Review and New Reporting Guidelines. *Journal of Human-Robot Interaction*, 1(1).

John Rogers Searle. 1969. *Speech acts: An essay in the philosophy of language*. Cambridge University Press.

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)*, pages 104–111, Miyazaki, Japan. European Language Resources Association (ELRA).

David R. Traum. 1999. Speech acts for dialogue agents. In Anand Rao and Michael Wooldridge, editors, *Foundations of Rational Agency*, pages 169–201. Kluwer.