Exploring Variation of Natural Human Commands to a Robot in a Collaborative Navigation Task

Matthew Marge[1], Claire Bonial[1], Ashley Foots[1], Cory Hayes[1], Cassidy Henry[1],
Kimberly A. Pollard[1], Ron Artstein[2], Clare R. Voss[1], and David Traum[2]

[1]U.S. Army Research Laboratory, Adelphi, MD 20783
[2]USC Institute for Creative Technologies, Playa Vista, CA 90094
matthew.r.marge.civ@mail.mil

Abstract

Robot-directed communication is variable, and may change based on human perception of robot capabilities. To collect training data for a dialogue system and to investigate possible communication changes over time, we developed a Wizard-of-Oz study that (a) simulates a robot's limited understanding, and (b) collects dialogues where human participants build a progressively better mental model of the robot's understanding. With ten participants, we collected ten hours of human-robot dialogue. We analyzed the structure of instructions that participants gave to a remote robot before it responded. Our findings show a general initial preference for including metric information (e.g., *move forward 3 feet*) over landmarks (e.g., *move to the desk*) in motion commands, but this decreased over time, suggesting changes in perception.

1 Introduction

Instruction-giving to robots varies based on perception of robots as conversational partners. We present an experiment designed to elicit robot-directed language that is a happy medium between existing natural language processing capabilities and fully natural communication. The data elicited will be used to train a dialogue system in the future, and it provides insights into what communication strategies people use when instructing robots. In this paper, we begin to examine how people vary their strategies as they build a progressively more accurate mental model of the robot and its capabilities. To simulate a robot's limited understanding of its environment, we employ the Wizard-of-Oz (WOz) method, where humans simulate robot intelligence and actions without participant awareness. With ten participants, we collected ten hours of human-robot dialogue. We are currently undertaking corpus curation and plan to make the data freely available in the next year.

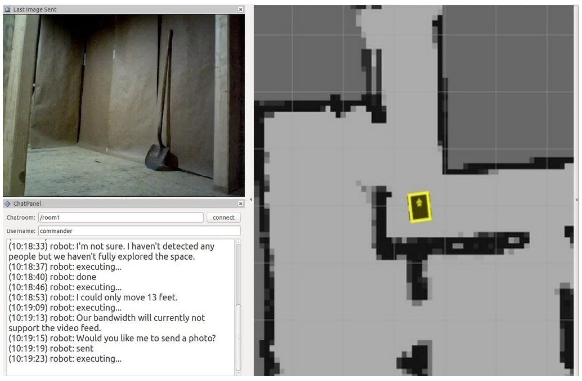
In this experiment, a human and robot engage in a series of *transactions* (Carletta et al., 1997) where an instruction is issued, and wizards acting on behalf of the robot either perform a task or prompt for clarification until the requested task is completed or abandoned. We propose a new term, *instruction unit* (IU), to identify all commands within a transaction issued before the robot generates a response. IUs were analyzed both in structure and variation. Our findings suggest a general, initial preference for including metric information over landmarks in motion commands, but this decreased over time. Results will assist in future work adapting robot responses to varied instruction styles.

In the sections to follow, we first give needed background: the experiment setup and our approach to eliciting natural, robot-directed language. We then describe the annotations we have undertaken thus far to explore communication strategies. In our results section, we provide some statistics on the data collected thus far as well as noted changes in communication strategies. We provide a discussion of these results and comparison to related work and close with a summary and description of future work.

2 Background

2.1 Collaborative Exploration Task

The domain testbed for our work is collaborative exploration in a low-bandwidth environment (Marge et al., 2016). This testbed mimics what can be found in a reconnaissance or search-and-navigation operation, wherein a human *Com-*



```
(10:18:33) robot: I'm not sure. I haven't detected any
people but we haven't fully explored the space.
(10:18:37) robot: executing...
(10:18:40) robot: done
(10:18:46) robot: executing...
(10:18:53) robot: I could only move 13 feet.
(10:19:09) robot: executing...
(10:19:13) robot: Our bandwidth will currently not
support the video feed.
(10:19:15) robot: Would you like me to send a photo?
(10:19:19) robot: sent
(10:19:23) robot: executing...
```

Figure 1: Interface showing robot text responses (*lower left*) to the Commander's verbal instructions, map (*right*), and the last still image sent (*upper left*).

mander verbally instructs a robot at a remote location, guiding the robot to move around and explore a physical space. The sensors and video camera on-board the robot populate a map as it moves, enabling it to describe that environment and send photos at the Commander's request, but the communications bandwidth prohibits real-time video streaming or direct teleoperation. The robot is assumed capable of performing low to intermediate level tasks, but not more complex tasks involving multiple or quantified goals, without clear directions or plans for ordering subgoals. The physical implementation of the testbed is an indoor environment, containing several rooms and connecting hallways, located in a separate building from the Commander. We use a Clearpath Robotics Jackal, fitted with an RGB camera and LIDAR sensors, to operate in the environment.

The Commander sees the following information from the robot's sensor data: a 2D occupancy grid with the robot's current position and heading streamed within the grid (i.e., map), and the last still image captured by its front-facing camera. In addition, the Commander can speak to the robot and see the robot's text responses. Figure 1 shows the information made available to the Commander.

2.2 Experiment Design

In each session, a (Commander) participant engaged the robot in collaborative search-and-navigation tasks. A session was comprised of three twenty-minute phases: a training phase and two main task phases (main phase 1 and 2). Training may voluntarily end when participants were comfortable with controls. Each phase focused on a slightly different search task and started in a distinct location. Experiment tasks were developed to encourage the participant to use the robot as a teammate to search for certain objects in the environment. The participant needed to use their real-world knowledge in order to answer questions that required analysis of the observed environment. The robot didn't know common words for target objects, which required participants to consider word choice as they addressed the robot. An example search task was to locate *shoes* in an environment, relying on robot-provided images. An example analysis task was to consider whether the explored space was suitable as a headquarters-like environment. All phases situated the robot in an unfamiliar indoor environment, unlike canonical scenes typically observed in homes and offices.

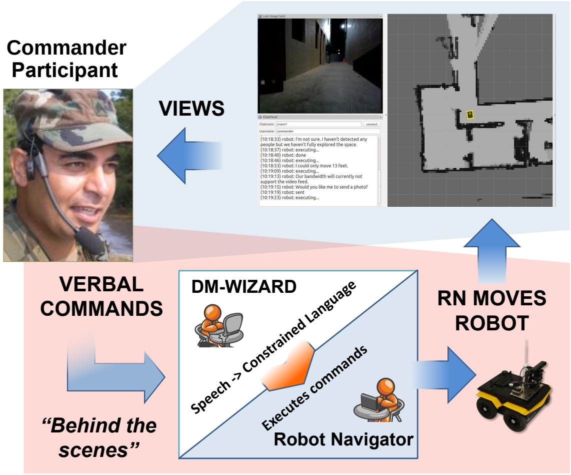
Preceding the study, participants received a list of robot capabilities (see Appendix A). They were told that the robot understood basic object properties (e.g., most object labels, color, size), relative proximity, some spatial terms, and location history. Participants were not given example instructions.

2.3 Wizard-of-Oz Setup

We use a WOz approach to allow for understanding of natural domain-specific instructions, in advance of collecting enough training data to implement an automated system. Our work expands on existing WOz approaches by incorporating multi-modal communication when the robot and human are not co-present – where information exchange of robot position, visual media, and dialogue is needed for collaborative exploration to succeed.

We use a multi-wizard setup to simulate the expected autonomous robot understanding and response. We use two wizards simultaneously for two reasons. First, a single wizard cannot type dialogue responses while teleoperating the robot with a joystick at the same time. Second, by design, we wish to decouple navigation behavior from dialogue behavior, as these will ultimately be separate modules in a fully-automated system.

A *Dialogue Manager* (DM-Wizard) listens to Commander speech and communicates directly with the Commander, using a chat window to type status updates and requests for clarification. When the Commander's instructions are executable in the current context, the DM-Wizard types in another chat window to pass a constrained, text in-



```
(10:18:33) robot: I'm not sure. I haven't detected any people but we haven't fully explored the space.
(10:18:37) robot: executing...
(10:18:40) robot: done
(10:18:46) robot: executing...
(10:18:53) robot: I could only move 13 feet.
(10:19:09) robot: executing...
(10:19:13) robot: Our bandwidth will currently not support the video feed.
(10:19:15) robot: Would you like me to send a photo?
(10:19:19) robot: sent
(10:19:23) robot: executing...
```

Figure 2: Wizard-of-Oz setup with wizards for dialogue management and robot navigation.

struction set to the *Robot Navigator* (RN), who teleoperates the robot. When hearing robot status updates directly from the RN, the DM-Wizard also communicates this information back to the participant. The DM-Wizard and RN roles were kept constant by having the same experimenters (female DM-Wizard, male RN) in those roles for the entirety of the study. Figure 2 presents our setup.

3 Approach: Eliciting Natural Language

One of the main research questions we seek to address with this experimental design is how to elicit natural communication, given that people may change strategies over time as they accommodate the robot's limited understanding. Like Chai et al. (2014) and Williams et al. (2015), we are interested in methods that robots can use to interpret and convey common ground in natural language interaction. Here, we describe how our DM-Wizard command-handling guidelines simulate a robot's limited understanding and the strategies that it could use to disambiguate phrases. Next, we introduce transaction and instruction units as a way to identify and measure possible variation in participant instructions.

3.1 DM-Wizard Guidelines

One way to elicit natural communications is to have the robot (in this case, the DM-Wizard) use strategies that mitigate its limited understanding, like offering suggestions or conveying its capabilities. We developed guidelines to determine when to employ such strategies and to ensure consistent dialogue decisions across participants. The guidelines governed the DM-Wizard's real-time decision-making. They first identify the minimal requirements for an executable command: each must contain both a clear action and respective endpoint. The guidelines provide response categories and templates, allowing for flexibility in exact response form, but with easily-remembered templates for elements of each response. Responses are broadly categorized into well-formed vs. unclear, problematic commands. The exchange in Figure 3 shows how a participant's problematic, open-ended instruction is handled under the guidelines.

Participant command (speech): *Move forward.*
Communication problem: Open-ended action (no endpoint specified)
Relevant template:
DESCRIBE PROBLEM + CAPABILITY
DM-Wizard response to participant (text): `How far? You can tell me to move to an object that you see or a distance.`
Participant response (speech): *Move to the yellow cone ahead of you.*

Figure 3: DM-Wizard guidelines for consistent dialogue behaviors (developed iteratively in piloting) applied in a sample exchange.

3.2 Dialogue Structure Annotation

In order to both study the question of what kinds of language, discourse, and dialogue strategies are used to give instructions, as well as to provide training data for automating the DM-Wizard functions, we annotated several aspects of dialogue structure. In this paper we focus on the former question, and examine how participants convey initial task intention to a robot before follow-on dialogue from the DM-Wizard. This analysis helps us understand the structure of instructions and anticipate possible task ranges required of a robot. We discuss four levels of dialogue semantics and structure below, from largest to smallest: *transaction units* (TUs), *instruction units* (IUs), *dialogue-moves*, and *parameters*. Each of these is defined and discussed below.

3.2.1 Transaction Units

Each dialogue is annotated as a series of higher-level *transaction units*. A TU is a sequence of utterances aiming to achieve task intention. The TUs document structures that appear within collected dialogues, while also providing emulatable interaction patterns for a dialogue manager. TUs

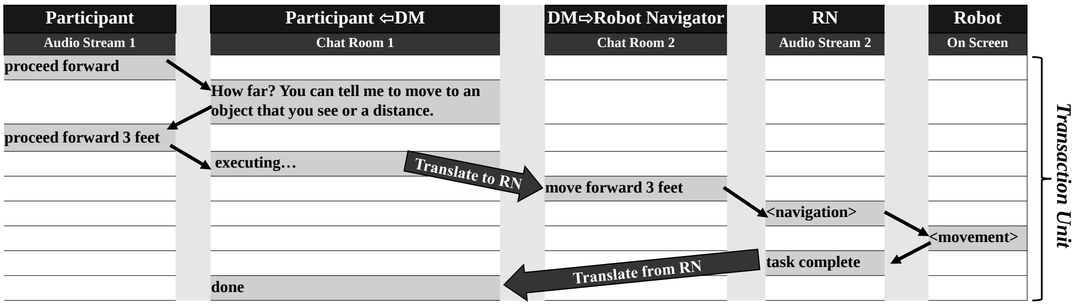


Figure 4: Two wizards manage the labor of robot intelligence. Dialogues divide into a series of *transactions* where a naive participant gives an instruction, a *Dialogue Manager* (DM-Wizard) decides how to handle it, and passes well-formed instructions to a *Robot Navigator* (RN) that moves the robot.

each contain a participant's initiating message and then subsequent messages by the participant and wizards to complete the transaction.

Figure 4 shows an example transaction in which a participant gives instructions, the DM-Wizard requests clarification, and the amended instructions are then passed to the RN, who completes the instructions.

3.2.2 Instruction Units

Within TUs, we marked *instruction units*. An IU comprises all participant speech to the robot within a transaction unit before robot feedback. Each IU belongs to exactly one TU, so that the start of each transaction (e.g., a new command is issued) marks a new IU. An IU terminates when the robot replies to the request, or when a new transaction is initiated. The relationships of IUs and TUs is shown in Figure 5.

3.2.3 Dialogue-Moves

To analyze internal IU structure, we annotated Commander-issued lower-level *dialogue-moves*. This annotation scheme is inspired by a prior approach to military dialogue that identified dialogue-moves in calls for artillery fire (Roque et al., 2006). Examples of a *command* type request are *command:drive* or *command:rotate*, that instruct the robot to perform certain motions. A dialogue-move list is provided in Appendix B.

Three annotators independently validated the dialogue-move set on 99 dialogue turns in our human-robot dialogue corpus. Annotators had high agreement ($\alpha = 0.92$; Krippendorf's α using the MASI distance measure (Passonneau, 2006)).

3.2.4 Parameters on Motion Commands

Some dialogue moves uniquely define the action that the robot should take, e.g. *command:stop* or *command:send-image*. Others require additional *parameters* to fully specify the complete action. Of particular interest to us is the information that participants chose to include in robot-directed motion requests. We focused on *command:drive* and *command:rotate* for variation in how participants communicated. We annotated motion-command parameters for their usage of *metric* (e.g., *move forward 2 feet; turn left 90 degrees*) and *landmark*-based points of reference (e.g., *move to the table; turn to face the doorway*) similar to *absolute* and *relative* steps in route instructions from Marge and Rudnicky (2010).

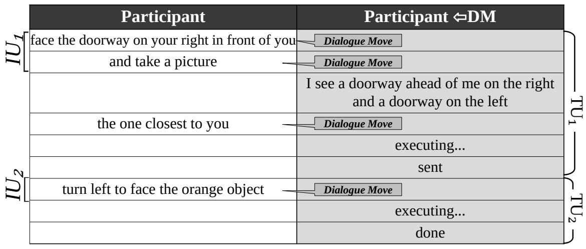


Figure 5: Annotation structures on human-robot dialogue, shown over participant and DM-Wizard streams.

3.3 Participants

This study recruited ten participants: two female, eight male. Ages ranged from 28 to 58 (mean = 44, s.d. = 10.6). Two participants reported one year or less of robotics research; others reported none.[1]

[1] We collected measures that were included in our statistical analysis but not presented in this paper. The Spatial Orientation Survey, part of the Guilford-Zimmerman Aptitude Survey (Guilford and Zimmerman, 1948), assesses spatial orientation perception. The HRI Trust Survey (Schaefer, 2013) measures subjective trust of the robot, based upon personal belief of the robot's capabilities.

3.4 Corpus Statistics

We collected approximately 10.5 hours of recorded Commander speech (approximately 1 hour per participant), and DM-Wizard text messages to participants and to the RN. All live video feed, map, and robot pose data, as well as task-relevant images requested by participants, were recorded. Language data was manually time-aligned. After transcription and annotation, the corpus yielded 858 IUs.

4 Results

Each IU in the corpus corresponded to a unique TU from participant-robot dialogue. To better understand the structure and possible instruction variation over time, we focused analysis on IUs, their respective dialogue-moves, and motion command parameters. We analyzed IUs based on measures of word count, dialogue-move, and parameters on motion commands. We assessed possible parametric differences on motion commands by experiment phase (training phase, main phase 1, main phase 2). For significance testing, we used a mixed-effects ANOVA (computing standard least square regression using reduced maximum likelihood (Harville, 1977)), where phase (a repeated measure), age, gender, and scores on the spatial orientation and HRI trust surveys were included as factors in the model. Participant ID was included as a random effect.

4.1 Instruction Units

To gauge instruction frequency, we observed the mean number of IUs issued in an experiment session. On average, each participant issued 86 IUs (s.d. = 24.7, min = 58, max = 126). The average IU length was 8 words (s.d. = 5.7, min = 1, max = 60). Three participants each issued over 112 IUs in total, while three issued 70 or fewer.

4.2 Dialogue-Moves in IUs

We analyzed the selection of dialogue-moves that participants issued in their IUs. Participants often issued more than one dialogue-move per IU (mean = 1.6 dialogue-moves per IU, s.d. = 0.88, min = 1, max = 8). Unsurprisingly, the *command* dialogue-move was in the most IUs (94% of all IUs). See Table 1 for the entire distribution. We report on notable exceptions in Section 5.2.

The most common functions observed in the instructions were *command* dialogue-moves to send a new image, rotate, and drive. As reported in Table 1, over half of IUs include an image request, followed by rotate and drive commands.

Dialogue-Move	Instruction Units	
	N	%
Command		94
Send-Image	443	52
Rotate	406	47
Drive	358	42
Stop	29	3
Explore	7	1
Request-Info	34	4
Feedback	28	3
Parameter	14	2
Describe	5	1

Table 1: Dialogue-move distribution over all IUs in the corpus (N=858). An IU may have one or more dialogue-moves.

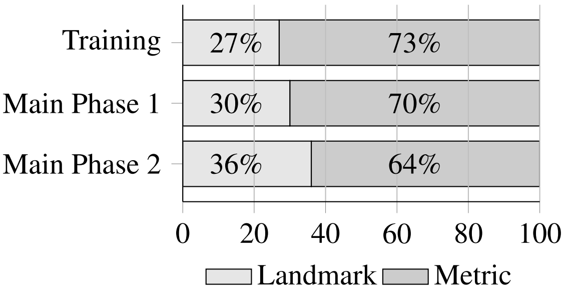


Figure 6: Proportions of landmark mentions to metric mentions within all *command* moves of subtype *drive* and *rotate* across experiment phases. There were 177, 333, and 316 occurrences of metric or landmark information in the training, main phase 1, and main phase 2 to compute proportions, respectively.

4.3 Parameters on Motion Commands

We delineate percentages of all IUs that involved motion requests for the robot (i.e., commands that were not image, stop, or exploration requests). 638 IUs contained a *drive* or *rotate* subtype request with a command parameter; 75% included metric units and 37% included landmarks (an IU could contain both). We tabulated all metric and landmark mentions in this IU subset.

We observed a substantial change in general participant strategy over time (Figure 6). In the training phase, participants began with a metric-dominant strategy that regressed in main phase 1, and further in main phase 2. The final phase experienced a 9% increase (absolute) in landmark references compared to the training phase, and a subsequent 9% decrease of metric references. A mixed-effects ANOVA test on the proportion of metric to landmark usage in commands found a main effect for phase (F[2, 627]=3.6, $p<0.05$). No other main effects were found. A Tukey HSD test found a significant difference between main phase 1 and 2 ($p<0.05$). We also tabulated instances of increased landmark usage by participant: six participants increased their proportion of landmark usage between main phase 1 and 2. Three used fewer landmarks in main phase 2, and one used the same proportion.

5 Discussion

This work seeks to elicit natural instruction-giving from participants, and to assess how communication strategies varied as people build an increasingly better mental model of a robot's understanding. Thus far, we've seen progress towards our goal in two main areas: (1) the experiment setup was workable; participants believed they were instructing an autonomous robot, and (2) we observed naturally occurring coordination efforts via changes in participant strategy over time. The latter area is discussed in more detail in the next subsection.

5.1 Metric vs. Landmark Usage

Our findings suggest possible changes in how participants perceived robot capabilities over time. This was highlighted by a significant decrease in metric usage between the two main experiment phases (main phase 1 and 2). This result suggests that participants became more comfortable in communicating with the robot through experience. Therefore, their communication styles become more "natural" and similar to human communication strategies, which tend to include landmark-based references (Clarke et al., 2015). This result has implications for language grounding and interpretation, in that developers should expect to handle both metric and landmark-based references.

We note that the dominant strategy overall was clearly the use of metric information. We identify several possible factors. One factor may be the participant interface: their situational awareness of the robot's environment is constrained to the most recent image of the robot's first-person perspective of a scene and the map displaying an occupancy grid of the surroundings. The indoor space is sparsely populated with objects, so a requested image might not return valuable visual information of an object of interest. The map, on the other hand, is visually salient and returns real-time information, including the presence of rooms, halls, and doorways. When landmark references are combined, the most frequent landmark used (139 out of 380 total landmark references) is "door," followed by "room," "hallway," and "wall." These are all landmarks recognizable in the map; other landmark types for navigation may be inhibited due to participant unawareness.

A second, somewhat related factor contributing to the use of metric references in general may be a misalignment of common ground between participant and robot, namely a lack of familiarity with the objects. Even when an object is returned in an image, the angle may not be conducive to object recognition. Participants are forced to either abandon the object as a landmark, or find another way of talking about it. For example, one participant describes a calendar hanging on the wall as *"the item on the right on the wall,"* while another describes a barrel as *"the round object."*

In addition, participants may be unsure if the robot can recognize an object by a given word. In training, participants were instructed that "the robot knows what some objects are but not all objects." They also know that the robot understands object features. Our intention was to encourage dialogue by making high-level search commands like *"find the shoe"* (a search-task target) outside the robot's capabilities. A side effect is that participants quickly became aware of this limitation, often as early as the training phase. When participants did try including a search-task target, without any additional descriptive information like color or shape, the DM-Wizard guidelines prompted for an alternative description. Some participants ignored the robot's request for a different description, and instead abandoned the landmark strategy in favor of metric instructions, which can be used in the absence of familiarity or knowledge of surrounding landmarks.

We note that the robot's surroundings are somewhat strange. They do not conform to canonical representations, disallowing use of lived experience of object expectations based on room type. Although an effort was made to group similar objects according to a room's possible function (e.g., kitchen items grouped together in one room and in a typical arrangement), the environment is sparsely filled with objects and is not in a finished state. These were practical limitations of laboratory resources, but in future work we plan to explore the effects of the environment further by varying it in a fully simulated version of the experiment.

5.2 Dialogue-Move Types

We found that most IUs contained *command* dialogue-moves, but with some exceptions. This was largely based on participants' assessment of robot capabilities. Two participants were responsible for 33 of the 34 occurrences of *request-info*. One participant issued requests like *"are you alone?"* and *"do you detect any threats?"* The other requested object identification, such as *"what's that object just to the left of the photo?"* This suggests an expectation for additional joint vision and language processing capabilities in these kinds of scenarios. *Feedback* dialogue-moves were largely experiment-specific start and end updates like *"I am ready."*

Our dialogue-move analysis of *commands* revealed a uniform strategy of consistent image requests shown in nearly half of all IUs. This is expected, as the bandwidth limitations of our experiment design prevented sending live video. More image requests are expected, but we found at least five phase runs where the robot "learned" to send images after receiving commands: occasionally the DM-Wizard would observe that a participant was requesting an image in every instruction, and as a result offered to remember to send images after each command.

6 Related Work

Our experiment setup and data collection effort resemble similar corpora, with some differences. The CReST (Eberhard et al., 2010), SCARE (Stoia et al., 2008), and GIVE (Gargett et al., 2010) corpora consist of search-and-navigation tasks, but are strictly human-human dialogue. We collected natural language interactions simulating fully autonomous dialogue processing, but without participant awareness that a human was simulating the robot responses. Participants assessed robot intelligence on their own when formulating instructions and follow-on responses.

6.1 Wizard-of-Oz Approach

By far the WOz method's most common use has been for handling natural language (Riek, 2012). Many studies use a wizard in automated dialogue system development (e.g., in virtual agent negotiation (Gandhe and Traum, 2007), time-offset storytelling (Artstein et al., 2015), and in-car personal assistants (Rieser and Lemon, 2008)).

Some researchers have considered a multi-wizard setup for multimodal interfaces. The SimSensei project (DeVault et al., 2014) used a two-wizard setup during the development stage; one controlling the virtual agent's verbal behaviors and another the non-verbal behaviors. Green et al. (2004) investigated using multiple wizards for dialogue processing and navigation capabilities for a robot in a home touring scenario, finding the multi-wizard approach effective when the robot and human were co-present.

6.2 Natural Language Interpretation

Traditional approaches to natural language interpretation for robots follow the methodology of *corpus-based robotics* (Bugmann et al., 2004), where some natural language, primarily route instructions, is collected. Route instruction interpreters dating back to MARCO (MacMahon, 2006), and more recently the robotic forklift (Tellex et al., 2011) and Tactical Behavior Specification grammar (Hemachandra et al., 2015; Boularias et al., 2016), rely on these initial route instructions to learn mappings to robot-executable procedures like path planning. Additionally, some use semantic parsers (e.g., (Chen and Mooney, 2011; Artzi and Zettlemoyer, 2013; Matuszek et al., 2013; Krishnamurthy and Kollar, 2013)) or translation (Matuszek et al., 2010) to map natural language to actions.

A gap in these works is bi-directional dialogue interaction, specifically cases where initial instructions are not well-formed and need additional clarification, or when participants grow to better grasp the robot's capabilities, varying instruction strategies over time. Our work collected instructions to a robot, but also included the dialogue and follow-on responses needed to establish or build common ground. This paper focused on analyzing initial robot-directed instructions, leaving analysis of responses during the dialogue to future work.

7 Summary and Ongoing Work

We presented a method for investigating changes in participant instruction strategies to a robot in a collaborative navigation task. We found an initial preference for metric information in motion commands, but this decreased over time as participants used more landmarks in their instructions.

We also note that the dataset under construction will provide value not only in the language collected, but also visual information. The accompanying images from the robot provide a unique resource with content that is both first-person and task-relevant for building situational awareness of a remote environment.

This work is a multi-stage effort to develop natural communication frameworks between humans and robots. In this work's next phase, automating language processing will begin, starting with language generation aspects. Rather than typing out full responses, wizards will use an interface to select responses following communicative guidelines. The Wizard-of-Oz interface allows template generation by filling in parameter values, if necessary. We expect a similar range of rich participant dialogue, but faster wizard response time, even for fairly complex strategies. Wizard selections will serve as training data for an automated dialogue manager.

Acknowledgments

We thank Susan G. Hill, Arthur William Evans, Stephanie Lukin, Su Lei, and Brendan Byrne for their contributions to this research program.

Appendix

A Robot Capabilities

These are, verbatim, the capabilities provided on a sheet to study participants:
"The robot can take a photo of what it sees when you ask. The robot has certain capabilities, but cannot perform these tasks on its own. The robot and you will act as a team.

Robot capabilities are:

- Robot listens to verbal instructions from you.
- Robot responds in this text box *(Experimenter points to instant messenger box on screen)* or by taking action
- Robot will avoid obstacles
- Robot can take photos directly in front of it when you give it a verbal instruction
- Robot will know what some objects are, but not all objects
- Robot also knows:
    - Intrinsic properties like color and size of objects in the environment
    - Proximity of objects like where objects are relative to itself and to other objects
    - A range of spatial terms like to the right of, in front of, cardinal directions like N, S
    - *History:* the Robot remembers places it has been
- Robot doesn't have arms and it cannot manipulate objects or interact with its environment except for moving throughout the environment
- Robot cannot go through closed doors and it cannot open doors, but it can go through doorways that are already open
- Robot can only see about knee height ($\sim$ 1.5 feet)."

B Dialogue-Move Annotation Set

Command Task-related instructions from the Commander to the robot are *command* dialogue-moves.

- *command:drive* Initiate/continue movement.
- *command:rotate* Initiate/continue a rotation.
- *command:explore* Explore an area via navigation using a target and/or direction as heading.
- *command:stop* End a drive or rotation.
- *command:send-image* Request an image.

Describe General statements from the Commander to the robot about a scene or plan are *describe* dialogue-moves.

- *describe:scene* Typically a description of what the Commander sees or thinks the robot should see.
- *describe:plan* Explication of the Commander's intention, not necessarily actionable.

Request-info *Request-info* dialogue-moves request information of the robot.

- *request-info:scene* Asking for information about what the robot sees, or confirmation for what the Commander thinks the robot should see.
- *request-info:map* Asking about robot's position or heading.
- *request-info:confirm* Confirm a proposed plan.

Feedback General domain-independent expressions from the Commander to the robot.

- *acknowledge* Acknowledgment of either a conversational move or an action (such as the sending of an image or map).
- *ready* Inform robot ready to do task.
- *yes* Simple positive response (yes).
- *no* Simple negative response (no).
- *standby* Inform robot to stand by or wait.

Standalone Instruction Content Provide further content for an existing instruction from the Commander to the robot.

- *direction* a heading (e.g., right, left)
- *distance* a unit of measure (e.g., feet, degrees)

References

Ron Artstein, Anton Leuski, Heather Maio, Tomer Mor-Barak, Carla Gordon, and David R Traum. 2015. How many utterances are needed to support time-offset interaction? In *Proc. of FLAIRS*.

Yoav Artzi and Luke Zettlemoyer. 2013. Weakly supervised learning of semantic parsers for mapping instructions to actions. *Transactions of the Association for Computational Linguistics* 1:49–62.

Abdeslam Boularias, Felix Duvallet, Jean Oh, and Anthony Stentz. 2016. Learning qualitative spatial relations for robotic navigation. In *Proceedings of IJCAI*.

Guido Bugmann, Ewan Klein, Stanislao Lauria, and Theocharis Kyriacou. 2004. Corpus-Based Robotics: A Route Instruction Example. In *Proc. of IAS-8*.

Jean Carletta, Stephen Isard, Gwyneth Doherty-Sneddon, Amy Isard, Jacqueline C Kowtko, and Anne H Anderson. 1997. The reliability of a dialogue structure coding scheme. *Computational linguistics* 23(1):13–31.

Joyce Y Chai, Lanbo She, Rui Fang, Spencer Ottarson, Cody Littley, Changsong Liu, and Kenneth Hanson. 2014. Collaborative effort towards common ground in situated human-robot dialogue. In *Proc. of HRI*.

David L Chen and Raymond J Mooney. 2011. Learning to Interpret Natural Language Navigation Instructions from Observations. In *Proc. of AAAI*.

Alasdair DF Clarke, Micha Elsner, and Hannah Rohde. 2015. Giving good directions: order of mention reflects visual salience. *Frontiers in psychology* 6:1793.

David DeVault, Ron Artstein, Grace Benn, Teresa Dey, Ed Fast, Alesia Gainer, Kallirroi Georgila, Jon Gratch, Arno Hartholt, Margaux Lhommet, et al. 2014. SimSensei Kiosk: A virtual human interviewer for healthcare decision support. In *Proc. of AAMAS*.

Kathleen M. Eberhard, Hannele Nicholson, Sandra Kübler, Susan Gundersen, and Matthias Scheutz. 2010. The Indiana "Cooperative Remote Search Task" (CReST) Corpus. In *Proc. of LREC*.

Sudeep Gandhe and David R Traum. 2007. Creating spoken dialogue characters from corpora without annotations. In *Proc. of Interspeech*.

Andrew Gargett, Konstantina Garoufi, Alexander Koller, and Kristina Striegnitz. 2010. The GIVE-2 Corpus of Giving Instructions in Virtual Environments. In *Proc. of LREC*.

Anders Green, Helge Huttenrauch, and Kerstin Severinson Eklundh. 2004. Applying the Wizard-of-Oz framework to cooperative service discovery and configuration. In *Proc. of ROMAN*.

Joy Paul Guilford and Wayne S Zimmerman. 1948. The Guilford-Zimmerman Aptitude Survey. *Journal of applied Psychology* 32(1):24.

David A. Harville. 1977. Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association* 72(358):320–338.

Sachithra Hemachandra, Felix Duvallet, Thomas M Howard, Nicholas Roy, Anthony Stentz, and Matthew R Walter. 2015. Learning models for following natural language directions in unknown environments. In *Proc. of ICRA*.

Jayant Krishnamurthy and Thomas Kollar. 2013. Jointly Learning to Parse and Perceive: Connecting Natural Language to the Physical World. *Transactions of the Association for Computational Linguistics* 1:193–206.

Matt MacMahon. 2006. Walk the Talk: Connecting language, knowledge, and action in route instructions. In *Proc. of AAAI*.

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 *Proc. of RO-MAN*.

Matthew Marge and Alexander I. Rudnicky. 2010. Comparing spoken language route instructions for robots across environment representations. In *Proc. of SIGdial*.

Cynthia Matuszek, Dieter Fox, and Karl Koscher. 2010. Following directions using statistical machine translation. In *Proc. of HRI*.

Cynthia Matuszek, Evan Herbst, Luke Zettlemoyer, and Dieter Fox. 2013. Learning to parse natural language commands to a robot control system. In *Experimental Robotics*.

Rebecca Passonneau. 2006. Measuring Agreement on Set-valued Items (MASI) for Semantic and Pragmatic Annotation. In *Proc. of LREC*.

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

Verena Rieser and Oliver Lemon. 2008. Learning Effective Multimodal Dialogue Strategies from Wizard-of-Oz Data: Bootstrapping and Evaluation. In *Proc. of ACL*.

Antonio Roque, Anton Leuski, Vivek Rangarajan, Susan Robinson, Ashish Vaswani, Shrikanth Narayanan, and David Traum. 2006. Radiobot-CFF: A Spoken Dialogue System for Military Training. In *Proc. of Interspeech*.

Kristin E Schaefer. 2013. *The perception and measurement of human-robot trust*. Ph.D. thesis, University of Central Florida.

Laura Stoia, Darla Magdalena Shockley, Donna K. Byron, and Eric Fosler-Lussier. 2008. SCARE: A Situated Corpus with Annotated Referring Expressions. In *Proc. of LREC*.

Stefanie A. Tellex, Thomas F. Kollar, Steven R. Dickerson, Matthew R. Walter, Ashis Banerjee, Seth Teller, and Nicholas Roy. 2011. Understanding natural language commands for robotic navigation and mobile manipulation. In *Proc. of AAAI*.

Tom Williams, Gordon Briggs, Bradley Oosterveld, and Matthias Scheutz. 2015. Going Beyond Literal Command-Based Instructions: Extending Robotic Natural Language Interaction Capabilities. In *Proc. of AAAI*.