SCOUT: A Situated and Multi-Modal Human-Robot Dialogue Corpus

Stephanie M. Lukin[1], Claire Bonial[1], Matthew Marge[2], Taylor Hudson[3], Cory J. Hayes[1], Kimberly A. Pollard[1], Anthony Baker[1], Ashley N. Foots[1], Ron Artstein[4], Felix Gervits[1], Mitchell Abrams[1], Cassidy Henry[1], Lucia Donatelli[5], Anton Leuski[4], Susan G. Hill[1], David Traum[4] and Clare R. Voss[1]

[1]DEVCOM Army Research Laboratory, [2]DARPA, [3]Oak Ridge Associated Universities,
[4]USC Institute for Creative Technologies, [5]Vrije Universiteit
stephanie.m.lukin.civ@army.mil

Abstract

We introduce the Situated Corpus Of Understanding Transactions (SCOUT), a multi-modal collection of human-robot dialogue in the task domain of collaborative exploration. The corpus was constructed from multiple Wizard-of-Oz experiments where human participants gave verbal instructions to a remotely-located robot to move and gather information about its surroundings. SCOUT contains 89,056 utterances and 310,095 words from 278 dialogues averaging 320 utterances per dialogue. The dialogues are aligned with the multi-modal data streams available during the experiments: 5,785 images and 30 maps. The corpus has been annotated with Abstract Meaning Representation and Dialogue-AMR to identify the speaker's intent and meaning within an utterance, and with Transactional Units and Relations to track relationships between utterances to reveal patterns of the Dialogue Structure. We describe how the corpus and its annotations have been used to develop autonomous human-robot systems and enable research in open questions of how humans speak to robots. We release this corpus to accelerate progress in autonomous, situated, human-robot dialogue, especially in the context of navigation tasks where details about the environment need to be discovered.

Keywords: human-robot interaction, corpus creation, situated dialogue, multi-modal, linguistic annotations

1. Introduction

For robots to team effectively with humans, a critical capability will be to use forms of natural communication like language. Moreover, these interactions must be bi-directional, as robots will need to provide status updates and ask for and receive clarification or help from teammates in challenging situations. Finally, the interactions must be situated in knowledge about the environment that the robot inhabits. In order to study these forms of communication and accelerate progress in the development of autonomous robot dialogue systems, collections of data should (1) focus on unconstrained, *robot-directed* dialogue in contrast to traditional human-human dialogue,[1] (2) exhibit natural diversity of communication strategies inherent in situated dialogue, and (3) be organized into a format that can be quickly labeled and used for training an autonomous dialogue system.

In this paper, we present **SCOUT, the Situated Corpus Of Understanding Transactions**, a multi-modal, human-robot dialogue corpus that meets these data criteria. SCOUT is a collection of human-robot dialogues within the task domain of collaborative navigation between a human and a remotely-located robot. Human participants assumed the role of *Commander* and collaborated with a remotely-located robot to explore and assess the robot's environment and locate objects of interest. To progress through the task, Commanders relied on a combination of overhead maps generated from streaming LIDAR (LIght Detection And Ranging) sensors, pictures from the robot's camera upon request, and text messages. Commanders spoke freely to the robot and were given no restrictions on how they formulated their language, allowing them to follow their natural tendencies for speaking to a robot when completing this task. The dialogues were collected in a Wizard-of-Oz (WoZ) experimental paradigm, wherein the robot's autonomy was controlled by two "wizard" experimenters: a *Dialogue Manager* to ground instructions to the robot's surroundings and select dialogue behaviors, and a *Robot Navigator* to move the robot and provide status updates (see Fig. 1).

Our construction of SCOUT brings together for the first time all the data streams and modalities collected in this task domain. SCOUT contains 89,056 utterances and 310,095 words from 278 dialogues lasting about 20-minutes each with an

[1] Humans instruct robots differently compared to instructing other humans (Mavridis, 2015; Marge et al., 2020, 2022).

LREC-COLING 2024, pages 14445–14458
20-25 May, 2024.

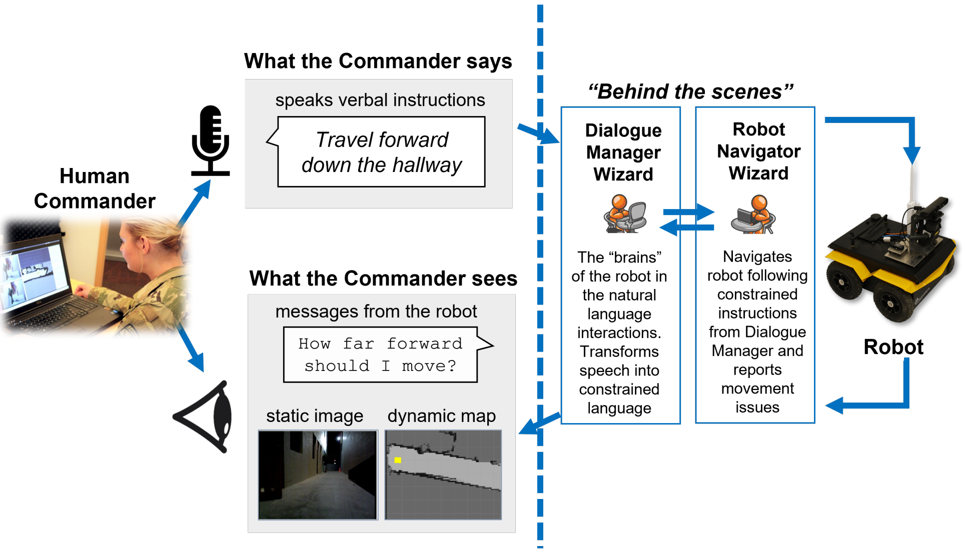

What the Commander says
speaks verbal instructions
Travel forward down the hallway
Human Commander
What the Commander sees
messages from the robot
How far forward should I move?
static image
dynamic map
"Behind the scenes"
Dialogue Manager Wizard
The "brains" of the robot in the natural language interactions. Transforms speech into constrained language
Robot Navigator Wizard
Navigates robot following constrained instructions from Dialogue Manager and reports movement issues
Robot

Figure 1: Wizard-of-Oz data collection design

average of 320 utterances per dialogue. 5,785 images were taken across the corpus and are linked to the moment they were taken in the dialogue. Thirty LIDAR maps are included with annotations that reflect the spaces the robot scanned in the environment during the dialogue (some of the maps include only subsets of the whole area).

The dialogues exhibit a situated and multi-modal nature with linguistic references to space, distance, and the physical world, providing for many opportunities to study dialogue dynamics. We apply four existing linguistic annotations to subsets of SCOUT to advance understanding of phenomena within human-robot dialogue. To this end, we annotate SCOUT with (1) Abstract Meaning Representation (AMR) (Banarescu et al., 2013) and (2) Dialogue-AMR (Bonial et al., 2019b) to assess interlocutor intents and meaning within an utterance. To track relationships between utterances, we annotate both (3) Transactional Units (TUs) (Carletta et al., 1996) and (4) the Relations between utterances within a TU; together these reveal patterns of the Dialogue Structure (Traum et al., 2018). We describe how these annotations on top of SCOUT have been subsequently used in action selection for automated robot navigation systems and automated dialogue systems along with other areas of human-robot analysis.

This paper offers the following contributions:

- A fully-compiled, novel corpus of multi-modal, robot-directed dialogue in a collaborative exploration task involving a remotely-located physical or simulated robot (Sec. 3 and 4).
- Annotations of SCOUT data using existing frameworks of AMR, Dialogue-AMR, and Dialogue Structure TUs and Relations (Sec. 5).
- Applications of SCOUT data and annotations, including the development of systems and analyses of the language and behaviors observed in the human-robot dialogue (Sec. 6).

Although portions of the dialogues have previously been shared via private data-sharing agreements to enable the annotations and applications described here, this paper presents the comprehensive compilation and curation process of the dialogues and multi-modal streams for public release under a Creative Commons Zero 1.0 Universal (CC0 1.0) license. We believe the corpus and its annotation will serve to immediately benefit the Human-Robot Interaction (HRI), Dialogue, and broader Robotics communities that aspire to use language as a way to interact with robots. SCOUT is available at `https://github.com/USArmyResearchLab/ARL-SCOUT`

2. Experiments and Data Collection

The experimental domain involved a collaborative human-robot exploration task in a low-bandwidth environment (Marge et al., 2016). The robot (a Clearpath Jackal with functionality implemented in ROS, the Robot Operating System (Koubâa, 2017)) entered an unexplored area and received instructions from a remotely-located human teammate. This human teammate (called ”Commander”) was given specific goals for the exploration, such as locating and counting doors and specific objects of interest, e.g., doorways, shovels, shoes. The Commander could not directly teleoperate the robot and had to provide verbal instructions to accomplish tasks with the robot (e.g., “Move forward five feet”, “Proceed through the doorway in front of you”). The Commander’s knowledge of the environment was based upon (1) a dynamic

	Experiment 1	**Experiment 2**	**Experiment 3**	**Experiment 4**
Dialogue Processing	WoZ + keyboard	WoZ + button GUI	WoZ + button GUI	ASR + auto-DM
Robot Behaviors	WoZ + joystick	WoZ + joystick	WoZ + joystick	WoZ + joystick
Robot & Environment	physical	physical	virtual	virtual

Table 1: Human-Robot Dialogue Experimentation. ASR: Automatic Speech Recognition

LIDAR map of the area built up in real time as the robot moved, (2) snapshot pictures from the robot's front-facing RGB camera, taken upon Commander request, and (3) text messages from the robot. The left-hand side of Fig. 1 depicts what the Commander saw and could do during the interaction while seated at their workstation.

The robot was controlled using a Wizard of Oz methodology to facilitate a data-driven understanding of how people talk to robots (Riek, 2012; DeVault et al., 2014). The experiments employed two Wizards; their division of labor is depicted on the right-hand side of Fig. 1. The Dialogue Manager Wizard (DM-Wizard) listened to the Commander's instructions and decided how the robot should proceed in the dialogue. Status updates and clarifications were sent to the Commander from the DM-Wizard in a chat window. When the DM-Wizard determined that instructions were executable in the current context, the instructions were passed in a constrained form to the Robot Navigator Wizard (RN-Wizard), who used a joystick to teleoperate the robot and who provided information of failures through speech or a button click back to the DM-Wizard as the task was being completed. The DM-Wizard passed status updates from the RN-Wizard to the Commander.

Four experiments varied the modes of dialogue processing and how the robot and environment in the experiments were represented (Table 1). **Experiment 1** had a DM-Wizard manually type messages to interact with the Commander in real-time, and a RN-Wizard control a physical robot in environments that resembled an alleyway or an indoor space of a house-like environment under construction. The house contained a variety of hallways, rooms, and objects that gave themes to different spaces, for example, a kitchen area, conference room, and office.

Experiment 2 automated the DM-Wizard's command handling and response generation with a click-button graphical user interface (GUI). The collection of messages uncovered from Experiment 1 was incorporated into the GUI for use by the DM-Wizard, substantially reducing typing and composition effort by the DM-Wizard while increasing response uniformity. The GUI design built in functionality to avoid inflexible situations by using open slots where the DM-Wizard could type in a value for well-defined templates, e.g., "I see a door on the left," "I see a door on the right," "I see a wall," etc. There were no entirely open response buttons; all buttons reflect, at a minimum, an observed template of responses like "I see ___". (Bonial et al., 2017). Experiment 2 took place in the same physical environments as Experiment 1.

Experiment 3 utilized the same DM-Wizard GUI and moved from a physical robot and environment, to a simulated one designed to be a 1–1 replica of the physical environments, including the objects and their placement. Gazebo, a high fidelity 3D simulator, was used to construct the environment and complete the experiments (Koenig and Howard, 2004). The simulated robot was programmed with the same capabilities as the physical one using ROS. From the Commander's perspective, the study was equivalent to the previous two, with the exception that the images from the camera were virtually rendered.

Finally, **Experiment 4** deployed a completely automated dialogue system trained on data collected from the prior experiments. Instead of a DM-Wizard listening to the Commander's speech and routing messages back and to the RN-Wizard, the dialogue system provided these capabilities. This auto-DM system was divided into the following components: (1) an automated speech recognition (ASR) component that would transcribe speech in real time from the Commander, (2) a dialogue manager including a classifier that would determine the Commander's intent from their speech using training data collected in the previous experiments, and determine whether to (3) translate instructions to the RN-Wizard, and/or (4) provide replies to or request clarification from the Commander. This experiment had only the RN-Wizard as a wizard experimenter to tele-operate the robot in response to instructions provided to it by the auto-DM system. The auto-DM system was trained on Experiment 1 and 2 data, and tested on a subset of Experiment 3 data. Several training methods were employed, and accuracy ranged from 61% - 75%. Over half of the incorrect responses would still be appropriate and advance the dialogue, as they were considered felicitous (appropriate responses that would have the same effect as the correct response) or approximate (responses that differed only slightly from the correct one, e.g., variation in turn radius or movement distance) (Gervits et al., 2019).

3. SCOUT Construction

In this section, we describe our methodology to isolate, extract, and verify the various data streams from the experimentation, and our process to compile the information into the human-readable and machine-processable formats that comprise SCOUT.

Commander data consists of the Commanders' verbal interactions with the robot. These were recorded in Mumble[2] using a push-to-talk button. Annotators were trained to manually transcribe the Commander speech from Experiments 1–3 using the Praat software (Boersma and Weenink, 1992–2019). Turns were segmented by silences, and then further by intents. For example, the turn "go to the map and take a picture" would be segmented into the utterances "go to the map" and "and take a picture" (Bonial et al., 2019a). The speech data from Experiment 4 was transcribed during the experiment by Google ASR or Kaldi. Key tokens were automatically normalized, for example, converting "five ft" to "5 feet" for consistency with the manual transcription coding schema in Experiments 1–3. Post-experimentation, manual verification was conducted to correct ASR results, for example "turn right for you to grieve" was corrected to "turn right 45 degrees" after listening to the utterance. The timestamp at which the utterance began (which was not necessarily the same as when the push-to-talk button was depressed) was saved in metadata by Mumble or the ASR for future alignment.

Dialogue Manager data are text messages sent by the DM to either the Commander or the RN-Wizard. Text messages were recorded as a modified `sensor_msgs/StringStamped` ROS topic in a ROS bag file—a file format also used to save diverse timestamped sensor data from the robot. Text messages to the Commander and to the RN-Wizard were differentiated from each other in the ROS bag file. The timestamps were extracted along with the utterances for alignment.

Robot Navigator data consists of the RN-Wizard's verbal or text messages to the DM. In Experiments 1–3, RN communications were spoken and transcribed following the same process using Praat, saving the timestamps from the metadata. In Experiment 4, communications were text messages the RN triggered by a button press on the joystick coded as the `sensor_msgs/Joy` ROS topic, or on the GUI coded as the `sensor_msgs/StringStamped` ROS topic. Text messages and their timestamps were extracted from the ROS bag file for alignment.

The transcribed speech and text utterances from all interlocutors were time-aligned into communication floors that reflect the passing of information during the dialogue as it occurred. The streams are shown in Table 2. Given that the DM served as an intermediary directing communications between the Commander and RN, the dialogue took place across two non-mutual conversational floors: the Left and Right floors. The Left floor includes communication between the Commander and what the Commander thinks of as "the robot," (really the DM, acting as front end), and contains streams "CMD" and "DM→CMD." The Right floor was between the DM and RN, and contains streams "DM→RN" and "RN." The DM would convey information across floors, as shown in lines 230-234 (Left to Right) and 238 (Right to Left). Timestamps were coded either as seconds since the dialogue started (Experiments 1 and 2) or with a unix timestamp (Experiments 3 and 4). Complications arose in synchronizing the timestamps across recordings from the three different machines used to run the experiments (one each for Commander, DM, and RN), and required scripts and manual verification to ensure that each utterance was inserted into the correct location in the transcript across the conversational floors.

The resultant *time-aligned transcripts* were compiled into .xlsx spreadsheets in the format of Table 2 (see Fig. 4 in Appendix for a screenshot), as well as a compacted tab delimited format to facilitate different methods of file-processing:

```
ID  time     stream    text
...
222 1054.31 CMD       "robot proceed
                      through the doorway"
223 1061.9  CMD       "turn a hundred and
                      eighty degrees to
                      the right"
224 1063.78 CMD       "and take a picture"
225 1070.54 DM->CMD "processing. . ."
...
```

In Experiment 4, the .xlsx spreadsheets contain additional columns for the raw ASR results and intermediary normalized forms, in addition to the final corrected utterance (see Fig. 5 in Appendix). The Commander text in these tab delimited formats is the corrected utterance.

Images were taken by the RN and recorded as `sensor_msgs/Image` ROS topics in ROS bag files. After extraction, each image per dialogue was given a unique id and inserted into a modified .xlsx spreadsheet at the moment the RN-Wizard had the robot take the image (see Fig. 6 in Appendix.) In the spreadsheet, clicking or control-clicking on the image name will open the .jpg image in the computer's default picture viewer program. This information is also available in a modified tab delimited format, where the stream value is "IMAGE" and the text value is the image filename:

[2] https://www.mumble.info/

		Left Conversational Floor		Right Conversational Floor	
#ID	**Timestamp**	**CMD**	**DM → CMD**	**DM → RN**	**RN**
222	1054.31	robot proceed through the doorway			
223	1061.9	turn a hundred and eighty degrees to the right			
224	1063.78	and take a picture			
225	1070.54		processing. . .		
226	1077.99		I see more than one doorway.		
227	1079.46		The one to my left?		
228	1081.34	the doorway to your left			
229	1085.72		processing. . .		
230	1103.95			move to Foyer - Kitchen doorway	
231	1107.81			then. . .	
232	1109.87			turn 180	
233	1111.06			then. . .	
234	1112.19			send image	
235	1114.76		moving. . .		
236	1121.65		turning. . .		
237	1134.48				done and sent
238	1135.29		done, sent		

Table 2: Navigation instruction initiated by the Commander (#222-244), clarification (#225-229), translation to a simplified form by Dialogue Manager (DM) to Robot Navigator (RN) (#230-234), status updates (#235-236), completion by the RN (#237), and notification of task completion to Commander (#238).

```
ID    time     stream  text
...
234  1112.19  DM->RN  "send image"
235  1114.76  DM->CMD "moving. . ."
236  1121.65  DM->CMD "turning. . ."
i019 1133.96  IMAGE   "frame019"
237  1134.48  DM->RN  "done and sent"
...
```

LIDAR maps were extracted from what the LIDAR had scanned by the end of each dialogue (Fig. 2a). These are snapshots from the last frame of the dialogue as captured in the screen recording of the Commander's monitor. Due to how the LIDAR visualization functionality was implemented, some snapshots show a subset of the environment rather than a complete view. One .png file was extracted for each dialogue in Experiment 1, with future plans to streamline the process and compile maps for the remaining dialogues.

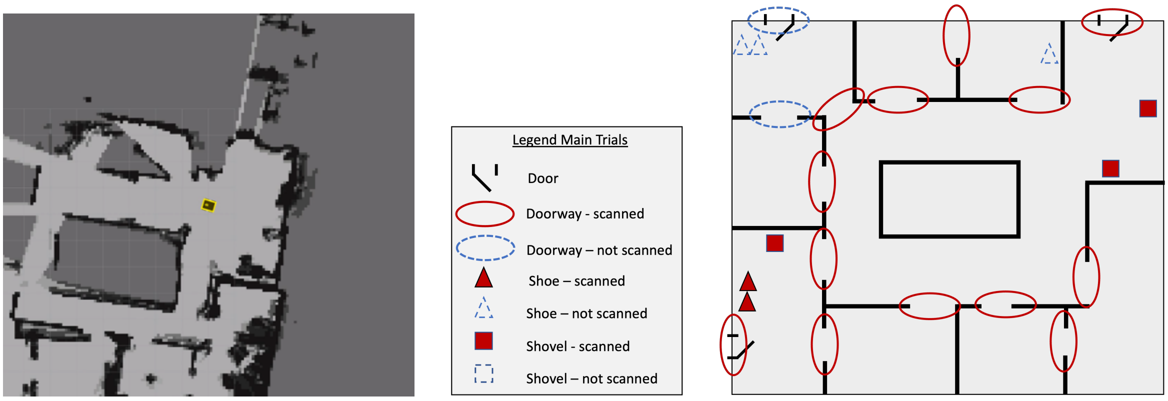
Rendered floor plans of the complete environment were manually created for each of the dialogues with LIDAR maps, showing the objects of interest scanned by the LIDAR (Fig. 2c, with legend in Fig. 2b). These determinations were made by manually comparing the LIDAR to the rendered floor plan. Dark gray spaces indicated the LIDAR had not scanned the area and informed the determinations. A text version of this annotated floor plan is also available, where each target entity (e.g., doorway, shovel, etc.) was mapped to a unique identifier (e.g., "door1") and denoted as scanned or not scanned:

```
door1    not-scanned
door2    scanned
...
```

4. SCOUT Statistics

The corpus contains data from 93 Commanders, where each Commander completed three dialogues—one training exercise in the alleyway and two main exercises in the house, starting at different locations with different objects of interest to count—for a total of 278 dialogues containing, on average, 320 utterances.[3] The corpus contains 89,056 utterances total, 310,095 words, and 5,785 images (Table 3). On average, Commanders requested 20.8 images per dialogue which they could use to assess the environment in their search and counting tasks. On a per individual and per task basis (training or main), the requests

[3] One Commander only completed the training and one main, thus SCOUT has 278 and not 279 dialogues.



(a) LIDAR map at the end of a dialogue. Dark gray is unscanned by LIDAR

(b) Legend for Fig. 2c

(c) Corresponding top-down floor plan of Fig 2a showing items' scanned status

Figure 2: One LIDAR map and annotated floor plan with items scanned or not scanned by the LIDAR marked. Floor plan and legend were not shown during the exercise.

Corpus Attribute	SCOUT Total
Commanders	93
Dialogues	278
Utterances	89,056
Words – All	310,095
Words – Unique	89,056
Images	5,785
Avg. Utterances per Dialogue	320
Avg. Images per Dialogue	20.8
Standard-AMR Sentences	569
Dialogue-AMR Sentences	569
Dialogue Structure TUs	13,663
Dialogue Structure Relations	69,430

Table 3: SCOUT corpus summary. Corpus attributes per experiment in Table 5 in Appendix.

ranged from 3–88 images. These statistics per experiment are given in Table 5 in the Appendix.

Table 4 shows the breakdown of corpus attributes by experiment and interlocutor per conversational floor. We also tabulated a subset of 30 randomly selected dialogues from Experiment 3 in order to show a more fair comparison across experiments due to the difference in participant pool size (shown in parenthesis next to the full counts in the Experiment 3 column.)

The Commanders spoke a total of 25,386 utterances (2,446 unique words) and the RN-Wizards spoke 13,805 utterances (526 unique words) as shown in "Total" column in Table 4. The difference in counts is likely due to their roles in the experiment. The primary vocabulary of the RN-Wizard comes from acknowledgements of DM-wizard requests and reports of failures, whereas the Commander vocabulary reflects their attempts to take initiative and issue the requests. As a result, we observe the Commander vocabulary is greater and more varied than the RN-Wizard.

The Dialogue Manager sent a total of 31,959 text messages (813 unique words) to the Commander (DM→CMD), and sent 17,906 text messages (537 unique words) to the RN (DM→RN). We observe that the variety in vocabulary drops from Experiment 1 to Experiment 2, which likely reflects the introduction of the GUI ("DM→CMD Words – Unique" 565 to 311; and "DM→RN Words – Unique" 349 to 141 in Table 4). Aspects of the DM-Wizard's dialogue processing were compared in Experiments 1 and 2 (i.e., keyboard vs. button GUI) for assessing the Commander's ability to work with the DM-Wizard to issue well-formed and executable instructions. In combination with the dialogue structure analysis in Section 5, we found that, compared to Experiment 1, more instructions were issued in Experiment 2 within the same 20 minute trial limit, and more instructions were successful, showing improvement in the speed of the interaction (Marge et al., 2018). The corpus attributes remain reasonably consistent within Experiment 1 and 2 values for the Experiment 3 Subset.

In Experiment 4, we observe a significant increase in DM→CMD frequency of words from the prior experiments ("DM→CMD Words – All" 5,550–10,923 in Exps 1–3Subset, up to 24,253 in Exp 4), perhaps due to the introduction of the auto-DM, while the Commander variety of words decreases ("CMD Words – Unique" 661–738 in Exps 1–3Subset, down to 320 in Exp 4). We suspect the rise in frequency signifies an increase in miscommunication with the auto-DM, and that the lack of Commander vocabulary variety, while still maintaining the same level of word frequency, is due to an attempt to revert to more simplified terms, or hesitation to 'try out' different ways of giving instructions due to the auto-DM's limitations; the word error rate in Experiment 4 was 25%.

Corpus Attribute		Exp. 1	Exp. 2	Exp. 3 (Subset)	Exp. 4	Total
	Dialogues	30	30	188 (30)	30	278
CMD	Utterances	1,819	2,161	18,206 (2,545)	3,200	25,386
	Words – All	9,883	10,923	85,549 (11,453)	10,633	116,988
	Words – Unique	738	661	2,078 (675)	320	2,446
	Avg. Words per Utterance	5.43	5.05	4.70 (4.50)	3.32	4.61
DM→CMD	Utterances	1,779	3,370	20,595 (2,987)	6,215	31,959
	Words – All	5,550	10,923	65,889 (10,485)	24,253	106,615
	Words – Unique	565	311	418 (326)	335	813
	Avg. Words per Utterance	3.12	3.24	3.20 (3.51)	3.90	3.34
DM→RN	Utterances	1,417	1,766	12,622 (1,688)	2,061	17,906
	Words – All	5,139	5,588	41,038 (5,568)	7,433	59,196
	Words – Unique	349	141	289 (158)	51	537
	Avg. Words per Utterance	3.63	3.16	3.25 (3.30)	3.61	3.31
RN	Utterance	1,082	1,124	8,436 (1,042)	3,163	13,805
	Words – All	2,246	1,647	14,528 (1,935)	8,875	27,296
	Words – Unique	253	39	349 (93)	105	526
	Avg. Words per Utterance	2.08	1.47	1.72 (1.86)	2.81	1.98

Table 4: SCOUT corpus statistics by experiment and interlocutor and conversational floor

5. Annotations on SCOUT

With SCOUT fully assembled, we applied existing linguistic annotation schemas in order to better understand how humans worked with the robot, namely through analyzing the form and content of their instructions. We include in SCOUT’s release Abstract Meaning Representation (AMR), Dialogue-AMR, and Dialogue Structure TU and Relation annotations (quantities shown in Table 3). Taken together, these make different levels of conversational patterns accessible to automated systems—propositional semantics of an utterance (AMR), the illocutionary force (Dialogue-AMR), the meso-level intentional structure of a set of utterances (Dialogue Structure TUs), and finally the individual relations of each subsequent utterance within a TU to an antecedent utterance (Dialogue Structure Relations)[4].

5.1. Standard-AMR and Dialogue-AMR Annotation

To distill a robot’s behavior primitives and their parameters from totally unconstrained natural language, we apply AMR, a formalism for sentence semantics that abstracts away many syntactic idiosyncrasies and represents sentences with directed, acyclic graphs (Banarescu et al., 2013). Below is the utterance “take a photo” represented in Standard-AMR form in PENMAN representation (Penman Natural Language Group, 1989); note that AMR abstracts away from the semantically light verb “take”, instead representing photographing semantics:

```
(p / photograph-01
        :ARG0 (y / you)
        :mode imperative)
```

Dialogue-AMR is an enhanced AMR representing not only the content of an utterance, but the illocutionary force behind it, as well as tense, aspect, and completability (Bonial et al., 2019b)—all aspects of meaning that are necessary for the robot to interpret and act upon the natural language instructions. The same utterance as above is represented in Dialogue-AMR form using a domain-specific action frame `send-image-99` which represents a robot’s photographing behavior:

```
(c / command-SA
        :ARG0 (c2 / commander)
        :ARG1 (s / send-image-99
            :ARG0 (r / robot)
            :ARG1 (i / in-front-of
                :op1 r)
            :ARG2 c2
            :completable +
            :time (a / after
                :op1 (n / now)))
        :ARG2 r)
```

We annotated subsets of utterances from Experiments 1 and 2 with Standard-AMR and Dialogue-AMR, and an entire dialogue from Experiment 3. These annotations are available in parallel .txt files where each annotated utterance is given a unique id corresponding to the dialogue and sentence ID from the .xlsx corpus files.

5.2. Dialogue Structure Annotation

To understand and make explicit the patterns of multi-floor conversation, we applied Dialogue Structure annotation to capture the relationships

[4] See references for details of annotation schemas.

```
ID  time    stream   text                                   TU  ant.  relation
...
222 1054.31 CMD      "robot proceed through the doorway"    28  None  None
223 1061.9  CMD      "turn a hundred and eighty degrees     28  222   continue
                     to the right"
224 1063.78 CMD      "and take a picture"                   28  223   continue
225 1070.54 DM->CMD  "processing. . ."                      28  224*  processing
226 1077.99 DM->CMD  "I see more than one doorway."         28  222   missing-info
227 1079.46 DM->CMD  "The one to my left?"                  28  222   req-clar
228 1081.34 CMD      "the doorway to your left"             28  227   clar-repair
...
```

Figure 3: Tab delimited format for Dialogue Structure. Transaction Unit (TU), antecedent (ant).

between utterances within and across the conversational floors. Each aligned transcript was divided into Transaction Units (TUs) defined in Traum et al. (2018) as clusters of utterances across conversational floors that sequentially work towards fulfilling the original speaker's intent. A TU may encompass multiple dialogue utterances, spanning multiple speaker turns, including requests for clarification and subsequent repairs, confirmations and various types of acknowledgments that instructions were heard, understood, and complied with. Each utterance was further annotated with the *Relation* and *Antecedent*—the ID of the most immediate direct relation between this utterance and a prior utterance (Traum et al., 2018). Any contextual information required for understanding the annotation was denoted.

Every dialogue in SCOUT was annotated with this formalism and recorded in new .xlsx spreadsheets (Fig. 7 in Appendix) and converted into the tab delimited format in Fig. 3. In this example, #222 is the start of a new TU and assigned no relation. The instruction is continued by the same interlocutor into #223 and #224 through the *continue* relation. In #226 the DM informs the Commander that information is missing (*missing-info*) for successful execution of the instruction, and thus their request for a clarification (*req-clar*), to which the Commander provides the appropriate repair (*clar-repair*) in #228.

6. Applications

SCOUT and its annotations provide for a variety of analyses in support of Robotics research, especially within the HRI and Dialogue communities. We briefly describe research directions making use of SCOUT for system development, and how the data have encouraged discovery of new questions on how humans speak to robots.

6.1. Systems Developed from SCOUT Annotations

The SCOUT annotations have been used to train and deploy fully autonomous dialogue and navigation prototypes. The *ScoutBot* system utilizes the Dialogue Structure annotations from Experiments 1–3 to enhance the auto-DM developed in Experiment 4, and further implement autonomous robot navigation through ROS twist messages that map to user intents in a simulated building (Lukin et al., 2018a; Gervits et al., 2019). The *MultiBot* system extends this auto-DM pipeline to a simulated urban outdoor environment with multiple robots, further integrating heuristics for goal-based navigation instructions (Marge et al., 2019). The AMR and Dialogue-AMR annotations have been utilized for designing a classifier for intents, and integrating with a Clearpath Husky Unmanned Ground Vehicle in the real-world in robot-directed navigation (Bonial et al., 2023). The SCOUT corpus was also used to train a dialogue structure parser (Kawano et al., 2023).

6.2. Analyses of Human-Robot Multi-Modal Communication

The diversity of Commanders and the open-ended nature of the communication gives rise to many questions about Commander instruction-giving and navigation preferences. Researchers have asked how Commanders' interactions with respect to time and trust affect their instruction-style, and found an increase in landmark instructions (e.g., "move to the door in front of you") over metric information (e.g., "move forward five feet") (Marge et al., 2017) as well as more verbose and compound instructions over time and with increasing trust (Lukin et al., 2018b). Moolchandani et al. (2018) used the navigation patterns observed in SCOUT to discover that humans prefer when the robot demonstrates a sense of self-safety and awareness of its environment. The multi-modal nature of the exchanges has been explored, finding a relationship between the success of item-counting and exploration and quantity of images taken (Lukin et al., 2023). The corpus has also allowed for study of different types of capabilities robots should have to conduct natural dialogue-based interactions with humans (Pollard et al., 2018), and the exchanges between the Commander and Dialogue Manager have been evaluated for various linguistic modalities (i.e., presence of modal expressions, negation, and quantifiers) (Donatelli et al., 2020).

The images of SCOUT represent an opportunity to develop new computer vision and language techniques. A subset of these low-resolution and dim images with atypical angles have been used in computational visual storytelling to represent diverse environments and presentation of imagery found to be lacking in other collections of visual storytelling data (Lukin et al., 2018a). In analyzing a human-authored story collection utilizing SCOUT images, Halperin and Lukin (2023) observed narrative biases with respect to the cultural and linguistic biases associated with what human-authors recognized in the images.

7. Related Work

There are many human-human situated corpora that exhibit the Director-Follower paradigm SCOUT follows. These corpora have been used to study referring expressions (Stoia et al., 2008; Liu et al., 2016; Hu et al., 2016), speaker intents (Narayan-Chen et al., 2019; Bonial et al., 2021), structure (Eberhard et al., 2010), and to develop autonomous robot systems for following directions (MacMahon et al., 2006; Chen and Mooney, 2011; De Vries et al., 2018; Suhr et al., 2019; Chen et al., 2019; Thomason et al., 2020; Padmakumar et al., 2021; Gervits et al., 2021). Other situated paradigms leverage knowledge about object affordances within the world, e.g., a cup observed to be on its side may roll, and in combination with human gestures, inform a robot's reasoning and actions within simulated spaces (Pustejovsky and Krishnaswamy, 2020). However, prior work has shown that the way humans instruct robots is different from how they instruct other humans (Mavridis, 2015; Marge et al., 2020, 2022). There is a critical need for corpora like SCOUT that capture the human-robot dynamic in a coordinated Director-Follower task.

The Multi-Woz corpus (Budzianowski et al., 2019) is the largest of several corpora (including Eric and Manning (2017)) collected in a WoZ crowd-sourcing paradigm proposed in Wen et al. (2016). Here, each crowd-worker acted as either a wizard or a user and supplied only a single turn after observing previous turns. While this paradigm allows for scaling to a larger training corpus, it is unclear that this turn-by-turn addition to the dialogue in text via an online portal can reveal naturalistic dialogue patterns or individual communication style differences. Therefore SCOUT presents a unique resource for studying multi-modal and situated dialogue within a more natural interaction modeled between a human and robot.

Recent zero-shot approaches using large language models show the ability to process robot-directed instructions and generate an executable plan without needing a corpus or annotations (e.g., Brohan et al. (2023)). Yet because these models are not trained on domain experience, they cannot afford the same rich semantic and structural knowledge supplied by SCOUT data and annotations.

8. Conclusion

SCOUT meets the characteristics outlined at the start of this paper for datasets to study human-robot dialogue. The corpus focuses on how humans would instruct a robot (rather than another human) to perform navigation tasks, and how robots could respond in a variety of situations. It explores the natural diversity of communication strategies in situated dialogue, ranging from complex, abstract-level instructions to lower-level basic control. The data and annotations have been used to advance our understanding of human-robot dialogue and to develop automated robotic systems.

We envision this corpus as providing critical annotation infrastructure and insight into multi-party cooperative tasks, for example, between heterogeneous human-robot teams. Instead of the three interlocutors (CMD, DM, RN) speaking across two conversational floors in SCOUT, a heterogeneous team of robots could organize communication with human participants in different ways, for example, unique conversational floors between CMD and each robot to avoid channel contention, or the robots communicating to each other within a conversational floor from which to then report back to the CMD. The corpus and its annotations thus represent one possible configuration out of many for multi-modal, multi-party human-robot interaction for studying how visual information, intents, and goal progress is tracked.

9. Acknowledgements

We thank many individuals for their past contributions to the project: Carlos Sanchez Amaro, David Baran, Jill Boberg, Brendan Byrne, Austin Blodgett, Taylor Cassidy, Jessica Ervin, Arthur William Evans III, Carla Gordon, Jason Gregory, Darius Jefferson II, Brecken Keller, Su Lei, Elia Martin, Pooja Moolchandani, John J Morgan, Reginald Hobbs, Brandon Perelman, John G Rogers III, Douglas Summers-Stay, Adam Wiemerslage, Volodymyr Yanov. We also thank our reviewers for suggestions on how to strengthen this paper.

Ethical Considerations

This data was collected following an approved IRB protocol with all participants signing a consent form. Personally identifiable information (PII) that was recorded during the study (i.e., participant speech and recordings of their face) are not released in SCOUT. Additional permissions to use PII, including presenting audio/video clips at conferences or publicly releasing the full audio/video, was agreed to with explicit permission.

Limitations

Due to the choice of experimental design, SCOUT data may not generalize to all scenarios. For instance, the vocabulary may be only representative of the search task assigned to the Commanders, and the low-lighting in the images may prove challenging for state-of-the-art computer vision algorithms trained on 'canonical' environments.

10. Bibliographical References

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. Association for Computational Linguistics (ACL).

Paul Boersma and David Weenink. 1992–2019. Praat: doing phonetics by computer [computer program]. Phonetic Sciences, University of Amsterdam. Version 6.1.01.

Claire 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. *International Conference on Computational Semantics (IWCS)*, page 173.

Claire Bonial, Julie Foresta, Nicholas Fung, Cory Hayes, Philip Osteen, Jacob Arkin, Benned Hedegaard, and Thomas Howard. 2023. AMR for Grounded Human-Robot Communication. In *Proceedings of the Designing Meaning Representation 2023 Workshop at IWCS 2023*.

Claire Bonial, Cassidy Henry, Ron Artstein, and Matthew Marge. 2019a. Transcription guidelines for the Army Research Laboratory (ARL) SCOUT human–robot dialogue corpus. Technical Report ARL-TR-8832, Army Research Laboratory.

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

Claire N Bonial, Lucia Donatelli, Jessica Ervin, and Clare R Voss. 2019b. Abstract meaning representation for human-robot dialogue. *Proceedings of the Society for Computation in Linguistics*, 2(1):236–246.

Anthony Brohan, Yevgen Chebotar, Chelsea Finn, Karol Hausman, Alexander Herzog, Daniel Ho, Julian Ibarz, Alex Irpan, Eric Jang, Ryan Julian, et al. 2023. Do as i can, not as i say: Grounding language in robotic affordances. In *Conference on Robot Learning*, pages 287–318. PMLR.

Paweł Budzianowski, Eric Mihail, Goel Rahul, Paul Shachi, Abhishek Sethi, Sanchit Agarwal, Shuyag Gao, and Dilek Hakkani-Tur. 2019. Research data supporting "MultiWOZ - a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling". Apollo - University of Cambridge Repository.

Jean Carletta, Amy Isard, Stephen Isard, Jacqueline Kowtko, Gwyneth Doherty-Sneddon, and Anne Anderson. 1996. HCRC dialogue structure coding manual. Technical Report 82, Human Communication Research Centre, University of Edinburgh.

David L. Chen and Raymond J. Mooney. 2011. Learning to interpret natural language navigation instructions from observations. In *AAAI'11: Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence*, pages 859–865.

Howard Chen, Alane Suhr, Dipendra Misra, Noah Snavely, and Yoav Artzi. 2019. Touchdown: Natural language navigation and spatial reasoning in visual street environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12538–12547.

Harm De Vries, Kurt Shuster, Dhruv Batra, Devi Parikh, Jason Weston, and Douwe Kiela. 2018. Talk the Walk: Navigating New York City through Grounded Dialogue. *arXiv preprint arXiv:1807.03367*.

David DeVault, Ron Artstein, Grace Benn, Teresa Dey, Edward Fast, Alesia Gainer, Kallirroi Georgila, Jonathan Gratch, Arno Hartholt, Margaux Lhommet, Gale Lucas, Stacy C. Marsella, Morbini Fabrizio, Angela Nazarian, Stefan Scherer, Giota Stratou, Apar Suri, David Traum, Rachel Wood, Yuyu Xu, Albert Rizzo, and Louis-Philippe Morency. 2014. SimSensei Kiosk: A virtual human interviewer for healthcare decision support. In *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*.

Lucia Donatelli, Kenneth Lai, and James Pustejovsky. 2020. A two-level interpretation of modality in human-robot dialogue. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4222–4238, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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

Mihail Eric and Christopher D Manning. 2017. Key-value retrieval networks for task-oriented dialogue. *arXiv preprint arXiv:1705.05414*.

Felix Gervits, Anton Leuski, Claire Bonial, Carla Gordon, and David Traum. 2019. 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.

Felix Gervits, Antonio Roque, Gordon Briggs, Matthias Scheutz, and Matthew Marge. 2021. How Should Agents Ask Questions For Situated Learning? An Annotated Dialogue Corpus. In *Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 353–359.

Brett A Halperin and Stephanie M Lukin. 2023. Envisioning narrative intelligence: A creative visual storytelling anthology. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–21.

Zhichao Hu, Gabrielle Halberg, Carolynn R Jimenez, and Marilyn A Walker. 2016. Entrainment in pedestrian direction giving: How many kinds of entrainment? *Situated dialog in speech-based human-computer interaction*, pages 151–164.

Seiya Kawano, Koichiro Yoshino, David Traum, and Satoshi Nakamura. 2023. End-to-end dialogue structure parsing on multi-floor dialogue based on multi-task learning. *Frontiers in Robotics and AI*, 10:949600.

Nathan Koenig and Andrew Howard. 2004. Design and use paradigms for Gazebo, an open-source multi-robot simulator. In *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)(IEEE Cat. No. 04CH37566)*, volume 3, pages 2149–2154. IEEE.

Anis Koubâa, editor. 2017. *Robot Operating System (ROS): The Complete Reference*, volume 1. Springer.

Kris Liu, Jean E Fox Tree, and Marilyn Walker. 2016. Coordinating Communication in the Wild: The Artwalk Dialogue Corpus of Pedestrian Navigation and Mobile Referential Communication. In *Language Resources and Evaluation Conference (LREC)*, pages 3159–3166.

Stephanie Lukin, Reginald Hobbs, and Clare Voss. 2018a. A pipeline for creative visual storytelling. In *Proceedings of the First Workshop on Storytelling*, pages 20–32.

Stephanie Lukin, Kimberly Pollard, Claire Bonial, Matthew Marge, Cassidy Henry, Ron Artstein, David Traum, and Clare Voss. 2018b. Consequences and Factors of Stylistic Differences in Human-Robot Dialogue. In *Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 110–118.

Stephanie M. Lukin, Kimbery A. Pollard, Claire Bonial, Taylor Hudson, Ron Artstein, Clare Voss, and David Traum. 2023. Navigating to success in multi-modal human-robot collaboration: Corpus and analysis. In *IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*.

Matt MacMahon, Brian Stankiewicz, and Benjamin Kuipers. 2006. Walk the talk: Connecting language, knowledge, and action in route instructions. In *AAAI'06: Proceedings of the 21st National Conference on Artificial Intelligence*, volume 2, pages 1475–1482.

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 *IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*.

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 RoboNLP: The First Workshop on Language Grounding for Robotics*.

Matthew Marge, Claire Bonial, Stephanie Lukin, Cory Hayes, Ashley Foots, Ron Artstein, Cassidy Henry, Kimberly Pollard, Carla Gordon, Felix Gervits, A Leuski, S Hill, C R Voss, and D Traum. 2018. Balancing efficiency and coverage in human-robot dialogue collection. In *AAAI Fall Symposium on Interactive Learning in Artificial Intelligence for Human-Robot Interaction*, Arlington, Virginia.

Matthew Marge, Carol Espy-Wilson, Nigel G. Ward, Abeer Alwan, Yoav Artzi, Mohit Bansal, Gil Blankenship, Joyce Chai, Hal Daumé III, Debadeepta Dey, Mary Harper, Thomas Howard, Casey Kennington, Ivana Kruijff-KorbayovÃ¡, Dinesh Manocha, Cynthia Matuszek, Ross Mead, Raymond Mooney, Roger K. Moore, Mari Ostendorf, Heather Pon-Barry, Alexander I. Rudnicky, Matthias Scheutz, Robert St. Amant, Tong Sun, Stefanie Tellex, David Traum, and Zhou Yu. 2022. Spoken language interaction with robots: Recommendations for future research. *Computer Speech & Language*, 71:101255.

Matthew Marge, Felix Gervits, Gordon Briggs, Matthias Scheutz, and Antonio Roque. 2020. Let's do that first! A comparative analysis of instruction-giving in human-human and human-robot situated dialogue. In *Proceedings of the 24th Workshop on the Semantics and Pragmatics of Dialogue (SemDial)*.

Matthew Marge, Stephen Nogar, Cory J. Hayes, Stephanie M. Lukin, Jesse Bloecker, Eric Holder, and Clare Voss. 2019. A Research Platform for Multi-Robot Dialogue with Humans. In *Conference of the North American Chapter of the Association for Computational Linguistics*, pages 132–137, Minneapolis, Minnesota. Association for Computational Linguistics (ACL).

Nikolaos Mavridis. 2015. A review of verbal and non-verbal human–robot interactive communication. *Robotics and Autonomous Systems*, 63:22–35.

Pooja Moolchandani, Cory J Hayes, and Matthew Marge. 2018. Evaluating robot behavior in response to natural language. In *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, pages 197–198.

Anjali Narayan-Chen, Prashant Jayannavar, and Julia Hockenmaier. 2019. Collaborative dialogue in Minecraft. In *ACL*, pages 5405–5415.

Aishwarya Padmakumar, Jesse Thomason, Ayush Shrivastava, Patrick Lange, Anjali Narayan-Chen, Spandana Gella, Robinson Piramuthu, Gokhan Tur, and Dilek Hakkani-Tur. 2021. TEACh: Task-driven Embodied Agents that Chat. *arXiv preprint arXiv:2110.00534*.

Penman Natural Language Group. 1989. The Penman user guide. *Technical report, Information Sciences Institute*.

Kimberly A. Pollard, Stephanie M. Lukin, Matthew Marge, Ashley Foots, and Susan G. Hill. 2018. How we talk with robots: Eliciting minimally-constrained speech to build natural language interfaces and capabilities. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 62, pages 160–164. SAGE Publications Sage CA: Los Angeles, CA, Human Factors and Ergonomics Society.

James Pustejovsky and Nikhil Krishnaswamy. 2020. Situated meaning in multimodal dialogue: human-robot and human-computer interactions. *Traitement Automatique des Langues*, 61(3):17–41.

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

Laura Stoia, Darla Magdalena Shockley, Donna K Byron, and Eric Fosler-Lussier. 2008. SCARE: a Situated Corpus with Annotated Referring Expressions. In *Language Resources and Evaluation Conference (LREC)*.

Alane Suhr, Claudia Yan, Jack Schluger, Stanley Yu, Hadi Khader, Marwa Mouallem, Iris Zhang, and Yoav Artzi. 2019. Executing Instructions in Situated Collaborative Interactions. In *Conference on Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing*, pages 2119–2130.

Jesse Thomason, Michael Murray, Maya Cakmak, and Luke Zettlemoyer. 2020. Vision-and-dialog navigation. In *Conference on Robot Learning*, pages 394–406. PMLR.

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

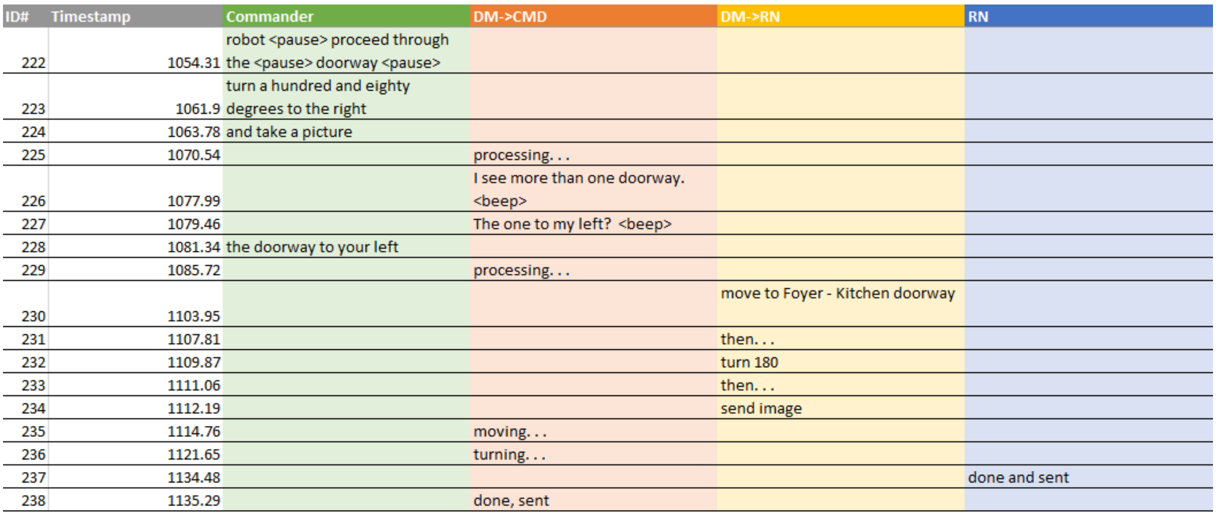
Tsung-Hsien Wen, David Vandyke, Nikola Mrksic, Milica Gasic, Lina M Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2016. A network-based end-to-end trainable task-oriented dialogue system. *arXiv preprint arXiv:1604.04562*.

Appendix

The corpus statistics from Table 3 are further broken up by experiment here in Table 5. Screenshots of the .xlsx formatted files are shown in Figures 4-7.

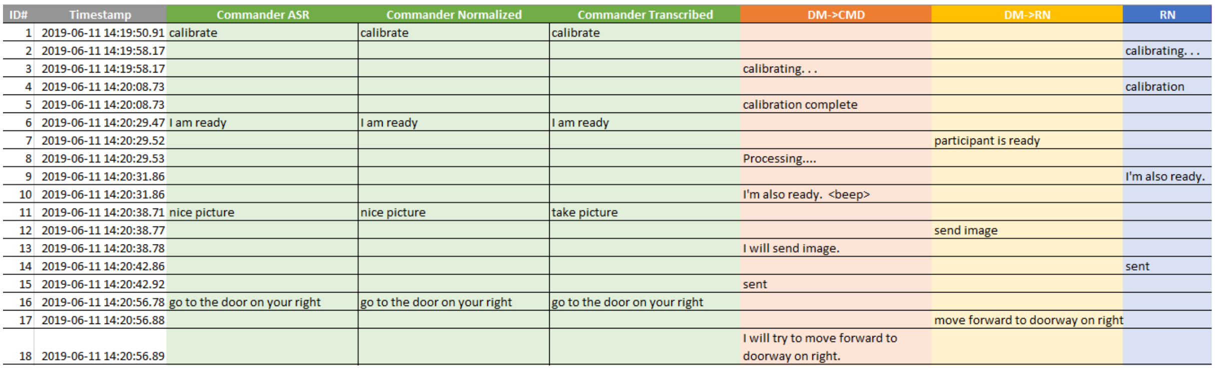
Corpus Attribute		Exp. 1	Exp. 2	Exp. 3 (Subset)	Exp. 4	Total
SCOUT Totals	Commanders	10	10	63 (10)	10	**93**
	Dialogues	30	30	188 (30)	30	**278**
	Utterances	6,097	8,421	59,859 (8,262)	14,639	**89,056**
	Words – All	17,818	29,081	207,004 (29,441)	51,194	**310,095**
	Words – Unique	1,905	1,172	3,134 (1,252)	811	**4,322**
	Images	835	565	3,694 (519)	691	**5,785**
	Avg. Utterance per Dialogue	203	280	318 (275)	487	**320**
	Avg. Images per Dialogue	27.8	18.8	19.5 (17.3)	23	**20.8**
Annotation Totals	Standard-AMR Sentences	52	212	305	—	**569**
	Dialogue-AMR Sentences	52	212	305	—	**569**
	Dialogue Structure TUs	1,005	1,243	9,169 (1,216)	2,246	**13,663**
	Dialogue Structure Relations	4,700	6,728	45,648 (6,406)	12,354	**69,430**

Table 5: SCOUT corpus statistics by experiment



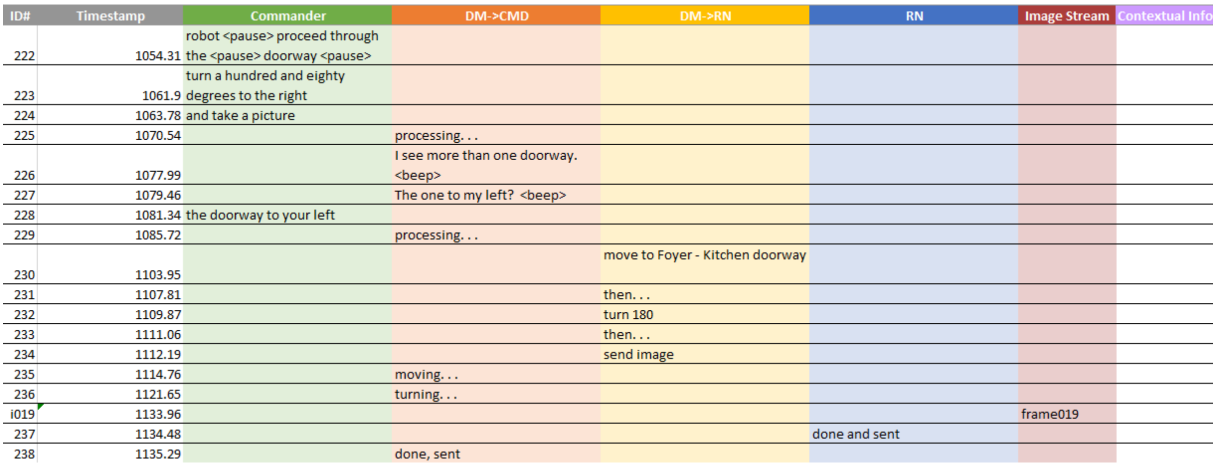
ID#	Timestamp	Commander	DM->CMD	DM->RN	RN
222	1054.31	robot <pause> proceed through the <pause> doorway <pause>			
223	1061.9	turn a hundred and eighty degrees to the right			
224	1063.78	and take a picture			
225	1070.54		processing. . .		
226	1077.99		I see more than one doorway. <beep>		
227	1079.46		The one to my left? <beep>		
228	1081.34	the doorway to your left			
229	1085.72		processing. . .		
230	1103.95			move to Foyer - Kitchen doorway	
231	1107.81			then. . .	
232	1109.87			turn 180	
233	1111.06			then. . .	
234	1112.19			send image	
235	1114.76		moving. . .		
236	1121.65		turning. . .		
237	1134.48				done and sent
238	1135.29		done, sent		

Figure 4: Aligned .xlsx transcript screenshot format for Experiments 1-3



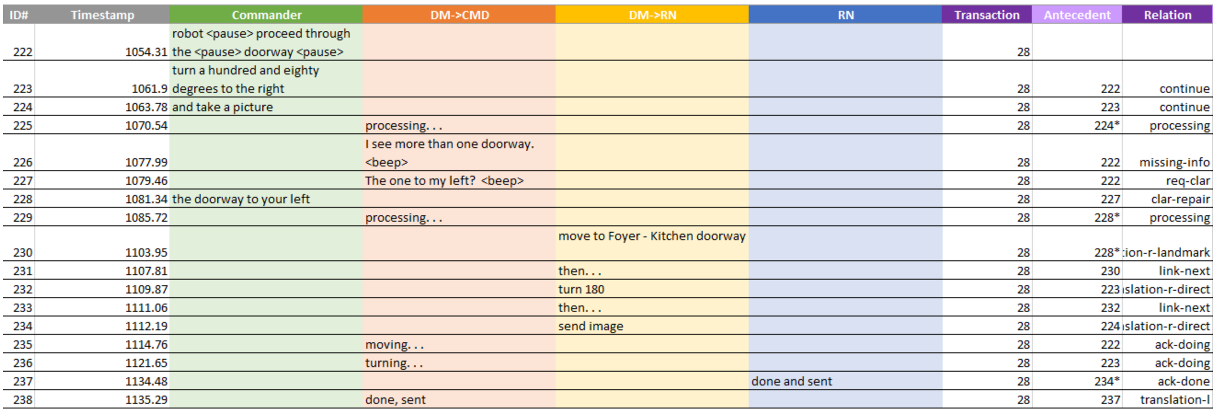
ID#	Timestamp	Commander ASR	Commander Normalized	Commander Transcribed	DM->CMD	DM->RN	RN
1	2019-06-11 14:19:50.91	calibrate	calibrate	calibrate			
2	2019-06-11 14:19:58.17						calibrating. . .
3	2019-06-11 14:19:58.17				calibrating. . .		
4	2019-06-11 14:20:08.73						calibration
5	2019-06-11 14:20:08.73				calibration complete		
6	2019-06-11 14:20:29.47	I am ready	I am ready	I am ready			
7	2019-06-11 14:20:29.52					participant is ready	
8	2019-06-11 14:20:29.53				Processing....		
9	2019-06-11 14:20:31.86						I'm also ready.
10	2019-06-11 14:20:31.86				I'm also ready. <beep>		
11	2019-06-11 14:20:38.71	nice picture	nice picture	take picture			
12	2019-06-11 14:20:38.77					send image	
13	2019-06-11 14:20:38.78				I will send image.		
14	2019-06-11 14:20:42.86						sent
15	2019-06-11 14:20:42.92				sent		
16	2019-06-11 14:20:56.78	go to the door on your right	go to the door on your right	go to the door on your right			
17	2019-06-11 14:20:56.88					move forward to doorway on right	
18	2019-06-11 14:20:56.89				I will try to move forward to doorway on right.		

Figure 5: Aligned .xlsx transcript screenshot format for Experiment 4 with the ASR results and intermediary normalized forms



ID#	Timestamp	Commander	DM->CMD	DM->RN	RN	Image Stream	Contextual Info
222	1054.31	robot <pause> proceed through the <pause> doorway <pause>					
223	1061.9	turn a hundred and eighty degrees to the right					
224	1063.78	and take a picture					
225	1070.54		processing. . .				
226	1077.99		I see more than one doorway. <beep>				
227	1079.46		The one to my left? <beep>				
228	1081.34	the doorway to your left					
229	1085.72		processing. . .				
230	1103.95			move to Foyer - Kitchen doorway			
231	1107.81			then. . .			
232	1109.87			turn 180			
233	1111.06			then. . .			
234	1112.19			send image			
235	1114.76		moving. . .				
236	1121.65		turning. . .				
i019	1133.96					frame019	
237	1134.48				done and sent		
238	1135.29		done, sent				

Figure 6: Aligned .xlsx transcript screenshot format with image references



ID#	Timestamp	Commander	DM->CMD	DM->RN	RN	Transaction	Antecedent	Relation
222	1054.31	robot <pause> proceed through the <pause> doorway <pause>				28		
223	1061.9	turn a hundred and eighty degrees to the right				28	222	continue
224	1063.78	and take a picture				28	223	continue
225	1070.54		processing. . .			28	224*	processing
226	1077.99		I see more than one doorway. <beep>			28	222	missing-info
227	1079.46		The one to my left? <beep>			28	222	req-clar
228	1081.34	the doorway to your left				28	227	clar-repair
229	1085.72		processing. . .			28	228*	processing
230	1103.95			move to Foyer - Kitchen doorway		28	228*	:ion-r-landmark
231	1107.81			then. . .		28	230	link-next
232	1109.87			turn 180		28	223	ıslation-r-direct
233	1111.06			then. . .		28	232	link-next
234	1112.19			send image		28	224	ıslation-r-direct
235	1114.76		moving. . .			28	222	ack-doing
236	1121.65		turning. . .			28	223	ack-doing
237	1134.48				done and sent	28	234*	ack-done
238	1135.29		done, sent			28	237	translation-l

Figure 7: Aligned .xlsx transcript screenshot format with Dialogue Structure annotations (the three rightmost columns)