Miscommunication Recovery in Physically Situated Dialogue

Matthew Marge[*†]
[*]Army Research Laboratory
Adelphi, MD 20783
matthew.r.marge.civ@mail.mil

Alexander I. Rudnicky[†]
[†]Carnegie Mellon University
Pittsburgh, PA 15213
air@cs.cmu.edu

Abstract

We describe an empirical study that crowdsourced human-authored recovery strategies for various problems encountered in physically situated dialogue. The purpose was to investigate the strategies that people use in response to requests that are referentially ambiguous or impossible to execute. Results suggest a general preference for including specific kinds of visual information when disambiguating referents, and for volunteering alternative plans when the original instruction was not possible to carry out.

1 Introduction

Physically situated dialogue differs from traditional human-computer dialogue in that interactions will make use of reference to a dialogue agent's surroundings. Tasks may fail due to dependencies on specific environment configurations, such as when a robot's path to a goal is blocked. People will often help; in navigation dialogues they tend to ask proactive, task-related questions instead of simply signaling communication failure (Skantze, 2005). They supplement the agent's representation of the environment and allow it to complete tasks. The current study establishes an empirical basis for grounding in physically situated contexts. We had people provide recovery strategies for a robot in various situations.

The focus of this work is on recovery from *situated grounding problems*, a type of miscommunication that occurs when an agent fails to uniquely map a person's instructions to its surroundings (Marge and Rudnicky, 2013). A *referential ambiguity* is where an instruction resolves to more than one possibility (e.g., "Search the room on the left" when there are multiple rooms on the agent's left); an *impossible-to-execute* problem fails to resolve to any action (e.g., same instruction but there are no rooms on the agent's left). A common strategy evidenced in human-human corpora is for people to ask questions to recover from situated grounding problems (Tenbrink et al., 2010).

Dialogue divides into two levels: that of managing the actual dialogue—determining who has the floor, that an utterance was recognized, etc.—and the dialogue that serves the main *joint activities* that dialogue partners are carrying out, like a human-robot team exploring a new area (Bangerter and Clark, 2003). Most approaches to grounding in dialogue systems are managing the dialogue itself, making use of spoken language input as an indicator of understanding (e.g., (Bohus, 2007; Skantze, 2007)). Situated grounding problems are associated with the main joint activities; to resolve them we believe that the recovery model must be extended to include planning and environment information. Flexible recovery strategies make this possible by enabling dialogue partners to coordinate their joint activities and accomplish tasks.

We cast the problem space as one where the agent aims to select the most efficient recovery strategy that would resolve a user's intended referent. We expect that this efficiency is tied to the cognitive load it takes to produce clarifications. Viethen and Dale (2006) suggest a similar prediction in their study comparing human and automatically generated referring expressions of objects and their properties. We sought to answer the following questions in this work:

- How good are people at detecting situated grounding problems?
- How do people organize recovery strategies?
- When resolving ambiguity, which properties do people use to differentiate referents?
- When resolving impossible-to-execute instructions, do people use active or passive ways to get the conversation back on track?

Proceedings of the SIGDIAL 2015 Conference, pages 22–31,
Prague, Czech Republic, 2-4 September 2015.

We determined the most common recovery strategies for referential ambiguity and impossible-to-execute problems. Several patterns emerged that suggest ways that people expect agents to recover. Ultimately we intend for dialogue systems to use such strategies in physically situated contexts.

2 Related Work

Researchers have long observed miscommunication and recovery in human-human dialogue corpora. The HCRC MapTask had a direction giver-direction follower pair navigate two dimensional schematics with slightly different maps (Anderson et al., 1991). Carletta (1992) proposed several recovery strategies following an analysis of this corpus. The SCARE corpus collected human-human dialogues in a similar scenario where the direction follower was situated in a three-dimensional virtual environment (Stoia et al., 2008).

The current study follows up an initial proposal set of recovery strategies for physically situated domains (Marge and Rudnicky, 2011). Others have also developed recovery strategies for situated dialogue. Kruijff et al. (2006) developed a framework for a robot mapping an environment that employed conversational strategies as part of the grounding process. A similar study focused on resolving misunderstandings in the human-robot domain using the Wizard-of-Oz methodology (Koulouri and Lauria, 2009). A body of work on referring expression generation uses object attributes to generate descriptions of referents (e.g., (Guhe and Bard, 2008; Garoufi and Koller, 2014)). Viethen and Dale (2006) compared human-authored referring expressions of objects to existing natural language generation algorithms and found them to have very different content.

Crowdsourcing has been shown to provide useful dialogue data: Manuvinakurike and DeVault (2015) used the technique to collect game-playing conversations. Wang et al. (2012) and Mitchell et al. (2014) have used crowdsourced data for training, while others have used it in real time systems (Lasecki et al., 2013; Huang et al., 2014).

3 Method

In this study, participants came up with phrases that a search-and-rescue robot should say in response to an operator's command. The participant's task was to view scenes in a virtual environment then formulate the robot's response to an operator's request. Participants listened to an operator's verbal command then typed in a response.

Figure 1: An example trial where the operator's command was "Move to the table". In red is the robot (*centered*) pointed toward the back wall. Participants would listen to the operator's command and enter a response into a text box.

Scenes displayed one of three situations: *referential ambiguity* (more than one possible action), *impossible-to-execute* (zero possible actions), and *executable* (one possible action). The instructions showed some example problems. All situations involved one operator and one robot.

3.1 Experiment Design

After instructions and a practice trial, participants viewed scenes in one of 10 different environments (see Figure 1). They would first watch a fly-over video of the robot's environment, then view a screen showing labels for all possible referable objects in the scene. The participant would then watch the robot enter the first scene. The practice trial and instructions did not provide any examples of questions.

The robot would stop and a spoken instruction from the operator would be heard. The participant was free to replay the instruction multiple times. They would then enter a response (say an acknowledgment or a question). Upon completion of the trial, the robot would move to a different scene, where the process was repeated.

Only self-contained questions that would allow the operator to answer without follow-up were allowed. Thus generic questions like "which one?" would not allow the operator to give the robot enough useful information to proceed. In the instructions, we suggested that participants include some detail about the environment in their ques-

Trial Group	#PARTIC	#AMB	#IMP	#EXE
1	15	9	9	7
2	15	16	6	3
Total	30	25	15	10

Table 1: Distribution of stimulus types across the two trial groups of participants (PARTIC). Trials either had referential ambiguity (AMB), were impossible-to-execute (IMP), or executable (EXE).

tions.

Participants used a web form[1] to view situations and provide responses. We recorded demographic information (gender, age, native language, native country) and time on task. The instructions had several attention checks (Paolacci et al., 2010) to ensure that participants were focusing on the task.

We created fifty trials across ten environments. Each environment had five trials that represented waypoints the robot was to reach. Participants viewed five different environments (totaling twenty-five trials). Each command from the remote operator to the robot was a route instruction in the robot navigation domain. Trials were assembled in two groups and participants were assigned randomly to one (see Table 1). Trial order was randomized according to a Latin Square.

3.1.1 Scenes and Environments

Scenes were of a 3D virtual environment at eye level, with the camera one to two meters behind the robot. Camera angle issues with environment objects caused this variation.

Participants understood that the fictional operator was not co-located with the robot. The USARSim robot simulation toolkit and the UnrealEd game map editor were used to create the environment. Cepstral's SwiftTalker was used for the operator voice.

Of the fifty scenes, twenty-five (50%) had referential ambiguities, fifteen (30%) were impossible-to-execute, and ten (20%) were executable controls. The selection was weighted to referential ambiguity, as these were expected to produce greater variety in recovery strategies. We randomly assigned each of fifty trials a stimulus type according to this distribution, then divided the list into ten environments. The environments featured objects and doorways appropriate to the trial type, as well as waypoints.

[1] See http://goo.gl/forms/ZGpK3L1nPh for an example.

Referential Ambiguity We arranged the sources of information participants could use to describe referents, to enable analysis of the relationship between context and recovery strategies. The sources of information (i.e., "situated dimensions") were: (1) *intrinsic properties* (either color or size), (2) *history* (objects that the robot already encountered), (3) *egocentric proximity* of the robot to candidate referents around it (the robot's perspective is always taken), and (4) *object proximity* (proximity of candidate referents to other objects). Table 2 provides additional details.

Scenes with referential ambiguity had up to four sources of information available. Information sources were evenly distributed across five trial types: one that included all four sources, and four that included all but one source of information (e.g., one division excluded using history information but did allow proximity, spatial, and object properties, one excluded proximity, etc.).

Impossible-to-Execute The impossible-to-execute trials divided into two broad types. Nine of the fifteen scenes were impossible because the operator's command did not match to any referent in the environment. The other six scenes were impossible because a path to get to the matching referent was not possible.

Executable Ten scenes were executable for the study and served as controls. The operator's command mentioned existing, unambiguous referents.

3.1.2 Robot Capabilities

Participants were aware of the robot's capabilities before the start of the experiment. The instructions said that the robot knew the locations of all objects in the environment and whether doors were closed or open. The robot also knew the color and size of objects in the environment (*intrinsic properties*), where objects were relative to the robot itself and to other objects (*proximity*), when objects were right, left, in front, and behind it (*spatial terms*), the room and hallway locations of objects (*location*), and the places it has been (*history*, the robot kept track of which objects it had visited). The robot could not pass through closed doors.

3.2 Hypotheses

We made five hypotheses about the organization and content of participant responses to situated grounding problems:

Dimension	Property	#Scenes
Intrinsic (aka "perceptual feature")	On no dimension does the target referent share an intrinsic property value with any other object of its type. The two intrinsic properties are color and size.	20
History (aka "conceptual feature")	The robot already visited the referent once.	14
Object Proximity (aka "functional relation")	The referent has a unique, nearby object that can serve as a "feature" for reference purposes.	21
Egocentric Proximity (aka "spatial relation")	The referent has a unique spatial relationship relative to the robot. The relation is prototypical, generally falling along a supposed axis with the robot.	20

Table 2: Ambiguous scene referent description space. Number of scenes was out of 25 total. We relate the current terms to general types defined by Carlson and Hill (2009).

- *Hypothesis 1:* Participants will have more difficulty detecting impossible-to-execute scenes than ambiguous ones. Determining a robot's tasks to be impossible requires good *situation awareness* (Nielsen et al., 2007) (i.e., an understanding of surroundings with respect to correctly completing tasks). Detecting referential ambiguity requires understanding the operator's command and visually inspecting the space (Spivey et al., 2002); detecting impossible commands also requires recalling the robot's capabilities and noticing obstacles. Previous research has noted that remote teleoperators have trouble establishing good situation awareness of a robot's surroundings (Casper and Murphy, 2003; Burke et al., 2004). Moreover, obstacles near a robot can be difficult to detect with a restricted view as in the current study (Alfano and Michel, 1990; Arthur, 2000).
- *Hypotheses 2a and 2b:* Responses will more commonly be single, self-contained questions instead of a scene description followed by a question (2a for scenes with referential ambiguity, 2b for scenes that were impossible-to-execute). This should reflect the principle of *least effort* (Clark, 1996), and follow from Carletta's (1992) observations in a similar dataset.
- *Hypothesis 3:* Responses will use the situated dimensions that require the least cognitive effort when disambiguating referents. Viethen and Dale (2006) suggest that minimizing cognitive load for the speaker or listener would produce more human-like referring expressions. We predict that responses will mention visually salient features of the scene, such as color or size of referents, more than history or object proximity. Desimone and Duncan (1995) found that color and shape draw more attention than other properties in visual search tasks when they are highly distinguishable.
- *Hypothesis 4:* In cases of referential ambiguity where two candidate referents are present, responses will confirm one referent in the form of a yes-no question more than presenting a list. Results from an analysis of task-oriented dialogue suggests that people are efficient when asking clarification questions (Rieser and Moore, 2005). Additionally, Clark's *least effort* principle (Clark, 1996) suggests that clarifying one referent using a yes-no confirmation would require less effort than presenting a list in two ways: producing a shorter question and constraining the range of responses to expect.
- *Hypothesis 5:* For impossible-to-execute instructions, responses will most commonly be ways for the robot to proactively work with the operator's instruction, in an effort to get the conversation back on track. The other possible technique, to simply declare that the problem is not possible, will be less common. This is because participants will believe such a strategy will not align with the task goal of having the robot say something that will allow it to proceed with the task. Skantze found that in human-human navigation dialogues, people would prefer to look for alternative ways to proceed rather than simply express non-understanding (Skantze, 2005).

3.3 Measures

The key independent variable in this study was the stimulus type that the participant viewed (i.e., referential ambiguity, impossible-to-execute, or executable). Dependent variables were observational measurements, presented below. We report Fleiss' kappa score for inter-annotator agreement

between three native English speaking annotators on a subset of the data.

Correctness ($\kappa = 0.77$): Whether participants correctly determined the situation as ambiguous, impossible, or executable. Annotators labeled correctness based on the content of participant responses. This measure assessed participant accuracy for detecting situated grounding problems. Either *correct* or *incorrect*.

Sentence type ($\kappa = 0.82$): Either *declarative*, *interrogative*, *imperative*, or *exclamatory* (Cowan, 2008).

Question type ($\kappa = 0.92$): Sentences that needed an answer from the operator. The three types were *yes-no questions*, *alternative questions* (which presented a list of options and includes *wh-* questions that used sources from Table 2), and generic *wh- questions* (Cowan, 2008).

Situated dimensions in response ($\kappa = 0.75$): The capability (or capabilities) that the participant mentioned when providing a response. The types were *intrinsic* (color or size), *object proximity*, *egocentric proximity*, and *history*.

Projected belief (impossible-to-execute trials only, $\kappa = 0.80$): The participant's belief about the next task, given the current operator instruction (projected onto the robot). The types were *unknown* (response indicates participant is unsure what to do next), *ask for more* (ask for more details), *propose alternative* (propose alternative object), *ask for help* (ask operator to physically manipulate environment), and *off topic*.

3.4 Participation

We recruited 30 participants. All participants completed the web form through the Amazon Mechanical Turk (MTurk) web portal[2], all were located in the United States and had a task approval rate $\geq$95%. The group included 29 self-reported native English speakers born in the United States; 1 self-reported as a native Bangla speaker born in Bangladesh. The gender distribution was 15 male to 15 female. Participants ranged in age from 22 to 52 (*mean:* 33 years, *std. dev.:* 7.7). They were paid between $1 and $2 for their participation. We collected a total of 750 responses.

[2]https://www.mturk.com

Problem Type	Sample Crowdsourced Responses
Referential Ambiguity	▷ *Do you mean the table in front of me?* ▷ *Should I go to the small or big table?*
Impossible-to-Execute	▷ *There is not a lamp behind me. Would you like for me to go to the lamp in front of me?* ▷ *Do you mean the lamp in front of me?*

Table 3: Participants composed recovery strategies in response to operator commands that were referentially ambiguous or impossible-to-execute.

4 Results

We analyzed the measures by tabulating frequencies for each possible value. Table 3 presents some example responses.

4.1 Correctness

In general, participants were good at detecting situated grounding problems. Out of 750 responses, 667 (89%) implied the correct scene type. We analyzed correctness across actual stimulus types (ambiguous, impossible-to-execute, executable) using a mixed-effects analysis of variance model[3], with participant included as a random effect and trial group as a fixed effect.

Hypothesis 1 predicted that participants will do better detecting scenes with referential ambiguity than those that were impossible-to-execute; the results support this hypothesis. Actual stimulus type had a significant main effect on correctness ($F[2, 58] = 12.3$, $p < 0.001$); trial group did not ($F[1, 28] = 0.1$, $p = 0.72$). Participants had significantly worse performance detecting impossible-to-execute scenes compared to ambiguous ones ($p < 0.001$; Tukey HSD test). In fact, they were four times worse; of the impossible-to-execute scenes, participants failed to detect that 22% (50/225) of them were impossible, compared to 5% (17/375) of scenes with referential ambiguity. Of the 150 instructions that were executable, participants failed to detect 11% (16/150) of them as such.

4.2 Referential Ambiguity

We analyzed the 358 responses where participants correctly detected referential ambiguity.

[3]This approach computed standard least squares regression using reduced maximum likelihood (Harville, 1977).

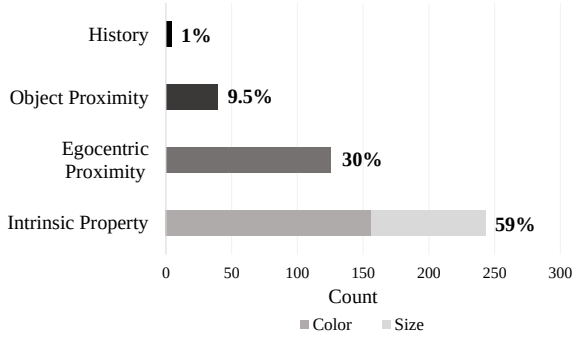

Figure 2: Counts of situated dimensions in recovery strategies for scenarios with referential ambiguity.

Hypothesis 2a predicted that participants would more commonly ask single, self-contained questions instead of describing the scene and asking a question. We assessed this by counting sentence types within a response. Responses that had both a declarative sentence and an interrogative would fit this case. The results confirmed this hypothesis. Only 4.5% (16/358) of possible responses had a declarative and an interrogative.

Hypothesis 3 predicted that participants would use the situated dimensions that require the least cognitive effort when disambiguating referents. More specifically, the most common mentions will be those that are visually apparent (intrinsic properties like color and size), while those that require more processing would have fewer mentions (history and to a lesser extent object proximity and egocentric proximity). We measured this by tabulating mentions of situated dimensions in all 358 correct participant responses, summarized in Figure 2. Multiple dimensions could occur in a single response. The results support this hypothesis. By far, across all ambiguous scenarios, the most mentioned dimension was an intrinsic property. More than half of all situated dimensions used were intrinsic (59%, 242/410 total mentions). This was followed by the dimensions that we hypothesize require more cognitive effort: egocentric proximity had 30% (125/410) of mentions, object proximity 9.5% (39/410), and history 1% (4/410). Of the intrinsic dimensions mentioned, most were only color (61%, 148/242), followed by size (33%, 81/242), and using both (5%, 13/242).

Hypothesis 4 predicted that participants would ask yes-no confirmation questions in favor of presenting lists when disambiguating a referent with exactly two candidates. The results suggest that the opposite is true; people strongly preferred to list options, even when a confirmation question about one would have been sufficient. Of the 285 responses that were correctly detected as ambiguous and were for scenes of exactly two possible referents, 74% (212/285) presented a list of options. Only 14% (39/285) asked yes-no confirmation questions. The remaining 34 questions (12%) were generic wh-questions. These results held in scenes where three options were present. Overall 72% (259/358) presented a list of options, while 16% (58/358) asked generic wh-questions and 11% (41/358) asked yes-no confirmations.

Projected Belief	**Count**	**Percentage**
Propose Alternative	72	41%
Unknown	56	32%
Ask for More	42	24%
Ask for Help	5	3%
Total	175	100%

Table 4: Projected belief annotations for the 175 correct detections of impossible-to-execute stimuli.

4.3 Impossible-to-Execute

We analyzed the 175 responses where participants correctly identified impossible-to-execute situations.

Hypothesis 2b predicted that participants would more often only ask a question than also describe the scene. Results confirmed this hypothesis. 42% (73/175) of responses simply asked a question, while 22% (39/175) used only a declarative. More than a third included a declarative as well (36%, 63/175). The general organization to these was to declare the problem then ask a question about it (89%, 56/63).

Hypothesis 5 predicted that responses for impossible-to-execute instructions will more commonly be proactive and make suggestions, instead of simply declaring that an action was not possible. Table 4 summarizes the results, which confirmed this hypothesis. The most common belief that participants had for the robot was to have it propose an alternative referent to the impossible one specified by the operator. The next-most common was to have the robot simply express uncertainty about what to do next. Though this belief occurred in about a third of responses, the remaining responses were all proactive ways for the robot to get the conversation back on track (i.e., propose alternative, ask for more, and ask for help).

5 Discussion

The results largely support the hypotheses, with the exception of Hypothesis 4. They also provide information about how people expect robots to recover from situated grounding problems.

Correctness Participants had the most trouble detecting impossible-to-execute scenes, supporting Hypothesis 1. An error analysis of the 50 responses for this condition had participants responding as if the impossible scenes were possible (62%, 31/50). The lack of good situation awareness was a factor, which agrees with previous findings in the human-robot interaction literature (Casper and Murphy, 2003; Burke et al., 2004). We found that participants had trouble with a specific scene where they confused the front and back of the robot (9 of the 31 impossible-executable responses were for this scene). Note that all scenes showed the robot entering the room with the same perspective, facing forward.

Referential Ambiguity Results for Hypothesis 2a showed that participants overwhelmingly asked only a single, self-contained question as opposed to first stating that there was an ambiguity. Participants also preferred to present a list of options, despite the number of possible candidates. This contradicted Hypothesis 4. Rieser and Moore (2005) found that in task-oriented human-human dialogues, clarification requests aim to be as efficient as possible; they are mostly partially formed. The results in our study were not of real-time dialogue; we isolated specific parts of what participants believed to be human-computer dialogue. Moreover, Rieser and Moore were observing clarifications at Bangerter and Clark's (2003) dialogue management level; we were observing them in service of the joint activity of navigating the robot. We believe that this difference resulted in participants using caution by disambiguating with lists.

These results suggest that dialogue systems should present detection of referential ambiguity implicitly, and as a list. Generic *wh-* questions (e.g., "which one?" without presenting a follow-on list) are less desirable because they don't constrain what the user can say, and don't provide any indication of what the dialogue system can understand. A list offers several benefits: it grounds awareness of surroundings, presents a fixed set of options to the user, and constrains the range of linguistic responses. This could also extend to general ambiguity, as in when there are a list of matches to a query, but that is outside the scope of this work. Lists may be less useful as they grow in size; in our study they could not grow beyond three candidates.

The data also supported Hypothesis 3. Participants generally preferred to use situated dimensions that required less effort to describe. Intrinsic dimensions (color and size) had the greatest count, followed by egocentric proximity, object proximity, and finally using history. We attribute these results to the salient nature of intrinsic properties compared to ones that must be computed (i.e., egocentric and object proximity require spatial processing, while history requires thinking about previous exchanges). This also speaks to a similar claim by Viethen and Dale (2006). Responses included color more than any other property, suggesting that an object's color draws more visual attention than its size. Bright colors and big shapes stand out most in visual search tasks; we had more of the former than the latter (Desimone and Duncan, 1995).

For an ambiguous scene, participants appear to traverse a *salience hierarchy* (Hirst et al., 1994) whereby they select the most visually salient feature that also uniquely teases apart candidates. While the salience hierarchy varies depending on the current context of a referent, we anticipate such a hierarchy can be defined computationally. Others have proposed similar processes for referring expression generation (Van Der Sluis, 2005; Guhe and Bard, 2008). One way to rank salience on the hierarchy could be predicted mental load; we speculate that this is a reason why history was barely mentioned to disambiguate. Another would be to model visual attention, which could explain why color was so dominant.

Note that only a few dimensions were "competing" at any given time, and their presence in the scenes was equal (save for history, which had slightly fewer due to task design constraints). Egocentric proximity, which uses spatial language to orient candidate referents relative to the robot, had a moderate presence. When intrinsic properties were unavailable in the scene, responses most often used this property. We found that sometimes participants would derive this property even if it wasn't made prototypical in the scene (e.g., referring to a table as "left" when it was in front and

off to the left side of the robot). This suggests that using egocentric proximity to disambiguate makes a good fallback strategy when nothing else works. Another situated dimension emerged from the responses, disambiguation by location (e.g., "Do you mean the box in this room or the other one?"). Though not frequent, it provides another useful technique to disambiguate when visually salient properties are not available.

Our findings differ from those of Carlson and Hill (2009) who found that salience is not as prominent as spatial relationships between a target (in the current study, this would be the robot) and other objects. Our study did not direct participants to formulate spatial descriptions; they were free to compose responses. In addition, our work directly compares intrinsic properties for objects of the same broad type (e.g., disambiguation of a doors of different colors). Our findings suggest the opposite of Moratz et al. (2003), who found that when pointing out an object, describing its position may be better than describing its attributes in human-robot interactions. Their study only had one object type (cube) and did not vary color, size, or proximity to nearby objects. As a result, participants described objects using spatial terms. In our study, we explored variation of several attributes to determine participants' preferences.

Impossible-to-Execute Results supported Hypothesis 2b. Most responses had a single sentence type. Although unanticipated, a useful strategy emerged: describe the problem that makes the scene impossible, then propose an alternative referent. This type of strategy helped support Hypothesis 5. Responses for impossible scenes largely had the participant proactively presenting a way to move the task forward, similar to what Skantze (2005) observed in human-human dialogues. This suggests that participants believed the robot should ask directed questions to recover. These questions often took the form of posing alternative options.

5.1 Limitations

We used the Amazon Mechanical Turk web portal to gather responses in this study. As such we could not control the participant environment when taking the study, but we did include attention checks. Participants did not interact with a dialogue system. Instead we isolated parts of the interaction that were instances of where the robot would have to say something in response to an instruction. We asked participants to provide what they think the robot should say; there was no ongoing interaction. However, we maintained continuity by presenting videos of the robot navigating through the environment as participants completed the task. The robot was represented in a virtual environment, which prevents us from understanding if there are any influencing factors that may impact results if the robot were in physical form or co-present with the participant.

6 Conclusions

Recovery strategies allow situated agents like robots to recover from misunderstandings by using the human dialogue partner. We conducted a study that collected recovery strategies for physically situated dialogue with the goal of establishing an empirical basis for grounding in physically situated contexts. We crowdsourced 750 written strategies across 30 participants and analyzed their situated properties and how they were organized.

We found that participants' recovery strategies minimize cognitive effort and indicate a desire to successfully complete the task. For disambiguation, there was a preference for strategies that use visually salient properties over ones that require additional mental processing, like spatial reasoning or memory recall. For impossible-to-execute scenes, responses more often presented alternative referents than just noting non-understanding. We should note that some differences between our findings and those of others may in part rest on differences in task and environment, though intrinsic variables such as mental effort will likely persist over different situations.

In future work, we intend to use these data to model salience ranking in similar contexts. We will further assess the hypothesis that participants' preferences in this study will enhance performance in a spoken dialogue system that deploys similar strategies.

Acknowledgments

The authors thank Prasanna Kumar Muthukumar and Juneki Hong for helping to annotate recovery strategies. We also thank Taylor Cassidy, Arthur William Evans, and the anonymous reviewers for their valuable comments.

References

Patricia L. Alfano and George F. Michel. 1990. Restricting the field of view: Perceptual and performance effects. *Perceptual and Motor Skills*, 70(1):35–45.

Anne H. Anderson, Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, Catherine Sotillo, Henry S. Thompson, and Regina Weinert. 1991. The HCRC Map Task Corpus. *Language and Speech*, 34(4):351–366.

Kevin Wayne Arthur. 2000. *Effects of field of view on performance with head-mounted displays*. Ph.D. thesis, University of North Carolina at Chapel Hill.

Adrian Bangerter and Herbert H. Clark. 2003. Navigating joint projects with dialogue. *Cognitive Science*, 27(2):195–225.

Dan Bohus. 2007. *Error Awareness and Recovery in Conversational Spoken Language Interfaces*. Ph.D. thesis, Carnegie Mellon University.

Jennifer L. Burke, Robin R. Murphy, Michael D. Coovert, and Dawn L. Riddle. 2004. Moonlight in Miami: Field study of human-robot interaction in the context of an urban search and rescue disaster response training exercise. *Human–Computer Interaction*, 19(1-2):85–116.

Jean Carletta. 1992. Planning to fail, not failing to plan: Risk-taking and recovery in task-oriented dialogue. In *Proc. of the 14th Conference on Computational Linguistics: Volume 3*, pages 896–900. Association for Computational Linguistics.

Laura A. Carlson and Patrick L. Hill. 2009. Formulating spatial descriptions across various dialogue contexts. In K. R. Coventry, T. Tenbrink, and J. Bateman, editors, *Spatial Language and Dialogue*. Oxford University Press.

Jennifer Casper and Robin R. Murphy. 2003. Human-robot interactions during the robot-assisted urban search and rescue response at the world trade center. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 33(3):367–385.

Herbert H. Clark. 1996. *Using Language*. Cambridge University Press.

Ron Cowan. 2008. *The Teacher's Grammar of English with Answers: A Course Book and Reference Guide*. Cambridge University Press.

Robert Desimone and John Duncan. 1995. Neural mechanisms of selective visual attention. *Annual Review of Neuroscience*, 18(1):193–222.

Konstantina Garoufi and Alexander Koller. 2014. Generation of effective referring expressions in situated context. *Language, Cognition and Neuroscience*, 29(8):986–1001.

Markus Guhe and Ellen Gurman Bard. 2008. Adapting referring expressions to the task environment. In *Proc. of the 30th Annual Conference of the Cognitive Science Society (CogSci)*, pages 2404–2409.

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.

Graeme Hirst, Susan McRoy, Peter Heeman, Philip Edmonds, and Diane Horton. 1994. Repairing conversational misunderstandings and non-understandings. *Speech Communication*, 15(3-4):213 – 229.

Ting-Hao K. Huang, Walter S. Lasecki, Alan L. Ritter, and Jeffrey P. Bigham. 2014. Combining non-expert and expert crowd work to convert web apis to dialog systems. In *Proc. of Second AAAI Conference on Human Computation and Crowdsourcing*.

Theodora Koulouri and Stanislao Lauria. 2009. Exploring miscommunication and collaborative behaviour in human-robot interaction. In *Proc. of SIGdial'09*, pages 111–119.

Geert-Jan Kruijff, Hendrik Zender, Patric Jensfelt, and Henrik I. Christensen. 2006. Situated dialogue and understanding spatial organization: Knowing what is where and what you can do there. In *Proc. of ROMAN'06*, pages 328–333.

Walter S. Lasecki, Rachel Wesley, Jeffrey Nichols, Anand Kulkarni, James F. Allen, and Jeffrey P. Bigham. 2013. Chorus: a crowd-powered conversational assistant. In *Proc. of the 26th Annual ACM Symposium on User Interface Software and Technology*, pages 151–162. ACM.

Ramesh Manuvinakurike and David DeVault. 2015. Pair me up: A web framework for crowd-sourced spoken dialogue collection. In *Proc. of IWSDS'15*.

Matthew Marge and Alexander I. Rudnicky. 2011. Towards overcoming miscommunication in situated dialogue by asking questions. In *Proc. of AAAI Fall Symposium Series - Building Representations of Common Ground with Intelligent Agents*, Washington, DC.

Matthew Marge and Alexander I. Rudnicky. 2013. Towards evaluating recovery strategies for situated grounding problems in human-robot dialogue. In *Proc. of ROMAN'13*, pages 340–341.

Margaret Mitchell, Dan Bohus, and Ece Kamar. 2014. Crowdsourcing language generation templates for dialogue systems. In *Proc. of INLG'14*.

Reinhard Moratz, Thora Tenbrink, John Bateman, and Kerstin Fischer. 2003. Spatial knowledge representation for human-robot interaction. In *Spatial Cognition III*, pages 263–286. Springer.

Curtis W Nielsen, Michael A Goodrich, and Robert W Ricks. 2007. Ecological interfaces for improving mobile robot teleoperation. *IEEE Transactions on Robotics*, 23(5):927–941.

Gabriele Paolacci, Jesse Chandler, and Panagiotis G. Ipeirotis. 2010. Running experiments on amazon mechanical turk. *Judgment and Decision Making*, 5(5):411–419.

Verena Rieser and Johanna D. Moore. 2005. Implications for generating clarification requests in task-oriented dialogues. In *Proc. of the ACL'05*, pages 239–246.

Gabriel Skantze. 2005. Exploring human error recovery strategies: Implications for spoken dialogue systems. *Speech Communication*, 45(3):325–341.

Gabriel Skantze. 2007. *Error Handling in Spoken Dialogue Systems: Managing Uncertainty, Grounding and Miscommunication*. Ph.D. thesis, KTH Royal Institute of Technology.

Michael J. Spivey, Michael K. Tanenhaus, Kathleen M. Eberhard, and Julie C. Sedivy. 2002. Eye movements and spoken language comprehension: Effects of visual context on syntactic ambiguity resolution. *Cognitive Psychology*, 45(4):447–481.

Laura Stoia, Darla M. Shockley, Donna K. Byron, and Eric Fosler-Lussier. 2008. Scare: A situated corpus with annotated referring expressions. In *Proc. of LREC'08*, Marrakesh, Morocco.

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

Ielka Francisca Van Der Sluis. 2005. *Multimodal Reference, Studies in Automatic Generation of Multimodal Referring Expressions*. Ph.D. thesis, University of Tilburg.

Jette Viethen and Robert Dale. 2006. Algorithms for generating referring expressions: Do they do what people do? In *Proc. of INLG'06*, pages 63–70.

William Yang Wang, Dan Bohus, Ece Kamar, and Eric Horvitz. 2012. Crowdsourcing the acquisition of natural language corpora: Methods and observations. In *Proc. of SLT'12*, pages 73–78.