Let's do that first! A Comparative Analysis of Instruction-Giving in Human-Human and Human-Robot Situated Dialogue

Matthew Marge[1], **Felix Gervits**[1*], **Gordon Briggs**[2], **Matthias Scheutz**[3], **Antonio Roque**[3]
[1]U.S. Army Research Laboratory, Adelphi, MD 20783
[2]U.S. Naval Research Laboratory, Washington, DC 20375
[3]Tufts University, Medford, MA 02155
matthew.r.marge.civ@mail.mil, gervif@gmail.com,
gordon.briggs@nrl.navy.mil, {matthias.scheutz, antonio.roque}@tufts.edu

Abstract

We present an annotation scheme that captures the structure and content of task intentions in situated dialogue where humans instruct robots to perform novel action sequences and sub-sequences. This representation identifies patterns and structural differences between human-human and human-robot communications. We find that humans engage in more dialogue about updating beliefs with other humans, while they are significantly more direct in their intentions with robots, incrementally instructing physical actions. Additionally, humans talk significantly less about plans with robots compared to other humans.

1 Introduction

Robots will inevitably be placed in open-world contexts and be expected to accommodate substantial changes in environment and task ability. Moreover, task-based communications between humans and robots will likely involve instructing robots to perform novel actions as it is impractical if not impossible for robots to have pre-defined knowledge that accommodates all situations. However, it is currently not clear whether and to what extent such task-based instructions differ dependent on whether the instructee is a robot or a human. To shed light on this question, we examine task-based instructions given by human instructors to either human or robotic instructees using a novel annotation scheme that captures the structure and content of task intentions in situated dialogue. Ultimately, this scheme will help improve the types of situated dialogues that robots are capable of handling, especially in terms of beliefs and plans.

To explore similarities and possible differences in instruction-giving, we analyze a task-based corpus where the task is for a human instructor to provide an instructee (human or robot) a series of instructions that require the movement of objects situated in the physical world. In advance of the interaction, the instructor prepares a diorama representing a miniature configuration of full-size objects in the space, and verbally tasks the instructee to move objects such that they match the diorama. Instructors participate in two trials, one with another human and one with an anthropomorphic robot as the instructee. The robot is controlled using a "Wizard of Oz" methodology, where wizard experimenters control the robot's behaviors without the instructor's awareness. This dataset, the *Diorama corpus* (originally collected in Bennett et al. (2017)) provides an opportunity to investigate whether human instructors teach robot instructees differently from human ones.

The annotation scheme we present in this paper aims to capture the content of **dialogue moves** in a human-robot interaction task. In doing so, we seek to better understand how humans will frame sequences of instructions to robots. Previous schemes for dialogue acts (e.g., the ISO standard (Bunt et al., 2017, 2020)) focus on observations in human-human dialogue and general-purpose dialogue structures. In contrast, our scheme is concerned with robot-directed language and describing dialogue acts in the context of a specific domain, e.g., performed as part of a conversational game (Carletta et al., 1996) or performed while updating an information state (Traum and Larsson, 2003). While the scheme focuses on the Diorama corpus task, it also builds on established dialogue move taxonomy work (Marge et al., 2017).

While the dialogue move taxonomy we present in this paper can be applied to human-human as well as human-robot dialogue, we consider our taxonomy to be *robot-centric* because at the highest level it is organized around the type of responses that might be expected from a robot. We are partic-

[*] Most contributions were performed as a graduate student at Tufts University.

ularly interested in *instructions*, defined as instructor utterances that include one or more dialogue moves meant to achieve a task-related physical, verbal communicative, or belief update task intention. A *task intention* is an instructor goal that the instructor wants the instructee to adopt (Cohen and Levesque, 1990); it can include a plan consisting of a sequence of one or more actions. Thus we classify dialogue moves based on the type of response that the instructor intends the instructee to perform, given the current context.

We divide the content of dialogue moves into (1) the dialogue move **type**, which indicates whether the move conveys a physical action intention, verbal communicative action intention, or belief update intention, and (2) the dialogue **move** itself, which identifies the specific actions or plans that the instructor intends that the instructee perform in the environment.

Along with this type and move analysis, instructions are further assessed in terms of their **structure**: whether they include an individual or more than one instruction, and if there is more than one instruction, whether or not the instructions are specified as needing to be in a certain sequence.

Our analysis is performed in this way to help identify notable differences between human-directed and robot-directed instructions in the Diorama corpus. Our contributions are the following:

- We introduce a novel dialogue move taxonomy that captures the structure and content of task intentions in situated dialogue.
- We present a comparative analysis of communications involving the instructing of novel action sequences in human-human and human-robot dialogues conducted in a physical environment, with results that indicate that humans convey more direct intentions to robots compared to other humans.

2 Related Work

2.1 Instruction-Giving in Human-Robot Dialogue

Extensive prior work has explored the core properties of instruction-giving in human-robot dialogue, as well as differences in how humans communicate with robots compared to other humans. Much of this work has focused on the task domain of *navigation* (Bugmann et al., 2004; Koulouri and Lauria, 2009; Marge and Rudnicky, 2011; Marge et al., 2017), where a desired capability for robots, especially in teaming with humans, is to move around the physical world (e.g., to support collaborative exploration, search-and-rescue, or delivery tasks). In a corpus of route instructions, Bugmann et al. (2004) found that robot instructees only succeeded in following the human-provided routes 63% of the time compared with 83% for human instructees. The difference lay in the ability of human instructees to detect and correct errors autonomously.

Other work has explored how error-handling can be achieved through dialogue interaction and how such strategies differ in human-human versus human-robot interactions. In a kitchen scenario, Gieselmann (2006) found that humans largely use "achievement" strategies for error-handling regardless of whether they are talking to a human or a robot. Examples of such strategies include paraphrases, repetition, and requesting missing information.

Other comparative analyses have identified differences in instruction-giving to humans versus robots. Koulouri and Lauria (2009) examined spatial language in the IBL corpus of route instructions in a Wizard-of-Oz task. They found that most instructions to a robot were in the form of simple actions, and that there were more of these kinds of instructions when the human could monitor the robot's actions versus when they could not. Without monitoring, people used more landmark references in their instructions to ensure grounding. Tenbrink et al. (2010) compared the ways in which route instructions were generated in human-human versus human-computer interaction. They found that human-provided utterances to a human partner were more complex than those provided to a computer partner. Specifically, this included more spatial language, more perspective shifting, more location references, and more complex syntax. On the other hand, when the partner was a robot, the instructions tended to be simplified, turn-by-turn commands. As a whole, these results highlight important differences in instruction-giving behavior for navigation tasks, and suggest that humans often simplify their language when instructing robots.

Other task domains have also been explored, including language-guided assembly and manipulation. In a dyadic collaborative task, Bennett et al. (2017) found that human instructors are polite regardless of whether their conversational partner is a human or robot. In particular, they found that humans commonly used *indirect speech acts (ISAs)* (Searle, 1969) to give instructions, and that there was no difference in the frequency of ISAs for human versus robot instructees. In a similar task, Briggs et al. (2017) also found high rates of ISAs in human-generated instructions to robots. The tasks surveyed so far are novel ones without conventionalized social norms, but Williams et al. (2018) found that ISAs are increasingly used in scenarios involving conventional social norms, e.g., restaurant ordering. These results suggest that in non-navigation tasks, people speak to robots much like they do to humans, highlighting the need for robots to represent politeness norms and interpret ISAs in task-oriented dialogue.

2.2 Teacher-Learner Dialogue with Robots

Limited work has explored differences in human-human versus human-robot communication in tasks involving the teaching of novel action sequences. One study by Schreitter and Krenn (2014) investigated the language of instruction-giving in a teacher-learner assembly task. They found that people used a broader vocabulary when instructing humans compared to robots, suggesting that teachers may adapt to the perceived capabilities of a learner. Some work has also explored the role of a robot teaching a human. For example, Torrey et al. (2013) found that human and robot assistants that used more hedges (e.g., "kind of") or discourse markers (e.g., "basically") in a cooking domain were perceived as more likable, more considerate, and less controlling. Moreover, robots were perceived as less controlling than humans even when they used identical discourse markers. These findings were replicated in Strait et al. (2014) and shown to be influenced by additional factors including robot appearance and interaction distance.

While the literature has explored important differences in how humans communicate with robots compared to other humans, what is missing is an understanding of the ways in which humans structure and communicate novel concepts and action sequences. This is important for the development of mechanisms to enable robots to identify plans and actions from dialogue interaction (Laird et al., 2017; Appelgren and Lascarides, 2019). The present work fills this gap by introducing a scheme to capture human task intention and using it to annotate a corpus of human-robot dialogue to better understand the properties of human instruction-giving that robots will need to process.

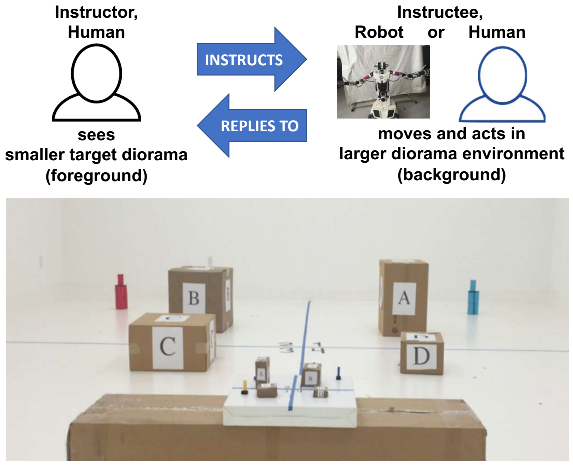

Figure 1: Instructors manipulate a miniature diorama (foreground) to a desired configuration, then command the instructee, who is either human or a robot, to move objects in the larger diorama environment (background) to match. During the task, the instructee cannot directly observe the miniature diorama.

3 Diorama Corpus Overview

The Diorama corpus (Bennett et al., 2017) consists of dialogues between human instructors and either a human or robot instructee. Instructors would teach the human and robot in separate trials a new "skill" (i.e., how to rearrange a set of objects *however they wanted*). As depicted in Figure 1, instructors reconfigured a miniature version of the room (i.e., a *diorama*) from an initial start state, then verbally conveyed a sequence of actions that would be required to move objects to match. The human instructee was a confederate experimenter, while the robot instructee was an anthropomorphic robot.

Instructors (naïve participants) told an instructee situated directly in the environment to perform this reconfiguration. In the segment of the corpus that we analyzed, an experimenter informed the instructor that the robot was fully autonomous. No specific capabilities of the robot were described. Two other experimenters used a Wizard-of-Oz interface to control the robot's movements and verbal responses, which were synthesized as speech.

3.1 Diorama Task

In all trials, the environment was divided into two parts, a *teaching* area where the instructor (seated) would have a diorama, and a larger *experiment* area. The instructor would reconfigure the diorama of miniature objects from an initial start state, then tell an instructee (human or robot) to replicate their configuration with the full-sized objects in the experiment area.

Beyond minor restrictions, instructors were free to rearrange the objects however they desired. The diorama and experiment area were comprised of four quadrants with seven movable objects: four cardboard boxes labeled with letters and three color-coded towers (in the diorama, towers were composed of Lego blocks while in the experiment area they were metal cans). After configuring their diorama, an instructee would enter the experiment area and introduce themselves. They would engage in spoken dialogue so the instructee would move the objects in the experiment area.

3.2 Interaction Design

Here we briefly outline the interaction design of the Diorama experiments (for additional details, see Bennett et al. (2017)). Example dialogues from the same instructor are presented in Figure 2. During the trials, both instructees had a limited set of dialogue responses. Their responses were constrained to “Okay”, “Yes”, and “No”, with the exception of a few clarification questions regarding status; for example, they could ask “Are there any more instructions?” to verify the trial was over.

The robot was controlled by an experimenter with an interface that permitted the robot to pick up and put down objects with its arms. At the end of the trial, the instructee would say “Goodbye” and leave the experiment area. The order in which the instructor interacted with either the robot or human was randomized.

3.3 Corpus Statistics

The entire corpus consists of thirty-three participants (21 female, 12 male; aged 18 to 25, *M*=20.85, *SD*=1.37), but of these, seventeen interacted with a robot they were told was fully autonomous. The focus of our analysis was on this subset because we are interested in the dynamic of perceived robot autonomy and instruction-giving. Not counting standard introductory and closure exchanges, participants said anywhere between 2 and 27 utterances to the instructee in a trial (*M*=8.7, *SD*=4.8). Overall, each dialogue between the participant and the instructee ranged from between 6 and 50 task-related utterances (*M*=17.1, *SD*=8.8).

Human Instructee Condition	
Instructor:	Um so, I’m going to ask you to move Box D forward...
Instructor:	I guess forward once you turn around, and to the left so that it’s right on the number 4.
Instructee:	Okay.
Instructee:	Should I do that now?
Instructor:	Yes please.
Instructee:	Okay.

Robot Instructee Condition	
Instructor:	Please move Box D backwards and to the left so that it sits on top of the number 4.
Instructee:	Okay.
Instructor:	Next, please move Box C backwards and to the right so that it rests on top of the number 3.
Instructee:	Okay.

Figure 2: Dialogue excerpts for an Instructor in Human and Robot Instructee conditions.

4 Corpus Annotation Scheme

We annotated the corpus on three levels: dialogue move type, dialogue move, and instruction structure. To calculate inter-annotator agreement, dialogues from seven randomly-selected participants (totaling 299 utterances) were annotated by two annotators. We report agreement for each level below.

4.1 Dialogue Move Types

Our taxonomy is meant to apply to human-human as well as human-robot interactions, but we think of the taxonomy as being *robot-centric* because its dialogue moves are organized into *types* based on the nature of response that is expected from a robot, for the given dialogue move and the given context. Informally, a robot might be expected to perform a physical action (such as moving a block), or to perform a verbal communicative action (such as describing its capabilities), or to update a belief without further action (such as understanding a plan that the instructor is describing).

More formally, we classify the **dialogue move type** based on *the primary reaction*[1] *that the instructor intends*[2] *the instructee to perform given*

[1] When we say *primary* reaction, we recognize that, for example, when the instructee performs a physical or verbal action, the instructee will typically also update their beliefs (i.e., that the action has been committed, and that the world has changed because of it). However, for the purposes of classification, the *primary* intention of the instructor is that the physical or verbal action be performed.

[2] When we refer to the instructor’s *intention*, we use the term in the sense of an action goal that the instructor wants the instructee to adopt (Cohen and Levesque, 1990): to perform a physical action, a verbal action, or a belief update action.

p-action	move-box	*move box D into square 1*
	knock-over-tower	*knock over tower number 2*
	rotate-box	*move box B so its facing that way*
	switch-boxes	*switch box B with box D*
	go-to	*walk over to the yellow tower*
	turn	*turn to your left*
	stop	*okay, stop*
	start-plan	*let's do that first*
	other-action	*yeah, put the yellow ones back up*
	non-task-related	*turn towards me (before we get started)*
v-action	check-task-ability	*see the tallest one over there to your left?*
	check-capability	*can you move those cans with your hands?*
	request-info	*what do I need to do?*
	non-task-related	*how's it going?*
b-update	plan	*then we'll do the same for the last box*
	past-action	*do the same thing with box A*
	update-plan	*we'll just forget about the towers*
	acknowledge	*that's good*
	share-info	*I didn't realize the towers are separate cans*
	non-task-related	*my name is Alex*

Figure 3: Taxonomy of Dialogue Move Types and Dialogue Moves, with examples.

atomic	*move to quadrant 2*	
non-atomic	compound	*knock down the blue and red towers*
	chain	*return to quadrant 1 and grasp the blue tower object*
	complex	*knock down the blue and red towers, then push the fallen blocks to quadrant 2*

Figure 4: Taxonomy of Instruction Structure Types, with examples.

the current context[3]. As shown in Figure 3 and as further described in Section 4.2, the three dialogue move types we identified are:

- **p-actions**: physical action intentions
- **v-actions**: verbal communicative action intentions
- **b-updates**: belief update intentions

There was substantial inter-annotator agreement for dialogue act types with 91.5% raw agreement (Cohen's unweighted $\kappa = .79$).

The dialogue move type is not determined by the full range of possible follow-up actions, responses, or updates that the instructee may perform; if the instructor says, "We'll just forget about the yellow tower," the instructee may legitimately move on to another tower (p-action), answer "okay" (v-action), or just update their beliefs (b-update), but clearly the utterance should not be classified as all three types. As defined above, the specific dialogue move type is defined by the speaker's intentions for the hearer, which is determined based on the context. This definition is motivated by our experience designing interactions for robots, during which one of the main concerns is determining what is expected of the robot at any given moment. Although there are numerous dialogue moves that may vary based on domain and scenario, we maintain that there are a far more limited number of *types* of expectations that those moves represent: expectations that the robot perform an action, that the robot say something, or that the robot do neither of those but instead just add something to its knowledge base of beliefs. From those types of expectations, we identified the three dialogue move types presented here.

[3] When we say the classification should consider the *current context*, this refers to the common ground (Clark and Schaefer, 1989), i.e., mutually understood material.

4.2 Dialogue Moves

We distinguish our *dialogue moves* from established schemes for describing *dialogue acts* (e.g., the ISO standard (Bunt et al., 2017, 2020)) by defining dialogue moves in the context of a specific domain, e.g., performed as part of a conversational game (Carletta et al., 1996) or performed while updating an information state (Traum and Larsson, 2003).

To define the range of dialogue moves that we would study, we began with an existing taxonomy of robot dialogue moves that had been developed in a collaborative navigation task (Marge et al., 2017). We used this as a basis for an initial dialogue move annotation of two representative interactions (four dialogues total) to develop the set of dialogue moves shown in Figure 3, which are divided into three types as described in Section 4.1. There was also substantial inter-annotator agreement for dialogue moves, with raw agreement of 81.6% (Cohen's unweighted $\kappa = .77$).

One type of dialogue move (*p-actions*) involves requests for physical actions such as manipulating a box in various ways, knocking over towers, and performing motions. A qualitatively different move, *start-plan*, indicates that a mutually understood plan (presumably established through belief update dialogue moves) should be initiated.

Another type of dialogue move (*v-actions*) involves requests that the instructee perform a verbal communicative action. For example, *check-task-ability* requests a verbal action that would provide information about the instructee's current ability (e.g., ability to see something, or reach something, or if there is a path somewhere). In contrast, *check-capability* requests a verbal action that would provide information about the instructee's general, potential capabilities (e.g., capability of bending over, or of grasping something in one hand).

The third type of dialogue move (*b-updates*) involves requests that the instructee update one or more of their beliefs without further physical or verbal action. This may involve, for example, utterances in which task-related plans are provided before the plans are actually enacted, utterances that update mutual beliefs while a plan is underway, and maintaining mutual beliefs. *Acknowledgment* dialogue moves provide a general expression of understanding or approval, usually intended as positive feedback. *Plan* dialogue moves refer to future or possible actions in more than a general way; *update-plan* dialogue moves revise previously-established plans, and *past-action* dialogue moves involve references to past actions in any other way – they are unusual in that they include p-action dialogue moves as a parameter, although they are not commands for action in themselves. *B-updates* highlight how the *primary* intended reaction is used to classify the dialogue move type because in a way, all dialogue moves involve updating beliefs. However, b-updates do not have any additional requirement for physical or verbal action.

All three dialogue move types include a catch-all dialogue move. For example, if an annotator encounters a physical action intention that is not explicitly defined by a dialogue move, that would be classified as an *other-action*. An undefined verbal action intention would be a *request-info*, and an undefined belief update would be a *share-info*.

All three dialogue move types also include a *non-task-related* move for utterances that are not part of the Diorama task. For the purpose of this analysis, we consider *preparing for the session* and *providing session setup instructions* to be different tasks than *performing the Diorama task*. As shown in Figure 3, a *p-action:non-task-related* dialogue move might be a request for the instructee to move towards them before beginning the Diorama task; only one example of this, "turn towards me," was seen in the corpus. Similarly, a *v-action:non-task-related* dialogue move might be a general off-topic question; the example in Figure 3 is notional, as no examples were seen in the corpus. Finally, a *b-update:non-task-related* dialogue move might be introductions or a farewell expression.

4.3 Instruction Structure

Along with the type and move annotations described above, instructions are further assessed in terms of their **structure**, as summarized in Figure 4.

An *atomic* instruction is an individual low-level instruction, while *non-atomic* instructions are made up of more than one atomic instruction. This is similar to the distinction between *minimal* and *extended* instructions made in Lukin et al. (2018) and Traum et al. (2018), but we expand upon the non-atomic/extended case in the following ways.

A *compound* non-atomic instruction has no constraints on the execution order of the atomic instructions that make it up. Consider Figure 4: in

		Instructee			
		Human		Robot	
Move Type	Move	%	N	%	N
p-action**		21 %	31	52.7 %	79
	move-box	13.6 %	20	27.3 %	41
	knock-over-tower	4.1 %	6	12 %	18
	rotate-box	0 %	0	1.3 %	2
	switch-boxes	1.4 %	2	2 %	3
	go-to	0 %	0	5.3 %	8
	turn	0 %	0	2 %	3
	stop	0 %	0	2.7 %	4
	start-plan	1.4 %	2	0 %	0
	other-action	4.8 %	7	4 %	6
v-action		5.4 %	8	5.3 %	8
	check-task-ability	2 %	3	4.7 %	7
	check-task-capability	0 %	0	0 %	0
	request-info	3.4 %	5	<1 %	1
b-update**		73.5 %	108	42 %	63
	plan*	32.7 %	48	11.3 %	17
	past-action	10.2 %	15	5.3 %	8
	update-plan	1.4 %	2	1.3 %	2
	acknowledge	9.5 %	14	6 %	9
	share-info	22.4 %	33	19.3 %	29
Utterance Count		100 %	147	100 %	150

* p < .05; ** p < .01

Table 1: Counts and percentages of dialogue move types and dialogue moves out of total number of instructor utterances. Significant differences between Instructee groups are marked by * $p < .05$, ** $p < .01$. Note that move type is only counted once per utterance, while each move may be counted more than once.

the example compound instruction, "knock down the blue and red towers," the towers in question can be knocked down in any order.

By contrast, a *chain* non-atomic instruction does have constraints on execution order; in the example chain instruction, "return to quadrant 1 and grasp the blue tower object," the instructee must return to quadrant 1 before grasping the blue tower object.

Finally, a *complex* utterance involves both compound and chain instructions. In the example complex instruction, "knock down the blue and red towers, then push the fallen blocks to quadrant 2," the blue and red towers can be knocked down in any order, but this knocking down must be performed before the fallen blocks can be pushed to quadrant 2.

We found a high rate of inter-annotator agreement for instruction structure, with a raw agreement of 97.1% (Cohen's unweighted $\kappa = 0.89$).

5 Results

The raw counts and percentages of dialogue move types and dialogue moves out of total instructor utterances are reported in Table 1. For statistical testing, we analyzed the proportion of move types, moves, and instruction structure to instructor utterances per trial using analyses of variance (ANOVAs) where instructee type (human or robot) and the order (robot first or not) were considered as fixed effects. Proportions were preferable in the analysis to raw counts because they mitigate the possible variations in dialogue length per participant[4]. After confirming that there were no main or interaction order effects[5], we performed ANOVAs on the proportions by instructee type. All *non-task-related* moves were excluded from our analysis.

		Instructee			
		Human		Robot	
Instruction Structure	Structure Type	%	N	%	N
atomic		78 %	60	88 %	88
non-atomic		22 %	17	12 %	12
	chain	9 %	7	7 %	7
	compound	13 %	10	5 %	5
	complex	0 %	0	0 %	0
Instruction Count		100 %	77	100 %	100

Table 2: Counts and percentages of instruction structure types out of total number of instructions.

5.1 Dialogue Move Types

When analyzing the proportion of dialogue move types to the total number of utterances by the instructor, several significant differences were found. A significant main effect for instructee type was found for both physical action (*p-action*) and belief update (*b-update*) proportions. Instructors used significantly more direct physical action intentions in their dialogues with robots (*M*=.48, *SD*=.36) compared to humans (*M*=.22, *SD*=.22), *F*[1,32] = 6.64, $p < .01$. In contrast, instructors used significantly more belief updates in their dialogues with humans (*M*=.71, *SD*=.27) compared to robots (*M*=.46, *SD*=.31), *F*[1,32] = 6.01, $p < .01$.

Of the dialogue moves that participants used, the most meaningful contrast was the use of the *plan* move. There were significantly greater proportions of plans in instructor utterances with the human instructee (*M*=.33, *SD*=.30) than the robot instructee (*M*=.13, *SD*=.26), *F*[1,32] = 4.40, $p < .05$. Among the other moves, the p-action moves *move-box* and *knock-over-tower* were markedly higher, though not significant, in dialogues with robots over humans.

5.2 Instruction Structure

We report the raw counts and percentage of instruction structure types out of total number of

[4] A secondary ANOVA-based analysis of raw counts confirmed all significant differences found with proportions.

[5] No significant main effects for order were found.

instructions in Table 2. No significant main effects were found for instructee type when comparing proportions of *atomic* instruction structure to total instructions per trial. The proportion of atomic instructions was similar for both instruction-giving to robots (*M*=.80, *SD*=.28) and to humans (*M*=.72, *SD*=.35).

6 Discussion

The work presented in this paper aims to identify similarities and differences in how humans instruct new action sequences to robots compared to other humans. Our findings suggest several fundamental differences both in the intentions for and the content of instructions.

6.1 Dialogue Move Types and Dialogue Moves

We found that instructors formulate instructions with different intentions when the task is to have an instructee perform actions in the physical world: they tend to engage in more dialogue about updating beliefs with humans, while more directly instructing physical actions with robots. We observed significantly higher proportions of belief update moves in human-human dialogues, while there were significantly higher proportions of physical action moves in human-robot dialogues. This result suggests that the underlying human mental models of the instructee differ with respect to what humans (intuitively) believe needs to be addressed: with humans, instructors know that their instructees know what to do and how to carry out actions if their beliefs are aligned with the instructor, whereas with robots, instructors assume that robots will do what they tell them to do.

Our results also support the conclusions of other studies presented in Section 2.1 where humans formulate less complex instructions to robots compared to other humans. This was further supported by a much higher use of plans in utterances directed to humans compared to robots. Given the complex nature of teaching a set of novel actions to a dialogue partner, these results suggest that humans prefer to rely on belief updating all at once when instructing other humans which is both more succinct and less burdensome for human instructees (because human instructees are aware of all the required actions) than instructing each action directly and incrementally.

In the case of robots, human instructors do not typically know whether their instructees are capable of doing the right thing if given high-level belief states; in fact, human instructors might not believe that robots can understand those types of belief utterances or have those types of belief states. This may be in part due to the fact that instructors were not given any details on the robot's capabilities, forcing them to rely on their own prior notions of what such a robot could do and possibly inferences based on their perceptions of the robot.

6.2 Instruction Structure

Looking at how many actions instructors chose to include in an utterance, we found no significant difference in the proportion of *atomic* to *non-atomic* instructions to either the human or robot instructee. In general, there was an overall heavy use of atomic instructions (at least 70% for both conditions). We believe this is in part due to the shared gaze permitted by the study setup: while they were in the same space, without any limitations on how many instructions they could give, the principle of *least collaborative effort* (Clark and Schaefer, 1989) suggests the least costly way of doing the task is to teach one chunk of the larger task at a time. We suspect that if the instructor and instructee were in separate spaces, or if there were limitations imposed on how many instructions they could give to the instructee, there could be substantial differences.

6.3 Limitations

A complication we identified while performing the annotations was that we received only the utterance-speaker pairs of the corpus, without information about timing, prosody, or visually co-occurrent actions. In the few cases where the instruction-giver appeared to repeat themselves, timing information would have confirmed whether this was actually a self-correction or a re-issuing of an instruction that was not acknowledged by the instructee. Additionally, a few utterances (e.g., "Like this?") referred to a gesture that the instructee was performing. While it was clear a physical action was underway, identifying the intention could be made clearer with access to visual recordings. We recommend in future studies that investigate such dialogue phenomena that visual and timing information be collected in addition to the spoken dialogue.

We also note that some of the differences in instruction-giving to human and robot instructees are because humans generally perceive robots differently from other humans (Bartneck et al., 2020). This fundamentally implies that there will be differences in the instructions given to robots; our results confirm this general concept.

7 Conclusions

In this paper, we presented a novel annotation scheme that enables comparisons between how humans instruct either a human or a robot instructee to perform previously unknown actions. The scheme centers on capturing the structure and content of task intentions in situated dialogue. We annotated a corpus of instruction-giving dialogues (Diorama corpus), and found substantial differences in how humans instruct robots compared to other humans. Humans tend to use dialogue to update beliefs of their instructees, as also indicated by a marked increase in the use of plans. In contrast, human instructors were more direct with robots, instructing physical actions in the space.

The results highlight the need to better understand people's implicit mental models of robots and what effects these mental models have on how humans formulate instructions for robots. This is particularly critical for the design of natural language understanding systems as well as cognitive architectures for future, more sophisticated robots. Such robots will likely be confronted with more humanlike instructions, and so will need to understand high-level beliefs and translate them into executable actions, which requires an understanding of the links between high-level goals, purposes, and how to realize them – all capabilities that current robotic systems lack. Future work should then explore how human instructors relate to perceived or known robot capabilities, and what level of sophisticated instructions robots would then be able to handle based on the changes to human mental models of robots.

Acknowledgments

This research was sponsored by the Basic Research Office of the U.S. Department of Defense with a Laboratory University Collaboration Initiative Fellowship awarded to the first and third authors and by a NASA Space Technology Research Fellowship (NSTRF) awarded to the second author. The authors would like to thank the anonymous reviewers for their helpful comments.

References

Mattias Appelgren and Alex Lascarides. 2019. Coherence, symbol grounding and interactive task learning. In *Proc. of SemDial*.

Christoph Bartneck, Tony Belpaeme, Friederike Eyssel, Takayuki Kanda, Merel Keijsers, and Selma Šabanović. 2020. *Human-Robot Interaction: An Introduction*. Cambridge University Press.

Maxwell Bennett, Tom Williams, Daria Thames, and Matthias Scheutz. 2017. Differences in interaction patterns and perception for teleoperated and autonomous humanoid robots. In *Proc. of IROS*.

Gordon Briggs, Tom Williams, and Matthias Scheutz. 2017. Enabling robots to understand indirect speech acts in task-based interactions. *Journal of Human-Robot Interaction*, 6(1):64–94.

Guido Bugmann, Ewan Klein, Stanislao Lauria, and Theocharis Kyriacou. 2004. Corpus-based robotics: A route instruction example. In *Proc. of Intelligent Autonomous Systems*.

Harry Bunt, Volha Petukhova, Emer Gilmartin, Catherine Pelachaud, Alex Fang, Simon Keizer, and Laurent Prévot. 2020. The ISO standard for dialogue act annotation, second edition. In *Proc. of LREC*.

Harry Bunt, Volha Petukhova, David Traum, and Jan Alexandersson. 2017. Dialogue act annotation with the ISO 24617-2 standard. In *Multimodal interaction with W3C Standards*. Springer.

Jean Carletta, Amy Isard, Jacqueline Kowtko, Gwnyeth Doherty-Sneddon, and Anne Anderson. 1996. *HCRC Dialogue Structure Coding Manual*. Human Communication Research Centre.

Herbert H. Clark and Edward F. Schaefer. 1989. Contributing to discourse. *Cognitive science*, 13(2):259–294.

Philip R. Cohen and Hector J. Levesque. 1990. Intention is choice with commitment. *Artificial intelligence*, 42(2-3):213–261.

Petra Gieselmann. 2006. Comparing error-handling strategies in human-human and human-robot dialogues. In *Proc. of KONVENS*.

Theodora Koulouri and Stanislao Lauria. 2009. A corpus-based analysis of route instructions in human-robot interaction. In *Proc. of Towards Autonomous Robotic Systems (TAROS)*.

John E. Laird, Kevin Gluck, John Anderson, Kenneth D. Forbus, Odest Chadwicke Jenkins, Christian Lebiere, Dario Salvucci, Matthias Scheutz, Andrea Thomaz, Greg Trafton, Robert E. Wray, Shiwali Mohan, and James R. Kirk. 2017. Interactive task learning. *IEEE Intelligent Systems*, 32(4):6–21.

Stephanie Lukin, Kimberly Pollard, Claire Bonial, Matthew Marge, Cassidy Henry, Ron Artstein, David Traum, and Clare Voss. 2018. Consequences and factors of stylistic differences in human-robot dialogue. In *Proc. of SIGdial*.

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 *Proc. of the First Workshop on Language Grounding for Robotics*.

Matthew Marge and Alexander Rudnicky. 2011. The TeamTalk corpus: Route instructions in open spaces. In *Proc. of the RSS Workshop on Grounding Human-Robot Dialog for Spatial Tasks*.

Stephanie Schreitter and Brigitte Krenn. 2014. Exploring inter- and intra-speaker variability in multimodal task descriptions. In *Proc. of RO-MAN*.

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

Megan Strait, Cody Canning, and Matthias Scheutz. 2014. Let me tell you! Investigating the effects of robot communication strategies in advice-giving situations based on robot appearance, interaction modality and distance. In *Proc. of HRI*.

Thora Tenbrink, Robert J Ross, Kavita E Thomas, Nina Dethlefs, and Elena Andonova. 2010. Route instructions in map-based human–human and human–computer dialogue: A comparative analysis. *Journal of Visual Languages & Computing*, 21(5):292–309.

Cristen Torrey, Susan R Fussell, and Sara Kiesler. 2013. How a robot should give advice. In *Proc. of HRI*.

David Traum, Cassidy Henry, Stephanie Lukin, Ron Artstein, Felix Gervits, Kimberly Pollard, Claire Bonial, Su Lei, Clare Voss, Matthew Marge, Cory J. Hayes, and Susan G. Hill. 2018. Dialogue structure annotation for multi-floor interaction. In *Proc. of LREC*.

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

Tom Williams, Daria Thames, Julia Novakoff, and Matthias Scheutz. 2018. “Thank you for sharing that interesting fact!” Effects of capability and context on indirect speech act use in task-based human-robot dialogue. In *Proc. of HRI*.