

# Enabling Robots to Understand Indirect Speech Acts in Task-Based Interactions

Gordon Briggs

NRC Postdoctoral Fellow, U.S. Naval Research Laboratory

Tom Williams

Human-Robot Interaction Laboratory, Tufts University

and

Matthias Scheutz

Human-Robot Interaction Laboratory, Tufts University

---

An important open problem for enabling truly taskable robots is the lack of task-general natural language mechanisms within cognitive robot architectures that enable robots to understand typical forms of human directives and generate appropriate responses. In this paper, we first provide experimental evidence that humans tend to phrase their directives to robots indirectly, especially in socially conventionalized contexts. We then introduce pragmatic and dialogue-based mechanisms to infer intended meanings from such *indirect speech acts* and demonstrate that these mechanisms can handle all indirect speech acts found in our experiment as well as other common forms of requests.

*Keywords:* human-robot dialogue, human perceptions of robot communication, robot architectures, speech act theory, intention understanding

---

## 1. Introduction

Two key challenges at the intersection of artificial intelligence (AI), robotics, and human-robot interaction (HRI) need to be addressed in order to enable truly taskable robots: (1) the capability challenge of developing robotic agents that are able to both algorithmically and physically perform the desired tasks, and (2) the interaction challenge of developing agents that can be instructed by humans through *natural language* (NL) in natural and intuitive ways (e.g. Scheutz, Schermerhorn, Kramer, & Anderson, 2007) to perform the desired tasks and appropriately respond to these instructions. We will focus exclusively on the second challenge.

Specifically, we will focus on how robots can understand *directives*: utterances issued with the intention that the addressee will perform some task for the speaker. For example, for an utterance intended as a *question*, the speaker intends the addressee to provide an informative response; for an utterance intended as a *command*, the speaker intends the addressee to perform some general action. Several capabilities are required to understand even the simplest of directives, from *speech recognition*, to *syntactic and semantic analysis*, to *pragmatic understanding*. What is more, robotic

---

Authors retain copyright and grant the Journal of Human-Robot Interaction right of first publication with the work simultaneously licensed under a Creative Commons Attribution License that allows others to share the work with an acknowledgement of the work's authorship and initial publication in this journal.

architectures that provide these capabilities must take into account the many *social norms* that NL-enabled agents are expected to obey. For example, in order to adhere to social norms such as *politeness*, people frequently use so-called *indirect speech acts* (ISAs), in which the speech act’s *literal* meaning does not match its *intended* meaning. For example, since it is often rude to use a direct command such as “Bring me coffee,” one might instead use the indirect request “Could you get me a coffee?”, which is literally a request for information yet indirectly communicates the speaker’s true intention: a request to be brought coffee. In such cases, human listeners automatically and without contemplation understand the indirect interpretation to be the intended one. Hence, if humans were to use the same kinds of indirect speech acts with robots—as we will show—robots interacting with humans will need to take social norms into account if they are to properly understand the intentions implied by human utterances.

We believe that one of the primary obstacles for enabling truly interactive taskable robots is the lack of general, integrated, and architectural mechanisms for understanding the intentions behind directives, regardless of how these directives are expressed. In this paper, we present mechanisms integrated into a cognitive robotic architecture that make strides toward addressing this obstacle and demonstrate that the proposed mechanisms can handle all common indirect speech acts found in an empirical study specifically designed to probe the extent to which ISAs will actually be used in human-robot dialogue. To ensure terminological clarity, we will use the following definitions throughout the paper.

**Indirect speech act:** an utterance whose literal meaning does not match its intended meaning.  
**Direct speech act:** an utterance whose literal and intended meanings match.  
**Illocutionary point:** the category of an utterance, such as *statement*, *question*, *suggestion*, or *command*. An utterance has both a *literal* illocutionary point (which is directly reflected in the utterance’s form) and an *intended* illocutionary point. For direct speech acts, these match. For indirect speech acts, they may or may not.  
**Directive:** an utterance intended to cause the addressee to perform some action.  
**Direct request:** a *direct* directive whose literal illocutionary point is that of a question.  
**Direct command:** a *direct* directive whose literal illocutionary point is that of a command.  
**Indirect request:** any *indirect* directive. Thus, an indirect request is an indirect speech act with the literal illocutionary point of a statement, question, or suggestion, and the intended illocutionary point of a question or command.

The rest of the paper will proceed as follows. In Section 2, we begin by discussing the computational challenge of indirect speech act understanding. In Section 3, we then present the results of an HRI study intended to probe the extent to which humans will use indirect speech acts in their NL interactions with robots. The results of this experiment provide the first evidence for the need to develop mechanisms in cognitive robot architectures for handling indirect requests. After discussing these results, we use them to present design recommendations for robot architecture designers. In Section 4, we introduce an architectural framework that makes significant progress toward addressing the interaction challenge for taskable robots. Specifically, we discuss data representations and inference algorithms that use contextual knowledge to allow a robot to understand commands and indirect requests. In Section 5, we demonstrate the proposed mechanisms integrated in a cognitive robotic architecture and show how they handle the variety of utterances observed in our human-subject experiment. Finally, in Sections 6 and 7, we discuss the significance of our results and propose directions for future work.

## 2. Computational Motivation

The capabilities necessary for understanding indirect speech acts (or any other form of speech act) can be thought of as a subset of those necessary for achieving *common ground*, (i.e., mutual understanding, between interactants). Theoretical work in conversation and dialogue has construed the establishment of common ground as a multi-stage process (Clark, 1996). In the *attentional stage*, interactants must attend to each other in a conversational context. In the *perceptual stage*, the addressee must successfully perceive a communicative act directed to him/her by the speaker. In the *semantic understanding stage*, the addressee must abduce some literal meaning from this perceived act. Finally, in the *intentional understanding stage*, which Clark 1996 terms *uptake*, the addressee must abduce some intention from this literal meaning given the joint context.

While Clark’s multi-stage model of establishing common ground is valuable in conceptualizing the challenges involved, it can be even further refined. Schlöder 2014 proposes that uptake be divided into both weak and strong forms. *Weak uptake* can be associated with Clark’s intentional understanding process, whereas *strong uptake* denotes the stage where the addressee may either accept or reject the proposal implicit in the speaker’s action. A proposal is not strongly “taken up” unless it has been accepted as well as understood (Schlöder, 2014). This distinction is important as the addressee can certainly understand the intentions of an indirect request such as “Could you deliver the package?”, but this does not necessarily mean that the addressee will actually agree to the request and carry it out. In order for the proposal to be accepted, a set of *felicity conditions* must hold. For a directive, the addressee must have the *knowledge, capacity, obligation, and permission* necessary to carry out the intended action before he or she (or it) will accept that directive.

Uptake in human-robot interaction requires mechanisms that either implicitly or explicitly deal with the previous stages in Clark’s model of mutual understanding (attentional, perceptual, and semantic), and there have indeed been several prior attempts at implicitly enabling human-robot interaction at these levels. For instance, work at the attentional level includes the development of mechanisms to detect when an interactant is engaged with the robot (e.g. Rich, Ponsler, Holroyd, & Sidner, 2010). Work at the perceptual level includes projects seeking to enable robust speech and gesture recognition (e.g. Gomez, Kawahara, Nakamura, & Nakadai, 2012). Finally, work at the semantic understanding level includes developing mechanisms to tackle challenges, such as reference resolution (e.g. Tellex et al., 2013). A much smaller body of work, however, has focused on explicit intentional understanding in human-agent interactions (weak uptake), which we discuss next.

In order to handle common linguistic forms of directives, robot architectures require a number of additional mechanisms, as directives can come in both literal and non-literal forms. Often these non-literal forms (i.e., indirect requests) are considered to be tightly associated with their intended meanings, as is the case in *conventionalized* (or *idiomatic*) ISAs (Clark & Schunk, 1980; Searle, 1975). Conventionalized ISAs include questions used as *pre-requests* to remove potential obstacles toward a desired action or outcome (e.g., “Can I get a coffee?”) (Gibbs Jr, 1986) and assertions of needs or desires (e.g., “I would like a coffee”).

Some robot architectures have enabled ISA understanding by using rule-based systems that reason over conventionalized forms (Wilske & Kruijff, 2006). This approach stands in contrast to the plan-reasoning, or *inferential* approach, where each utterance is viewed as an action within the speaker’s larger dialogue plan (Perrault & Allen, 1980).

There are advantages and disadvantages to both strategies. The idiomatic approach, while less computationally expensive, can only detect and handle indirect speech acts that have (known) conventionalized forms. The inferential approach, while more general, requires computationally expensive goal and plan abduction mechanisms. As such, some researchers have proposed hybrid architectures, which combine the idiomatic and inferential forms of pragmatic reasoning (Briggs &

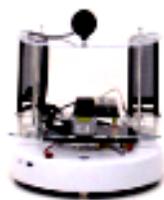
Scheutz, 2013; Hinkelman & Allen, 1989). These approaches first attempt to identify whether an utterance fits a conventionalized form given the current context. If the utterance does not fit any known conventionalized form, a more expensive plan-reasoning process is utilized.

The choice to develop ISA-understanding mechanisms presupposes that ISAs will be used in human-robot dialogues because they are used in human-human dialogues. However, it is not immediately obvious whether this assumption is warranted. It would be reasonable for one to suspect, for example, that humans might talk to robots the way they typically talk to dogs: by issuing simple, direct commands. This should be especially true for simple tasks that lack their own conventionalized social norms. And although we have some evidence from previous HRI studies suggesting that the opposite might be true—that people make frequent use of indirect speech acts when instructing robots, even in the simplest of tasks—there is currently no formal HRI study that verifies this transfer of human social norms to robot interactants.

In the next section, we thus present the results of a human-subjects experiment intended to investigate the extent to which humans will actually use ISAs when interacting with robots through natural language in task-based settings. The context of the experiment is a simple, novel task, for which conventionalized social norms do not exist beyond those of everyday life, and which is simple enough that it can be achieved solely through simple, direct commands. If, indeed, ISAs come so natural to people that they will even use them in a novel, non-conventionalized task as the one we will explore in the experiment, then this will provide firm evidence for the need to handle ISAs in robotic architecture capable of task-based natural language interactions with humans.

### 3. Experiment: ISA Use in Simple Task-Based Dialogues

In our experiment, we investigated whether subjects would use ISAs at all in a simple, novel task where the robot could be easily instructed using direct commands. Furthermore, we investigated whether subjects would use ISAs when the robot repeatedly demonstrated an inability to understand ISAs.



*Figure 1.* Augmented iRobot Create used in the experiment.

#### 3.1 Design

For this experiment, we chose one of the simplest tasks we could find in the HRI literature; a human instructor must command a robot to knock over colored towers built out of cans (Briggs & Scheutz, 2014). This is an interaction so simple that one would only expect direct language to be used by participants; the task can be easily accomplished using only instructions of the form “knock over the <color> tower.”

Participants were told that the experimenters were developing natural language interaction capabilities for robots and that their task would be to interact with a “tower-toppling robot.” Participants

were given a list of three towers (i.e., “Red tower,” “Yellow tower,” and “Blue tower”) and were told that, after being introduced to the tower-toppling robot, they were to command the robot to knock over those three towers, one at a time, in whatever order they wished.

After this briefing, the experimenter left the room through one door, and the tower-toppling robot (an iRobot Create, as seen in Fig. 1) entered through a different door and introduced itself. This robot was teleoperated by a trained confederate through a *Cognitive Wizard-of-Oz (WoZ)* interface (Baxter, Kennedy, Senft, Lemaignan, & Belpaeme, 2016) created with the ADE implementation (Scheutz et al., 2013) of the DIARC architecture (Schermerhorn, Kramer, Middendorff, & Scheutz, 2006), with video data being streamed to the confederate through a GoPro camera affixed to the robot. Utterances made by the robot were prerecorded using the open source MaryTTS text-to-speech package. A simple diagram of this setup can be seen in Fig. 2.

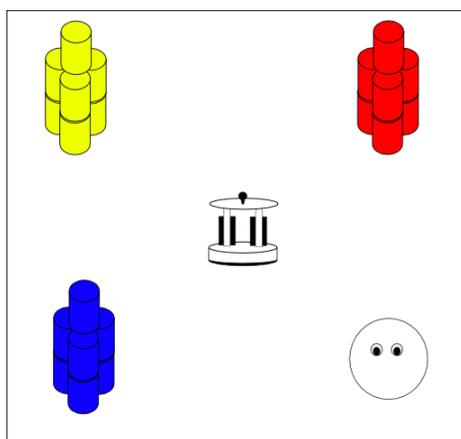


Figure 2. Diagram demonstrating the experimental setup and positioning of participant, robot, and towers in the tower-toppling experiment.

Participants were randomly assigned to two experimental conditions: an *understanding* condition and a *misunderstanding* condition, which determined how the robot responded to participants when they used indirect speech acts. In both conditions, if the participant used a direct speech act to command the robot (e.g., “Knock down the red tower”) or a bare noun phrase (e.g., “Red tower”), the robot would simply knock over the denoted tower. Similarly, in the understanding condition, the robot would also knock over the denoted tower if the participant used an indirect speech act (e.g., “Could you knock down the red tower?”). In the misunderstanding condition, however, if the participant used an indirect speech act, the robot took their utterance at face value and responded according to Table 1. This table shows the responses that are given for ISAs with different combinations of *direct illocutionary point*, *condition of focus*, and *direction of focus*, aspects derived from Searle’s *Speech Act Theory* (Searle, 1975, 1976). For example, “Could you knock down the red tower?” has the illocutionary point of a question, focuses on a preparatory condition (i.e., *ability* of the robot to perform the desired action), and focuses on the agent (i.e., the robot performing the action, as opposed to the tower that is the patient of the action).

### 3.2 Measures

In order to assess the extent to which participants used ISAs, participants’ utterances were recorded and transcribed; all *task-relevant utterances* were classified by independent annotators as either direct or indirect. Difference in rate of ISA use in task relevant utterances was analyzed using

Point	Cond	Dir	Example	Response
Q	P	A	“Could you $X$ ?”	“Yes, I am able to do that. Please tell me your order.”
S	S	A	“I need you to $X$ .”	“Thank you for sharing that interesting fact. Please tell me your order.”
S	P	A	“You can $X$ .”	“Thank you, but I am already aware of my capabilities. Please tell me your order.”
S[Su]	P	A	“You should $X$ .”	“Thank you for your suggestion. Please tell me your order.”
Q	P	P	“Could $X$ be knocked down?”	“Yes, that is permissible. Please tell me your order.”
S	S	P	“I’d like you to $X$ .”	“Thank you for sharing that interesting fact. Please tell me your order.”
S	P	P	“ $X$ will be knocked down.”	“Thank you for sharing that interesting prediction. Please tell me your order.”
S[Su]	P	P	“ $X$ should be knocked down.”	“Thank you for your suggestion. Please tell me your order.”

Table 1: Responses given for different categories of ISAs. Direct (Direct Illocutionary) Point: Q = Question, S = Statement, S[Su] = Suggestive Statements; Cond(ition): P = Preparatory, S = Sincerity; Dir(ection): A = Agent, P = Patient.

a Welch two-sample TOST with dialogue condition (understanding vs misunderstanding) as the independent variable. In addition, analyses of variance (ANOVAs) were used to rule out age and gender effects.

### 3.3 Population

Participants were recruited online and through flyers posted near a university campus. All participants (11 male, 13 female) were between the ages of 19 and 69 ( $M = 32.04, SD = 15.96$ ). Participants were paid \$10 each for their participation and provided informed written consent before beginning the experiment. While most participants were beyond college age, we asked them for their current or previous college major, if any. Three reported studying mathematics, computer science, or engineering; six reported studying a natural science or medicine; five reported studying a social science; four reported studying a branch of the arts or humanities; two reported studying some other field; and four reported no previous or current major.

### 3.4 Results

The demographic ANOVAs revealed no age ( $F(1, 46) = 2.24, p > .1$ ) or gender ( $F(1, 47) = 0.79, p > .3$ ) effects. Given that there were no gender effects, we can treat the subjects as a uniform group that has sufficient size for the purposes of this experiment.

We hypothesized that even in this simple scenario, some participants would use ISAs. In fact, over half of all participants ( $n=14$ ) used ISAs and 27.97% of all task-relevant utterances were ISAs. We found this striking, particularly because of the task’s lack of conventionalized social norms that would have necessitated the use of ISAs. What is more, participants used roughly this number of ISAs even when the robots repeatedly demonstrated an inability to understand ISAs: ISAs comprised 29.41% of task-relevant utterances in the understanding condition, and 26.87% of task-relevant utterances in the misunderstanding condition. Mean ISA use was found to be statistically equivalent using a Welch two-sample TOST analysis (Schuirmann, 1981; Westlake, 1981) ( $p < .001$ ).

...

**Subject:** Can you knock over the yellow tower please?

**Robot:** Ok. ... I have knocked down the yellow tower.

**Subject:** Thank you. Can you please knock down the blue tower now, please?

**Robot:** Ok. ... I have knocked down the blue tower.

**Subject:** And can you knock down the red tower now?

**Robot:** Ok. ... I have knocked down the red tower.

Figure 3. In-Task Human-Robot Dialogue

Overall, one might expect participants to rely on direct commands for a variety of reasons: (1) the task used in this experiment is a novel and unfamiliar one for which there exist no conventionalized social or politeness norms; (2) the task is straightforward, and there is nothing of interest for participants to talk to the robot about beyond the task at hand, and thus, casual conversation or small-talk is unlikely; and (3) the task is simple enough that it could be completed solely by issuing low-level metric commands (e.g., “drive forward one meter”).

However, our results show that not only did most participants use ISAs in this scenario, but that some relied almost exclusively on them, as seen in Fig. 3. In order to more fully examine this phenomenon, we decided to examine ISA use across a set of experiments we have previously run. In addition to the experiment presented in this paper, we examined two additional interaction corpora collected from previously conducted but unpublished pilot experiments: (1) an experiment similar to the understanding condition of the presented experiment, differing only in that an Aldebaran Nao was used, and (2) an experiment similar to the misunderstanding condition of the presented experiment, but in which the iRobot Create delivered meals instead of demolishing towers.

While the data from those experiments cannot be directly compared to the data from the presented experiment due to different experimental designs, we can broadly compare ISA use across the interaction data from all three experiments, to get a sense as to whether or not the ISA use seen in the presented experiment should be regarded as typical. As shown in Table 2, ISA use was common across all three experiments. Fig. 4 shows a dialogue drawn from the first of these two interaction corpora (Row 2 in Table 2); Fig. 5 shows a dialogue drawn from the second of these two interaction corpora (Row 3 in Table 2).

Table 2: Meta-Analysis of Indirect Speech Act Use Across Experiments.

#	Robot Used	Understanding	n	Total Task-Relevant Utterances	% Indirect
1	iRobot Create	Conditional	24	118	28.0%
2	Aldebaran Nao	Yes	13	167	10.2%
3	iRobot Create	No	21	160	42.0%
			58	351	33.33%

### 3.5 Discussion and Implications for NL-Enabled Robots

Our results showed that most participants used ISAs, and that they were used a significant percent of the time, even in an unconventional, non-social, simple scenario, and even when the robot repeatedly displayed an inability to understand ISAs. Furthermore, our meta-analysis of these results within the context of a set of similar experiments suggests that these results are not out of the ordinary.

**Subject:** Oh let's see here. I have to do something. I would like you to go to the blue tower of cans to your left.  
**Robot:** Ok.  
**Subject:** And I would like you to knock down the top can on the blue tower.  
**Robot:** Ok.  
**Subject:** You can knock down all of the cans if you like.  
**Robot:** Ok.  
**Subject:** Very good, thank you so much. Nice job!

Figure 4. In-Task Human-Robot Dialogue from Additional Interaction Corpus 1 (Aldebaran Nao capable of understanding ISAs in a tower toppling scenario)

**Robot:** Hello my name is Waiterbot. I am taking care of you today. Please tell me what drink I should bring you.  
**Subject:** Uh, I'd like a water.  
**Robot:** That is nice. Please tell me what drink I should bring you.  
**Subject:** Uh could I have a water?  
**Robot:** You could order that. Please tell me what drink I should bring you.  
**Subject:** Um I would like to have a glass of water please.  
**Robot:** That is interesting. Please tell me what drink I should bring you.  
**Subject:** Uh, I need a water.  
**Robot:** Okay. Please tell me what drink I should bring you.  
**Subject:** Um could I have a water?  
**Robot:** You could order that. Please tell me what drink I should bring you.  
**Subject:** Water.

Figure 5. In-Task Human-Robot Dialogue from Additional Interaction Corpus 2 (iRobot Create incapable of understanding ISAs in a restaurant scenario)

Broadly, this suggests that participants are likely to bring their social norms into interactions with robots. What is more, participants will bring these *politeness* norms into contexts even when the robot's sole purpose is to fulfill participant's requests (i.e., it does not purport to have its own goals, desires, and intentions), and when the robot is distinctly non-humanoid. We suspect that the frequency of ISA usage might have been even higher had the robot expressed its *own* goals, which the participants would have been overriding with their requests, or if a more humanoid robot had been used (given that previous research has suggested that humans treat humanoid robots more politely than they treat mechanical robots (Hinds, Roberts, & Jones, 2004)). Overall, the results of this experiment suggests some high-level design and application principles, which, if followed by robot architecture designers, should improve task-based dialogue interactions of natural-language enabled robots with humans.

**Language-enabled robots engaging in dialogue-based human-robot interactions must be able to understand ISAs in some application areas and situations:**

If a language-enabled robot is expected to be used in any situation with dialogue-based interaction with humans, designers should expect the robot to misinterpret upwards of 10% of commands if they are unable to understand ISAs. What is more, this level of miscomprehension is likely to occur even with non-humanoid robots, with robots under clear obligations to satisfy interactants' requests, in

contexts for which conventionalized social norms do not exist, and even when the robot repeatedly demonstrates an inability to understand indirect speech acts.

If a robot is expected to interact with naïve users, this error rate is clearly unacceptable: In such cases, we believe that it would thus be inappropriate to use a language-enabled robot incapable of understanding at the very least, common conventionalized ISAs such as those concerned with capabilities, permissions, and desires. In cases where interaction with naïve users is not expected to be common, this error rate may be less problematic, as users may be explicitly or implicitly trained to avoid using indirect language. But this avoidance of natural, polite communication is likely to come at a cost with respect to humans' perceptions of the robot: If it is important to robot designers that human teammates be able to engage in natural, human-like dialogue with a robot, then this constrained communication style and its associated interaction costs may prove to be unacceptable. We thus suggest that language-enabled robots engaging in dialogue-based human-robot interactions must be able to understand ISAs if the robots are expected to commonly engage with naïve users, or if natural, human-like dialogue is of paramount importance.

**Language-enabled robots engaging in dialogue-based human-robot interactions should be able to learn new ISA forms:**

We have thus far suggested that language-enabled robots expected to be used in dialogue-based interactions should be able to understand ISAs. However, this does not mean that robot designers are expected to explicitly design rules to capture every possible way in which one might use indirect speech acts. Instead, it may be sufficient for a robot to be able to learn new ISA forms as they are encountered. Robot designers may expect human perception of robots without the ability to learn new ISA forms to suffer. We would thus suggest that it would be useful for researchers to develop mechanisms allowing language-enabled robots to automatically learn new ISA forms. It is outside the scope of this paper to present and evaluate methods for learning such new ISA forms; however, we argue that some pragmatic reasoning system with a generalized representation of direct and indirect speech acts would enable such learning to occur.

Finally, we can analyze the types of ISAs observed in these experiments in order to determine how robot designers might predict the types of ISAs they may expect in their own experiments. According to Searle's theory of speech acts, an illocutionary act has four components: (1) its *propositional content*, (2) its *essential condition* (i.e., what it "counts as"), (3) its *sincerity condition* (e.g., for a request, that the speaker actually wants the listener to perform the requested action), and (4) a set of *preparatory conditions* (e.g., for a request, that the hearer is able to perform the requested action, that the speaker believes the hearer able to perform the requested action, and that it is not obvious to both the speaker and hearer that the hearer is already planning to perform the requested action).

From the range of ISAs found in our experiment, as well as those observed in our additional interaction corpora (as seen in the previously presented dialogues), we can infer how ISAs are commonly constructed: by simultaneously calling attention to (1) either the preparatory or sincerity condition of the intended utterance form, and (2) part of the requested action (e.g., the action's agent or patient). Examples of each observed combination of literal illocutionary point, condition of focus, and action aspect can be seen in the table below.<sup>1</sup> All indirect requests observed in the presented experiment or in the additional interaction corpora can be accounted for by the taxonomy represented by this table.

Notice that there exists no column for the combination of either *Question* or *Suggestion* as literal illocutionary point and *Sincerity* as condition of focus. This is because it generally does not make

<sup>1</sup>The final parenthetical item is a form we did not observe in our experiment or additional interaction corpora, but which fits the presented framework.

Table 3: Taxonomy of Observed Indirect Requests

Direct Point	Question	Statement	Statement	Suggestion
Condition	Preparatory	Sincerity	Preparatory	Preparatory
Agent	Could you $X$ ?	I need you to $X$ .	You can $X$ .	You should $X$ .
Patient	Could $X$ happen?	I'd like $X$ .	$X$ will happen.	( $X$ should happen.)

sense to draw attention to your own mental states by asking what they are, as your interlocutor cannot assess them, or to draw attention to your own desires by suggesting what they can be, as your interlocutor cannot change them.

Similarly, it does not always make sense to make statements about the abilities of others, especially when they are the presumed domain experts. Consider the third column of Table 3. Here, the examples seen in the second row call attention to the preparatory condition of requests concerning whether or not the action is going to happen anyway. Another subcategory of such patient-directed preparatory statements would be to call attention to the preparatory condition of capability (e.g., “The red tower can be knocked down”). While such an utterance makes sense, it runs the risk of coming off as rude if the hearer is the presumed domain, as it appears to assert that the speaker knows something that the hearer does not. Calling attention to either capability or inevitability for agent-directed preparatory statements runs a similar risk; the speaker calling attention to capability seems to presume a lack of knowledge on the hearer’s part, whereas calling attention to inevitability runs the risk of asserting dominance.

The discussion in this section suggests that robot designers should consider (at least) the following criteria when deciding what types of ISA forms their system must be prepared to handle: (1) The likely illocutionary points users will need to convey (e.g., requests, suggestions, and statements); (2) the relationship between agent and patient in actions users might desire to be performed; and (3) the relationships between the robot and user which might make some utterance forms presumptive or rude.

#### 4. Pragmatic Interpretation Mechanisms

We have identified different categories of ISAs used in actual HRI scenarios and made the case that in many contexts, language-enabled robots will need to understand such indirect speech acts if they are to successfully interact with humans. How this capability could be realized is demonstrated in the subsequent sections. As previously discussed, language understanding is a multi-stage process. Successfully understanding ISAs is an *intention understanding* challenge, which builds upon other necessary capabilities such as speech recognition and semantic processing. In this section, we present a rule-based framework for understanding both direct and indirect speech acts within a given context. Because this rule-based framework derives meaning based on both observed utterances and context, it implements a form of *pragmatic interpretation*. We begin by discussing our utterance representation.

##### 4.1 Utterances

For defining pragmatic rules and inference mechanisms, we adopt the representations from (Briggs & Scheutz, 2011) and consider utterances of the following form:

$$U = \text{UtteranceType}(\alpha, \beta, X, M)$$

where *UtteranceType* denotes the speech act classification,  $\alpha$  denotes the speaker,  $\beta$  denotes the

addressee,  $X$  denotes an initial semantic analysis, while  $M$  denotes a set of sentential modifiers (e.g., “now,” “still,” “really,” “please”).

Below we specify the different utterance types that have been currently implemented in the dialogue component:

$Stmt(\alpha, \beta, X, M)$  - denotes a statement by  $\alpha$  to  $\beta$  asserting that  $X$  is true.

$Ack(\alpha, \beta, X, M)$  - denotes an acknowledgment by  $\alpha$  to  $\beta$  with additional semantics that  $X$  is true.

$AskYN(\alpha, \beta, X, M)$  - denotes a question by  $\alpha$  to  $\beta$  inquiring whether or not  $X$  is true.

$AskWH(\alpha, \beta, X, M)$  - denotes a question by  $\alpha$  to  $\beta$  asking  $\beta$  to resolve the reference specified by  $X$  (e.g., location, identity, etc.).

$ReplyY(\alpha, \beta, X, M)$  - denotes a positive response by  $\alpha$  to  $\beta$  to a yes-no question with additional semantics  $X$ .

$ReplyN(\alpha, \beta, X, M)$  - denotes a negative response by  $\alpha$  to  $\beta$  to a yes-no question with additional semantics  $X$ .

$Instruct(\alpha, \beta, X, M)$  - denotes an instruction by  $\alpha$  to  $\beta$  to perform action or obtain world state  $X$ .

$Accept(\alpha, \beta, X, M)$  - denotes an instruction acceptance by  $\alpha$  to  $\beta$  with additional semantics  $X$  (e.g., acceptance reason).

$Reject(\alpha, \beta, X, M)$  - denotes an instruction rejection by  $\alpha$  to  $\beta$  with additional semantics  $X$  (e.g., rejection reason).

Note that each of the utterance types we described also contains a set of sentential modifiers  $M$ . These modifiers do not generally alter the core semantics of the utterance (primarily indicated by the utterance type and  $X$ ), but rather, they usually alter other facets of how the utterance is understood or chosen during generation, including politeness information and additional semantics regarding the belief state of the interlocutor.

## 4.2 Pragmatic Rules

For the pragmatic rule representation, we build on Briggs and Scheutz (2013):

$$\boxed{[[U]]_C := \langle \mathbf{B}_{lit}, \mathbf{B}_{int}, \theta \rangle}$$

Each rule associates a particular utterance form  $U$  in context  $C$  with a tuple containing the set of beliefs  $\mathbf{B}_{int}$  to be inferred based on the intended meaning of the utterance, the set of beliefs to be inferred based on the literal meaning of the utterance  $\mathbf{B}_{lit}$ , as well as the degree  $\theta$  to which the utterance can be considered a face-threatening act (FTA) in context  $C$  (Brown & Levenson, 1987). The double-brackets  $[[...]]$  with subscript denote the context of the utterance, such that  $[[U]]_C$  can be interpreted as, “the utterance  $U$  in the context described by  $C$ .” To distinguish the implication from other, stricter forms (e.g., logical), we use “:=” to denote pragmatic implication. As such, the rule form can be read in the following way, “The utterance  $U$  in context  $C$  can be pragmatically interpreted as entailing the set of beliefs  $\mathbf{B}_{int}$ .” Both belief sets are represented in order to determine whether or not the pragmatic rule corresponds to a literal form, which affects NLG modulation based on politeness considerations.

It is sometimes the case that interlocutors expect that both the literal and non-literal aspects of utterances to be addressed and acknowledged. In addition to appropriately reacting to the intended illocutionary point of an utterance, there are also expectations regarding the linguistic form of the response. Utterances with the literal illocutionary point of a *question*, for instance, generally demand answers at some point, while utterances with the literal illocutionary point of a *command* require either acceptance or rejection. Consider an ISA such as, “May I have a coffee?”, which has the literal illocutionary point of a question (i.e., as to the permissibility of the speaker obtaining and having coffee) but has the intended illocutionary point of a command to the addressee to serve coffee to the speaker. It is clear that in this context the addressee should serve coffee to the speaker, but in addition, the addressee is prompted by discourse obligations to provide a yes or no answer to the query. Indeed, previous studies on politeness have demonstrated that people prefer and consider it more polite when both the literal and intended illocutionary points of an indirect request are attended to in responses (Clark & Schunk, 1980; Gibbs Jr. & Mueller, 1988).

As such, there should be at least two pathways of interpretation in the dialogue component. The updates to the agent’s beliefs about the world and the intentions of the speaker should be modified only by an utterance’s intended meaning, although the literal meaning should also be tracked and handled.

### 4.3 Example Rules

To more easily specify rules, we use some notational shorthand that reduces the amount of information that needs to be specified for each rule. For instance, in cases where the intended meaning and the literal meaning are equal, it would be redundant to have to specify two equal sets of conjoined belief predicates. Therefore, in the case where two sets of semantics are not specified, it is assumed that  $\mathbf{B}_{\text{int}} = \mathbf{B}_{\text{lit}}$ . And because  $\theta$  is primarily used for NLG purposes beyond the scope of this paper, we elide it from the rule representations found in the following sections.

#### 4.3.1 Direct Utterances

A direct command from agent  $\alpha$  to agent  $\beta$  to do  $\phi$  (e.g., “Get me a coffee!”) can be represented by the following context-independent (i.e.,  $C = \emptyset$ ) rule:

$$\boxed{[[\text{Instruct}(\alpha, \beta, \text{do}(\beta, \phi), \{\})]]_{\emptyset} := \langle \{\text{want}(\alpha, \phi)\} \rangle}$$

Likewise, direct statements and questions can also be simply represented. A direct statement from agent  $\alpha$  to agent  $\beta$  that  $\phi$  is true (e.g., “The coffee is in the breakroom”) can be represented by the following context-independent rule:

$$\boxed{[[\text{Stmt}(\alpha, \beta, \phi, \{\})]]_{\emptyset} := \langle \{\text{want}(\alpha, \text{bel}(\beta, \phi))\} \rangle} \quad (1)$$

A direct question from agent  $\alpha$  to agent  $\beta$  asking whether or not  $\phi$  is true (e.g., “Is the coffee in the breakroom?”) can be represented by the following context-independent rule, where  $\text{itk}(\alpha, \phi)$  denotes the *intention* of agent  $\alpha$  to *know* the true value of  $\phi$ :

$$\boxed{[[\text{AskYN}(\alpha, \beta, \phi, \{\})]]_{\emptyset} := \langle \{\text{itk}(\alpha, \phi)\} \rangle} \quad (2)$$

#### 4.3.2 Indirect Requests

Below we present several examples of pragmatic rules designed to handle different ISA forms. When a speaker  $\alpha$  asks an addressee  $\beta$  whether or not  $\beta$  has a name, this has the literal interpretation that

$\alpha$  wants to know whether or not  $\beta$  possess a name. The non-literal (i.e., intended) interpretation is that  $\alpha$  wants to know  $\beta$ 's name.

$$\boxed{\begin{aligned} AskYN(\alpha, \beta, have(\beta, name), \{\}) := \\ \langle \{bel(\beta, itk(\alpha, have(\beta, name)))\}, \\ \{bel(\beta, itkRef(\alpha, nameOf(\beta)))\} \rangle \end{aligned}}$$

When a speaker  $\alpha$  asks an addressee  $\beta$  whether or not  $\beta$  “can turn  $\phi$ ,” this has the literal interpretation that  $\alpha$  wants to know if  $\beta$  has the ability of turning in direction  $\phi$ , whereas the non-literal interpretation is that  $\alpha$  wants  $\beta$  to turn in direction  $\phi$ :

$$\boxed{\begin{aligned} AskYN(\alpha, \beta, capableOf(\beta, turn(\beta, \phi)), \{\}) := \\ \langle \{bel(\beta, itk(\alpha, capableOf(\beta, turning(\beta, \phi))))\}, \\ \{bel(\beta, want(\alpha, turning(\beta, \phi)))\} \rangle \end{aligned}}$$

#### 4.4 Dialogue Mechanisms

The rule-based framework and its constituent representations for pragmatic interpretation presented above must next be embedded in a larger dialogue component responsible for handling and responding to incoming utterances (as well as necessary turn-taking behaviors). Fig. 6 illustrates how both the literal and intended meanings of an incoming utterance are handled as part of dialogue interactions. In addition to asserting the belief updates based on the intended meaning (which may or may not differ from the literal meaning of the utterance) in the robot’s belief state, a robot’s dialogue component must also handle and generate responses based on expectations generated by the literal illocutionary point of the utterance.

Different response processes to literal illocutionary point occur for: generating acknowledgments, generating answers to yes-no questions, generating answers to non-polar questions, and finally, for generating responses to explicit commands/instructions.

*Generating Acknowledgments:* When an acknowledgment is expected (e.g., after a statement or after a response to a directive), a simple acknowledgment can be generated (e.g., “okay”).

*Generating Yes-No Responses:* The intention of the speaker to know whether or not a proposition  $\phi$  holds is represented as an intention-to-know predicate  $itk(\alpha, \phi)$ , which the dialogue component can use to query the belief component to see if  $\phi$  holds. If  $\phi$  holds a *ReplyY* utterance is constructed and communicated, otherwise a *ReplyN* utterance is constructed and communicated.

*Generating Answers:* In this case, the speaker’s intention is to know some information denoted by a referring expression (e.g., location of the robot). This can be represented by the intention of the speaker to want the robot to inform him or her as to the information denoted by the referring expression (e.g.,  $want(\alpha, informref(\beta, \alpha, \rho))$ , where  $\beta$  denotes the robot,  $\alpha$  denotes the human interlocutor, and  $\rho$  denotes the reference), or an intention-to-know-reference predicate  $itkRef(\alpha, \rho)$ . The dialogue component contains a function that searches belief for an appropriate fact that satisfies the referring expression specified by  $\rho$ . If such a fact can be found, it is communicated in a statement. In the case when no such answer can be found in belief, a general statement that states the ignorance of the robot is generated (e.g., “Sorry, I do not know that”).

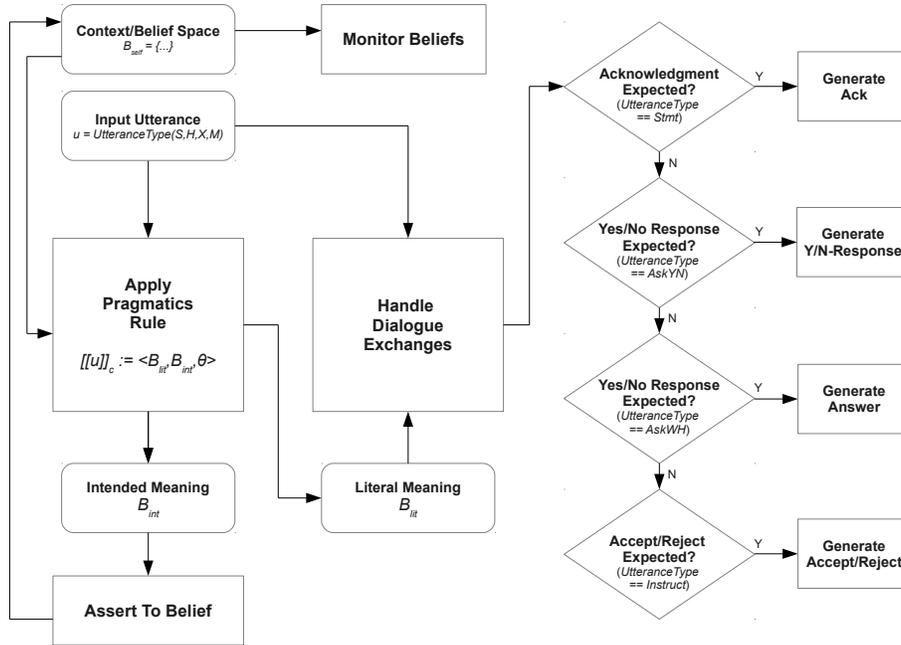


Figure 6. Process diagram for handling and responding to utterances in the dialogue component that handles the extended pragmatic representation, addressing both the literal and non-literal aspects of incoming utterances.

*Generating Responses to Directives:* When people give directives, they expect feedback as to whether these directives are either accepted or rejected. However, to answer appropriately, a robot must be able to reason about the appropriateness of adopting the goal/directive. This depends on reasoning about the various *felicity conditions* that need to be satisfied in order to accept a proposed course of action (the *strong uptake* process). This reasoning process is beyond the scope of this paper but is discussed in (Briggs & Scheutz, 2015).

## 5. Evaluation

In Section 3, we presented evidence of robot-directed ISAs from a human-subject experiment. We then analyzed the indirect requests seen in this experiment as well as those found in two additional interaction corpora, producing the taxonomy seen in Table 3. In this section, we verify that the computational mechanisms introduced in Section 4 can handle the utterance forms associated with each category in that taxonomy. We first show in Section 5.1 how the mechanisms can be integrated into the natural language procession system of a robot architecture and then demonstrate how these mechanisms can handle the ISAs in our experimental data. Specifically, in Section 5.2, we verify coverage of utterance forms seen in Experiment 1, and in Section 5.3, we verify coverage of utterance forms seen in two additional interaction corpora. Finally, in Section 5.4, we verify coverage of utterance forms captured by the proposed taxonomy but not observed in the experimental or additional corpus data.

### 5.1 Architectural Integration

In this section, we give an overview of the various architectural components related to NL understanding and generation in the DIARC architecture (Schermerhorn et al., 2006; Scheutz et al., 2013). However, we should note that the rule-based pragmatics framework introduced in Section 4 can be implemented in any belief-desire-intention (BDI)-paradigm architecture (Bratman, 1987). The proposed pragmatic and dialogue mechanisms were integrated into the natural language processing system of the DIARC architecture, as shown in Fig. 7.

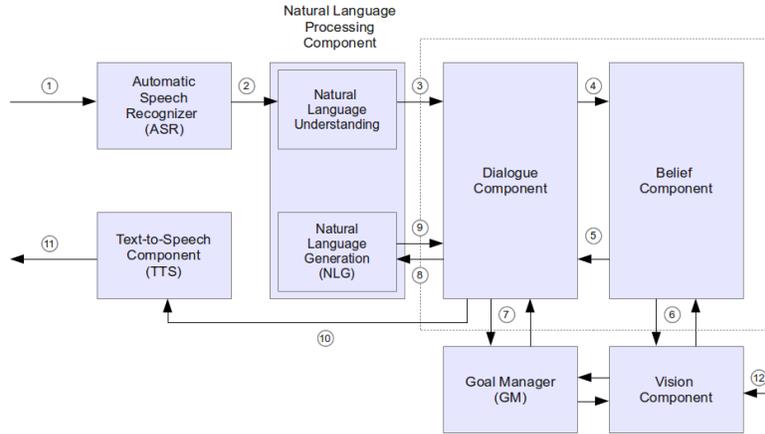


Figure 7. Architectural diagram highlighting components involved in the DIARC NL channel, where nodes in the dotted region are the focus of this work. The significance of each component is described in Table 4. Here, the Vision component is included as an example of how non-linguistic information enters the architecture.

The NL information channel in DIARC consists of a sequence of components that allow for both the understanding (Fig. 7, Component 1) and generation of spoken language (Component 11). The key components in the NL information channel that allow for NL understanding are as follows: the automatic speech recognition (ASR) component, the natural language processing (NLP) component, the dialogue component, and the belief component. The ASR sends detected natural language text to the NLP component (see 2), which performs parsing, reference resolution, and initial semantic analysis. The results are sent to the dialogue component (see 3), which makes pragmatic inferences to ascertain speaker intent and passes the final semantic analysis to the robot’s belief component (see 4). These results specify how the robot should update its own beliefs about the world, as well as beliefs about other agents and their beliefs.

It is worth noting how the component divisions in DIARC appear to nicely correspond to Clark’s theoretical stages of joint understanding. The ASR component is responsible for the process of *perceptual* understanding, while the NLP component is responsible for the process of *semantic* understanding. Weak update (*intentional* understanding) is carried out primarily by the pragmatics reasoning process in the dialogue component (which also factors in contextual knowledge stored in the belief component), whereas *strong uptake* is a process that is started in the belief component but involves interactions between a number of components.

Table 4: Description of the principle data being communicated between components.

Diagram Number	Data Description
1	Acoustic signal from microphone.
2	Recognized text string (e.g., “walk forward”).
3	Inferred surface semantics and speech act classification (e.g., $Instruct(S, H, do(H, walk(H, forward)), \{\})$ ).
4	Inferred intentional semantics/belief updates (e.g., $\{want(S, walking(H, forward))\}$ ).
5	Contextual knowledge. Interaction relevant beliefs: (e.g., $goal(self, walking(self, forward))$ )
6	Belief updates regarding visual perception (e.g., $see(self, face)$ ).
7	Predicates representing goals to be adopted (to GM). NLG requests (from GM).
8	Utterance form (similar to 3) to be translated to string.
9	Surface realization of utterance form (string, as in 2).
10	Text to be spoken by TTS system.
11	Speech output.
12	Visual input from camera.

In summary, in the context of DIARC, the process of NL understanding can then be thought of as a multi-stage process that transduces an acoustic speech signal into a set of first-order-like logical predicates, representative of the interlocutor’s intention, which are asserted into the belief component. Once this predicate is asserted into the belief component, it is up to a variety of other mechanisms to determine the appropriate reaction. In the following section, we describe a set of pragmatic rules that allow the described architectural components to achieve full coverage over the set of task-relevant utterance forms observed in Section 3.

## 5.2 Coverage of Utterance Forms Observed in the Presented Experiment

As described in Section 3, the presented experiment used a simple tower-toppling scenario in which a robot was obligated to find and knock down towers for a human participant. To verify coverage of utterance forms observed in this experiment, we thus provided the robot with (1) the contextual knowledge that, in the current tower toppling scenario, the robot is in the role of the *tower-toppler* and the human interlocutor is in the role of the *instructor* (represented in the robot’s initial set of beliefs by predicates  $role[self, towerToppler]$  and  $role[commX, towerInstructor]$ , respectively), (2) rules defining the role-based obligations found in the tower toppling scenario:

$$\boxed{\begin{aligned} &role(\alpha, towerToppler) \wedge role(\beta, towerInstructor) \wedge want(\beta, \gamma) \\ &\wedge \gamma \in \mathbf{A}_{tower} \Rightarrow oblR_{tower}(\alpha, \beta, \gamma) \end{aligned}}$$

where  $\mathbf{A}_{tower}$  denotes a set of task-relevant actions/effects for the tower toppling task (e.g.  $knockedDown(\alpha, \tau)$  – that a tower  $\tau$  is knocked down by agent  $\alpha$ ), and (3) rules defining when an agent is *potentially* obligated to perform an action or achieve some goal state, such as:

$$\boxed{role(\alpha, towerToppler) \wedge \phi \in \mathbf{A}_{tower} \Rightarrow potentiallyObl(\alpha, \phi)}$$

Note, we distinguish between whether or not an agent is *potentially* obligated to perform an action or achieve a goal state (*potentiallyObl*( $\alpha, \phi$ )) from whether they are *currently* obligated (e.g. *obl<sub>RedTower</sub>*( $\alpha, \beta, \gamma$ ) as seen above). This is because current obligation requires conditions to be met such as the speaker wanting the addressee to perform some action, which will not be inferred in the system until after the directive has undergone pragmatic interpretation as an ISA. As such, determining whether or not an utterance should be tried as an ISA should depend on whether or not the addressee is potentially, rather than currently, obligated.

To use the nomenclature defined in Section 3.5, participants in this experiment used *agent-directed preparatory-pertaining questions* (i.e., “Could/can/would/will you  $X$ ”) and *agent-directed sincerity-pertaining statements* (i.e., “I need/want/would like you to  $X$ ”). Given the information described above, the following rules were sufficient to handle the observed set of utterance forms in a contextually appropriate manner.

### 5.2.1 Agent-Directed Preparatory-Pertaining Questions

Questions of the form “Can you  $X$ ” were handled by the following rule, where  $c = \{\textit{potentiallyObl}(H, X)\}$ :

$$\boxed{[[\textit{AskYN}(S, H, \textit{can}(H, X), \{\})]]_c := \textit{want}(S, X)}$$

The general non-ISA interpretation case (no contextual constraints) is handled by:

$$\boxed{[[\textit{AskYN}(\alpha, \beta, \textit{can}(\alpha, \phi), \{\})]]_{\emptyset} := \textit{itk}(\alpha, \textit{capableOf}(\alpha, \phi))}$$

Using these rules, the utterance “Can you knock down the red tower?” is processed in the following manner:

#### Parse Result:

DIARC’s NLP component performs syntactic analysis, semantic analysis, and literal illocutionary point recognition to transduce the given sentence into the following utterance form:

$$\textit{AskYN}(\textit{commX}, \textit{self}, \textit{can}(\textit{self}, \textit{knockedDown}(\textit{self}, \textit{red}(\textit{tower}))), \{\})$$

#### Derived Belief Updates and Goals:

If  $\textit{bel}(\textit{self}, \textit{potentiallyObl}(\textit{self}, \textit{knockedDown}(\textit{self}, \textit{red}(\textit{tower}))))$ :

$$\textit{want}(\textit{commX}, \textit{knockedDown}(\textit{self}, \textit{red}(\textit{tower})))$$

$$\Rightarrow \textit{goal}(\textit{self}, \textit{knockedDown}(\textit{self}, \textit{red}(\textit{tower})))$$

If  $\neg \textit{bel}(\textit{self}, \textit{potentiallyObl}(\textit{self}, \textit{knockedDown}(\textit{self}, \textit{red}(\textit{tower}))))$ :

$$\textit{itk}(\textit{commX}, \textit{capableOf}(\textit{self}, \textit{knockedDown}(\textit{self}, \textit{red}(\textit{tower}))))$$

Questions of the form “Could you  $X$ ” were handled through the following rule, where  $c = \{\textit{potentiallyObl}(H, X)\}$ :

$$\boxed{[[\textit{AskYN}(S, H, \textit{could}(H, X), \{\})]]_c := \textit{want}(S, X)}$$

The general non-ISA interpretation comes from the rule:

$$\boxed{[[\textit{AskYN}(\alpha, \beta, \textit{could}(\alpha, \phi), \{\})]]_{\emptyset} := \textit{itk}(\alpha, \textit{per}(\alpha, \phi))}$$

Using these rules, the utterance “Could you knock down the red tower?” is processed in the following manner:

**Parse Result:**

$$AskYN(commX, self, could(self, knockedDown(self, red(tower))), \{\})$$
**Derived Belief Updates and Goals:**

If  $bel(self, potentiallyObl(self, knockedDown(self, red(tower))))$ :  
 $want(commX, knockedDown(self, red(tower)))$   
 $\Rightarrow goal(self, knockedDown(self, red(tower)))$

If  $\neg bel(self, potentiallyObl(self, knockedDown(self, red(tower))))$ :  
 $itk(commX, per(self, knockedDown(self, red(tower))))$

## 5.2.2 Agent-Directed Sincerity-Pertaining Statements

Statements such as “I need you to  $X$ ” were handled through the following rule:

$$[[[Stmt(S, H, need(S, X), \{\})]]_0 := want(S, bel(H, want(S, X)))]$$

Using this rule, the utterance “I need you to knock down the red tower” is handled in the following manner:

**Parse Result:**

$$Stmt(commX, self, need(commX, knockedDown(self, red(tower))), \{\})$$
**Derived Belief Updates and Goals:**

$want(commX, knockedDown(self, red(tower)))$   
 $\Rightarrow goal(self, knockedDown(self, red(tower)))$

## 5.3 Coverage of Utterance Forms Observed in Additional Interaction Corpora

As described in Section 3, we examined ISA use in a set of corpora previously collected in the course of conducting a set of unpublished pilot experiments. These experiments either used tower-toppling contexts or a restaurant scenario in which a robot was obligated to bring requested dishes to the human participant. To verify coverage of utterance forms observed in these experiment, we thus provided the robot with, depending on the context, either the previously discussed tower-toppling contextual knowledge and rules, or (1) the contextual knowledge that, in the current service scenario, the robot is the role of the *server* and the human interlocutor is in the role of the *customer* (represented in the robot’s initial set of beliefs by predicates  $role[self, server]$  and  $role[commX, customer]$ , respectively), (2) rules defining the role-based obligations found in the restaurant scenario:

$$\begin{aligned} &role(\alpha, server) \wedge role(\beta, customer) \wedge want(\beta, have(\beta, \gamma)) \\ &\wedge onMenu(\gamma) \Rightarrow oblR_{foodService}(\alpha, \beta, have(\beta, \gamma)) \end{aligned}$$

$$\begin{aligned} &role(\alpha, server) \wedge role(\beta, customer) \wedge want(\beta, served(\alpha, \beta, \gamma)) \\ &\wedge onMenu(\gamma) \Rightarrow oblR_{foodService}(\alpha, \beta, served(\alpha, \beta, \gamma)) \end{aligned}$$

where  $\gamma$  denotes the item being requested, and (3) rules defining when an agent is *potentially* obligated to perform an action or achieve some goal state, such as:

$$\boxed{\begin{aligned} &role(\alpha, server) \wedge role(\beta, customer) \wedge onMenu(\gamma) \\ \Rightarrow &potentiallyObl(\alpha, served(\alpha, \beta, \gamma)) \wedge potentiallyObl(\alpha, have(\beta, \gamma)) \end{aligned}}$$

In the additional interaction corpora, participants additionally used *agent-directed preparatory-pertaining suggestions* (i.e., “You should  $X$ ”), *patient-directed preparatory-pertaining questions* (i.e., “Could/can/may I get/have  $X$ ”), *patient-directed preparatory-pertaining statements* (i.e., “I’ll have/get/take  $X$ ”), and *patient-directed sincerity-pertaining statements* (i.e., “I’d like/love  $X$ ”, “I need/want  $X$ ”). *agent-directed preparatory-pertaining suggestions* (i.e., “Why don’t you  $X$ ”, “You’re going to have to  $X$ ”, “Maybe  $X$ ”, and “Let’s try  $X$ ”), *agent-directed preparatory-pertaining statements* (i.e., “Now you can  $X$ ”, “We’re going to  $X$ ”).

Given the context provided to the robot, the following rules were sufficient to handle the observed set of utterance forms in a contextually appropriate manner.

### 5.3.1 Agent-Directed Preparatory-Pertaining Statements

Statements such as “You can  $X$ ” were handled through the following rule, where  $c = \{potentiallyObl(H, X)\}$ :

$$\boxed{[[Stmt(S, H, can(H, X), \{\})]]_c := want(S, X)}$$

The general, non-ISA interpretation is handled by Rule 1 in Section 4. Using these rules, the utterance “You can knock down the red tower” is processed in the following manner:

<p><b>Parse Result:</b>  <math>Stmt(commX, self, can(self, knockedDown(self, red(tower))), \{\})</math></p> <p><b>Derived Belief Updates and Goals:</b>                  If <math>bel(self, potentiallyObl(self, knockedDown(self, red(tower))))</math>:  <math>want(commX, knockedDown(self, red(tower)))</math>  <math>\Rightarrow goal(self, knockedDown(self, red(tower)))</math></p> <p>If <math>\neg bel(self, potentiallyObl(self, knockedDown(self, red(tower))))</math>:  <math>want(commX, bel(self, can(self, knockedDown(self, red(tower))))</math></p>
---

Statements such as “We’re going to  $X$ ” were handled through the following rule, where  $c = \{potentiallyObl(H, X)\}$ :

$$\boxed{[[Stmt(S, H, goingTo(we, X), \{\})]]_c := want(S, X)}$$

The general, non-ISA interpretation is again handled by Rule 1 in Section 4. Using these rules, the utterance “We’re going to knock down the red tower” is processed in the following manner:

<p><b>Parse Result:</b>  <math>Stmt(commX, self, goingTo(we, knockedDown(self, red(tower))), \{\})</math></p> <p><b>Derived Belief Updates and Goals:</b>                  If <math>bel(self, potentiallyObl(self, knockedDown(self, red(tower))))</math>:</p>
--

$want(commX, knockedDown(self, red(tower)))$   
 $\Rightarrow goal(self, knockedDown(self, red(tower)))$

If  $\neg bel(self, potentiallyObl(self, knockedDown(self, red(tower))))$ :  
 $want(commX, bel(self, goingTo(we, knockedDown(self, red(tower))))))$

### 5.3.2 Agent-Directed Preparatory-Pertaining Suggestions

Suggestions such as “You’re going to have to  $X$ ” were handled through the following rule, where  $c = \{potentiallyObl(H, X)\}$ :

$[[Stmt(S, H, goingTo(H, obl(H, X)), \{\})] := want(S, do(H, X))]$

The general literal interpretation comes from a general pragmatic rule of the form:

$[[[Stmt(S, H, X, \{\})]]_{\emptyset} := want(S, bel(H, X))]$

Using these rules, the utterance “You’re going to have to knock down the red tower” is processed in the following manner:

**Parse Result:**  
 $Stmt(commX, self, goingTo(self, obl(self, knockedDown(self, red(tower))))), \{\})$

**Derived Belief Updates and Goals:**  
 If  $bel(self, potentiallyObl(self, knockedDown(self, red(tower))))$ :  
 $want(commX, knockedDown(self, red(tower)))$   
 $\Rightarrow goal(self, knockedDown(self, red(tower)))$

If  $\neg bel(self, potentiallyObl(self, knockedDown(self, red(tower))))$ :  
 $want(commX, bel(self, goingTo(self, knockedDown(self, red(tower))))))$

Suggestions such as “Why don’t you  $X$ ” were handled through the following rule, where  $c = \{potentiallyObl(H, X)\}$ :

$[[AskWH(S, H, why(not(X)), \{\})]_c := want(S, X)]$

The general, literal interpretation comes from a general pragmatic rule of the form:

$[[AskWH(S, H, X, \{\})]_{\emptyset} := want(S, informref(H, S, X))]$

Using these rules, the utterance “Why don’t you knock down the red tower” is processed in the following manner:

**Parse Result:**  
 $AskWH(commX, self, why(not(knockedDown(self, red(tower))))), \{\})$

**Derived Belief Updates and Goals:**

If  $bel(self, potentiallyObl(self, knockedDown(self, red(tower))))$ :

$want(commX, knockedDown(self, red(tower)))$

$\Rightarrow goal(self, knockedDown(self, red(tower)))$

If  $\neg bel(self, potentiallyObl(self, knockedDown(self, red(tower))))$ :

$want(commX, informref(self, commX, why(not(knockedDown(self, red(tower))))))$

Suggestions such as “Maybe  $X$ ?” were handled through the following rule, where  $c = \{potentiallyObl(H, X)\}$ :

$$[[AskYN(S, H, maybe(X), \{\})]]_c := want(S, X)$$

The general, non-ISA interpretation is generated by Rule 2 in Section 4. Using these rules, the utterance “Maybe knock down the red tower?” is processed in the following manner:

**Parse Result:**

$AskYN(commX, self, maybe(knockedDown(self, red(tower))), \{\})$

**Derived Belief Updates and Goals:**

If  $bel(self, potentiallyObl(self, knockedDown(self, red(tower))))$ :

$want(commX, knockedDown(self, red(tower)))$

$\Rightarrow goal(self, knockedDown(self, red(tower)))$

If  $\neg bel(self, potentiallyObl(self, knockedDown(self, red(tower))))$ :

$itk(commX, maybe(knockedDown(self, red(tower))))$

Suggestions such as “Let’s try  $X$ ?” were handled through the following rule, where  $c = \{potentiallyObl(H, X)\}$ :

$$[[Stmt(S, H, let(us, try(X)), \{\})]]_c := want(S, X)$$

The general, non-ISA interpretation is handled by Rule 1 in Section 4. Using these rules, the utterance “Let’s try knocking down the red tower?” is processed in the following manner:

**Parse Result:**

$Stmt(commX, self, let(us, try(knockedDown(self, red(tower))), \{\})$

**Derived Belief Updates and Goals:**

If  $bel(self, potentiallyObl(self, knockedDown(self, red(tower))))$ :

$want(commX, knockedDown(self, red(tower)))$

$\Rightarrow goal(self, knockedDown(self, red(tower)))$

If  $\neg bel(self, potentiallyObl(self, knockedDown(self, red(tower))))$ :

$want(commX, bel(self, let(us, try(knockedDown(self, red(tower))))))$

### 5.3.3 Patient-Directed Preparatory-Pertaining Statement

Statements of the form “I’ll have  $X$ ” were handled through the following rule, where  $c = \{potentiallyObl(H, X)\}$ :

$$\boxed{[[Stmt(S, H, will(S, have(S, X)), \{\})]]_c := want(S, have(S, X))}$$

The general, non-ISA interpretation is again handled by Rule 1 in Section 4. Using these rules, the utterance “I will have a coffee” is processed in the following manner:

<p><b>Parse Result:</b>  <math>Stmt(commX, self, will(commX, have(commX, coffee)), \{\})</math></p> <p><b>Derived Belief Updates and Goals:</b>                  If <math>bel(self, potentiallyObl(self, have(commX, coffee)))</math>:  <math>want(commX, have(commX, coffee))</math>  <math>\Rightarrow goal(self, have(commX, coffee))</math></p> <p>If <math>\neg bel(self, potentiallyObl(self, have(commX, coffee)))</math>:  <math>want(commX, bel(self, will(commX, have(commX, coffee))))</math></p>
--

Statements of the form “I’ll get  $X$ ” were handled through the following rule, where  $c = \{potentiallyObl(H, X)\}$ .

$$\boxed{[[Stmt(S, H, will(S, get(S, X)), \{\})]]_c := want(S, have(S, X))}$$

The general, non-ISA interpretation is again handled by Rule 1 in Section 4. Using these rules, the utterance “I will get a coffee” is processed in the following manner:

<p><b>Parse Result:</b>  <math>Stmt(commX, self, will(commX, get(commX, coffee)), \{\})</math></p> <p><b>Derived Belief Updates and Goals:</b>                  If <math>bel(self, potentiallyObl(self, have(commX, coffee)))</math>:  <math>want(commX, have(commX, coffee))</math>  <math>\Rightarrow goal(self, have(commX, coffee))</math></p> <p>If <math>\neg bel(self, potentiallyObl(self, have(commX, coffee)))</math>:  <math>want(commX, bel(self, will(commX, get(commX, coffee))))</math></p>
--

Statements of the form “I’ll take  $X$ ” were handled through the following rule, where  $c = \{potentiallyObl(H, X)\}$ :

$$\boxed{[[Stmt(S, H, will(S, take(S, X)), \{\})]]_c := want(S, have(S, X))}$$

The general, non-ISA interpretation is again handled by Rule 1 in Section 4. Using these rules, the utterance “I will take a coffee” is processed in the following manner:

**Parse Result:**

$Stmt(commX, self, will(commX, take(commX, coffee)), \{\})$

**Derived Belief Updates and Goals:**

If  $bel(self, potentiallyObl(self, have(commX, coffee)))$ :

$want(commX, have(commX, coffee))$

$\Rightarrow goal(self, have(commX, coffee))$

If  $\neg bel(self, potentiallyObl(self, have(commX, coffee)))$ :

$want(commX, bel(self, will(commX, take(commX, coffee))))$

5.3.4 *Agent-Directed Preparatory-Pertaining Questions*

Questions of the form “Will you  $X$ ” were handled through the following rule, where  $c = \{potentiallyObl(H, X)\}$ :

$[[AskYN(S, H, will(H, X), \{\})]]_c := want(S, X)$

The general, non-ISA interpretation is handled by Rule 2 in Section 4. Using these rules, the utterance “Will you get me a coffee?” is processed in the following manner:

**Parse Result:**

$AskYN(commX, self, will(self, served(self, commX, coffee)), \{\})$

**Derived Belief Updates and Goals:**

If  $bel(self, potentiallyObl(self, served(self, commX, coffee)))$ :

$want(commX, served(self, commX, coffee))$

$\Rightarrow goal(self, served(self, commX, coffee))$

If  $\neg bel(self, potentiallyObl(self, served(self, commX, coffee)))$ :

$itk(commX, will(self, served(self, commX, coffee)))$

5.3.5 *Agent-Directed Preparatory-Pertaining Suggestions*

Suggestions such as “You should  $X$ ” were handled through the following rule, where  $c = \{potentiallyObl(H, X)\}$ :

$[[Stmt(S, H, obl(H, X), \{\})]]_c := want(S, X)$

The general, non-ISA interpretation is handled by Rule 1 in Section 4. Using these rules, the utterance “You should get me a coffee” is processed in the following manner:

**Parse Result:**

$Stmt(commX, self, obl(self, served(self, commX, coffee)), \{\})$

**Derived Belief Updates and Goals:**

If  $bel(self, potentiallyObl(self, served(self, commX, coffee)))$ :

```
want(commX,served(self,commX,coffee))
⇒goal(self,served(self,commX,coffee))

If ¬ bel(self,potentiallyObl(self,served(self,commX,coffee))):
    want(commX,bel(self,obl(self,served(self,commX,coffee))))
```

### 5.3.6 Patient-Directed Preparatory-Pertaining Questions

Questions of the form “Can I X” were handled through the following rule, where  $c = \{potentiallyObl(H, X)\}$ :

$$\boxed{[[AskYN(S, H, can(S, X), \{\})]]_c := want(S, X)}$$

The general, non-ISA interpretation is handled by Rule 2 in Section 4. Using these rules, the utterance “Can I have a coffee?” is processed in the following manner:

**Parse Result:**  
*AskYN(commX,self,can(commX,have(commX,coffee)),\{\})*

**Derived Belief Updates and Goals:**  
 If *bel(self,potentiallyObl(self,have(commX,coffee)))*:  
     *want(commX,have(commX,coffee))*  
     ⇒*goal(self,have(commX,coffee))*

If  $\neg$  *bel(self,potentiallyObl(self,have(commX,coffee)))*:  
     *itk(commX, capableOf(commX, have(commX, coffee)))*

Questions of the form “Could I X” were handled through the following rule, where  $c = \{potentiallyObl(H, X)\}$ :

$$\boxed{[[AskYN(S, H, could(S, X), \{\})]]_c := want(S, X)}$$

The general, non-ISA interpretation is handled by Rule 2 in Section 4. Using these rules, the utterance “Could I have a coffee?” is processed in the following manner:

**Parse Result:**  
*AskYN(commX,self,could(commX,have(commX,coffee)),\{\})*

**Derived Belief Updates and Goals:**  
 If *bel(self,potentiallyObl(self,have(commX,coffee)))*:  
     *want(commX,have(commX,coffee))*  
     ⇒*goal(self,have(commX,coffee))*

If  $\neg$  *bel(self,potentiallyObl(self,have(commX,coffee)))*:  
     *itk(commX, could(commX, have(commX, coffee)))*

when parsed as  $AskYN(commX, self, could(commX, have(commX, coffee)), \{\})$ , resulted, when  $potentiallyObl(self, have(commX, coffee))$  is true, in the following goal being adopted:  $goal(self, have(commX, coffee))$ , and resulted, when this potential obligation does not exist, in the following belief being adopted:  $itk(commX, per(self, have(commX, coffee)))$ .

Questions of the form “May I  $X$ ” were handled through the following rule, where  $c = \{potentiallyObl(H, X)\}$ :

$$[[AskYN(S, H, may(S, X), \{\})]]_c := want(S, X)$$

The general, non-ISA interpretation is handled by Rule 2 in Section 4. Using these rules, the utterance “May I have  $X$ ” is processed in the following manner:

**Parse Result:**

$AskYN(commX, self, may(commX, have(commX, coffee)), \{\})$

**Derived Belief Updates and Goals:**

If  $bel(self, potentiallyObl(self, have(commX, coffee)))$ :

$want(commX, have(commX, coffee))$

$\Rightarrow goal(self, have(commX, coffee))$

If  $\neg bel(self, potentiallyObl(self, have(commX, coffee)))$ :

$itk(commX, may(commX, have(commX, coffee)))$

5.3.7 *Patient-Directed Sincerity-Pertaining Statements*

Statements of the form “I need  $X$ ” were handled through the following rule:

$$[[Stmt(S, H, need(S, X), \{\})]]_{\emptyset} := want(S, bel(H, want(S, X)))$$

Using this rule, the utterance “I need coffee” is processed in the following manner:

**Parse Result:**

$Stmt(commX, self, need(commX, have(commX, coffee)), \{\})$

**Derived Belief Updates and Goals:**

$want(commX, have(commX, coffee))$

$\Rightarrow goal(self, have(commX, coffee))$

Statements of the form “I want  $X$ ” were handled through the literal statement case (Rule 1 in Section 4).

Using this rule, the utterance “I want a coffee” is processed in the following manner:

**Parse Result:**

$Stmt(commX, self, want(commX, have(commX, coffee)), \{\})$

**Derived Belief Updates and Goals:**

$want(commX, have(commX, coffee))$

$\Rightarrow goal(self, have(commX, coffee))$

Statements of the form “I would like  $X$ ” and “I would love  $X$ ” were handled through the following rules, where  $c = \{potentiallyObl(H, X)\}$ :

$$\begin{aligned} & [[Stmt(S, H, would(want(S, X)), \{\})]_c := want(S, bel(H, want(S, X))) \\ & [[Stmt(S, H, would(like(S, X)), \{\})]_c := want(S, bel(H, want(S, X))) \\ & [[Stmt(S, H, would(love(S, X)), \{\})]_c := want(S, bel(H, want(S, X))) \end{aligned}$$

The general, non-ISA interpretation is handled by Rule 1 in Section 4. Using these rules, the utterance “I would like a coffee” is processed in a similar manner to the above cases.

#### 5.4 Coverage of Unobserved Utterance Forms

Finally, while we did not observe any *Patient-Directed Preparatory-Pertaining Suggestions*, these are a natural category of ISA to expect and complete the taxonomy defined in Section 3.5; we will thus describe how such ISAs are handled by our approach. Suggestions such as “ $X$  should be  $Y$ .” were handled through the following rule, where  $c = \{potentiallyObl(H, Y(X))\}$ :

$$[[Stmt(S, H, shouldBe(X, Y(X)), \{\})]_c := want(S, Y(X))$$

Using this rule, an utterance such as “The red tower should be knocked down” is processed in the following manner:

<p><b>Parse Result:</b>  <math>Stmt(commX, self, shouldBe(red(tower), knockedDown(red(tower))), \{\})</math></p> <p><b>Derived Belief Updates and Goals:</b>                  If <math>bel(self, potentiallyObl(self, knockedDown(self, red(tower))))</math>:  <math>want(commX, knockedDown(self, red(tower)))</math>  <math>\Rightarrow goal(self, knockedDown(red(tower)))</math></p> <p>If <math>\neg bel(self, potentiallyObl(self, knockedDown(self, red(tower))))</math>:  <math>want(commX, bel(self, shouldBe(red(tower), knockedDown(red(tower))))</math></p>
---

## 6. General Discussion

Indirect speech acts are an integral part of human-human communication. The ability to communicate our intentions indirectly allows us, as just one example, to better achieve our goals through the help of others, without straining our social relationships. As robots’ capabilities increase, so too will their status as agents be increased in perception. With this increase in perceived agency, it will become increasingly difficult for us to avoid carrying over behaviors such as ISAs from our social interactions with humans into our interactions with robots. And it will be just as hard not to make pejorative inferences about robots when they speak or act in ways that unknowingly violate those social conventions.

While it has been suspected for some time that we are rapidly approaching this point, the experimental work we have presented provides the first empirical evidence that we have reached it. For many years, the lack of good speech recognition has been the primary obstacle on the path to natural human-robot dialogue. This year, for the first time, word error rates on the Switchboard Corpus are dipping below double digits (Saon, Sercu, Rennie, & Kuo, 2016). While the state-of-the-art word error rate of 6.9% is still too high for natural human-robot dialogue, it is low enough that speech

recognition can no longer be considered the main source of error in natural language understanding. In our second experiment, a robot participating in a restaurant scenario (a domain not far removed from the domains of interest for many HRI researchers) suffered a 28% *utterance error rate* due to its inability to understand indirect speech acts. We take this as evidence that natural-language capable robots employed in realistic task-based environments will increasingly find not speech recognition errors, but semantic and pragmatic errors, to be the dominant source of error in their interactions.

### 6.1 Empirical and Design Contributions

While the space of utterances a human *could* say during a task-based human-robot interaction is truly vast, conventionalized indirect requests take up a large proportion of the space of utterances they are *likely* to say. We have shown how a simple taxonomy can be used by roboticists to determine the types of conventionalized indirect requests likely to be used in their task domain, and how a relatively small number of contextually appropriate pragmatic rules designed with the aid of this taxonomy enable full coverage over an example task domain.

In Section 3.5, we used our experimental results to present a set of design recommendations for robot architecture designers: specifically, that language-enabled robots (1) must be able to understand ISAs if they are to engage in dialogue-based human-robot interactions, and (2) should be able to learn new ISA forms. In the following subsections, we will discuss the extent to which our technical contributions allow robot architecture designers to comply with these principles.

### 6.2 Technical Contributions

The presented algorithms and architectural mechanisms enable robot architecture designers to comply with our first recommendation, as those mechanisms allow robots to engage in the reasoning processes needed to appropriately recognize both commands and indirect requests. To demonstrate these mechanisms, we introduced several examples of reasoning rules and showed how these rules can be instantiated in an integrated robot architecture to achieve the desired capabilities in a simple HRI scenario. However, it is important to point out that these rules are only for demonstration purposes and not the principle contribution of the present work; rather, the contribution consists in the algorithms and architectural mechanisms (and their interplay), which together constitute a framework for *explicit reasoning* about joint interaction between the robot and its interaction partners.

We emphasize this to draw a contrast with many of the popular approaches being transferred over from dialogue systems to social robotics, such as POMDP-based dialogue/action management systems (e.g., (Gopalan & Tellex, 2015; Young, S., Gašić, M., Keizer, S., Mairesse, F., Schatzmann, J., Thomson, B., & Yu, K., 2010)) in which much of the reasoning we describe is made *implicitly* (if present at all). Such systems are not only tied to the data they learned from but are limited in their introspective capabilities due to the lack of structured representations about dialogue structure, intended meanings, and interlocutor’s mental states. And while it is outside the scope of this paper to argue for a mixture of statistical and structured representation (as they are employed in the proposed system compared to statistical approaches without structured representations), it is important to point out that the present system allows for such feats as explicit dialogues about goals, obligations, and permissions that essentially rely on the system’s introspective capabilities allowing it to access its structured representations. Below, we contrast our approach with other systems found in the literature.

#### 6.2.1 Other HRI NL Systems

There exist a variety of approaches to enabling robots to have basic task-based NL interactions with human interactants. Some of these approaches include POMDP-based methods (Tellex et al., 2013). Other architectures include a variety of belief-desire-intention (BDI)-paradigm NL understanding

and generation systems that facilitate task-based human-robot interactions (Lemaignan, Ros, Sisbot, Alami, & Beetz, 2012; She et al., 2014). Some even focus on issues of mental modeling and perspective-taking (Warnier, Guitton, Lemaignan, & Alami, 2012). However, these architectures do not address issues such as ISA understanding.

The architectures that do attempt to enable robots to understand ISAs (Williams, Briggs, Oosterveld, & Scheutz, 2015; Wilske & Kruijff, 2006), however, do not recognize dialogue obligations generated by the literal/surface forms of the received utterances in addition to those generated by non-literal semantics. Instead, they seek only to detect non-literal directives and then respond accordingly (either answering only the literal non-directive or the non-literal directive). For example, Wilske and Kruijff (2006) also utilize this behavior pattern to avoid having the robot say “I’m unavailable/busy” in the case when it receives a non-literal directive when it is in “non-servant” mode. Instead, the robot simply responds to the literal form of the utterance in order to at least satisfy discourse obligations (Wilske & Kruijff, 2006).

### 6.2.2 TRIPS

In contrast with many NL approaches implemented in robotic architectures that either ignore or make *implicit* many of facets of NL interaction, one of the most prevalent NL architectures that engage in *explicit* reasoning about dialogues is the TRIPS architecture. However, making direct comparison with TRIPS is somewhat tricky, as there are often no particular rules or algorithmic commitments specified in the relevant literature. For instance, (Allen et al., 2001) describes the handling of an ISA in which both the literal surface form and non-literal interpretation are processed. However, the only details given regarding how the literal semantics are processed are that the system recognizes and processes an obligation to “RESPOND-TO” the surface utterance. Nonetheless, we believe we can identify at least two points of relative strength.

The two components within the TRIPS architecture responsible for handling ISAs and determining the type of response entailed by the received utterance are the Interpretation Manager (IM) and Discourse Context Component (DCC), respectively (Allen et al., 2001). According to the TRIPS website<sup>2</sup>, the rules found in the TRIPS-IM are described in more detail in Hinkelman and Allen (1989). The rules described in Hinkelman and Allen (1989) for identifying potential conventionalized ISAs are primarily rules pertaining to the surface features of the utterance (e.g., Is “please” used? Does the utterance align with one of the Searle’s ISA forms?) However, other pieces of evidence can indicate ISAs. For instance, repeated use of the “Can you  $X$ ?” construction, despite having been shown that the robot is indeed capable of performing an action such as  $X$ . As such, it would make sense for the pragmatic interpreter to have a rule that formalizes the notion of, “If  $A$  asks  $B$  whether or not  $B$  is capable of doing  $X$ , and  $A$  already believes  $B$  is capable of doing  $X$ , then interpret this as a directive.” Yet, it is not clear that TRIPS currently has the mechanisms to do this at the initial, rule-based level. In our architecture, a rule dependent on a belief about the speaker’s belief is no different than any other rule, as the set of contextual constraints for the rule  $C$  can include terms that indicate the beliefs of other agents (and the dialogue component will query the belief modeling component to ensure these constraints are satisfied). This direct connection with the architecture’s belief component also would enable perceptual modulation of ISA understanding as well. Additionally, the TRIPS architecture does not appear to model the sociolinguistic aspects of utterances, modeled in our architecture by the  $\theta$  values. These values are primarily utilized during NL generation (influencing whether or not to generate literal or non-literal directives) and are not the focus of this paper (see Briggs and Scheutz (2016) for more information about utterance selection and politeness values).

<sup>2</sup>[https://www.cs.rochester.edu/research/cisd/projects/trips/architecture/interpretation\\_manager.html](https://www.cs.rochester.edu/research/cisd/projects/trips/architecture/interpretation_manager.html)

### 6.2.3 Summary

In this section, we have compared our rule-based pragmatics framework to other approaches that seek to enable robotic agents to understand ISAs. While there are various specific and subtle differences between our architecture and implementation relative to other approaches, we view our key contribution to be a framework and set of representations that bind utterances (speech act representations) with a meaning representation that is much richer than what is found in other systems and allows for reversing the direction of information flow in a straightforward manner. Not only are we interested in simply solving the problem of associating the right intended meaning with the observed utterance in the present context, but we are interested in associating the utterance with additional semantic information that allows for distinguishing and ranking different utterance forms during NL generation. The rules and reasoning processes used in other approaches to identify and interpret ISAs do not inform the generation process (e.g., TRIPS contains a rule to interpret utterances with “please” as directives, but this does not inform when the system should or should not use this politeness softener).

### 6.3 Limitations and Future Work

This integrated architectural approach advances the state of the art in taskable robots in HRI and opens up several directions for future research. For example, given the pragmatic framework introduced in this paper, it is now possible to develop mechanisms for learning new pragmatic rules and interpretations that can influence where and when the robot will generate both direct and indirect representations and when it will prefer one over the other (our second design requirement). As a robot may not be certain of the rules it learns (or, indeed, of the utterances it hears or the contexts it operates under), the proposed approach may need to be adapted to represent uncertainty and ignorance. And indeed, we have already begun to examine this possibility (Williams et al., 2015). In that work we demonstrated, in part, how pragmatic rules such as those used in this paper can, when paired with a Dempster-Shafer theoretic uncertainty representation framework (c.f. Shafer 1976), be adapted based on new information. By integrating this adaptation mechanism into the algorithms presented in this paper, and by further developing mechanisms allowing entirely new pragmatic rules to be learned, we should be able to make progress toward our second, learning-oriented, design goal.

Our integrated approach also enables robots to be aware of the explicit use of politeness markers and thus enables them to use this information to make various inferences about their interlocutor (e.g., their social roles). Furthermore, the framework provides the building blocks for handling more advanced interpretations of utterances that require the integration of social norms together with context-dependent modulators (such as the interlocutor’s social standing).

## 7. Conclusion

We have presented experimental evidence that humans make frequent use of ISAs when communicating with robots, even in the simplest of task-based interaction contexts. This use of ISAs, however, poses a significant challenge for future cognitive robotic architectures as they will have to be able to make sense of such utterances. To address this challenge, we have presented mechanisms for automatically understanding ISAs in different HRI contexts. Our findings provide both justification for previous and ongoing research in AI and HRI on interpreting ISAs, as well as motivation for future research. First, it will be important to continue development of ISA understanding algorithms that increase both the breadth of handled ISA forms as well as the efficiency and accuracy of intention understanding. Second, it will also be important to develop mechanisms that allow robots to automatically *learn* ISA understanding rules, based on an analysis of previous dialogue, reinforcement signals, and direct explanations. Third, future work should determine how to best generate clarification requests when a robot is unsure if it correctly interpreted an utterance, and how such

a request might be phrased so as to maximize the information that can be gleaned from an interlocutor's response. Finally, HRI researchers should investigate whether the frequency or variety of ISAs will differ when using robots with different morphologies, varying capabilities, or in different experimental contexts.

## Acknowledgments

This work was supported in part by ONR grants #N00014-14-1-0149 and #N00014-11-1-0493.

## References

- Allen, J. F., Byron, D. K., Dzikovska, M., Ferguson, G., Galescu, L., & Stent, A. (2001). Toward Conversational Human-Computer Interaction. *AI Magazine*, 22, 27–37.
- Baxter, P., Kennedy, J., Senft, E., Lemaignan, S., & Belpaeme, T. (2016). From characterising three years of HRI to methodology and reporting recommendations. In *the Eleventh ACM/IEEE International Conference on Human Robot Interaction* (pp. 391–398). Christchurch, NZ. doi:10.1109/HRI.2016.7451777.
- Bratman, M. (1987). *Intention, plans, and practical reason*. Stanford, CA: Center for the Study of Language and Information.
- Briggs, G., & Scheutz, M. (2011). Facilitating Mental Modeling in Collaborative Human-Robot Interaction through Adverbial Cues. In *Proceedings of the sigdial 2011 conference* (pp. 239–247). Portland, Oregon: Association for Computational Linguistics.
- Briggs, G., & Scheutz, M. (2013). A hybrid architectural approach to understanding and appropriately generating indirect speech acts. In *Proceedings of the Twentieth-Seventh AAAI Conference on Artificial Intelligence* (pp. 1213–1219). Bellevue, WA.
- Briggs, G., & Scheutz, M. (2014). How robots can affect human behavior: Investigating the effects of robotic displays of protest and distress. *International Journal of Social Robotics*, 6, 343–355. doi:10.1007/s12369-014-0235-1.
- Briggs, G., & Scheutz, M. (2015). “Sorry, I can’t do that”: Developing mechanisms to appropriately reject directives in human-robot interactions. In *AAAI Fall Symposium Series: Artificial Intelligence for Human-Robot Interaction* (pp. 32–36). Arlington, VA.
- Briggs, G., & Scheutz, M. (2016). The pragmatic social robot: Toward socially-sensitive utterance generation in human-robot interactions. In *AAAI Fall Symposium Series: Artificial Intelligence for Human-Robot Interaction* (pp. 12–15). Arlington, VA.
- Brown, P., & Levenson, S. C. (1987). *Politeness: Some universals in language usage*. Cambridge, MA: Cambridge University Press.
- Clark, H. H. (1996). *Using language*. New York, NY: Cambridge University Press. doi:10.1017/cbo9780511620539.
- Clark, H. H., & Schunk, D. H. (1980). Polite responses to polite requests. *Cognition*, 8, 111–143. doi:10.1016/0010-0277(80)90009-8.
- Gibbs Jr, R. W. (1986). What makes some indirect speech acts conventional? *Journal of Memory and Language*, 25, 181–196. doi:10.1016/0749-596x(86)90028-8.
- Gibbs Jr, R. W., & Mueller, R. A. (1988). Conversational sequences and preference for indirect speech acts. *Discourse Processes*, 11, 101–116. doi:10.1080/01638538809544693.
- Gomez, R., Kawahara, T., Nakamura, K., & Nakadai, K. (2012). Multi-party human-robot interaction with distant-talking speech recognition. In *Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction* (pp. 439–446). Boston, MA. doi:10.1145/2157689.2157835.
- Gopalan, N., & Tellex, S. (2015). Modeling and solving human-robot collaborative tasks using POMDPs. In *Robotics: Science and Systems: Workshop on Model Learning for Human-Robot Communication*. Rome, Italy.
- Hinds, P. J., Roberts, T. L., & Jones, H. (2004). Whose job is it anyway? A study of human-robot interaction in a collaborative task. *Human-Computer Interaction*, 19, 151–181. doi:10.1207/s15327051hci1901&2.7.

- Hinkelman, E. A., & Allen, J. F. (1989). Two constraints on speech act ambiguity. In *Proceedings of the 27th Annual Meeting on Association for Computational Linguistics* (pp. 212–219). Vancouver, Canada. doi:10.3115/981623.981649.
- Lemaignan, S., Ros, R., Sisbot, E. A., Alami, R., & Beetz, M. (2012). Grounding the interaction: Anchoring situated discourse in everyday human-robot interaction. *International Journal of Social Robotics*, 4, 181–199. doi:10.1007/s12369-011-0123-x.
- Perrault, C. R., & Allen, J. F. (1980). A plan-based analysis of indirect speech acts. *Computational Linguistics*, 6, 167–182.
- Rich, C., Ponsler, B., Holroyd, A., & Sidner, C. L. (2010). Recognizing engagement in human-robot interaction. In *Proceedings of the Fifth ACM/IEEE International Conference on Human-Robot Interaction* (pp. 375–382). Osaka, Japan. doi:10.1109/hri.2010.5453163.
- Saon, G., Sercu, T., Rennie, S. J., & Kuo, H. J. (2016). The IBM 2016 English Conversational Telephone Speech Recognition System. *CoRR*, abs/1604.08242, 1–5.
- Schermerhorn, P., Kramer, J. F., Middendorff, C., & Scheutz, M. (2006). DIARC: A testbed for natural human-robot interaction. In *Proceedings of the AAAI Conference on Artificial Intelligence* (pp. 1972–1973).
- Scheutz, M., Briggs, G., Cantrell, R., Krause, E., Williams, T., & Veale, R. (2013). Novel mechanisms for natural human-robot interactions in the DIARC architecture. In *Proceedings of the AAAI Workshop on Intelligent Robotic Systems*.
- Scheutz, M., Schermerhorn, P., Kramer, J., & Anderson, D. (2007). First steps toward natural human-like HRI. *Autonomous Robots*, 22, 411–423.
- Schlöder, J. J. (2014). *Uptake, clarification and argumentation*. Unpublished master's thesis, Universiteit van Amsterdam.
- Schuurmann, D. (1981). On hypothesis-testing to determine if the mean of a normal-distribution is contained in a known interval. *Biometrics*, 37(3), 617–617.
- Searle, J. R. (1975). Indirect speech acts. *Syntax and Semantics*, 3, 59–82.
- Searle, J. R. (1976). A classification of illocutionary acts. *Language in Society*, 5, 1–23. doi:10.1017/s0047404500006837.
- Shafer, G. (1976). *A mathematical theory of evidence*. Princeton, NJ: Princeton University Press.
- She, L., Yang, S., Cheng, Y., Jia, Y., Chai, J. Y., & Xi, N. (2014). Back to the blocks world: Learning new actions through situated human-robot dialogue. In *Proceedings of the Fifteenth Annual Meeting of the Special Interest Group on Discourse and Dialogue* (pp. 89–97). Philadelphia, PA.
- Tellex, S., Thaker, P., Deits, R., Simeonov, D., Kollar, T., & Roy, N. (2013). Toward information theoretic human-robot dialog. In *Robotics: Science and Systems VIII* (pp. 409–416). Cambridge, MA: The MIT Press. doi:10.15607/rss.2012.viii.052.
- Warnier, M., Guitton, J., Lemaignan, S., & Alami, R. (2012). When the robot puts itself in your shoes. Managing and exploiting human and robot beliefs. In *Proceedings of the 21st IEEE International Symposium on Robot and Human Interaction Communication* (pp. 948–954). Paris, France.
- Westlake, W. (1981). Bioequivalence testing—a need to rethink. *Biometrics*, 37, 589–594.
- Williams, T., Briggs, G., Oosterveld, B., & Scheutz, M. (2015). Going beyond command-based instructions: Extending robotic natural language interaction capabilities. In *Proceedings of AAAI Conference on Artificial Intelligence* (pp. 1388–1393). Austin, TX.
- Wilske, S., & Kruijff, G.-J. M. (2006). Service robots dealing with indirect speech acts. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems* (pp. 4698–4703). Beijing, China.
- Young, S., Gašić, M., Keizer, S., Mairesse, F., Schatzmann, J., Thomson, B., & Yu, K. (2010). The hidden information state model: A practical framework for POMDP-based spoken dialogue management. *Computer Speech & Language*, 24, 150–174.

---

Gordon Briggs, Naval Research Laboratory, Washington, DC, USA. Email: gordon.briggs.ctr@nrl.navy.mil; Tom Williams, Tufts University, Medford, MA, USA. Email: williams@cs.tufts.edu; Matthias Scheutz, Tufts University, Medford, MA, USA. Email: matthias.scheutz@tufts.edu