

Collaborative Effort towards Common Ground in Situated Human-Robot Dialogue

Joyce Y. Chai, Lanbo She, Rui Fang, Spencer Ottarson, Cody Littley,
Changsong Liu and Kenneth Hanson

Department of Computer Science and Engineering
Michigan State University
East Lansing, MI 48824, USA

{jchai, shelanbo, fangrui, ottarso5, littley1, cliu, hanson54}@cse.msu.edu

ABSTRACT

In situated human-robot dialogue, although humans and robots are co-present in a shared environment, they have significantly mismatched capabilities in perceiving the shared environment. Their representations of the shared world are misaligned. In order for humans and robots to communicate with each other successfully using language, it is important for them to mediate such differences and to establish common ground. To address this issue, this paper describes a dialogue system that aims to mediate a shared perceptual basis during human-robot dialogue. In particular, we present an empirical study that examines the role of the robot's collaborative effort and the performance of natural language processing modules in dialogue grounding. Our empirical results indicate that in situated human-robot dialogue, a low collaborative effort from the robot may lead its human partner to believe a common ground is established. However, such beliefs may not reflect true mutual understanding. To support truly grounded dialogues, the robot should make an extra effort by making its partner aware of its internal representation of the shared world.

Categories and Subject Descriptors: H.5.2 [Information Interfaces and Presentation]: User Interfaces

General Terms: Algorithms, Experimentation

Keywords: Common Ground, Collaboration, Human-Robot Dialogue

1. INTRODUCTION

In human-human communication, when we are co-present, we have the same perceptual access to the shared environment [4] and generally do not have any problem understanding each other's references (e.g., what "the green cup" refers to in the environment). What makes the communication between us successful is the *common ground* we have: mutual beliefs and knowledge about the shared environment [4].

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
HRI'14, March 3–6, 2014, Bielefeld, Germany.
Copyright 2014 ACM 978-1-4503-2658-2/14/03 ...\$15.00.
<http://dx.doi.org/10.1145/2559636.2559677>.

However, in human-robot dialogue, although a human and a robot are co-present in a shared environment, they have significantly mismatched perceptual capabilities. The "cup" that is easily recognized by the human may not be recognizable to the robot. In addition, the human has much more knowledge about the environment and can infer and represent the environment symbolically and thus be able to communicate with others using language. However the robot's representation of the same environment may not be symbolic and may contain lower level numerical features (such as bounding boxes or color distributions associated with the perceived objects). These phenomena contribute to a gap between the human representation and the robot representation of the shared environment. Therefore, the shared perceptual basis is missing. In order for humans and robots to communicate with each other successfully using language, it is important for them to mediate such differences and establish a common ground.

Despite recent efforts [9, 13, 20, 21], research on situated human-robot dialogue is still in its infancy. It is not clear how the mismatched perceptual basis may affect the use of language and how to mediate such differences. To address this issue, this paper presents a dialogue system implemented in a NAO robot. Our robot can carry natural language dialogue with its human partner and utilize the collaborative discourse to acquire knowledge about the environment and strive to establish common ground. In particular, we present an empirical study that examines the role of the robot's collaborative effort in this process. A *low effort* refers to the robot's minimum effort in accepting or rejecting a presentation from its partner through explicit confirmation. A *high effort* refers to the robot's extra effort in proactively describing what it perceives from the shared environment in addition to an explicit confirmation. Our empirical results have shown that, a low collaborative effort from the robot may lead its human partner to believe a common ground is established. However, such beliefs may not reflect true mutual understanding. To establish a true common ground, the robot should make an extra effort by making its partner aware of its internal representation of the shared world.

2. RELATED WORK

Common ground is essential to the success of communication [24]. It is defined as a collection of "mutual, common, or joint knowledge, beliefs, and suppositions." [5]. In conversation, participants coordinate their mental states based



(a) Situated Setup

H1:	The green cup is called Bill.
R1:	Ok, the green cup is Bill. (point to the inferred object)
H2:	Do you see a car?
R2:	Yes, I see a car.
H3:	Do you see a blue can behind it?
R3:	I don't see a blue can. But I see a blue unknown object there. Is that the can you are talking about? (point to the inferred object)
H4:	Yes, it's called Mary.
R4:	Got it, the blue can is Mary.

(b) Sample Dialogue

Figure 1: An example of situated setup and human-robot dialogue

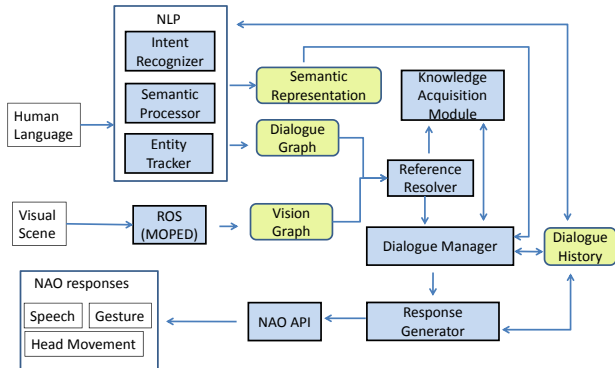


Figure 2: The overall architecture

on their mutual understanding of their intentions, goals, and current tasks [4]. The process to establish common ground is called *grounding*. Clark and colleagues have further characterized this process by defining *contributions*, the basic elements of a conversation [6]. Each contribution has two phases: a *presentation* phase where the speaker presents an utterance to the addressee and an *acceptance* phase where the addressee provides evidence of understanding to the speaker. This grounding process is viewed as a collaborative process where the speaker and the addressee make extra efforts to collaborate with each other to reach a mutual understanding [7]. The notion of common ground and communication grounding have been investigated extensively from human-human communication [5, 10], to computational models for spoken dialogue systems [30], and more recently to human-robot interaction [12, 18, 27, 28]. In particular, the recent work that is most relevant here has investigated symbol grounding [17] and feedback and adaptation [2] in human-robot dialogue.

Different from the above previous works, this paper presents a human-robot dialogue system with a specific focus on collaborative efforts. To infer the common ground, *shared bases* such as joint personal experiences, events, or episodes play an important role [4]. However, in situated human-robot dialogue, the human and the robot have disjoint perceptual experiences although they are physically co-present. Therefore, this paper intends to demonstrate how the mismatched perceptual basis may affect the joint tasks in human-robot dialogue, and how collaborative efforts may influence dialogue grounding.

3. HUMAN-ROBOT DIALOGUE SYSTEM

Suppose a human and a robot are co-present as shown in Figure 1(a). The human needs to communicate the secret names of several objects to the robot so that the robot knows which object has what name. Figure 1(b) shows a sample dialogue between the human and the robot. As shown in the example, instead of directly saying “the blue can is called Mary”, the human starts with *the car* (i.e., H_2) (right in front of the robot) and makes sure the robot recognizes it and from there goes to the target object (H_3 , H_4). This is one type of collaborative behavior referred to as *installment* [6]. The robot accepts the human’s presentation by pointing to the intended object. In addition, the robot also describes to the human what it perceives from the environment (e.g., R_3). Through these mechanisms, the human and the robot strive to reach a common ground about the names of objects in the shared world. This is the kind of dialogue our system aims to support. The overall architecture is shown in Figure 2. Next we illustrate several key components of the system.

3.1 NLP Modules

Human language (i.e., the output from an automatic speech recognizer) is processed by a set of NLP modules to create a semantic representation that captures the meaning of each human utterance (e.g., intent of the utterance, focus of attention, etc.) as shown in Figure 3.

More specifically, an **Intent Recognizer** is applied to identify the human intent (represented by **Act**) such as whether it is a request for the robot to identify certain objects (i.e., **Act:RequestIdObject**) or an acceptance of the robot’s presentation (i.e., **Act:Accept**). Currently, we only implemented a simple rule-based approach for intent recognition. Since our domain is quite narrow (mainly involving negotiation about the shared environment), it appears sufficient.

Besides human intent, a **Semantic Processor** is applied to identify the focus of attention (represented by **Focus**), as well as linguistic entities (i.e., **Refs**) and their properties mentioned in the utterance (i.e., the corresponding logic forms). Since human utterances can be informal, fragmented, and contain various types of disfluencies, we apply a partial parser based on combinatory categorial grammar (CCG) [26]. We have defined a set of basic CCG lexicon rules, which covers key vocabulary in our domain, such as object colors, shapes, spatial relations and so forth. Given a human utterance, our CCG parser repeatedly searches for the longest sentence covered by the grammar until the end of the utterance. As a result, the **Semantic Processor** creates a list of

H1: Human Act: DescribeObjectName [Type: Statement]; Focus: x1 Refs: ['x1'] Sem: Color(x1,green) ^ Type(x1,cup) ^ Name(x1,bill)	H2: Human Act: RequestIdObject [Type: Question(y/n)]; Focus: x2 Refs: [x2] Sem: Type(x2,car)	H3: Human Act: RequestIdObject [Type: Question(y/n)] Focus: x3 Refs: [x3, x4 x4->y2] Sem: Color(x3,blue) ^ Type(x3,can) ^ Spatial-Behind(x3,x4)	H4: Human Act: Accept [Type: Explicit(yes)]; Act: DescribeObjectName [Type: Statement] Focus: x5 Refs: [x5 x5->y3] Sem: Name(x5,Mary)
R1: Robot Act: Accept [Type: Explicit]; Act: DescribeObjectName [Type: Statement] Focus: y1 Refs: [y1 y1->x1] Sem: Color(y1,green) ^ Type(y1,cup) ^ Name(y1,bill)	R2: Robot Act: Accept [Type: Explicit(yes)]; Act: DescribeObjectProperty [Type: Statement] Focus: y2 Refs: [y2 y2->x2] Sem: Type(y2,car)	R3: Robot Act: Reject [Type: Explicit]; Act: DescribeObjectProperty [Type: Statement]; Act: RequestConfirm [Type: Explicit] Focus: y3 Refs: [y3 y3->x3] Sem: Color(y3,blue)	R4: Robot Act: Accept [Type: Explicit]; Act: DescribeObjectName [Type: Statement] Focus: y4 Refs: [y4 y4->x5] Sems: Type(y4,can) ^ Color(y4,blue) ^ Name(y4,Mary)

Figure 3: Examples of semantic representations

linguistic entities introduced by the utterance, and a list of first order predicates specifying the properties and relations between these entities.

An **Entity Tracker** is used to manage the linguistic entities from the conversation discourse. It determines whether a linguistic entity is newly introduced into the discourse or co-refers to an entity that already exists in the discourse. To arrive at such a decision, the **Entity Tracker** combines linguistic information (e.g., the types of linguistic expressions), recency information (e.g., how long ago linguistic entities are introduced into the discourse), and semantic compatibility between various entities to make a prediction. For example, for H_3 in Figure 3, the linguistic entity x_4 (introduced by the pronoun “it”) is predicted to corefer with a linguistic entity created by the robot (i.e., y_2 introduced by the noun phrase “a car” in R_2).

3.2 Reference Resolution

Through entity tracking, the coreferring entities (since they all refer to a same object in the physical world) are merged together to form a node in the **Dialogue Graph** as shown in Figure 4(a). The attributes of a node capture semantic properties of the corresponding linguistic entity. Each edge in the **Dialogue Graph** captures the semantic relations between linguistic entities.

The visual scene received by the robot is processed by a real-time object recognition system MOPED (<http://personalrobotics.ri.cmu.edu/projects/moped.php>). For example, Figure 4(b) shows what is perceived by the robot from the shared environment. The processing results are captured by the **Vision Graph** as shown in Figure 4(c). The **Vision Graph** captures the robot’s internal representation of the shared environment. Each node in the vision graph represents a perceived object. The lower level visual properties such as color distributions, size of bounding boxes, etc. are captured as attributes of the nodes.

Given these two graphs, a **Reference Resolver** applies an inexact graph-matching algorithm to match the dialogue graph to the vision graph to infer what objects in its own representation of the world are referred to or described by the human during conversation. Different matching algorithms and issues related to graph-matching have been described in our previous work [14, 15].

3.3 Knowledge Acquisition

Since the dialogue graph represents the human’s knowledge of the shared world and the vision graph represents the robot’s own internal representation of the perceived world, graph matching allows to bridge the gap between the symbolic representations and lower level numerical features. Based on such matching, a **Knowledge Acquisition Module** attempts to acquire knowledge from the matched nodes and relations to enrich the robot’s own representation of the environment. However, graph-matching alone is not sufficient. The robot needs to acquire *correct* descriptions, in other words, descriptions that are mutually agreed upon. Therefore, the **Knowledge Acquisition Module** not only takes the results from the **Reference Resolver**, but also utilizes the **Dialogue Manager** to arrange *confirmation* sub-dialogue with the human. Only when the confirmation is accepted by the human partner, is the corresponding symbolic knowledge added to the robot’s internal representation. Because of such confirmation, the knowledge acquired is considered *common ground knowledge*. For example, in Figure 5, the name, color, and type information concerning n_5 and n_7 is acquired and added to the robot’s internal representation. In addition, the symbolic description *Behind* is also added to the representation to describe the spatial relation between n_7 and n_0 . Through this way, the robot acquires knowledge through dialogue to build up the common ground.

3.4 Dialogue Management

The goal of the **Dialogue Manager** is to decide what actions the robot should take during conversation. Our **Dialogue Manager** has three parts: a representation of the state of the dialogue; an action space for the robot; and a dialogue policy that describes which action to take under which state.

The dialogue states are characterized based on human intent, focus of attention, graph matching results from the **Reference Resolver**, previous actions of the robot maintained in the **Dialogue History**, and pending knowledge to be acquired. The current space of actions allow the robot to greet, explicitly or implicitly confirm a request, accept or reject a presentation from its partner, describe what it perceives, ask for confirmation, and ask for specific information. Our dialogue policy consists of 17 (dialogue state, action) pairs. At execution time, the dialogue manager will go through the dialogue policy and identify a dialogue state

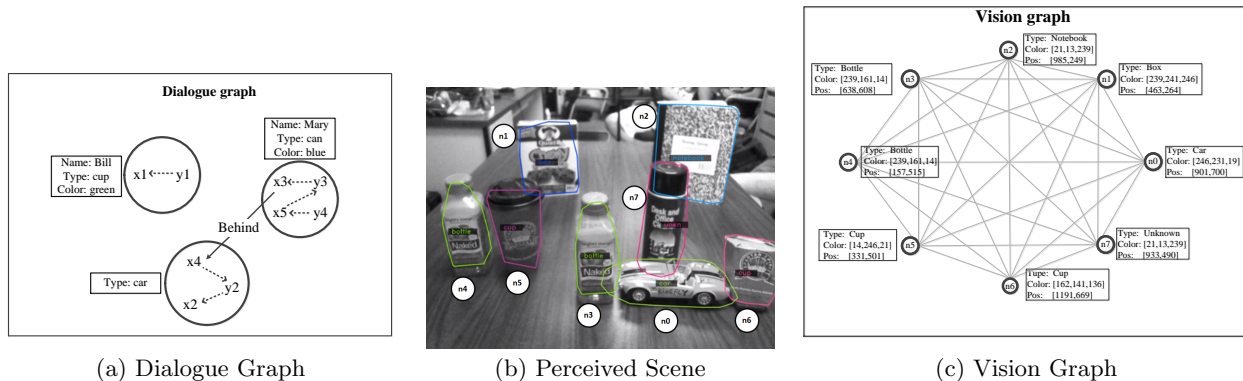


Figure 4: Examples of graph representation

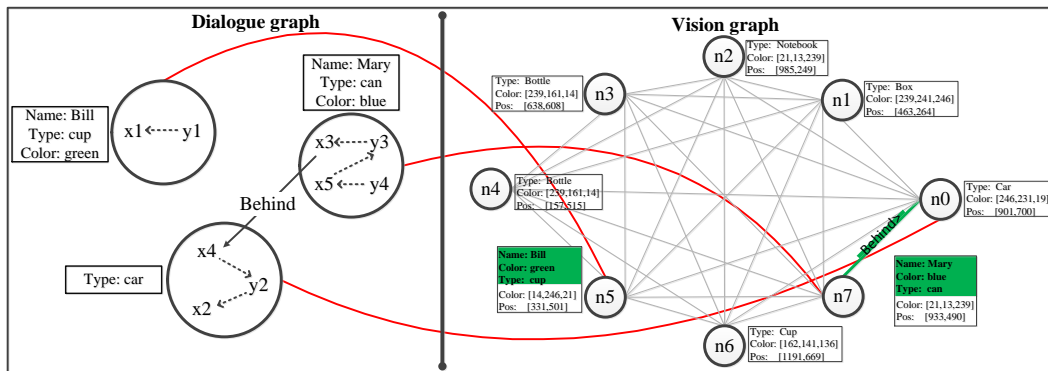


Figure 5: An example of knowledge acquisition through graph matching

that best matches the current state and then take the corresponding actions.

3.5 Response Generation

Once an action is chosen by the Dialogue Manager, the Response Generator decides how to realize such action using speech, gesture, and head movement. Non-verbal modalities such as gesture and gaze have been investigated extensively in HRI literature, for example, for engagement [1, 19, 23], and turn taking and participation [16]. These are all very important aspects in human-robot dialogue. Currently, we have not addressed these issues in our robot. Similar to previous work [25], we only use hand gesture and head movement to indicate the inferred object (i.e., the object understood by the robot) and the intended object (i.e., the object the robot describes). Occasionally, we also use head nodding or shaking to accept or reject a presentation from the human. For speech responses, several templates are predefined which support parameter passing at the run time. The correspondence between an action and a combination of speech, gesture, and head movement are predefined as a behavior policy. During the run time, given an action passed by the Dialogue Manager, the Response Generator will check the policy and generate the specified movements.

4. EMPIRICAL STUDIES

Using the dialogue system described above, we conducted an empirical study. The goal was to examine how the robot’s

collaborative effort may help mediate a shared perceptual basis and establish a common ground.

4.1 Design Factors

Our experiments were designed to take into consideration the following three main issues:

Shared Perceptual Basis. This factor aims to address to what degree mismatched perceptions between humans and robots will affect communication grounding in human-robot dialogue. More specifically, we selected a collection of everyday objects which were easily recognizable by human subjects for our experiments. We trained our robot to recognize these objects. By controlling the recognition results, we were able to manipulate two levels of mismatched perceptual basis:

- **Low-Mismatch** situation, where the robot can recognize most of the objects in the shared environment as its human partner. Either 10% or 30% of the objects in the scene are not recognizable to the robot. Thus the perceptual differences between the human and the robot are low.
- **High-Mismatch** situation, where the robot can *not* recognize most of the objects (either 60% or 90% of the objects are not recognizable). Thus the perceptual differences between the human and the robot are high.

Robot’s Collaborative Effort. Our previous work based on human-human dialogue [15] has shown that, when the shared perceptual basis is missing, conversation partners

Confidence	Low Effort	High Effort
[80%..1]	Accept "Got it"	Accept and Describe "Sure, the orange bottle is Eric"
[25%..80%]	Weak Accept "I think I see it, is that correct?"	Weak Accept and Describe "I see something there, is that orange bottle Eric? "
[0..25%]	Reject "I don't see it"	Reject and Describe "I don't see it, but I see an orange bottle there"

Table 1: Definitions of two levels of collaborative effort

tend to make extra effort to collaborate with each other to mediate shared basis. In particular, a prominent collaborative strategy adopted by the robot players (with simulated lower perceptual capability) is to proactively present what he/she perceives from the environment. Such presentation often provides an anchor to his/her partner (with higher perceptual capability) to further expand towards a mutual understanding. Motivated by our earlier findings, we are interested in whether and to what degree the extra effort made by the robot may help with communication grounding. Therefore, in our experiments, we specifically modeled two levels of collaborative effort associated with the robot’s acceptance:

- **Low-Effort** represents the minimum effort taken by the robot. It only accepts or rejects a presentation from its human partner explicitly, nothing more.
- **High-Effort** represents an extra collaborative effort taken by the robot. After it accepts or rejects a presentation, the robot proactively describes what it perceives from the environment based on its internal representation of the inferred object.

The decision on whether to accept or to reject is made by the **Dialogue Manager** based on the robot’s confidence at identifying the intended object (as a result of graph-matching). Table 1 shows the definition of these two levels of effort, together with examples of the robot’s verbal responses. These verbal responses are also accompanied by pointing gestures to the inferred or intended objects.

Automated NLP Processing. The ability to reach a common ground largely depends on how well the robot understands human language. To examine how automated NLP performance may affect dialogue grounding, we also experimented with two different settings:

- **Automatic Run** refers to the condition where all NLP modules are fully automated (as shown in Figure 2) during dialogue.
- **Wizard Run** refers to the condition where the results from NLP modules are verified and possibly updated by a human wizard to ensure correct interpretation.

In both settings, the downstream modules in Figure 2 remain fully automated.

4.2 Experimental Tasks

We designed a naming game for our experiments. Similar to the set up shown in Figure 1(a) a human subject and our NAO robot were positioned in front of a table which held ten everyday objects. The subject was given a printout of the environment at the beginning of each trial/task. The

printout showed the secret names of six objects in the environment. The human subject’s goal was to communicate these names to the robot so that the robot became aware which object has what name. The subject was also given a pen and instructed to keep track of his/her progress on the sheet. The task is considered *completed* if the subject **believes** he/she has successfully communicated these names to the robot and he/she and the robot have reached a mutual understanding about the object names. In other words, the task is considered completed if the human subject believes he/she and the robot have come to a common ground. So this is a subjective measurement, which we call *perceived task completion*, *perceived common ground*, or *perceived dialogue grounding*.

To avoid interaction deadlock that may result from errors in automated processing, we also defined a *task abort* criterion. During interaction, if the robot failed to acquire any knowledge (i.e., no knowledge updating) as described in Section 3.3 for three minutes, the task would be aborted and considered *incomplete*.

In each trial, the subject was instructed to keep his/her dialogue going with the the robot until either the subject believed the task was completed or the on-going dialogue met the abort criterion.

4.3 Procedures

The perceptual differences were implemented by four levels: 10% and 30% unrecognizable objects for Low-Mismatch, and 60% and 90% for High-Mismatch. To create an experimental configuration, we combined two of the four levels with a low collaborative effort and the other two with a high collaborative effort. For example, one configuration was 10% with a low effort, 30% with a high effort, 60% with a low effort, and 90% with a high effort. Thus there were six different configurations in total. Each of these configurations specified four treatment conditions (i.e., a combination of perception and effort) assigned to one subject.

We also designed four scenes with ten objects each where six objects were assigned secret names. These four scenes were independent from each other (e.g., an apple in two different scenes would have two different names). To each subject, an experimental configuration was first assigned. Each of the four treatment conditions specified in the configuration was randomly assigned to each of the four scenes. The objects from each scene that were deemed un-recognizable were randomly chosen based on the distribution specified by the perception factor.

We recruited a total of 24 human subjects from the MSU campus to participate in our study. These subjects were not aware of any part of the system or the study design. During the experiments, each subject, based on the assigned configuration, performed four naming tasks for four scenes. The order of the scenes was randomized. These four tasks were first performed under the condition of *Automatic Runs*, meaning that the human interacted with a fully automated robot¹. After the automatic runs, if there were any aborted tasks, we conducted further experiments under the condition of *Wizard Run*. We asked the subjects to take a break and

¹Speech recognizers (e.g., Sphinx 4) are integrated in our robot. However, to focus our current effort on language understanding and dialogue grounding, in our user studies, human speech was transcribed on the fly by a researcher behind the scene.

	Low-Mismatch		High-Mismatch	
	L-Effort	H-Effort	L-Effort	H-Effort
Num. of Dialogues	24	24	24	24
Perceived Grounding	17	6	14	6
Rate	0.708	0.25	0.583	0.25

Table 2: Rate of perceived common ground

then repeated those aborted tasks in a random order with the same subject. The wizard’s responses were reasonably fast. We did not observe any noticeable delay.

4.4 Empirical Results

Our experiments resulted in a total of 96 dialogues with the fully automated robot (i.e., under the condition of *Automatic Run*). Among these, 43 dialogues led to *perceived task completion* or *perceived common ground*. The remaining 53 incompleting tasks were performed again with the help from the wizard, which led to an additional 53 dialogues under the condition of *Wizard Run*.

4.4.1 Perceived Common Ground

Table 2 shows the rate of perceived grounding for all 96 dialogues during automatic runs. There is a significant difference among different settings ($\chi^2 = 15.9649$, $p = 0.001$). Interestingly, the dialogues where the robot made an extra collaborative effort (i.e., *H-Effort*) have a lower rate compared to those where the robot made a low effort (i.e., *L-Effort*). A Tukey post-hoc test reveals that such difference is significant for the *Low-Mismatch* setting (at $p < 0.05$) and marginal for the *High-Mismatch* setting.

Although it may appear counter-intuitive at first glance, this result demonstrates the role of collaborative effort in the perception of common ground. Note that the *perceived common ground* is defined from the human subject’s perspective concerning whether he/she *believes* a common ground is reached. In the case of *L-Effort*, a simple acceptance from the robot (without further describing its internal states) could often lead the human subjects to believe that a mutual understanding has been reached. This is especially the case since the space is crowded and pointing to or looking at an object sometimes cannot be precise. In the case of *H-Effort*, the robot also proactively describes its internal representation of the inferred object. This extra information provides feedback to the human partners on what exactly is grounded. This information can potentially reveal incorrect grounding and thus lower the human’s beliefs of mutual understanding.

4.4.2 True Common Ground

A perceived common ground does not mean a true common ground. In the context of our naming game, a true common ground is where the name acquired for each object by the robot at the end of dialogue is indeed the correct name intended by the human for that object. In the best case, the robot should acquire all six names correctly to reach a common ground with its partner. Thus we take a closer look at those dialogues with perceived common ground. These include 43 out of 96 dialogues from the automatic runs and 28 out of 53 dialogues from the wizard runs.

First, we examine the average number of turns in these dialogues as shown in Table 3. For automatic runs (the row marked with “A”), a two-way ANOVA did not find any significant main effects from mismatched perception, collaborative effort, or their interaction. However, for the wiz-

	Low-Mismatch		High-Mismatch	
	L-Effort	H-Effort	L-Effort	H-Effort
A	29.12 ± 10.81	28.33 ± 7.94	31.43 ± 11.08	29.50 ± 8.94
W	17.50 ± 1.00	21.82 ± 4.77	23.60 ± 8.65	25.78 ± 6.28

Table 3: The average number of turns from dialogues with perceived common ground

	Low-Mismatch		High-Mismatch	
	L-Effort	H-Effort	L-Effort	H-Effort
A	3.47 ± 1.13	3.67 ± 1.03	2.36 ± 1.39	3.67 ± 1.37
W	4.75 ± 0.50	4.36 ± 0.51	3.25 ± 1.26	4.50 ± 0.71

Table 4: The average number of correctly acquired names from dialogues with perceived common ground

ard runs (the row marked with “W”), a two way ANOVA found a significant main effect of mismatched perception ($F(1, 25) = 4.59$, $p < 0.05$, partial $\eta^2 = 0.16$) on the number of turns. This indicates that a higher perceptual mismatch requires a larger number of dialogue turns in order to reach a perceived common ground.

Next we examine the average number of correctly acquired names (i.e., an objective measure of common ground) for those dialogues as shown in Table 4. For the automatic runs, through a two-way ANOVA, the main effect of collaborative effort on true common ground is marginal ($F(1, 39) = 3.17$, $p = 0.083$, partial $\eta^2 = 0.08$). For the wizard runs, a two-way ANOVA found a significant main effect of perception ($F(1, 25) = 5.363$, $p = 0.029$, partial $\eta^2 = 0.18$) on true common ground. We also found a significant interaction of perception and effort ($F(1, 25) = 7.723$, $p = 0.010$, partial $\eta^2 = 0.236$). The interaction plot is shown in Figure 6. For the low-mismatched perception, whether a low effort or a high effort is applied does not seem to make much difference. However, for the high-mismatched perception, a high effort leads to much better grounding performance than a low effort. This result indicates that while a low effort may work fine when the perceptual basis is mostly aligned, it will not be sufficient when there is a significant perceptual difference between the human and the robot.

4.4.3 Automatic NLP Processing

We conducted a further analysis on how the automatic NLP performance may affect dialogue grounding. Particularly, we are interested in semantic processing and entity tracking since their performance determines the quality of dialogue graphs, which further links to graph matching, essential to our task on grounding names. Among 96 dialogues, the overall accuracy for semantic processing is 81.4%. This means that 81.4% of properties associated with linguistic entities extracted from human utterances are correct. The accuracy on entity tracking is 64.6%. This means that, given a linguistic entity from a human utterance (a total of over 1600 in our data), 64.5% of time the **Entity Tracker** can correctly predict whether it introduces a new entity or it corefers with an existing entity. The performance is further broken down as shown in Table 5.

There is no significant difference in semantic processing performance across different conditions ($\chi^2 = 4.303$, $p = 0.231$). However, there is a significant difference regarding entity tracking performance ($\chi^2 = 20.889$, $p < 0.001$). In particular, a Tukey’s post-hoc comparison test indicates that

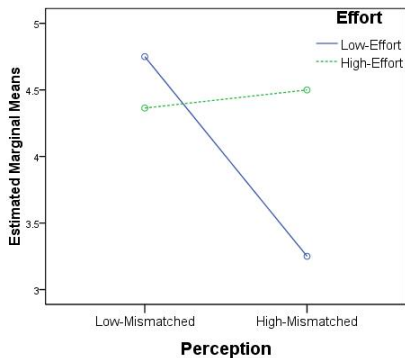


Figure 6: Interaction plot between perception and effort

	Low-Mismatch		High-Mismatch	
	L-effort	H-effort	L-effort	H-effort
semantic processing	0.780	0.814	0.835	0.826
entity tracking	0.651	0.605	0.742	0.605

Table 5: Performance of automated NLP processing

under the high mismatched perception, the entity tracking component performed significantly worse in the high-effort setting compared to the low-effort setting (at $p < 0.05$). This is mainly because in the high-effort setting, the robot proactively describes what it perceives from the environment. Therefore, the robot can introduce new entities into the discourse. So a linguistic entity now can also refer to the entities introduced by the robot (as opposed to the entities introduced only by the human as in the low-effort case). This may add some complexity to our simple entity tracking algorithm. We are currently developing better algorithms for entity tracking based on the collected data.

To understand how much NLP performance may affect dialogue grounding, we compared the 53 incompleting tasks under the condition of *Automatic Run* with the same tasks under the condition of *Wizard Run*. We examine the average number of correctly acquired names from these tasks as shown in Figure 7. It shows a significant main effect of semantic interpretation on grounding ($F(1, 98) = 33.25$, $p < 0.0001$, partial $\eta^2 = 0.253$). A follow up T-test has shown that, wizard runs significantly outperform the automatic runs under both the low mismatch ($t = -4.231$, $p < 0.0001$, two-tailed) and the high mismatch ($t = -4.608$, $p < 0.0001$, two-tailed) perception, as well as the low collaborative ($t = -2.578$, $p = 0.015$, two-tailed) and the high collaborative ($t = -5.970$, $p < 0.0001$, two-tailed) effort.

These results confirm that better NLP performance will lead to better common ground. In addition, since the average number of correctly acquired names from the wizard runs (with close-to ground-truth NLP processing) is still short from being perfect (i.e., 6), these results further emphasize the need to improve other modules such as dialogue management and knowledge acquisition.

5. DISCUSSION

It has been long established that common ground is key to communication success. A significant body of work has investigated common ground in human-human communication, for example, how the shared culture background [4] and

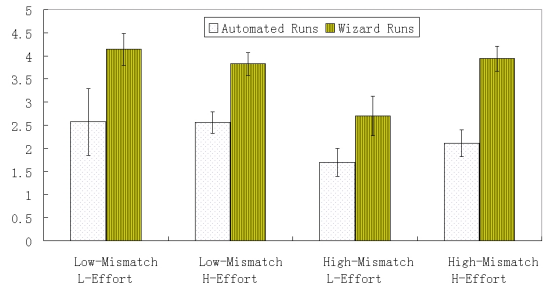


Figure 7: Average number of correctly acquired names from automatic runs and wizard runs

spatial reasoning capabilities [22] affect human-human communication. In spoken dialogue systems, previous work has addressed how automated speech recognition and language understanding affect common ground [11]. Compared to human-human communication or spoken dialogue systems, the issue of common ground is more complex in human-robot dialogue. This is because it involves an additional layer of common ground - a lower level that concerns perception (e.g., visual and tactile) and symbol grounding. Although fundamental to human-robot dialogue, this level generally is not concerned in human-human communication. Thus, as in recent work [2, 17], this paper pays a special attention to this lower level of common ground.

One of the important characteristics of situated interaction is the embodiment. Humans and robots have physical bodies in the shared environment and they can presumably utilize non-verbal modalities and actions to communicate and establish common ground. However, our current work has focused on mediating a shared visual perceptual basis mainly using language. Especially, we have not looked into collaborative non-verbal behaviors such natural gestures or actions (e.g., pointing gestures and picking up an object to teach the robot different attributes) from human partners. These actions will be even more important for a dynamic environment. Our future work will address this limitation by integrating embodiment with language based communication and investigate its role in establishing common ground.

The acquired common ground knowledge will provide a basis for follow-up communications. Furthermore, as knowledge accumulates, it will allow the robot to automatically learn new models, for example, word models for grounded semantics [8]. Our current work has only focused on acquiring knowledge towards common ground through collaborative dialogue. We have not explored the direction where the robot actively learns about the environment [3, 29]. Our future work will extend the current models to address robot learning through dialogue to establish common ground.

6. CONCLUSION

This paper describes a dialogue system that attempts to mediate a shared perceptual basis between the human and the robot through automatic knowledge acquisition. As conversation proceeds, the robot first matches human descriptions to its internal representation of the shared world. It then automatically acquires and confirms through dialogue common ground knowledge about the shared environment. The acquired knowledge is used to enrich the robot’s representation of the shared world. In particular, this paper examines the role of the robot’s collaborative effort in this

grounding process. Our results have shown that an extra effort from the robot to make its human partner aware of its internal representation of the shared world contributes to better common ground. Our future work will build upon the current system and integrate non-verbal behaviors and active learning for common ground knowledge acquisition.

7. ACKNOWLEDGMENTS

This work was supported by N00014-11-1-0410 from the Office of Naval Research and IIS-1208390 from the National Science Foundation.

8. REFERENCES

- [1] D. Bohus and E. Horvitz. Models for multiparty engagement in open-world dialog. In *Proceedings of the SIGDIAL 2009 Conference*, pages 225–234, 2009.
- [2] H. Buschmeier and S. Kopp. Co-constructing grounded symbols - feedback and incremental adaptation in human-agent dialogue. *Künstliche Intelligenz*, 27:137–143, 2013.
- [3] M. Cakmak and A. L. Thomaz. Designing robot learners that ask good questions. In *Proceedings of the HRI12*, pages 17–24, 2012.
- [4] H. H. Clark. *Using language*. Cambridge University Press, Cambridge, UK, 1996.
- [5] H. H. Clark and S. E. Brennan. Grounding in communication. In L. B. Resnick, R. M. Levine, and S. D. Teasley, editors, *Perspectives on socially shared cognition*, pages 127–149. 1991.
- [6] H. H. Clark and E. F. Schaefer. Contributing to discourse. In *Cognitive Science*, number 13, pages 259–294. 1989.
- [7] H. H. Clark and D. Wilkes-Gibbs. Referring as a collaborative process. In *Cognition*, number 22, pages 1–39. 1986.
- [8] R. Fang, C. Liu, and J. Y. Chai. Integrating word acquisition and referential grounding towards physical world interaction. In *Proceedings of the 14th ACM International Conference on Multimodal Interaction, ICMI '12*, pages 109–116, 2012.
- [9] B. Fransen, V. Morariu, E. Martinson, S. Blisard, M. Marge, S. Thomas, A. Schultz, and D. Perzanowski. Using vision, acoustics, and natural language for disambiguation. In *Proceedings of the HRI07*, pages 73–80, 2007.
- [10] D. Gergle, R. E. Kraut, and S. R. Fussell. Using visual information for grounding and awareness in collaborative tasks. *Human Computer Interaction*, 28:1–39, 2013.
- [11] J. Gordon, R. Passonneau, and S. Epstein. Learning to balance grounding rationales for dialogue systems. In *Proceedings of the SIGDIAL 2011 Conference*, pages 266–271, 2011.
- [12] S. Kiesler. Fostering common ground in human-robot interaction. In *Proceedings of the IEEE International Workshop on Robots and Human Interactive Communication (ROMAN)*, pages 729–734, 2005.
- [13] G.-J. M. Kruijff, P. Lison, T. Benjamin, H. Jacobsson, and N. Hawes. Incremental, multi-level processing for comprehending situated dialogue in human-robot interaction. In *Symposium on Language and Robots*, 2007.
- [14] C. Liu, R. Fang, and J. Y. Chai. Towards mediating shared perceptual basis in situated dialogue. In *Proceedings of the SIGDIAL 2012 Conference*, pages 140–149, 2012.
- [15] C. Liu, R. Fang, L. She, and J. Chai. Modeling collaborative referring for situated referential grounding. In *Proceedings of the SIGDIAL 2013 Conference*, pages 78–86, Metz, France, 2013.
- [16] B. Mutlu, T. Shiwa, T. Kanda, H. Ishiguro, and N. Hagita. Footing in human-robot conversations: How robots might shape participant roles using gaze cues. *Proceedings of the HRI09*, pages 61–68, 2009.
- [17] J. Peltason, H. Rieser, S. Wachsmuth, and B. Wrede. On grounding natural kind terms in human-robot communication. *Künstliche Intelligenz*, 27:107–118, 2013.
- [18] A. Powers, A. Kramer, S. Lim, J. Kuo, S.-L. Lee, and S. Kiesler. Common ground in dialogue with a gendered humanoid robot. In *Proceedings of ICRA*, 2005.
- [19] C. Rich, B. Ponsleur, A. Holroyd, and C. L. Sidner. Recognizing engagement in human-robot interaction. In *Proceedings of the HRI10*, pages 375–382, 2010.
- [20] P. E. Rybski, K. Yoon, J. Stolarz, and M. M. Veloso. Interactive robot task training through dialog and demonstration. In *Proceedings of the HRI07*, pages 49–56, 2007.
- [21] M. Scheutz, P. Schermerhorn, J. Kramer, and D. Anderson. Incremental natural language processing for hri. In *Proceedings of the HRI07*, 2007.
- [22] M. F. Schober. Spatial dialogue between partners with mismatched abilities. In *Spatial Language and Dialogue*, pages 23–39. Oxford University Press, 2009.
- [23] C. Sidner, C. Lee, C. D. Kidd, N. Lesh, and C. Rich. Explorations in engagement for humans and robots. In *Artificial Intelligence*, volume 166(1-2), pages 140–164, 2005.
- [24] R. Stalnaker. Common ground. *Linguistics and Philosophy*, 25:701–721, 2002.
- [25] M. Staudte and M. W. Crocker. Visual attention in spoken human-robot interaction. In *Proceedings of the HRI09*, pages 77–84, 2009.
- [26] M. Steedman and J. Baldrige. Combinatory categorial grammar. *Non-Transformational Syntax Oxford: Blackwell*, pages 181–224, 2011.
- [27] K. Stubbs, P. Hinds, and D. Wettergreen. Autonomy and common ground in human-robot interaction. In *IEEE Intelligent Systems*, pages 42–50, 2007.
- [28] K. Stubbs, D. Wettergreen, and I. Nourbakhsh. Using a robot proxy to create common ground in exploration tasks. In *Proceedings of the HRI08*, pages 375–382, 2008.
- [29] S. Tellex, R. Deits, P. Thaker, D. Simeonov, T. Kollar, and N. Roy. Clarifying commands with information-theoretic human-robot dialog. *Journal of Human-Robot Interaction*, 1:78–95, 2012.
- [30] D. Traum. *A Computational Theory of Grounding in Natural Language Conversation*. PhD thesis, University of Rochester, 1994.