Situated Natural Language Interaction in Uncertain and Open Worlds

Thomas E. Williams

A dissertation submitted
in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy
Computer Science: Cognitive Science

Tufts University
May 2017

Reading Committee

Matthias Scheutz (Advisor)
Anselm Blumer
Jan de Ruiter
Candace Sidner

Abstract

As intelligent agents become integrated into our society, it becomes increasingly important for them to be capable of engaging in natural, human-like human-agent interactions. A key aspect of such interactions is the ability to engage in *pragmatically appropriate* natural language dialogues. That is, intelligent agents must be able to understand and generate natural language expressions in a way that is sensitive to their current environmental context, social context, and dialogue state.

This problem is especially difficult in the *uncertain and open worlds* common to typical *human-robot interaction* scenarios, in which a robot cannot be expected to have *perfect* or *complete* knowledge of its environment. What is more, many of the approaches that have been developed to facilitate human-robot dialogues are tailored to specific knowledge representation schemes or particular domains of information that prevent them from being generally applicable across robot architectures or across application domains.

To address these concerns, I have developed a set of algorithms for understanding and generating natural language in uncertain and open worlds, and a set of *general frameworks* and *architectural mechanisms* that allow these algorithms to be agnostic to representational format and application domain whenever possible. The algorithms and architectural mechanisms presented in this dissertation represent an interdisciplinary approach to artificial intelligence, in which cognitive science is drawn upon to provide theoretical frameworks (e.g., *Speech Act Theory*, the *Givenness Hierarchy*), and cognitive models (e.g. the *Incremental Algorithm*), and in which computer science is drawn upon to provide computational frameworks (e.g., *Multi-Agent Systems*, *Integrated Robot Architectures*) and techniques (*Dempster-Shafer Theory*, *logical inference*, *search*).

In this dissertation, I demonstrate how these algorithms and architectural mechanisms can be integrated into a single natural language processing pipeline within an integrated robot architecture. What is more, I show how this integrated system extends the state of the art in domains such as *natural language enabled wheelchairs* when implemented on robot hardware.

Contents

List of Figures

List of Tables

List of Videos

For the reader's benefit, here are the videos referenced in this dissertation.

1. Dialogue from *Star Trek: The Next Generation, "Interface"*:
   `https://www.youtube.com/watch?v=39XnZAEl1Z4`
2. Context-Sensitive Indirect Speech Act Understanding and Generation:
   `https://www.youtube.com/watch?v=wDrz44YyI58`
3. Covert Robot-Robot Communication: Report Condition 1 (SILENT):
   `https://www.youtube.com/watch?v=t_MLNBReoic`
4. Covert Robot-Robot Communication: Report Condition 2 (VERBAL):
   `https://www.youtube.com/watch?v=y9WODq30Nrk`
5. Intelligent Wheelchair Demonstration
   `https://www.youtube.com/watch?v=eSU1YWdSfpk`

Preface

All work presented in this dissertation was conducted at the Human-Robot Interaction Laboratory at Tufts University under the supervision of Dr. Matthias Scheutz. All human subject experiments conducted as part of this dissertation were approved by the Tufts Institutional Review Board.

This dissertation presents work that was previously presented in conference papers, journal articles, and book chapters, some of which is presented here, verbatim.

- **Chapter Two** contains text and/or content from a paper to be published in the Proceedings of the International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS) (Williams, Johnson, Scheutz, & Kuipers, 2017).
- **Chapter Three** contains text and/or content from a book chapter to be published by Oxford University Press (Williams & Scheutz, 2017), and from papers published in the Proceedings of the AAAI Conference on Artificial Intelligence (AAAI) (Williams, Cantrell, Briggs, Schermerhorn, & Scheutz, 2013; Williams & Scheutz, 2016a), the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (Williams & Scheutz, 2015b), the Annual Meeting of the Cognitive Science Society (COGSCI) (Williams & Scheutz, 2015a).
- **Chapter Four** contains text and/or content from papers published in the Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI) (Williams, Acharya, Schreitter, & Scheutz, 2016) and the AAAI Fall Symposium on Artificial Intelligence for Human-Robot Interaction (AI-HRI) (Williams, Schreitter, Acharya, & Scheutz, 2015).
- **Chapter Five** contains text and/or content from a paper (currently) under review at the Annual Meeting of the Association for Computational Linguistics (ACL).

- **Chapters Six and Seven** contain text and/or content from papers published in the Proceedings of the AAAI Conference on Artificial Intelligence (Williams, Briggs, Oosterveld, & Scheutz, 2015), the IEEE Symposium on Robot and Human Interactive Communication (RO-MAN) (Williams, Briggs, Pelz, & Scheutz, 2014) the Ibero-American Conference on Artificial Intelligence (IBERAMIA) (Williams, Núñez, et al., 2014), and the AAAI Fall Symposium on Artificial Intelligence for Human-Robot Interaction (Williams & Scheutz, 2016b), and in the Journal of Human-Robot Interaction (JHRI) (Williams, Briggs, & Scheutz, 2015; Briggs, Williams, & Scheutz, 2017).

- **Chapter Eight** contains text and/or content from papers to be published in the Proceedings of the International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS) (Williams, Johnson, Scheutz, & Kuipers, 2017) and under review at the Journal of Robotics and Autonomous Systems.

  Furthermore, content throughout this dissertation was previously described in the proceedings of the AAAI Doctoral Consortium (Williams, 2016), HRI Pioneers (Williams, 2015) and the Young Researchers' Roundtable on Spoken Dialogue Systems (YRRSDS) (Williams, 2014).

Acknowledgments

They say it takes a village to raise a graduate student. Well, at least I've said so. Surely we can agree that there exists at least one person p such that p says that it takes a village to raise a graduate student. Existence proof: myself. I would also count myself as an existence proof for this (now well established) aphorism, as there are simply too many people who deserve thanks for their inspiration, assistance, and camaraderie over the past six years.

I must begin, of course, by thanking my advisor, Matthias Scheutz, for making me the researcher, writer, and presenter I am today; for keeping me on track, while supporting me when my path diverged; for giving me the opportunity to travel the world; and for giving me a world of opportunities. It has been an absolute honor to work with him, and I consider it enormously and improbably lucky that he took a chance on me and took me into his laboratory.

I would also like to thank the other members of my dissertation committee; Anselm Blumer for his mentorship as I begin my career as an educator, and Candy Sidner and J.P. de Ruiter both for pushing me to make this dissertation the best it can be. All four of my committee members have been truly inspirational to me, and have each motivated me to be rigorous in my research, with respect to my writing and argument (Matthias) my cognitive grounding (Candy), my algorithms and theory (Anselm), and my methodology and analysis (J.P.).

I couldn't ask for better labmates than I've had in the Human-Robot Interaction Lab. While all of my current and former labmates have been a pleasure to work with, I'd especially like to thank Gordon Briggs, Evan Krause, and Matt Dunlap, for their conversation, collaboration, commiseration, constant inspiration, and of course for the carbonation and colonization, which is to say, for the well-hopped board gaming sessions.

Outside of the lab[1], I would like to thank Mike Shah, who has been a great friend on this amazing journey. On the other side of the pond, I'd like to thank Michael Beetz and the members of the Institute for Artificial Intelligence at Uni-Bremen, especially Fereshta Yazdani, for one of the most interesting, adventurous, and information-dense months of my life. I would like to give a shout out to my AI-HRI colleagues, including Rick Freedman, Ross Mead, Brad Hayes, and Patrícia Alves-Oliveira, who have been a pleasure to work with, who have motivated me to become deeply engaged with our research community, and who have simply become great friends. Finally, I would like to thank Out of Range for the best two hours of my week, every week. It is hard for me to express how great a blow it is for me to leave them; I don't want to stop at all.

None of this would have been possible without the love and support of my family. I'd especially like to thank my mom – who has been my biggest cheerleader – not only for believing in me, but for always making sure that I believed in myself. Finally, I'd like to thank Rachel for helping me find the appropriate balance between sanity and insanity, for making sure I remember to eat and come home, and, well, for everything. I am so excited for this next step in our life together.

This work was funded in part by ONR grants #N00014-11-1-0289, #N00014-11-1-0493, #N00014-10-1-0140 #N00014-14-1-0144, #N00014-14-1-0149, #N00014-14-1-0751, and NSF grants #1111323, #1038257.

[1]Contrary to popular belief, graduate students do on occasion leave their natural laboratorial habitat.

Dedication

To my parents, for their unwavering and constant support.

Chapter 1

Introduction

We are entering the age of intelligent agents. It has long been a dream of humanity to integrate artificially intelligent agents into our society, and we are now discovering that we not only desire, but need, such agents: An aging population suffers from the expense and lack of caretakers; disaster relief workers are confronted with disaster zones too dangerous to enter; astronauts find themselves in need of assistance in the vacuum of space.

And while their need may be less dire, the general public yearns for such agents as well. Personal Assistants such as Apple's Siri and Amazon's Alexa are increasingly used to help manage schedules and lifestyles and access information. And such assistants are promising to enter our homes in the form of robots like Cynthia Breazeal's Jibo and ASUS' Zenbo.

As intelligent agents come into the field and onto shelves, the need for these agents to be *socially* and *contextually* intelligent increases. Perhaps most critical is the need for such agents to be able to engage in truly *natural* communication, not only for capricious consumers, but for those in eldercare facilities, disaster areas, and the depths of space, for whom the training or use of other control methods may be too cognitively or physically taxing.

In eldercare robotics and education robotics, it may simply be too *cognitively burdensome* for the target population to learn to interact with their would-be caregiving or educational assistants through some other modality. In space robotics and urban search-and-rescue robotics, it may be too *physically burdensome* for the target population to interact with their would-be assistants or rescuers, due to, e.g., lack of gravity, or trapped limbs. In urban search-and-rescue environments, victims are also not likely to have the time or inclination to learn another control modality to interact with their would-be rescuers. It is thus important that robots operating in these and other domains be taskable through control modalities, like natural language, that the general public is already familiar and proficient with.

In addition, natural language is an attractive interaction modality even for users without physical or cognitive limitations. Natural language is infinitely flexible, requires only the simplest of hardware additions, is lightning-fast and interactive, and allows for learning from a single example.

While there has been some progress on enabling natural-language based human-robot interaction (HRI) (Mavridis, 2015), most natural language enabled robots rely on highly scripted interactions, keyword spotting, and shallow natural language processing techniques. For many applications, these methods may be sufficient to achieve the desired behavior, which may be restricted to a small class of tasks. Such methods, however, are not helpful for the development of robots that are generally and flexibly taskable, that can learn about new entities and concepts on the fly, and that are capable of engaging in truly natural *human-like* human-robot interactions; that is, the robots that are the ultimate goal of human-robot interaction (Fong, Thorpe, & Baur, 2001; Goodrich & Schultz, 2007; Dautenhahn, 2007).

What is more, even natural-language enabled robots designed to handle more natural, flexible dialogue typically operate under a set of assumptions that severely restrict the types of language they are prepared to handle. To see why this is the case, consider the following scenario.

1.1 An Illustrative Example

Imagine a robot named Cindy and a human named Bob. Cindy and Bob are working together in a disaster relief scenario, and have just left a room containing a refrigerator, and two medical kits: one on a table, and one on a counter. After driving down the hallway for a few minutes, Bob turns to Cindy and says "The commander needed that medical kit from the kitchen! He's in the cafeteria on the second floor. Could you bring it to him?"

In the following subsections, I will step through a variety of implicit assumptions made by most current natural language enabled robots, and how those assumptions may be violated in this example.

1.1.1 Knowledge is Certain

Many language enabled robots make an assumption of *certain knowledge*: that there are a set of entities known to the robot, each with a set of known features. However, in most realistic human-robot interaction scenarios, the robot's knowledge will almost certainly be *uncertain*, especially since robots do not have perfect perception of their worlds. Here, for example, even if Cindy perfectly perceives her world, it may be difficult for her to ascertain the identity of the referent described as 'the kitchen'. The fact that the room the robot recently left contained a refrigerator, a table, and a counter certainly provides evidence that it may be a suitable candidate referent, but it is unlikely that this information will be sufficient for the robot to be completely *certain* that that room was a kitchen. Uncertain knowledge may also be troublesome if a robot needs to refer to some entity but is uncertain of its properties.

1.1.2 Knowledge is Complete

With very few exceptions, natural language enabled robots operate under a closed world assumption: that natural language utterances made by interlocutors will only refer to entities known of *a priori*. In many cases, it is even assumed that interlocutors will only refer to entities that are *currently visible*. However, in many realistic human-robot interaction scenarios, robots cannot be assumed to have full knowledge of every entity which could possibly be referenced. Here, for example, it may be the case that Cindy has never actually seen the cafeteria on the second floor. This should not prevent Cindy from discussing or reasoning about this location, and it may even be the case that Cindy is able to infer where the cafeteria may be located, and travel to it accordingly. We would thus state that robots must be able to operate under an *open world* rather than *closed world* assumption.

1.1.3 Knowledge is Homogeneous in Domain

Many natural language enabled robots only attempt to handle referents from a single domain, e.g., objects, or locations. However, in many realistic human-robot interaction scenarios, interlocutors may be expected to refer to referents from a variety of domains. Here, for example, Bob refers to objects, locations, and people. Similarly, it may be most natural for Cindy to refer to some object based on its relation to some location, person, or event.

1.1.4 Knowledge is Centralized, and Homogeneous in Representation

Many natural language enabled robots assume that all information regarding candidate referents is contained within a single, centralized knowledge base, and that all information regarding those entities is represented in a single format (an assumption that goes hand in hand with an assumption of a single relevant domain of knowledge). However, in many realistic robot architectures, information may be decentralized for a variety of reasons. First, a centralized knowledge base knowledge base may become a computational bottleneck when multiple processes need to access it frequently. Not only does querying become more expensive due to the size and complexity of the central knowledge base, which must accrue and bind together information from different architectural components, but computational resources become focused onto a single "stress point" rather than balanced across the architecture's components.

Knowledge may also be decentralized due to heterogeneity of knowledge representations; information about entities recognized by a vision component will likely be stored in a substantially different manner than the map produced by a mapping component. Moreover, only the vision component must deal with low-level visual features such as pixels, textures, edges, etc. and thus it makes sense to keep such information local to where it is processed and needed.

Here, for example, it is entirely plausible that information regarding the medkit may be stored as a visual memory object in the architecture's vision component, information regarding the kitchen and the second floor may be stored as elements of a hybrid metric-topological map in the architecture's mapping component, and information regarding the commander may be stored in a Prolog knowledge base elsewhere in the architecture.

1.1.5 All Utterances are Commands

Many natural language enabled robots make a command-based assumption: that they have been designed purely for task-based interactions, and that thus all utterances made by human interlocutors will be for the purpose of providing some command to the robot. Even if we assume that Cindy was designed for task-based interactions[1], it is entirely plausible that Bob might issue questions or statements in service of that task. Here, for example, Bob uses two statements and an utterance that may be a command, or may be a question. This leads directly to our next assumption.

[1]In fact, in many of our examples, we assume Cindy to be a Mobile Dexterous Social (MDS) robot, designed by Breazeal et al. (2008): a robot that can certainly be used in task-based scenarios, but whose design prioritizes social concerns.

1.1.6 Utterances are Expressed Directly

Most natural language enabled robots make an assumption of *direct expression*: that the *intended* meaning of a sentence is directly derivable from its *semantics*, which reflect the literal, direct meaning of the sentence. However, in most human-robot interaction scenarios, it is reasonable to assume that humans will use a high volume of so-called *indirect speech acts*, whose *literal* meanings mismatch their *intended* meanings. Here, for example, Bob's final utterance is *literally* a question regarding Cindy's abilities. But it is *more likely* that Bob's true intention is to express a command that he believed would be impolite to express directly (i.e., "Bring it to him.")

1.1.7 Utterances are Contextually Invariant

Many natural language enabled robots make an assumption of *contextual invariance*: that the meaning of a particular utterance is always the same (an assumption that goes hand in hand with an assumption of direct expression). However, in most human-robot interaction scenarios, it is reasonable to assume that humans will use a variety of utterances whose intended meanings depend on the current context. Here, for example, Cindy may need to decide whether Bob's last utterance is intended as a question or a command based on her current model of Bob. Does Bob really not know whether this is something she is able to accomplish? Is Bob in a social role that grants him the right to issue her commands? And so forth.

What is more, many utterances in human-robot dialogue may contain referring expressions that can *only* be resolved with respect to the current dialogical or environmental context. Here, for example, Bob uses a variety of *anaphoric* expressions, such as 'he', 'it', and 'him' whose meanings are entirely dependent on the current dialogue state, as well as the demonstrative adjective phrase 'that medical kit', in which 'that' uses the current context to narrow the scope of possible referents. In addition, Bob uses 'you', a *deictic* expressions whose meaning additionally depends on the immediate spatio-temporal context.

1.2 Summary

In order to advance the state of the art of natural language based human-robot interactions, we must develop natural language enabled robots that challenge the assumptions explicitly or implicitly made by most other such robots. We must develop natural language enabled robots that can (1) handle uncertain and open worlds; (2) make use of distributed knowledge that is heterogeneous in domain and in representation; (3) process a wide variety of utterance forms and referring expression forms; and (4) process such utterances in a context sensitive manner.

Over the course of the next nine chapters, I will describe the algorithms and architectural mechanisms I have developed in service of these goals, from the ground up, starting with architectural concerns, then discussing referring expression understanding and generation, followed by pragmatic understanding and generation, and finishing with a discussion of applications in which the full understanding and generation pipeline has been employed.

To be specific, I provide in Chapter 2 a short overview of the robot architecture in which these algorithms and mechanisms have been developed. In Chapter 3 I present a set of reference resolution algorithms as well as architectural mechanisms that they facilitate or are facilitated by: Section 3.2 presents *SPEX*, the *Spatial Expert*, an architectural component responsible for performing spatial reference resolution in open worlds; Section 3.3 presents *REX*, the *Referential Executive*, an architectural component responsible for a broader class of referential activities, including domain-independent reference resolution of definite noun phrases in uncertain and open worlds. In Chapter 4, I describe how *REX* is used within the context of a broader *Givenness Hierarchy theoretic* framework in order to additionally resolve anaphoric and deictic expressions in a context sensitive manner. In Chapter 5, I move beyond referring expression *understanding*, and show how *REX* is used for referring expression *generation* as well.

Next, I move on to discuss *pragmatic reasoning*: in Chapter 6, I present experimental evidence demonstrating the extent of indirect speech act usage in human-robot interaction, and then present a Dempster-Shafer theoretic algorithm for understanding indirect speech acts in a context sensitive manner under uncertainty and ignorance. In Chapter 7, I move beyond indirect speech act *understanding*, and show (1) how our Dempster-Shafer theoretic framework is used for indirect speech act *generation* as well; and (2) how this framework can be used to generate clarification requests to resolve pragmatic and referential ambiguity. Finally we move beyond the pragmatics of *human*-robot communication, and discuss the pragmatics of *robot*-robot communication.

In Chapter 8 I discuss the application of the presented algorithms and architectural mechanisms to assistive robotics, by providing a comprehensive survey of natural language enabled wheelchairs and then demonstrating how the algorithms presented in this dissertation advance the state of the art by integration of our robot architecture (configured to use the presented algorithms and architectural mechanisms) with the Vulcan intelligent wheelchair. Finally, in Chapter 9 I discuss directions for future work, including applications of the presented algorithms and architectural mechanisms outside of robotics.

Chapter 2

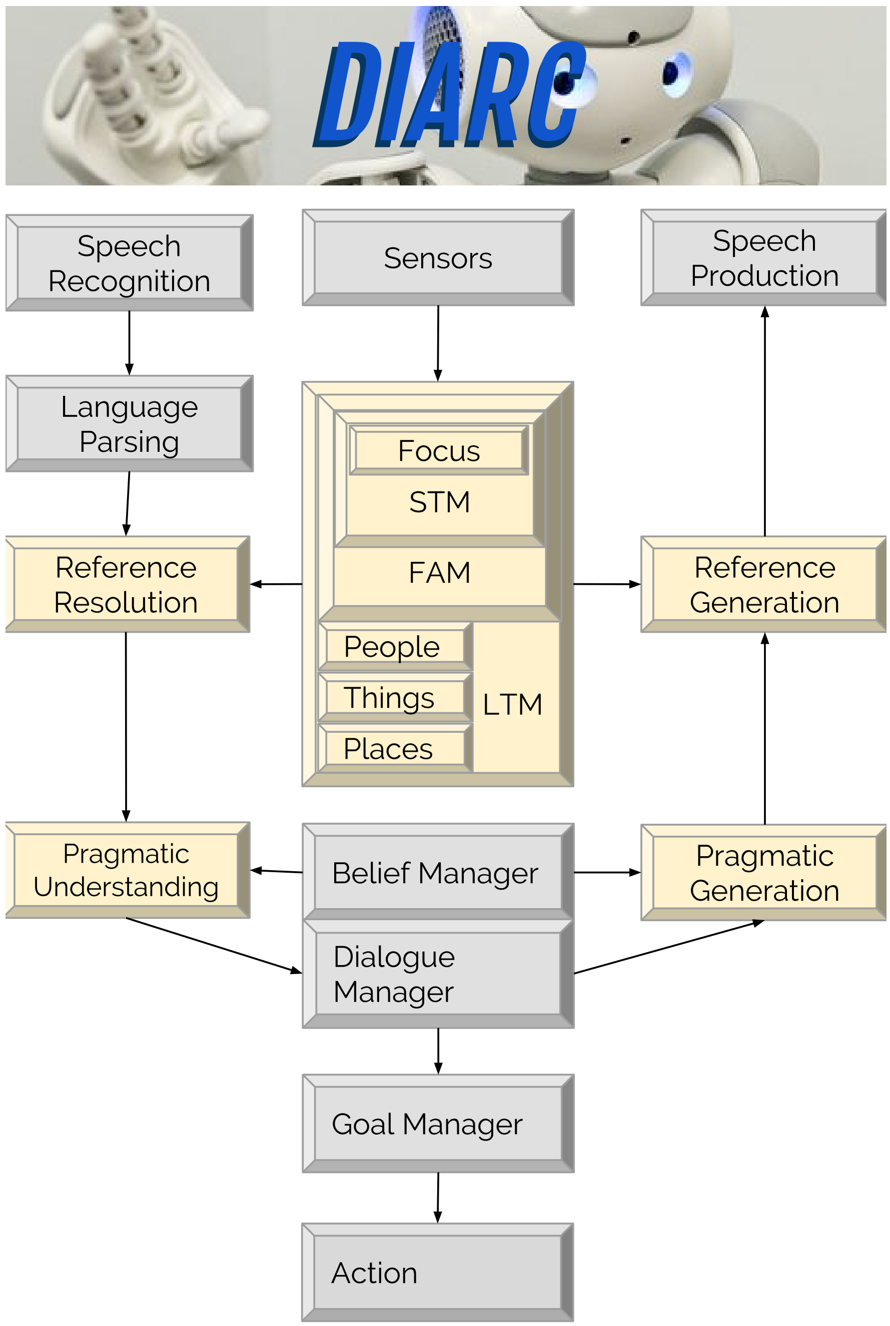
Architectural Background

In Chapter 1, I described a set of capabilities necessary for *natural, human-like human-robot interaction.* The majority of this dissertation concerns natural language processing algorithms I have designed in order to facilitate these capabilities. These algorithms do not exist in a vacuum: in order for my algorithms to facilitate meaningful capabilities, they must interact with computational mechanisms that dictate how a robot reasons about and travels within its environment. As such I owe a great intellectual debt to the researchers who have enabled the host of capabilities that my algorithms make use of and interact with, both linguistic and non-linguistic.

But perhaps even more important to acknowledge is the cognitive scaffold into which this host of algorithms is integrated. The computational mechanisms discussed in this dissertation are implemented as part of a larger *robot architecture* that conceptualizes how a robot's cognitive processes can and should interact. This robot architecture, *DIARC*, is not merely an implementation detail, but is instead explicitly leveraged by the presented algorithms. *DIARC* will make appearances in every chapter of this dissertation, under a variety of different configurations, and as such it is important to understand its motivations, principles, and points of variance and invariance. The partial architecture diagram shown in Figure 2.1 serves as a roadmap for the technical contents of this dissertation, as described in the caption beneath that figure.

This chapter is divided into two sections. In Section 2.1 I will describe *DIARC*, the robot architecture into which the presented algorithms are integrated. In Section 2.1, I will describe *ADE*, the multi-agent infrastructure in which *DIARC* is implemented.

Figure 2.1: Architectural Diagram



Partial Architecture Diagram showing the *primary* architectural configuration built up over the course of this dissertations. Specifically, the nodes highlighted in yellow represent the technical contributions of this dissertation: (from left to right, top to bottom) Reference Resolution (Chapters 3-4), the GH-theoretic hierarchical memory model, including long term memory model comprised of *distributed heterogeneous knowledge bases* (Chapters 3-4 as well), Referring Expression Generation (Chapter 5), Pragmatic Understanding (Chapter 6), and Pragmatic Generation (Chapter 7).

2.1 DIARC

The Distributed, Integrated, Affect, Reflection, Cognition (*DIARC*) architecture is a robot architecture designed for robots that interact naturally with humans (P. W. Schermerhorn, Kramer, Middendorff, & Scheutz, 2006). As described by Scheutz, Schermerhorn, Kramer, & Anderson (2007), *DIARC* is in part designed to satisfy three requirements of natural, human-like, human-robot interaction: social behavior, goal-oriented cognition, and robust intelligence.

First, *DIARC* prioritizes teleological (i.e., goal-related) capabilities, in the form of explicit goal representations and far-reaching goal and task management mechanisms that appear in the majority of architectural configurations. These representations and architectural mechanisms are prioritized because *DIARC* recognizes the importance of *bidirectional intentionality ascription* for natural human-like human-robot interaction. In order for robots to successfully interact with humans, robots must not only be able to infer and communicate about the intentions of their human interlocutors; they must also be able to communicate their own intentions to humans, and generally promote *theory of robot minds*. That is, humans should be prompted to regard robots as having their own beliefs, desires, and intentions, in order to allow them to better predict those robots' behaviors.

Second, unlike most other robot architectures (cf., e.g., Quigley et al., 2009), *DIARC* also prioritizes the natural-language capabilities required for social interaction. While the precise architectural configuration used may differ on a case-by-case basis, *DIARC*'s architectural constituents typically provide capabilities for speech recognition and synthesis, syntactic and semantic processing, as well as mechanisms for introspecting upon and communicating about the dialogical and aforementioned teleological structures required for natural linguistic interactions (Grosz & Sidner, 1988; Lochbaum, Grosz, & Sidner, 1990; Grosz & Sidner, 1990). And, although such mechanisms will not appear in this dissertation, *DIARC* at one time made consistent use of mechanisms for recognizing and generating affective and non-verbal interaction cues.

Finally, *DIARC* makes a set of theoretical commitments designed to ensure *robust* intelligence, including the following, which are discussed by Scheutz et al. (2013):

1. Processing is distributed across a set of *architectural components*.
2. All processing performed in architectural components is asynchronous with respect to that performed in other components.

3. Each component operates on its own, possibly multi-threaded, cognitive cycle.
4. Control is decentralized, and no central executive or "homunculus" may be employed.
5. Goals and their associated primitive skills are represented in a form that includes preconditions, post-conditions, and expected utilities.
6. Actions are selected by priority, and based on availability of resources constrained based on relations of mutual exclusivity.
7. No single architectural learning mechanism is prescribed.
8. No single architectural knowledge representation is prescribed within individual components.
9. Logical formulae are used as "common currency" for inter-component communication whenever possible.
10. Architectural components are capable of introspecting on the multi-agent middleware in which they are implemented.

While in principle *DIARC* may be implemented in any sufficiently flexible multi-agent system middleware, this dissertation will only discuss its implementation within the *Agent Development Environment*.

2.2 ADE

The *DIARC* architecture is implemented in the *Agent Development Environment* (ADE) multi-agent system middleware. *ADE* is an architectural *framework* (Kramer & Scheutz, 2007) that builds on previous work from multi-agent systems(Bellifemine, Poggi, & Rimassa, 1999; Sycara, Paolucci, Van Velsen, & Giampapa, 2003) in order to support the development of individual agent architectures using distributed multi-agent system computing infrastructure.

As such, the *DIARC* architecture is implemented in *ADE* as a set of distributed *architectural components* which satisfy the theoretical commitments of *DIARC* listed above. The majority of these components are responsible for providing discrete cognitive capabilities (e.g., speech recognition, syntactic processing, dialogue processing, goal management, speech synthesis) or for providing access to a robot's sensors and effectors, although components also exist to provide debugging interfaces, provide simulated environments, or facilitate Wizard-of-Oz-style[1] control over real or simulated robots. In addition, every instantiation of the *ADE* framework must make use of at least one *Registry* component with which other components must *register* in order to be part of the architecture. The registry (and by extension, *ADE*) treats architectural components as autonomous software agents in order to facilitate dynamic system configuration, fault tolerance and recovery, distributed computation, and autonomic computing(Scheutz, 2006; Andronache & Scheutz, 2006; Kramer & Scheutz, 2006).

ADE is primarily implemented in Java, with inter-agent communication facilitated by Java RMI. The use of a JVM language provides two main advantages: first, this allows for portability between different architecture platforms; second, it allows for developer flexibility, as code written in a variety of programming paradigms (i.e., through Java, Clojure, or Scala) can be seamlessly and richly integrated.

A comparison of *ADE* to similar robot development environments (e.g. Carmen (Montemerlo, Roy, & Thrun, 2003), Player/Stage (Gerkey, Vaughan, & Howard, 2003)) was previously presented by Kramer & Scheutz (2007). However, in the time since that publication, a number of other robot development environment environments have been created, most principally ROS (Quigley et al., 2009). Because *ROS* is widely used throughout the robotics community at this time, I believe it is important to provide a comparison between it and ADE. For the purposes of this dissertation, it is sufficient to identify three points of contrast, beyond which the capabilities of the two frameworks are generally comparable.

1. There is a general philosophical difference between *ADE* and *ROS* regarding inter-component communication. While ROS components tend to communicate using high volumes of low-level messages, *ADE* components tend to communicate using low volumes of high-level messages (in keeping with *DIARC* commitment 9).

[1] Wizard-of-Oz (WoZ): an experimental paradigm in which a human interacts with a robot or other intelligent agent which they are led to believe is autonomous, when in fact some or all of its behavior is controlled by an experimenter, i.e., "Wizard". The WoZ paradigm was most famously discussed by Dahlbäck, Jönsson, & Ahrenberg (1993) in the context of natural language dialogue systems, and was effectively surveyed within the field of human-robot interaction by Riek (2012). An important delineation is drawn by Baxter, Kennedy, Senft, Lemaignan, & Belpaeme (2016) between *Perceptual* WoZ, in which the Wizard replaces some (potentially unreliable) *input* capability (e.g., speech recognition), and *Cognitive* WoZ, in which the Wizard provides some *cognitive* capability which the robot (may have) otherwise lacked.

2. This philosophical difference is reflected in the mechanisms *ADE* and ROS provide and rely on for inter-component communication. *ROS* provides three mechanisms: *topics* (a publish-subscribe mechanism), *services* (a blocking remote procedure call mechanism used for short-term information requests), and *actions* (a non-blocking remove procedure invocation mechanism used for longer-term task requests). Practically, however, ROS implementations *principally* rely on its publish-subscribe mechanisms, given the tendency towards high volumes of low level messages. In contrast, *ADE* only (currently) provides remote procedure call mechanisms equivalent to *ROS's* services and actions, due to its tendency towards low volumes of high level messages.

3. Finally, as a practical detail, *ADE* and *ROS* differ with respect to distribution and ease-of-use. ROS is readily available in Linux repositories, and makes it easy to distribute and install ROS packages providing new architectural components; however, ROS is notoriously difficult to install, especially outside of the Ubuntu operating system. In contrast, *ADE* is not easily available outside of Tufts University, and does not provide mechanisms for simplifying the distribution or installation of packages providing new architectural components; but *ADE* is far easier to install, and essentially only requires Java as a dependency. This makes *ADE* much easier to get started using, and thus potentially easier to integrate into the classroom as a tool for teaching robotics.

In this chapter, I have provided a high level overview of the *DIARC* architecture and the *ADE* middleware in which it is implemented. In the next chapter, I will begin the discussion of my work within this framework by considering the problem of *reference resolution*.

Chapter 3

Reference Resolution

A crucial aspect of natural language communication is the ability to *refer* (G. M. Green, 1996). That is, we (humans) commonly use expressions that serve to "pick out" some entity about which we want to make some claim, request some information, or issue some command. These so-called *referring expressions* come in a variety of forms (Strawson, 1950), including demonstrative pronouns (e.g., 'this' and 'that'), personal and impersonal pronouns (e.g., 'I', 'you', 'he', 'it'), proper names (e.g., 'Tufts University', 'Thelonious Monk'), and definite and indefinite noun phrases (e.g., "I have eaten *the plums that were in the icebox*", "There is *a house in New Orleans*").

In order for robots to be able to engage in natural, human-like human-robot interactions, they must thus be able to both *understand* and *generate* such referring expressions. In this chapter, I will focus on the task of referring expression *understanding*.

Perhaps the most popular approach towards enabling this capability is to use *co-reference resolution* (Ng, 2010; Soon, Ng, & Lim, 2001), in which new referring expressions are "linked" with previously heard referring expressions. For example, for the sentence pair "The commander needs the medical kit. He says that he left the medkit in the atrium", a co-reference resolution system should identify that [The commander], [He], and [he] all co-refer, as do [the medical kit] and [the medkit]. Determining what referring expressions belong to the same co-reference cluster can be highly informative, especially in text-centered fields such as question answering (McCarthy & Lehnert, 1995) and document summarization (Morton, 2000), in which the co-reference resolution problem has enjoyed significant attention[1].

[1]Although perhaps not as much attention as it *could* attract, as discussed by Versley et al. (2008)

In robotics, however, this is not typically sufficient: the referring expressions which must be understood by robots typically refer to entities in "the real world". The robotics community has thus emphasized the problem of identifying what real-world entities are the referents of referring expressions. This problem goes by many names, including "language grounding" (Steels & Hild, 2012), "reference resolution" (Popescu-Belis, Robba, & Sabah, 1998), and "entity resolution" (Meyer, 2013). While these names are sometimes used to denote the same concept, they carry different connotations, and we will use them to refer to distinct concepts:

Language Grounding:

The problem of associating a referring expression with a continuously (i.e. sub-symbolically) represented percept[2]. Language Grounding (as we cast it) can be broken into two sub-parts:

Reference Resolution: The problem of associating a referring expression with a discretely (i.e., symbolically) represented entity.

Symbol Grounding: The problem of associating a discretely represented entity with a continuously represented percept(Harnad, 1990).

Note that while *referring* is something assumed to happen between linguistic expressions and *real world entities* (i.e., the *referents* of those referring expressions), we take the problem of reference resolution to be the identification of the *mental representations* that may or may not actually be associated with real world entities. This is necessary in that robots (at least those operating purely in the "real world" rather than in virtual or mixed-reality environments) must typically *infer* the existence of "real world" entities, having direct access only to their own internal mental representations. Variants of the reference resolution problem can be further subdivided in two ways. First, they can be subdivided along existential lines:

[2]This is not to be confused with H. H. Clark & Brennan (1991)'s concept of *Grounding in Language*, which refers instead to the process by which communicating agents come to arrive at common ground. For work on that topic, the interested reader is directed to surveys provided by (Baker, Hansen, Joiner, & Traum, 1999; Traum, 1999; M. J. Clark & Liggins, 2012).

Closed World Resolution:

Local Resolution: The reference resolution problem under the assumption that all candidate referents are *perceivable* at resolution time.

Global Resolution: The reference resolution problem under the assumption that all candidate referents are *known* at resolution time.

Open World Resolution:

The reference resolution problem under which it is neither assumed that candidate referents will be perceivable nor that they will be known[3].

Second, variants of the reference resolution problem can be subdivided along ontological lines:

Domain-Dependent Resolution:

The reference resolution problem under the assumption that all candidate referents are drawn from a single "domain", such as objects, or locations.

Domain-Independent Resolution:

The reference resolution problem under the assumption that candidate referents may be drawn from multiple domains[4].

In this chapter I will specifically focus on *open-world reference resolution of definite noun phrases.*

Rest assured, however, that this does not constitute the entire treatment of reference within this dissertation. In Chapter 4 I will expand this discussion to reference resolution for a wider class of referring expressions; and in Chapter 5 I will discuss the process of referring expression *generation*.

There are a number of unique challenges that present themselves to robots seeking to understand referring expressions, due to robots' status as

[3]What is important here is that a solution to this problem may need to create new mental representations if it is determined that a referring expression does not refer to any previously known entity.

[4]What is important here is that a solution to this problem should not perform any sort of domain-specific processing.

situated agents: agents (entities capable of autonomously acting to achieve their own goals (Jennings, 2000)) that are embedded in an environment that is perceivable and manipulable by themselves and other agents with whom they can interact (G. J. Smith & Gero, 2005). While a software entity operating within a non-situated domain such as text mining or document summarization may need to associate entities referenced in a text with previous portions of that text, a robot must instead associate entities referenced in dialogue with its own *mental representations* resulting not only from dialogue and inference, but also from interpretation of sensory data gathered by its perceptual systems.

In this chapter, I will examine the problem of reference resolution over the course of five subsections. In Section 3.1, I will begin with an introduction to the process of reference resolution, from the perspective of psycholinguistics. In Section 3.2 I will discuss my initial work in reference resolution, and will introduce the concept of *open world* reference resolution, in which it is not assumed a priori that the robot knows all possible entities that could be referenced. In Section 3.3, I will present an architectural framework to facilitate referential processing within *DIARC*, and a set of algorithms which make use of that framework in order to perform domain independent reference resolution in uncertain and open worlds. In Section 3.4 I will present analysis and evaluation of these algorithms. Finally, in Section 3.5 I will contrast this work to previous work on reference resolution in robotics.

3.1 Psycholinguistic Motivations

How do people determine what entities are being referred to in natural language utterances? This problem, known as *reference resolution*, is among the most basic facets of language comprehension. In this section, I will examine the theories presented over the past 30 years to account for this process. I will begin by discussing reference resolution in the context of traditional sentence processing models. I will then discuss alternative "propose and dispose" models, and how they led to constraint-based models, the dominant modeling paradigm. Next, I will discuss these constraint-based models in depth, including various implementations of such models. Finally, I will discuss recent Bayesian models of reference resolution.

3.1.1 Modular Theories

"The horse raced past the barn fell" is one of the most famous sentences in the sentence processing literature. Such sentences are known as "garden path" sentences, because they lead the reader or hearer down one path of interpretation up until the end, at which point the reader or hearer realizes that interpretation is incorrect, and must backtrack in order to determine the correct interpretation (Traxler, 2011).

In this case, upon encountering 'fell' it finally becomes known that 'the horse' refers to whichever entity is best described as 'the horse that was raced past the barn', and not, as would previously have been assumed, whichever entity is best described as 'the horse'.

In order to explain this phenomenon, researchers developed *modular* theories of sentence processing (e.g. Frazier, 1987; Britt, 1994; Perfetti, 1990; Pritchett, 1992; Mitchell, Corley, & Garnham, 1992). Such theories, the most prominent of which was Frazier's aptly named "Garden Path" theory (Frazier, 1987), hypothesized that sentences were first analyzed by a *Syntactic Processing Module*, which would build up the most likely parse of a given sentence according to some core set of principles (e.g., *Minimal Attachment* and *Late Closure*) and send the resulting parse tree to the second component, a *Semantic Processing Module*.

The *Semantic Processing Module* would then determine the literal meaning of the sentence by performing subtasks such as assigning thematic roles to the entities referenced in the sentence. If this could not be achieved, i.e., if the sentence was not meaningful given the syntactic structure produced by the previous module, then the Semantic Processing Module would signal the Syntactic Processing Module that a new, better parse was needed, prompting the process to start over.

These sentences predict that semantic information should have no initial effect on sentence interpretation. This prediction, however, does not hold out. Consider the following utterance:

(1) The burglar blew up the safe with the rusty lock.

Upon reading this sentence, it is unclear whether the burglar blew up a safe *that had* a rusty lock, or whether he blew up a safe *using* a rusty lock using some McGuyver-like ingenuity. It should be clear that the first interpretation is more likely, but it takes readers a second or two to come to this conclusion. Under the Garden-Path Theory, nothing should be able to change this, as in all cases the same syntactic structure should be created by the Syntactic Processing Module. However, if readers are provided with initial context, e.g., "The burglar had to choose whether to blow up a new safe or a safe which had sat in the rain for ten years" before Example 1, then readers will have no trouble reading the sentence, and will *not* slow down as the Garden Path theory would expect.

3.1.2 "Propose and Dispose" Theories and Constraint-Based Models

In order to explain this, other researchers (primarily Altmann and colleagues) developed an alternate theory, often referred to as "Propose and Dispose". Models falling under this theory (e.g. Crain & Steedman, 1985; G. Altmann, 1987; Steedman, 1986; G. Altmann & Steedman, 1988; Ni & Crain, 1990) suggested that all possible parse trees were proposed *in parallel* by the Syntactic Processing Module, and that the Semantic Processing Module then uses contextual information to rapidly winnow down this set of options. While this casting of the problem avoids the problem described above, it is not clear whether it's plausible to assume that the thousands of syntactically plausible parses for a given sentence are all generated and stored (see also Church & Patil, 1982). For this and other reasons, "Propose and Dispose" models were quickly superseded by "constraint-based" models, which remain the dominant paradigm today. Constraint-based models suggest that a large number of factors influence sentence processing from the first moments of processing, serving to constrain the syntactic and referential ambiguities *while the sentence is parsed*, such that by the time parsing is finished, the meaning of the sentence (at least its literal, direct meaning) is understood. Thus two stages are no longer needed; a single stage will suffice.

3.1.3 The Visual World Paradigm

While evidence in favor of constraint-based models had already started to accumulate throughout the early nineties (e.g. Boland, Tanenhaus, Garnsey, & Carlson, 1995; MacDonald, Pearlmutter, & Seidenberg, 1994; M. J. Spivey-Knowlton, Trueswell, & Tanenhaus, 1993; M. Spivey-Knowlton & Tanenhaus, 1994; M. Spivey-Knowlton, Tanenhaus, Eberhard, & Sedivy, 1995; Spivey & Tanenhaus, 1998; Tanenhaus & Trueswell, 1995), the best evidence came with the development of the *Visual World Paradigm* by Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy (1995).

In the *Visual World Paradigm* (Huettig, Olivers, & Hartsuiker, 2011), participants hear utterances while looking at a display. This display is typically a set of line drawings depicted on a computer screen (e.g. Allopenna, Magnuson, & Tanenhaus, 1998), or a set of objects arranged in a physical display (e.g. Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995). The utterances heard by participants usually include references to objects in the scene, and come in the form of instructions to follow or general comments to listen carefully to. These utterances are typically heard a few seconds after the objects become visible in the scene. The eye gaze of participants is tracked while they listen to utterances. This is what makes the Visual World Paradigm so useful with respect to related paradigms; participants tend to look at objects when they are mentioned (as well as to objects that constrain the interpretation of other objects), perhaps as a way of relating an utterance to the world around them, so that auditory and visual information may inform each other (G. T. Altmann & Kamide, 2007). Eye tracking data gathered in this way may thus be used to examine the time course of reference resolution[5]. Eye tracking data is analyzed by examining the relative likelihood of looking at different regions during different time intervals, through either analysis of variance, t-tests, logistic regression, log-linear regression, and growth-curve analysis (Huettig, Olivers, & Hartsuiker, 2011).

This paradigm has been useful for evaluating a wide variety of hypotheses associated with constraint-based models. Under constraint-based models, words impose various constraints, which immediately affect the activation levels of various competitors within a referential domain (e.g., by downgrading their likelihoods[6] (Weber & Crocker, 2012))in order to facilitate ambiguity resolution. In order to test what factors serve as constraints, researchers vary the items depicted in a *visual world* and determine whether the factor under investigation affects eye gaze in the time region following a particular word of interest (200ms afterwards, typically).

This has shown that in general, when context is available and salient, it is used in the incremental interpretation of referring expressions, and that goal-relevant reference in particular is established as quickly as possible (Eberhard, Spivey-Knowlton, Sedivy, & Tanenhaus, 1995). Furthermore, this paradigm has been used to show that two-stage language models are implausible because syntactic and referential constraints interact on a word-by-word basis; Sedivy (2002) demonstrates how referential constraints effect parsing choices, by showing how the use of the word ‘only’ affects incremental interpretation of sentences such as “Only businessmen loaned money at low interest were told to record their expenses”. In such sentences, the use

[5]Note that it is unclear to what extent these eye gaze patterns reflect *automatic* versus *deliberate* processing; an important factor for the interpretation of results from this paradigm.

[6]While constraints are typically seen as *downgrading* rather than *eliminating* referential competitors, it is important to note that only a limited number of candidates are typically considered: those within the *referential domain* circumscribed by various contextual factors. For the sake of space I will not address this facet of reference resolution, but it is both interesting and important (Grosz, 1977; Arnold, Tanenhaus, Altmann, & Fagnano, 2004; Dahan & Tanenhaus, 2004; Brown-Schmidt, Campana, & Tanenhaus, 2001; Brown-Schmidt & Tanenhaus, 2008; Louwerse & Bangerter, 2010).

of ‘only’ *immediately* establishes contrast and causes hearers to anticipate a modifier (e.g., ‘loaned money at low rates’), thus supporting the hypothesis that referential factors are able to bias parsing *on-line* in order to pursue otherwise dispreferred syntactic structures.

In fact, the Visual World Paradigm has been used to provide evidence for a wide variety of information sources which work from the first moments of language processing to constrain interpretations, thus facilitating resolution of ambiguity, whether lexical, referential, syntactic, or otherwise.

Pronoun Gender

Perhaps the most straightforward extension of the referential constraints described above is the finding that *pronoun gender* serves to constrain possible sentence interpretations(Arnold, Eisenband, Brown-Schmidt, & Trueswell, 2000).

Participants were shown a scene containing Donald Duck and either Mickey or Minnie Mouse, and were read a sentence such as “Donald is bringing some mail to [Mickey/Minnie]. He’s sauntering down the hill while a violent storm is brewing. [He’s/She’s] carrying an umbrella, and it looks like they’re both going to need it.”

Arnold et al. found that when the choice of pronoun was disambiguating (i.e., when Minnie was depicted), participants’ fixations rapidly converged on the referent soon after the pronoun’s offset, whereas when the choice of pronoun was not disambiguating (i.e., when Mickey was pictured), participants did *not* rapidly converge to one of the two candidates.

Semantics

Researchers have also shown how the presence of semantically related distractors can bias reference resolution through spreading activation. Specifically, Yee and Sedivy showed that objects semantically related to the target of resolution receive more activation than do other distractors, as measured through fixation patterns(Yee & Sedivy, 2006). For example, upon hearing the word ‘lock’ when viewing a scene containing a lock, key, deer, and apple, participants were shown to be more likely to fixate on the key than the apple or deer.

Scalar Implicature

One of the classic examples of pragmatic reasoning is the processing of *scalar implicatures* such as ‘some’. For example, hearing “/Some/ of the apples are green" suggests that some *but not all* of the referenced apples are green. Breheny, Ferguson, & Katsos (2013) showed participants videos in which a person moved items into two boxes. In different conditions, the video showed the person moving different numbers of objects into the two boxes. Participants then heard phrases such as "The woman put a spoon into Box A and a spoon and a fork into Box B." Results showed that participants rapidly constrained the set of considered candidates as soon as conjunctions disambiguated the sentence, even if this required the interpretation of a so-called scalar implicature. Similar findings have been found by others(Y. T. Huang & Snedeker, 2011; Degen & Tanenhaus, 2015). These findings are especially important as they suggest that pragmatic inferences are computed incrementally, i.e., that a complete literal meaning is not computed before determining an utterance's indirect meaning.

Common-Sense Reasoning and Affordance

A related aspect of pragmatic reasoning is the consideration of *object affordances*. For example, a cup *affords* placing things inside of it, whereas a plate does not. Previous models of affordance-reasoning in language assumed that such reasoning only occurred at utterance boundaries (as in Centering theory (Grosz, Weinstein, & Joshi, 1995)), or on a sentence-by-sentence basis (as in Mental Model accounts (Johnson-Laird, 1983; Morrow, Bower, & Greenspan, 1989) or Dynamic Semantic Interpretation accounts (Stalnaker, 1978; Groenendijk & Stokhof, 1991)). In contrast, Chambers et al. showed that when prepositions such as 'inside' are used, attention is immediately restricted to containers, specifically to those large enough to hold the referenced object(Chambers, Tanenhaus, Eberhard, Filip, & Carlson, 2002).

Researchers have also demonstrated that constraints for other types of Common-Sense Reasoning are similarly employed. Kamide et al. showed that verbs can be used to anticipate compatible themes and goals, verbs and agents can combine to anticipate compatible themes, and that in head-final constructions, constraints from pre-verbal arguments can anticipate other pre-verbal arguments(Kamide, Altmann, & Haywood, 2003). Hanna & Tanenhaus (2004) showed that if a listener is asked to hand a speaker an object, and the speaker *has one or more empty hands*, attention is rapidly restricted to out-of-reach items, whereas if the speaker has her hands full, attention is not restricted.

Prosody

Prosodic features of utterances have also been shown to constrain reference resolution. Dahan, Tanenhaus, & Chambers (2002) found that deaccented nouns biased resolution towards prominent entities, while accented nouns biased resolution towards less prominent entities. Arnold et al. found that when disfluent articles were used (e.g., "Thee…uh…"), resolution was biased towards new or unfamiliar entities(Arnold, Tanenhaus, Altmann, & Fagnano, 2004; Arnold, Kam, & Tanenhaus, 2007); Bosker, Quené, Sanders, & de Jong (2014) went on to show that this only occurs when the speaker is a *native speaker* (i.e., when the disfluency can't be easily explained by a language barrier).

Common Ground

Perhaps the deepest investigation of reference-constraining factors has been the work on *common ground*. Suppose you are part of a team exploring a building in a disaster relief scenario. You and a teammate walk past a door marked 'kitchen' in a debris-filled hallway, and then reach an intersection. You take one hallway-fork, your teammate takes the other fork. As you walk down your hallway fork, you see another door marked 'kitchen'. At this moment, your partner says over your walkie-talkie that he's going to come back and check the kitchen. Which room is he talking about? *You* know of two kitchens, but as far as you know, your partner only knows of the single kitchen you had previously passed together. It is thus far more likely that your partner is referring to that previous kitchen. That first kitchen is said to be in *common ground* between you and your partner, whereas the second kitchen is said to be in your own *privileged ground*. Determining that the kitchen in common ground is the likely referent of your teammate's expression requires you to consider the scene from *your teammate's perspective*. This is known as *perspective taking*.

Early work on perspective taking (e.g. Keysar, Barr, Balin, & Paek, 1998; Keysar, Barr, Balin, & Brauner, 2000; Keysar, Lin, & Barr, 2003; Keysar, 2007) worked under the assumption that listeners initially process utterances *egocentrically* (as a heuristic to avoid the cognitive expense of perspective taking), and then, in a second stage, *adjusted* the results of that process to account for what is or is not in common ground. Evidence for this view, known as *strategic egocentrism*, came from studies showing that in certain cases, listeners would look to or even reach for items in *privileged* ground over acceptable referential candidates in common ground.

However, evidence from visual world studies performed by constraint-based modelers (i.e. Hanna, Tanenhaus, & Trueswell, 2003) showed that this was only the case when the item in privileged ground was the *best perceptual match*. If an equally viable candidate was present in common ground, the privileged ground competitor would not typically be considered. Hannah et al. explained this as the interaction of two constraints operating from the earliest moments of processing: a *common ground* constraint, and a *perceptual match* constraint.

The perspective adjustment theorists suggested that perhaps this bias of common ground on early resolution processes was simply the result of *anticipatory effects*; that is, that perhaps reference resolution really is egocentric, but that the activation levels at the start of the reference resolution processed may be implicitly biased towards entities in common ground (as people may be already looking at those items predictively), resulting in selection of items in common ground (Barr, 2008). This is known as the *autonomous activation* or *anticipation integration* account.

Constraint-based modelers rebutted with evidence showing that information in common ground is typically ignored if it can be ruled out *in relation* to information in privileged ground. Heller, Grodner, & Tanenhaus (2008) showed participants scenes containing objects of contrasting sizes, some of which were clearly only visible to themselves (in privileged ground), and some of which were clearly also visible to a confederate (in common ground). Heller et al. tracked participants' eye gaze within such scenes while hearing sentences such as "Pick up the big duck". If a participant heard such an utterance while regarding a scene containing a big box and a big duck in common ground and a small box in privileged ground, all objects should receive equivalent activation before the sentence is heard. If the *anticipation integration* account is correct, then upon hearing 'big', people should be equally likely to look at the big box as the big duck. But in fact, their data suggested that people were far more likely to look (for example) at the big duck than the big box, suggesting that the small box's privileged status was used from the earliest moments of processing to rule out the use of 'big' to refer to the big box.

Other experimenters have since investigated precisely when and how common ground is used to constrain reference resolution. In a series of studies, Brown-Schmidt showed that listeners quickly attend *to privileged information* rather than common ground when *WH-Questions* (e.g., 'Who', 'What', and 'Where') are used (Brown-Schmidt & Tanenhaus, 2009) (although this depends on the contour of one's speech (Brown-Schmidt & Fraundorf, 2015)), that insensitivity to common-ground information in other situations may be due to failures in inhibitory control (Brown-Schmidt, 2009), (see also Grodner, Dalini, Pearlstein-Levy, & Ward, 2012; H. J. Ferguson & Breheny, 2012), and that otherwise the use of an egocentric vs. non-egocentric perspective may be a strategic choice which is made based on the strength of various cues (Brown-Schmidt & Hanna, 2011; Brown-Schmidt, 2012).

Kuhlen & Brennan (2013) and Bezuidenhout (2013) came to similar conclusions: Kuhlen & Brennan (2013) suggest that the degree of believability to a participant that an confederate is *really not aware* of what is in the participant's privileged ground could be one such constraint; Bezuidenhout (2013) suggest that degree of motivation given to participants to actively consider the privileged nature of privileged-ground information could be another. Evidence also shows that people automatically "choose" the correct strategy from the earliest moments of processing, and have trouble switching an entire scene to another perspective afterwards (Ryskin, Brown-Schmidt, Canseco-Gonzalez, Yiu, & Nguyen, 2014). However, other evidence suggests that while users are not initially egocentric with respect to knowledge of the relevant items in a scene, they may be egocentric with respect to the *identities* of those objects when those objects could be ambiguous in how they're interpreted (Mozuraitis, Chambers, & Daneman, 2015).

More recently, Heller, Parisien, & Stevenson (2016) has suggested that cues may not lead to the "choosing" of one strategy over the other after all. Rather, they suggest that people simultaneously consider meanings from multiple interpretations, probabilistically weighting and integrating information from the two perspectives. However, it is not clear how this would account for the difficulty in shifting from one perspective to another.

Conceptual Pacts

Another relevant aspect of common ground is the use of *conceptual pacts*. Over time, interlocutors tend to converge on specific words and phrases to refer to certain entities. This is known as *lexical entrainment*. Brennan & Clark (1996) view this as interlocutors converging on a mutual *conceptualization* of a conversationally relevant entity, that is, establishing a *conceptual pact* with respect to that entity. Such entrainment has been found to be highly partner-specific (Metzing & Brennan, 2003). In line with this theory, Brown-Schmidt (2009) found that participants use early, on-line, partner-specific information to quickly resolve referents. However, other research (i.e. Yoon & Brown-Schmidt, 2013) suggests that while interlocutors are sensitive to conceptual pacts when *generating* referring expressions, they are (relatively) insensitive to those pacts when *interpreting* referring expressions.

Thus while interlocutors *can* make early, on-line, use of such pacts during reference resolution, they often may not.

3.1.4 Implementations of Constraint Based Models

Finally, while constraint-based models have been implemented in a variety of ways (McRae & Matsuki, 2013), most such implementations are *connectionist models*, typically Recurrent Neural Networks such as Elman Nets (Elman, 1990; Elman, Hare, & McRae, 2004). In this subsection, I will briefly describe several prominent implementations.

Competition Integration

The simplest implementation of a constraint-based model is the *competition-integration* model (M. J. Spivey-Knowlton, 1996; McRae, Spivey-Knowlton, & Tanenhaus, 1998). This model simply uses two layers of units: an input layer of "constraint" nodes, and an output layer of "interpretation" nodes. During each model cycle, values within each input node are normalized, and interpretation node activations are calculated by the sum of facilitative constraint activations scaled by connection weights. Constraint nodes then receive feedback proportional to activation. This model suffers in that it cannot generatively construct potential interpretations, is not necessarily scalable, and does not compute the meanings of utterances.

Visitation Set Gravitation

The *visitation set gravitation* model (Tabor & Tanenhaus, 1999) uses simple recurrent networks (i.e., Elman Nets (Elman, 1990)) with input, output, and hidden units. Over time, the hidden units converge to a set of clusters of similar patterns. The centers of these clusters are known as *attractors*. For a given sentence, this network moves through a multi-dimensional space of mental states. As it does so, the hidden layer is analyzed in order to generate reading time predictions, where the predicted time to read a particular word is the time taken to gravitate from the network's current focal point in mental state space to a particular attractor. When a sentence is ambiguous, the focal point in mental space hovers between multiple attractors.

Coordinated Interplay Account

The *coordinated interplay account* is also based on recurrent neural networks (Knoeferle & Crocker, 2006). It differs from visitation set gravitation in that it explicitly takes visual context into account: included in its input layer is a set of units through which visual input is coded in terms of "event constituents". This network then operates in three stages. First, an incoming word is integrated into the interpretation of the sentence. Second, the interpretation and associated expectations guide attention to referents in the visual world or working memory. Finally, information from the word comes into the network through the aforementioned visual context units, in order to update the linguistic interpretation of the sentence. This model has been used to show that online comprehension is affected by contextual effects such as goal of comprehension and speaker gaze (Kreysa & Knoeferle, 2013).

3.1.5 Bayesian Approaches

The majority of models discussed thus far have been concerned with modeling the cognitive processes of reference resolution and the time course thereof. However, in recent years, there have also been a number of *Bayesian models* which seek to model reference resolution at the *computational level* (cf. Marr, 1982). Frank & Goodman (2012) model reference resolution with the following equation:

$$P(r \mid w, C) = \frac{P(w \mid r, C) \cdot P(r)}{\sum_{r_i \in C} P(w \mid r_i, C) \cdot P(r_i)} \tag{3.1}$$

Here, the prior probability $P(r)$ of referring to potential referent r is defined to be the *contextual salience* of object r, which combines the perceptual, social, and conversational salience of object. Frank & Goodman (2012) assume that the prior probability of selecting a particular referent from within a scene does not depend on the presence or absence of the distractors in that scene, rendering the choice of referent r and the context C independent. Due to this independence assumption, $P(r) = P(r, C)$, resulting in the abbreviated prior in the equation above. The likelihood term, $P(w \mid r, C)$ is defined as one divided by the number of referents in context C that "could be referred to" by word w. An algorithmic implementation of this model was then presented by Engonopoulos, Villalba, Titov, & Koller (2013).

This model was formalized by Goodman & Stuhlmüller (2013) into the *Rational Speech Act Theory*. According to this theory, listeners assume that speakers choose their utterances *approximately optimally*, and use Bayesian inference to invert their speaker model. Furthermore, this model assumes that people try to choose utterances that maximize *surprisal* for listeners so as to be maximally informative. This model has since evolved to handle a more general class of linguistic phenomena, as described by Goodman & Frank (2016). Finally, Heller, Parisien, & Stevenson (2016) have presented a Bayesian model of reference resolution which extends this model to account for perspective, as described in the previous subsection. Note that Bayesian models are not necessarily incompatible with the other models presented in this section; rather, they operate at different levels of analysis.

However, the validity of these models has been contested by some. Marcus & Davis (2013) points out that most Bayesian models of cognition fall into one of three categories. First, some models work on specific domains where Bayesian models happen to perform well, resulting in a disproportionate reporting of success, and an inability to account for related domains where people routinely make errors not predicted by such models. Second, some models present data in a way which obscures poor performance. Finally, some models make assumptions about prior distributions which are demonstrably false, or which could have just have easily been replaced with other, equally valid alternatives which would have yielded different results.

Marcus and Davis specifically point to the aforementioned work of Frank & Goodman (2012) as an example of the first group of models, as their model would not have been successful if they had chosen a winner-take-all decision strategy, which would have been consistent with their hypotheses but produced different results. Furthermore, Marcus & Davis (2015) cast Frank & Goodman (2012) as an example of the third group of models, due to their assumption that hearers know of all word choices available to the speaker. Finally, Frank & Goodman (2012) have also been critiqued by Gatt, van Gompel, van Deemter, & Kramer (2013) due to their inability to account for factors such as referential overspecification.

3.1.6 Future Directions: Beyond the Visual World

Reference resolution is a fascinating area of study, sitting at the intersection of language, memory, attention, and other areas of cognition. However, the majority of work surveyed in this section has looked at reference resolution in scenarios in which the intended referent is in the scene presented to the agent. Meanwhile, related work in the *visual search* literature has examined how people *search* for described objects in a visual scene(Huettig, Olivers, & Hartsuiker, 2011; Huettig, Rommers, & Meyer, 2011; Andersson, Ferreira, & Henderson, 2011). Neither of these areas cover how people resolve references to entities out of the field of view, such as 'the kitchen on the floor above us', or how people resolve references to entities which *cannot* be seen, such as those encountered in hypothetical statements such as “Imagine a red box; the box has a handle”, or in statements that refer to abstract entities such as sentences and events: such utterance types have typically been considered from a linguistics perspective.

Finally, there has been a dearth of research on *open-world* reference resolution, which occurs when the listener does not know of all referenced entities. For examine, imagine that you are told “My office is the room across from the kitchen on the second floor.” Upon hearing this sentence, you must determine not only the referents of *known* entities within this sentence, but also determine which entities are as yet *unknown* to you, and for which you must hypothesize new entities. Previously, Schlangen, Baumann, & Atterer (2009) presented an incremental Bayesian model of reference resolution in which a default decision of “undecided” is maintained until a candidate with posterior probability above some adaptive threshold is found. In the following sections, I will discuss alternative approaches I have developed.

3.2 Open World Spatial Reference Resolution

In Chapter 1, I discussed the importance of challenging the assumptions of certain, complete, centralized, homogeneous knowledge. As this chapter progresses I will demonstrate how all of these assumptions are challenged – but in this chapter, I will begin by presenting my initial work towards this goal, which challenged *specifically* the assumption of *complete knowledge*, while still holding to assumptions of certain, centralized, homogeneous knowledge. This research was performed within the area of *location-based spatial language grounding*, in which a robot must travel to described locations such as rooms, hallways, and doors.

Most approaches to location-based spatial language grounding have focused on traveling to described locations that are *known a priori* (e.g. S. Hemachandra, Kollar, Roy, & Teller, 2011; Kollar, Tellex, Roy, & Roy, 2010; Zender, Kruijff, & Kruijff-Korbayová, 2009; Shimizu & Haas, 2009; Chen & Mooney, 2011; Matuszek, Fox, & Koscher, 2010). These approaches all use some form of map to represent the robot’s environment which is either provided beforehand or built on the fly through exploration or a guided tour of the environment. For all of these approaches, it is assumed that any referenced location will be found in this map. What is more, most of these maps are “static” in nature, i.e., they will not (and are not allowed to) change from the point when reference resolution is performed.

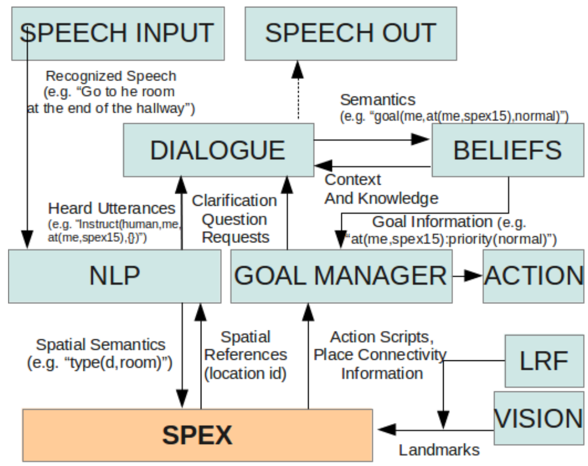
At the time at which this research was conducted, only Matuszek, Herbst, Zettlemoyer, & Fox (2012) were able to also travel to previously *unknown* locations, by parsing natural language utterances directly into action sequences which would bring the robot to the described location. However, this approach is a member of a subset of these previous approaches that must be categorized not as an approach to *reference resolution*, but rather as an approach to *directive grounding* – the objective is not to ascertain the identities of the referenced entities, but rather to successfully follow the provided instructions. This means that using this approach, a robot is not able to learn about new locations without being asked to travel to them; and the ascertained information about the location of a new location from a route instruction may be only of value at the location where it was provided. Furthermore, like the majority of the other mentioned approaches, that taken by Matuszek, Herbst, Zettlemoyer, & Fox (2012) is only able to handle natural language *commands*, and not, for example, statements or questions.

In this section, I will describe algorithms for location-based spatial reference resolution which are integrated into the aforementioned *DIARC* cognitive robotic architecture; the remainder of the work in this chapter is related to these algorithms in spirit and motivation, although that later work does not *directly* extend this previous work. These algorithms significantly improved on previous approaches by: (1) systematically creating representations for previously unknown referent locations, allowing robots to meaningfully communicate about unknown locations without having to first discover their exact location (e.g., by way of navigating there); (2) systematically creating representations for previously unknown *observed* locations, allowing robots to have natural language interactions about new environmental features discovered while navigating to a previously unknown place; and (3) generating action sequences only when they are actually needed to visit a referenced location (instead of immediately generating and following such sequences, the proposed approach stores the information in a location-independent form, which allows a robot to learn a map entirely through dialogue).

3.2.1 Algorithms, Architecture and Implementation

As previously stated, the algorithms presented in this section are integrated into the *ADE* implementation of the *DIARC* architecture described in Chapter 2. Figure 3.1 depicts the components within this *DIARC* configuration that are relevant to the processing of natural language utterances (hereafter

Figure 3.1: Spatial Cognition Architectural Diagram



Partial architecture diagram isolating the components and interactions relevant to natural language spatial cognition.

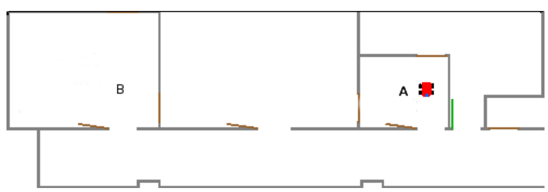
denoted by small caps such as NLP for "Natural Language Processing"), with a focus on the SPatial EXpert (SPEX), which employs all algorithms proposed in this section. SPEX receives information about landmarks from perceptual components in order to build a map of its environment. NLP queries SPEX in order to perform reference resolution on utterances it receives from the speech recognition component. NLP then sends utterance semantics to the dialogue manager (DIALOGUE). DIALOGUE uses contextual information from the belief modeling component (BELIEF) to perform pragmatic analysis on received semantics. BELIEF uses these semantics to inform the Goal Manager (GM) of new goals. The GM uses connectivity information from SPEX for path planning, and uses action scripts generated by SPEX when a destination has only been described to the robot and not yet visited (and thus might not have a precisely known location). It is necessary for the GM to rely on SPEX in these circumstances, as the best way to formulate a plan to these locations is by exploiting knowledge gained through dialogue interactions.

Spatial Semantics

Producing spatial descriptions entails identifying relationships between objects and modifiers, whether that is expressed in predicate semantics (as we do it), or in some other form (e.g., spatial description clauses (Tellex et al., 2011b)). To identify these relationships, we use a data-driven dependency parser which produces semantic predicates (Cantrell, Scheutz, Schermerhorn, & Wu, 2010).

Components such as SPEX interpret these semantic predicates in ways that are typically more limited than humans would interpret them. For example, a relational term like $to\_right(X, Y)$ [7] is interpreted by SPEX in the context of a route description in the following way: when the robot is past a landmark Y (e.g., a room or hallway), X is always to the right from the robot's perspective (whereas a human might more flexibly entertain additional interpretations that are not egocentric). To illustrate, imagine that the robot is located as shown in Figure 3.2.

Figure 3.2: Simulation Environment



A: Robot's initial position, B: The room at the end of the hallway.

A refers to the robot's current location, specifically the room in which it is located (as opposed to its location within that room). The robot is told "The room at the end of the hall is to the right", or $to\_right(B, A)$ where B is the name of the room at the end of the hall. The robot will assume that, as it is exiting A, B is to its right. So far, that is in accord with human intuition, and this description can apply to the labeled room on the map. However, imagine it is instead standing in the hallway and is given the same utterance. Now it thinks that B is to its right *as it is exiting the hallway*. This is no longer true of the labeled room, and does not match the human interpretation of the utterance.

With this in mind, let us examine the utterance "Go to the room at the end of the hall down to the right." *The room* is the only noun to which *at the end of the hall* could be attached; there is no ambiguity. By contrast, it is

[7] All formulae/predicates/terms/constraints in this dissertation will be represented as here, with uppercase letters representing variables and lowercase letters representing constants.

not clear whether *the hall* or *the room* is the noun to which *down to the right* should be attached. The parser chooses the most syntactically and lexically likely relationship. However, based on SPEX's interpretation of *to_right*, only the attachment to *room* will be successful. If it is attached to *hall*, the robot will assume that, as it is exiting A into the hallway, it will still need to find a second hallway that is to its right and which is connected to the room in question. Thus, the only acceptable semantic representation for the phrase "the room at the end of the hall down to the right" is $(endroom(X) \wedge hall(H) \wedge connected\_to(X, H) \wedge to\_right(X, A))$. If the attachment were to the hallway, the semantics would instead be $(endroom(X) \wedge hall(H) \wedge connected\_to(X, H) \wedge to\_right(H, A))$. Note that SPEX and NLP together resolve such ambiguities in a fixed manner, in that "higher attachments" are always preferred in the parser. In other words, because *at the end of the hall* is already attached to (is dependent on) *the room*, attachment to *the room* is preferred over attachment to *the hall*.

Spatial Reference Resolution and Exploratory Route Suggestion

SPEX aggregates information about its environment in order to build a hierarchical map M stratified into two layers, similar to the mapping approaches presented by Kruijff, Lison, Benjamin, Jacobsson, & Hawes (2007) and Kuipers (2000). The *top layer* $M_{top} = (V_{top}, E_{top})$ is a graph with vertices V_{top} and edges E_{top} where each vertex $v \in V_{top}$ is a *large-scale* place such as a room or hallway, and where each edge $e \in E_{top}$ represents a means to travel between such places (i.e., through a doorway). The *bottom layer* $M_{bot} = (V_{bot}, E_{bot})$ is a graph with vertices V_{bot} and edges E_{bot} where each vertex $v \in V_{bot}$ is a *small-scale* place: a specific location in a room or hallway, or a landmark such as a door, and where each edge $e \in E_{bot}$ represents a path between these places. A vertex v in either graph is indexed by a uniquely referring identifier, and contains an adjacency list of connecting places' identifiers and a list of properties held by the represented place. The primary difference between the two levels is that M_{top} is only concerned with whether or not its vertices connect (e.g., whether or not a room is accessible from a given hallway), while M_{bot} is additionally concerned with the details of where and how its vertices connect. For this reason, the metric positions of M_{bot}'s vertices are stored when known, while such information is not maintained for vertices in M_{top}. Topological information, such as an ordering of places within a hallway, can be extracted from M_{bot} using the coordinates of V_{bot}. Each large-scale place in V_{top} also stores a list of places from V_{bot} that it contains. This stratification is conceptually motivated by the fact that natural language references to spatial locations are typically concerned with *large-scale* places, whereas sensing and planning systems are typically concerned with the *small-scale* places contained within.

Information used to augment this map can come from both perceptual components and dialogue. As the robot travels through its environment, SPEX actively requests information from various perceptual components, e.g, it requests information about landmarks from the Laser Range Finder (LRF) component. If LRF has detected a landmark, it returns its coordinates to SPEX, as well as coordinates necessary for establishing the landmark's orientation or for navigating through traversable landmarks such as doors. SPEX then uses the algorithm summarized in *detectLandmark* (Algorithm 1) to process this information.

detectLandmark seeks to determine whether a landmark has been seen before, and if not, to build a new representation for the landmark and any new locations its observation entails (e.g., rooms on the other sides of observed doors). To do so, it requires the coordinates of the observed landmark (D), the coordinates that will allow one to approach (T) and depart (O)the landmark. In this algorithm, it is assumed that landmarks are *doors* that separate large-scale topological spaces (e.g., rooms and hallways), and thus the set of coordinates T and O are assumed to be in the current and adjoining topological space, respectively. In addition, *detectLandmark* requires a list of predicates P describing any other properties the landmark might have. For example, if a camera is being used, visual characteristics of a landmark might be included.

detectLandmark then operates as follows: if the landmark can be identified as a known landmark based on its coordinates, nothing more must be done, and the subroutine immediately returns (Algorithm 1, Lines 2-3). Otherwise, *detectLandmark* constructs a list of locations that are *known* but *unobserved* (such as those previously described in natural language) whose properties include those which the landmark is observed to have (Lines 5-8). If there are no such known but unobserved locations, then the landmark is assumed to represent an as-yet-unknown location, and new representations are created and added to M_{bot} for the three small-scale locations implied by this observation, as well as any large scale places entailed, such as a room assumed to be on the other side of an observed door (Lines 11-13). SPEX determines whether such an adjoining space is a room or hallway using the simplifying heuristic that rooms only connect to hallways and not to other rooms. A more general solution would be to postpone this decision until this property can be verified through exploration. This would require more sophisticated exploration strategies, which I will later discuss.

Algorithm 1 detectLandmark(D, T, O, P)

1: (D,T,O are Coordinates, P is a list of Predicates)
2: **if** D is a known landmark **then**
3:     **return**
4: **end if**
5: $m = \emptyset$
6: **for all** $\phi \in M_{bot}$ such that $loc(\phi) == \emptyset$ **do**
7:     **if** $P \subset \phi.properties$ **then**
8:         $m \leftarrow m \cup \{\phi\}$
9:     **end if**
10: **end for**
11: **if** $m == \emptyset$ **then**
12:     add places ϕ_D, ϕ_T, ϕ_O at coordinates D, T, O to M_{bot}
13:     $properties(\phi_D) \leftarrow properties(\phi_D) \cup P$
14: **else**
15:     $c \leftarrow$ the room the robot is currently in
16:     $t \leftarrow$ the place connected to n in c
17:     $o \leftarrow$ the other place connected to n
18:     $loc(m[0]) \leftarrow D; loc(t) \leftarrow T$; $loc(o) \leftarrow O$
19:     $children(c) \leftarrow children(c) \cup \{t\}$
20:     $connect(c, parent(o))$
21: **end if**

If, however, there *do* exist some number of previously known yet unobserved locations with properties that match those of the observed landmark, then the first of these is selected, and is grounded using the set of coordinates D. Furthermore, the small-scale locations assumed to be on either side of this now-grounded landmark are also grounded using T and O, and the representation for the large-scale location associated with the newly grounded landmark is connected to the robot's current large-scale location at both the large-scale and small-scale levels (Lines 14-20).

When NLP sends SPEX information regarding a received utterance, it is as a list of predicates P representing the semantics of the utterance that convey some information about the structure of the environment. This can be seen in *processSemantics* (Algorithm 2). SPEX separates these predicates into three categories: *type predicates* (predicates that express the type of an entity, such as $room(X)$), *descriptive predicates* (predicates that describe an entity, such as $color(X, green)$), and *relational predicates* (predicates that describe relations between entities, such as $connects(X, Y)$).

This separation is useful since each type of predicate plays a different role in the process of reference resolution. When SPEX receives a predicate list

Algorithm 2 processSemantics(P)

1: (P is a list of Predicates)
2: $M \leftarrow$ new Map
3: **for all** type predicate $p \in P$ **do**
4:     $M(vars(p)[0]) \leftarrow buildBindings(p)$
5: **end for**
6: **for all** descriptive predicate $p \in P$ **do**
7:     $pruneBindings(p, M)$
8: **end for**
9: $P^R = relationalpredicates \in P$
10: $pruneRelationBindings(P^R, M)$
11: **if** $(\exists L \in M | L == \emptyset)$ **then**
12:     **return** $createPlaces(P, M)$
13: **else if** $(\exists L \in M | size(L) > 1)$ **then**
14:     **return** $ambiguous$
15: **end if**
16: **return** $M(args(P[0])[0])$

from NLP, it attempts to determine the identities of any locations referenced in the utterance (i.e., specified in a type predicate with name 'room', 'door' or 'hall'). For each landmark or large-scale location referenced by a type predicate in P, SPEX constructs a new list of candidate identifiers (Algorithm 2, Line 4). For example, the type predicate $room(R)$ will result in the creation of a list which initially contains the identifiers of all known rooms[8].

SPEX then uses the descriptive and relational predicates to eliminate bad candidate locations from these lists. First, descriptive predicates are used by the *pruneBindings* routine to quickly eliminate incorrect candidates from M (Line 7). At this point, the only constraints left to consider are the relational constraints. These are used by *pruneRelationBindings* to reduce each list in M to its smallest size (Line 10). This process can be perhaps best be understood by casting it as a constraint satisfaction problem (CSP, see also Kumar, 1992) with variables $X = keys(M)$ such that the set of domains $D = values(M)$, and set of constraints P^R, i.e., the set of relational constraints. The process effected by *pruneRelationBindings* can then be viewed as equivalent to the process of finding all solutions to that CSP.

Once SPEX has reduced each list to its smallest size (i.e., by finding the set of CSP solutions), it examines each list (Algorithm 2, Line 11). The size of the lists can be used to classify the level of ambiguity in the utterance. If a list

[8]While this approach was acceptable for our initial foray, it is important to realize that it may become unacceptable as the number of locations grows

Algorithm 3 getScript(S, D)

```
(S is the id of the source, D is the id of the destination)
A is a new action script
if small-scale(S)AND small-scale(D)AND same-or-adjacent(par(S), par(D)) then
   if locationKnown(D) then
      if adjacent(par(S), par(D)) then
         A = A ∪ (move through doorway towards D)
      end if
      A = A ∪ (go to D)
   else
      useCluesToPlanMotion(S, D, A)
   end if
else
   D′ = nearestGrounded(D)
   I = planSmallScaleRoute(S, D′)
   if D′ ≠ D then
      I = I ∪ planLargeScaleRoute(D′, D)
   end if
   for all i ∈ I do
      A = A ∪ (go to i)
   end for
end if
return A
```

is empty, then the description corresponds to a previously unknown place. In practice, this may be incorrect since the description could actually refer to a known location in a way that is not currently determinable (e.g., if the robot knows of a certain door, but does not know that the room beyond it is the cafeteria, then it will not be able to automatically resolve a reference to 'the cafeteria'), but this discrepancy will need to be resolved during exploration or during further dialogue. If the received semantics came in the context of an assertion about the world as opposed to a query regarding its structure, SPEX adds a new entry to the appropriate map level, and then adds to this new entry any relevant properties from P. In the case of the room at the end of the hall, the list corresponding with the referenced room is empty, so SPEX creates a new large-scale place to represent the room, along with small-scale places in M_{bot} representing the door which must connect it to the hallway and the places on either side of this door. SPEX then gives the newly created door the property $end\_of\_hall(d, h)$, indicating where in the hall it is located. This is necessary since SPEX would otherwise not be able to identify the door as being at the end of the hall, as its coordinates are as of yet unknown.

If a list contains a single identifier, such as in the case of the list associated with the hallway, SPEX assumes this is the identifier of the referenced location and modifies that place's connections and properties accordingly.

This brings up one of the difficult problems which SPEX must deal with. If the robot is given a description of a place whose location it does not know, SPEX needs to create a representation of that place without specifying its metric location. If it is informed of some series of connected rooms and hallways, their topological representations should be linked to the known map only if their locations are known relative to some known place.

If the robot is later able to determine the precise location of one of the rooms, its child locations in M_{bot} can then be given metric positions. In the case of the room at the end of the hall, SPEX creates place representations in M_{bot} for the door and the points of access on either side of it, but sets a property in each of these representations indicating that its coordinates are unknown. The identifiers for these places are then placed into a list of unknown places which is considered whenever a new place is seen.

If there are multiple candidate identifiers for a described place, SPEX informs NLP that more information is needed to disambiguate between the candidates (Line 14) – determining whether and how to choose between multiple candidates is an interesting topic. SPEX could, for example, return a distribution representing the relative likelihoods of the various candidates, and allow NLP to decide for itself how to resolve this ambiguity, or it could create and assert the conveyed information for all candidate locations, along with some diminished confidence value. This would also be useful if NLP needed to partially assert two alternate semantic interpretations of a received utterance. These are issues I will return to in subsequent chapters.

Another important capability addressed by SPEX is the generation of actions to reach locations whose metric locations are unknown. Since the semantics in the case of "the room at the end of the hall down to the right" involve clues about the location of the place (i.e., it is at the end of a hallway, in a room "to the right" of the current one), SPEX is able to produce a possible action sequence to reach the target location. This process is summarized in Algorithm 3. When SPEX is asked for an action script which, when executed, will take the robot between two locations S and D, it creates a new action script A, and adds actions to this script based on the properties and connectivity of S and D.

If S and D are small scale locations, in the same or adjoining large scale locations, and D's locations are known, the script returned by SPEX only has to contain an instruction to travel to D (Line 8), and, in the latter case, an instruction to move through the doorway connecting the two rooms (Line 6). If one of these relationships hold but D's location is unknown (Line 10), SPEX uses a set of rules of thumb to determine if it knows how to travel towards D (i.e., whether it is known that one location is 'to the left' of the other, with respect to some landmark). It should be noted that these rules of thumb were designed to operate specifically within simple room-and-hallway networks.

If, on the other hand, neither of these relationships hold, SPEX uses a standard search algorithm to plan a route through small-scale space from the starting location S to the nearest grounded location D' to the destination D (Line 14). If this location is not D itself, i.e., if D's location is unknown, then appended to this route is the shortest path from D' to D through large-scale topological space (Line 16). An instruction to travel to each point along this full route is appended to A (Line 19). Finally, A is returned.

For the example sentence of the room at the end of the hall, the action sequence is formalized in this manner:

```
[moveTo, self, exitposition]
[exitRoom, self]
[moveTo, self, entryposition]
[moveTo self, currentroom]
[turnRel, self, ang]
[traverse, self]
[informSpexEnd, AtEndOfHall]
[moveTo, self, destination]
```

Note that since SPEX is unable to directly detect "end-of-hall-ness", the created script includes a request `informSpexEnd` to be informed when the exploration of the hallway is completed. SPEX finally alerts *DIARC's* Goal Manager (GM) component of any new places it learned of and the connections between them.

When the GM receives the above script and and issues execution, the robot exits the room, turns in the direction indicated, and starts driving down the hall. When the robot reaches the end of the hallway, the GM informs SPEX that the exploration has finished. SPEX checks whether any nearby location is close enough to be construed as being "at the end of the hall." It then examines all places connected to the current hallway, and checks to see if any of them have "end-of-hall-ness" listed among their properties (in this case, the previously described room does). Assuming one place fits this description (at this stage in our research we did not attempt to handle ambiguity resolution), SPEX consolidates its representations of the recently encountered place and the described place, placing into a consolidation map the identifier of the place that is consolidated away, in case the old reference is used by some other component.

3.2.2 Evaluation

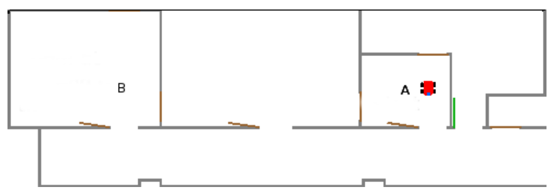
We ran three sets of evaluations of SPEX. In the first two, SPEX alone was evaluated, and in the third, the integrated architecture was tested. In order to abstract away from the concerns of other architectural components such as NLP, SPEX was provided with a starting location and gold standard semantics for the utterance being tested which uniquely identified a location; the robot was not asked, for example, to go to "the room at the end of the hall" in an environment in which several rooms existed at the hallway's end. If these types of requests and environments had been included in testing, performance would have decreased.

In the first evaluation, SPEX was given a full map of an environment, and 64 resolution tests, which represented all ways that a set of utterances (such as "the room to your immediate left when exiting the break room", "the room at the right end of the hallway" and "the third room on the right facing left from your current position") could be successfully resolved in the environment. For example, "the room to your immediate left" was evaluated from all starting points that had a room on their immediate left. SPEX generated the correct reference for 64/64 (100%) of the tests.

In the second evaluation, SPEX was given a partial map of the same environment; 44% of the large-scale locations were removed, along with all contained small-scale locations and any connecting doors. SPEX was then given all 34 tests from the original set of tests whose starting location was still known. Since some destinations were unknown in this set, success in the case of an unknown destination was qualified as generating a new place representation and returning a plan which would successfully take the robot from its current location to the location. SPEX passed 34/34 (100%) of these tests.

Finally, the complete architecture using SPEX was tested in a simulated environment on a set of utterances. We used a simulated MobileRobots Pioneer robot, although the remainder of the architecture ran in the same configuration that it would on the real robot. We first gave the command "Go to the room at the end of the hallway down to the right" to the robot in the simulated environment pictured again in Figure 3.3. The robot exited the room and proceeded to the right end of the hallway. Examining SPEX's map showed that SPEX had successfully consolidated its representations of the rooms the robot had heard referenced in natural language and observed at the end of the hall. Thus, the original reference was successfully resolved to its physical location.

Figure 3.3: Simulation Environment



A: Robot's initial position, B: The room at the end of the hallway.

We also evaluated some basic exploratory functionality for resolving ambiguous statements. Consider the command "Go to the room at the end of the hallway." In an unknown environment, this will result in the GM asking SPEX for an action script, which will need to be formed using the *useCluesToPlanMotion* (Algorithm 3, Line 10) function. When this function tries to determine which end of the hallway it needs to send the robot to, it will determine that the room could be at either end of the hallway. It thus chooses one of the ends and adds the necessary instructions to the action script. It then creates a new script to return to the choice point and travel to the other end of the hallway, and stores this second script in an "alternate plan" list indexed by the destination point. SPEX then returns the first action script. When the system follows this script and travels to the first end of the hallway, the last action it will execute will be to move to the destination point. If the reference is successfully resolved, it will move to that point. If it is not, the GM will once again ask SPEX for an action script. SPEX will check its alternate plan list and see that there is a plan waiting for that destination, and will remove and return it to the GM. Assuming the robot's interlocutor did not give an instruction to go to a nonexistent location, this plan will lead it to the target location. We tested this in the manner of the first two steps of evaluation and achieved successful results, as evidenced in the produced action scripts:

```
[moveTo, self, spex12]
[turnRel, self, -1.5708]
[traverse, self]
[informSpexEnd, AtEndOfHall]
[moveTo, self, spex14]

[moveTo, self, spex12]
[traverse, self]
[informSpexEnd, AtEndOfHall]
[moveTo, self, spex14]
```

3.2.3 Discussion

The above evaluations showed that SPEX was able to successfully resolve spatial references to both known and unknown locations as long as the spatial semantics picks out places uniquely. Storing the information gleaned from natural language and through exploration in a location-independent format affords the robot improved capabilities. Specifically, it allows the robot to (1) travel to previously described locations, (2) describe how two unknown locations are positioned relative to each other, (3) pause an action sequence and then later resume it from another location, and (4) return to a known location after visiting an unknown one. Finally, augmenting the robot's world model based only on descriptions allows a robot to learn a map purely through dialogue if it is able to extract sufficiently accurate semantics representations, while none of the approaches mentioned in the introduction would able to learn a map of their environment without physical exploration from dialogue alone.

Despite these improvements, SPEX had several shortcomings, specifically in situations where the attempt to resolve a spatial reference produces either no candidate places, or several appropriate candidates. Consider the instruction "Go to the cafeteria": if the robot knows of no cafeterias, what heuristics should it use to determine where to explore? Clearly, unless the robot has some notion of where cafeterias are usually located (e.g., in buildings like the current one), this will be very challenging. One strategy might be to simply ask a human for help. If that is not feasible or not allowed, another strategy might be for the robot to start exploring its environment, even when it has no notion of the goal location (cf. Hawes et al., 2011). Sometimes a combination of strategies may be called for – identifying the best strategy for a given situation is in itself a challenging open research problem.

Another condition in which SPEX was not able to resolve references was when the robot could identify several candidate referents. For example, if the robot is told to go to the cafeteria and it knows of several cafeterias, how is it to determine the intended one? In the final part of our evaluation, I presented one possible approach. Later in this dissertation, I will present another – to ask for clarification. The exploration-based technique presented in this work could be improved in a variety of ways, such as prioritization of exploration based on relative likelihood, the use of other experts (e.g., an Episodic Memory Expert), or the modeling of the beliefs and knowledge of other agents.

Furthermore, in our evaluation of SPEX, we did not require SPEX to handle underspecified descriptions of locations, which typically happen in natural interactions. A place could easily be described in a way which fails to mention important details that are necessary for determining its location. In such a case, a representation for the described place would have been added to the map, but SPEX would either have been unable to generate a plan to reach it, or it would have never been able to recognize the place when it was encountered. There are additional complications that would impact the performance of SPEX, for example, the environmental complexity (including multiple intersecting hallways with loop closure, multi-level spatial layouts with connections among the levels, and others). Finally, our evaluation also assumed reliable perceptual information, but this is rarely the case in practice. For example, if the robot is sent to the third room in a hallway but fails to notice one of these doors, many problems will arise. In the second of our evaluations, we counted a test case as successful if SPEX was able to generate an appropriate action plan, but did not check whether the robot made mistakes while carrying out those plans as this would require additional action monitoring mechanisms to detect action failures and mechanisms to recover from them.

In this section, I presented SPEX, an architectural component capable of resolving references to unknown locations in an indoor environment in a manner that allows a robot to discuss and reason about such locations without having to visit them first. I discussed how SPEX's capabilities are facilitated by its interaction with other components of the *DIARC* architecture. However, I also identified a number of serious shortcomings, including assumptions of certain, unambiguous knowledge in a single domain and representation scheme. In the next section, I present a more general framework for referential processing that eliminates these shortcomings.

3.3 Probabilistic Open World Reference Resolution

In the previous section, I presented an algorithm for open-world location-based spatial reference resolution. But as was shown, it was subject to a number of assumptions that limited its viability in realistic human-robot interaction scenarios. In this section, I will present a more general architectural framework and set of algorithms to perform reference resolution within that framework. In order to do so, let us first reexamine the reasons why reference resolution is particularly difficult for agents such as robots: reasons both external (i.e., relating to their environments), and internal (i.e., relating to their architectures).

Externally, robots must contend with environments that are uncertain and unknown. Suppose a robot is asked "Can you bring this mug to the kitchen and put it in the sink?" Because the robot's knowledge of what locations may or may not be kitchens is likely inferred from evidence provided by noisy sensors, it may be *uncertain* as to whether a particular location is or is not a kitchen; but it should still be able to identify and assess candidate referents. Similarly, because the robot likely does not have perfect knowledge of every object in the building in which it works (let alone other buildings), it may simply never have seen a sink that's in the building's kitchen; but it should still be able to discuss, reason about, and travel to the described sink, perhaps by creating a new internal representation of the described sink.

Internally, these difficulties are exacerbated by the organizational properties of integrated robot architectures. Most previous algorithms for reference resolution (and, as well, for referring expression generation, as I will discuss in Chapter 5), assume that information about possible referential candidates is stored in a single, easily accessible knowledge base. This is usually a reasonable assumption given the context in which most such algorithms are developed: most reference resolution algorithms, for example, have traditionally focused on resolution *within a single domain*, such as large-scale topological locations (as we did in Section 3.2) or objects perceived in a visual scene (as have many of the psycholinguistic models covered in Section 3.1). Unfortunately, this is *not* a reasonable assumption in most of the use cases for which robot designers seek to enable natural language capabilities, such as search and rescue robotics or assistive robotics. In these use cases, it is likely that a robot will need to resolve references to entities from a wide variety of domains, including people, places, objects, and utterances.

This is problematic because many integrated robot architectures used on robots operating in such domains, such as ROS or our own *DIARC*, store information about entities from these various domains in a distributed fashion across architectural components, rather than in a single centralized knowledge base. Furthermore, these distributed knowledge bases may use a variety of representational formats that facilitate domain-specific reasoning: information about locations may be stored as a hybrid metric-topological map in a mapping component; information about objects may be stored as a scene graph in a vision component; information about people may be stored in a Prolog knowledge base in a social knowledge component; and information about recent utterances may be stored as a discourse tree in a dialogue component.

Before reference algorithm algorithms can be designed that overcome the limitations suffered by SPEX, a new architectural framework must be developed within *DIARC* to facilitate such algorithms.

3.3.1 Architectural Framework

In our architecture, the set of architectural components capable of providing information about entities that might be referenced by a robot or its interlocutors can be viewed as a set of *distributed, heterogeneous knowledge bases* (DHKBs). These are viewed as abstract *knowledge bases* because they are assumed to represent and reason about information of interest to the robot; they are viewed as *heterogeneous* because each such knowledge base may have its own unique means of representing and providing access to such information; and they are viewed as *distributed* because each such knowledge base may be physically located on a different machine.

DHKBs

Due to the heterogeneous nature of these knowledge bases, we cannot assume a unified means of access to the information stored in a given knowledge base. What is more, it cannot be assumed that all information in a given knowledge base should even be accessible through introspection by an agent's higher level cognitive processes. In humans, for example, it is uncontroversial that the set of information introspectable by higher level cognitive processes is much smaller than the total amount of information processed by the brain (Nisbett & Wilson, 1977). Similarly, it is possible that only a small subset of the information stored in a given knowledge base will be introspectable by a robot's higher level cognitive processes.

In order for a DHKB to be useful, however, we assume that some subset of its information *can* be introspectively accessed. Furthermore, we make the following assumptions about such information:

1. We assume that, for a given DHKB, there exists some set of *entities* known of by that DHKB *about which* information is accessible through introspection, and that these entities can be enumerated using a set of unique identifiers.

2. We assume that all information that can be introspectively accessed can be *described* using *positive arity predicate symbols*, i.e., *properties* (unary predicate symbols) and *relations* (polyadic predicate symbols (i.e., predicate symbols with arity greater than one))[9].

3. We assume that all information that can be introspectively accessed can be *assessed* (i.e., it is possible to determine *to what extent* a given property or relation holds for a given entity or set of entities).

4. We assume that all information that can be introspectively accessed can be *imagined* (i.e., it is possible for the DHKB to store, on command, information about a new hypothetical entity that has an arbitrary set of properties and relations

For example, consider the case of a DHKB taking the form of a *hybrid metric-topological map*. The DHKB may store information in a variety of forms, e.g., occupancy grids used at the metric mapping level. But it may be the case that the only information that can be introspected upon are the concepts of *Rooms*, *Hallways*, and *Connectivity*. This DHKB may have a set of large-scale *places* {p1,p2} it knows of, which constitute its set of entities. It must be possible to assess the degree to which it is believed that each of {room(p1),room(p2),hallway(p1),hallway(p2),connects(p1,p2),connects(p2,p1)} hold, and it must be able to *imagine* a new entity (e.g., p3) such that, for example, each of {room(p3),connects(p1,p3),connects(p3,p1)} can be consciously willed to be true to any degree desired.

Consultants

Our assumptions given DHKBs are exploited using a set of *consultants*, each of which provides an interface to a single *type* of information stored in an

[9]It is crucial to understand, however, that (1) no assumption is made that information is *stored* in this form – it is merely *describable* in this form, and that (2) no assumption is made that such descriptions are made with perfect certainty.

agent's DHKBs. Typically, each consultant serves as the interface to a particular DHKB. The interface provided by a consultant c must be capable of the following four functions:

1. Providing the set c_{domain} of atomic entities about which information is assessable through introspection in its associated DHKB;

2. Advertising a list $c_{constraints}$ of *constraints* (i.e., typed positive-arity predicate symbols) that can be assessed with respect to known entities from c_{domain} or asserted with respect to imagined entities. These constraints are assumed to be listed in descending order of *preference* (see also Dale & Reiter, 1995);

3. Assessing the extent to which constraints from $c_{constraints}$ can be said to be true about entities from c_{domain}; and

4. Asserting newly imagined entities into c_{domain} such that some arbitrary set of constraints from $c_{constraints}$ are true to a desired extent (an action known hereafter as *hypothesization*).

Above, I stated that the constraints advertised by a consultant are *typed* positive-arity predicate symbols. By this, I mean that a consultant must specify the consultants responsible for handling each of the entities that can be bound to the arguments of an advertised constraint. For example, a consultant named *obj* responsible for knowledge regarding objects may advertise $in(X : obj, Y : obj)$, indicating that it can, for example, assess whether one entity from obj_{domain} can be considered to be 'in' another entity from obj_{domain} (for example, whether a given book is in a given box)[10]. This is important for two reasons.

First, we assume that multiple consultants may advertise constraints that are identical in name and arity, but that differ with respect to type information (but that no consultants advertise constraints that are identical with respect to name, arity, *and* type information). For example, a consultant named *loc* responsible for knowledge regarding locations may advertise $in(X : loc, Y : loc)$, indicating that it can, for example, assess whether one entity from loc_{domain} can be considered to be 'in' another entity from loc_{domain} (for example, whether a given room is in a given building).

[10]Note that here we informally use the word 'object' to refer to small, everyday objects, and use the word 'locations' to refer to large-scale spatial locations such as rooms or hallways. It is of course possible to view a room as an object or the inside of a cup as a location, but here, and throughout this dissertation, we will use the more everyday sense of these words, and define consultants and algorithms with this sense in mind.

Second, for a constraint q advertised by consultant c with arguments q_{args}, we only assume that *at least* one argument in q_{args} must be bound to an entity from c_{domain}. This relaxed assumption allows *obj* to advertise, for example, $in(X : obj, Y : loc)$, indicating *obj* can assess whether a given object is in a given *location* about which more information can be found by querying consultant *loc*. In integrated agent architectures in which information is distributed across multiple DHKBs, connections often exists enabling communication of information between those components. For example, when a new object is seen, a vision component may query a mapping component for the current location, so it can be stored along with the new object's representation.

Referential Executive

The final facet of our architectural framework is *REX*, the Referential Executive. REX is a *DIARC* component that manages and makes use of the *consultants* available within a robot architecture. Whenever a new component that fills the role of *consultant* registers with the *DIARC* architecture, REX is automatically notified, at which point REX records that consultant's name and the constraints it advertises.

3.3.2 Notation

Before I move on, I introduce the notation used throughout this chapter.

Λ	The set of logical formulae (or *constraints*) $\lambda_0 \ldots \lambda_{\vert\Lambda\vert}$ associated with the current variable of interest in the semantic *connotation* of an incoming utterance.
$\Lambda[t]$	The aforementioned set of logical formulae under typing t.
Λ^V	The set of variables used in all logical formulae in Λ.
λ^V	The set of variables $\{\lambda^{v_0}, \ldots, \lambda^{v_{\vert\Lambda^V\vert}}\}$ used in logical formula λ.
$V[t]$	The *ordered list* of free variables found in Λ under a typing t mapping variables to consultants.
C	A set of *consultants* $\{c_0, \ldots, c_{\vert C\vert}\}$.
c^Q	A set of *query templates* $\{c_0^q, \ldots, c^{q_{\vert c^q\vert}}\}$ advertised by consultant c.
C^Q	The set of *query templates* $\{{c_i}^Q \cup \cdots \cup {c_{\vert C\vert}}^Q\}$ advertised by all consultants.
M	A robot's *world model* of entities $\{m_0 \ldots m_{\vert M\vert}\}$ found in the domains provided by the robot's various consultants.
$M_{v[t]}$	The subset of M provided by the consultant associated with variable v under typing t.
Γ	A binding hypothesis, i.e., a set of bindings from variables in Λ^V to entities in M, denoting the semantic *denotation* of the current variable of interest of an incoming utterance.
$\tilde{\Gamma}$	A set of binding hypotheses $\{\Gamma_0, \ldots, \Gamma_{\vert\tilde{\Gamma}\vert}\}$ under consideration.
Γ^V	The set of variables $\{\Gamma^{v_0}, \ldots, \Gamma^{v_{\vert\Gamma^V\vert}}\}$ used in binding hypothesis Γ.
Γ^Λ	The set of formulae $\{\Gamma^{\lambda_0}, \ldots, \Gamma^{\lambda_{\vert\Gamma^\Lambda\vert}}\}$ that have not yet been considered for binding hypothesis Γ.
Φ	A *satisfaction* variable which is *True* iff all formulae in Λ *hold* when bound using Γ.
Γ^P	The probability of satisfaction for binding Γ; Superscript notation is used for consistency and to distinguish this possibly partially computed probability value from the final value $P(\Phi \mid \Gamma, \Lambda)$ obtained for a complete binding Γ, after considering all $\lambda \in \Lambda$.

Here, I use Mill (1884)'s terminology of connotation and denotation for simplicity; this should not be taken as a commitment to any particular theory of meaning.

3.3.3 Mechanisms to Facilitate Reference Resolution

We are now ready to examine how reference resolution is facilitated using the presented architectural framework. When a robot receives an utterance, it is parsed into an *utterance structure* with a set of *additional semantic content*. For example, the utterance "The medkit is on the shelf in the breakroom" may be parsed into the utterance structure $Statement(speaker, self, on(X,Y))$ with additional semantic content $\Lambda = \{medkit(X), shelf(Y), breakroom(Z), in(Y,Z)\}$. The additional semantic content associated with each variable found in the utterance structure are assumed to be distinct, and are dealt with separately. For each variable found in the utterance structure, its associated additional semantic content is sent to REX for resolution, along with an *ordering* over the variables found in those semantics, based on considerations such as prepositional attachment. For example, when processing the second clause of the above utterance, REX would receive $\Lambda = \{shelf(Y), breakroom(Z), in(Y,Z)\}$ and the ordering $\{Y, Z\}$, determined because entity Y is being described *with respect to* entity Z.

Before REX can perform reference resolution, it must determine which consultant should be used to provide information about each variable. That is, REX must find the most probable mapping from variables to consultants, given the provided (untyped) constraints and the the (typed) constraints advertised by REX's consultants.

The process of associating a consultant with each variable is viewed as the process of finding the optimal mapping $t : \Lambda^V \rightarrow C$ from variables in Λ^V to consultants in C, drawn from set of possible mappings T:

$$\operatorname*{argmax}_{t \in T} \prod_{\lambda \in \Lambda} P(t|\lambda).$$

Here, $P(t|\lambda)$ represents the probability that mapping t correctly maps variables to consultants given that λ appears in Λ. This can be calculated in one of two ways. If a training corpus is available, $P(t|\lambda)$ can be calculated by consulting the learned conditional distribution $P(T|\lambda)$. Otherwise a uniform distribution may be assumed, and $P(t|\lambda)$ can be calculated as:

$$P(t|\lambda) = \begin{cases} 0, & \text{if } \delta = 0. \\ 1/\delta, & \text{otherwise.} \end{cases}$$

$$\text{where } \delta = \sum_{c \in C} \sum_{q \in c^Q} |matches(q, \lambda)|$$

.

Here, $|matches|$ is the number of query templates in c^Q that can be unified with λ.

Once a constraint typing has been chosen, and used to transform the set of constraints associated with a variable into a set of *typed constraints*, those constraints are provided to the *DIST-POWER* reference resolution algorithm. In the following three subsections I will describe how the process of reference resolution may be modeled within the presented architectural framework, and how that model is implemented by the *DIST-CoWER* and *DIST-POWER* algorithms.

3.3.4 Model

Using the previously provided notation, we can create models of both *closed world* and *open world* reference resolution. Closed world reference resolution is the problem of finding the optimal association from referenced entities to known entities, that is, the optimal set of bindings $\Gamma\star$ from variables in Λ^V to entities in M. We define optimality as highest probability of maximizing a satisfaction variable Φ, and assume independence among constraints such that the probability of a set of variable bindings satisfying a set of constraints is equivalent to the product of the probabilities of those variable bindings satisfying each individual constraint, *as assessed by REX's Consultants*. This produces the following model for closed-world reference resolution:

$$\Gamma\star = \underset{\Gamma\in\tilde{\Gamma}}{\text{argmax}} P(\Phi \mid \Gamma, \Lambda[t]) = \underset{\Gamma\in\tilde{\Gamma}}{\text{argmax}} \prod_{i=0}^{|\Lambda[t]|} P(\phi_i \mid \Gamma, \lambda[t]_i) \tag{3.2}$$

However, in an *open world* in which new entities may be introduced through dialogue, it may be impossible or inappropriate to associate *all* variables in Λ^V to entities in M, as some subset of those variables should perhaps instead be associated with new, previously unknown entities. Thus, open world reference resolution can be characterized as the problem of finding both (1) the optimal division of variables in Λ^V into those that should be resolved to existing representations and those for which new representations should be created, and (2) the optimal set of bindings from variables in the former subset to entities in M. In order to help find the optimal division of variables, it is assumed that the variables in Λ^V are listed in increasing order of likeliness of being part of this subset, as I describe in Williams & Scheutz (2015b). Using this assumption, the first half of the problem can be cast as that of finding the longest *suffix* of Λ^V such that the probability of satisfaction for the best *closed world* solution using that suffix is above some threshold τ_{cower} [11]. Let ${\Lambda^V}^i$ be the suffix of Λ^V containing its last i elements, let $\Lambda[t]^i$ be the subset of (typed) constraints that only involve those variables, let Γ^i be the space of candidate variable bindings from variables in ${\Lambda^V}^i$ to entities in M, let $\Gamma^i\star$ be the solution to the closed world reference resolution problem for Γ^i, and let $\Gamma^i\star^P$ be the probability associated with that best solution, and let $complete()$ be a function that completes a partial solution by hypothesizing new representations for unbound references. The problem of open-world reference resolution can then be modeled as:

$$\Gamma\star = complete\left(\underset{\Gamma^i\star\in\{\Gamma^0\star,\ldots,\Gamma^{|\Lambda^V|}\star\}}{\operatorname{argmax}} \begin{cases} i, & \text{if } \Gamma^i\star^P > \tau_{cower} \\ 0, & \text{otherwise} \end{cases}\right) \tag{3.3}$$

Here, it is important to understand that i is the *index* of each intermediate solution. There may indeed be multiple intermediate solutions that are sufficiently probable (i.e., whose associated probability is above τ_{cower}). By choosing from these the intermediate solution with the highest index, we achieve the affect of choosing the intermediate solution for which the fewest variables have been removed from consideration, and thus, the intermediate solution for which the fewest new representations must be created.

I will now discuss how this model is algorithmically realized. The space of possible bindings Γ is quite large, namely $O(d^{|\Lambda^V|})$, where $d = \max_{v\in\Lambda^V}|M_{v[t]}|$, i.e., the size of the largest domain provided by a consultant associated with a variable in Λ^V. As it will be prohibitively expensive to examine all such bindings, I have developed the *DIST-POWER* algorithm to efficiently search this space, as discussed in this section.

I will first discuss the *Distributed, probabilistic, Closed World Entity Resolution* algorithm (DIST-CoWER,[12] Algorithm 4), which implements Equation 3.2, and I will then discuss our *Distributed, Probabilistic, Open World Entity Resolution* algorithm (*DIST-POWER*, Algorithm 5), which implements Equation 3.3, using DIST-CoWER as a subroutine.

[11] A note about τ: Several of the algorithms in this dissertation make use of thresholds denoted by *tau* – a symbol which does not appear in the definitions for those algorithm. This is because these thresholds are intended to be set by the developer at implementation time, to values chosen at the developer's discretion.

[12] I use "Entity" Resolution here as a way of emphasizing the domain-independent nature of our reference resolution algorithms; I use "Probabilistic" to emphasize that the algorithm incrementally computes the probabilities of binding hypotheses.

3.3.5 DIST-CoWER

The *DIST-CoWER* algorithm, generally stated, performs a best-first search through the space of possible assignments from variables to known entities, in which (incrementally constructed) partial assignments are eliminated when their (incrementally computed) joint probability falls below a given threshold; this probability function also serves as the priority function that dictates the order in which these partial assignments are considered.

To be specific, DIST-CoWER begins with a set of typed variables $V[t]$, a list of typed formulae $\Lambda[t]$, an initial set of candidate hypotheses (i.e., partial assignments, each of which is paired with a set of formulae left to examine, and a probability value that serves as its priority), and a set of consultants C. DIST-CoWER then does the following until some termination condition is met (Algorithm 4, Line 2). Here, different termination conditions may be preferred depending on context of use: a termination condition of $(|\tilde{\Gamma}| = 0)$ will find all sufficiently probable solutions; a termination condition of $(|\tilde{\Gamma}| = 0 \ or \ |\Gamma\star| = 1)$ will find the best sufficiently probable solution; a termination condition of $(|\tilde{\Gamma}| = 0 \ or \ |\Gamma\star| = n)$ will find the top n candidate solutions.

DIST-CoWER first pops the highest-priority (i.e., most probable) hypothesis Γ off of $\tilde{\Gamma}$ (Line 3). If this hypothesis Γ is a solution (i.e., a sufficiently probable hypothesis that accounts for all formulae in $\Lambda[t]$), it is added to the set of solutions (Lines 19-22). Otherwise, one of two actions is taken. If the first formula λ not yet accounted for in the hypothesis Γ contains a variable v that Γ does not include a binding for, a set of copies of Γ are put back into the queue, each of which binds v to a different candidate entity m provided by the consultant responsible for v (as determined by v's typing) (Lines 6- 10). If, on the other hand, the first formula λ not yet accounted for in hypothesis Γ does not include any new variables, DIST-CoWER finds a consultant c to "consult with" for λ, acquires from c the probability that λ is satisfied under hypothesis Γ (using the function *apply*, provided by each consultant in compliance with consultant capability 3, multiplies this with Γ's prior probability to produce a new posterior probability, and if this new probability is above threshold τ_{cower}, removes λ from Γ's set of unconsidered constraints, and adds Γ back onto the queue (Lines 11-18).

Thus, as previously stated, DIST-CoWER uses best-first search through an incrementally computed tree of partial solutions to find some subset (depending on chosen terminating condition) of sufficiently probable binding hypotheses.

I will now discuss how DIST-CoWER is used in the context of DIST-

Algorithm 4 DIST-CoWER($V[t], \Lambda[t], \tilde{\Gamma}, C$)

$\Gamma\star = \emptyset$
while (not done) **do**
  $\Gamma = pop(\tilde{\Gamma})$
  if $\Gamma^{\Lambda} \neq \emptyset$ **then**
    $\lambda = \Gamma^{\lambda_0}$
    if $(\exists v \in \lambda^{V} \mid v \notin \tilde{\Gamma}^{V})$ **then**
      for all $m \in M_{v[t]}$ **do**
        $\gamma = (v \rightarrow m)$
        $\tilde{\Gamma} = \tilde{\Gamma} \cup (\Gamma \cup \gamma, \Gamma^{\Lambda}, \Gamma^{P})$
      end for
    else
      $c = find\_consultant(C, \lambda)$
      $\Gamma^{P} = h^{P} \cdot apply(c, \lambda, \Gamma)$
      $\Gamma^{\Lambda} = \Gamma^{\Lambda} \setminus \lambda$
      if $(\Gamma^{P} > \tau_{cower})$ **then**
        $\tilde{\Gamma} = \tilde{\Gamma} \cup \Gamma$
      end if
    end if
  else
    $\Gamma\star = \Gamma \star \cup \Gamma$
  end if
end while
return $\Gamma\star$

POWER to implement equation 3.3 and effect *open-world* reference resolution.

3.3.6 DIST-POWER

While DIST-CoWER finds the optimal set of bindings from a set of variables to entities, DIST-POWER finds the longest suffix of a list of variables such that the set of bindings produced by DIST-CoWER to those variables is nonempty. DIST-POWER begins with the same input as DIST-CoWER: a set of typed variables $V[t]$, a list of typed formulae $\Lambda[t]$, an initial set of candidate hypotheses, and a set of consultants C. DIST-POWER's behavior can then be split into three stages: pre-processing (Algorithm 5 Lines 1- 9), resolution (Lines 10- 14), and hypothesization (Lines 15- 18).

During the pre-processing step, DIST-POWER first populates initial priority queue of hypotheses $\tilde{\Gamma}$ with the space of possible bindings to the first variable found in the first formula in $\Lambda[t]$, unless some partial set of bindings is set a priori. DIST-POWER then creates copies of $\Lambda[t]$ and $V[t]$, which may be pruned during iterations of the resolution loop.

During this resolution stage, DIST-POWER attempts to use DIST-CoWER to perform resolution; each time DIST-CoWER fails to produce results, DIST-POWER tries again with a smaller set of variables and formulae.

Finally, during the hypothesization step, DIST-POWER instructs its consultants to hypothesize new representations to associate with any variables that were "set aside" during the resolution stage (using the function *posit*, provided by each consultant in compliance with consultant capability 4. Because this may involve asserting that certain relations hold between known entities and these new entities, if hypothesization must be performed, all but the most probable hypothesis are discarded. If hypothesization is performed, DIST-POWER uses the newly created entities to produce a new hypothesis binding all variables in $V[t]$ to either previously known or newly hypothesized entities.

Algorithm 5 DIST-POWER$(V[t], \Lambda[t], \tilde{\Gamma}, C)$

if $\tilde{\Gamma} = \emptyset$ **then**
    $v = \lambda[t]_0^{v_0}$
    for all $m \in M_{v[t]}$ **do**
        $\gamma = (v \to m)$
        $\tilde{\Gamma} = \tilde{\Gamma} \cup \{\Gamma \cup \gamma, \Lambda[t], 1.0\}$
    end for
end if
$\Lambda' = \Lambda[t]$
$V' = V[t]$
repeat
    $\Gamma\star =$ DIST-CoWER$(V', \Lambda', \tilde{\Gamma}, C)$
    $\Lambda' = [\lambda \in \Lambda | head(V') \notin \lambda^V]$
    $V' = tail(V')$
until $\Gamma\star \neq \emptyset$ or $V' = \emptyset$
if $V' \neq V[t]$ **then**
    $\Gamma\star = posit(\Gamma\star_0, \Lambda[t], C)$
end if
return $\Gamma\star$

Currently, it is assumed that after a DIST-POWER query in which hypothesization is performed, a subsequent, identical DIST-POWER query should be able to identify that the newly hypothesized entities are the target referents. Because this requires asserting properties and relations involving those hypothesized entities, DIST-POWER may be unable to simultaneously identify ambiguities and perform hypothesization. For example, if one were to say "Could you go to the room at the end of the hallway" and the robot knows of two hallways, and does not know of a room at the end of either, it will choose the large-scale location more likely to be a hallway and hypothesize a room at its end. In the future, it may be more appropriate to move this hypothesization stage into a separate action such that resolution can be *interleaved* with, rather than purely precede, clarification request generation mechanisms. In addition, as described, DIST-POWER ends up performing some queries up to Λ^V times; this may be remedied (at the cost of additional memory usage) using traditional dynamic programming techniques.

3.4 Evaluation

In this section, I will present a multi-faceted evaluation of the previously presented algorithms.

1. First, I will provide a proof-of-concept demonstration of the algorithm's behavior.

2. Second, I will compare the performance of the proposed algorithm to that of *humans* (using human data collected through a crowdsourcing experiment), to demonstrate the extent to which it might serve as a model of human open-world reference resolution.

3. Third, I will discuss the *worst-case* time and space complexity of DIST-CoWER and DIST-POWER.

4. Fourth, I will compare the performance of the proposed algorithm *in practice* relative to a baseline (in which all relevant knowledge is contained within a single centralized knowledge base rather than distributed) in order to assess the efficiency gains effected by *DIST-POWER's* use of distribution and variable typing.

5. Fifth, I will examine how the performance of the proposed algorithm is *in practice* affected by variations in constraint ordering (e.g., whether it is more efficient to sort constraints based on arity, cost, or some other metric).

6. Finally, I will discuss the proposed algorithm's adherence to the theoretical commitments of *DIARC*.

3.4.1 Proof of Concept Demonstration

In this section we present a proof of concept demonstration of our proposed algorithm and framework. The purpose of this demonstration is two-fold: First, we will demonstrate that *DIST-POWER* behaves as intended, that is, that it allows resolution to be performed when the requisite information is distributed across various knowledge bases, and it allows resolution to be performed without knowledge (on the part of the algorithm itself) as to (1) the format of the knowledge stored in each knowledge base, and (2) the techniques necessary for extracting the relevant knowledge from each knowledge base. Second, we will demonstrate that *DIST-POWER* has been fully integrated into our cognitive robotic architecture (i.e, *DIARC*) in order to perform tasks natural to human-robot interaction scenarios.

As previously described, *DIARC* uses a distributed heterogeneous knowledge representation scheme: the architecture has *Belief*, *Goal*, and *Dialog* management components that track, as their names suggest, information about the beliefs of other agents, the robot's goals, and the human-robot dialogue state; but information about visual targets, for example, is localized in the *Vision* component, and information about spatial entities is localized in the *Spatial Expert* component. To implement the proposed framework, a set of "consultants" were implemented to interface with knowledge bases of known objects, locations, and people. Each consultant performed four functions:

1. Each consultant advertised the types of queries it handled by exposing a list of formulae such as $in(W - objects, Y - locations)$. This formula, for example, states that the consultant which advertises it is able to assess the degree to which some entity from the `objects` knowledge base is believed to be in an entity from the `locations` knowledge base.

2. Each consultant provided a method which returned a set of numeric identifiers of the atomic entities in its associated knowledge base.

3. Each consultant provided a method which, given formula p (e.g., $in(X - objects, Y - locations)$) and mapping m from variable names to numeric identifiers, (e.g., from X and Y to 22 and 25) would return the probability that relationship p held under the variable bindings specified in m. In this example, the appropriate consultant would return the degree to which it believed object 22 to be in location 25.

4. Each consultant provided a method which, given a set of formulae with some unbound variables, would posit new representations to associate

with those unbound variables, store the knowledge of their properties represented by those formulae, and return new variable bindings accounting for the newly posited entities.

In addition, a Referential Executive component provided a *DIST-POWER* method which, given a set of formulae Λ, calculated optimal mapping t and executed $DIST\text{-}POWER(V[t], \Lambda[t], \tilde{\Gamma}, C)$. As a proof of concept demonstration, we examined the behavior of a robot running this *DIARC* configuration, when given the utterance “Jim would like the ball that is in the room across from the kitchen” (assumed to be uttered by an agent named ‘Bob’). This utterance is represented as:

$$Stmt(bob, self, and(wouldlike(jim, X), ball(X), in(X, Y), room(Y), acrossfrom(Y, Z), kitchen(Z))).$$

This utterance is a statement from ‘Bob’ to the robot (i.e., ‘self’), where the head of the *and* list (i.e., $\{wouldlike(jim, X)\}$) represents the literal semantics of the sentence, and the tail of the *and* list represents the properties which must be passed to the Resolver for resolution.

We will now describe the behavior of the Referential Executive (REX) as it follows the *DIST-POWER* algorithm (with termination condition $|\tilde{\Gamma}| = 0$), detailing the state of REX’s hypothesis queue $\tilde{\Gamma}$ at several points throughout the trace of the algorithm. In order to provide an easily describable example, we limited the number of entities in the initial domain of each knowledge base to three or four entities. The robot’s knowledge base of locations contained a hallway and several rooms, including a kitchen, and a room across from it which only contained, to the robot’s knowledge, a table. The robot’s knowledge base of objects contained the table and several boxes and balls. We will use o as shorthand for `objects` and l as shorthand for `locations`.

REX first calculates optimal mapping t, and returns $\{X : o, Y : l, Z : l\}$. Next, REX determines that the first constraint to be examined will be $ball(X : o)$. REX thus instantiates its hypothesis queue by requesting a set of candidate entities for X from the consultant associated with knowledge base o, which produces candidate set $\{o_1, o_2, o_3, o_4\}$ [13]. REX then requests from o the probability of each of $\{ball(o_1), ball(o_2), ball(o_3), ball(o_4),\}$ being true, and receives back, respectively, $0.82, 0.92, 0.0, 0.0$. Since $0.0 < 0.1$ (the chosen value of τ_{cower}), the hypotheses with mappings $X : o_3$ and $X : o_4$ are thrown out, and the other two hypotheses are returned to H, resulting in hypothesis queue:

Γ	Γ^Λ	Γ^P
$\{X : o_2\}$	$\{room(Y : l), kitchen(Z : l), in(X : o, Y : l),$ $acrossfrom(Y : l, Z : l)\}$	0.92
$\{X : o_1\}$	$\{room(Y), kitchen(Z : l), in(X : o, Y : l),$ $acrossfrom(Y : l, Z : l)\}$	0.82

The next constraint to be considered is $room(Y : l)$. Since $\{X : o_2\}$ does not contain a candidate identifier for Y, REX requests the initial domain of Y from l, receives the set of candidates $\{l_1, l_2, l_5, l_6\}$, and replaces the first hypothesis with a set of four hypotheses which each have a different binding for Y but share the original Γ^P priority and set of unconsidered constraints Γ^Λ. The probability resulting from $apply(l, in(X : o, Y : l), \{X : o_2, Y : l_i \dots\})$ is then computed for each of these four hypotheses, resulting in, respectively, $0.82, 0.92, 0.0, 0.6$. The third hypothesis is thrown out and the others are returned to Γ with updated probabilities, resulting in hypothesis queue:

Γ	Γ^Λ	Γ^P
$\{X : o_2, Y : l_2\}$	$\{kitchen(Z : l), in(X : o, Y : l),$ $acrossfrom(Y : l, Z : l)$	0.846
$\{X : o_1\}$	$\{room(Y : l), kitchen(Z : l), in(X : o, Y : l),$ $acrossfrom(Y : l, Z : l)\}$	0.820
$\{X : o_2, Y : l_1\}$	$\{kitchen(Z : l), in(X : o, Y : l),$ $acrossfrom(Y : l, Z : l)\}$	0.754
$\{X : o_2, Y : l_6\}$	$\{kitchen(Z : l), in(X : o, Y : l),$ $acrossfrom(Y : l, Z : l)\}$	0.736

[13]Note that this is the set of *all entities* known of by consultant o, and *not* the set of entities that o believes could plausibly be a ball. A simple extension of this process that would likely effect significant speedups would be to provide the query or some subset of the query to the consultant, and ask for the set of *locally sufficiently likely* candidates rather than the set of *all* candidates. This would, however, result in the bulk of the work involved in reference resolution being pushed onto the individual consultants, and thus, for the sake of generality, I do not consider such an extension in this dissertation, but instead leave it aside, to be considered in future work.

As the hypothesis with binding $\{X : o_2, Y : l_2\}$ is then the most likely hypothesis and the next constraint to consider is $kitchen(Z : l)$, Z is expanded with candidate locations, each checked for the $kitchen(Z : l)$ property. As only location 2 is known to be a kitchen, the first hypothesis is replaced with a single new hypothesis, with probability 0.762. This causes the hypothesis with binding $\{X : o_1\}$ to become the most probable hypothesis, resulting in the above process being repeated for that hypothesis, resulting in hypothesis queue:

Γ	Γ^{Λ}	Γ^{P}
$\{X : o_2, Y : l_2, Z : l_2\}$	$\{in(X : o, Y : l),$ $acrossfrom(Y : l, Z : l)\}$	0.762
$\{X : o_2, Y : l_1\}$	$\{kitchen(Z : l), in(X : o, Y : l),$ $acrossfrom(Y : l, Z : l)\}$	0.754
$\{X : o_1, Y : l_2\}$	$\{kitchen(Z : l), in(X : o, Y : l),$ $acrossfrom(Y : l, Z : l)\}$	0.754
$\{X : o_2, Y : l_6\}$	$\{kitchen(Z : l), in(X : o, Y : l),$ $acrossfrom(Y : l, Z : l)\}$	0.736
$\{X : o_1, Y : l_1\}$	$\{kitchen(Z : l), in(X : o, Y : l),$ $acrossfrom(Y : l, Z : l)\}$	0.672
$\{X : o_1, Y : l_6\}$	$\{kitchen(Z : l), in(X : o, Y : l),$ $acrossfrom(Y : l, Z : l)\}$	0.656

When the next best hypothesis is examined, it will be eliminated, as o_2 is not known to be located in l_2. Indeed, as no ball is known to exist in a room across from a kitchen, all hypotheses are systematically eliminated. Once this has finished, *DIST-POWER* removes the head of its variable list $V\prime$ and tries the entire above process again, with $V\prime = \{Y : l, Z : l\}$ and $\Lambda\prime = \{room(Y : l), acrossfrom(Y : l, Z : l), kitchen(Z : l))\}$. The elimination of $X : o$ from these sets suggests that X refers to an object that is not yet known to the robot. This time, after considering the first formula in $\Lambda[t]$ (i.e., $room(Y : l)$), the initial hypothesis queue is:

Γ	Γ^{Λ}	Γ^{P}
$\{Y : l_2\}$	$\{kitchen(Z : l), acrossfrom(Y : l, Z : l)\}$	0.92
$\{Y : l_1\}$	$\{kitchen(Z : l), acrossfrom(Y : l, Z : l)\}$	0.82
$\{Y : l_6\}$	$\{kitchen(Z : l), acrossfrom(Y : l, Z : l)\}$	0.8

After going through the same resolution process, the final hypothesis queue will be:

Γ	Γ^{Λ}	Γ^{P}
$\{Y : l_1, Z : l_2\}$	$\{\}$	0.702

DIST-POWER then instructs the `objects` consultant to create a new representation for X. It does so and returns a new identifier for it, which is used to update the hypothesis queue:

Γ	Γ^{Λ}	Γ^{P}
$\{X : o_5, Y : l_1, Z : l_2\}$	$\{\}$	0.702

DIST-POWER then instructs both the `objects` and `locations` consultants to maintain consistency with $\Lambda[t]$ under the bindings Γ of the one remaining hypothesis, h. This results in the `objects` consultant asserting into its KB that o_5 is a ball, and the `locations` consultant asserting into its KB that l_1 contains o_5.

REX then uses $\Gamma\star$ to convert $\mathit{wouldlike}(\mathit{jim}, X)$ into $\mathit{wouldlike}(\mathit{jim}, o_5)$. The utterance $\mathit{Stmt}(\mathit{bob}, \mathit{self}, \mathit{wouldlike}(\mathit{jim}, o_5))$ is then returned to the process that requested resolution.

In Section 6.4, we will describe how the *intentions* behind this type of utterance could be inferred. In our proof-of-concept demonstration, this process (which we will leave until that section to explain) results in the following goal being inferred: $\mathit{goal}(\mathit{self}, \mathit{bring}(\mathit{self}, o_5, \mathit{jim}))$, The robot thus determines that bob wants it to bring object o_5 (which is in room l_1) to jim. The robot responds "Okay" and drives to l_1 to retrieve object o_5.

Before moving on, we would like to emphasize that this proof-of-concept demonstration is of course not a quantitative evaluation, nor is it intended to be. In the next section, we present the first of three quantitative evaluations which fill just this purpose.

3.4.2 Evaluation and Cognitive Model

In this section, we present a human-subject experiment designed to achieve two purposes: (1) to serve as an evaluation as to how successful *DIST-POWER* would be in resolving references relative to humans (as in (Williams & Scheutz, 2015b)), and (2) to investigate how well *DIST-POWER* models the products of human reference resolution in uncertain and open worlds, at the *computational* level of analysis (Marr, 1982) (as in (Williams & Scheutz, 2015a)). This experiment does not exploit *DIST-POWER*'s distributed reasoning capabilities, evaluation of which is examined in subsequent sections.

Experimental Design

In this experiment, participants were asked to consider three sets of referential statements. For each of the three sets of statements, they were provided with the corresponding third of the following knowledge base shown in Table 3.1.

Table 3.1: Knowledge Base Provided to Participants

ID	Name	Description
1	Jim Nelson	Doctor (**pretty sure**). Friends with Sam Greene.
2	Sam Greene	friends with Jim Nelson. **Probably** male.
3	Jim Cruz	?
4	Mary Greene	Sister of Sam Greene.
5	Frank Roberts	Jon says he's a painter, **but** Craig says he's an author **. . . ?** Lives next door to Nicolas.
6	Martin Francis	Painter, lives next door to Heidi.
7	Kristy Roberts	**Might** be the daughter of Frank Roberts. **Unsure**.
8	Heidi Wilkerson	Chemist, lives next door to Martin.
9	Nicolas Morris	Chemist, lives next door to Frank.
10	Craig Horton	Chemist, **might** work with Heidi? **Probably** doesn't work with Nicolas, **but who knows**.
11	Ted Wells	Baker. **Possibly** brothers with Phillip and/or Troy.
12	Phillip Wells	Brewer. **Possibly** brothers with Ted and/or Troy.
13	Troy Wells	Byron's friend. **Possibly** brothers with Phillip and/or Ted.
14	Laurie Rodgers	Byron's friend. Girlfriend of **one of** the Wells brothers.
15	Sally Owens	Teacher. Sibling of Willie Owens. Laurie's neighbor.
16	Willie Owens	Customs officer. **Possibly** female. Sibling of Sally Owens.
17	Byron Todd	**Could be** a podiatrist **. . . or maybe** a pediatrician.

In bold are words indicating uncertain information.

Participants were told that their siblings were planning a party, and that the aforementioned list was a list of people their sister had invited. Each participant was then given a second list corresponding to each third of the second column of Table 3.2, and were told that each description in this list represented a description given by their brother of someone *he* wanted invited to the party, that anyone mentioned in a description needed to be invited as well, and that it was their job to determine, for each person mentioned in one of their brother's descriptions, whether or not that person already appeared on their sister's list and if so who that person was. The sixteen referring expressions used in this evaluation specifically probed 16 conditions we will now describe.

We identify four types of uncertainty which may arise when resolving referring expressions. (1) In cases of *incomplete knowledge* (IK), an utterance might seem to refer to an entity not yet known to the robot. (2) In cases of *uncertain knowledge* (UK) an utterance might use properties to describe an entity which a robot is not *sure* actually has those properties. (3) In cases of *ambiguous knowledge* (AK) an utterance may seem to be equally likely to refer to multiple known entities. (4) And of course, an utterance may seem to uniquely identify an entity (a case we refer to as *certain knowledge* (CK)).

In the resolution of a referring expression, these categories can apply either to the *target* (i.e., the intended referent) of a referring expression or to one or more of its *anchors*. For example, in the referring expression "The uncle of the doctor's brother", *the uncle* is the target, and *the doctor* and *the doctor's brother* are the anchors. Similarly, when considering the subclause *the doctor's brother*, *the brother* is the target, and *the doctor* is the anchor.

Sixteen classes of uncertainty can thus be created by classifying referring expressions into four classes (i.e., IT, UT, AT, CT) based on the uncertainty status of the referring expression's *target*, crossed by four classes (i.e., IA, UA, AA, CA) based on the uncertainty status of the referring expression's *anchors*[14].

The sixteen referring expressions we used to probe these sixteen classes of uncertainty are listed, along with their uncertainty class, in Table 3.2.

For each expression, *DIST-POWER*'s consultant was provided the same knowledge encoded in logical form, with confidences attached to each statement indicative of any uncertainty associated with that statement. For example, the consultant was informed that Kristy was the daughter of Frank with probability 0.5. All terms used to choose these probability values are highlighted in Table 3.1. *DIST-POWER* was then provided with the same referring expressions as were given to participants, encoded into logical form, with hand-annotated variable orderings.

[14]Note, however, that for the referring expression associated with the combination of a certain target and a certain anchor, the qualifier 'Pretty Sure' was used. This was originally intended to indicate a *low degree* of uncertainty, but obviously could be construed as suggesting uncertainty. If this paradigm is used in the future, we would suggest omitting this qualifier, which in retrospect was an experimental oversight.

Table 3.2: Evaluation Cases

Condition	Description given to participant
CA:CT	The doctor's friend's sister
AA:CT	Jim's friend
AA:IT	Jim's daughter
IA:IT	Tabitha's mother
AA:AT	The chemist's neighbor
UA:IT	Craig's coworker's neighbor's son
IA:CT	Marion's daughter Kristy
UA:UT	Craig's coworker's neighbor's daughter
CA:UT	Troy's girlfriend
CA:AT	The baker's brother
IA:AT	The chemist, Billie's father
IA:UT	Michelle's daughter
CA:IT	Sally's wife
AA:UT	The Wells boy's girlfriend
UA:CT	Troy Wells, the podiatrist's friend
UA:AT	The podiatrist's friend

(1) Each condition, (2) the referring expression used to probe that condition.

Participation

For this experiment, participants were recruited using Amazon Mechanical Turk. The pool of subjects who finished the task consisted of 40 participants (18 Male, 22 Female) with mean age 34.75. Participants were paid $2.00 to perform the task.

Results

The results of this experiment are summarized in Columns 3-5 of Table 3.3. Here, Column 3 shows the most frequent human response given for each referring expression, and the result or set of equally-likely results returned by *DIST-POWER* are shown in Column 4. In both cases, referents deemed not already on the guest-list are denoted '?'. For those referents, the model added new entries to the knowledge base and updated existing entries appropriately.

Column 5 of Table 3.3 shows the percentage of participants whose response aligned with each of *DIST-POWER* responses, with conditions in which the most frequent human response matched one of *DIST-POWER*'s

Table 3.3: Evaluation Cases and Results

Condition	Description given to participant	Most Frequent Human Response	Model Responses	%
CA:CT	The doctor's friend's sister	(Sister:4, Friend:2, Doctor:1)	(Sister:4, Friend:2, Doctor:1)	80.0
AA:CT	Jim's friend	(Friend:2, Jim:1)	(Friend:2, Jim:1)	60.0
AA:IT	Jim's daughter	(Daughter:?, Jim:1)	(Daughter:?, Jim:1)	47.5
			(Daughter:?, Jim:3)	37.5
IA:IT	Tabitha's mother	(Mother:?, Tabitha:?)	(Mother:?, Tabitha:?)	90.0
AA:AT	The chemist's neighbor	(Neighbor:6, Chemist:8)	(Neighbor:6, Chemist:8)	22.5
			(Neighbor:5, Chemist:9)	15.0
UA:IT	Craig's coworker's neighbor's son	(Son:?,Nei.:6, Co.:8,Craig:10)	(Son:?,Nei.:6,Co.:8, Craig:10)	65.0
IA:CT	Marion's daughter Kristy	(Kristy:7,Marion:?)	(Kristy:?,Marion:?)	18.5
UA:UT	Craig's coworker's neighbor's daughter	(Daug.:?,Nei.:6, Co.:8,Craig:10)	(Daug.:?,Nei.:6, Co.:8,Craig:10)	50.0
CA:UT	Troy's girlfriend	(Girlfriend:14, Troy:13)	(Girlfriend:14,Troy:13)	55.0
CA:AT	The baker's brother	(Brother:12,Baker:11)	(Brother:12,Baker:11)	70.0
			(Brother:13,Baker:11)	5.0
IA:AT	The chemist, Billie's father	(Father:?,Billie:?)	(Father:?,Billie:?)	97.5
IA:UT	Michelle's daughter, Willie	(Willie:16,Michelle:?)	(Willie:?,Michelle:?)	5.0
CA:IT	Sally's wife	(Wife:?,Sally:15)	(Wife:?,Sally:15)	95.0
AA:UT	The Wells boy's girlfriend	(Girlfriend:14,Wells boy:13)	(Girlfriend:14,Wells boy:13)	5.0
			(Girlfriend:14,Wells boy:12)	2.5
			(Girlfriend:14,Wells boy:11)	2.5
UA:CT	Troy Wells, the podiatrist's friend	(Troy Wells:13, Podiatrist:17)	(Troy Wells:13, Podiatrist:17)	85.0
UA:AT	The podiatrist's friend	(Friend:13, Podiatrist:17)	(Friend:13, Podiatrist:17)	27.5
			(Friend:14, Podiatrist:17)	20.0

(1) Each condition, (2) the referring expression used to probe that condition, (3) the most frequent human response for that referring expression, (4) the responses provided by *DIST-POWER* for that referring expression (with multiple rows used when multiple responses were returned), and (5) the percentage of human participants who provided the same answer as the model for each model response. Cases in which the most frequent human response matched a response provided by *DIST-POWER* are bolded in Column 1.

responses displayed in bold in Column 1.

The results show that in 14 of the 16 conditions (87.5%), *DIST-POWER* gave a response that was most frequent among human participants.

Overall, these results suggest that *DIST-POWER* was successful at producing results that align well with human performance. We take this as evidence that *DIST-POWER* serves as a successful model of human reference resolution in uncertain and open worlds, at the computational level of analysis. We will now turn our attention towards those few cases where human and algorithmic responses did not align: IA:CT and IA:UT.

Both are examples of *false negatives*, in which *DIST-POWER* failed to find a match it thought sufficiently probable. These are strictly better than the *false positives* which would have been unavoidable had the algorithm not accounted for open-world operation. False negatives are strictly better in part because they can be more easily recovered from: if it is later established that a posited hypothetical entity is in fact the same as some known, grounded entity, those two representations may be consolidated. Recovering from the discovery of an error of mistaken identity is much harder, as it would require source tracking whenever information is added to a knowledge base.

We also compare world model modifications suggested by participants with those made by *DIST-POWER*. Modifications made by *DIST-POWER* were straightforward: if *DIST-POWER* believed a referenced person did not yet exist in the knowledge base, it added a new representation for that person. For example, for 'Jim's friend', *DIST-POWER* created a new representation and gave it a property indicating it was friends with Jim. In all but one condition, the most common human suggestions for world-model modification followed this pattern, and thus the most frequent human response for world model modification matched that of *DIST-POWER* in 13 of the 16 conditions (81.25%).

We will now examine the conditions in which *DIST-POWER* produced incorrect results. In condition **IA:CT**, *DIST-POWER* produces an incorrect response due to, we believe, a violation of its assumption that unknown entities are always referenced with respect to known entities. This type of violation occurs when a speaker makes incorrect assumptions about their addressee's beliefs.

We believe that *DIST-POWER* would be able to handle this condition if it was extended to (1) consider whether newly posited anchors were highly probable matches to other known entities, (2) generate a clarification request as to whether those matches were valid, and (3) consolidate the relevant representations if an affirmative response is returned.

In condition **IA:UT**, participants seem to have assumed, as in condition **IA:CT**, some failure in belief modeling on the part of the speaker. In this condition, however, this assumption was made despite high uncertainty as to whether the *known* Willie was even of the same gender as the *described* Willie, perhaps due to the relative uniqueness of the name.

In order for *DIST-POWER* to successfully handle this condition it would need to acknowledge that certain properties, such as being named 'Willie', are relatively unique, perhaps by modeling properties' prior probabilities.

DIST-POWER's world model modifications differed from those suggested by human participants in both **IA:UT** and **IA:CT**, as would be expected. However, *DIST-POWER*'s modifications also differed from humans in condition **UA:UT**, probed by the utterance 'Craig's coworker's neighbor's daughter'. In this condition, the response that no modification of the list was needed was more popular (by a single participant) than the response which aligned with that given by *DIST-POWER* (i.e., that 'Craig's coworker's neighbor's daughter' or 'Martin's daughter' should be added to the list). One may wonder why, for this question, the most popular human response for world model modification did not align with the most popular human resolution response. Curiously, several participants reported that 'the daughter' did not already appear in the list, yet responded that no modification of the list was necessary. If these inconsistent responses are ignored, than the most popular human response aligns with the response provided by *DIST-POWER*. We would thus argue that for this condition, *DIST-POWER* provided a more appropriate response than that provided by human participants.

Before moving on, it is important to note that technically in this experiment, the "knowledge base" provided to participants reflected the knowledge and uncertainties of *the participant's (fictional) brother*, and not, strictly, of the participant themselves. However, as the information in this knowledge base was the only information participants were (at least initially) given about people assumed to exist within the party domain, we see no reason why they should not have adopted this knowledge, and its uncertainties, as their own. It would of course be an interesting direction for future work to examine both the extent to which this uptake actually occurs, as well as the extent to which the existence or non-existence of this uptake actually matters; it is possible that there is no difference in assessing ones' own knowledge and uncertainty and assessing the knowledge and uncertainty of another, when this is the only available information. For the time being, the fact that the entirety of both the brothers' knowledge and the participants' knowledge are taken as indistinguishable and held in common prevents the need for any Theory-of-Mind belief modeling, which has been shown by H. H. Clark & Marshall (2002) to need to infinitely recurse in order to achieve exact inference (see also Van Deemter, 2016).

3.4.3 Complexity Analysis

In this section I will discuss the time and space complexity of DIST-CoWER and DIST-POWER. This is meant to paint a general picture of the complexities of these algorithms, and is not intended as a formal theoretical analysis, which would be beyond the scope of this dissertation.

The worst case for DIST-CoWER will occur when the termination condition is finding all candidate bindings, and in which DIST-CoWER is given a set of j *uninformative* \$k\$-arity single-domain predicates, i.e., predicates that do not actually rule out any candidates and require all hypotheses to contain candidate bindings for the maximum of k variables all from a single domain. When this is the case, DIST-CoWER will need to build a hypothesis queue with m^k entries (each comprised of k variable bindings), where m is the number of candidates in the singly handled domain. Because in this scenario all predicates involve all k variables, they will not be evaluable until after the hypothesis queue is completed, rendering a time complexity of $O(j \cdot m^k)$ and space complexity of $O(k \cdot m^k)$.

DIST-POWER has the same time and space complexity as DIST-CoWER, even though DIST-POWER runs DIST-COWER up to k times. The time complexity of DIST-POWER is identical to that of DIST-CoWER because each additional use of DIST-COWER is executed with a smaller number of variables, and thus with a smaller value of k. Since k dominates each individual run, and the runs are combined additively, the $O(k \cdot m^k)$ of the initial run is guaranteed to dominate the complexity of the subsequent additive runs. The space complexity of DIST-POWER is identical to that of DIST-CoWER because, for the same reasons, the most space will be used in the first run of DIST-POWER. And moreover, because information is not retained between runs of DIST-CoWER, *no* additional space will be used after the first run.

But of course, these analyses do not reflect the *expected* complexity of DIST-CoWER and DIST-POWER in practice. First, most predicates used will likely be of low arity (i.e., unary or binary). But second, and more importantly, the use of uninformative predicates will be uncommon, as this would be a violation of Grice's Maxims. We would not say, as a Rational Speech Act Theorist might, that predicates will be designed to be *optimally* informative, as speakers may indeed use phrases like 'the thing' which are near-completely uninformative, or properties like color for reasons of preference rather than informativity (a topic we will come back to in Chapter 5). But the majority of properties and relations used *are* expected to be informative, which will have the effect of drastically reducing the number of hypotheses in the hypothesis queue between DIST-CoWER's rounds of hypothesis queue expansion. In the next two sections we thus investigate complexity concerns from a more realistic perspective, using empirical simulations rather than worst-case analysis.

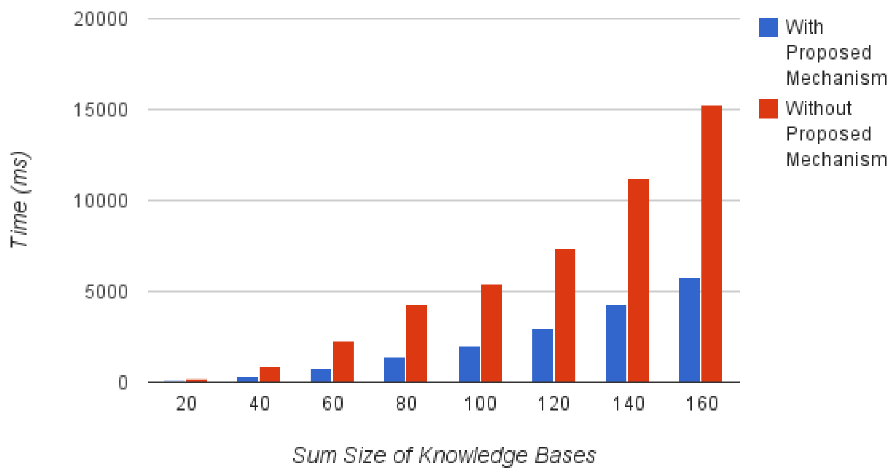
3.4.4 Performance Differences by use of Typing

In this section we evaluate the performance of *DIST-POWER* (see also (Williams & Scheutz, 2016a)) with that of our previous *non-distributed* algorithm (*POWER*, see also (Williams & Scheutz, 2015b)). This primarily serves to evaluate the efficiency gains made by using typed variables. To facilitate this evaluation we generated forty knowledge bases: five each of sizes $n = 20, 40, 60, \ldots, 160$ where n indicates the total number of entities stored in each knowledge base. In each knowledge base, half of the entries were locations in a randomly generated floor plan (i.e., rooms, halls, intersections and floors) with various properties with randomly assigned likelihoods, and the other half were objects (i.e., balls, boxes and desks), each of which was randomly assigned various properties and assigned to a randomly chosen room. Baseline performance was then evaluated by measuring the time taken by *DIST-POWER* to evaluate the query associated with the referring expression 'the box in the room' for each knowledge base. The times for each set of five knowledge bases were then averaged.

We then generated forty additional pairs of knowledge bases: five pairs each of sizes $(n_1, n_2) = (10,10), (20,20), (30,30), \ldots, (80,80)$ such that the first knowledge base dealt with all information pertaining to locations and the second knowledge base dealt with all other (i.e., object-related) knowledge. Performance of *DIST-POWER* was then established by measuring the time taken to evaluate the query associated with the referring expression 'the box in the room' for each *pair* of knowledge bases. The times for each set of five pairs of knowledge bases were then averaged.

Figure 3.4 shows the results of this experiment: along the horizontal axis are the sum sizes of knowledge bases used in each test case (e.g., '40' refers to the knowledge base containing 40 entities used when evaluating performance *without* the proposed mechanism, and the two knowledge bases containing 20 entities each used when evaluating performance *with* the proposed mechanism.) Along the vertical axis is the average time taken, for each set of

Figure 3.4: Performance Differences



knowledge bases of each size, to perform the simple query described. From these results one may observe the performance improvement effected through use of the proposed algorithm: up to 3x speedup among the examined cases.

One will notice that both algorithms show performance exponential in the number of stored entities, due to the use of best-first search over, e.g., beam search. However, the complexity of both algorithms when used in the real world would likely be substantially reduced, for several reasons. First, the consultants used by *DIST-POWER* did not use any heuristics when returning the set of initial candidates to consider. While these would certainly be used in practice, but using them would have conflated the performance of the algorithm with the performance of those heuristics, which is beyond the scope of this dissertation.

Second, complexity would be significantly reduced by tracking the entities in, e.g., the robot's short term memory, and checking against those entities before querying the robot's knowledge bases. In fact, as we will describe in Chapter 4, we have integrated *DIST-POWER* into a larger resolution framework inspired by J. K. Gundel, Hedberg, & Zacharski (1993)'s *Givenness Hierarchy* , which substantially reduces complexity and allows a robot to resolve references occurring in a wider variety of linguistic forms.

We also note that in order to have a consistent evaluation, the *POWER* and *DIST-POWER* algorithms were provided with information represented in the same way. However, one of the primary advantages of the *DIST-POWER* algorithm is that information need not be represented in a single format; the information stored in the `locations` knowledge base could just as easily have been represented in a topological map rather than as a database of formulae. In fact, this was the case for our proof-of-concept demonstration (Section 3.4.1)

3.4.5 Performance Differences by Constraint Ordering

As previously described, the core of the *DIST-POWER* algorithm is a best-first search over a set of relational assignments: given a set of **constraints** (e.g., $box(X)$, $on(X,Y)$), *DIST-POWER* uses each constraint to search over the set of possible bindings of known entities to the variables found in those constraints (e.g., X, Y). In the presentation of *DIST-POWER*, however, the order in which these constraints are examined is left as an open question; it is possible that the order in which constraints are examined could greatly affect the performance of the algorithm, by lowering branching factor or "failing fast", similar to the variable-ordering heuristics used by constraint-satisfaction problem solvers(Kumar, 1992; Minton, Johnston, Philips, & Laird, 1992; B. M. Smith & Grant, 1997). Furthermore, it's possible that the effectiveness of a particular constraint-ordering heuristic may depend on the properties of the constraints being ordered, or upon the knowledge base in which those constraints will be assessed. For example, perhaps it is better to consider more expensive constraints first, while there are few resolution hypotheses. On the other hand, maybe this depends on exactly how costly those constraints are to evaluate, or the degree to which the costliness of evaluation differs from constraint to constraint. In this section, we investigate the effect of knowledge base characteristics and constraint-ordering heuristics on the *DIST-POWER* algorithm through a set of controlled experiments.

Procedure

In order to evaluate the effects of knowledge base characteristics, we first devised a system in which random knowledge bases could be generated, populated in accordance with a knowledge base "template". Each template contains information on ten different constraints. For each constraint, the template specifies: (1) the name of the constraint, (2) its arity, (3) how long it takes to evaluate, (4) what proportion of knowledge-base items it should be applied to, and (5) the range in probability which should be returned for a "successful" query. Each template thus varied with respect to the distribution of arity of constraints, the distribution over constraint costs, the distribution over coverage frequencies, and the distribution over constraint certainties.

We first defined a default knowledge base template. This knowledge base handled ten different relations: five unary relations (u1-u5), three binary relations (b1-b3), and two ternary relations (t1-t2). Each relation was stated to occur with 30% probability (if a binary or ternary constraint is applied, other entities in the knowledge base are chosen at random to be the other members of that constraint). Evaluating whether a relation held for an entity or set of entities was stated to take one millisecond. If the relation was judged to hold, the probability of it holding was stated to always be 1.0.

We then defined knowledge base templates which varied according to each of the four dimensions noted above:

Arity: We defined three additional knowledge base templates which differed from the default template with regards to the number of unary, binary, and ternary relations (seven, two and one, respectively; three, six and one, respectively; or three, two and five respectively).

We were interested in investigating varying arity, because, in practice, evaluating polyadic constraints forces expansion of the search space to account for new variables if any of a constraint's members have not yet been considered, and can thus be costly. It is for this reason why in previous work arity-ordering has been the default constraint-ordering strategy.

Coverage: We defined four additional knowledge base templates which differed from the default template with regards to how common each relation was (each relation having 10% probability of being applied; each relation having 50% probability of being applied; relations having 10,20,30,40 or 50% probability of being applied, with probability monotonically increasing with respect to arity; or relations having 2,8,16,32,or 64% probability of being applied, with probability monotonically increasing with respect to arity).

We were interested in varying coverage because it determines the branching factor of resolution. It may thus be more prudent to evaluate rarer constraints (e.g., President of the United States) before evaluating constraints that will preserve high branching factor (e.g., whether two people are friends).

Cost: We defined four additional knowledge base templates which differed from the default template with regards to how long each relation took to apply (1,2,3,4 or 5 ms, with time either monotonically increasing or monotonically decreasing with respect to arity; or 0.2, 0.4, 0.8, 1.6 or 3.2 ms, with time either monotonically increasing or monotonically decreasing with respect to arity).

We were interested in varying cost because, in practice, evaluating certain constraints may take a significant amount of time, which could be a waste if other constraints would have led to hypothesis pruning if they had been evaluated first. For example, consider the referring expression "the room across from the kitchen on the second floor". Suppose that a robot has a fast-accessible list of known locations, their floors, and their types, and a metric map specifying their exact positions, which is more costly to access. If the robot can quickly rule out that it doesn't know of a kitchen on the second floor (or that it knows of exactly one) before it begins the costly process of assessing the across-from-ness of all pairs of known rooms, it may save itself a large amount of time.

Uncertainty: We defined four additional knowledge base templates which differed from the default template with regards to how certain the knowledge base would be of applied properties (30 $\pm$ 5% probable; 70 $\pm$ 5% probable; 60 to 100% probable with probability monotonically increasing with respect to arity; or 20 to 100% probable with probability monotonically increasing to arity).

We were interested in varying uncertainty because, in practice, increased uncertainty may lead to faster search-space pruning, and some constraints (e.g., whether a given room could be considered to be an office) may naturally have more uncertainty associated with them than others (e.g., whether a given person is male).

Provided with these knowledge base templates, we now had the ability to systematically generate knowledge bases that differed with respect to key characteristics. Since we were also interested in evaluating the effect of constraint-ordering heuristics, we next define a set of seven simple heuristics.

Random: Using this heuristic, constraints are considered in a random order. This provides a baseline performance level.

Arity: Using this heuristic, constraints are considered in *increasing* order of arity. This is the heuristic typically used in the *DIST-POWER* algorithm. This is expected to perform well as it allows candidate bindings for one variable to be pruned out before expanding hypotheses to account for bindings to other variables.

Reverse Arity: Using this heuristic, constraints are considered in *decreasing* order of arity. This is expected to perform poorly because it will result in the hypothesis list rapidly expanding before it can be effectively pruned.

Coverage: Using this heuristic, constraints are considered in increasing order of frequency of coverage. This is expected to perform well as it allows rare constraints to rapidly prune the search space before more common constraints are considered.

Reverse Coverage: Using this heuristic, constraints are considered in decreasing order of frequency of coverage. This is expected to perform poorly as those constraints most likely to rule out candidates will be considered last.

Cost: Using this heuristic, Constraints are considered in increasing order of cost. This is expected to perform well as expensive constraints are put off until last, at which point they may no longer need to be evaluated.

Reverse Cost: Using this heuristic, constraints are considered in decreasing order of cost. This is expected to perform poorly as expensive constraints are considered first, which may not turn out to be necessary if the hypotheses on which they are evaluated would have otherwise been pruned away.

Finally, we defined a set of evaluation queries with differing levels of complexity: ({u1(X)}; {u1(X), u2(Y), b1(X,Y)}; {u1(X), u2(Y), u3(Z), t1(X,Y,Z)}; {u1(X), u2(Y), u3(Z), b1(X,Y), b2(Y,Z), t1(X,Y,Z)}).

Altogether, the above considerations result in 448 meaningful combinations of query, knowledge base template and heuristic (Q, T, H). For each such combination, we generated five random knowledge bases according to template T and then recorded the mean time taken to evaluate query Q using heuristic H.

Results

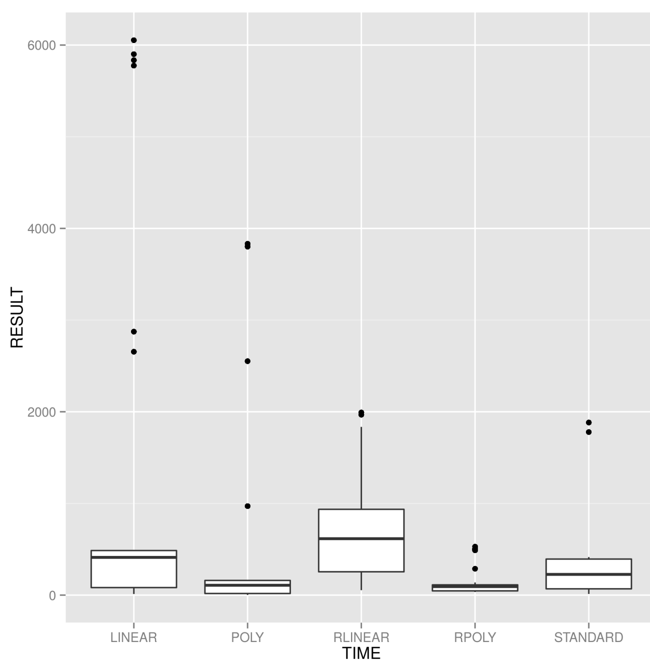
In this section, we will describe the results of our experiment. First we will discuss the results of varying knowledge base characteristics (i.e., arity, cost, coverage and uncertainty distributions). We will then discuss the results of varying query complexity. Next, we will discuss the results of varying the choice of constraint ordering heuristic. Finally we will discuss the interaction effects found between knowledge base characteristics and the choice of constraint ordering heuristic.

1. Knowledge Base Characteristics

    (a) **Arity:** Upon examining the degree to which the distribution of unary, binary, and ternary relations would affect resolution performance, no significant differences were observed.

    (b) **Cost:** As seen in Figure 3.5, *DIST-POWER* performed better in knowledge bases where some relations (especially polyadic relations) are cheap to evaluate, even when other relations are slightly more costly to evaluate.

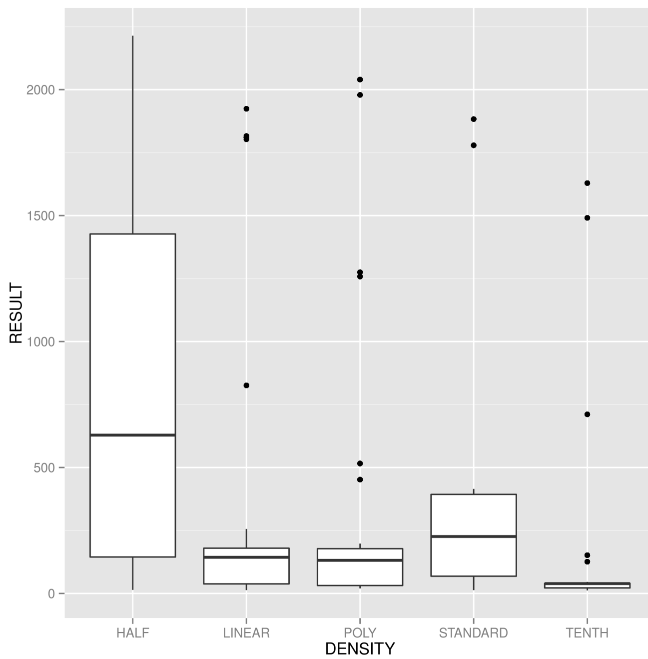
Figure 3.5: Effect of Cost



Policy labels indicate relation between constraint application cost and arity: linearly increasing, polynomially increasing, linearly decreasing, polynomially decreasing, and constant.

(c) **Coverage:** As seen in Figure 3.6, *DIST-POWER* performed better in knowledge bases where relations tend to apply to a smaller percentage of entities.

Figure 3.6: Effect of Coverage



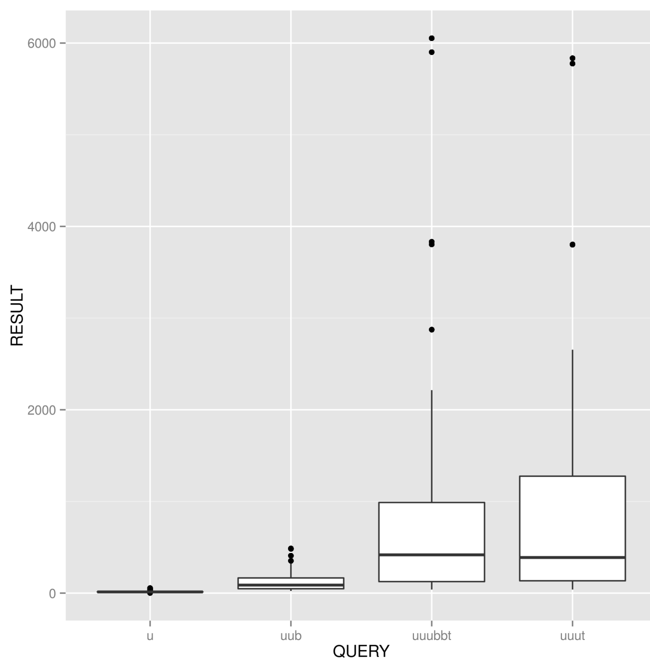
Policy labels indicate probabilities for whether predicates hold for particular entities: 50% probability, probability linearly increasing with arity, probability polynomially increasing with arity, the standard 30% probability, and 10% probability.

(d) **Uncertainty:** Upon examining the degree to which relation uncertainty would affect resolution performance, no significant differences were found.

2. Query Complexity

   As seen in Figure 3.7, *DIST-POWER* performed better with less complex queries.

Figure 3.7: Effect of Query Complexity

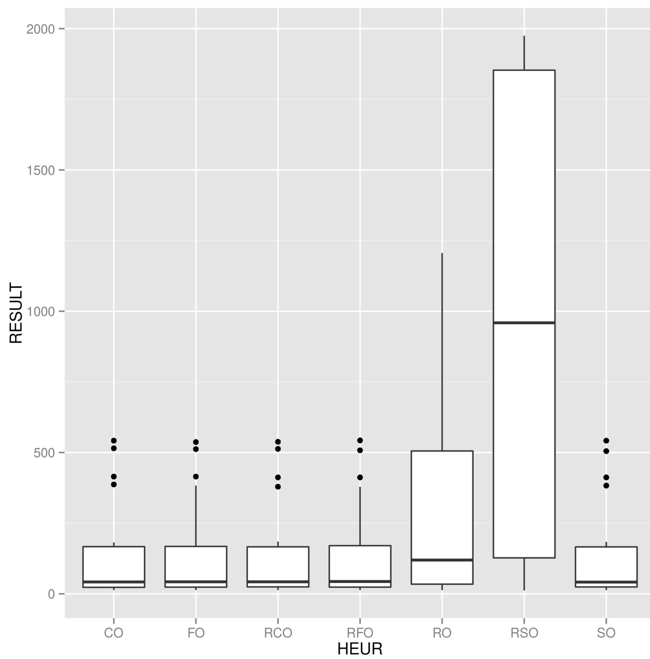


Here, *u* denotes a query containing a single unary predicate; *uub* denotes a query containing two unary predicates and a single binary predicate. *uubbt* denotes a query containing three unary predicates, two binary predicates, and a single ternary predicate. *uuut* denotes a query containing three unary predicates and a single ternary predicate.

3. Constraint-Ordering Heuristics

   As seen in 3.8, *DIST-POWER* performed better when constraints were sorted by arity or coverage, with reverse coverage ordering and cost ordering coming in close behind. It should not be surprising that the performance of coverage, cost, and reverse-coverage ordering is similar to that of arity ordering: the majority of experimental data was collected in circumstances where all relations had the same cost and coverage, in which cases those heuristics performed identically to arity-ordering.

Figure 3.8: Effect of Heuristics



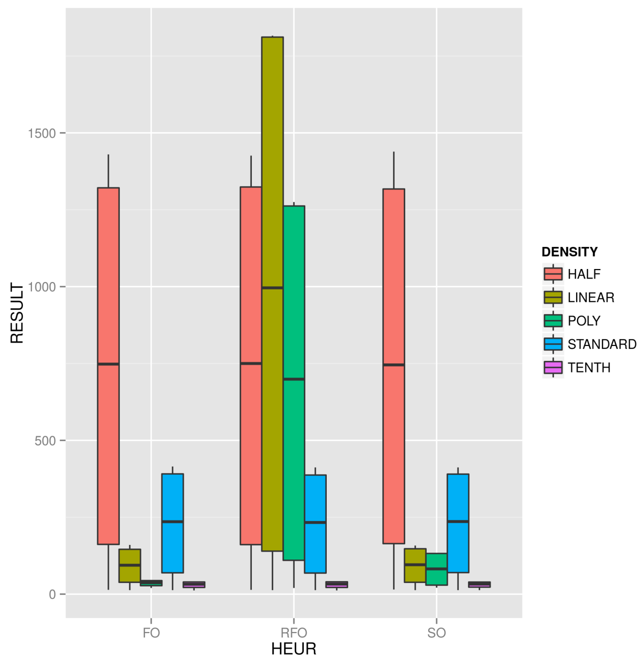
Here, *CO* denotes the cost-ordering heuristic, *FO* the coverage-ordering heuristic, *RCO* the reverse cost-ordering heuristic, *RFO* the reverse coverage-ordering heuristic, *RO* the random ordering heuristic, *RSO* the reverse arity-ordering ordering heuristc, and *SO* the arity-ordering heuristic.

We thus decided to examine the interaction between cost, coverage, and heuristic choice.

4. Interaction between Knowledge Base Characteristics and Constraint-Ordering Heuristics

   As seen in 3.9, reverse-coverage ordering heuristic was simply a bad choice under certain conditions (i.e., when a few (but not all) constraints have high coverage) but that these three heuristics did not in general *DIST-POWER* performs better when constraints have low coverage. This is to be expected as this simply translates into a lower number of possible solutions.

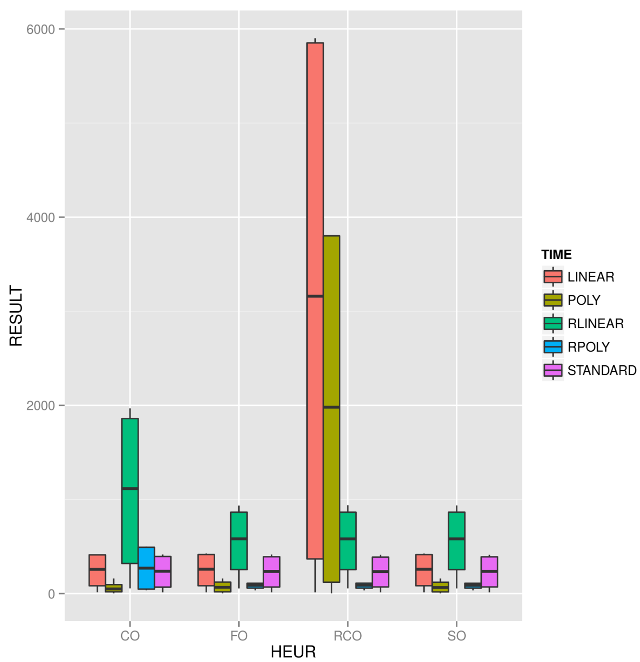
Figure 3.9: Interaction of Coverage and Heuristics



Here, FO denotes the coverage-ordering heuristic, RFO the reverse coverage-ordering heuristic, and SO the arity-ordering heuristic. Density indicates probabilities for whether predicates hold for particular entities: 50% probability, probability linearly increasing with arity, probability polynomially increasing with arity, the standard 30% probability, and 10% probability.

As seen in Figure 3.10, coverage and arity ordering performed significantly better than cost or reverse-cost ordering. A significant interaction effect between cost and heuristic was found ($F(12,60)=2.263$, $p<0.02$), suggesting that this difference in heuristic occurs specifically when polyadic predicates are more expensive than others. Finally, a marginal effect of cost was found, suggesting, as expected, that *DIST-POWER* performs better when constraints have lower costs.

Figure 3.10: Interaction of Cost and Heuristics



Here, *FO* denotes the coverage-ordering heuristic, *CO* the cost-ordering heuristic, *RCO* the reverse cost-ordering heuristic, and *SO* the arity-ordering heuristic. Time indicates whether time taken to apply constraints increased linearly, increased polynomially, decreased linearly, decreased polynomially, or held constant with respect to constraint arity.

Discussion

As expected, sorting in *increasing* order of coverage, cost, and arity proved to be significantly more efficient than sorting in a decreasing or random order. And, as shown in Figure 3.10, for example, knowledge base characteristics can make certain heuristics more effective than others. That being said, it appears that sorting by arity is almost always the best choice; Figure 3.9 shows small gains for other heuristics (i.e., coverage ordering) under certain circumstances, but those gains were not statistically significant, and it thus seems that regardless of knowledge base characteristics, arity-ordering should be used.

While we did not investigate these aspects in this work, it would be useful in the future to further investigate performance under a different set of modifications, such as when knowledge bases vary in size, or when more complicated heuristics are used. If more sophisticated heuristics were found to result in knowledge base-dependent results, it may be useful to train a classifier to learn optimally efficient query orderings.

3.4.6 Discussion of Architectural Desiderata

Finally, we would like to discuss how the experiments demonstrate the architectural commitments of *DIARC* facilitated by *DIST-POWER*. First, the architecture does not prescribe any single knowledge representation. This is facilitated by distributing information amongst knowledge bases of *heterogeneous* representation. Second, the architecture uses formulae for *inter-component communication* whenever possible. This is facilitated by accepting queries represented as sets of formulae. Finally, architectural components should perform processing asynchronously, with components possibly spread across multiple computers. This is facilitated by allowing information and processing to remain localized in separate components, rather than enforcing consolidation into a single knowledge base. However, the proposed algorithm is neither incremental nor parallelized, aspects which would yield tighter adherence to this architectural commitment. These topics are further discussed in Section 9.4.1.

3.5 Previous Work in Robotics

At this point, it is useful to take a step back, and compare the work presented thus far to previous work on reference resolution in robotics. As previously discussed, we are primarily interested in open-world reference resolution. While there has been significant work on open-world *directive grounding*(Matuszek, Herbst, Zettlemoyer, & Fox, 2012; MacMahon, Stankiewicz, & Kuipers, 2006), in which natural language utterances are translated directly into *action sequences* (thus bypassing the need to ground constituent noun phrases) these has been relatively little work in open-world *reference resolution*. In this section we will thus discuss both closed-world and open-world approaches.

Work on reference resolution in robotics can be traced back to Terry Winograd's SHRDLU system(Winograd, 1971), in which a simulated robot used a *procedural semantics* approach to natural language understanding in order to carry out commands in a simple environment. Under this approach, each lexical item was associated with a short procedure, such as a search through objects in the scene, which was executed when that lexical item was encountered.

Many research efforts over the following decades took inspiration from SHRDLU and employed a similar approach. For example Gorniak and Roy associated lexical items with procedures that greedily applied composable continuous *visual models* in order to produce a solution to the full language grounding problem (i.e. Gorniak & Roy, 2004; D. Roy, Hsiao, Mavridis, & Gorniak, 2003). This was a significant improvement over SHRDLU's approach, in which objects' properties were assessed by checking a database of hand-assigned symbolic properties.

Kruijff et al. also take a SHRDLU-like approach, employing a set of *comparators* that assess whether certain entities satisfy certain properties (Kruijff, Lison, Benjamin, Jacobsson, & Hawes, 2007). Unlike Gorniak and Roy, however, Kruijff et al. use these comparators to address the reference resolution half of the language grounding problem – the symbol grounding half is addressed using a separate process that binds information from different modalities into composite representations in a centralized knowledge base.

Other researchers have also taken a "knowledge-based" approach in which properties are assessed based on the information stored in a centralized knowledge base. For example, Lemaignan uses a semantic parser to translate utterances into lists of RDF triples (Klyne & Carroll, 2006); for example, 'the yellow banana' is translated into {((?obj type banana) (?obj hasColor yellow))}. These triples can then be used to query a central knowledge base populated by input from perception systems, thus producing the set of entities in that knowledge base that satisfy the conjunction of triples (Lemaignan, Ros, Alami, & Beetz, 2011).

Similarly, Zender et al., who focus on reference resolution in the domain of large-scale topological spaces such as rooms and hallways (as opposed to the domain of objects used by the previous approaches), parse utterances into SPARQL queries (Prud'Hommeaux & Seaborne, 2008) (a particular form of RDF query) (Zender, Kruijff, & Kruijff-Korbayová, 2009). This approach also differs from the approach used by Lamaignan through the use of a dedicated *co-reference resolution* step, which attempts to add the references found in an utterance to clusters of references found in past utterances – a step which results in resolution of some anaphoric expressions.

Meyer uses a pair of tightly coupled co-reference resolution and reference resolution algorithms in order to jointly resolve anaphoric and non-anaphoric references (Meyer, 2013). The reference resolution algorithm used by Meyer uses a Markov Logic Network whose weights are learned based on the connections between lexical items and the taxonomic classes of possible referents.

Chai et al. also use a co-reference resolution pre-processing step. After this step, Chai et al. use incoming utterances and perceived deictic gestures to build up a graph representing the relations between the entities mentioned in conversation, and perform reference resolution by finding the best partial match between this graph to a similar graph that represents the relations between entities observed in the world (J. Y. Chai, Hong, & Zhou, 2004; Fang, Liu, & Chai, 2012; C. Liu, Fang, She, & Chai, 2013; J. Y. Chai et al., 2014).

A different approach is taken by Fasola & Matarić (2013), through their work on semantic fields. Fasola and Mataric use a simple reference resolution procedure in which a knowledge base of labels is checked when particular nouns are used – their approach is interesting, however in how they process relations. When a noun is ambiguous, if that noun is a constituent of a prepositional phrase it is disambiguated using a semantic field: a data-driven model of the preposition that produces a probability distribution over coordinates in the environment; the referent whose location has the highest probability value according to this distribution is selected as the referent.

A probabilistic approach is also taken by a number of Bayesian modelers. Kennington and Schlangen present an incremental Bayesian model in which each word is used to modulate the probability of reference for each entity in a scene (Kennington & Schlangen, 2017). Similarly, Tellex and Kollar's Generalized Grounding Graph (G^3) approach uses utterances (after a co-reference resolution pre-processing step) to instantiate probabilistic graphical models that are used to resolve references (Tellex et al., 2011a, 2012). This approach has been extended by Tellex and Kollar's colleagues through the Hierarchical Distributed Correspondance Graph approach, which differs in that it uses the "type" associated with each observed noun to restrict the set of possible values associated with each noun-node in the resulting graphical model (Chung, Propp, Walter, & Howard, 2015).

Finally, similar to all three of these approaches, Matuszek et al. present an approach in which utterances are parsed into lambda expressions associated through training with a set of visual classifiers used to identify objects, each of which returns a probability value representing its confidence that a given object fits a given property.

Each of the approaches mentioned thus far in this section addresses, at the least, the *classic reference resolution problem*[15]: given a definite description, a set of candidate referents from a common domain, and a set of

[15] We use this nomenclature to draw a parallel to the *classic REG problem*(Van Deemter, 2016) which is very much the counterpart of the classic reference resolution problem.

properties held by each of those referents, determine the candidate referent associated with each entity mentioned in the definite description.

But solutions to this "classic" problem framing are not sufficient for robots operating in realistic human-robot interaction scenarios, for a number of reasons discussed thus far in this dissertation, which I will now lay out again: First, robots cannot assume that referring expressions will always come in the form of definite descriptions: interlocutors may use *anaphoric* expressions (e.g., 'it') that reference entities previously mentioned in dialogue; or they may use *deictic expressions* (e.g., 'this') that reference entities based on their joint situated perspective with the robot. Second, robots cannot assume that candidate referents will be drawn from a single domain; interlocutors may refer in a single utterance to some combination of locations, objects, people, utterances, ideas, actions, and so on. Third, robots cannot assume that candidate referents will even be known *a priori*; interlocutors may refer to entities that were previously unknown to the robot. Finally, robots cannot assume that they will have perfect knowledge regarding the properties of objects: they may only have confidence *to some extent* that a certain property or relation holds for a certain object or set of objects. In this section, we will analyze the previous approaches and assess the extent to which they address each of these four additional concerns.

3.5.1 Anaphoric and Deictic Reference

Many of the discussed approaches handle anaphoric reference to at least a limited extent. Winograd (1971) associated anaphoric expressions such as 'it' with special procedures that gave preference to elements considered to be 'in focus' (see also Mitkov, 1999); a simpler procedure is used by Gorniak & Roy (2004). Kruijff, Lison, Benjamin, Jacobsson, & Hawes (2007) also select items based on focus when 'this' is used, and use occurrences of 'it' to constrain search to the domain of objects. Lemaignan, Ros, Alami, & Beetz (2011) handle anaphora by replacing anaphoric references with the last entity in the dialogue history that matches the animacy and gender constraints imposed by that referent (which serves to restrict uses of 'it', for example, to objects, and uses of 'he', for example, to men) – an approach also taken by Fasola & Matarić (2014).

As previously discussed, the approaches of Zender, Kruijff, & Kruijff-Korbayová (2009); Meyer (2013); J. Y. Chai et al. (2014); Tellex et al. (2012) handle anaphora through dedicated co-reference resolution pre-processing stages. Kennington & Schlangen (2017) handle anaphora by attributing a special property to entities selected in dialogue, and then statistically associating pronouns with that special property through training.

Fewer approaches handle deictic references. Kruijff, Lison, Benjamin, Jacobsson, & Hawes (2007) use deictic references to impose *preference orderings* over candidate referents. Lemaignan, Ros, Alami, & Beetz (2011) resolve deictic expressions to the last entity in the dialogue history that was the focus of simultaneous eye gaze and gesture. J. Y. Chai, Hong, & Zhou (2004) incorporate gestural information into their dialogue graph structures. Kennington & Schlangen (2017) handle deixis and gaze by linearly combining the probability of reference given an utterance with the probability of reference given gaze and the probability of reference given gesture; an approach also taken by Matuszek, Fitzgerald, Zettlemoyer, Bo, & Fox (2012).

3.5.2 Domain Independence

The majority of the examined approaches are dependent on a particular domain. The approaches of Winograd (1971); Gorniak & Roy (2004); Lemaignan, Ros, Alami, & Beetz (2011); J. Y. Chai et al. (2014); Fasola & Matarić (2013); Kennington & Schlangen (2017); Matuszek, Fitzgerald, Zettlemoyer, Bo, & Fox (2012) were designed to operate in the domain of objects in a visual scene.

Zender, Kruijff, & Kruijff-Korbayová (2009), on the other hand, operates in the domain of large-scale topological locations (similar to SPEX, Section 3.2).

Meyer (2013) appears to consider objects and units of time, with entities from both domains stored in a single, centralized knowledge base. Similarly, Kruijff, Lison, Benjamin, Jacobsson, & Hawes (2007)'s approach understands references to both objects and small-scale locations (i.e., local points in space), with information from both domains stored in a single, centralized knowledge base (but informed by a set of independent sensory systems).

The approach of Tellex et al. (2011a), like our own initial work on *DIST-POWER*, makes steps forward with respect to these previous approaches. The approach presented by Tellex et al. is not hand tailored to a particular domain, but appears to handle references to entities from whatever data set it is trained on – so long as they are physically extant and can be grounded to coordinates in Cartesian space. We expect that this assumption is also true of the work presented by Chung et al. (2015). Similarly, our approach is a domain-independent framework into which domain-dependent algorithms can be used as "consultants". Our approach, however, does not make any assumptions about the physical existence of candidate entities.

3.5.3 Operation in Uncertain Worlds

It is important to note that the Bayesian approaches do not handle uncertainty in the way we describe: the approaches presented by Kennington and Schlangen, Tellex and Kollar, and Chung et al. represent uncertainty with respect to the relationship between words and features, but not the uncertainty in whether certain entities have certain features. And in fact, representing this uncertainty would undermine the features of some of these algorithms. Chung et al., for example, use entity type to restrict the values that need to be considered for each noun-node in the probabilistic graphical models instantiated by their approach. This approach would need modification if there was uncertainty as to an entity's type.

While the Semantic Fields approach does not appear able to handle uncertain properties, it does handle uncertain spatial relations (Fasola & Matarić, 2013). Finally, Fang, Liu, & Chai (2012) describes how Chai et al.'s approach handles uncertain properties by incorporating an *extent of compatibility* measure into their graph-matching scoring functions; the approach taken by Matuszek, Fitzgerald, Zettlemoyer, Bo, & Fox (2012) is able to represent the uncertainty in the properties of the objects it reasons about, based on classifier confidences; and the *DIST-POWER* framework is specifically designed to use domain-specific "consultants" that provide probability values of just this sort.

3.5.4 Operation in Open Worlds

Of the discussed approaches, only work from two groups begins to address operation in open worlds. Recent work from Duvallet et al. in the G^3 framework allows a robot to handle references to previously unknown objects described in relation to previously known objects (Duvallet et al., 2014). This approach is limited, however, to spatially situated objects: the pose of the new object is sampled with respect to the other object according to a learned distribution. The *DIST-POWER* framework is also able to hypothesize new entities, but is domain independent in nature, and thus does not have this limitation (see also Williams & Scheutz, 2015b).

3.6 General Discussion

We have argued that a robot operating in natural human-robot interaction scenarios must use a domain-independent reference resolution algorithm capable of handling not only definite descriptions, but also anaphoric and deictic expressions, and must do so in both uncertain and open worlds. The algorithms presented in this chapter (see also (Williams, Cantrell, Briggs, Schermerhorn, & Scheutz, 2013; Williams & Scheutz, 2015a,b, 2016a)) make progress towards this goal, but fall short in two main respects: (1) they are unable to handle the wide variety of expressions that occur in situated dialogue *beyond* definite noun phrases; and (2) they require all entities known of by a robot's various consultants to be considered when resolving any referring expression, leading to an explosion of computational complexity. In the next chapter, we present a new set of algorithms that address these concerns by embedding *DIST-POWER* into a larger framework inspired by a principled linguistic framework known as the *Givenness Hierarchy* (GH).

Chapter 4

Reference Resolution in Context

In the previous chapter, we presented algorithms for resolving references occurring in definite noun phrases. Those algorithms were designed to handle the open worlds (Williams, Cantrell, Briggs, Schermerhorn, & Scheutz, 2013; Williams & Scheutz, 2015a,b) and uncertain contexts (Williams & Scheutz, 2015a,b) commonplace in natural human-robot interaction scenarios.

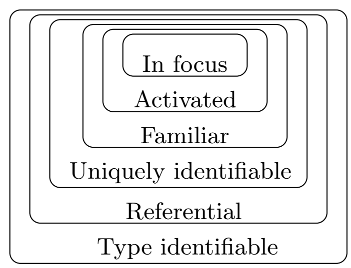
In this chapter, we present open-world reference resolution algorithms that can handle a wider array of linguistic forms by using the *Givenness Hierarchy* (GH) (J. K. Gundel, Hedberg, & Zacharski, 1993), a linguistic framework which associates the form of a referential expression (e.g., pronominal, definite noun phrase, indefinite noun phrase) with a presumed "cognitive status" (e.g., focus of attention, short term memory, long term memory). This significantly advances the state of the art of natural-language based HRI, by (1) increasing the breadth and complexity of referring expressions understandable by robots, (2) allowing robots to understand such expressions in *open* and *uncertain* worlds, and (3) bringing robot natural language understanding closer in line with an established linguistic framework (i.e., the GH). What is more, it is significant in its extension of the GH itself, through the addition of guidelines which clarify how the GH should be computationalized.

The rest of the chapter proceeds as follows. In Section 4.1 we present our linguistic motivations, and describe the basics of the GH. In Section 4.2 we discuss previous work in computationalizing the GH, and explain how those previous implementations might be improved if clear guidelines for using the GH could be crafted. In Section 4.3 we suggest such guidelines for the

GH, and present and evaluate GH-POWER: an algorithm which uses those guidelines to improve on previous approaches. In Section 4.4 we present GROWLER: an algorithm that expands on GH-POWER by accounting for a notion of conversational relevance.

4.1 Linguistic Motivations

Figure 4.1: The Givenness Hierarchy



The GH (J. K. Gundel, Hedberg, & Zacharski, 1993) is comprised of six hierarchically nested tiers of cognitive status, as seen in Figure 4.1. If a candidate referent is marked as having one of these statuses, the hierarchical nature of this framework means that it also has all statuses lower in the hierarchy. For example, a candidate referent that is *familiar* is also *uniquely identifiable*, *referential*, and *type identifiable*. It is *possible* that the candidate referent is also *activated*, or even *in focus*, but a higher status cannot be inferred from a lower status. Each level of the GH is "cued" by a set of linguistic forms, as seen in Table 4.1 for English. For example, the second row of the table shows that when the definite 'this' is used, one can assume that *the speaker assumes* the referent of 'this' to be *at least* activated for their interlocutor.

Table 4.1: Cognitive Status and Form in the GH

Cognitive Status	Linguistic Form
In focus	*it*
Activated	*this,that,this* N
Familiar	*that* N
Uniquely identifiable	*the* N
Referential	indefinite *this* N
Type identifiable	*a* N

The GH is attractive to computational researchers not only because it suggests a clear mapping between linguistic form and cognitive status, but because, due to its focus on *means of access* rather than *salience*, each status evokes a particular *mnemonic actions* (i.e., actions involving selecting or creating mental representations) upon an agent's cognitive structures.

When the linguistic form of an expression explicitly signals that its referent is type identifiable or referential (but not necessarily uniquely identifiable), this suggests the action of *hypothesization*: creating a *new* mental representation, and then selecting that representation as the target referent.

When the linguistic form of an expression signals that its referent can also be uniquely identified (but is not necessarily familiar), this suggests *either* the action of hypothesizing a referent *or* selecting an existing referent from memory. When the linguistic form of an expression signals that its referent is also familiar, this suggests that the referent should be able to be found by searching through memory and selecting an existing representation.

When the linguistic form of an expression signals that its referent is also activated or in focus, this suggests that the referent should be able to be found by searching through a subset of memory (the subset of activated entities and the subset of activated entities that are in focus, respectively) and selecting a referent from that subset.

The GH can directly solve certain computational problems: To determine the cognitive status ascribed to a candidate referent, one need only check which forms explicitly encode which statuses on the GH in a given language (see also the Coding Protocol provided by J. K. Gundel, Hedberg, et al. (2006)). And, when Speaker S uses linguistic form F to refer to entity E when speaking to hearer H, it is easy to determine the *most restrictive* status that H can rationally assume S to ascribe to E. For example, when S uses 'it', we can assume that S believes E to be in the subset of H's memory that is *in focus*: any information that could *not* plausibly be *in focus* can be ruled out, as such an interpretation *would not be possible* given the cognitive status conventionally signaled by 'it'; when S uses 'this', we can assume that S believes E to be at least in the subset of H's memory that is currently *activated*. E may also be in the subset of those entities that are *in focus*, but we can not assume this; and in fact, it is unlikely that S believes E to be in that subset, as otherwise S could have used the more informative 'it'. Furthermore, information that could not plausibly be in the *activated* subset of H's memory can be ruled out, as such an interpretation *would not be possible* given the cognitive status conventionally signaled by 'this'.

However, within the GH framework, choices among referents that meet cognitive status restrictions are made through *interaction* of the GH with general pragmatic principles operative in language interpretation, such as Grice's Maxims(Grice, 1970) or Relevance theory(Sperber & Wilson, 1986). As a result, there are other computational problems for which the GH itself can only *facilitate*, but not *directly produce* solutions:

1. The *referring expression generation* task:
   "When S wishes to refer to E when speaking to H, what linguistic form F should be used?" (cf. Van Deemter, 2016; Krahmer & Van Deemter, 2012)

2. The *reference resolution* task:
   "When S uses linguistic form F when speaking to H, what entity E is most likely being referenced?" As discussed, this is the task for which we wish to use a GH-theoretic approach.

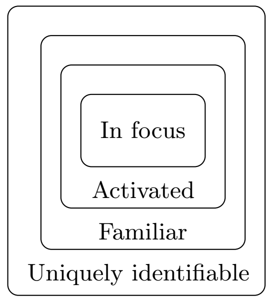
It is thus unsurprising that there have been several attempts to use the GH to inform reference resolution algorithms in the fields of Human-Robot and Human-Agent Interaction. We will now describe the two implementations which, until now, have made the most extensive use of the GH.

4.2 Previous Computational Implementations of the Givenness Hierarchy

The first implementation of the GH that we will examine is that presented by Kehler (2000), in which they propose the modified hierarchy seen in Figure 4.2. There, Kehler omits the last two levels of the GH, due to a primary interest in interfaces with which it is unlikely for one to refer to unknown or hypothetical entities. Kehler used his modified hierarchy to craft four rules (presented here verbatim) capable of resolving all references he encountered:

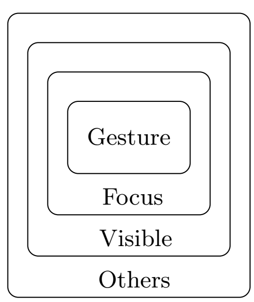
1. If the object is gestured to, choose that object
2. Otherwise, if the currently selected object meets all semantic type constraints imposed by the referring expression (i.e., "the museum" requires a museum referent; bare forms such as "it" and "that" are compatible with any object), choose that object.
3. Otherwise, if there is a visible object that is semantically compatible, then choose that object (this happened three times; in each case there was only one suitable object).
4. Otherwise, a full NP (such as a proper name) was used that uniquely identified the referent.

Figure 4.2: Kehler's Modified Hierarchy



The second GH implementation we will examine, presented by J. Chai, Prasov, & Qu (2006), expands on Kehler's approach in two important ways: First, Chai et al.'s implementation can identify and resolve ambiguities (Kehler's first rule is problematic if the target of a gesture is ambiguous, and Kehler's third rule is problematic if a referring expression is ambiguous). Second, Chai et al.'s implementation makes it possible to handle utterances containing multiple referential expressions or gestures. To make these advancements, Chai et al. combine a subset of the GH with Grice's theory of Conversational Implicature (Grice, 1970) to produce the modified hierarchy seen in Figure 4.3.

Figure 4.3: Chai's Modified Hierarchy



Chai et al.'s modified hierarchy contains four tiers: (1) "Gesture", containing entities gestured toward (because a gesture *intentionally* singles out entities), (2) "Focus", combining Gundel's *in focus* and *activated* tiers, (3) "Visible", combining Gundel's *familiar* and *uniquely identifiable* tiers, and (4) "Others", combining Gundel's *referential* and *type identifiable* tiers, although this tier does not appear to be used, perhaps due to the lack of hypothetical entities in graphical interfaces.

Chai et al. present a greedy reference resolution algorithm using their hierarchy. This algorithm first assigns a score between each referring expression X in an utterance and each entity N contained in a set of vectors (Gesture, Focus, Visible), calculated by multiplying (1) the probability of selecting N from its vector, (2) the probability of selecting that tier given the form of X, and (3) the "compatibility" between X and N. Compatibility is 1 if the N has all properties mentioned in X, is of the type mentioned in X (if any), has the name mentioned in X (if any), and was gestured towards when X was uttered (if any), and 0 otherwise; it is thus *binary* in nature and cannot account for uncertainty.

After scoring all visible entities, the algorithm greedily assigns reference-entity matches, moving downward through the hierarchy of vectors.

This approach does not address all aspects of reference resolution found in typical human-robot dialogues (nor does any other current approach). There are, in particular, five aspects of human-robot dialogue not captured by this approach.

First, the algorithm assumes complete certainty as to entities' properties. In realistic HRI scenarios, an agent may only be able to say that an entity has a certain property *with some probability.* Furthermore, an agent could be aware that it simply *does not know* whether an entity has a certain property.

Second, consider the following command:

"Get **my laptop** from **my office**, and if you see **a charger** bring that too."

The three bolded referring expressions present issues for Chai et al.'s approach. **My laptop** is (presumably) not currently visible, a condition common in many HRI scenarios, but one which cannot currently be handled using Chai et al.'s algorithm. **My office** is also (presumably) not currently visible. And, it is not an object, per se, and cannot be gestured towards in the same way as can be objects or icons. It is unclear whether Chai et al.'s modified hierarchy could handle references to locations, which are common in many HRI scenarios. **A charger** is also (presumably) not currently visible. And, it is not even known to exist, as it is *hypothetical.* In order to resolve such references, one must assume an *open world* in which new entities may be added through experience or dialogue. While many HRI scenarios are open-world in nature, Chai et al.'s algorithm operates in a closed world.

Third, a robot may need to resolve references to events, speech acts, or other entities that *cannot* physically exist, as seen in Examples 2 and 3. However, Chai et al.'s algorithm cannot handle references to nonexistent entities.

Fourth, because Chai's modified hierarchy combines the first two levels of the GH, Chai et al.'s algorithm cannot distinguish between Examples 2 and 3 even if it *could* handle references to physically nonexistent entities.

(2) Can you repeat it?

(3) Can you repeat that?

When Example 2 is used to respond to the utterance "I'm sorry, but I failed to complete the task", 'it' unambiguously refers to 'the task'. However, this is not the case when Example 3 is used. The GH predicts that when a form associated with the *activated* level is used, one should prefer an activated referent (such as a speech act) to an in-focus referent (such as the focus of the previous sentence), because if the speaker had meant to refer to an in-focus entity she could have used an in-focus-cueing form (e.g., 'it'). Thus, while Example 3 could refer to either the speech act or failed task, the speech act should be preferred [1].

Fifth, natural human-robot dialogues may contain complex noun phrases such as "Do you see the red block on that blue block?" Because Chai et al.'s algorithm uses a greedy approach (instead of, e.g., the graph matching approach used in their previous work), it may choose an incorrect referent for the first considered referential expression, and may thus be unable to successfully resolve subsequent referring expressions. Chai et al. argue that using a greedy approach is advantageous because it allows significant pruning of the search space. However, their algorithm scores all entities against all referring expressions *before* employing its greedy approach. In a realistic HRI scenario, this may not be practical, as a robot may know of hundreds or thousands of entities. Furthermore, the process of checking whether certain properties hold for all entities may be cost prohibitive. For example, while determining whether a given person is a man may be accomplished by a simple database look-up, determining whether two rooms are across from each other may require more expensive computation. An algorithm which performed such assessments lazily (i.e., only when needed, perhaps as the search space was pruned) could be much more efficient.

Thus far, we have described reasons for extending Chai et al.'s modified hierarchy and algorithm. But to make the needed extensions, we must first

[1] Gundel et al. have empirically verified that these two hierarchical levels are distinguished between in a wide variety of languages beyond English, including Eegimaa, Kumyk, Ojibwe, and Tunisian Arabic (each of which is genetically and typologically unrelated to the other three.) (J. K. Gundel, Bassene, Gordon, Humnick, & Khalfaoui, 2010)

extend the GH itself: each extension we have discussed thus far can be related to an area for which the GH lacks clear usage guidelines. No existing GH-based approach can handle uncertain information, perhaps because the GH neither specifies how uncertainty is handled nor provides guidelines for how intra-tier ambiguity is resolved. GH-based approaches must be extended to better resolve multiple referring expressions occurring in the same utterance, in order to avoid incorrect greedy decisions. This is because the GH does not provide guidelines for how multiple related referents are simultaneously resolved.

Chai et al.'s approach cannot handle references to entities that are unknown, hypothetical, intangible or not present. This is the result of Chai et al.'s omission and combination of GH tiers, and their use of a purely top down traversal. This may have been avoided if clear guidelines had existed for traversing the tiers of the GH and for guiding intra-tier search using salience arising from linguistic, visual or gestural factors. We thus believe that a GH-based reference resolution algorithm for human-robot dialogue requires the following:

Clear guidelines for:

1. Determining the order in which to peruse the tiers of the Givenness Hierarchy that allow gestured-towards or gazed-upon entities to take some degree of precedence.
2. Resolving complex referring expressions.
3. Choosing between candidates found within a given tier.

Assumptions of:

1. Uncertain information (i.e., the properties of an entity may not be certain or known)
2. An open world (i.e., the *existence* of an entity may not be certain or known)
3. Global resolution (i.e., a referring expression may refer to an entity which is not currently visible)
4. Domain independence (i.e., a referring expression may refer to *any* entity, regardless of type or tangibility).

To satisfy these needs, we presented the GH-POWER algorithm(Williams, Acharya, Schreitter, & Scheutz, 2016), which we outline in the next section.

4.3 GH-POWER

The GH-POWER reference resolution algorithm dictates how the referent of a referring expression should be searched for, given a memory model organized in a specific, hierarchical way that parallels the organization of the GH. In this section, we will first discuss the memory structure used in our approach. Next, we will discuss the *between-structure processes* by which GH-POWER algorithm chooses which structures to search. We will then discuss the *within-structure processes* by which GH-POWER selects suitable referents from a given structure.

4.3.1 The GH-POWER Memory Model

The memory model used by *GH-POWER* aligns well with Nelson Cowan's conceptualization of working memory(Cowan, 1998). According to Cowan, working memory and long-term memory are not disjoint structures. Rather, working memory can be regarded as the subset of entities in long term memory that are currently activated. Cowan further posits an additional substructure, the focus of attention, which is a subset of those activated entities that is limited in size to at most four elements, comprised of those items of which an agent is consciously aware. There is clearly a strong parallel between Cowan's *Focus of Attention* ⊂ *Set of Activated Entities* ⊂ *Long Term Memory* structures and Gundel's *In Focus* ⊂ *Activated* ⊂ *Familiar* statuses, and observing this connection will facilitate understanding the connection between our own memory structure and the statuses of the GH.

Our approach consists of *four* hierarchically nested data structures: the *Focus of Attention* (FoA), *Set of Activated Entities* (ACT), *Set of Familiar Entities* (FAM) and *Long Term Memory* (LTM). These four data structures are hierarchically organized such that FoA ⊂ ACT ⊂ FAM ⊂ LTM. At the *computational level* of analysis(Marr, 1982), the FoA, ACT, and LTM data structures are identical to Cowan's three memory structures. But in a robot architecture, all of a robot's knowledge is not typically located in a single, monolith knowledge base. Instead, it may be distributed across a *set* of knowledge bases that may be located on different machines, may use different knowledge representation schemes, and may have different ways of accessing and modifying the knowledge contained within them. Thus, at the *algorithmic level*, our LTM data structure is really a set of domain-specific *distributed, heterogeneous knowledge bases.* Because LTM is not a single coherent knowledge base, the FoA and ACT also must differ at the algorithmic level; instead of being literal subsets of the mental representations distributed across LTM, the FoA and ACT instead contain *memory traces* that allow access to certain of those mental representations. Note that these three structures are not intended to serve as the agent's *actual* cognitive structures; instead, they serve to model what an interlocutor might *believe* to be in those structures, and thus as a model of *common ground*.

Finally, for the sake of convenience and efficiency, we introduce the FAM structure, a minor point of departure from both the GH and Cowan's model of Working Memory, which we make for practical rather than theoretical reasons. FAM contains memory traces for entities in LTM that are likely to be referenced, such as entities mentioned at some point in the robot's current dialogue, recently visited locations, and recently visited objects, including all entities in ACT (and by extension, in the FoA). Because searching all of LTM is potentially expensive, when LTM needs to be searched for an entity that matches some criteria, that search is preempted by a search of FAM: if a match can be found there, LTM need not be searched.

To summarize, our model consists of four hierarchically nested data structures: a distributed LTM data structure containing mental representations of known entities, and three smaller data structures that contain memory traces allowing fast access to entities in LTM (i.e., FoA, ACT, and FAM). These three data structures are populated periodically (e.g., after an utterance is processed) according to rules inspired by the GH Coding Protocol. In the next two sections, we will describe how these structures are used during reference resolution. We will begin by providing a high-level overview of the GH-POWER algorithm: how the linguistic form of a referring expression is used by the GH-POWER algorithm to determine *which* of these structures to examine, and how GH-POWER chooses whether a particular candidate referent within one of those structures is the target referent. We will then provide an *algorithmic* description of this algorithm, and discuss its integration into our robot architecture.

4.3.2 Process-Level Description

Between-Structure Processes

The GH alone does not specify how cognitive structures are selected for perusal during reference resolution. For example, suppose Speaker S uses the pronoun 'that' to refer to entity E when speaking with Hearer H. The GH suggests that H can assume that S *assumes* that E is at least in H's ACT, and thus *may or may not* also be in H's FoA.

Several strategies could be used to search ACT and the FoA. The agent could consider entities in the FoA, then out-of-focus entities in short term memory (a top down approach), or she could consider out-of-focus entities in ACT, then in-focus entities (a bottom up approach).

While some previous approaches (e.g. J. Chai, Prasov, & Qu, 2006) have used a global top-down approach, this may violate certain predictions of the GH. For example, the Givenness Hierarchy framework (i.e., the GH when working in conjunction with general cognitive principles such as Grice's Maxim of Quantity) suggests that in the example above, while the referent of 'that' *could be* assumed to be in H's FoA, it is more likely to be assumed to be in H's ACT *but not in H's FoA*, as otherwise S could have used 'it' to refer to the referent. If a purely top-down approach is used, this effect may not be captured. On the other hand, consider the utterance "Pick up the box". The bottom-up approach would inappropriately prioritize inactive boxes from LTM over an activated box in front of the listener. Since neither a purely top-down or purely bottom-up approach seems adequate, we developed a hybrid approach, in which a unique search strategy is used for each GH tier. These strategies, refinements of those we previously presented (Williams, Acharya, Schreitter, & Scheutz, 2016) are seen in Table 4.2. In that table, FoA denotes a search through memory traces found in the FoA; ACT denotes a search through memory traces found in ACT *but not in* the FoA; FAM denotes a search through memory traces found in FAM *but not in ACT* LTM denotes a search through all of LTM; HYP denotes hypothesization. We will now explain the rationale for each strategy.

Table 4.2: Search Plans for Complete GH

Level	Search Plan
in focus	FoA
activated	ACT → FoA
familiar	ACT → FoA → FAM → LTM
uniquely identifiable	ACT → FoA → FAM → LTM → HYP
referential	ACT → FoA → HYP
type identifiable	HYP

1. In Focus

   In the case of an "in focus" cuing form (e.g., 'it'), we only search the FoA, as it would be otherwise inappropriate to use such a form.

2. Activated Entities

In the case of an "activated" cuing form (e.g., 'this'), search is expanded to include out-of-focus entities in ACT. For the reasons discussed above, we proceed bottom-up, first searching the out-of-focus entities in ACT, then searching the FoA. However, this process is modified in the case of 'This N', as we discuss below.

3. Familiar Entities

   In the case of a "familiar" cuing form (e.g., 'that N'), search is expanded to include all entities in memory. As it is inappropriate to *prioritize* entities in LTM over those in ACT, we still perform our search through ACT and the FoA first, and then move on to search through LTM. As previously discussed, we first search through FAM, the subset of most probable referents in LTM (not including those referents also found in ACT), and only search *all* of LTM if this search fails, using the previously described *DIST-POWER* algorithm, which has two main features relevant to GH-POWER: its ability to simultaneously resolve all parts of a complex definite description, and a feature discussed in the following subsection.

4. Uniquely Identifiable

   In the case of a "uniquely identifiable" cuing form (e.g., 'the N'), search is extended to allow for the possibility that the speaker is referencing a previously unknown entity. This search process begins by searching through the four tiers of the GH-POWER memory model, as performed with familiar entities. However, when searching through LTM, we take advantage of *DIST-POWER's* second important feature *DIST-POWER's* "hypothesization mode". When run in this mode, if *DIST-POWER* is unable to find a mental representation that satisfies all semantic criteria of a definite description, it attempts to find a subset of that description that it *can* successfully resolve, and automatically hypothesizes representations for remaining entities.

5. Referential

   Gundel et al. suggest that the *indefinite* form of 'this N' (as in "This dog I saw was enormous!") cues the "referential" tier[2], resulting in the hypothesization of a representation. As a simplification (i.e., so that we do not need to decide whether each use of 'This N' is definite or indefinite), GH-POWER deals with both forms at the referential tier. To

[2]In fact, this form, which is only used colloquially, is the only form in English that overtly cues the referential status.

do so, we begin with the standard "activated" search strategy (i.e., a bottom-up search starting from ACT), and hypothesize a representation only if this search fails. We acknowledge that there may be cases in which our strategy for handling the referential form of 'this N' may not produce the correct behavior. For example, if one says "This dog I saw was enormous!" while standing in front of a dog, 'This dog' may be incorrectly resolved to the co-present canine.

6. Type Identifiable

   In the case of a linguistic form that *only* cues the Type Identifiable tier (e.g., 'a N'), we immediately hypothesize a representation in the same way as is performed in the previous subsection. Note that the decision to hypothesize a new representation does not necessarily imply that the robot does not yet have a representation for the intended referent. For example, suppose the robot is looking at a box, and its interlocutor says to it remotely, "You should see a box: Bring it to me." In this case, the robot's interlocutor actually intends to refer to a particular box, and the robot in fact already knows of this box. Even in such a case, we still create a new mental representation for a new box. It will be up to subsequent processing stages to recognize the meaning of the sentence, find the two representations, verify that they match, consolidate them into a single representation, and of course, bring the box to the interlocutor.

7. Complex Referring Expressions

   The GH framework also does not specify how to resolve syntactically complex referring expressions, i.e., referring expressions containing multiple referents described in relation to each other, such as those in Example 4:

   (4) *Scene: A table upon which sits a large green block and a large blue block (towards the front of the table), and a greenish-yellow block on a bluish-purple block (in a far corner of the table).*

      a. Pick up the green block that is on the blue block.

      b. Pick up the one on the blue block.

   Chai et al. resolve references of this sort using a a *greedy algorithm* in which locally optimal choices are sequentially made for each subexpression. However, in cases like that seen in Example 4a, this is likely to incorrectly resolve whichever referring expression is considered first, due to the decreased salience, prototypicality, and proximity of the targets. Greedily resolving Example 4b will likely be even less successful due to the underspecification of 'the one'.

We would thus argue that syntactically complex referring expressions should not be considered greedily in a GH-theoretic reference resolution algorithm. How, then, should search plans (i.e., from Table 4.2) for an expression's constituent parts be jointly examined? We decided to handle this problem by "crossing" the search plans for the constituent parts, that is, considering all possible combinations of search plans sorted in search plan order. For example, crossing $ACT \rightarrow FoA \rightarrow HYP$ with $ACT \rightarrow FoA$ yields Table 4.5.

Table 4.3: Sample Joint Search Plan Table

Y	X
ACT	ACT
ACT	FoA
FoA	ACT
FoA	FoA
HYP	ACT
HYP	FoA

The rows of this table are successively examined until a sufficiently probable solution is found or the table is exhausted. Two decisions were made in designing this subroutine. First, while rows are considered in left-to-right order, the action of hypothesization (denoted HYP) is postponed until the search process is successfully terminated; a new representation should only be generated if sufficiently probable referents are found for all other entries in a row, halting the search process. Second, because our implementation's LTM queries (1) automatically include hypothesization when necessary and (2) operate simultaneously on all eligible references (e.g., 'the green box' and 'the blue box'), all [LTM → HYP] steps are executed simultaneously, and *after* all other steps in a given row.

Within-Structure Processes

The GH does not specify how candidates are selected from *within* cognitive structures during reference resolution. Despite what is often assumed (cf. Brown-Schmidt, Byron, & Tanenhaus, 2005), Gundel et al. state that the

GH is *not* a hierarchy of salience or accessibility, and that it is necessary to model salience *independently* of tier of cognitive status (J. K. Gundel, 2010). We will now describe how the proposed model handles degree of salience and uncertainty, and how these measures are used to select candidates.

1. Focus of Attention and Activated Entities

   In order to account for salience without relying on, e.g., a dedicated gestural tier (cf. J. Chai, Prasov, & Qu, 2006), GH-POWER uses a multi-modal salience score to assign a "degree of activation" to entities contained in the FoA and ACT. The entities returned by the *assess* methods associated with the FoA and ACT structures are then the set of all *sufficiently probable* entities within those tiers, ordered by activation such that the most salient candidate will be chosen if multiple are available.

2. Familiar Entities and Long Term Memory

   In the proposed model, the Set of Familiar Entities is equivalent to a "highly salient" LTM cache; we would argue that the "Familiar" and "Uniquely Identifiable" tiers can be viewed as different means of accessing the same structures, with different worst-case conditions. This is consistent with J. K. Gundel (2010)'s that:

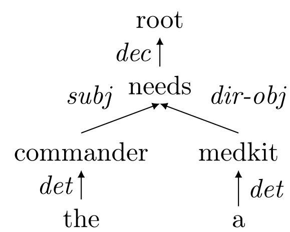
   > '[F]orms that encode cognitive status on the GH are not markers of *degree* of accessibility. Rather, they provide procedural information about *manner* of accessibility, how and where to mentally access an appropriate representation.'

   The entities returned by the *assess* method associated with the FAM are its *sufficiently probable* entities, ordered in *reverse chronological order*; the entities returned by the *assess* method associated with LTM are its *sufficiently probable* entities, ordered in decreasing order of *likelihood.*

4.3.3 Algorithmic Description

In this section, we will describe the GH-POWER algorithm from a computational perspective, discussing the realities of its implementation and integration into our robot architecture. We will first discuss how utterances are parsed and analyzed, and then describe the data structures we use and how they are updated. Finally, we will describe how those data structures are

Figure 4.4: Example Parser Output



used to resolve references in parsed utterances. All capabilities described in these sections are performed by components of the Distributed, Integrated, Affect, Reflection and Cognition (*DIARC*) architecture (Scheutz, Schermerhorn, Kramer, & Anderson, 2007), as implemented in the Agent Development Environment (ADE) (Scheutz, 2006; Scheutz et al., 2013).

Parsing

Each utterance is first sent to the C&C parser (S. Clark & Curran, 2007), which uses the Combinatory Categorical Grammar formalism (Steedman, 2000) to generate a dependency graph. That graph is then converted into a tree such as that seen in Figure 4.4, which shows the tree produced for "The commander needs a medkit."

From the structure of this tree one may extract: (1) a set of formulae representing the surface semantics of the utterance, (2) a set of "status cue" mappings for each referenced entity, and (3) the *type* of utterance which was heard. From the tree shown in Figure 4.4, for example, one would extract:

1. The set of formulae $\{needs(X,Y) \wedge commander(X) \wedge medkit(Y)\}$.

2. The set of status cue mappings $\{X \rightarrow \text{uniquely id'able}, Y \rightarrow \text{type id'able}\}$.

3. The utterance type "STATEMENT" (indicated by the label "dec" on the arc pointing to the root node).

Data Structure Population

We now describe how the GH's data structures (i.e., FoA, ACT, FAM, LTM) are populated, as summarized in Table 4.4. Lines marked with a star denote information which is not yet included in each data structure, representing future work.

Table 4.4: Contents of Relevant Data Structures

Level	Contents
FoA	Main clause subject of clause $n-1$ Syntactic focus of clause $n-1$ * Event denoted by clause $n-1$
ACT	* Entities visible in int.'s region of attention All other entities referenced in clause $n-1$ * Focus of int.'s gesture, if any * Focus of int.'s sustained eye gaze, if any * Speech act associated with clause $n-1$ * All propositions entailed by clause $n-1$
FAM	All entities referenced in clause $n-1$ * The robot's current location
LTM	All declarative memory (including contents of clause n)

Starred items are the subject of current or future work.

Before clause n of some natural language utterance is processed, the contents of *FoA* and *ACT* are reset (*FAM* is reset after each dialogue, and *LTM* is never reset). *FoA*, *ACT* and *FAM* are then updated using the rules listed in Table 4.4. Linguistically, this updating process entails placing the main clause subject, syntactic focus, and event denoted by clause $n-1$ into *FoA* (each of which may be extracted from the syntactic representation of clause $n-1$), placing the speech act and any propositions entailed by clause $n-1$ into *ACT*, and placing all entities referenced at all in clause $n-1$ into both *ACT* and *FAM*. In addition, each location visited by the robot and its interlocutor should be placed into *FAM*, and any entities within the interlocutor's region of attention should be placed into *ACT*.

Each data structure is then sorted according to a "relevance score" or "salience score" R-score. Although the ideal scoring function would account for a variety of extra-linguistic factors, in this work we use the function $R(e) = \alpha_1 \cdot m(e) + \alpha_2 \cdot s(e) + \alpha_3 \cdot r(e)$ where $m(e) \in [0, 1]$ represents whether e is in a main clause, $s(e) \in [0, 1]$ measures the syntactic prominence of e, $r(e) \in [0, 1]$ measures the recency of mention of e, and $\alpha_1, \alpha_2, \alpha_3$ are monotonically decreasing coefficients (i.e., $\alpha_1 > \alpha_2 > \alpha_3$) prioritizing the three measures.

Reference Resolution

We will now describe how GH-POWER is used in practice to perform reference resolution. To resolve the references in a given clause, that clause is first viewed as a graph whose vertices and edges are the variables and formulae used to represent the semantics of that clause[3]. This graph is then partitioned into connected components. For each partition, Alg. 9 (*GH-POWER*) is used to resolve all references found in that partition, producing a set of variable-entity bindings.

GH-POWER takes four parameters: (1) Λ (the semantics of clause n), (2) M (the status cue mappings for clause n), (3) GH (containing FoA, ACT, and FAM), and (4) REX: The aforementioned Referential Executive, which provides access to the distributed LTM framework.

GH-POWER first collects the variables appearing in Λ and sorts them with respect to the tier they are cued towards. For example, if $X \rightarrow$ *in focus* and $Y \rightarrow familiar$ appear in M, then X will appear before Y (Alg. 9 line 2). *GH-POWER* then initiates cache-table C which stores a memoized list of variable-to-entity bindings for each combination of variables in V and tiers in $\{FoA, ACT, FAM, HYP\}$ (line 3).

Before *GH-POWER* begins trying different variable-entity assignments, it must determine in which data structures[4] to look for those entities, determined by the *plan* associated with each level of the hierarchy seen in Table 4.2.

(5) The ball in this red box

To handle multi-variable expressions, *GH-POWER* creates a table Θ, storing all multi-variable plan combinations.

For example, if the referring expression seen in Example 5 is parsed as:
$\{ball(X) \wedge box(Y) \wedge red(Y) \wedge in(X,Y)\}$
with status cue mappings

[3]To properly handle declarative and imperative utterances, we omit the formula associated with the main clause verb from consideration. As later discussed, future work will include consideration of the main clause verb using common-sense reasoning.

[4]As described above in Table 4.4, these data structures will have been recently populated with recently referenced entities, as well as, possibly, other situatedly relevant entities. When the architecture is configured such that salient co-present objects are included in the set of entities used to populate the ACT or FAM data structures, then *GH-POWER* will automatically be able to resolve deictic referring expressions, due to the target referent's elevated status.

Algorithm 6 GH-POWER(Λ, M, GH, REX)

1: S: set of formulae, M: set of status cue mappings, GH: FoA, ACT, and FAM data structures, REX: the Referential Executive
2: $V = [v | v \in \Lambda^V]$ sorted by $M(v)$
3: $C = create\_cache\_table(V, \{\text{FoA,ACT,FAM,HYP}\})$
4: $\Theta = create\_plan\_table(M)$
5: $\Gamma\star = \emptyset$
6: **for all** $\theta \in \Theta$ **do**
7: $\quad \theta_d = [p | p \in \theta, tier(p) = LTM]$
8: $\quad V_\theta =$ new list
9: $\quad$ **for all** $p \in (\theta \setminus \theta_d)$ **do**
10: $\quad\quad (v, t) = (var(p), tier(p))$
11: $\quad\quad$ **if** $C[v, t] == \emptyset$ **then**
12: $\quad\quad\quad$ **if** $(t == HYP)$ **then**
13: $\quad\quad\quad\quad C[v, t] = \{((v \rightarrow \text{'?'}) \rightarrow 1.0)\}$
14: $\quad\quad\quad$ **else**
15: $\quad\quad\quad\quad C[v, t] =$ *GH-ASSESS*(S, v, t, REX)
16: $\quad\quad\quad$ **end if**
17: $\quad\quad$ **end if**
18: $\quad\quad V_\theta = v \cup V_\theta$
19: $\quad\quad \Gamma\star =$ *GH-ASSESS-ALL*$(\Lambda, V_\theta, (\Gamma \star \times C[v, t]), REX)$
20: $\quad\quad$ **if** $\Gamma\star == \emptyset$ **then**
21: $\quad\quad\quad$ BREAK
22: $\quad\quad$ **end if**
23: $\quad$ **end for**
24: $\quad$ **if** $\theta_d ! = \emptyset$ **then**
25: $\quad\quad$ **for all** $\Gamma \in \Gamma\star$ **do**
26: $\quad\quad\quad \Gamma = resolve(REX, bind(\Lambda, \Gamma), order(vars(\theta_d)))$
27: $\quad\quad$ **end for**
28: $\quad$ **end if**
29: $\quad \Gamma\star = [\Gamma | \Gamma \in \Gamma\star, \Gamma^P >= \tau_{resolve}]$
30: $\quad$ **if** $\Gamma\star \neq \emptyset$ **then**
31: $\quad\quad$ BREAK
32: $\quad$ **end if**
33: **end for**
34: **if** $|\Gamma\star| \neq 1$ **then**
35: $\quad$ **return** $\Gamma\star$ // AMBIGUOUS or UNRESOLVABLE
36: **else**
37: $\quad$ **return** $assert(REX, bind(\Lambda, \Gamma \star [0]))$
38: **end if**

$\{X\rightarrow$ *uniquely id'able,* $Y\rightarrow$ *referential*$\}$,

then Table 4.5 of joint search plans will be created, as previously described.

After Θ is created (line 4), an empty set of candidate hypotheses $\Gamma\star$ is created. *GH-POWER* then examines Θ one row at a time until a solution is

Algorithm 7 GH-ASSESS(Λ, V, T, REX)

Λ: set of formulae, V: variable of interest, T: tier of interest, REX: the Referential Executive
$\Lambda_V = [\lambda | \lambda \in \Lambda, \lambda^V = \{V\}]$
$\Gamma\star = \emptyset$
for all $t \in members(T)$ sorted by $R(t)$ **do**
  $\Gamma = \{(V \rightarrow t)\}$
  $\Gamma^P = \prod_{\lambda \in \Lambda_v} assess(REX, bind(\{\lambda\}, (V \rightarrow t)))$
  if $\Gamma^P >= \tau_{assess}$ **then**
    $\Gamma\star = \Gamma \star \cup \Gamma$
  end if
end for
return $\Gamma\star$

Algorithm 8 GH-ASSESS-ALL($\Lambda, V, \tilde{\Gamma}, REX$)

Λ: set of formulae, V: variables of interest, $\tilde{\Gamma}$: set of hypotheses, REX: the Referential Executive
$\Lambda_v = [\lambda | \lambda \in \Lambda, head(V) \in \Lambda^V, [\exists v \in tail(V) | v \in vars(\lambda)]]$
$\Gamma\star = \emptyset$
for all $\Gamma \in \tilde{\Gamma}$ **do**
  $\Gamma^P = \Gamma^P \cdot \prod_{\lambda \in \Lambda_v} assess(REX, bind(\{\lambda\}, \Gamma))$
  if $\Gamma^P >= \tau_{assess}$ **then**
    $\Gamma\star = \Gamma \star \cup \Gamma$
  end if
end for
return $\Gamma\star$

Table 4.5: Sample Joint Search Plan Table

Y	X
ACT	ACT
ACT	FoA
ACT	FAM
ACT	(LTM $\rightarrow$ HYP)
FoA	ACT
FoA	FoA
FoA	FAM
FoA	(LTM $\rightarrow$ HYP)
HYP	ACT
HYP	FoA
HYP	FAM
HYP	(LTM $\rightarrow$ HYP)

found or the end of the table is reached.

For each table entry θ, *GH-POWER* first separates variables for which it must query LTM from all other variables (line 7). It then initializes an empty list V_θ to hold variables that have been examined thus far for entry θ (line 8). Next, it iterates over each (variable, tier) pair in that row, as we now describe.

Consider row one of Table 4.5. *GH-POWER* would first examine the first entry in this row, which says to look for Y's referent in ACT. If C does not already contain hypotheses for $var(p)$ and $tier(p)$ (i.e., Y and ACT), a new one is created: if $tier(p) = HYP$, this hypothesis binds $var(p)$ to "?". Otherwise, *GH-POWER* uses *GH-ASSESS* to search $tier(p)$ for the most likely entity to assign to $var(p)$ (line 15).

GH-ASSESS takes four parameters: (1) Λ (the set of formulae), (2) V (the variable of interest), (3) T (the tier in which to look for possible referents for V), and (4) REX. GH-ASSESS creates, for each entity $t \in T$, a new hypothesis which maps V to t, with probability equal to the product of probabilities of each formula $\lambda \in \Lambda$ which only refers to V (Alg. 7 lines 2-6). For example, if Example 5 is heard and there is one entity in ACT (e.g., obj_{13}), *GH-ASSESS* would consult REX, which uses *DIST-POWER* to see to what degree obj_{13} could be considered to be a box, and to what degree it could be considered to be red, and then create a hypothesis mapping Y to obj_{13} with probability equal to the product of the two probabilities returned by *DIST-POWER*.

Once all formulae containing only $var(p)$ are examined, all those containing both $var(p)$ *and* any other previously examined variables are considered (line 19) using Alg. 8 (*GH-ASSESS-ALL*). For Example 5, this would involve inquiring to what degree the candidate entities for X could be considered to be "in" each candidate entity for Y. After each variable is considered, all candidate bindings whose likelihoods fall below a certain threshold are removed. If this leaves no hypotheses with probability above τ_{assess}, *GH-POWER* breaks out of its loop and considers the next row of the table.

For example, if resolving Y produces hypothesis list
$\{((Y \rightarrow obj_{13}) \rightarrow 0.8), ((Y \rightarrow obj_{12}) \rightarrow 0.75)\}$,
and resolving X produces the hypothesis list
$\{((X \rightarrow obj_5) \rightarrow 0.9)\}$,
these are combined into:

$$\{((Y \rightarrow obj_{13}, X \rightarrow obj_5) \rightarrow 0.72),$$
$$((Y \rightarrow obj_{12}, X \rightarrow obj_5) \rightarrow 0.675)\}.$$

If GH-ASSESS determines that $in(X, Y)$ has probability 0.2 for the first of these hypotheses and 0.9 for the second, the two hypotheses are updated to

$$\{((Y \rightarrow obj_{13}, X \rightarrow obj_5) \rightarrow 0.144),$$
$$((Y \rightarrow obj_{12}, X \rightarrow obj_5) \rightarrow 0.6075)\}.$$

If τ_{assess} is set to 0.6, for example, then the first of these hypotheses would be removed.

GH-POWER now considers all variables which were previously set aside because they were to be searched for in LTM. If any such variables exist, *GH-POWER* considers each candidate binding in H (line 26). For each, S is bound using γ's variable bindings, and an ordering of the variables V_h to be queried in LTM is created based on the prepositional attachment observed in Λ. The bound semantics and variable ordering are then used by the *DIST-POWER* algorithm discussed in Chapter 3.3 to determine (1) whether any of the variables in V_h refer to unknown entities, and (2) which entities in LTM are the most probable referents for each other variable in V_h. The set of hypotheses H is then updated using these results.

Finally, once a solution is found or all table rows are exhausted, the *number* of remaining hypotheses is examined. If more or less than one hypothesis was found, *GH-POWER* returns the set of solutions. This signifies that the referring expression was either ambiguous or unresolvable. If only one hypothesis remains, *GH-POWER* uses that hypothesis' variable bindings to update the set of semantics Λ, and then uses *DIST-POWER* to assert a new representation for each variable bound to "?" (line 37). For example, if resolving Example 5 produces a single hypothesis with probability 0.7 in which X is bound to obj_4 and Y is bound to "?", *DIST-POWER* will create a new object (perhaps with identifier 5) with properties $\{box(obj_5), red(obj_5), in(obj_4, obj_5)\}$ and return $\{((Y \rightarrow obj_5, X \rightarrow obj_4) \rightarrow 0.7)\}$. Once all partitions have been processed in this way, the results are combined into a comprehensive set of candidate binding hypotheses.

4.3.4 Validation and Evaluation

In this section, we verify that the proposed algorithm and GH extensions do indeed improve on previous approaches, and then perform an experimental evaluation on real-world human-human and human-robot dialogues collected by Schreitter et al. (Schreitter & Krenn, 2014). In those dialogues, human *instructors* demonstrated to human or robot *listeners* how to connect two sections of tubing and then affix the tubing to a box.

Validation

We first evaluated several test cases within the previously described experimental context, to demonstrate the success of the proposed approach in addressing our concerns with previous GH-based approaches to reference resolution:. In each case, the algorithm was provided with a knowledge base containing information about the robot's environmental and task context (possibly modified according to that case), and was incrementally fed the relevant utterances for that case.

(1) Previous approaches could not handle *uncertainty*. We confirmed that when the robot believed there was 70% probability that one tube could be referred to as flexible, and 40% probability that the other tube could be referred to as flexible, the algorithm resolved 'The flexible tube' to the first tube.

(2) Previous approaches could not handle *open worlds*. We confirmed that when the robot only knew of red and yellow markers, the algorithm posited a new entity when resolving "Find the blue marker."

(3) Previous approaches could not handle references to hypothetical entities. We confirmed that when the robot knew of a box on a table in front of it and was then asked to resolve the utterances "Imagine a box." and "Describe the box", 'the box' was resolved to the imaginary box and not the observed box.

(4) Previous approaches could not resolve references to unobservable entities. We confirmed that when the robot believed it was learning a task, the algorithm resolved 'the task' in "Describe the task".

(5) Previous approaches have been subject to errors when resolving *complex noun phrases*. We confirmed that when a tube on a triangular table was in "familiar" and a tube on a round table was in "activated", the algorithm successfully resolved 'the tube' in "Pick up the tube that is on the triangular table."

Evaluation

In addition to validating that the proposed algorithm significantly extended the set of cases handled compared to previous algorithms, we evaluated it on the corpus of human-human and human-robot dialogues collected by Schreitter et al. As participants' utterances in that experiment were originally in German, these were first translated to English. As we are not currently attempting to handle disfluencies, these utterances were then "cleaned up", removing disfluencies and parenthetical statements. For example, an utterance with word-for-word translation "So then put you the grasp you here at the marker at the red and yellow one" was "cleaned up" to "So then you grasp here at the red and yellow marker."

A knowledge base containing the relevant properties of the 16 objects and agents involved in the task was constructed and provided to *GH-POWER*. Then, each task-relevant utterance (excepting, e.g., "Hello.") was provided to *GH-POWER* in sequence, and the results of resolution were compared against "gold standard" resolution results provided by human annotators. The human-robot corpus contained 32 task-relevant utterances, the human-human corpus contained 110.

It will also be illustrative to provide the distribution over linguistic forms found across this corpus. Human annotators identified 304 referring expressions across all dialogues. As shown in Figure 4.5, In Focus forms comprised 16.1% of all REs, Activated forms 15.8%, Familiar forms 0.7%, Uniquely Identifiable forms 65.8%, and Type Identifiable forms 1.6%, with Referential forms not being used at all. It is interesting that Familiar and Type Identifiable forms were used with such higher frequency in the Human-Robot dialogues, although the size of the dataset makes it difficult to draw any precise conclusions from this phenomena. In future work it will be necessary to examine this, as well as all results presented in this section, using a larger, more comprehensive dataset[5].

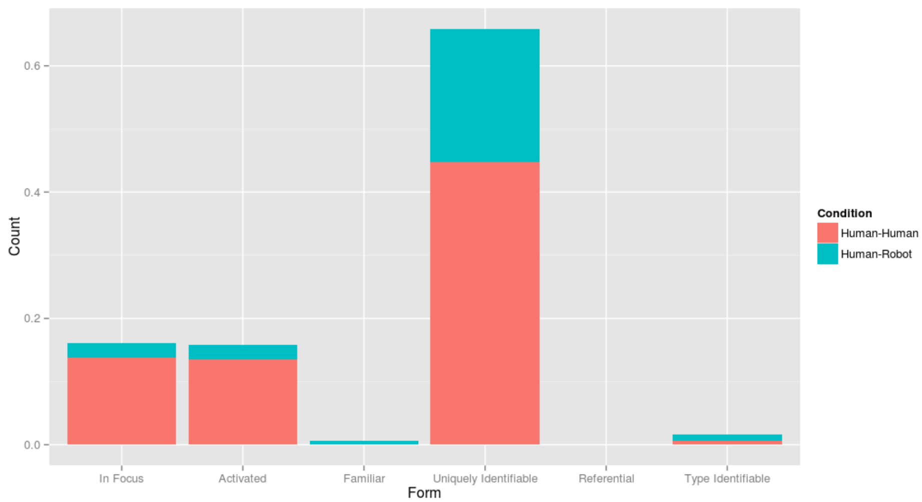
Overall, *GH-POWER* correctly resolved 48 of the 98 (48.98%) references found by the C&C parser in the human-robot dialogues (HRDs), and 121 of the 270 (44.81%) in the human-human dialogues (HHDs), for a net 45.92% accuracy. However, 17.93% of references found by C&C (14.29% in HRDs, 19.26% in HHDs) were not references at all, but artifacts or parse errors. For example, the parser frequently decided that utterances like "Right, so" referred to entities on the right. Discarding these parse errors, *GH-POWER* correctly resolved 55.96% of references (57.14% in HRDs, 55.50% in HHDs). The remaining 44.04% of references could not be resolved due to a variety of reasons, shown in Figure 4.6:

4.97% of references (2.38% in HRDs, 5.96% in HHDs) were *plurals* (e.g. 'the tubes'). *GH-POWER* was unable to resolve these as it is designed to handle *singular* references. Future work will be needed to generate likely groupings of entities to which plurals might be resolved.

10.60% of references (10.71% in HRDs, 10.55% in HHDs) referred to non-

[5]One of the challenges in doing so thus far has been the difficulty of finding a corpus that provides both the ground truth of referring expressions *as well as* information about the uncertainty of relationships between candidate referents within the environment.

Figure 4.5: Distribution of Forms



discrete entities, e.g., regions or sections of tube. Future work will be needed to generate likely *regions* or *portions* of entities to which such references might be resolved.

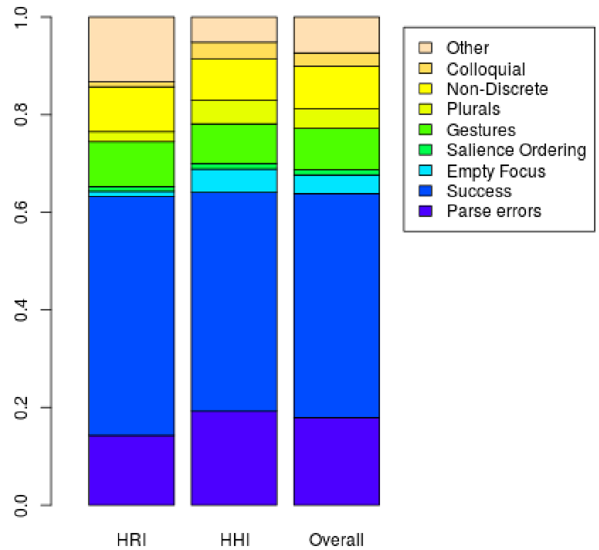
10.26% of references (10.71% in HRDs, 10.09% in HHDs) needed gestural information to be disambiguated; while it is an explicit design aim for *GH-POWER* to handle this facet of multi-modal interaction, we do not yet make use of such information. Future work will be needed to use gesture and eye gaze to correctly bias entities' salience scores.

4.64% of references (1.19% in HRDs, 5.96% in HHDs) were incorrectly resolved due to inconsistencies regarding the "beginning" of the task. For example, participants sometimes started interactions with utterances similar to "I will now describe *it* to you." Because speaker and listener shared a joint context at the start of the task, *the task* may have been in the listener's focus of attention. However, in the evaluation, the system never "heard" the experimenter giving instructions, and thus 'the task' was considered at most activated.

3.31% of references (1.19% in HRDs, 4.13% in HHDs) were idiomatic or colloquial. For example, "That was it" was understood to indicate task completion, but *GH-POWER* was not privy to such information. This suggests that reference resolution may need tighter integration with pragmatic inference.

1.32% of references (1.19% in HRDs, 1.32% in HHDs) were incorrectly

Figure 4.6: Reference Resolution Results



resolved because the linguistic salience score we used did not sufficiently boost the target. Future work will be needed to investigate other salience scoring functions.

The remaining 8.94% of references (15.48% in HRDs, 6.42% in HHDs) were incorrectly resolved for various other reasons. For example, some participants referred to some concepts we were unprepared to handle (e.g., "The problem here..."), and some participants used indefinite noun phrases in ways we did not anticipate (e.g., "There is a pipe there").

4.3.5 Discussion

In the previous section, we demonstrated how GH-POWER was able to resolve the majority of references occurring in a corpus of human-human and human-robot team tasks (as originally presented in (Williams, Acharya, Schreitter, & Scheutz, 2016)). While there were a number of cases that GH-POWER was unable to handle, it *was* able to capture several aspects of the GH missing from previous GH-theoretic approaches. Consider, for example, the following example presented by J. K. Gundel (2010):

(6) a. Alice: I failed my linguistics course.
b. Bob: Can you repeat that?

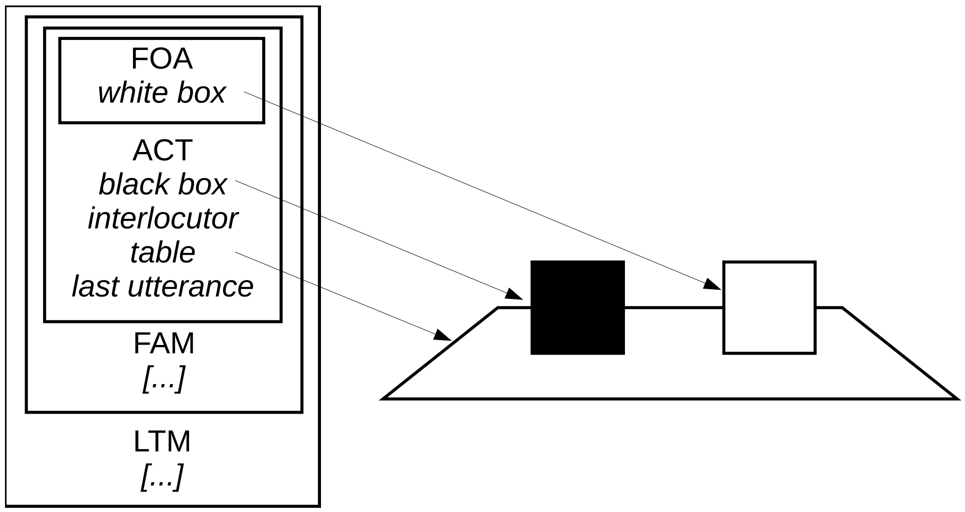
Before resolving 'that'[6], the referent of "my linguistics course" should be in the agent's FoA, while the utterance itself should be in the agent's ACT, but not in the agent's FoA since, as J. K. Gundel, Hedberg, & Zacharski (1993) note, speech acts are activated, but not brought into focus just by being uttered. Gundel et al. suggest that if Bob had meant to refer to the course, he would have used *it* instead of *that*, because 'it' explicitly picks out an in focus referent, whereas 'that' only signals that the referent is activated and therefore could be in focus, and thus the course should be dispreferred to the sentence itself. This effect is captured through GH-POWER's between-structure processes: When 'that' is used, ACT is first checked; and because the utterance is in ACT, it is chosen. FoA is not even examined, because any options residing therein should be dispreferred. However, consider Example 7:

(7) *Scene: A table on which sits a black box and a white box*
a. Bob: Look at the white box
b. Bob: Pick that up

Before resolving 'that', the white box should be in the agent's FoA, and the black box is likely to only be in the agent's ACT, as depicted in Figure 4.7. Following the logic of Example 6, if Bob had meant to refer to the white box, he could have used *'it'* instead of *'that'*, and thus the white box should be *dispreferred*. Yet while 'it' would have been more natural in referring to the white box, choosing the white box as the referent of 'that' is clearly not

[6]While it is beyond the scope of this dissertation, the use of 'that' actually presents much more information than we are making use of in this work (J. Gundel, Hedberg, & Zacharski, 1988). Previous work has shown that 'that' is sometimes used to refer to objects that are either *physically* removed (Xueping & Rong, 2009; Stevens & Zhang, 2013) or *referentially* removed (see also C. Sidner, 1986). Interestingly, some of this work (Stevens & Zhang, 2013) has suggested that this distance only plays a role in selection of 'that' or 'this' in cases of joint eye gaze between speaker and listener, suggesting that Theory-of-Mind may play an important role in this process (see also J. K. Gundel, Ntelitheos, & Kowalsky, 2006), and that it is only the linguistic or semantic notion of distance that plays a role, and not literal visual distance (Kemmerer, 1999). In future work, it would be interesting to try to incorporate different notions of 'thisness' (Poesio & Modjeska, 2002) and 'thatness' into either the relevance assessment process or the property assessment process itself. That is, 'this' and 'that' could be construed as yet-other-constraints, and candidates could be supported or discounted based on the level to which they exhibit an appropriate physical, temporal, or conceptual distance.

Figure 4.7: GH-POWER Contents



Contents of GH-POWER's Cognitive Structures during Hypothetical Algorithm Run. Structures are arranged to depict their hierarchical nature (i.e., an entity in one structure is also in all lower structures). [...] indicates the wide variety of entities contained in the set of familiar entities and in long term memory which are not immediately relevant to this example.

wrong and probably preferred in this situation in the absence of any gesture indicating a shift in attention.

In this scenario, GH-POWER errs for two reasons. First, it treats hierarchical preference as *absolute*, whereas dispreferred entities should be just that: dispreferred, not removed from consideration. Second, GH-POWER does not take *conversational relevance* into account. These factors were initially overlooked because the GH does not specify how relevance factors influence search. GH-POWER checks whether resolution candidates are suitable, i.e., whether they satisfy all described properties, and only moves on to consider entities in another cognitive structure if no resolution candidates in the current structure are deemed *suitable*. In this case, however, this is insufficient. In order for GH-POWER to perform correctly in this scenario, it should recognize not only that both boxes are *suitable*, but that the white box is conversationally more relevant than the black box; there is no clear reason why the agent would be asked to look at the white box and then pick up the black box.

In order to address this issue, GH-POWER should operate in the following way: When ACT is examined, the low conversational relevance of the black box should be noted. This should result in search extending to the FoA while retaining the black box as a resolution candidate. The white box should then be selected using an equation that takes relevance, suitability, and other factors into account.

To be precise, at least three factors must be considered in the within-structure processes of future versions of GH-POWER: (1) suitability (i.e., the agent's certainty that a candidate holds all described properties), (2) relevance, (i.e., the agent's certainty that reference to a candidate would not violate, e.g., Grice's Maxim of Relevance (Grice, 1970)) and (3) common-sense judgments (here, e.g., the agent's certainty that a candidate *can be picked up*). Note that each of these factors may be used differently: while a candidate must score highly on all three factors for the search to cease, only low suitability will likely result in a candidate's complete removal from consideration. Furthermore, to respect the Theory-of-Mind considerations of the GH, this process must consider the extent to which the speaker would have been cognizant of each of these factors. In the next section, we present a new algorithm that takes these factors into account.

4.4 GROWLER

In this section we present *Givenness- and Relevance-theoretic Open WorLd Entity Resolution*, or GROWLER: a new reference resolution algorithm that seeks to address the concerns listed in the previous section. GROWLER uses the same memory model as GH-POWER, but uses different between-structure processes.

Here, the following new notation is used:

Λ_v	The list of logical formulae *that use* variable v.
Θ_v	The list of mnemonic actions *associated* with variable v.
C_v	The list of candidate referents *associated* with variable v.

While GH-POWER proceeded through a list of variable-tier combinations until a sufficiently probable solution was found, GROWLER takes a different

Algorithm 9 GROWLER (Λ, Ξ)

Λ: set of formulae, Ξ: set of status cue mappings
$\Theta = create\_plan\_maps(\Xi)$
$C = create\_candidate\_maps(\Lambda^V)$
for all $v \in \Lambda^V$ **do**
    while $(\nexists c \in C_v \mid R(c) >= \bar{R}) \wedge (\Theta_v \setminus LTM \neq \emptyset)$ **do**
        $grow(\Lambda_v, C_v, \Theta_v)$
    end while
end for
$\Gamma = BuildAndAssessTable(C)$
$R = [v \in \Lambda^V | CanAndShouldExpand(v, \Gamma)]$
while $R \neq \emptyset$ **do**
    for all $v \in R$ **do**
        $C' = (C \setminus C_v) \cup grow(\Lambda_v, C_v, \Theta_v)$
        $\Gamma = \Gamma \cup BuildAndAssessTable(C')$
        $C_v = C_v \cup C'_v$
    end for
    $R = [v \in \Lambda^V | CanAndShouldExpand(v, \Gamma)]$
end while
$\Gamma\star = ASSESS\text{-}LTM(\Lambda, Q)$
if $|\Gamma\star| \neq 1$ **then**
    return $relevantPrefix(\Gamma)$
else
    return $assert(bind(\Lambda, \Gamma \star [0]))$
end if

Algorithm 10 grow(Λ,C,Θ)

for all $e \in domain(head(\Theta))$ **do**
    $P(e) = ASSESS(e, \Lambda)$
    if $P(e) > \bar{P}$ **then**
        $C = C \cup \langle e, P(e), R(e) \rangle$
    end if
end for
$pop(\Theta)$
return C

approach, as shown in Algorithm 9[7]. GROWLER takes two arguments: (1)

[7]In the presented pseudocode we abstract away many technical details for clarity and brevity, including optimizations that can be made through dynamic programming techniques.

a set of logical formulae Λ encoding the surface semantics of an utterance (excepting the predicate associated with the verb), and (2) a set of "status cue mappings" Ξ associating each variable found in Λ to its presumed cognitive status.

Given these arguments, *GROWLER* begins by associating with each variable v: (1) a sequence Θ_v of *mnemonic actions* (i.e., data structure searches and hypothesizations) to perform (Line 2) when searching for the referent to bind to that variable, and (2) an (initially empty) list C_v of *candidate referents* for that variable (Line 3).

GROWLER then finds for each variable v a set of candidate referents that sufficiently satisfy the unary predicates in Λ that involve v (Lines 4–8). For each variable v, GROWLER successively considers mnemonic actions in Θ_v. Each action may yield new candidates that have varying levels of satisfaction probability (as assessed by the *grow* algorithm (Algorithm 10, which makes use of the same *ASSESS* subroutine used by *GH-POWER* (Algorithm 7))) and conversational relevance (assessed using a function $R(x)$, such as that seen in Equation 4.3.3). While all sufficiently *probable* candidates are added to C_v, regardless of relevance, new mnemonic actions are only considered while until a sufficiently *relevant* candidate (as assessed with respect to threshold $\bar{R}$) is found (or until there are no such actions left to consider beyond LTM queries, which are saved until the end of the resolution process). Another way to view this is to say that during this process, all insufficiently *probable* candidates are *removed from consideration*, while insufficiently *relevant* candidates are not removed from consideration, but do not suffice to stop the search process.

Now, *GROWLER* must consider polyadic predicates. To do so, it begins by building a table of hypotheses Γ containing *combinations* of variable assignments (similar to the similarly-named structures used by *DIST-POWER* and *GH-POWER*), and makes use of the same *ASSESS-ALL* subroutine used by *GH-POWER* (Algorithm 8) in order to weed out hypotheses that are no longer sufficiently probable when polyadic predicates are considered.

Because this process may result in all sufficiently relevant referential candidates being eliminated, *GROWLER* now goes through a second cycle of potential expansion-by-mnemonic-action. *GROWLER* goes through this cycle so long as it is determined that it *can and should* perform such expansion for some set of variables V. A variable v *can* be expanded if Θ_v is nonempty (excepting LTM queries), and *should* be expanded if there does not already exist a hypothesis in Γ that binds v to a sufficiently relevant candidate referent.

During each expansion cycle, *GROWLER* performs the following actions for each such variable v. First, *GROWLER* creates a copy of C (i.e., C') that differs from C in that the candidate referents associated with v are replaced by the new candidate referents discovered through a round of expansion effected by the *grow* subroutine (i.e., C'_v). *GROWLER* then creates a new hypothesis table using C', and adds all new sufficiently probable hypotheses found in this table to Γ. Finally, *GROWLER* adds all new sufficiently probable bindings for v (i.e., C'_v) to C_v.

Now, *GROWLER* must deal with *LTM* queries that were previously set aside, using the *ASSESS-LTM* subroutine. For any variables that still *should* be expanded, and that *can* be expanded if LTM-querying is viewed as an acceptable mnemonic action, *ASSESS-LTM* uses *DIST-POWER* to effect such *LTM* queries and update the results stored in $\Gamma\star$.

Finally, once the final set of hypotheses is found, the *number* of such hypotheses is examined. If more or less than one hypothesis was found, *GROWLER* returns the *most relevant* subset of hypotheses. This may be effected either by returning the top n most relevant hypotheses, all hypotheses, the set of hypotheses that are considered sufficiently relevant when viewed with respect to the most relevant of the final set of hypotheses, or, in the cases that $\Gamma\star$ is empty, an empty set. If exactly only one hypothesis remains, however, *GROWLER* uses that hypothesis' variable bindings to update the set of semantics Λ, and then uses *DIST-POWER* to assert a new representation for each variable bound to "?".

For each candidate remaining in the set of hypotheses returned at the end of this process, C_v will contain an entry $\langle ID, P, R\rangle$, where ID is a unique identifier representing a memory trace allowing access to an entity in LTM, P is the probability that entity ID satisfies unary predicates involving v, and R is the *relevance* of entity ID: a combined measure of its visual, linguistic, and conversational salience.

4.5 General Discussion

In Chapter 3, I introduced the problem of reference resolution, and presented the *DIST-POWER* algorithm for resolving references found in definite noun phrases, an an architectural framework that facilitates the use of that algorithm. In this chapter, I showed how this architectural framework can be expanded using Givenness Hierarchy theoretic principles, in order to facilitate a new reference resolution algorithm (*GH-POWER*) capable of resolving a wider class of referring expressions, including indefinite noun phrases, and anaphoric and deictic expressions.

I then identified shortcomings of this approach, and how they might be addressed by a modified algorithm, *GROWLER*. But while we have informally verified that *GROWLER* performs correctly in a handful of test cases, we have not yet performed a systematic evaluation of the form seen in Section 4.3.4. Future work must involve just such an evaluation, as well as parameter tuning to choose appropriate values for $\bar{P}$ and $\bar{R}$ and determine how best to assess the relevance of possible candidate referents. For now, however, we will turn away from referring expression *understanding*, and show how its inverse, the problem of referring expression *generation*, can also be handled by the presented reference-processing architecture.

Chapter 5

Referring Expression Generation

For situated agents to effectively engage in natural-language interactions with humans, they must not only be able to *understand* referring expressions. They must be able to *generate* them as well. For natural language enabled robots, for example, this capability is crucial for clarification request generation: imagine a robotic wheelchair that is asked by its user "Could you bring me to the kitchen?" in a building containing multiple kitchens. In order for the wheelchair to carry out the command, it must determine which kitchen is intended by its user, and thus may decide to *ask* which of these kitchens is correct. Such a robot may be most successful in generating an effective clarification request if it can sufficiently describe the various options between which it is arbitrating, e.g., "Do you mean *the large kitchen on the first floor* or *the small kitchen on the second floor*?"

This task, known as *referring expression generation (REG)* is typically split into two sub-tasks: *content determination*, in which the agent decides which properties to use to describe the target referent, and *linguistic realization*, in which lexical items (i.e., words) are chosen to communicate those properties (Krahmer & Van Deemter, 2012). In keeping with traditional nomenclature, we will refer to algorithms that solve the content determination stage of the REG problem as REG algorithms.

Traditionally, the task of *content determination* is carried out by an algorithm that uses a domain (comprised of target referent m and a set of distractors X), where each entity in that domain is represented by an *attribute set* of properties and relations that hold for that entity (Dale & Reiter, 1995). The most traditionally successful algorithm of this form has been Dale and

Reiter's *Incremental Algorithm (IA)*, which additionally takes a *preference ordering* P in which attributes are to be considered.

A variety of factors prevent many situated agents from using algorithms from this tradition. Most crucially, to check whether a given attribute holds for a given entity, *IA* simply checks whether that attribute is a member of that entity's attribute set, which always produces a clear and unambiguous answer. But for many intelligent agents (e.g., robots), it is imperative to represent the *uncertainty* associated with an agent's knowledge. The knowledge bases of these agents may thus be *unable* to definitively state whether or not a given attribute holds for a given entity.

While there have been previous approaches to generating referring expressions (REs) under uncertainty, those algorithms have been explicitly designed to refer to *objects* in visual scenes, and as such are tightly integrated with visual classifiers (Zarrieß & Schlangen, 2016; D. K. Roy, 2002; Meo, McMahan, & Stone, 2014). This is problematic for least two reasons: First, intelligent agents may need to generate REs for a much wider class of entities than those appearing in a visual scene (e.g., agents, locations, ideas, utterances), which may not be possible if an REG algorithm is tightly coupled with *visual* classifiers. Second, due to this tight coupling, the evaluation of these algorithms conflates the performance of the REG algorithms themselves with the performance of the visual classifiers they employ. For these reasons, previous REG algorithms have only been evaluated relative to different versions of themselves, and not to other algorithms or to humans. We believe that it is important to be able to talk separately about the design, efficacy, and integration of REG algorithms and the design, efficacy, and integration of property classifiers used by those algorithms. As we will discuss, we present in this section an REG algorithm that is *not* tightly integrated with specific property classifiers, but is easily extensible to allow for arbitrary property classifiers to be utilized within a general framework.

In addition to these two primary concerns, we raise a third, which is specific to the realities of modern integrated architectures; a concern we previously rose in the context of referring expression *understanding*. In many integrated robot architectures (e.g., *DIARC* and *ROS*), information may be distributed across a number of architectural components, rather than being stored in a single centralized knowledge base, meaning that no central attribute set is necessarily ready and available to use by an REG algorithm. While there has been much research on merging disparate knowledge bases (Lin, 1996; Liberatore & Schaerf, 1998; Konieczny, 2000), this is not always feasible in integrated robot architectures, as their distributed knowledge bases may use "lazy evaluation" (e.g., a mapping component may not have precomputed information about whether the *acrossFrom* relation holds between all sets of places), and may not store information as attribute-value pairs (e.g., a mapping component may represent spatial information using a hybrid metric-topological map).

In this chapter, I address three main research challenges. First and most crucially, I address the need for REG algorithms that take into account the generator's uncertainty regarding entities' attributes, and which are not tied to a particular domain (e.g., visible objects). Second, I address the lack of a rigorous evaluation framework for systematically evaluating such algorithms, in a way that allows REG algorithms in this class to be compared to one another (as well as to humans). Third and finally, I address the need for such algorithms to take into account the realities of the distributed knowledge representation schemes used by modern integrated architectures.

To address these challenges, I present *DIST-PIA* – an *IA*-inspired REG algorithm designed to operate within the referential processing framework presented in Section 3.3. Furthermore, I present a novel two-stage evaluation framework in which human participants first assess the uncertainty that various attributes hold within a domain and to generate novel REs, and then evaluate the effectiveness of REs created from both human- and machine-generated sets of properties within that domain.

I will begin by discussing previous work on REG, particularly REG under uncertainty. Next, I present *DIST-PIA*, a novel REG algorithm facilitated by the previously presented referential processing framework. Next, I present a novel evaluation framework used to evaluate *DIST-PIA*. Finally, I conclude by discussing possible directions for future work.

5.1 Previous Work

"Referring" has been referred to as the "fruit fly" of language due to the amount of study it has attracted (Van Deemter, 2016). And in truth an enormous amount of research has been done on the topic of REG in the past few decades. The bulk of this work has focused on the content determination stage of REG, cast as what Van Deemter refers to as the *classic REG task*. As previously noted, classic REG algorithms (e.g., *Full Brevity*, the *Greedy Algorithm* (Dale, 1989), and the aforementioned *Incremental Algorithm* (Dale & Reiter, 1995)) operate under a number of simplifying assumptions (such as completely certain knowledge on the part of both speaker and listener) that are not tenable in realistic interaction scenarios.

In this section, I will not attempt to survey the full scope of REG algorithms developed in the past few decades (for an excellent primer, I recommend Van Deemter's recent book on the subject (Van Deemter, 2016)), but will instead focus on REG algorithms that have relaxed the constraint of completely certain knowledge.

The first REG algorithm that explicitly sought to relax the assumption of completely certain knowledge may be that of Horacek (Horacek, 2005). Specifically, Horacek presents an algorithm that reasons about the certainty that *a listener* will be able to recognize that a target referent has certain attributes. The algorithm thus attempts to choose an utterance that minimizes recognition failure: an example of *audience design* in which the *listener's* knowledge and capabilities are taken into account. Horacek's algorithm does not, however, take into account the uncertainty of the agent's *own* knowledge. When an agent's own knowledge is uncertain, it is inappropriate to employ audience design without first taking that agent's own uncertainty into account. We thus focus in this work upon handling the agent's own uncertainty; once this is accomplished, audience design mechanisms can be naturally integrated in future work.

In the graph-based approach presented by Sadovnik (Sadovnik, Gallagher, & Chen, 2013), the confidences of computer vision classifiers are used not only to facilitate audience design, but also as a measure of how well a candidate referent (or an *anchor* with respect to which the referent could be described) matches the attribute associated with that classifier. In their approach, which is targeted towards generating expressions referring to people in photographs, if the algorithm cannot generate an RE which it believes is sufficiently likely to disambiguate the target, the algorithm is re-run using the attributes of both the target and one of its neighbors. While this approach relaxes the assumption of completely certain knowledge, it imposes a number of assumptions of its own, as it is specifically tailored to use computer vision techniques to inform probability judgments and pick anchors. Furthermore, by giving equal weight to the attributes of the target and the attributes of possible anchors, the algorithm appears to generate REs that curiously under-describe the target relative to anchors (e.g., "*The person* on the right of a person who is not Asian and has eye glasses and is smiling and has bangs and whose mouth is not closed").

A similar approach is the graph-based approach presented by Fang, Liu, She, & Chai (2013); Fang, Doering, & Chai (2014). In work from 2013, Fang et al. present an approach which more systematically handles attributes by expanding a hypergraph of properties until a hypergraph-matching algorithm determines that the selected properties can be used to disambiguate the target referent. When choosing how to expand this hypergraph, Fang et al. choose the attribute of minimal cost, using cost functions that take into account the uncertainty of the agent's knowledge as well as the preference of that attribute (in the same sense as that used by the Incremental Algorithm).

In work from 2014, Fang et al. extend this approach so that REs can be generated one clause at a time, facilitating *collaborative* human-agent dialogue. While these approaches are a move in the right direction, they operate under similar assumptions to those imposed by Sadovnik: that probabilities come from the confidence values of applied computer vision classifiers, that the target referent is an object in a visual scene (as are all distractors and candidate anchors), and that information about all such entities is stored in a single, centralized data structure. As previously discussed, these assumptions do not hold in most integrated robot architectures, where information about a wide variety of entities is distributed across multiple heterogeneous knowledge bases.

5.2 The Distributed Probabilistic Incremental Algorithm

I will now describe how Referring Expression *Generation* is performed using our architectural framework. For this task, we use a modified version of the *Incremental Algorithm (IA)* (Dale & Reiter, 1995). The *IA* incrementally proceeds through an ordered list of potential properties that could be used to describe a referent. For each such property p, if p is in the list of properties attributed to the target referent, *IA* checks whether p is *not* true of any distractors (initially all other possible referents). If any distractors are ruled out in this way, p is added to the *description*, i.e., the list of properties to communicate, and those ruled-out distractors are removed from the set of distractors. This process iterates until the algorithm has either ruled out all distractors or run out of properties to consider.

When information is uncertain and distributed across multiple DHKBs as it is in our architecture, however, the application of this algorithm is not straightforward: one cannot assume that there will exist a precomputed set of sufficiently probable properties that hold for either the target or its distractors, meaning that instead of a simple set-membership check, one must assess whether each property holds for the target or a given distractor by making a query to the relevant consultant. We thus present *DIST-PIA*, the Distributed, Probabilistic Incremental Algorithm, in which *REX* uses *DIST-CoWER* to do just that.

In this section, we use the same notation as used in Section 3.3, re-presented here within the context of referring expression generation.

C	A set of *consultants* $\{c_0, \ldots, c_{\vert C\vert}\}$.
c^Q	A set of *query templates* $\{c^q_0, \ldots, c^{q_{\vert c^q\vert}}\}$ advertised by consultant c.
M	A robot's *world model* of entities $\{m_0 \ldots m_{\vert M\vert}\}$ found in the domains provided by the robot's various consultants.
Γ	A set of bindings from variables to entities in M.
λ	A semantic constraint which, under a particular variable binding, makes a claim regarding a property held by the entities in that variable binding.

The *DIST-PIA* algorithm takes a target referent m, and a set of consultants C. *DIST-PIA* begins by initializing a new empty description D and a new queue of referents Q for which sub-descriptions must be generated, which starts out containing only m (Algorithm 11, Lines 1- 2). Until this queue is empty, *DIST-PIA* repeatedly does the following: first, *DIST-PIA* pops the first referent m' off of Q (Line 4), and uses *DIST-PIA-HELPER* to generate a sub-description for that referent. Next, any entities mentioned in that subdescription for which subdescriptions have not yet been generated (including m') are pushed onto Q so that they too can be described (Lines 5- 9). When this loop terminates, the completed description D is returned (Line 11).

Algorithm 11 *DIST-PIA*(m, C)

1: D = new Map()
2: Q = new Queue(m)
3: **while** $Q \neq \emptyset$ **do**
4:     m' = pop(Q)
5:     d = *DIST-PIA-HELPER*(m', C)
6:     $D = D \cup \{m \rightarrow d\}$
7:     **for all** $m'' \in d^M \setminus keys(D)$ **do**
8:         $push(Q, m'')$
9:     **end for**
10: **end while**
11: **return** D

The bulk of the REG process, however, is done by *DIST-PIA-HELPER*,

Algorithm 12 *DIST-PIA-HELPER*(m, C)

1: $d = \emptyset$
2: $X = M \setminus m$
3: $P = [\forall \lambda \in c_m^{\Lambda} : (\lambda, \emptyset)]$
4: **while** $X \neq \emptyset$ and $P \neq \emptyset$ **do**
5:     $(\lambda, \Gamma) = pop(P)$
6:     $V = find\_unbound(\lambda, \Gamma)$
7:     **if** $|V| > 1$ **then**
8:         **for all** $\Gamma\prime \in cross\_bindings(\lambda, \Gamma, C)$ **do**
9:             $push(P, (\lambda, \Gamma\prime))$
10:         **end for**
11:     **else if** $apply(c_m, \lambda, \Gamma \cup (v_0 \rightarrow m)) > \tau_{dph}$ **then**
12:         $\bar{X} = [x \in X \mid apply(c_x, \lambda, \Gamma \cup (v_0 \rightarrow x)) > \tau_{dph}]$
13:         **if** $\bar{X} \neq \emptyset$ **then**
14:             $d = d \cup (\lambda, \Gamma \cup (v_0 \rightarrow m))$
15:             $X = X \setminus \bar{X}$
16:         **end if**
17:     **end if**
18: **end while**
19: **return** d

which is responsible for crafting the sub-descriptions that comprise D. Like *DIST-PIA*, *DIST-PIA-HELPER* takes as arguments a target referent m and set of consultants C. *DIST-PIA-HELPER* begins by initializing an empty sub-description d, a set of distractors X, and a stack of properties P to consider (i.e., the unbound formulae c_m^{Λ} advertised by the constraint c_m responsible for target referent m (Algorithm 12, Lines 1- 3)). *DIST-PIA-HELPER* then does the following until either all distractors in X have been eliminated or until there are no properties left in P to consider adding to the description:

First, *DIST-PIA-HELPER* pops from P the first unconsidered property, in the form of a pair (λ, Γ), where λ is a formula and Γ is a partial set of bindings for that formula (Line 5). Next, *DIST-PIA-HELPER* finds all variables V in λ^V that do not have bindings in Γ (Line 6). If there is more than one such variable, *DIST-PIA-HELPER* does not immediately consider whether or not to add the property. Instead, it uses *cross_bindings* to find all possible *partial* bindings from entities known of in C to variables in λ^V (respecting type restrictions). These bindings are *partial* in that each leaves exactly one variable unbound to which m could potentially be bound (once again, respecting type restrictions). New versions of the property, each of which uses one of these possible partial bindings, are then pushed back onto

P (Lines 7- 10).

Otherwise, i.e., if there is exactly one unbound variable in V (i.e., because λ is a unary predicate or because it has gone through the *cross_bindings* process), *DIST-PIA-HELPER* considers whether the property actually applies to m, and if so, whether it rules out any remaining distractors in X. If so, the property is added to sub-description d, after being modified such that its set of bindings Γ includes a binding from the formula's previously unbound variable v_0 to m (Lines 11- 17). Finally, once this loop terminates, the completed sub-description d is returned (Line 19).

5.3 Algorithm Analysis

In this section, I will discuss the guarantees and complexities of *DIST-PIA* and *DIST-PIA-HELPER*. This is meant to paint a general picture of these guarantees and complexities, and is not intended as a formal theoretical analysis, which would be beyond the scope of this dissertation.

DIST-PIA's performance is straightforward. Because new referents from M are only added to Q if they do not yet have associated subdescriptions in D, the algorithm is guaranteed to terminate in no more than $O(M)$ time – the worst case circumstance in which describing the target referent requires describing every known entity. Similarly, the worst case space complexity is $O(M)$ because in the worst case D will contain an entry for each $m \in M$, and because Q will never contain a referent already appearing in M. However, these linear bounds may be deceiving, as they represent the time taken to generate and store a subdescription for each possible referent, which may in fact be quite computationally expensive, depending on the complexity of *DIST-PIA-HELPER*.

In the worst case, all candidate properties provided by the consultant responsible for m will apply both to m and to all distractors in X, in which case all known entities in M will have to be examined for each property that is considered. If only the literal set of properties P needed to be considered, the time complexity of *DIST-PIA-HELPER* would be $O(MP)$. However, because *DIST-PIA-HELPER* expands each *relation* r into a set of $M^{arity(r)-1}$ bound relations (whose cost of evaluation is equivalent to that of properties). Thus, the true worst case is when each property is in fact a relation of arity k, in which case the true time complexity is $O(M^{k-1} \cdot MP) = O(M^k P)$. With respect to space complexity, the consideration of relations also significantly increases complexity. If all predicates in P are properties, the space complexity will be $O(P)$ (in the case that all properties either have yet to be considered or have already been moved into subdescription d). In the worst case, however, all predicates in P will in fact be relations. But because only one relation at a time may be "expanded" into property equivalents, the largest P will ever be will be immediately after examining the first predicate if it happened to be a \$k\$-arity relation, in which case the size of P will be $|P| - 1 + M^{k-1}$. This space complexity is also contributed to by the size of the subdescription d. In the worst case, it would seem that d could contain $O(P \cdot M^{k-1})$ properties. However, this added complexity is mitigated due to the condition that a property is only ever included in the description if it rules out at least one distractor. This means that there could only ever be at most $O(M)$ elements in d, one for each distractor. Thus the true space complexity is at worst $O(P + M^{k-1} + M)$.

Analysis of *DIST-PIA-HELPER* may be facilitated by discussion of the $find\_unbound$ subroutine. $find\_unbound$ is guaranteed to return a nonempty list. First, every consultant is required to advertise a set of query templates, each of which by definition is represented as a predicate containing *at least* one typed, unbound variable. Line 3 will thus populate P with a nonempty set of pairs whose first argument is a predicate with at least one typed, unbound variable, and whose second argument is an empty binding list. If $find\_unbound$ examines a pair for which this second argument is still empty, it must find at least one unbound variable, since the predicate that is the pair's first argument must have at least one unbound variable, and because there are no bindings in the pair's second argument that override this unbound variable. The only circumstance in which $find\_unbound$ will examine a pair for which this second argument is *not* empty is if the pair was pushed onto P on Line 9. In this case, however, this second argument is nonempty because *all but one* previously unbound variable was newly bound on Line 8, as described above. The fact that all but one previously unbound variable was newly bound means that, by definition, there must still be at least (and in fact exactly) one unbound variable for $find\_unbound$ to find. Thus, it is ensured that $find\_unbound$ will return a nonempty list.

It is important to note, however, that *DIST-PIA-HELPER* will only find a successfully discriminating subdescription *if one exists*. *DIST-PIA-HELPER* continues to add properties to d so long as there are both properties left to consider and distractors left to eliminate. This means that if the properties to be considered run out before all distractors are eliminated, then the description returned by *DIST-PIA-HELPER* will not uniquely describe the target. This will only happen, however, when the target *cannot* be uniquely described using a conjunction of the properties provided by the target's consultant.

5.4 Algorithm Walkthrough

Imagine a robot with three consultants: *ppl*, *locs*, and *objs*, that is instructed to construct a set of properties referring to entity ppl_5. *DIST-PIA* will begin by creating empty description $D = \emptyset$ and referent queue $Q = \{ppl_5\}$. Next, *DIST-PIA* will pop ppl_5 off of Q, and because a description for ppl_5 does not appear in D, will call *DIST-PIA-HELPER*$(ppl_5, \{ppl, locs, objs\})$. *DIST-PIA-HELPER* will first initialize sub-descriptor $d = \emptyset$, and set of distractors $X = \{ppl_1, ppl_2, ppl_3, ppl_4\}$, assuming for simplicity that *ppl* only knows of five people. Next, let's suppose that *ppl* advertises the following properties, which may be all the properties it is able to handle, or may only be the properties which it knows currently hold for *some* entity it knows about:
$\{jim(X - ppl),$
$jill(X - ppl),$
$man(X - ppl),$
$woman(X - ppl),$
$lives\text{-}in(X - ppl, Y - locs)\}.$

First, *DIST-PIA-HELPER* will consider $jim(X - ppl)$ (with empty set of variable bindings Γ). This predicate has exactly one unbound variable. *DIST-PIA-HELPER* will thus use *DIST-POWER's apply* method to ask how probable it is that ppl_5 has property $jim(ppl_5)$. Suppose the returned probability is above some threshold, say 60%. *DIST-PIA-HELPER* will thus determine if this property also weeds out distractors. For each referent ppl_x in X, *DIST-PIA-HELPER* will use *DIST-POWER's apply* method to ask how probable it is that ppl_x has property $jim(ppl_x)$. Suppose that it is only sufficiently probable that ppl_2 has this property. The set of eliminated distractor $\bar{X}$ will thus equal $\{ppl_1, ppl_3, ppl_4\}$. Because this is nonempty, $jim(ppl_5)$ will be added to subdescription d and $\{ppl_1, ppl_3, ppl_4\}$ will be removed from X.

DIST-PIA-HELPER will next consider $jill(X - ppl)$ (with empty set of variable bindings Γ). This predicate has exactly one unbound variable. *DIST-PIA-HELPER* will thus use *DIST-POWER's apply* method to ask how probable it is that ppl_5 has property $jill(ppl_5)$. Suppose the returned probability is below 60%. *DIST-PIA-HELPER* will thus move on, to consider $man(X - ppl)$ (with empty set of variable bindings Γ). This predicate has exactly one unbound variable. *DIST-PIA-HELPER* will thus use *DIST-POWER's apply* method to ask how probable it is that ppl_5 has property $man(ppl_5)$. Suppose the returned probability is above 60%. *DIST-PIA-*

HELPER will thus determine if this property also weeds out distractors. For each referent ppl_x in X, *DIST-PIA-HELPER* will use *DIST-POWER's apply* method to ask how probable it is that ppl_x has property $man(ppl_x)$. Suppose that this is sufficiently probable for the lone remaining distractor, ppl_2. Because the set of eliminated distractor $\bar{X}$ is empty, *DIST-PIA-HELPER* will not add this property to subdescription d, but will instead move on.

DIST-PIA-HELPER will next consider $woman(X - ppl)$ (with empty set of variable bindings Γ). This predicate has exactly one unbound variable. *DIST-PIA-HELPER* will thus use *DIST-POWER's apply* method to ask how probable it is that ppl_5 has property $woman(ppl_5)$. Suppose the returned probability is below 60%. *DIST-PIA-HELPER* will thus move on, to consider $lives\text{-}in(X - ppl, Y - locs)$ (with empty set of variable bindings Γ). This predicate has two unbound variables. *DIST-PIA-HELPER* will thus use *cross_bindings* to come up with partial bindings to those variables that leave exactly one unbound *ppl*-associated variable unbound. Because only X is associated with *ppl*, this consists of finding the set of candidate bindings to Y. Suppose *locs* knows of three locations: $\{locs_1, locs_2,$ and $locs_3\}$. *DIST-PIA-HELPER* will thus add the following properties to P:

$(lives\text{-}in(X - ppl, Y - locs), \{Y \rightarrow locs_1\})$,
$(lives\text{-}in(X - ppl, Y - locs), \{Y \rightarrow locs_2\})$, and
$(lives\text{-}in(X - ppl, Y - locs), \{Y \rightarrow locs_3\})$.

DIST-PIA-HELPER will next consider $lives\text{-}in(X - ppl, Y - locs)$ (with variable bindings $\Gamma = \{Y \rightarrow locs_1\}$). This predicate has exactly one unbound variable. *DIST-PIA-HELPER* will thus use *DIST-POWER's apply* method to ask how probable it is that ppl_5 has property $lives\text{-}in(ppl_5, locs_1)$. Suppose the returned probability is above 60%. *DIST-PIA-HELPER* will thus determine if this property also weeds out distractors. For each referent ppl_x in X, *DIST-PIA-HELPER* will use *DIST-POWER's apply* method to ask how probable it is that ppl_x has property $lives\text{-}in(ppl_x, locs_1)$. Suppose that this is not sufficiently probable for the lone remaining distractor, ppl_2. The set of eliminated distractors $\bar{X}$ will thus equal $\{ppl_2\}$. Because this is nonempty, $lives\text{-}in(X - ppl, Y - locs)$ will be added to subdescription d and $\{ppl_2\}$ will be removed from X. Because X is empty, $ppl_5 \rightarrow \{jim(ppl_5), lives\text{-}in(ppl_5, locs_1)\}$ will be returned to *DIST-PIA*.

Now, *DIST-PIA* will add all entities mentioned in this set of properties other than ppl_5 (i.e., $locs_1$) to Q. Next, *DIST-PIA* will pop $locs_1$ off of Q, and because a description for $locs_1$ does not appear in D, will call *DIST-PIA-HELPER*$(loc_1, \{ppl, locs, objs\})$. *DIST-PIA-HELPER* will first initialize sub-descriptor $d = \emptyset$, and set of distractors $X = \{locs_2, locs_3\}$, assuming for simplicity that *locs* only knows of three people. Next, let's suppose that *locs* advertises the following properties, which may be all the properties it is able to handle, or may only be the properties which it knows currently hold for *some* entity it knows about:

$\{somerville(X - locs),$
$cambridge(X - locs),$
$massachusetts(X - locs),$
$in(X - locs, Y - locs)\}.$

First, *DIST-PIA-HELPER* will consider $somerville(X - locs)$ (with empty set of variable bindings Γ). This predicate has exactly one unbound variable. *DIST-PIA-HELPER* will thus use *DIST-POWER's apply* method to ask how probable it is that $locs_1$ has property $somerville(locs_1)$. Suppose the returned probability is above some threshold, say 60%. *DIST-PIA-HELPER* will thus determine if this property also weeds out distractors. For each referent $locs_x$ in X, *DIST-PIA-HELPER* will use *DIST-POWER's apply* method to ask how probable it is that $locs_x$ has property $somerville(locs_x)$. Suppose that it is not sufficiently probable that any of the distractors have this property. The set of eliminated distractor $\bar{X}$ will thus equal $\{locs_2, locs_3\}$. Because this is nonempty, $somerville(locs_1)$ will be added to subdescription d and $\{locs_2, locs_3\}$ will be removed from X. Because X is empty, $locs_1 \rightarrow \{somerville(locs_1)\}$ will be returned to *DIST-PIA*. Because Q is empty, *DIST-PIA* will return:
$\{ppl_5 \rightarrow \{jim(ppl_5), lives\text{-}in(ppl_5, locs_1)\},$
$locs_1 \rightarrow \{somerville(locs_1\},$
with the expectation that natural language generation will use these properties to craft a referring expression along the lines of "Jim, who lives in Somerville".

5.5 Evaluation

REG algorithms generate sets of attributes to use to describe target referents, in particular contexts, under particular assumptions. The goal of evaluating an REG algorithm is to see how well the attributes chosen by it align with those used in human-generated REs. Traditional REG evaluation metrics (e.g. *Dice* (Gatt, van der Sluis, & van Deemter, 2007) and *MASI* (Passonneau, 2006)) do so by measuring the distance (e.g., set difference) between machine-generated attribute sets and those used in human-generated REs. Recently, however, this methodology has come under criticism, as the semantic similarity of two attribute sets does not imply similarity between those two sets with respect to *effectiveness*, that is, how well each allows a target referent to be picked out by a hearer, which is presumably the purpose of an REG algorithm in the first place (Van Deemter & Gatt, 2009). Recently, there has thus been a shift towards *task-based* evaluations (e.g., (Byron et al., 2009; Koller et al., 2010; Viethen & Dale, 2006)), in which algorithms are compared by how well they allow some task to be achieved.

The previously discussed uncertainty-handling REG algorithms have mainly used task-based evaluations in which an image provided to participants is also provided directly to the REG algorithm. However, this necessarily conflates the evaluation of the REG algorithm with the evaluation of the visual classifiers used to process that image. Furthermore, it does not allow the REG algorithm to be directly compared to either other REG algorithms (unless they use identical classifiers) or to humans (who certainly do not use identical classifiers). It is thus imperative to develop a *new* evaluation framework that allows an REG algorithm to receive information about how uncertain a human would be regarding the attributes of various entities in an environment, without having to visually process the scene.

In this section we present an evaluation framework that achieves this goal, and use it to evaluate our algorithm. It is our hope that in the future it will be used as a general framework for evaluating REG algorithms designed for uncertain situated contexts. Our evaluation framework is comprised of two stages. In the first stage, participants are shown an environment, and are asked to provide (1) an RE referring to a particular entity in the environment, and (2) probability judgments that particular attributes hold for particular entities in in the environment. The probability judgments can be used to train REG algorithms to assess whether various attributes hold without committing to a particular domain (i.e., that for which visual classifiers are needed), and in a way that should allow machine-generated and human-generated REs to be directly comparable. In the second stage, participants are shown the same environments shown in the first stage, along with either "human-driven REs" (REs created using the properties used in human-generated REs) or "machine-driven REs" (REs created using the properties chosen by human-data-trained REG algorithms), and are asked to select the entity in the environment that matches the RE. This framework thus allows REG algorithms to be compared to both other algorithms as well as to humans in uncertain situated contexts. In the following sections we describe how we employ this framework to evaluate *DIST-PIA*.

5.5.1 Stage One

In the first stage of the evaluation, participants were each shown three images of rooms (positioned in a random order), where one image contained a red bounding box surrounding an object within that room. The three rooms depicted a kitchen, an office, and a large empty white room; each contained three objects which were candidates for targeting. In each scene, one such object was an object that only appeared in that scene; a second was an object for which an identical object appeared in a different scene; the third was an object for which an object of the same type (but, for example, of a different color) appeared in a different scene. This resulted in five task-relevant objects in each scene. Each image also contained around five salient irrelevant objects as well. And, because each participant was simultaneously shown three scenes, the rooms themselves also serve as anchors with respect to which participants could describe their target referents. These scenes are shown in Figure5.1.

Figure 5.1: Scenes Shown to Participants

In the scene to the left, the possible target referents in the two evaluation stages were the waterbottle, headphones, and mug; in the middle scene, these were the laptop, chair, and notebook; in the right scene, these were the briefcase, book, and marker.

Participants were told to imagine that in a subsequent experiment, another participant would tour the three rooms shown in the pictures, and then receive a description of an object to put a sticker on. Participants were told to write the description that they should receive. Participants were told that the later participant would only walk through the rooms and not see the exact images, and that they would tour the rooms in a random order, and thus could receive the description in any room. This was to prevent participants from referring to the positioning of the three images.

After each participant provided a description, they were asked to evaluate how well the target object matched each of twenty attributes (randomly selected from a total of 52[1]), such as "is blue", "is a marker", and "is in the kitchen" by re-positioning [0-100] sliders that were originally set to 50.

Participants (56 male, 33 female; mean age 35 (sd=11.65)) were recruited through Amazon Mechanical Turk. Each participant was shown a random target object, providing us with an average of 10 REs per target object, as well as an average of 3.8 probability judgments for each attribute for each object. Gold standard semantic parses (i.e., sets of logical formulae representing properties and relations) were then crafted for each such human-generated RE.

Next, two *DIST-POWER* consultants were created, one for objects, and one for locations, which were provided with, respectively, a subset of the objects and locations found in the three scenes. Specifically, the location consultant was provided with information regarding the three rooms, while the objects consultant was provided with information regarding each object referenced by pilot participants, as well as each possible distractor (i.e., objects of the same type as an object referenced by a pilot participant).

When these consultants are asked for the probability that an entity has a particular attribute, they return the mean probability judgment provided by participants, so long as that value is above a threshold of 0.1. For example, if in this first experiment, some subset of participants were asked how strongly they would agree that a certain object was a chair, the mean rating was stored as the probability that that object had that property. While we collected data on target objects, we did not collect data on distractors: for distractors, we gave the consultants information regarding the properties they shared with the target objects (using the same probability values), as well as certain knowledge of the dimensions on which they differed (e.g., color). In addition, the consultants were provided with certain knowledge that each room had the property $room(X)$, that each object had the property $object(X)$, and information about what room each object was located in, if and only if that information had not been specified by participants. Finally, each *DIST-POWER* consultant was provided with a preference ordering over properties. While this ordering was hand constructed, we would eventually like to learn similar orderings from data.

The *DIST-PIA* algorithm was then used to generate attribute sets for each of the nine target referents, as shown in Table 5.1. We then combined these with the attribute sets derived from human utterances, and removed duplicates in order to yield an average of 9.56 (sd=3.13) unique sets of at-

[1]Before Stage One, we informally collected object descriptions from colleagues. The attributes used here were the entire set of attributes used in those descriptions.

tributes per target object. For each of these attribute sets, we then crafted one RE, using a consistent template to generate all REs. This conversion from REs to logical form and back allows us to control for phrasing so that all utterances were consistently worded. Because one RE was produced for each unique set of properties, this thus resulted in an average of 9.56 (sd=3.13) REs per target object.

5.5.2 Stage Two

In the second stage, a new set of participants were shown the same images, but without bounding boxes, and were told to imagine that previous participants had walked through the pictured scenes, and had written descriptions of object that they were to click on. For each of the nine target referents, participants were shown a randomly selected human- or machine-driven RE for that referent, and asked to click on the described object. After each image, participants were notified as to whether they had clicked on the correct object.

Participants were recruited through Amazon Mechanical Turk (62 male, 46 female; mean age 35.07 (sd=10.14)) [2]. Each of the 85 unique REs was thus shown to an average of 11.44 participants. Recall, however, that these utterances were crafted based on either property sets chosen by *DIST-PIA* or based on property sets extracted from the utterances collected from participants in Stage One. Because some of these property sets were identical, each of the unique REs in this section really corresponds to a *cluster* of semantically identical human- or machine-driven property sets. For each cluster, we computed the accuracy of that cluster's chosen properties in allowing the true target referent to be picked out by second-stage participants. This allowed us, for each target referent, to compute an accuracy ranking over clusters, in turn allowing us to calculate an *accuracy percentile* for *DIST-PIA*.

5.5.3 Results and Discussion

Overall, *DIST-PIA* allowed successful identification in 91.37% of cases; on average, it achieved the 45.67th accuracy percentile (sd=23.94) among the

[2]While filters were set up intended to prevent participants from Stage One to participate in Stage Two, these did not work as intended, and this set of 108 participants included five from Stage One, a fact that was only caught months later. While unfortunate, we do not believe this set of repeated participants was large enough to prompt serious concern, especially given the fact that utterances were standardized with respect to noun phrasing between Stages One and Two, and given the low likelihood of any given participant in experiment Two being shown their own utterance phrasing from Stage One.

Table 5.1: Properties Chosen by *DIST-PIA*

ID	Properties	Translation	Acc.	Rank
1	notebook(X)	The notebook	80%	11th of 11
2	Dell(X),laptop(X),blue(Y), chair(Y),in-front-of(Y,X)	The Dell laptop that the blue chair is in front of.	85.7%	5th of 10 (Tie)
3	blue(X),chair(X)	The blue chair.	97.8%	2nd of 3
4	laptop-bag(X)	The laptop-bag.	100%	1st of 5
5	textbook(X),laptop-bag(Y), behind(Y,X)	The textbook that the laptop-bag is behind.	63.6%	5th of 10
6	red(X),whiteboard-marker(X)	The red whiteboard-marker.	86.7%	6th of 10
7	headphones(X)	The headphones.	100%	1st of 13 (Tie)
8	shaker-bottle(X), headphones(Y),next-to(X,Y)	The shaker-bottle next to the headphones.	85.7%	3rd of 13 (Tie)
9	coffee-mug(X)	The coffee-mug.	100%	1st of 10 (Tie)

Properties chosen by *DIST-PIA* to refer to each target object, with accompanying NL translation. "Accuracy" denotes the percent of participants who clicked on the correct object when provided with the machine-driven RE. "Rank" compares this percentage with human-driven REs. For example, when "The notebook" was used, 80% of participants clicked on the correct object, but all other REs for that object yielded higher accuracy rate. In contrast, "The textbook that the laptop-bag is behind" had only a 63.6% accuracy rate, but this was a higher accuracy rate than was achieved by all but four of the unique human-driven REs for that object. "Tie" indicates that a machine-driven REs had the same accuracy rate as at least one human-driven RE. For example, in row 9, "the coffee-mug" had a 100% success rate, but so did the human-driven, "the white mug near the headphones on the table on which are the green marker and the blue-topped cup."

competing RE for a given object. That is, on average, the REs crafted based on the properties chosen by *DIST-PIA* were as or more successful than those crafted based on the properties chosen by 45.67% of human participants. This demonstrates that *DIST-PIA* not only allowed for successful disambiguation of the target object, but the attributes chosen by *DIST-PIA* allowed nearly the same degree of accuracy and efficiency as the attributes used in the average human-driven RE. Note that the 50th percentile indicates *super-human* performance – any algorithm achieving in at least the 50th percentile when adjudicated relative to humans alone would be performing, on average, better than humans at referring expression generation. Thus, while not super-human in nature, we would view anything above the 40th percentile as generally successful.

DIST-PIA's performance would further improve if given more sophisticated consultants. In order to fairly evaluate *DIST-PIA*, we provided it with consultants that only made judgments based on the attributes used in our pilot study. Because pilot participants referred to the blue chair as being in front of the laptop, but did not refer to the laptop as being behind the blue chair, we did not collect data on the extent to which people believed the laptop to be behind the blue chair. Because the consultants were not given any extra knowledge (e.g., as to the symmetry of in-front-of and behind), the algorithm produced properties in some cases that were less natural than they could have been, as seen in Table 5.1.

Furthermore, *DIST-PIA* had to handle occasionally nonsensical probability judgments, likely produced by participants who did not feel compelled to take the task seriously (a problem common to crowdsourcing experiments). For example, participants asked about the black laptop bag gave an average confidence rating of 35 out of 100 that that object was actually a dry-erase marker; when asked about the green book gave an average confidence rating of 75/100 that it was a chair. Accepting such errors is necessary, however, to prevent the performance of the algorithm from being conflated with the performance of specific classifiers. This raises an important point: because of the generality of *DIST-PIA*, it does not compete with the classifiers used by other algorithms (e.g. Zarrieß & Schlangen, 2016; D. K. Roy, 2002; Meo, McMahan, & Stone, 2014). In the future, it would be interesting to integrate such classifiers into the *DIST-POWER* framework as part of special consultants. This would also allow for direct comparison of disparate classifiers when used with the *DIST-PIA* algorithm.

We must also comment on the visual nature of our evaluation, and how this relates to our previous claim that *DIST-PIA* is domain independent and not restricted to visual information. To be clear, there is no reason to restrict the consultants to those that handle visual information, a fact we tried to account for by including a spatial classifier which handled the properties of the larger rooms containing the objects of interest to the experiment. It is certainly possible to further exploit this domain-independent nature in additional evaluations. We needed participants to be able to unambiguously select referents in an easily understandable, online, static environment. We could have provided participants with information in other modalities: a video recording of a fly-through of the environment in order to show spatial information, audio recordings of conversations to provide dialogue-based information, and so forth. However, we did not want participants to get overwhelmed, and wanted all information that they would need to make their decisions to be instantly and easily accessible and assessable. While location-based data was presented visually to participants, a robot could easily acquire location-based information using laser readings rather than camera data; and abstract topological data and conversational data can of course be represented in a much more readily accessible and assessable means to a robot than it could have to human participants in our experiments.

Finally, *DIST-PIA* did not "nicely overspecify" in some conditions where humans did. For example, *DIST-PIA*'s choice of simply $notebook(X)$ for the first target object achieved 80% success rate, but had the lowest ranking for that object, in part because most humans used descriptions involving $red(X)$, which allows the eye to be drawn away from distractors like the green book. The traditional *IA* captures this effect by placing colors at a high priority. However, unlike the traditional *IA*, we chose to have the object's "type" (e.g., "bottle") and variants thereof (e.g., "waterbottle") be handled as properties just like any other, so that we would not need to specify an additional mandatory *DIST-POWER* capability (i.e., the ability to provide the "type" of a candidate object). This required us to place these type-like properties at the top of the preference orderings, in order to make sure that a type-like property was always used. In this case, when presented the trade-off, we chose generality of our architectural mechanisms over possible performance gain.

5.6 General Discussion

In this chapter, we discussed three main research contributions made through the design and evaluation of the *DIST-PIA* algorithm. First, we have presented an REG algorithm which, unlike previous algorithms, accounts for the generator's uncertainty *and* is domain independent. This algorithm operates within a general reference framework which allows architectural components providing different types of information to be easily integrated together. We believe that it is crucial for researchers to separate the problems of referring expression generation and reference resolution from the task of *property assessment*, both for algorithm development and for evaluation of those algorithms. Our presented approach makes great strides towards this goal; in the future, it would be interesting to take visual classifiers previously presented as tightly-integrated facets of previous referring expression generation and reference resolution frameworks, and integrate them into more general frameworks such as that which was used in this chapter.

Second, we have presented a novel evaluation framework which allows REG algorithms designed for uncertain situated contexts to be evaluated relative to both other algorithms and to humans, without conflating the performance of the *algorithm* with the performance of the *classifiers used by the algorithm*. This evaluation showed that the performance of *DIST-PIA* was comparable to that of humans. And finally, we have taken the realities of modern integrated agent architectures into account by using the *DIST-POWER* framework, which allows information to be distributed across multiple heterogeneous knowledge bases. Yet, *DIST-PIA* still represents only a first step for algorithms within such architectures. In the future, we would like to improve *DIST-PIA* in a variety of ways.

First, *DIST-PIA* should be modified to consider not only whether the probability of a particular referent is above a certain threshold but also its probability relative to those of other referents, and the choice of threshold should be learned from data. Second, we would like to incorporate audience design considerations, similar to Horacek (2005), as well as perspective-taking considerations. Third, we would like to use Givenness-Hierarchy Theoretic mechanisms similar to those seen in Chapter 4, in conjunction with a multi-modal reference model to generate deictic and anaphoric REs. Finally, we would like to modify our approach to use a Dempster-Shafer Theoretic uncertainty representation, in order to better handle ignorance – an approach we have used in the context of *pragmatic* understanding and generation, as we will discuss in the next two chapters.

Chapter 6

Pragmatic Understanding

In Chapters 3-5, I presented algorithms that enable *referential* capabilities (i.e., referring expression understanding and generation) in robots. In the following two chapters, I move inwards and upwards, to discuss the tasks that immediately succeed referring expression understanding and immediately precede referring expression generation; tasks that take place at a higher level of abstraction. Specifically, I will talk about the *pragmatic reasoning* tasks of *pragmatic inference* and *pragmatic generation*.

When seeking to understand an interlocutor's utterance, it is not enough to know *what entities* are being communicated about; it is just as important to discern the general *intentions* which one's interlocutor is trying to communicate. Unfortunately, this problem is made challenging by the fact that humans, at least, rarely communicate their intentions directly, and instead tend to use linguistic forms whose intended meanings must be *inferred*. For example, when a human asks "Could you get me a coffee?", her interlocutor may infer from goal-based, task-based, or other context-based information that she is not really asking a question, but is instead making a request. Such non-literal utterances are known as *indirect speech acts (ISAs)* (Searle, 1975), and are used in order to achieve a variety of socio-cultural goals (e.g., politeness) (Lakoff, 1973). While the use of ISAs differs between individuals and between cultures (Tannen, 1981), their use is generally accepted as a common feature of natural human dialogue. In this chapter, I will discuss the importance of understanding ISAs in human-*robot* dialogue, and present a set of mechanisms for doing so. Throughout this chapter, I will use the following terminology:

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 causing 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 this chapter will proceed as follows. In Section 6.1, I present a treatment of the philosophical foundations of our work (i.e., the theories of meaning in human-*human* dialogue postulated by Grice and Searle). In Section 6.2, I present experimental work justifying the study of indirect speech act understanding in human-*robot* dialogue, and design recommendations suggested by our experimental findings. In Section 6.3, I discuss previous work seeking to enable indirect speech act understanding capabilities in robotics. Finally, in Section 6.4, I present a novel computational approach to indirect speech act understanding.

6.1 Philosophical Motivations

In this section I will present the philosophical motivations behind our work. I will start by discussing Grice's theory of natural and non-natural meaning. I will then discuss Searle's response to this theory, as found in his theory of speech acts. Finally, I will discuss how this theory is extended through Searle's theory of *indirect* speech acts.

6.1.1 Grice's Theory of Natural and Non-Natural Meaning

In his 1957 article entitled *Meaning* (Grice, 1957), Paul Grice provides an account (which is later expanded upon in Grice (1968, 1969)) of two notions of meaning: *natural* and *non-natural* meaning. The *natural* meaning M of some sign Θ is the set of facts, statements, or belief revisions *entailed* from such a sign: a binary relation $means_N(\Theta, M)$. In contrast, the *non-natural* meaning M of some sign U (e.g., an utterance) is the set of facts, statements, or belief revisions which the issuer S of that sign intended to be produced in the mind of hearer H by receiving (e.g., hearing or viewing) that sign *by recognizing this intention on the part of the speaker*: a quaternary relation $means_{NN}(S, H, U, M)$.

That is, Grice suggests that to say that Speaker S meant M by utterance or sign U (directed towards some hearer H) is to say that S intended U to produce M as an *effect* in the mind of H, and that S intends U to produce M by virtue of H recognizing this intention. Grice argues that the non-natural or *speaker meaning* of a sign is determined from the natural or *semantic* meaning through a set of conversational maxims (described here as laid out by Davis (2011):

Maxim of Quality Make your contribution true; so do not convey what you believe false or unjustified.

Maxim of Quantity Be as informative as required.

Maxim of Relation Be relevant.

Maxim of Manner Be perspicuous; so avoid obscurity and ambiguity, and strive for brevity and order.

6.1.2 Searle's Theory of Speech Acts

In his 1969 book *Speech Acts* (Searle, 1969), John Searle presents two main critiques of Grice's account of non-natural meaning. First, Searle argues that under Grice's account, there is no clear link between sentence and meaning, and that a better account of meaning would show how rules can be used to connect what a sentence *means* with *what a speaker means* by that sentence. Specifically, Searle argues that Grice's framework must be reformulated so that the meaning of a sentence is not *randomly* related to its meaning, but is rather derived from the sentence using a set of *rules* arising from *cultural convention*. In this new account, a speaker performing a speech act intends his interlocutor to recognize his intention to produce some effect by that speech act, using a set of shared conventionalized rules which map an utterance form, under a particular context, to a produced effect.

Here, the speech acts performed and recognized are assumed to be *illocutionary acts*. This is a term originating in Austin's three-tier speech act framework (Austin, 1975):

1. A **locutionary act** is the physical performance of an utterance with a particular propositional content.
2. An **illocutionary act** is a speech act with a particular illocutionary point. Asserting, suggesting, demanding, promising, and vowing are all illocutionary acts.
3. A **perlocutionary act** is a speech act which actually produces an effect. Persuading, scaring, and inspiring are all perlocutionary acts.

This leads to Searle's second, primary (see also Searle, 2007) criticism of Grice's account: Searle argues that because Grice defines meaning in terms of intended effects, his account confuses illocutionary with perlocutionary acts. Searle uses an example of the following form as evidence of this point:

Suppose you are an American soldier captured by Italian forces, who are aligned with the Welsh. You do not believe the Italians speak any Welsh, and wish to convince them that you are a Welsh soldier. You think back to your days as a child singing in *Cymanfaoedd Canu* and recite the only phrase you can remember, the first line of Calon Lân: *"Nid wy'n gofyn bywyd moethus, Aur y byd na'i berlau mân"*

Under Grice's account, the *meaning* of this utterance is "I am a Welsh soldier". Searle, on the other hand, would say that the meaning of this sentence is its *illocutionary effect*, i.e., "I do not ask for a luxurious life, the world's gold or its fine pearls" (or, rather, the propositional content associated with that sentence), whereas convincing the interlocutor of Welsh heritage is the intended *perlocutionary effect* of the utterance. Searle argues that under Grice, one intends to perform a *perlocutionary act* when they mean something by an utterance, whereby, he would say, one really intends to perform an *illocutionary* act. Searle presents three reasons why this is the case:

First, many utterances have no perlocutionary effect. For example, when you say "Hello", you are not trying to evoke some perlocutionary effect, e.g., persuading your interlocutor that they are being greeted; you just *want them to know* that they are being greeted. Second, you can say something without actually caring if you convince your interlocutor, instead only caring whether your interlocutor has *heard* you, e.g., if you feel that you have to announce something out of duty. Finally, when you say something to someone, you do not generally believe that they will accept it solely because they will understand that you want them to accept it.

Searle's solution is to provide an account of meaning which suggests that to say that S meant M by utterance U (directed towards hearer H) is to say that S intended to produce the knowledge that the culturally accepted meaning of U obtains (an *illocutionary effect* IE in the mind of H) because H will recognize this intention (because H is aware of the cultural conventions C that will result in this inference).

Searle posits five main categories of illocutionary acts (i.e., assertives, directives, commissives, expressives, declarations), each of which contains several subcategories of acts. Under Searle's framework, utterance understanding is effected using constitutive rules which specify the conditions under which an utterance can be used to "count as" an attempt to perform one of these types of illocutionary acts.

Under this framework, Speaker S uttering a sentence and meaning M thus additionally involves S intending H to use *conventionalized rules* to recognize that S wants H to know that S *means* M. A rule set for a particular illocutionary act act_X follows the following overall form:

1. Propositional content condition: Utterance U is only to be uttered in the context that it predicates propositional content $prop_X$.

2. Preparatory conditions: Utterance U is only to be uttered if the *preparatory conditions* $prep_X$ hold.

3. Sincerity condition: Utterance U is only to be uttered if the speaker intends for the effects of act_X to happen.

4. Essential Effect: When conditions 1-3 apply, Utterance U counts as an attempt to perform act_X, and for the effects of act_X to happen.

For example, consider the case of the illocutionary act of *requesting*.

1. Propositional content condition: Utterance U is only to be uttered to hearer H in the context that it predicates *H doing action A in the future.*

2. Preparatory conditions: Utterance U is only to be uttered by Speaker S to Hearer H if *H is able to do A, and S believes this.*

3. Sincerity condition: Utterance U is only to be uttered to Hearer H if *S wants H to do A.*

4. Essential Effect: When conditions 1-3 apply. Utterance U counts as S attempting to get H to do A in the future.

Thus, if Alice says "Get me a beer" to Bob and intends this to be a request, then she expects Bob to recognize, by virtue of these rules, that she is requesting Bob to get her a beer, i.e., because Bob should have no problem believing that:

1. Alice's utterance involves Bob getting her a beer (and is not being issued in a quotational or narrative context),

2. Alice believes that Bob is able to get her a beer,

3. Alice wants Bob to get her a beer

All three of these conditions are truly important to the correct interpretation of an utterance. If Alice issues some utterance, and Alice believes that Bob is able to get her a beer, there is no reason for her to believe that Bob will recognize that she is requesting Bob to get her a beer if (a) the utterance did not involve Bob getting her a beer (e.g., if the words Alice used were "Have you seen Westworld?") or if (b) Bob previously had reason to believe that Alice explicitly did not want to be given a beer (e.g., if previous conversation had established that all the beer in the house had been brewed using Isinglass or Gelatin, and that as a strict vegan, Alice viewed it as morally wrong for the beer to be consumed).

6.1.3 Searle's Theory of Indirect Speech Acts

Searle's Speech Act Theory is not (on its own) sufficient to account for utterances such as "Could you pass the *Bara brith*?" whose *literal* meanings do not match their *intended meanings*, i.e., whereby one illocutionary act is performed *by way of* performing some other illocutionary act. In order to handle such utterances, Searle introduces a theory of *Indirect Speech Acts* (Searle, 1975).

As an example, Searle presents an example of the following form:

(8) a. Alice: Let's go to the Eisteddfod!
b. Bob: I have to study for an exam.

In Example 8b, Searle would say that Bob is performing the *primary, nonliteral* illocutionary act of rejection *by way of* performing the *secondary, literal* illocutionary act of making a statement. Searle would posit, then that Alice should (implicitly) go through the following reasoning steps:

1. Because I made a proposal, and because I believe Bob is being cooperative (i.e., relevant[1]), the appropriate responses would be: acceptance, rejection, counterproposal, further discussion, etc.

2. Because his response was not one of these, I must assume that his statement is only his *secondary* illocutionary point.

3. I can infer based on his statement that one of the preparatory conditions of an acceptance to my proposal (i.e., his being free to accept it) is not met; thus, it is most likely that the statement is intended as a rejection.

In the example above, the agent must infer that one of the preparatory conditions of the *most likely response* is not met. However, the majority of indirect speech acts are constructed by calling attention to (by asking about or stating) the preparatory, sincerity, or propositional conditions of one's intended primary illocutionary act, or by calling attention to (by asking about or stating) some fact from which it might be inferred that the preparatory, sincerity, or propositional conditions of such an act might be met. For example, "Could you wash the car?" calls attention to the fact that the listener satisfies a salient preparatory condition of directives: that the hearer is *able* to carry out the directive.

It is important to note, however, that one *cannot* perform an indirect speech act by asking about their own mental states (e.g., "Do I want you to wash the car?"), as this attempts to call attention to information the hearer cannot have access to, or by making statements about their interlocutors'

[1]Here, Searle makes implicit reference to Grice's Conversational Maxim of Relation, as presented in Grice's "Logic and Conversation"(Grice, 1970), which appeared in the same volume as Searle's "Indirect Speech Acts". This Maxim is also listed above, in Section 6.1.1.

mental states (e.g., "You want to wash the car"), which presumes the speaker to make claims about information they cannot have access to.

The use of indirect speech acts is common (at least in English) due to social concerns such as politeness; since the above method (of calling attention to conditional requirements while respecting theory-of-mind restrictions) is the most common means of constructing indirect speech acts, the most common constructions of this form have become *conventionalized*, such that the inference process described above (known as the *inferential* approach to indirect speech act understanding) is no longer needed. For example, if an interlocutor says to you "Could you get me a coffee?" it is not necessary in most circumstances to reason about what they are attempting to call attention to and why: it is automatically understood that their use of this utterance form implies that they are trying to politely direct you to acquire a coffee for them. The process of following this direct association rather than deriving it through inference is known as the *idiomatic* approach to indirect speech act understanding.

Thus far, I have provided an appropriate philosophical account of pragmatic reasoning (i.e., indirect speech act usage) in human-human communication. Before applying this account to human-*robot* communication, however, we must experimentally verify that this phenomena is not restricted to human-human communication, and that the social considerations that lead to indirect speech act usage in humans will carry over to human-robot interaction and lead to indirect speech act usage therein as well.

6.2 Experimental Motivations

Due to the social benefits of politeness, there has been a significant body of human-robot interaction research investigating strategies for *using* politeness strategies such as indirect speech, and the effects of those strategies on human perceptions of robots (e.g. Briggs & Scheutz, 2013; Castro-González et al., 2016; Dautenhahn et al., 2005; Kennedy, Baxter, & Belpaeme, 2014; Nomura & Saeki, 2009; Salem, Ziadee, & Sakr, 2013, 2014; Strait, Canning, & Scheutz, 2014; Torrey, Fussell, & Kiesler, 2013; Torrey, Powers, Marge, Fussell, & Kiesler, 2006).

Furthermore, a variety of work over the past few decades (e.g. Briggs & Scheutz, 2013; Deits, Tellex, Kollar, & Roy, 2013; Hinkelman & Allen, 1989; Wilske & Kruijff, 2006) (See also Section 6.3) has investigated mechanisms by which robots and other intelligent agents might automatically *understand* (i.e., infer the intentions behind) indirect speech acts.

But while the ability to understand indirect speech acts may allow an intelligent agent to understand a wider range of human language, it is not clear to what extent humans will actually *use* indirect speech acts when communicating with such an agent. One could argue that humans are not held to the same social contracts and pressures when interacting with intelligent agents such as robots as they are when interacting with other humans. It is thus important to investigate the extent to which indirect language will actually be used, as well as the extent to which an intelligent agent could get away *without* the ability to understand indirect speech acts.

Specifically, in this section we seek to examine the following hypotheses:

H1 Indirect speech acts are central to human-robot dialogue patterns regardless of task context.

H2 Indirect speech acts are central to human-robot dialogue patterns regardless of whether or not they are actually understood by robot interlocutors.

H3 While indirect speech acts are central to human-robot dialogue patterns regardless of task context, they are *more frequently* used in highly conventionalized scenarios.

H4 Because ISAs are central to human-robot dialogue patterns, a human interacting with a robot unable to understand ISAs will be less efficient in accomplishing a desired task than will a human interacting with a robot able to understand ISAs.

H5 Because ISAs are central to human-robot dialogue patterns, a robot unable to understand ISAs will be perceived less favorably by its human interlocutor than a robot that is able to understand ISAs.

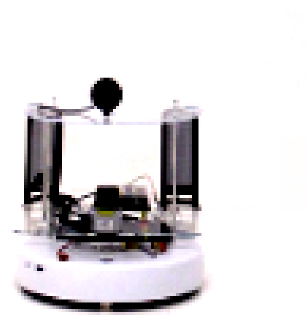
6.2.1 Methodology

In order to investigate these five hypotheses, we conducted a Wizard-of-Oz experiment in which humans had to interact with a robot in one of two contexts. The first was a restaurant scenario: a context in which humans readily use indirect speech acts (e.g., "Could I get a coke?") to give orders to employees obligated to fulfill their requests. The second was a tower-toppling scenario in which participants needed to command a robot to knock down colored towers of cans (see Briggs & Scheutz, 2014a): a context which participants would be unfamiliar with, and for which there should be no conventionalized social norms requiring the use of indirect speech acts (although they could still be used felicitously, e.g., "Could you knock down the red tower"). In each of these contexts, participants interacted with a robot that either understood, or was clearly unable to understand, indirect speech acts.

Procedure

Participants were randomly assigned to one of four experimental conditions, in which the task context was either that of a *restaurant* or a *demolition*, and in which the robot either *understood* or *misunderstood* indirect speech acts. In both conditions, participants were seated in the corner of a small experiment room. In the restaurant scenario, the room was empty; in the demolition scenario, the room contained three colored towers of aluminum cans.

Figure 6.1: Augmented iRobot Create used in our Experiment



Participants were told that the experimenters were in the process of developing natural language interaction capabilities for robots, and that their task would be to interact with either a robot waiter or a tower-toppling robot in a simulated restaurant or demolition scenario: after being introduced to the robot, they were to provide the robot with their first order (in the restaurant scenario, they were provided with a list of three "courses" which they could request to be delivered in any order; in the demolition scenario, they were provided with a list of three towers which they could request to be knocked down in any order); once the robot had completed the first order, they were to provide the robot with their second order; once the robot had completed the second order, they were to provide the robot with their third order.

The robot then entered the room and introduced itself. The robot used in this experiment was an iRobot Create augmented with a Raspberry Pi computer, Hokuyo Laser Range Finder, speakers, and webcam, as shown in Figure 6.1. The robot was teleoperated through a Wizard of Oz interface by a trained confederate in a nearby room, using the *ADE* implementation of the *DIARC* architecture (see Chapter 2). The robot's voice was an ungendered voice produced through the MaryTTS text-to-speech system (Schröder & Trouvain, 2003).

The human and robot then engaged in each of the three sub-interactions, which proceeded as follows: if the participant used a *direct speech act*, such as "Knock down the red tower" or "Bring me a salad", or used a bare noun phrase, such as "red tower" or "salad", the robot carried out the command. In the restaurant scenario, this consisted of driving into a nearby room where a confederate placed a card corresponding with the requested dish onto the top of the robot, after which the robot drove back into the experiment room, delivered the card to the participant, and requested their next order. In the demolition scenario, this consisted of driving into the requested tower of cans until it had fallen over.

If the participant used an *indirect speech act*, such as "Could you knock down the red tower" or "I need a salad", and the participant was in the UNDERSTANDING condition, the robot carried out their request as if they had used an equivalent direct form. If the participant used an indirect speech act and the participant was in the MISUNDERSTANDING condition, the robot took their utterance at face value, issuing a response according to the Table 6.1. This table shows the forms of indirect speech acts we would expect to observe in our experiments, based on combinations of *direct illocutionary point* (i.e., statements, questions and suggestive statements), *condition of focus* (i.e., sincerity condition or preparatory condition), and *direction of focus* (i.e., agent or patient). For example, "Could you bring me a salad?" has the illocutionary point of a question, focuses on a preparatory condition of questions (i.e., the condition that the addressee is *able* to perform the desired action), and focuses on the agent (i.e., the addressee bringing the salad, as opposed to themselves receiving the salad) – facets derived from Searle's *Speech Act Theory* (Searle, 1969, 1975, 1976) (see also Section 6.1.2).

Participation

Participants were recruited online and through fliers posted near a university campus. Before beginning the experiment, participants were given a short demographic survey regarding their prior experience with robots and their

Table 6.1: Responses Given for ISAs Categories

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 I get X?"	"Yes, that is permissible. Please tell me your order."
S	S	P	"I'd like X."	"Thank you for sharing that interesting fact. Please tell me your order."
S	P	P	"I'll have X."	"Thank you for sharing that interesting prediction. Please tell me your order."
S[Su]	P	P	"X should occur."	"Thank you for your suggestion. Please tell me your order."

Direct (Direct Illocutionary) Point: Q=Question, S=Statement, S[Su]=Suggestive Statements; Cond(ition): P=Preparatory, S=Sincerity; Dir(ection): A=Agent, P=Patient.

use of technology: Of the participants, 21 were male, and 28 were female. Participants varied in age between 19 and 69 (m = 34.1, sd = 16.2). 12 participants were initially assigned to each experimental condition; one additional participant was later recruited for the RESTAURANT/UNDERSTANDING condition after a previous participant forgot to answer a large number of survey questions. While the majority of participants were beyond college age, we asked them for information regarding their current or previous college major, if any. 6 reported studying mathematics, computer science or engineering; 15 reported studying a natural science or medicine; 10 reported studying a social science; 7 reported studying a branch of the arts or humanities; 5 reported studying some other field; 6 reported no previous or current major. A minority (12/49 = 24%) reported playing video games. The vast majority (45/49= 92%) reported having seen a robot in a movie, but fewer than half (22/49 = 45%) had seen a robot in real life, and a minority reported having interacted with a robot before (11/49 = 22%). Participants were paid $10 each for their participation and provided informed written consent before beginning the experiment.

Measures

The dependent variables used in this study involved both *behavioral* and *subjective* measurements.

Behavioral In order to assess Hypotheses **H1-H4**, utterances made by participants during the experiment were recorded and transcribed afterward. Annotators then classified all task-relevant utterances as either *direct* or *indirect*.

Subjective In order to assess Hypothesis **H5**, participants took a 60-item post-survey immediately following the experiment, assessing their perceptions of the robot, as well as their beliefs about the types of utterances that would have been appropriate and effective. The majority of this study (regarding perception of the robot) was drawn from the survey first used in (P. Schermerhorn, Scheutz, & Crowell, 2008).

Expectations

If our five hypotheses are correct, we would expect the following results:

E1 If ISAs are central to human-robot dialogue patterns regardless of task context, we would expect a significant number of ISAs to be used in both task contexts (RESTAURANT and DEMOLITION). "Significant" is an obviously subjective term – but for the purposes of this experiment, we will consider a frequency of ~12% (the current speech recognition state of the art on the Switchboard corpus) to be a reasonable threshold.

E2 If indirect speech acts are central to human-robot dialogue patterns regardless of whether or not they are actually understood by robot interlocutors, we would expect a significant number of ISAs to be used in both dialogue conditions (UNDERSTANDING and MISUNDERSTANDING)

E3 If indirect speech acts are *more frequently* used in highly conventionalized scenarios, we would expect a statistically significantly greater number of ISAs to be used in the RESTAURANT task context than in the DEMOLITION task context.

E4 If a human interacting with a robot unable to understand ISAs is less efficient in accomplishing a desired task than a human interacting with a robot able to understand ISAs, we would expect that a statistically significantly greater number of task-relevant utterances will be required to accomplish the task in the MISUNDERSTANDING scenario than in the UNDERSTANDING scenario.

E5 If a robot unable to understand ISAs is perceived less favorably by its human interlocutor than a robot that is able to understand ISAs, we would expect to find a number of dimensions in our post-survey in which robots' assessments are statistically significantly lower in valence for participants in the MISUNDERSTANDING scenario than in the UNDERSTANDING scenario.

6.2.2 Results

Both behavioral and subjective measures were analyzed using 2x2 analyses of variance (ANOVA) with task context (RESTAURANT vs DEMOLITION) and dialogue condition (UNDERSTANDING vs MISUNDERSTANDING) as independent variables. In order to use this methodology to analyze Likert items, we assumed that responses to these items would be normally distributed (Gombolay & Shah, 2016).

Behavioral Results

We hypothesized (**H1**) that ISAs would be central to human-robot dialogue patterns across task scenarios, but more frequently used (**H3**) in highly conventionalized scenarios. We thus expected (**E1**) ISAs to comprise at least 12.5% of task relevant utterances in both the RESTAURANT and DEMOLITION task contexts, but (**E3**) for this composition to be significantly higher in the RESTAURANT task context than in the DEMOLITION task context.

In fact, ISAs comprised the *majority* (52%) of the 295 observed task-relevant utterances, the majority of participants ($34/49 = 69\%$) used at least one ISA, and the average proportion of task-relevant utterances coded as ISAs was 46%. This strongly supports our first hypothesis (**H1**). Furthermore, this average was twice as high in the RESTAURANT condition (M=63%, SD=39%) as it was in the DEMOLITION condition (M=28%, SD=34%), supporting ($F(1,47)=11.05, p<.01$) our third hypothesis (**H3**).

In addition, we hypothesized (**H2**) that indirect speech acts would be central to human-robot dialogue patterns regardless of whether or not they were actually understood by robot interlocutors, but that a human interacting with a robot unable to understand ISAs would be less efficient in accomplishing a desired task than a human interacting with a robot able to understand ISAS (**H4**). We thus expected (**E2**) ISAs to comprise at least 12.5% of task relevant utterances in both the UNDERSTANDING and MISUNDERSTANDING dialogue conditions, but (**E4**) that a statistically significantly larger number of task relevant utterances would be used in the MISUNDERSTANDING dialogue condition than in the UNDERSTANDING dialogue condition.

In fact, ISAs were used about twice as frequently as this threshold in both experimental conditions (M=24%,SD=23% in the UNDERSTANDING condition, M=24%,SD=20% in the MISUNDERSTANDING condition), supporting our second hypothesis (**H2**). Furthermore, participants in the MISUNDERSTANDING dialogue condition did indeed use significantly ($F(1,47)=7.99$ ($p<.01$)) more task-relevant utterances (M=8.08,SD=7.42) to complete the task than did participants in the UNDERSTANDING dialogue condition (M=4.04,SD=1.21), supporting our fourth hypothesis (**H4**).

Subjective Results

We hypothesized (**H5**) that a robot unable to understand ISAs would be perceived less favorably by its human interlocutor than a robot that is able to understand ISAs. We thus expected (**E5**) to find a number of dimensions in our post-survey in which robots' assessments were statistically significantly lower in valence for participants in the MISUNDERSTANDING scenario than in the UNDERSTANDING scenario.

Table 6.2 lists the significant effects of dialogue condition on perception of the robot. The data suggest that participants found the robot to be harder to interact with in the MISUNDERSTANDING dialogue condition (Rows 1-3), as well as less responsive (Row 4) and cooperative (Row 5). The data also suggest that participants found the robot in the UNDERSTANDING condition to be less annoying (Row 6), to better understand their commands (Row 7), and to have better matched their expectations (Row 8). Taken together, these results support our fifth hypothesis (**H5**).

A number of context-dependent effects were also found, as shown in Table 6.3: Participants in the RESTAURANT context rated the task to be more difficult (Row 1), and the robot to be less responsive (Row 2), less

Table 6.2: Subjective Results: Effects of Dialogue Condition

	Question	F	p	M: m (sd)	U: m (sd)
1	The robot was easy to interact with (from 1 to 10, 'strongly disagree' to 'strongly agree')	16.02	<.001	7.67 (2.35)	9.60 (0.76)
2	How would you rate the ease of interacting with the robot? (from 1 to 10, "easy" to "hard")	17.00	<.001	3.33 (2.60)	1.24 (0.52)
3	How would you rate the difficulty of the task? (from 1 to 10, "easy" to "hard")	4.19	.047	3.08 (2.32)	1.96 (1.85)
4	The robot was responsive to my commands (from 1 to 10, 'strongly disagree' to 'strongly agree')	6.83	.01	8.83 (2.08)	9.56 (1.41)
5	The robot was cooperative (from 1 to 10, 'strongly disagree' to 'strongly agree')	9.65	<.001	8.67 (1.90)	9.80 (0.50)
6	The robot was annoying (from 1 to 10, 'strongly disagree' to 'strongly agree')	5.95	.02	2.75 (2.29)	1.52 (1.12)
7	Did you feel that the robot understood what you were saying? (from 1 to 10, "no" to "yes")	7.15	.01	7.38 (2.72)	9.08 (1.89)
8	Did the robot meet your expectations? (from 1 to 10, "no" to "yes")	4.37	.042	7.73 (2.73)	8.80 (2.22)

All results are for F(1, 45). M = MISUNDERSTANDING; U = UNDERSTANDING.

cooperative (Row 3), and having poorer comprehension (Rows 4-5) than did participants in the DEMOLITION context.

Finally, we asked participants towards the end of their questionnaire if they would do anything differently if they had to do the study again. Out of the 49 participants, 31 participants said they would, and said that they would change the way that they phrased their commands. Of these 31, seven indicated that they would make their commands more direct, nine indicated that they would make their commands less direct, and 15 were not clear

Table 6.3: Subjective Results: Effects of Task Context

	Question	F	p	R: m (sd)	D: m (sd)
1	How would you rate the difficulty of the task? (from 1 to 10, “easy” to “hard”)	7.98	.007	3.28 (2.48)	1.71 (1.40)
2	The robot was responsive to my commands (from 1 to 10, ’strongly disagree’ to ’strongly agree’)	6.79	.01	8.36 (2.34)	9.58 (0.83)
3	The robot was cooperative (from 1 to 10, ’strongly disagree’ to ’strongly agree’)	5.27	.02	8.84 (1.80)	9.67 (0.92)
4	How would you rate the robot’s level of comprehension? (from 1 to 10, “low” to “high”)	5.66	.02	6.72 (3.08)	9.38 (1.83)
5	Did you feel that the robot understood what you were saying? (from 1 to 10, “no” to “yes”)	4.35	.04	7.60 (2.99)	8.92 (1.56)

All results are for F(1, 45). R = RESTAURANT; D = DEMOLITION.

(e.g., one participant said they would “ask for items in a different manner”). What is striking is that only five participants used more than five utterances – and of those five participants, four indicated that they would be less direct. One such participant said “I would use concise commands, because it gets confused when I talk like a normal casual person.” Another participant said they would “cut out unnecessary words used for politeness to be more direct in asking for what I want” Yet another said “I would speak to the robot very literally, not like a human.” These participants attempted to speak to the robot in *normal*, *polite*, *human-like* ways, and acknowledged that they would have to give up such a way of speaking in order to have successful interactions in the future. In contrast, participants who said they would be more indirect gave reasons that seemed more exploratory, likely as a result of the previous set of survey questions: One participant said, for example, “I would use different sentence structure and make my words confusing.” – another said “I would experiment with different commands, phrases, attempt to talk more freely [with] the robot both before and during its tasks.”

Figure 6.2: In-Task Human-Robot Dialogue (RESTAURANT context, MISUNDERSTANDING dialogue condition)

Robot	Hello my name is Waiterbot. I will be helping you today.
Participant	Great, can I order something?
Robot	Please tell me your order.
Participant	Um, can I have one water?
Robot	Yes, that is permissible.
Participant	Great.
Robot	Please tell me your order.
Participant	Can I have one water?
Robot	Yes, that is permissible.
Participant	Great, I'll take one water.
Robot	Thank you for sharing that interesting prediction. Please tell me your order.
Participant	Can I order one water?
Robot	Yes, that is permissible.
Participant	Okay, can you bring me one water?
Robot	Yes, I am able to do that.
Participant	May you bring me one water?
Robot	Yes, I am able to do that.
Participant	I would like -
Robot	Please tell me your order.
Participant	I would like to have one water.
Robot	Thank you for sharing that interesting fact. Please tell me your order.
	16-Second Pause
Participant	Can you bring me one water?
Robot	Yes, I am able to do that. Please tell me your order.
Participant	May you please bring me one water?
Robot	Yes, I am able to do that.
Participant	So can you do it?
Robot	Yes. Please tell me your order.
Participant	Can you go inside and get the water for me?
Robot	Yes, I am able to do that. Please tell me your order.
Participant	My order is water-
Robot	Please tell me your order.
Participant	My order is water.
Robot	Okay.
	Robot retrieves meal card for participant
Robot	Here is your drink.
Participant	Thank you.
Robot	You're welcome. Please tell me your order.
Participant	My order is two cakes.
Robot	Okay.
	Robot retrieves meal card for participant
Robot	Here is your dessert.
Participant	Thank you.
Robot	You're welcome. Please tell me your order.
Participant	My order is done.
Robot	Please tell me your order.
Participant	My order is noodles.
Robot	Okay.
	Robot retrieves meal card for participant
Robot	Here is your meal.
Participant	Thank you.
Robot	You're welcome. Thank you for helping me. We are now done.

Figure 6.3: In-Task Human-Robot Dialogue

	...
Participant	Can you knock over the yellow tower please?
Robot	OK. ... I have knocked down the yellow tower.
Participant	Thank you. Can you please knock down the blue tower now, please?
Robot	OK. ... I have knocked down the blue tower.
Participant	And can you knock down the red tower now?
Robot	OK. ... I have knocked down the red tower.

Figure 6.4: In-Task Human-Robot Dialogue from Additional Interaction Corpus 1 (Aldebaran Nao capable of understanding ISAs in a DEMOLITION scenario)

Participant	Oh lets see here. I have to do something. I would like you to go to the blue tower of cans to your left.
Robot	OK.
Participant	And I would like you to knock down the top can on the blue tower.
Robot	OK.
Participant	You can knock down all of the cans if you like.
Robot	OK.
Participant	Very good, thank you so much. Nice job!

6.2.3 Supplemental Analysis

Our results showed that not only did most participants use ISAs in this scenario, but that some relied almost exclusively upon them, as seen in Figure 6.3. 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 section, we examined two additional interaction corpora collected from previously conducted but unpublished experiments: (1) an experiment similar to the DEMOLITION x UNDERSTANDING condition of the presented experiment, differing only in that an Aldebaran Nao was used, and (2) an experiment similar to the RESTAURANT x MISUNDERSTANDING condition of the presented experiment.

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

Figure 6.5: In-Task Human-Robot Dialogue from Additional Interaction Corpus 2 (iRobot Create incapable of understanding ISAs in a RESTAURANT scenario)

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

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 6.4, ISA use was common across all three experiments, albeit slightly higher in this larger sample study than in the two smaller sample studies. Figure 6.4 shows a dialogue drawn from the first of these two interaction corpora (Row 2 in Table 6.4); Figure 6.5 shows a dialogue drawn from the second of these two interaction corpora (Row 3 in Table 6.4).

Table 6.4: Meta-Analysis of Indirect Speech Act Use Across Experiments.

#	Robot Used	Context	Understanding	n	TRUs	ISAs
1	iRobot Create	Both	Conditional	49	295	51.5%
2	Aldebaran Nao	DEMOLITION	Yes	13	167	10.2%
3	iRobot Create	RESTAURANT	No	21	160	42.0%
				83	622	40.5%

6.2.4 Discussion

In this section, we will first summarize the results of our experiment with respect to our experimental hypotheses. We will then discuss the theoretical implications of these results and what they suggest for for robot designers. Next, we will examine how the ISAs used in both our primary experiment and secondary interaction corpora can be categorized in order to produce a taxonomy of observed ISAs, and discuss the implications for how robot designers might proactively predict what ISA forms are likely to occur for their own application domains. Finally, we will conclude with some possible directions for future experimental work.

Justification of Experimental Hypotheses

The results of our experiment supported each of our five hypotheses. We hypothesized (**H1**) that ISAs are central to human-robot dialogue patterns regardless of task context, and our results showed that ISAs were frequently used in both task contexts. Conversations such as those shown in Figure 6.2 demonstrate participants' general reluctance to command the robot using *direct* commands.

We hypothesized (**H2**) that ISAs are central to human-robot dialogue patterns regardless of whether or not they are actually understood by robot interlocutors, and our results showed that ISAs were frequently used in both dialogue conditions.

We hypothesized (**H3**) that while indirect speech acts are central to human-robot dialogue patterns regardless of task context, they are *more frequently* used in highly conventionalized scenarios, and our results suggested that ISAs were used significantly more often in our highly conventionalized task context than in our novel task context.

We hypothesized (**H4**) that because ISAs are central to human-robot dialogue patterns, a human interacting with a robot unable to understand ISAs will be less efficient in accomplishing a desired task than will a human interacting with a robot able to understand ISAs, and our results suggested that participants that interacted with a robot unable to understand ISAs needed to use significantly more task relevant utterances to accomplish their task than did participants that interacted with a robot able to understand ISAs.

Finally, we hypothesized (**H5**) that because ISAs are central to human-robot dialogue patterns, a robot unable to understand ISAs will be perceived less favorably by its human interlocutor than a robot that is able to understand ISAs, and we found that robots unable to understand ISAs were rated less favorably on a number of dimensions, relative to robots that *were* able to understand ISAs. However, we acknowledge that the large number of interesting questions we sought to examine in the data led to a large amount of analysis, increasing the possibility of Type I errors. While the likelihood of Type I errors could be decreased through the use of a Bonferoni correction, this technique has a number of deleterious consequences, including an increased likelihood of Type II errors (Perneger, 1998). Instead, we simply direct the reader to consider the presented effect sizes when assessing the reported results.

Theoretical Implications

These results suggest 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 domain-independent 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:
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. This will be particularly true in contexts with conventionalized social norms for which we observed an ISA usage rate of 69%.

Furthermore, our results imply that even if robot designers are willing to accept such a high error rate, the perception of their robot will be severely impaired: users may find the robot difficult to interact with, unresponsive, uncooperative, annoying, and uncomprehending – at least in comparison to how it would have otherwise been perceived.

And while the *specific* scenarios used in this experiment are simple enough that a robot could perform effectively without ISA understanding capabilities (i.e., with simple word-spotting algorithms) if the robot was *designed specifically for use in that scenario*, we emphasize that the goal of this dissertation is to develop mechanisms for, and study interaction with, robots that are *generally taskable* and capable of working in *any* application domain.

We thus suggest that language-enabled robots engaging in domain-independent 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.

Ontological Analysis and Taxonomy

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. As discussed in Section 6.1.2, Searle's theory of speech acts suggests that 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, 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) a constituent 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[2]. 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.

Table 6.5: Taxonomy of Observed Indirect Requests

Direct Point Condition	Question Preparatory	Statement Sincerity	Statement Preparatory	Suggestion 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.)

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 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

[2]The final parenthetical item is a form we did not observe in our experiment or additional interaction corpora, but which fits the presented framework.

abilities of others, especially when they are the presumed domain experts. Consider the third column of Table6.5. 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 anyways. 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. If 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, 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.

Future Experimental Work

We foresee a number of interesting directions for future work. In this experiment, we observed participants continuing to use ISAs even when the robots clearly and repeatedly failed to understand indirect language: future research should examine whether participants would be able to refrain from using ISAs if explicitly told not to, or explicitly told that a robot was unable to comprehend them. We would also like to examine how facets of a robot affect linguistic interaction patterns with that robot. In this experiment, we used a single robot with a single voice. It is unclear whether ISA usage patterns would have differed if the robot had a different (e.g., more human) morphology, of if the robot had used a clearly gendered voice. Future research should also develop mechanisms whereby robots can *automatically* learn to understand ISAs in general, or to understand specific newly encountered ISA forms. Finally, the issues discussed in this section are of import to a wide variety of intelligent agents beyond robots, and it will be important to investigate the extent to which the effects found in this section depend upon the type of agent used, and whether these patterns hold for non-embodied and non-situated agents.

In this chapter, thus far, I have discussed our philosophical motivations, focusing on Searle's theory of indirect speech acts. I then presented experimental evidence that indirect speech acts are commonplace in human-*robot* interaction and demonstrated how one might analyze the *types* of indirect speech acts that might be found in a human-robot interaction domain, in order to determine what inference rules might need to be written to appropriately handle likely indirect speech acts. Before discussing how we actually go about writing such rules, and the algorithms we have created to make use of such rules, I will briefly discuss previous approaches to indirect speech act understanding and generation.

6.3 Previous Computational Work

Over the past 40 years, there has been much research on indirect speech acts in philosophy and linguistics(Searle, 1975; Pinker, Nowak, & Lee, 2008), and much computational work on ISA understanding (outside the context of robot architectures). Within robotics, there has been a significant amount of work on enabling natural language capabilities in integrated robot architectures (To name a few illustrative examples from the past few years alone, Lemaignan, Warnier, Sisbot, & Alami, 2014; Kruijff et al., 2010; J. Y. Chai et al., 2014; Deits, Tellex, Kollar, & Roy, 2013; Scheutz et al., 2013; Jing, Finucane, Raman, & Kress-Gazit, 2012), but very little work on enabling any sort of pragmatic analysis (e.g., indirect speech act understanding or generation)

Indeed, there seems to be a disconnect between the language-capable agents currently being commercialized and the robots of the future which we seek to develop. In this dissertation, I have made clear that our goal is to develop robots capable of *natural human-like human-robot interaction* that are generally taskable, and not crafted for particular scenarios. In contrast, the language-capable agents that are currently being commercialized (mostly phone-based personal assistants such as Siri, Cortana, and Google Now, but also robots such as Jibo) are intended to operate within restricted domains for which so-called "deep" linguistic processing is unnecessary[3], as "shallow" natural language processing (NLP) methods such as keyword spotting will

[3]Although in truth, the terminology of "deep" vs. "shallow" linguistic processing is not always terribly informative. It is illustrative, for example, to consider approaches such as that of DeVault & Stone (2009), who claim to determine the "deep semantics" of incoming utterances, but do not actually ascertain the intended *meanings* of such utterances (as one would if using an indirect speech act understanding algorithm), which is presumably a yet-deeper level of meaning.

typically be sufficient. Furthermore, even outside of robots actively intended for commercialization, many robot designers seem to make an assumption that all utterances that will be provided to robots will be *commands*, and that future natural language understanding (NLU) systems will only need to solve the problem of identifying the verb used in such a command and its arguments (e.g. Tanaka, Tokunaga, & Shinyama, 2004).

Shallow processing techniques are *not* sufficient, however, for many robots intended to interact with humans, who may need to understand a richer class of natural language expressions, and for whom the consequences of misunderstanding may be more dire. Many designers of such natural language enabled robots, however, seem to have tacitly assumed that human interlocutors will restrict themselves to direct commands, and that an ability to understand and generate indirect speech acts is simply unnecessary. As we have shown, however, this assumption is unwarranted. Not only are indirect speech acts *used* in simple, task-based human-robot dialogues: they are *commonplace*. We must thus consider previous work on indirect speech act understanding and generation, to see how it could be applied or extended for use in robotics.

6.3.1 Inferential vs. Idiomatic Approaches

Computational approaches to indirect speech act understanding and generation has tended to fall into two categories. First, there has been a set of *inferential* approaches, in which indirect speech act understanding is regarded as a *plan reasoning* problem(Allen, 1979; Allen & Perrault, 1979, 1980; P. R. Cohen & Perrault, 1979; P. R. Cohen & Levesque, 1985; Perrault & Allen, 1980). By using rich knowledge about interlocutors' beliefs, desires, and intentions, these approaches are able to infer the intended meanings behind complicated *unconventionalized* indirect speech acts. For example, such a system might be able to infer that "It's cold in here!" was issued in order to command the hearer to close an open window.

Also residing beneath the umbrella of "inferential" approaches would be work such as that of Herzig & Longin (2002), who present a cognitive robotics oriented logic intended to allow robots to infer when assertions made to a robot (e.g., "I intend to know whether X holds") are intended to be interpreted as either yes-no questions or more generally as requests, rather than as simple assertions. As we will later describe, this is something of an inverse to the types of inferences we would desire to make.

Similarly, this class includes work on inferring the intentions behind indirect *responses*. N. Green & Carberry (1999); de Marneffe, Manning, & Potts (2010) present computational models for understanding and generating indirect *responses* to questions. For example, "I can't drive" is an *indirect response* to the question "You'll probably get a car when you arrive?" Such indirect responses are often used by humans in order to answer implicit questions, to comply with social norms, to provide explanations, or to provide clarification.

Unfortunately, all approaches in this category tend to be very expensive computationally. As such, several researchers have turned their attention toward *idiomatic* approaches that cannot handle this broad class of *unconventionalized* indirect speech acts, but instead focus on handling the restricted set of *conventionalized* indirect speech acts, such as "Could you X?", "I need \$X\$", "Will you X?" "Didn't you want to \$X\$", and "Y would be happy if you did \$X\$", which are so commonplace that we automatically understand what is truly meant without the need for sophisticated reasoning processes (Searle, 1975) (see also the ontological analysis and taxonomy presented in Section 6.2.4). The idiomatic approach is attractive because conventionalized ISAs are the most common ISAs observed in dialogue, and can be understood and generated with much greater efficiency than can unconventionalized ISAs.

For example, Wilske & Kruijff (2006) present an idiomatic approach capable of handling some conventionalized ISAs, by mapping indirect *requests* directly to intended actions. This approach is also advantageous with respect to the inferential approaches in that it attempts to handle *uncertainty*, and is able to generate *clarification requests* when it is unsure how to interpret an utterance. Unfortunately, this approach is limited to handling only indirect *commands*. Furthermore, the adaptations this approach is able to make with response to clarification request responses is restricted to all-or-nothing changes, which can lead to shifts in belief of unwarranted magnitude. Wilske and Kruijff attempt to rectify this problem by always allowing a chance for the robot to ask for clarification, so unwarranted belief shifts can be reversed. However, this can lead to superfluous questions (when the agent is fairly certain) and incorrect interpretations (when the agent has a belief that is certain and incorrect).

Finally, there have been a few *hybrid* approaches that have sought to employ both plan reasoning mechanisms to handle unconventionalized ISAs, *and* mechanisms for quickly handling conventionalized ISAs. Briggs & Scheutz (2013), for example, presents an approach within our own DIARC architecture that does just this, for the purposes of both understanding and generation. Similarly, Hinkelman & Allen (1989) introduced a hybrid approach that was later implemented by Allen et al. (2001) within the TRIPS system (G. Ferguson & Allen, 1998) (although to the best of our knowledge this approach was never implemented in an integrated *robotic* architecture). These approaches(see also Litman & Allen, 1987), however, like the aforementioned inferential approaches, are unable to handle any sort of uncertainty.

6.3.2 Algorithmic Desiderata

In the next section, I will present a new algorithm for understanding indirect speech acts, which builds directly off of the idiomatic half of the work presented by Briggs & Scheutz (2013). In developing this algorithm, we sought to facilitate three primary capabilities, which we view as necessary for robust understanding of conventionalized ISAs in realistic human-robot interaction scenarios:

C1: Uncertainty. An agent must not assume perfect knowledge of the contexts in which an indirect interpretation applies. The conventionalized meaning of an ISA is not always the intended meaning; sometimes "I'd love some cake" is simply a statement expressing a desire, and not an indirect request for someone to give you cake. Since an agent might not always be able to determine the true intended meaning of an utterance, it should ascribe a level of confidence to each of its interpretations, based on the contextual factors that provide evidence for each interpretation. Furthermore, it is important that an agent be able to represent and reason about its own uncertainty and ignorance, and be able to act appropriately when uncertainty is identified.

C2: Adaptation. Since an agent should be able to learn new ISAs, and since it may not know the precise scenarios in which new ISAs should be used, an agent should be able to learn and adapt new rules, using feedback from interlocutors to adjust its beliefs as to when the rules it knows apply. For example, consider the dialogue shown in Figure 6.6.

In the space of this short dialogue (which can be viewed at `https://www.youtube.com/watch?v=39XnZAEl1Z4`), an agent (i.e., the android "Data") must make several adaptations. First, he must alter his beliefs about the ISA "just passing by" based on feedback from Geordi that the ISA's literal meaning had been the correct interpretation. Then, he must at least partially revert to his previous beliefs, as well as alter his belief as to when $said(X,Y) \rightarrow means(X,Y)$.

Figure 6.6: Short dialogue from *Star Trek: The Next Generation. "Interface"*

DATA Are you certain you do not wish to talk about your mother?

GEORDI Why do you ask that?

DATA You are no doubt feeling emotional distress as a result of her disappearance. Though you claimed to be "just passing by", that is most likely an excuse to start a conversation about this uncomfortable subject. Am I correct?

GEORDI Well, no. Sometimes "just passing by" means "just passing by."

DATA Then I apologize for my premature assumption...

GEORDI Data, maybe you gave up too fast.

DATA I do not understand.

GEORDI When I said "just passing by" means "just passing by", I didn't really mean it.

DATA My initial assumption was correct. You do wish to speak of your mother.

C3: Belief modeling. An agent should be able to model interlocutors' beliefs: the interpretation of an ISA *uttered by an interlocutor* should be based not on the *robot's* beliefs about, for example, its capabilities and obligations, but rather on its *interlocutor's* beliefs.

6.4 A Dempster-Shafer Theoretic Approach to Pragmatic Inference

Enabling a robot to understand a broad coverage of human speech acts requires a number of mechanisms within a robot architecture. In this section, we introduce novel algorithms based on Dempster-Shafer (DS) theory (Shafer, 1976) for inferring intentions I from utterances U in contexts C, and, conversely, for generating utterances U from intentions I in contexts C. We select more general DS-based representations over single-valued probabilities because the probability-based Bayesian inference problem to calculate $P(I|U,C)$ in terms of $P(U|I,C)$ is not practically feasible, for at least two reasons: (1) we do not have access to distributions over an agent's intentions (as we cannot look inside its head), and (2) we would need a table containing priors on all combinations of intentions and contexts. Instead, we employ rules of the form $u \wedge c \rightarrow_{[\alpha,\beta]} i$.

This serves to capture the intentions behind utterances in particular contexts, where $[\alpha, \beta]$ is a confidence interval contained in [0,1] which can be specified for each rule independently (e.g., based on social conventions, or corpora statics when available). These rules are very versatile in that they can be defined for individual utterances and contexts or whole classes of utterances and contexts. Most importantly, we can employ DS-based modus ponens to make uncertain deductive and abductive inferences which cannot be made in a mere Bayesian framework.

We start with background information regarding basic Dempster-Shafer theoretic concepts, and then introduce our proposed algorithm for pragmatic inference.

6.4.1 Basic Notions of Dempster-Shafer Theory

Since the proposed algorithms and architecture will use DS-theoretic representations of uncertainty, we briefly review the basic concepts of this framework for reasoning about uncertainty, which is a generalization or extension of the Bayesian framework (Shafer, 1976).

Frame of Discernment:

A set of elementary events of interest is called a *Frame of Discernment* (FoD). A FoD is a finite set of mutually exclusive events $\Theta = \{\theta_1, ..., \theta_N\}$. The power set of Θ is denoted by $2^\Theta = \{A : A \subseteq \Theta\}$.

Basic Belief Assignment:

Each set $A \in 2^\Theta$ has a certain weight, or *mass* associated with it. A *Basic Belief Assignment* (BBA) is a mapping $m_\Theta(\cdot) : 2^\Theta \rightarrow [0,1]$ such that $\sum_{A\subseteq\Theta} m_\Theta(A) = 1$ and $m_\Theta(\emptyset) = 0$. The BBA measures the support assigned to the propositions $A \subseteq \Theta$ only. The subsets of A with non-zero mass are referred to as *focal elements* and comprise the set F_Θ. The triple $E = \{\Theta, F_\Theta, m_\Theta(\cdot)\}$ is called the *Body of Evidence* (BoE).

Belief, Plausibility, and Uncertainty:

Given a BoE $\{\Theta, F_\Theta, m_\Theta(\cdot)\}$, the *belief* for a set of hypotheses A is $Bel(A) = \sum_{B \subseteq A} m_\Theta(B)$. This belief function captures the total support that can be committed to A without also committing it to the complement A^c of A. The *plausibility* of A is $Pl(A) = 1 - Bel(A^c)$. Thus, $Pl(A)$ corresponds to the total belief that does not contradict A. The *uncertainty* interval of A is $[Bel(A), Pl(A)]$, which contains the true probability $P(A)$. In the limit case with no uncertainty, we get $Pl(A) = Bel(A) = P(A)$. Thus, we see how information regarding the probability of some event can be gathered from the Dempster-Shafer theoretic notions of belief and plausibility, and how these notions themselves are derived from the masses m_Θ ascribed to specific hypotheses.

Inference and Fusion:

Uncertain logical inference can be performed using DS-theoretic modus ponens (denoted $\odot$) (Tang, Hang, Parsons, & Singh, 2012). We will use the DS-theoretic AND (denoted $\otimes$) to combine BoEs on different FoDs (Tang, Hang, Parsons, & Singh, 2012), and Yager's rule of combination (denoted $\bigcap$) to combine BoEs on the same FoD (Yager, 1987). We choose to use Tang's models of modus ponens and AND over other proposed models due to the counter-intuitive results of those models, and because those models do not allow uncertainty to be multiplicatively combined. Yager's rule of combination is chosen because it allows uncertainty to be pooled in the universal set, and due to the counter-intuitive results produced by Dempster's rule of combination (as discussed in (Zadeh, 1979)).

Logical AND:

For two logical formulae ϕ_1 (with $Bel(\phi_1) = \alpha_1$ and $Pl(\phi_1) = \beta_1$) and ϕ_2 (with $Bel(\phi_2) = \alpha_2$ and $Pl(\phi_2) = \beta_2$), applying logical AND yields $\phi_1 \otimes \phi_2 = \phi_3$ with $Bel(\phi_3) = \alpha_1 * \alpha_2$ and $Pl(\phi_3) = \beta_1 * \beta_2$.

Modus Ponens:

For logical formulae ϕ_1 (with $\mathrm{Bel}(\phi_1) = \alpha_1$ and $Pl(\phi_1) = \beta_1$) and $\phi_{\phi_1 \rightarrow \phi_2}$ (with $Bel(\phi_{\phi_1 \rightarrow \phi_2}) = \alpha_R$ and $Pl(\phi_{\phi_1 \rightarrow \phi_2}) = (1 - \beta_R)$), the corresponding model of modus ponens is $\phi_1 \odot \phi_{\phi_1 \rightarrow \phi_2} = \phi_2$ with $Bel(\phi_2) = \alpha_1 \cdot \alpha_R$ and $Pl(\phi_2) = (1 - \beta_R)$.

Measuring Uncertainty:

We will use the "uncertainty measure" λ discussed in (Williams, Núñez, et al., 2014) to compare the uncertainties associated with formulae ϕ and their respective confidence intervals $[\alpha, \beta]$:

$$\lambda(\alpha, \beta) = 1 + \frac{\beta}{\gamma} log_2 \frac{\beta}{\gamma} + \frac{1-\alpha}{\gamma} log_2 \frac{1-\alpha}{\gamma} \tag{6.1}$$

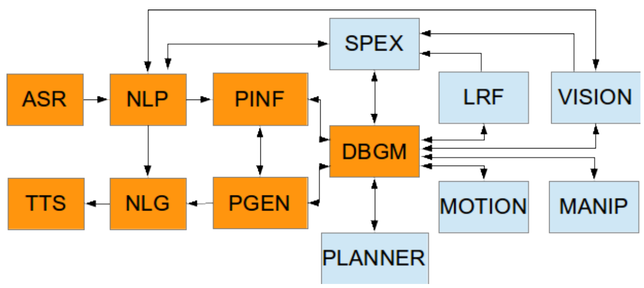
$$\text{where } \gamma = 1 + \beta - \alpha.$$

Here, ϕ is deemed more uncertain as $\lambda(\alpha, \beta) \to 0$. We introduce an "uncertainty threshold" Λ (set to 0.1) where utterances with $\lambda(\alpha, \beta) < \Lambda$ will require clarification from an interlocutor.

6.4.2 Algorithm and Architecture

A cognitive robotic architecture capable of going beyond direct command-based instructions needs several high-level components in addition to typical NL components (such as speech recognizers, parsers, etc.) that work in concert to extract intended meanings. Figure 6.7 depicts how these new components are integrated into *DIARC*.

Figure 6.7: Pragmatic Reasoning Architectural Diagram



Partial architecture diagram. Highlighted are the components that form the natural language pipeline: Automatic Speech Recognition (ASR), Natural Language Processing (NLP) Pragmatic Inference (PINF), and Dialogue, Belief and Goal Management (DBGM). Also shown are language generation components that will be covered in the next chapter (Text-to-Speech (TTS), Natural Language Generation (NLG), and Pragmatic Generation (PGEN)), as well as relevant components that interact with the DBGM: the SPatial EXpert (SPEX), Task Planner (PLANNER), Motion Planner (MOTION), Manipulation (MANIP), Laser Range Finder (LRF), and Vision (VISION).

When an interlocutor speaks to the robot, speech is processed via the standard NL pipeline (speech recognizer, syntactic and semantics parser) resulting in candidate semantic expressions ϕ, each with its own uncertainty interval $[\alpha, \beta]$ attached. While a typical command-based system (e.g. Dzifcak, Scheutz, Baral, & Schermerhorn, 2009) would attempt to act on the semantic interpretation with the highest confidence (and fail if it is not actionable), in the proposed architecture semantic representations are further processed in a pragmatic inference component, which attempts to apply modulatory pragmatic rules to utterance and semantic representations to infer the intentions of the speaker.

The semantic interpretation is passed to our new component for Pragmatic Inference (PINF), which uses contextual and general knowledge to determine the *intention* underlying the literal semantics. By using pragmatic rules indexed by utterance and context, PINF can determine, for example, that asking if one knows the time should be interpreted not as a "yes-no question", but as an indication that the speaker would like to be told what time it is. The resulting intention or intentions can then be returned to *DIARC's* 'Dialogue, Belief, and Goal Manager' (DBGM), which is responsible for dialogue management (Briggs & Scheutz, 2012), storing beliefs in its knowledge base, performing inference on those beliefs, tracking and managing goals (Brick, Schermerhorn, & Scheutz, 2007; Scheutz & Schermerhorn, 2009), and determining what actions to take in pursuit of its goals[4]. Next, we will provide the details for the core PINF algorithm.

The goal of pragmatic analysis is to infer intentions based on (1) the semantics of incoming utterances, (2) the robot's current context, and (3) the robot's general knowledge. This process is depicted in Algorithm 13, which takes three parameters: (1) a BoE of candidate utterances $\{\Theta_U, m_u\}$ provided by NLP, (2) a BoE of relevant contextual items $\{\Theta_C, m_c\}$ provided by the DBGM, and (3) a table of pragmatic rules R. Each rule $r_{uc \to i}$ in R is indexed by an utterance u and a set of contextual items c, and dictates the mass assigned to $Bel(i)$ and $Pl(i)$ when the robot believes that utterance u was heard and that contextual items c are true. Here i is a logical formula representing the intention the interlocutor was expressing through utterance u. When these contextual items involve the shared context of the robot and its interlocutor, they are couched in terms of the *interlocutor's beliefs*. This is critical, as the intentions of the robot's interlocutor are dependent not on the robot's beliefs, but on his or her own beliefs. This allows the robot to correctly interpret its interlocutor's intentions when cognizant of discrepancies between its own beliefs and its interlocutor's beliefs, and to identify information of which its interlocutor may want to be informed. This is important for both pragmatic inference and generation, as this paradigm implicitly assumes that the robot's interlocutor communicates according to the same table of rules known to the robot (however, it is straightforward to keep separate rule tables for individual interlocutors if required).

When an utterance is heard, each rule $r_{uc \to i} \in R$ is examined (line 5), and m_{uc} is determined by performing $m_u \otimes m_c$ (line 6), where m_u specifies the degree to which utterance u is believed to be heard, and m_c specifies the

[4]In truth, this is comprised of three components: the Dialogue Manager, the Belief Manager, and the Goal Manager. However, for the sake of this section, it is simpler to consider them as a single component with multiple responsibilities.

degree to which each of the rule's associated contextual items is believed to be true. DS-based modus ponens is then used to obtain m_i from $m_{uc\to i}$ and m_{uc} (line 6).

Algorithm 13 getIntendedMeaning($\{\Theta_\mathrm{U}, m_\mathrm{u}\}$,$\{\Theta_\mathrm{C}, m_\mathrm{c}\}$,$R$)

1: $\{\Theta_U, m_u\}$: BoE of candidate utterances
2: $\{\Theta_C, m_c\}$: BoE of relevant contextual items
3: R: Currently applicable rules
4: $S = \emptyset$
5: **for all** $r \in R$ **do**
6: $\quad S = S \cup \{(m_u \otimes m_c) \odot m_{r=uc\to i}\}$
7: **end for**
8: $G = group(S)$
9: $\psi = \emptyset$
10: **for all** group $g_i \in G$ **do**
11: $\quad \psi = \psi \cup \{\bigcap_{j=0}^{|g_i|} g_{i_j}\}$
12: **end for**
13: **return** ψ

While previous approaches (e.g Briggs & Scheutz, 2013) look for a single applicable rule in order to produce a single likely intention, we instead consider all applicable rules. This is particularly important for complex contexts or abstract context specifications, where multiple rules might be applicable. Moreover, the robot might have rules that apply to its particular context as well as to a more general context, and it may be more appropriate to consider the combined implicatures of all applicable rules rather than only considering, for example, the most specific applicable rule. Since we may consider multiple rules, multiple intentions may be produced. And multiple rules may also produce the same intentions, possibly with different levels of belief or disbelief. To be able to generate a set of *unique* intentions implied by utterance u after considering all applicable pragmatic rules, we thus group intentions that have the same semantic content but different mass assignments (line 8) and use Yager's rule of combination (line 11) to fuse each group of identical intentions, adding the resulting fused intention to set ψ. This set then represents the set of intentions implied by utterance u and is returned to the DBGM.

It is important to note that, at least in the way that the algorithm is currently used in our architecture, the set of pragmatic rules provided to the pragmatics system must include rules to infer both non-literal *and literal* interpretations. This means, for example, that if a robot is provided with rules to interpret *Can you* utterances indirectly, it must also be provided with rules to infer the direct meaning if it is to be possible to inquire about the robot's capabilities.

6.5 Evaluation

In this section we present an evaluation of our algorithm and demonstrate the capabilities facilitated by our approach. The evaluation of a system at this stage of the natural language pipeline is difficult, as the performance of the algorithm is tightly coupled with the performance of components that precede it in the natural language pipeline (e.g., speech recognition, parsing, semantic analysis). Because of this, we take the same approach to evaluation as previous work, i.e., through a case study that demonstrates the behavior of our algorithm. We will now show how our algorithm works towards the capabilities necessary for robust understanding of conventionalized ISAs.

6.5.1 Handling Uncertainty

Consider a robot conversing with interlocutor Jim. Suppose Jim says to the robot: "I need coffee." From the robot's perspective, this is represented as the utterance $Stmt(jim, self, needs(jim, coffee))$. In this representation, the first argument represents the speaker of the utterance, in this case, Jim; the second argument represents the receiver, in this case, the robot ("self"); the last argument refers to the conveyed message, in this case, *Jim needs coffee*. Suppose the robot knows two pragmatic rules. In both rules, $[\alpha_{R_i}, \beta_{R_i}]$ represents the belief and plausibility of rule i.

First, if B believes A is a barista, then telling A that B needs coffee indicates that B wants A to believe that B wants A to get them coffee.

$$r^0_{[\alpha_{R_0},\beta_{R_0}]} = \frac{\begin{array}{c}(Context : believe(B, barista(A)))\\(Utterance : Stmt(B, A, need(A, coffee)))\end{array}}{(Intention : want(B, believe(A, want(B, get\_for(A, B, coffee)))))}$$

Second, if B believes C is thirst quenching, telling A that B needs C indicates that B wants A to believe that B is thirsty.

$$r^1_{[\alpha_{R_1},\beta_{R_1}]} = \frac{\begin{array}{c}(Context : believe(B, quenches(C, thirst)))\\(Utterance : Stmt(B, A, need(A, C)))\end{array}}{(Intention : want(B, believe(A, thirsty(B))))}$$

Our approach affords the first capability of an ideal system, i.e., the ability to handle uncertain contextual and dialogical information, and to recognize and reason about one's own ignorance. To demonstrate this, suppose the robot strongly believes the following:

(a) Jim believes coffee is thirst quenching:
$bel(jim, quenches(coffee, thirst))[1.0, 1.0]$,
(b) Jim just said he needs a coffee:
$Stmt(jim, self, need(jim, coffee))[0.9, 0.9]$, and
(c) Jim may or may not think the robot is a barista:
$bel(jim, barista(self)[\alpha_b, \beta_b]$.

Applying rules r^0 and r^1 will produce a BBA containing two consequents:

$$c_0 = want(jim, bel(self, want(jim, get\_for(self, jim, coffee))))$$
$$c_1 = want(jim, bel(self, thirsty(jim))).$$

The degree to which c_0 and c_1 are believed will depend on whether Núñezor Tang's fusion operators are used. The following table shows how belief in the two consequents changes depending on which set of fusion operators is used, the degree to which the robot believes the interlocutor believes the robot is a Barista (b), and the degree to which the robot believes the two rules r^0, r^1 hold.

Our approach can thus modulate its interpretation of utterances based on the certainty of the relevant utterance, contextual factors, and pragmatic rules. However, an ideal system should also explicitly reason about its own ignorance. Since we are using a DS-theoretic approach, we can use the consequents' uncertainty intervals to determine whether or not the robot needs to ask for clarification. Specifically, we use the ambiguity measure defined by Núñez et al. (2013) (Equation 6.1).

For example, for the interval $[0.6, 0.9]$,

Table 6.6: Comparison of operators under Tang and Núñez

	$b[\alpha,\beta]$	$r^0[\alpha,\beta]$	$r^1[\alpha,\beta]$	Fusion	$c_0[\alpha,\beta]$	$c_1[\alpha,\beta]$	λ_0	λ_1
1	[0.9,0.9]	[0.85,0.9]	[0.7,0.85]	Núñez	[0.85,0.90]	[0.70,0.85]	0.41	0.17
2	[0.9,0.9]	[0.85,0.9]	[0.7,0.85]	Tang	[0.69,0.90]	[0.63,0.85]	0.18	0.11
3	[0.1,0.1]	[0.1,0.1]	[0.5,0.5]	Núñez	N/A	[0.50,0.50]	N/A	0.00
4	[0.1,0.1]	[0.1,0.1]	[0.5,0.5]	Tang	[0.01,0.10]	[0.05,0.50]	0.56	0.07
5	[0.5,0.5]	[0.1,0.5]	[0.5,0.5]	Núñez	N/A	N/A	N/A	N/A
6	[0.5,0.5]	[0.1,0.5]	[0.5,0.5]	Tang	[0.05,0.5]	[0.45,0.5]	0.07	0.002
7	[0.002,0.002]	[0.99,0.99]	[0.99,0.99]	Núñez	[0.99,0.99]	[0.99,0.99]	0.92	0.92
8	[0.002,0.002]	[0.99,0.99]	[0.99,0.99]	Tang	[0.002,0.99]	[0.80,0.99]	0.00001	0.35

$$\lambda = 1 + \frac{0.9}{1.3} log_2 \frac{0.9}{1.3} + \frac{0.4}{1.3} log_2 \frac{0.4}{1.3} = 0.11.$$

$\lambda \to 0$ as uncertainty grows and as α and $1-\beta$ grow closer together. Using this equation, we generate a clarification request if $\lambda \leq 0.1$. This makes use of information that is unavailable to the Bayesian approach. As shown in Table 6.6, the results of the DS-theoretic approach are greatly dependent on which set of fusion operators is used. We will now briefly compare Tang and Núñez' fusion operators before discussing the other capabilities afforded by our approach.

One will notice that there are several cases in which Núñez' fusion operators do not return results. In fact, since the last fusion operation applied in our algorithm is Modus Ponens, using Núñez' fusion operators produces either no result, $[0, \beta_R], or [\alpha_R, \beta_R]$. While this ensures consistency with classical logic, it also means that the antecedent is largely ignored: the only effect of context is to cause no result to be returned if belief in the context is less than belief in the rule. This can lead to unwarranted confidence. Consider cases 7 and 8 in the table above. Here, the robot is highly confident in rule r^0, and is highly confident that its antecedent is false. Tang's logical operators report near total uncertainty, while Núñez' logical operators report near total certainty. This is a highly problematic result: the more confident a robot is in a rule, the less evidence it will need to be convinced in the rule's antecedent, even in the face of overwhelming evidence to the contrary. In light of this, we recommend the use of Tang's logical operators. While using Tang's operators avoids the aforementioned problems with Núñez' operators, we acknowledge that our use of them falls outside the scope of their intended use: Tang's operators were originally designed for use within an argumentation-theoretic context, and we are thus currently working on an argumentation-theoretic approach that allows Tang's operators to perform as intended; this exploration is outside the scope of this dissertation.

6.5.2 Adaptation

The second capability of an ideal system is the ability to adapt old rules and learn new ones. We currently assume that the initial beliefs and plausibilities of our rules and contextual items are given, but we do allow rules to be adapted based on user feedback. Upon receiving a corrected rule from a user, it is compared against all currently applicable rules[5]. Those whose antecedents and consequents are on the same frames as the antecedents and consequents of the new rule may be updated using the Conditional Update Equation (CUE) as defined by Wickramarathne, Premaratne, & Murthi (2012). For example, if rule r^i is on interval $[0.8, 0.8]$, and a correction states that in the current context, $[\alpha_{R_i}, \beta_{R_i}]^i$ should be $[0.5, 0.9]$, the CUE will update the rule's uncertainty to $r^i_{[0.53,1.0]}$ (a substantial increase in uncertainty). Although the proposed approach only allows for adaptation of rules, it could easily be extended to allow for the addition of new rules, which would initially have very high levels of uncertainty and would become less uncertain with exposure to applications of the rule.

We thus demonstrate that within our framework it is, at least in principle, possible to effect adaptations based on user feedback. However, the presented adaptation mechanisms have not been systematically integrated into our architectural framework, and we have not demonstrated how *new* rules would actually be learned, how old rules would be adapted when user feedback suggests a modification of an existing rule, how feedback warranting adaptation would actually be identified, how feedback or rules could be directly solicited, or how data-driven methods could be used to infer the need for rule learning or adaptation. These will all be important directions for future work.

[5]We would like to stress again that our architecture maintains a set of rules that are applicable in the current context, and that our ISA understanding algorithm only considers these rules for the sake of efficiency. In the future it would be interesting to investigate ways of probabilistically maintaining a more limited set of rules or for quickly retrieving a small set of appropriate rules when needed, in order to further reduce the complexity of inference.

6.5.3 Belief Modeling

The third capability of an ideal system is the ability to reason about other agents' beliefs. Rules such as r^0 and r^1 are formulated in terms of the *interlocutor's* beliefs; to determine what interlocutor J is trying to communicate, J's utterances must be evaluated in the context of J's beliefs. For example, if J says he needs coffee, the likelihood that he is trying to order a coffee should be modulated not by the *robot's* belief that it is a barista, but instead by J's beliefs; if J has no reason to think the robot is a barista, his statement should not be viewed as a coffee order even if the robot has barista training. Belief modeling also allows natural representation of interlocutors' beliefs about the robot's abilities and social roles. For example, the robot may need general rules (e.g., Equation 6.2) that suggest that a statement such as "I need a coffee" is only an indirect request if its interlocutor believes the robot to be able and obligated to get them coffee.

$$\frac{\begin{array}{c}(Context : bel(B, obligated(A, give(A, B, C))))\\ (Utterance : Stmt(B, A, would\_like(B, C)))\end{array}}{(Intention : want(B, bel(A, want(B, give(A, B, C)))))} \tag{6.2}$$

6.6 General Discussion

In this section, we have presented philosophical and empirical evidence motivating the need for indirect speech act understanding algorithm development in the field of human-robot interaction. We then presented a novel algorithm for understanding conventionalized indirect speech acts, and demonstrated how it satisfies several requirements we believe necessary for use in realistic human-robot interaction scenarios. The presented algorithm is not intended as an *alternative* to plan-based inferential approaches for understanding non-conventionalized ISAs, but rather as a *complement* to such approaches. Ideally, this approach would be integrated into a hybrid framework in order to efficiently handle conventionalized ISAs while still being able to fall back on plan reasoning in the case of non-conventionalized ISAs.

Just as we showed that the referential processing framework presented in Chapter 3 was useful not only for referring expression understanding, but for referring expression generation as well (in Chapter 5), we will similarly demonstrate in the next chapter that the pragmatic reasoning framework we have presented in this chapter can be used not only for pragmatic understanding, but for pragmatic generation as well.

Chapter 7

Pragmatic Generation

In the previous chapter, we introduced novel algorithms based on Dempster-Shafer (DS) theory (Shafer, 1976) for inferring intentions I from utterances U in contexts C. In this chapter, we present novel algorithms for generating *utterances* U from intentions I in contexts C, and show how these can be used to generate requests to disambiguate intended meanings and intended referents. We then demonstrate the operation of the algorithms in a detailed example showing how uncertainty is propagated at each stage of processing and can lead to different responses by the robot. Finally, in Section 7.3, I present experimental evidence motivating future work on the application of mechanisms such as those presented in this chapter to *robot-robot* communication.

7.1 A Dempster-Shafer Theoretic Approach to Pragmatic Generation

When a robot needs to communicate information, it must choose appropriate surface realizations of the semantic facts in intends to convey. However, for reasons of social convention such as politeness, it may be inappropriate to express semantic facts in the most direct manner. For example, one may find it rude if the robot were to say "I want to know what time it is. Tell me now." To allow the robot to generate socially acceptable utterances based on pragmatic considerations, we introduce an abductive inference algorithm called *pragmatic generation*, which, much like pragmatic inference, uses the robot's current context and its set of pragmatic rules to determine the best utterance to communicate intentions. The "best" utterance is determined to be the utterance that, according to the robot's set of pragmatic rules, would

Algorithm 14 getSemantics($\{\Theta_{\mathrm{I}},m_{\mathrm{i}}\}$,$\{\Theta_{\mathrm{C}},m_{\mathrm{c}}\}$,$R$)

1: $\{\Theta_i, m_i\}$: BoE of candidate intentions
2: $\{\Theta_C, m_c\}$: BoE of relevant contextual items
3: R: Currently applicable rules
4: $S = \emptyset$
5: **for all** $r \in R$ **do**
6: $\quad u = (m_i \otimes m_c) \odot m_{r=uc\rightarrow i}$
7: $\quad$ **for all** $(b_s, b_v) \in getBindings(u)$ **do**
8: $\quad\quad$ **if** $marked(b_v)$ **then**
9: $\quad\quad\quad u = adapt(u, getSemantics(buildBoE(b_s), \{\Theta_C,m_c\}, R))$
10: $\quad\quad$ **end if**
11: $\quad$ **end for**
12: $\quad u' = checkEffects(\ getIntendedMeaning(\{\Theta_U,m_u\}, \{\Theta_C,m_c\}, R))$
13: $\quad S = S \cup u'$
14: **end for**
15: **return** S

be most likely to communicate the given intention properly (e.g., without communicating any other information that the robot does not believe to be true). A DS-based approach is particularly useful here, because rule-based pragmatic inferences are determined by equations that relate the premise and rule to the consequent and can thus, exactly because they are equations, be used for inferences in both directions, deductive and abductive. We can thus infer the best utterance to convey a given intention in a given context from the same rules we use for inferring the best intention given an utterance in the same context. Moreover, we can perform pragmatic generation recursively: if a pragmatic rule matches the high-level structure of an utterance, it may be necessary to further abduce the best way to phrase individual clauses of the utterance that were left open by the high-level rule.

As with pragmatic inference, the pragmatic generation algorithm (see Algorithm 14) takes the robot's current context $\{\Theta_C,m_c\}$ and the set of currently applicable rules R. Instead of the BoE of possible incoming utterances $\{\Theta_U,m_u\}$, the algorithm takes a BoE of possible intentions desired to be communicated $\{\Theta_I,m_i\}$, as determined by the DBGM. For each applicable rule $r_{uc\rightarrow i} \in R$, the algorithm performs an uncertain modus ponens operation producing a BoE indicating which utterance would most likely generate the desired intention according to rule r (line 6).

The algorithm then examines the structure of the resulting utterance (line 7) to determine whether it should recurse on subsections of the utterance, recursing on the semantics b_s associated with each variable b_v marked as suitable for recursion. For example, for the utterance $Want(self, Know(self, or(X, Y)))$, it may be necessary to recurse on the portions of the utterance bound to X and Y. Once the results of any such recursions (line 9) are integrated into the representation of the utterance to communicate u, the set of intentions ψ that would be implied by utterance u are calculated (on line 12) by calling $getIntendedMeaning(\{\Theta_U, m_u\}, \{\Theta_C, m_c\}, R)$ (i.e., Algorithm 13) with the candidate utterance and the current context and rule set. The belief and plausibility of u are then modulated by $Bel(p_i)$ and $Pl(p_i)$ for $p_i \in \psi$. This prevents the robot from accidentally communicating some proposition that it does not actually believe to be true. Finally, the set of candidate utterances S is returned, from which an utterance is chosen to communicate, e.g., by choosing the candidate with the highest degree of belief.

In this section we have discussed the basics behind our pragmatic generation algorithm. In the next section we will discuss an important use for this algorithm: generating clarification requests.

7.2 Generating Clarification Requests

Imagine a robot named Cindy and a human named Bob. Cindy and Bob are working together in a disaster relief scenario, and have just left a kitchen containing two medical kits: one on a table, and one on a counter. After driving down the hallway for a few minutes, Bob turns to Cindy and asks "Can you go back to the kitchen and grab the medical kit?" In order for Cindy to successfully fulfill Bob's request, Cindy must resolve both the *pragmatic* and *referential ambiguity* inherent in his question. Bob's request is *pragmatically ambiguous* as it could be interpreted both *directly*, i.e., as a literal question as to Cindy's abilities, or *indirectly*, i.e., as a command to Cindy. Bob's request is *referentially ambiguous* because it could refer to either the medical kit on the table or the medical kit on the counter.

When humans are confronted with this sort of ambiguity, they typically resolve it using *clarification requests* such as "Do you want me to retrieve the medical kit that is on the counter or the medical kit that is on the table?" (Tenbrink, Ross, Thomas, Dethlefs, & Andonova, 2010). In this work, we seek to endow robots with this capability as well: We begin by discussing previous work on clarification request generation in human-robot interaction contexts. Then, we present a clarification request generation framework tailored to human-robot interaction scenarios. Next, we present the results of a human-subjects experiment in which previous findings regarding human preferences with respect to robot clarification request formulation are replicated and refined. Then, we present an approach to clarification request generation designed to align with human preferences. Next, we present a proof-of-concept demonstration of our approach, and evaluate our approach through human-subject experimentation. Finally, we discuss possible directions for future work.

7.2.1 Related Work

Clarification request generation has attracted a large amount of research overall(DeVault, Kariaeva, Kothari, Oved, & Stone, 2005; DeVault & Stone, 2007; Purver, Ginzburg, & Healey, 2003; Traum, 1994), but relatively little in *situated* contexts such as human-robot interaction. Recently, some researchers have used information-theoretic techniques to identify random variables which could have their entropy reduced if asked about. In such work, clarification requests have taken the form of yes/no questions about the properties of an object(Deits, Tellex, Kollar, & Roy, 2013; S. Hemachandra, Walter, & Teller, 2014; Purver, 2004) or as open-ended specification requests (e.g., "What do the words X refer to?")(Purver, 2004; Tellex et al., 2013).

Recent experimental evidence, however, suggests this may not be the right approach to take(Marge & Rudnicky, 2015). This evidence suggests that in human-robot interaction contexts, people prefer robots to list multiple options rather than asking for confirmation about a single referent with a yes/no question (cf. H. H. Clark, 1996). This is particularly striking as the evidence suggests that people maintain this preference even when a yes/no question would be more efficient (cf. S. Hemachandra, Walter, & Teller, 2014).

In contrast, Kruijff et al. present an approach in which robots can generate multiple-option clarification requests such as "Do you mean the blue or the red mug, Anne?" through a *continual planning* approach(Kruijff, Brenner, & Hawes, 2008). This approach, however, does not appear to be able to account for social context, uncertainty, or ignorance, and is only used for generation. The ability to handle social context is crucial for enabling natural human-robot interactions, and typical human-robot interaction scenarios are plagued by uncertainty and ignorance. An eldercare robot, for example, is not likely to be familiar with every object in the home of the elder it is assisting, nor is the robot likely to be familiar with every person who they might refer to. Furthermore, even for the objects and people that the robot *does* know of, it is unlikely to have uncertainty-free knowledge of all of the properties and relations involving those objects and people. We desire an approach that is able to account for these missing factors, and which can be used for both generation *and understanding*.

There has also been much previous work in developing general natural language generation (NLG) systems. For example, Reiter et al. present an NLG framework comprised of six stages: content determination, document structuring, aggregation, lexical choice, referring expression generation, and realization(Reiter, Dale, & Feng, 2000). It is unclear, however, whether such frameworks are well suited to *situated contexts*. In human-robot interactions, for example, NLG is often performed to *solicit* information, whereas in non-situated contexts it is more typically performed to *provide* information. For this reason, we propose the following alternate framework designed specifically for clarification request generation in human-robot interactions.

7.2.2 A Framework for Clarification Request Generation

We identify five stages necessary for successful clarification request generation, as shown in Figure7.1: (1) uncertainty identification, (2) decision to communicate, (3) utterance choice, (4) surface realization, and (5) speech synthesis. In this section we describe the actions necessary at each stage.

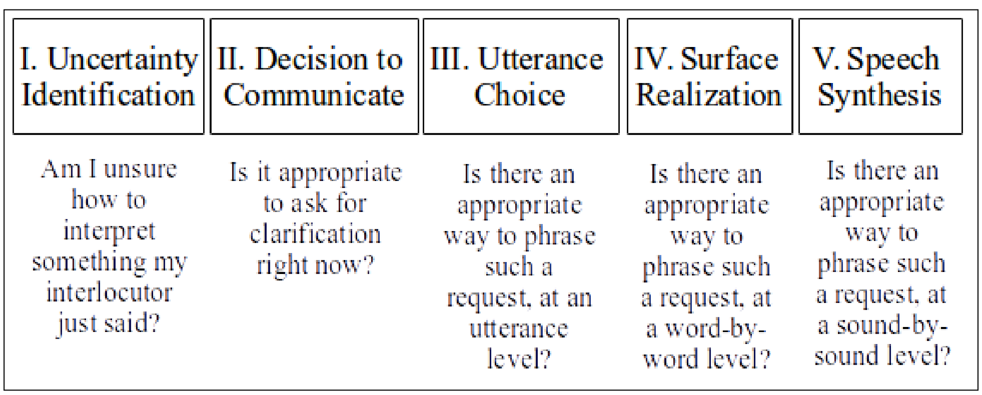


Figure 7.1: *Clarification Framework.*

Uncertainty Identification

Suppose that in our original example, Bob had asked Cindy "Can you grab the medkit?" During the stage of *uncertainty identification*, Cindy must determine if she is unsure how to interpret any part of this utterance. This may be uncertainty as to what entities are being *referenced*, e.g., *which* medkit Bob is referring to, or uncertainty as to the speaker's *intentions*, e.g., whether Bob wishes Cindy to bring him the medkit or whether he meant something else by the utterance. Furthermore, this uncertainty may take different forms (e.g. Stirling, 2010): the utterance may be *ambiguous* (e.g., if Cindy knows of multiple medkits) or the utterance may reveal *ignorance* (e.g., if Cindy knows of no medkits, or is unsure whether a particular object qualifies as a "medkit").

Decision to Communicate

If a robot has identified a point in need of clarification, it must decide whether it would be appropriate to actually ask for clarification. This decision will depend on a variety of factors: Is it permissible for the robot to ask for clarification[1]? Is the robot's interlocutor likely to be able to provide clarification? Would obtaining clarification really be the highest utility action at the current time (compared to, e.g., exploration)? For example, if Cindy determines there are actually two medkits that Bob could be referring to, but while coming to this decision Bob has already engaged another teammate in conversation, it may be necessary for Cindy to wait until this conversation finishes before asking for clarification. Alternatively, it may be the case that obtaining clarification for the point in question is of lower expected utility than obtaining clarification on some other point, obtaining clarification in some other way (e.g., exploration), or simply performing some other action.

Utterance Choice

Once a robot has decided to request clarification on a particular point, it must determine what utterance form to use to communicate its request: depending on the relationship between the robot and its interlocutor, and the obligations of each party, certain utterance forms may be more or less appropriate (Brown, 1987). For example, if Cindy is Bob's subordinate, it may be more appropriate to use an *indirect request* such as "Which medkit would you like?", whereas if Cindy is Bob's superior, it may be more appropriate to use a *direct request* such as "Tell me which medkit you would like."

[1]Cf. work from Traum & Allen (1994) in which dialogue moves are made based on *obligations* introduced by others' utterances.

Surface Realization

Once a robot chooses an utterance form to use, it must determine what words to use (Garoufi & Koller, 2014; Stone, 2003). For example, if Cindy decides to use an utterance of the form “Would you like [medkit$_1$]”, she must choose how to actually describe medkit$_1$, e.g., by referring to it as “the medkit in the kitchen” or perhaps as “the white medkit”. If one medkit is in front of Cindy, it may be more appropriate to point to it and use a deictic expression such as “this medkit.”

Speech Synthesis

Finally, once a robot determines what word to use, it must synthesize an appropriate sound pattern.

7.2.3 Experimental Motivations

In developing a new HRI-oriented approach to clarification request generation, our primary goal is to account for the factors missing from previous, non-HRI-oriented approaches. But we believe it is equally important to take *human preferences* into account as part of the design process. We believe that the previous work discussed thus far has not adequately considered what type of utterances humans *prefer* to use and be used.

Specifically, there are three categories of human preferences that we hypothesize humans will hold, and believe that these preferences should affect the design decisions made when developing HRI-oriented clarification request generation algorithms.

Design Hypotheses

1. Presentation of Options

   Marge & Rudnicky (2015)'s research suggests that people prefer that robots list options rather than ask yes/no questions. But clearly there are limits to this preference. If a robot is asked “Could you get me some ice cream?” It is unlikely that humans will prefer a robot that lists off all twenty-seven available flavors instead of just asking “Which flavor would you like?” We hypothesize **(H1)** that humans prefer options to be listed *only for a limited number of options.*

2. Demonstration of Intention Understanding

Similarly, many of the previous approaches have used clarification requests that do not indicate understanding of the *meaning* of the sentence. If a robot is asked "Could you get me some ice cream", a robot that replies "What do the words 'ice cream' refer to" or "Do you mean 'the chocolate ice cream' or 'the vanilla ice cream"' does not allow its interlocutor to discern whether their *intention* was understood (i.e., that they want ice cream *brought to them*), relative to utterances such as: " *Would you like me to get you* the chocolate ice cream or the vanilla ice cream?" In human-human dialogue, it is generally expected that speakers will use utterances that serve to create and maintain common ground with their conversational partners (see also Section 3.1.3). This "grounding" process by which conversational partners stay on the same page (H. H. Clark & Brennan, 1991) has been studied and exploited in many human-robot interaction studies and scenarios (e.g. Kiesler, 2005; Mutlu, Terrell, & Huang, 2013; Stubbs, Hinds, & Wettergreen, 2007). We hypothesize **(H2)** that humans prefer clarification requests that clearly demonstrate that their conversational partner has understood their intentions.

3. Pragmatic Appropriateness

   And finally, a robot that *does* generate clarification requests that reflect their understanding of human intentions will almost certainly need to use *indirect speech acts*(Searle, 1975) to create such clarification requests (e.g., *Would you like* me to get you the chocolate ice cream or would you like me to get you the vanilla ice cream?), as the direct alternatives (e.g., "I have an intention to know whether you want me to have a goal to bring you chocolate ice cream or a goal to bring you vanilla ice cream") are hard to express without being overly verbose. Humans may, paradoxically, have difficulty quickly inferring the intentions behind direct, explicit utterances such as these, both because they are so far removed from the more natural forms used in typical human-human dialogue, but also because they likely will fail to conform with the conceptual pacts humans will likely expect robots to follow within the local dialogue due to principles of lexical entrainment (see also Section 3.1.3). Furthermore, indirect phrasing is commonly used by humans in order to achieve various socio-cultural goals (e.g. politeness); much previous research has shown, as one might expect, that humans prefer robots that are polite (Castro-González et al., 2016; Dautenhahn et al., 2005; Nomura & Saeki, 2009; Salem, Ziadee, & Sakr, 2013, 2014) (see also Section 6.2). We hypothesize **(H3)** that humans prefer clarification requests that are indirectly rather than directly phrased, and that are thus more pragmatically appropriate.

We will now present the results of a human subjects experiment designed to test our three hypotheses.

Methodology

Participants were recruited using Amazon Mechanical Turk (20 Male, 10 Female, mean age 32.67). Each participant was asked seven simple questions, presented in a randomized order. Participants were told to imagine that they have commanded a robot to "Pick up the mug" in a scenario in which there are several mugs on a table. Each question then differed in the number of mugs (of different colors) that were on the table, and how the robot chose to ask for clarification; for each question, two ways of asking for clarification were presented, and participants were asked to indicate which of the two options they would prefer the robot to use.

The first five questions evaluated our first hypotheses: in each case, participants chose between an option that listed out all options (ranging from "Would you like the red mug or the orange mug?" to "Would you like the red mug or the orange mug or the yellow mug or the green mug or the blue mug or the purple mug?") and a catch-all option ("Which mug would you like?").

The sixth question evaluated our second hypothesis: participants chose between an option that indicated understanding of the speaker's goals ("Would you like the red mug or the green mug?") and an option that did not ("Do you mean the red mug or the green mug?").

The last question evaluated our third hypothesis: participants chose between a pragmatically appropriate option ("Would you like the red mug or the blue mug?") and a pragmatically inappropriate option ("I have an intention to know if you want me to have a goal to bring you the red mug or the blue mug.").

Results

We will now examine the results of this experiment, and the preferences those results suggest.

1. Presentation of Options

   Our results show that 70% of participants preferred options to be listed when there were only two options. But for more than two options, this number rapidly shrank. Only 20% of participants preferred options to be listed when there were three options, and preference for listing all options fell between 10 and 13% when more options were listed. This confirms but clarifies the previous findings of Marge & Rudnicky (2015), and suggests that robots likely do not need mechanisms for listing more than two options when there is referential ambiguity **(H1)**.

2. Demonstration of Intention Understanding

   Our results show that 80% of participants preferred the option that indicated understanding of their goals, supporting our second hypothesis **(H2)**.

3. Pragmatic Appropriateness

   Our results show that 93% of participants preferred the pragmatically appropriate option, supporting our third hypothesis**(H3)**.

Discussion

The results of this experiment suggest three design recommendations: **(D1)** When phrasing clarification requests, robots should not present all clarification options to their interlocutors unless there are only two such options; **(D2)** When phrasing clarification requests, robots should use phrasings that indicate that they understand the goals of their interlocutors; and **(D3)** When phrasing clarification requests, robots should use pragmatically appropriate phrasings.

7.2.4 Generating Clarification Requests to Resolve Intentional Ambiguity

In this section, we will demonstrate how the previously presented algorithms for pragmatic understanding and pragmatic generation can be used within the presented clarification request generation framework in order to fulfill the design recommendations suggested by the experiment presented in the previous section.

Approach

Recall that in the previous chapter, pragmatic inference resulted in a set of intentions implied by utterance u, which was returned to *DIARC's* DBGM.

Here, the level of uncertainty associated with these intentions must be assessed: for each intention $i \in \psi$ on uncertainty interval $[\alpha_i, \beta_i]$, a clarification request is generated if $\lambda(\alpha_i, \beta_i) < \Lambda$, where λ is the uncertainty measure originally presented in Section 6.4.1, and presented again here for clarity, and Λ is some threshold (e.g., 0.1).

$$\lambda(\alpha, \beta) = 1 + \frac{\beta}{\gamma} log_2 \frac{\beta}{\gamma} + \frac{1-\alpha}{\gamma} log_2 \frac{1-\alpha}{\gamma}$$
$$\text{where } \gamma = 1 + \beta - \alpha.$$

For example, consider a scenario in which the robot is unsure which of two contexts it is in. In the first context, a particular statement should be interpreted as a request for information, and in the second context, it should be interpreted as an instruction. In this case, the robot will ask "Should I <perform the intended action> or would you like to know <the intended information>?" This demonstrates the ability for the robot to exploit propagated uncertainty to identify and resolve uncertainties and ambiguities.

A similar process can also be employed directly *before* pragmatic inference: after NLP produces set of surface semantics s, those semantic interpretations are analyzed using the λ ambiguity measure. If $\lambda(\alpha_P, \beta_P) < \Lambda$ for semantic predicate p with uncertainty interval $[\alpha_P, \beta_P]$, a request to verify what was said is sent to NLG, which generates and communicates a realization of the form "Did you say that <s>" in which case the uncertain semantics are *not* passed on for pragmatic analysis.

Demonstration

Now that we have presented this use case for natural language generation, we are ready to present a proof-of-concept demonstration of our pragmatic generation algorithm, within that use case. To demonstrate the operation of the proposed inference algorithm for natural human-robot interactions, we consider a dialogue interaction that occurs as part of a Search-and-Rescue Task. The interaction starts with an interlocutor ("Jim") telling the robot "Commander Z needs a medical kit." The utterance and semantic representation produced by NLP for this statement is

$Statement(Jim, self, needs(commander\_z, medkit))[\alpha, \beta]$.

We will now examine how the dialogue between Jim and the robot plays out under two different combinations of values for α and β, corresponding with low and high of uncertainty accrued by the early NL components (up to semantic parsing). These two conditions are denoted[2]

U_{low} (with uncertainty interval $[0.95, 1.00]$),
U_{high} (with uncertainty interval $[0.62, 0.96]$).

Furthermore, we will assume three settings that differ with respect to the robot's assumptions regarding its interlocutor's beliefs about who is subordinate to whom. In the first case (denoted C_{jim}), the robot believes that Jim believes that the robot is subordinate to him. In the second case (denoted C_{robot}), the robot believes that Jim believes that he is subordinate to the robot. In the third case (denoted C_{unk}), the robot is unsure who Jim believes to be the subordinate between the pair of them. The differences in these scenarios are reflected in differences in the knowledge base of the robot at the start of the task:

$$
\begin{array}{ll}
C_{jim} & locationof(breakroom, medkit)[0.80, 0.90] \\
 & Believes(Jim, subordinate(self, Jim))[0.80, 0.90] \\
 & Believes(Jim, subordinate(Jim, self))[0.10, 0.20] \\
C_{robot} & locationof(breakroom, medkit)[0.80, 0.90] \\
 & Believes(Jim, subordinate(self, Jim))[0.10, 0.20] \\
 & Believes(Jim, subordinate(Jim, self))[0.80, 0.90] \\
C_{unk} & locationof(breakroom, medkit)[0.80, 0.90] \\
 & Believes(Jim, subordinate(self, Jim))[0.50, 0.60] \\
 & Believes(Jim, subordinate(Jim, self))[0.40, 0.50]
\end{array}
$$

In all conditions, the robot uses the following set of pragmatic rules[3]:

1. $Stmt(A, B, Want(A, bring(C, D, E))) \rightarrow$
   $Goal(C, bring(C, D, E))[0.95, 0.95]$
2. $AskWH(A, B, or(C', D')) \rightarrow$
   $ITK(A, or(C', D'))[0.95, 0.95]$
3. $Stmt(A, B, Want(A, Know(A, C))) \rightarrow$
   $ITK(A, C)[0.85, 0.85]$
4. $Instruct(A, B, C) \rightarrow$
   $Goal(B, C)[0.90, 0.90]$

[2]All beliefs and plausibilities listed in this section are rounded to two decimal places for the reader's convenience.

[3]Here intentions are represented as "Goal" and intentions to know are presented as "ITK(A,B)", e.g., (Perrault & Allen, 1980).

5. if $Bel(A, subordinate(B, A))$:
   $Stmt(A, B, needs(C, D)) \rightarrow$
   $Goal(B, bring(B, D, C))[0.80, 0.90]$
6. if $Bel(A, subordinate(B, A))$:
   $Stmt(A, B, needs(C, D)) \rightarrow$
   $not(ITK(A, locationof(E, D)))[0.80, 0.90]$
7. if $Bel(A, subordinate(A, B))$:
   $Stmt(A, B, needs(C, D)) \rightarrow$
   $ITK(A, locationof(E, D))[0.80, 1.00]$
8. if $Bel(A, subordinate(A, B))$:
   $Stmt(A, B, needs(C, D)) \rightarrow$
   $not(Goal(B, bring(B, D, C)))[0.80, 1.00]$

In both U_{high} and U_{low}, the semantics are passed to PINF, which yields the following intentions for each combination of U and C conditions:

	C_{jim}
U_{low}	$Goal(self, bring(self, medkit, commander\_z))[0.88, 0.95]$
	$ITK(Jim, locationof(X, medkit))[0.05, 0.12]$
U_{high}	$Goal(self, bring(self, medkit, commander\_z))[0.88, 0.93]$
	$ITK(Jim, locationof(X, medkit))[0.07, 0.12]$
	C_{robot}
U_{low}	$Goal(self, bring(self, medkit, commander\_z))[0.05, 0.12]$
	$ITK(Jim, locationof(X, medkit))[0.88, 0.95]$
U_{high}	$Goal(self, bring(self, medkit, commander\_z))[0.07, 0.12]$
	$ITK(Jim, locationof(X, medkit))[0.88, 0.93]$
	C_{unk}
U_{low}	$Goal(self, bring(self, medkit, commander\_z))[0.47, 0.67]$
	$ITK(Jim, locationof(X, medkit))[0.33, 0.54]$
U_{high}	$Goal(self, bring(self, medkit, commander\_z))[0.50, 0.62]$
	$ITK(Jim, locationof(X, medkit))[0.38, 0.50]$

These intentions are then passed to the DBGM, which performs different operations based on the uncertainty condition. In C_{unk}, the high level of uncertainty necessitates a clarification request, so the DBGM forms intention i:

$$ITK(self, or(ITK(Jim, locationof(X, medkit)),$$
$$Goal(self, bring(self, medkit, commander\_z))))[1.0, 1.0].$$

$Bel(i)$ and $Pl(i)$ are both 1.0, since the robot can be sure of its own intentions. Given i, PGEN produces:

$ITK(self, or(Want(Jim, Know(Jim, locationof(X, medkit))),$
$Want(Y, bring(self, medkit, commander\_z))))[0.95, 1.0].$

NLG then translates this intention to "Would you like to know where to find a medkit? or would you like me to bring commander z a medkit?" Suppose Jim responds "I'd like to know where to find one." PINF will produce:

$$\begin{array}{ll} U_{low} & ITK(Jim, locationof(X, medkit))[0.81, 1.0] \\ U_{high} & ITK(Jim, locationof(X, medkit))[0.52, 1.0] \end{array}$$

In U_{high}, the *intention* of the utterance resulting from PINF is deemed too uncertain since $\lambda(0.52, 1.00) < 0.1$, so the robot asks for clarification: "Would you like to know where to find a medkit?" In U_{low}, this intention is not deemed uncertain since $\lambda(0.81, 1.00) > 0.1$, so the intention is instead added to the robot's set of beliefs. This behavior, and the actions that follow, are identical to how the robot responds to the original utterance in scenario C_{robot}. Since Jim has not yet been provided an answer to his question, the robot attempts to answer him. The robot first queries its knowledge base to determine if it knows the answer. If it had not known the location of a medkit, it would have generated a response with the semantics

Stmt(self,Jim,not(Know(self,locationof(X,medkit))))[1.0,1.0].

In this scenario, the robot does know the answer as it has $locationof(breakroom, medkit)[0.80, 0.90]$ in its knowledge base, so it forms an utterance with semantics

Stmt(self,Jim, locationof(breakroom,medkit))[0.8,0.9].

NLG then translates this to "A medkit is located in the breakroom."

Suppose the robot's interlocutor instead responded to the initial clarification request by saying "Bring him one." PINF will produce:

$$\begin{array}{ll} U_{low} & Goal(self, bring(self, medkit, commander\_z))[0.86, 1.0] \\ U_{high} & Goal(self, bring(self, medkit, commander\_z))[0.55, 1.0] \end{array}$$

This intention is not deemed uncertain in either condition[4] so the intention is instead added to the robot's set of beliefs. This behavior, and the actions that follow, are identical to how the robot responds to the original utterance in scenario C_{jim}. The DBGM then determines which action will accomplish the goal and executes that action, setting forth to retrieve the medkit. A video of this interaction in operation on a Willow Garage PR2 robot can be viewed at `https://www.youtube.com/watch?v=wDrz44YyI58`.

The goal of the demonstration example on a real robot in a real-world setting was two-fold. First, we intended to show the potential of the proposed algorithms for making sound deductive and abductive pragmatic inferences based on human utterances and context that go beyond the direct interpretation of command-based instructions. And second, we wanted to demonstrate that the algorithms have been fully integrated into an operational cognitive robotic architecture. Yet, the demonstration is clearly not an evaluation and should not be taken as such. While an evaluation of the integrated system will eventually be critical, we believe that it would be premature at present given that we do not even know how to best evaluate such integrated systems (e.g., how many dialogue-based scenarios would we have to set up and how many pragmatic rules would we have to examine to be able to make a case about how well the system works and how could we be sure that the employed data was sufficient?). Instead, the current system can be seen as a proof-of-concept that the proposed algorithms do not only work in principle and isolation, but in real-time as part of an integrated robotic architecture.

As a next step towards a full evaluation in the future, we are interested in improving several aspects of the current system, including how pragmatic rules can be acquired in a way that does not require the system to learn from large data sets offline. Specifically, we are interested in using NL instructions to learn rules quickly and to use reinforcement methods (based on feedback from human interlocutors) to adapt the uncertainty values associated with the learned rules. This way of allowing for online learning of pragmatic interpretations will enable adaptive trainable systems that can quickly acquire new knowledge on the fly as is often required in human-robot interaction domains.

In addition, it is interesting to note (although not strictly germane to this dissertation) that while the Dempster-Shafer theoretic model $U^C \to I$ presented in this section was applied to indirect speech understanding, it also has applications to other areas as well. Recently, researchers from our

[4] One could argue that the uncertainty in U_{high} is high enough to warrant a clarification request. One may raise Λ to achieve such behavior, if so desired.

lab have taken this model and applied it to problems in affordance reasoning as well (Sarathy & Scheutz, 2015, 2016a,b).

We are also interested in extending PINF with plan reasoning capabilities so that it can better interpret non-idiomatic indirect speech acts, and extending PGEN so that it can use Grice's conversational maxims when choosing which utterance to communicate (e.g., analogous to (Briggs & Scheutz, 2013)).

7.2.5 Generating Clarification Requests to Resolve Referential Ambiguity

In the previous section, we described how our pragmatic reasoning system could be used to resolve pragmatic (i.e, *intentional*) ambiguity. In this section, we will discuss how it can also be used to resolve *referential ambiguity*. This is a critical capability when a reference resolution component such as that introduced in Chapters 3-4, as they may produce multiple candidate referential hypotheses for an incoming utterance.

Because this capability involves an integration of the work seen throughout this dissertation (reference resolution capabilities from Chapters 3-4, referring expression generation capabilities from Chapter 5, and pragmatic understanding and generation capabilities from Chapters 6-7, we will first step through the integration of these components, and explicitly describe how this integration implements each of the five stages of the framework described in Section 7.2.2. In doing so, we will also discuss special considerations which need to be made during this integration, due to the use of different uncertainty management frameworks, and due to the use of identifiers representing memory traces to referents within our utterance representations. Next, we will provide a proof-of-concept demonstration of the behavior of this integrated system. Finally, we will present the results of a human-subject experiment in order to evaluate our approach.

Implementation of Framework Stages

First, we will describe how the the five stages of our clarification request generation framework are implemented in the DIARC architecture.

1. Uncertainty Identification

   The first step in our clarification request generation framework is identifying whether or not there is uncertainty that needs to be clarified.

We achieve this step by using a reference resolution framework to determine the set of referential candidates and their respective levels of uncertainty, and then providing those candidates to a pragmatic inference component which produces a set of pragmatic interpretation candidates whose uncertainty levels depend on the uncertainty levels of the referential candidates. In this section, we will discuss this process at length, as well as the specific integration challenges which needed to be addressed.

Our approach uses the referential processing framework presented in Chapter 3 to facilitate access to information about the various entities known of by a robot. As previously described, this framework uses a set of "consultants" to integrate a central, domain-independent open-world reference resolution component with a set of heterogeneous knowledge bases distributed throughout a robot architecture, potentially residing on multiple computers. In our instantiation of this framework, we make use of GH-POWER: the *Givenness Hierarchy*-theoretic reference resolution algorithm presented in Chapter 4. Based on the theoretical linguistic framework presented by J. K. Gundel, Hedberg, & Zacharski (1993), GH-POWER treats *DIST-POWER's* distributed memory system as a Long Term Memory Store, and builds on top of this system a set of hierarchical caches representing models of the robot's Discourse Context, Short-Term Memory, and Focus of Attention. This allows GH-POWER to resolve a wide array of referring expressions, including anaphoric and deictic expressions. And, like the non-GH-theoretic version of the *DIST-POWER* algorithm, GH-POWER is designed to operate in both uncertain and open worlds.

GH-POWER uses the logical form of a referring expression to (1) hypothesize new representations for previously unknown referents, and (2) produce a distribution $P(\Phi \mid \Gamma, \Lambda)$; that is, the probability of successful satisfaction conditioned on binding hypotheses from variables to *known* referents:

$$\{\Gamma_0 = \{\gamma_{0_0} \ldots \gamma_{0_n}\} \ldots, \Gamma_m = \{\gamma_{m_0} \ldots \gamma_{m_n}\}\}$$

and semantic parse hypotheses[5]:

$$\{\Lambda_0 = \{\lambda_{0_0} \ldots \lambda_{0_n}\} \ldots, \Lambda_m = \{\lambda_{m_0} \ldots \lambda_{m_n}\}\}.$$

[5]While the actual parser we use only currently returns a single best parse, we use this notation to accommodate other parsers which might not share this restriction.

For example, suppose Bob asked Cindy "Can you grab the medical kit?" This may be parsed by Cindy into something like $QuestionYN(bob, self, can(self, grab(self, X)))$ with additional semantic content $\Lambda_i = \{medkit(X)\}$. If Cindy is 70% sure that the object with identifier m_5 is a medical kit, reference resolution will produce:

$$P(\Phi = True \mid \Gamma = \{X \rightarrow m_5\}, \Lambda = \{medkit(X)\}) = 0.7$$

All sufficiently probable referential hypotheses are then used to create a set of *bound utterances with supplemental semantics* (BUSSes) $\Psi = \{\psi_0 \ldots \psi_n\}$. Each $\psi_i \in \Psi$ is associated with a unique sufficiently probable binding γ_i from variables found in the parsed utterance form and its supplemental semantics to entities found in Long Term Memory. For example, the BUSS associated with form $QuestionYN(bob, self, can(self, grab(self, X)))$ semantics $\{medkit(X)\}$ and binding $\{X \rightarrow m_5\}$ would be:

$$\{QuestionYN(bob, self, can(self, grab(self, m_5))) \wedge medkit(m_5)\}.$$

One could then create a *distribution* over this set of BUSSes, where $P(\psi_i) = P(\Gamma_i, \Lambda_i \mid \Phi_i)$ using, e.g., Bayes' Rule, *if* the next component in the natural language pipeline used a Bayesian approach. In fact, the next component in the pipeline (i.e., the pragmatic reasoning component) uses the more general *Dempster-Shafer theoretic* approach described in this and the previous chapter, and thus another approach must be taken.

Of course, not all of a robot architecture's components are likely to be DS-theoretic. For some components, distributional information may be readily available, encouraging the use of a Bayesian approach. To allow each architectural component to use the knowledge representation and uncertainty management approaches most conducive to its own operation, we must thus develop mechanisms that allow those components to integrate seamlessly. In the rest of this section, we will (1) review how DS theory is used in our architecture, and (2) describe the technique we use for interoperability between our DS-theoretic pragmatic reasoning component and our probabilistic reference resolution component.

We take the following DS-theoretic approach. Let $\Theta = \{\theta_0, \dots, \theta_n\}$ be a *Frame of Discernment (FoD)* where each θ_i is a mutually exclusive singleton hypotheses described by ψ_i. Recall also that each ψ is a *bound utterance with supplemental semantics* that can partitioned into a set of semantics Λ a set of bindings Γ, and a decision variable Φ that takes the value *True* if and only if Λ is *satisfied* by set of bindings Γ.

Let $m(\cdot) : 2^{\Theta} \rightarrow [0, 1]$ be a *basic belief assignment* which assigns to each θ_i a mass:

$$\frac{1}{Z} P(\Phi_i \mid \Gamma_i, \Lambda_i), \tag{7.1}$$

where

$$Z = \sum_{j=0}^{|\Theta|-1} P(\Phi_j \mid \Gamma_j, \Lambda_j).$$

As mass is only assigned to singleton sets, $Bl(\theta_i) = Pl(\theta_i) = m(\theta_i)$. The confidence interval associated with each hypothesis according to this mass assignment is identical to $[Bl(\Gamma_i, \Lambda_i \mid \Phi_i), Pl(\Gamma_i, \Lambda_i \mid \Phi_i)]$ as calculated using Heendeni, Premaratne, Murthi, Uscinski, & Scheutz (2016)'s DS-theoretic equivalent to Bayes' Rule (Eq.7.2), assuming a uniform prior distribution $Bl(\Gamma, \Lambda) = Pl(\Gamma, \Lambda) = \frac{1}{|\Theta|}$.

$$\begin{aligned} Bl(A|B) &\geq \frac{Bl(B|A)Bl(A)}{Bl(B|A)Bl(A) + Pl(B|\bar{A})Pl(\bar{A})}; \\ Pl(A|B) &\leq \frac{Pl(B|A)Pl(A)}{Pl(B|A)Pl(A) + Bl(B|\bar{A})Bl(\bar{A})}. \end{aligned} \tag{7.2}$$

Before we move on, it is important to note that hypotheses with probabilities below a given threshold are pruned out during the resolution process, as described in our previous work described in Chapter 4. This has the effect of concentrating extra probability mass in the remaining hypotheses, leading, respectively, to higher beliefs and plausibilities.

The result of the above calculations is a Frame of Discernment whose singleton hypotheses can be described by the logical conjunctions (i.e., BUSSes) $\psi_0 \dots \psi_n$. Remember that each BUSS contains both a parsed utterance form and a set of supplemental semantics, bound using a single candidate variable binding. The next component in the *DIARC*

NL Pipeline (i.e., PINF) only uses the utterance form, however, and there may be multiple hypotheses in the resulting Frame of Discernment Θ that have the same utterance form but different supplemental semantics.

Note, however, that each ψ only uses those bindings in Γ_i associated with the utterance's root node (typically the formula representing the verb). There may be variables in V that had multiple possible bindings, but which do not appear in the utterance's root node, and thus there may be identical hypotheses within our frame of discernment.

For example, if Bob had asked "Can you grab the medkit that is near the book?", and one candidate medkit (m_1) is actually near two books (m_2 and m_3), we could have two hypotheses which can be described by BUSSes that have the same utterance form (e.g. $QuestionYN(bob, self, grab(self, m_1))$) but different supplemental semantics (e.g., $\{medkit(m_1) \wedge book(m_2) \wedge near(m_1, m_2)\}$ vs $\{medkit(m_1) \wedge book(m_3) \wedge near(m_1, m_3)\}$.

We thus cluster these hypotheses into sets $C_0, \ldots, C_n$ such that all hypotheses associated with each set are described by BUSSes that have the same utterance form. For example, if we have three singleton hypotheses $\{\theta_0, \theta_1 \theta_2\}$, and ψ_0 and ψ_1 have the same utterance form, $C = \{\{\theta_0, \theta_1\}, \{\theta_2\}\}$.

We can now split our Frame of Discernment Θ into a set of $\mid C \mid$ "binary" FoDs, one for each cluster C_i. Each binary FoD itself has two hypotheses: (1) that the utterance form describing all hypotheses in cluster C_i *does* represent what was communicated, and (2) that it does not. This splitting has no theoretical ramifications, but facilitates easier integration with PINF. Because each cluster is mutually exclusive from all other clusters, each binary FoD can be represented entirely by the *bound utterance structure*:

$$\langle utterance(\psi_i), Bl(\{C_{i_0} \ldots C_{i_n}\}), Pl(\{C_{i_0} \ldots C_{i_n}\})\rangle.$$

calculated as

$$\langle utterance(\psi_i), \sum_{j=0}^{|C_i|-1} m(C_{i_j}), \sum_{j=0}^{|C_i|-1} m(C_{i_j})\rangle \tag{7.3}$$

Suppose $\Theta = \{\theta_0, \theta_1, \theta_2\}$ and $\Psi = \{\psi_0, \psi_1, \psi_2\}$, where

$$
\begin{aligned}
\psi_1 =&(QuestionYN(bob, self, can(self, grab(self, m_1))) \\
&\wedge medkit(m_1) \wedge book(m_2) \wedge near(m_1, m_2)), \\
\psi_2 =&(QuestionYN(bob, self, can(self, grab(self, m_1))) \\
&\wedge medkit(m_1) \wedge book(m_3) \wedge near(m_1, m_3)), \\
\psi_3 =&(QuestionYN(bob, self, can(self, grab(self, m_4))) \\
&\wedge medkit(m_4) \wedge book(m_2) \wedge near(m_4, m_2)),
\end{aligned}
$$

and assume, for example, a DS-theoretic *Basic Belief Assignment(BBA)* that assigns probability masses to each hypothesis in Θ according to the following table, where Bl and Pl are belief and plausibility – upper and lower bounds on the expected probability of each hypothesis:

Hypothesis	Mass	Bl	Pl
$\emptyset$	0.0	0.0	0.0
$\{\theta_0\}$	0.2	0.2	0.2
$\{\theta_1\}$	0.3	0.3	0.3
$\{\theta_2\}$	0.5	0.5	0.5
$\{\theta_0, \theta_1\}$	0.0	0.5	0.5
$\{\theta_1, \theta_2\}$	0.0	0.8	0.8
$\{\theta_2, \theta_0\}$	0.0	0.7	0.7
$\{\theta_0, \theta_1, \theta_2\}$	0.0	1.0	1.0

Because ψ_0 and ψ_1 have the same utterance form, $C = \{\{\theta_0, \theta_1\}, \{\theta_2\}\}$. From this, the following set of bound utterance structures will be created:

$$
\begin{aligned}
&\{\langle QuestionYN(bob, self, can(self, grab(self, o_1))), \\
&\qquad\qquad\qquad\qquad Bl(\{\theta_0, \theta_1\}), Pl(\{\theta_0, \theta_1\})\rangle, \\
&\ \langle QuestionYN(bob, self, can(self, grab(self, o_4))), \\
&\qquad\qquad\qquad\qquad Bl(\{\theta_2\}), Pl(\{\theta_2\})\rangle\} = \\
&\{\langle QuestionYN(bob, self, can(self, grab(self, o_1))), 0.5, 0.5\rangle \\
&\ \langle QuestionYN(bob, self, can(self, grab(self, o_4))), 0.5, 0.5\rangle\}
\end{aligned}
$$

The set of bound utterance structures is sent to PINF, which uses context to determine the intentions underlying utterances (as described in the previous chapter), producing a set of intentional structures $\langle I, Bl(I), Pl(I)\rangle$. If Núñez' uncertainty measure (Núñez et al., 2013)) (Eq.7.4) returns a value below some threshold, such as 0.1 (which will become increasingly likely as the difference between $Bl(I)$ and $Pl(I)$ grows and as the center of the interval becomes closer to 0.5), intention I is deemed in need of clarification.

$$1+\frac{\beta}{K}log_2\frac{\beta}{K}+\frac{1-\alpha}{K}log_2\frac{1-\alpha}{K} \quad \text{where } K=1+\beta-\alpha. \tag{7.4}$$

PINF then formulates an intention-to-know (*itk*) which of these intentions is correct, denoted $itk(s, or(i_0, i_1, \ldots, i_n))$.

Before integration with *GH-POWER*, PINF only handled *pragmatic* uncertainty. But because PINF now receives a set of candidate utterance forms, each of which may have different argument bindings, it now automatically captures *referential* uncertainty as well.

Before we move on, we would like to point out that that because *DIARC's* reference resolution component handles *open worlds*, instances in which interlocutors refer to previously unknown entities do not automatically generate clarification requests. For example, if the robot is told "Go to the room at the end of the hall" and did not previously know of a room at the end of the hall, it will not ask for clarification, but will rather hypothesize a new location, and carry on.

We do not regard such situations as referentially ambiguous. Here, the robot knows what entity is being referred to: a previously unknown room at the end of the hall. It may, of course, be valuable for the robot to ask for more information about this location, but we believe such a decision is not appropriate at the stage of processing we discuss in this section.

2. Decision to Communicate

Currently, any *intention-to-know (itk)* formulated during the previously described stages of processing is automatically asserted into the robot's knowledge base, triggering a decision to communicate this intention once it is acceptable for the robot to accept the conversational turn. When this decision is made, the *itk* is passed to the pragmatic *generation* component for processing.

3. Utterance Choice

   The robot must now determine a contextually appropriate way to formulate its intention at the *utterance level*. This is accomplished once again by PGEN, which uses the same set of rules for generation as it uses for inference. In our first experiment, we observed that if there were more than two options, listing those options was dispreferred over a more general question. Thus, if we are to send a clarification request to PGEN that has semantics of the form $itk(self, or(option_1, ...option_n))$, we first check whether or not n is greater than the acceptable number of candidates to list, i.e., two. If $n = 2$, this intention is sent directly to PGEN. Otherwise, all options are unified into a single predicate whose only bound arguments are those that are identical for all options. For example, if $\{option_1, option_2, option_3\} = \{need(jim, objects_1), need(jim, objects_2), need(jim, objects_3)\}$, these will be unified into $need(jim, ?)$, and the intention $itk(self, need(jim, ?))$ will be sent to PGEN instead.

   Using DS-theoretic logical operators, PGEN is able to determine a set of candidate utterance forms, each of which is forward-simulated through pragmatic inference in order to ensure that the agent does not accidentally communicate anything it does not actually believe to be true as a side effect of communicating its primary illocutionary point. The best candidate utterance is then sent to NLG for surface realization.

   This processing step is not typically included in traditional NLG frameworks, which do not typically need to account for social context or dialogue context. They instead typically include a *document structuring* (cf. Reiter, Dale, & Feng, 2000) stage in which the agent determines the order in which to convey multiple utterances. Because clarification request generation in human-robot interaction *typically* only involves a single utterance, we do not currently handle this step, but it will be an important topic for future work. A robot may, for example, need to preface a clarification request by stating what parts of an utterance it *did* understand.

4. Surface Realization and Speech Synthesis

   Once an appropriately phrased utterance form is chosen by the pragmatic generation component, that utterance is sent to the *Natural Lan-*

guage Generation component for Surface Realization. First, that component chooses sets of properties to use to describe each of the utterance's referents, using the *DIST-PIA* algorithm presented in Chapter 5. For example, consider the utterance form

$$QuestionWH(self, bob, or(need(bob, grab(self, m_1)),\\ need(bob, grab(self, m_2)))).$$

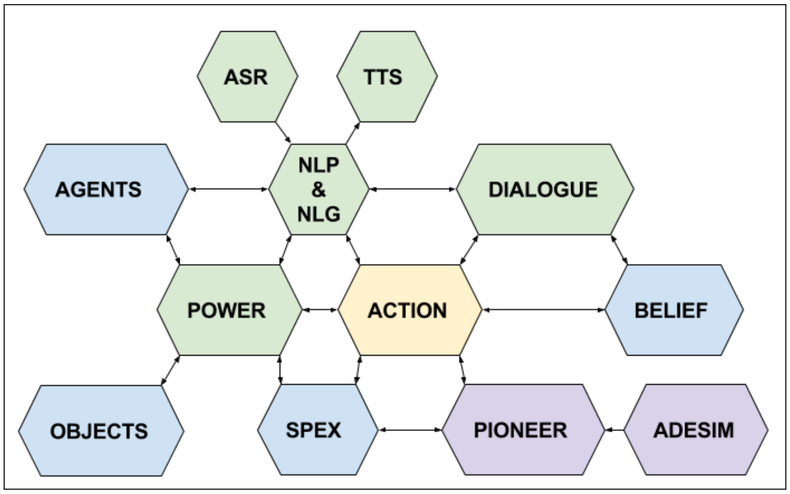
In this utterance form, there are two referents that must be described: m_1 and m_2. The referent m_1 may be described using the properties $\{mug(X) \wedge white(X)\}$, and m_2 may be described using the properties $\{mug(X) \wedge black(X) \wedge large(X)\}$.

The utterance form and sets of properties are then translated into raw text using the open source SimpleNLG package, producing, for example, "Do you need the white mug or do you need the large black mug?" when there are two referential candidates, and producing, for example, "Which one do you need?" in the case of a larger number of referential candidates. The open source MaryTTS package is then used to synthesize this text into an audio form that is produced by the robot.

Demonstration

To demonstrate the operation of the presented approach, we present a proof-of-concept interaction that occurs in a simulated environment. This demonstration highlights the full implementation of all stages of the clarification request generation framework through components of the *DIARC* architecture. Specifically, this demonstration uses the components of the *DIARC* architecture shown in Figure 7.2. In addition to components responsible for the simulation of a Pioneer robot within an office environment, our configuration used the following components: ASR (which performs simulated speech recognition), NLP (which uses the C&C parser within a GH-theoretic framework), POWER (i.e., *REX*, as described in Chapter 3), AGENTS, SPEX and OBJECTS (*GH-POWER* Consultants) providing information about people, places, and things), DIALOGUE (which, performs dialogue management, and includes PINF and PGEN as submodules), the BELIEF manager (which allows the DIALOGUE to assess its current context), and ACTION (i.e., the goal manager). Of these components, the POWER, NLP, NLG, and DIALOGUE components are central to the integrated approach presented in this section.

Figure 7.2: Clarification Request Architecture Diagram



Architecture Diagram. Knowledge base components are depicted in blue; linguistic components are depicted in green; simulation components are depicted in purple; the action manager is depicted in yellow.

The interaction begins with the speaker saying to the robot "I need the medkit" in an environment in which the robot knows of two medkits, one red and one white. ASR sends this sentence to NLP, which parses the utterance into the dependency tree:

[rootVB need [ncsubj I] [dobj medkit [det the]]].

From this tree, NLP extracts root semantic content $need(X1, X2)$, with utterance type *Statement*, additional semantic content: $\{speaker(X1) \wedge medkit(X2)\}$, and presumed cognitive statuses $\{X1 \rightarrow definite, X2 \rightarrow definite\}$.

Using this information, *GH-POWER* searches for the referents to bind to $X1$ and $X2$; for $X1$, *GH-POWER* finds a single probable candidate: $agents_1$, with probability 1.0; for $X2$, two candidates are found: $objects_1$, with probability of satisfaction 0.82, and $objects_2$, with probability of satisfaction 0.92. These bindings are then used to create the following bound utterances[5]:

[5]Here, $agent_1$ is changed to the name of that agent for the sake of dialogue processing.

$$\{Statement(bob, self, need(bob, objects_1)),$$
$$Statement(bob, self, need(bob, objects_2))\}$$

with corresponding probabilities[6] 0.82 and 0.92, respectively. These are normalized (see Eq.7.1) and used to create DS-theoretic bound utterance structures, which are passed to DIALOGUE:

$$\{\langle Statement(bob, self, need(bob, objects_1)), 0.471, 0.471\rangle,$$
$$\langle Statement(bob, self, need(bob, objects_2)), 0.529, 0.529\rangle\}$$

PINF and PGEN possess the rule:

$$\langle Statement(X, Y, need(Z, W)) \Rightarrow goal(Y, bring(Y, W, Z)), 0.9, 0.99\rangle, \quad (7.5)$$

indicating that the robot is between 90 and 99% confident in the rule; because the antecedent of this rule matches the utterance form of each bound utterance structure, uncertain Modus Ponens is applied in both cases, producing the set of intentional structures:

$$\{\langle goal(self, bring(self, objects_1, bob)), 0.424, 0.576\rangle,$$
$$\langle goal(self, bring(self, objects_2, bob)), 0.476, 0.524\rangle\}$$

Note that at this point, belief no longer equals plausibility: while the robot may not have encoded any ignorance with respect to what utterance was heard, ignorance encoded with respect to the context and rules the robot uses for pragmatic inference are reflected in the uncertainty intervals of the rules' consequents, thus painting a better picture of how much the robot truly knows about its interlocutor's intentions.

Núñez' uncertainty rule (see Eq.7.4) determines that both of these intentions are highly uncertain. DIALOGUE thus determines its own intention to know which is correct, encoded as the structure:

$$\langle itk(self, or(goal(self, bring(self, objects_1, bob)),$$
$$goal(self, bring(self, objects_2, bob)))), 1.0, 1.0\rangle$$

To decide how to communicate this intention, the bound utterance structure is passed through the pragmatics system in reverse, using a rule of the form:

$$\langle QuestionWH(X, Y, or(Z, W)) \Rightarrow itk(X, or(Z, W)), 0.95, 0.95\rangle, \quad (7.6)$$

[6] All beliefs and plausibilities in this section are rounded.

Our approach allows recursive generation, and thus Eq.7.6 is chained with Eq.7.5 to produce:

$$QuestionWH(self, bob, or(need(bob, objects_1), need(bob, objects_2))).$$

This utterance is then sent to our NLG component for generation of referring expressions for "bob", "objects_1" and "objects_2", and subsequent realization of the entire expression. This produces the text "Do you need the white medkit or do you need the red medkit?" which is then synthesized and output by the robot.

To evaluate our approach, we conducted a human-subject experiment similar to Experiment One. This experiment was comprised of two parts: (1) a data collection stage, and (2) an evaluation stage.

Data Collection Stage

First, I will describe the data collection stage of our human subjects experiment.

1. Experimental Design

   We first created a tabletop scene containing four differently colored waterbottles, four differently colored markers, and four differently colored notebooks, for a total of twelve objects, as seen in Figure 7.3. For each of these three object types, we took photographs of the scene in which zero, one, or two of that object type were taken away. This produced nine tabletop scenes, three of which contained identical object arrangements (i.e., those scenes in which no objects were removed).

   In our data collection experiment, each participant was shown one of these nine images at random, with a caption describing the participant's task, followed by a text box. For example, for the image in which three of the four waterbottles was shown, the following caption was used:

   "You have been told 'I need the bottle!' and would like to fulfill the speaker's request. However, as you can see, there are three bottles on the table: a silver bottle, a green bottle, and a blue bottle. Please type a sentence you would use to ask the speaker for clarification, so that you will know what bottle to pick up."

Table 7.1: Utterance forms generated in Experiment Two, Part One, and chosen between in Experiment Two, Part Two

Generator	#	Utterance Generated in Part One	Result
Robot	2	Do you need __ or do you need __?	9.4%
Human	2	Do you need __ or __?	**45.3%**
Human	2	What color __ do you need?	22.6%
Human	2	What color __ do you want?	22.6%
Robot	3	Which one do you need?	23.7%
Human	3	Which color __ do you need?	**33.9%**
Human	3	Which color __?	23.7%
Human	3	Which color __ would you like?	18.6%
Robot	4	Which one do you need?	20.0%
Human	4	What color __ do you need?	**24.3%**
Human	4	Which color __ would you like?	22.9%
Human	4	Which color __?	21.4%
Human	4	What color is the __?	11.4%

Column 1 indicates whether each utterance form was generated by the presented approach or by a human in Part One. *Column 2* indicates how many suitable referents existed in the scene for which each utterance was generated. *Column 3* indicates the generated utterance form, generalized across noun phrases. In Part Two, blanks were filled in with referring expressions generated by the REG algorithm presented in Chapter 5. For example, for scenes in which the initial utterance was "I need the bottle", the gaps in the first two rows were filled with "the green bottle" and "the silver bottle", and the remaining gaps were filled with "bottle". *Column 4* indicates the percentage of participants in Part Two who chose that utterance form as the best utterance form to use to ask for clarification.

Similar captions were used for the other images. Once the participant entered text into the text box, they were free to click to the next page, and end the experiment.

2. Participation

   Participants were recruited (53 Male, 39 Female) using Amazon Mechanical Turk. Participants ranged in age from 20 to 77 (M=33.15,SD=8.94), and were paid $0.30 to participate. As a total of 92 participants were recruited, an average of 30.7 utterances were collected for each grouping of scenes that had the same number of objects

Figure 7.3: Tabletop Environment used in Experiment Two.

removed.

3. Results

   All utterances collected in this stage were standardized with respect to noun phrasing. For example, “Do you want me to pick up the silver bottle or the blue bottle?” was reduced to “Do _ want _ to pick up _ or _ ?” All utterances within each cluster were grouped by identical phrasing, and the three most common phrasings for each cluster were selected (four in the case of a tie). The REG algorithm described above and presented in Chapter 5 was then used to generate noun phrases to fill into the previously created gaps, thus creating three to four utterances for each image. Next, an additional utterance was generated for each image using the approach presented in this section: for each image, knowledge of the objects in the image was provided to the robot architecture, and the utterance “I need the [name of object type]” was said to a robot running the architecture. Because the architecture also used the REG algorithm described above and presented in Chapter 5, the utterances generated by our robot architecture had the same noun-level phrasings as all other utterances, but a different utterance-level phrasing. Thus, the end result of this stage was a set of thirty-nine utterances that had unique phrasings at the utterance level but identical phrasings at the noun level. The thirteen utterances forms (before referring expressions were filled in) are shown in Table7.1, Column 3.

Evaluation Stage

Next, I will describe the evaluation stage of our human subjects experiment.

1. Experimental Design

   In this stage, each participant was shown one of the nine tabletop scenes created in the first stage, along with a caption such as: "Your friend Alex says to you, 'I need the bottle!' Which of the following sentences would be best to say to Alex, so that you will know which bottle to give her?"

   Each participant was then presented with the four to five utterances associated with the presented image, in the form of buttons. Clicking on one of the utterances moved the participant to the next page, and ended the experiment.

2. Participation

   Participants were recruited (94 Male, 88 Female) using Amazon Mechanical Turk. Participants ranged in age from 18 to 74 (M=34.55, SD=11.16), and were paid \$0.30 to participate. As a total of 182 participants were recruited, an average of 20.22 data points were collected for each scene.

3. Results

   Robot-generated requests were chosen only slightly less frequently than were human-generated requests: overall, robot-generated requests were chosen 18.13% of the time, whereas each form of human-generated request was chosen, on average, 24.67% of the time. Overall, this is a positive result as it suggests that the algorithm overall did not generate requests that were significantly worse than the requests that humans used most frequently. A request-by-request breakdown of participants' choices is shown in Table 7.1, Column 4.

   But in fact, robot-generated requests stands to perform significantly *better* than the majority of human-generated requests in the case where there were exactly two options to disambiguate (the first section of Table 7.1). At first glance, it appears that the robot-generated requests were chosen significantly less frequently than were human-generated requests. In this case, however, the robot-generated requests were nearly identical to the top performing human-generated requests: the robot-generated requests were simply more verbose, as they used a conjunction at the clause level rather than the noun-phrase level. This suggests that if our approach had been modified to use conjunctions at the phrase level, it would have outperformed the second- and third-best human-generated requests *combined*. And in fact, we have since made this modification, as we will later discuss.

Discussion

In our initial experiment (presented in Section 7.2.3), we observed that participants dispreferred clarification requests that were insensitive to pragmatic factors, that did not indicate understanding of an interlocutor's goals or intentions, that listed more than two options, or that did not list both options when there were only two likely candidates. These observations were confirmed in the evaluation stage of the experiment presented in this section. The most commonly chosen clarification requests were nearly identical to the clarification requests generated by our robot architecture. But in none of the three utterance groupings (i.e., the two-option, three-option, and four-option utterance groupings) were our chosen clarification requests *exactly* identical to the most commonly chosen clarification requests, and in fact, they differed from those requests in small but important ways.

As previously mentioned, when there was referential ambiguity between only two candidate referents, participants in the evaluation stage of our experiment did indeed greatly prefer clarification requests that listed all options. However, the specific phrasing used by our robot architecture was simply too verbose, as it failed to identify structural similarities and distribute appropriately. Since running this evaluation, we have added functionality to the NLG component that performs such distribution when structural similarity is detected, and our robot architecture thus now generates the exact utterance forms that were most preferred by humans (e.g., "Do you need _ or _ " rather than "Do you need _ or do you need _ ?"). When compared to the other human-generated REs in the 2-option case, the robot-generated RE could now have been chosen by between 45.3 and 54.7% of participants.

A more crucial difference, however, is observed when more than two options present themselves. It is striking to observe that all commonly-used human-generated utterances in these cases do not explicitly ask for disambiguation between bottles, but rather ask for information regarding a specific property that could be used to disambiguate between bottles. This suggests that in these cases, the optimal approach to clarification request generation likely lies somewhere between the approach presented in this section and the information-theoretic approaches seen in previous work(Deits, Tellex, Kollar, & Roy, 2013; S. Hemachandra, Walter, & Teller, 2014; Purver, 2004).

We predict that the ideal approach to clarification request generation may involve generation in a way quite similar to the approach used in this section, followed by a stage in which information-theoretic mechanisms are used to add differentiating modifiers.

It is also important to note, however, that in all three cases a significant percentage of participants did choose the less popular choices. When four options were presented, for example, "Which color _ would you like" was chosen by less than two percent fewer participants than was the most popular "What color _ do you need?". This suggests that it may be valuable in future work to develop models of human interlocutors that can predict which type of request they in particular would find most appropriate. While at first glance the difference between the alternate strategies may seem arbitrary, we suspect that they in fact represent different strategies that are either explicitly used, or which arise from differential weightings of pragmatic principles: Utterances such as "Which color _ *do you need*" may be used due to subconscious *lexical entrainment* or conscious *refashioning* in which speakers use the same phrasing as that used by their interlocutors(H. H. Clark & Schaefer, 1989; Brennan, Galati, & Kuhlen, 2010; Yoon & Brown-Schmidt, 2013); utterances such as "Which color _ *would you like*" and "Which color _ *do you want*" may be used due if the pragmatic value of a refashioned sentence is weighted lower than that of a more *conventionally indirect* utterance form(Searle, 1975); and utterances such as "Which color _ " may be used due to the interaction of either aforementioned pragmatic strategy with Grice's Third Maxim of Manner: "Be brief (avoid unnecessary prolixity)"(Grice, 1970).

7.2.6 Section Discussion

We have presented an integrated approach to clarification request generation for human-robot interaction contexts, and shown how this approach can identify and handle both pragmatic and referential ambiguity, and how our approach can be used in architectures where information about referents is distributed across multiple heterogeneous knowledge bases, as is often the case in cognitive robot architectures. The primary finding of this section is that a language-enabled robot's pragmatic reasoning component can track and address *referential* ambiguity when integrated with a probabilistic reference resolution component: a useful finding for designers of language-enabled robot architectures intended for use in human-robot interaction domains.

We have also demonstrated this approach as fully implemented on a simulated robot, and provided the results of a human-subject study showing that the theoretical commitments of our robot architecture align with human preferences, and that the clarification requests generated by our full NLG pipeline are comparable to the most highly preferred human-generated clarification requests.

Our findings suggest several directions for future work. First, research is needed on using information-theoretic mechanisms to adapt clarification requests generated by pragmatic reasoning components. Second, research is needed to develop speaker-specific models that can predict precisely what type of clarification request they would most likely prefer, based on their inferred weighting of pragmatic principles Third, future work should also further examine methods by which components using different frameworks for representing uncertainty can be optimally integrated. Finally, a tighter integration between pragmatic reasoning and reference resolution can be achieved: in previous work, we have shown how our pragmatic reasoning component can use contextual knowledge to abduce the most appropriate way to phrase an utterance; but this contextual knowledge is assumed to be stored in a robot's centralized belief and dialogue components. In future work, this should be extended to allow such contextual information to be distributed across the robot's distributed heterogeneous knowledge bases, when appropriate.

We will now turn our attention to a slightly different topic: that of pragmatic *robot-robot* communication. We showed in this section how robots can generate appropriate utterances when communicating with humans in natural language dialogues. We must now examine whether there is utility in using the presented algorithms to do so for the sake of robot-robot communication as well.

7.3 Pragmatic Robot-Robot Communication

In this dissertation, I have emphasized the importance of developing mechanisms to facilitate effective, natural human-robot interaction, especially mechanisms to facilitate natural language communication. Recent research within the field of HRI has investigated various social aspects of natural language interactions with robots, such as politeness (Briggs & Scheutz, 2014b), turn taking (Nadel, Revel, Andry, & Gaussier, 2004), affective speech (Scheutz, Schermerhorn, & Kramer, 2006), dialogue-appropriate facial movements (C. Liu, Ishi, Ishiguro, & Hagita, 2012), pragmatic analysis (Williams, Briggs, Oosterveld, & Scheutz, 2015), and collaborative control (Fong, Nourbakhsh, & Dautenhahn, 2003). Due to the difficulty of managing multi-party dialogue, this research has primarily focused on dialogue between a single human and a single robot, with a few exceptions: For example, some work has demonstrated multiple simultaneous conversations between a single human and several remote robots (Fong, Nourbakhsh, & Dautenhahn, 2003), and some work has demonstrated multi-party conversation between multiple co-located humans and a single robot (Foster et al., 2012; Matsuyama, Taniyama, Fujie, & Kobayashi, 2006).

However, little research to date has investigated the question of how robots that communicate with humans should communicate with *each other*. Some researchers have looked into wireless inter-robot communication protocols (Balch & Arkin, 1994; Fukuda & Sekiyama, 1994; J. W. J. Wang, 1994), and some researchers have developed mechanisms for managing conversation between a human and multiple co-located robots (Briggs & Scheutz, 2012), but such research does not examine how humans actually perceive such communication. *Should* robots communicate with each other in natural language, so as to be transparent to humans, or can they use whatever form of communication best suits their needs?[7] It seems clear that there cannot be a simple context-independent answer to this question. For example, consider the difference between cooperative vs. competitive contexts: In the first, humans and robots have to work together toward a common set of goals; in the second, humans and robots have competing, incompatible goals. Socially assistive robots and robots for search and rescue missions are examples of the former, while robots for robo-soccer or law enforcement are examples of the latter. It is clear that in the latter case, robots should not divulge their intentions and goals, as leaking knowledge about their plans and actions will benefit the adversary. It is less clear whether in the first case robots should *always* communicate in natural language. In some instances, keeping co-present humans "in the loop" will be advantageous, while in others, "communications overhead" might be unnecessary and distracting.

In this section, we set out to investigate human perceptions of robot-robot communication in the context of a mixed-initiative human-robot team, where the human commands two robots to perform a search and rescue task in a simulated disaster area. The main results of a set of two human-robot experiments in this domain suggest that it might be advantageous for robots to communicate in natural language with humans in the context of cooperative tasks, in order to avoid being viewed as unsettling or creepy by their cooperative, co-located human teammates. This suggests that the

[7] The authors find robots using the most effective means of communication to be non-controversial when humans are not present, telepresent, or otherwise in observation.

pragmatic generation mechanisms presented in this chapter may be valuable not only for human-robot dialogues, but for robot-robot dialogues as well.

The remainder of this chapter proceeds as follows: In Section 7.3.1, we briefly review previous work on human-robot and robot-robot communication, and then our hypotheses are described in Section 7.3.2. Sections 7.3.3 and 7.3.4 present the two HRI experiments we conducted to evaluate our four hypotheses. Section 7.3.5 discusses the implications of our findings, followed by a summary in Section 7.4.

7.3.1 Background

While there is a growing body of research on human perceptions of human-robot communication, very little work has investigated human perception of robot-robot communication. Two sets of studies have investigated verbal robot-robot communication by examining human perceptions of robots engaged in humorous banter or non-task-oriented conversation (Hayashi, Kanda, Miyashita, Ishiguro, & Hagita, 2008; Tsujimoto, Munekat, & Ono, 2013), and two sets of studies have investigated human perception of nonverbal robot-robot communication. As we are primarily concerned with human perception of silent robot-robot communication relative to verbal robot-robot communication, we will focus on these two studies.

In the first set of studies (Kanda, Ishiguro, Ono, Imai, & Mase, 2002; Kanda, Ishiguro, Ono, Imai, & Nakatsu, 2004), human participants observed two robots discussing a piece of artwork. The robots' manner of conversation fit one of three conditions: In the first condition, the robots conversed and gesticulated; in the second, they conversed without gesticulating; and in the third, the conversation was skipped altogether. In all three conditions, one of the two robots subsequently approached and spoke to the human observer. The human subjects were then asked about their comfort level when interacting with the robot. No adverse effects were found, suggesting that it is perfectly acceptable for robots to converse silently while observed by humans.

However, there are two important limitations to this study. For one, the experiment does not truly contrast verbal and silent behavior, as in the silent condition, no robot-robot conversation whatsoever took place from the participant's point of view. It would thus be more accurate to say that the study compares the comfort levels of participants who engage in conversation with robots that have been shown capable of conversation, and the comfort levels of participants who engage in conversation with robots that have *not* been shown capable of conversation. Moreover, participants in the experiment had no investment in the robots' conversation; the robots were not discussing anything the participants needed to know about, and thus there were no negative consequences to participants being kept "out of the loop" of the robots' conversation. In a human-robot team task, information communicated between robots could very well be crucial for human teammates.

In the second set of studies investigating human perception of nonverbal robot-robot communication (M. R. Fraune & Šabanović, 2014; M. Fraune & Šabanovic, 2014), participants completed surveys while robot activities in their vicinity unfolded according to one of four conditions: (1) three robots wandered pseudo-randomly, beeping occasionally; participants were told that the robots did not communicate with each other, (2) three robots wandered pseudo-randomly, beeping occasionally; participants were told that the robots communicated with each other over the Internet, (3) three robots wandered pseudo-randomly, beeping occasionally, and from time to time beeping loudly in sequence; participants were told that the robots communicated via beeps, and (4) a control condition with no robots present. The researchers were interested in whether the attribution of non-anthropomorphic communication styles to the robots would increase the salience of the robots' "out-group status", causing them to be viewed less favorably. Results showed that participants generally thought the robots were communicating aloud, even in conditions 1 and 2, where participants where either told that the robots were not communicating or that the robots were communicating over the Internet. Since no significant differences in human perception of robots among any of the four conditions were found, the researchers concluded that the robots were not attributed out-group status, and that communication style did not affect human perceptions of robots.

However, as with the previous set of studies, there are two important aspects of this study which significantly limit its applicability to other robot-robot communication scenarios, in particular, human-robot team tasks. First, it is not clear whether there was any reason for the participants to have felt left "out of the loop", or to have felt that the robots were uncooperative, untrustworthy, or unsettling, as the participants did not know what the robots were doing and the robots never communicated verbally. Had the subjects been given the opportunity to observe the robots communicate verbally, then the use of silent communication could have been cast as an *intentional choice* of the robots to prevent the humans from knowing the content of their communications. Furthermore, the lack of verbal communication may have reduced the degree to which humans perceived the robots as human-like, thus decreasing the effects communication style may have had on perception of intention-driven robot attributes, such as cooperativity. As with the first set of studies, the human observers had no investment in the robots' activities, with all the consequences previously described.

The above sets of studies are typical of a whole class of experiments in HRI, where humans are interaction observers rather than participants, and as such, they have no reason to be invested in the robots' activities or performance. Hence, any conclusions derived from such experiments are limited to interaction observation and cannot automatically be generalized to interaction participation.

7.3.2 Human Perceptions of Covert Robot Communication

Human subjects will have far lower investment in communication outcomes between robots they are merely *observing* compared to robots with which they are *interacting*. To address this lack of investment, we devised a joint human-robot team task where human participants (1) have to interact verbally with robots, (2) are able to verify when silent communication has occurred, and (3) have a vested interest in the accuracy of the robot-robot communication. To ensure that the participants would have an interest in the information communicated, we constructed a scenario in which participants needed one robot to relay instructions to another robot. In this way, participants depended on the robot interlocutor to communicate their instructions accurately to the other robot (in order for the scenario's task to be completed efficiently), and the robots depended on the participant to provide them with appropriate instructions. This paradigm allowed us to explore four important questions about robot-robot communication in human-robot team tasks:

1. Will robots be viewed as more or less **trustworthy** if they choose to communicate silently? A wide variety of factors can influence the degree of trust a human has for a robot teammate. One such factor is transparency: To engender trust, the motivation behind a robot's behavior should be transparent and easily understandable (Hancock, Billings, & Schaefer, 2011). If robot-robot communication is enacted silently, the motivation behind robot actions may be unclear, leading to distrust. Another factor influencing human-robot trust is similarity of mental models; to engender trust, teammates should endeavor to create and share mental models (Hancock, Billings, & Schaefer, 2011; Neerincx, 2007). If robot-robot communication is enacted silently, human teammates may not be able to appropriately update their mental

models. The resulting dissimilarity of mental models may lead to distrust. Given these concerns, we hypothesize that **(H1)** robots will be viewed as **less trustworthy** if they choose to communicate silently.

2. Will robots be viewed as more or less **cooperative** if they choose to communicate silently? We believe that the same factors that may cause a robot to be viewed as untrustworthy may also cause a robot to be viewed as uncooperative, as lack of transparency and dissimilarity in mental models are likely to lead to simple misunderstandings. For this reason, we hypothesize that **(H2)** robots will be viewed as **less cooperative** if they choose to communicate silently.

3. Will robots be viewed as more or less **unsettling** if they choose to communicate silently? Over the past few decades, a variety of fields have given increased attention to the "Uncanny Valley" (Mori, MacDorman (Translator), & Minato (Translator), 2005), a hypothesis stating that entities very close to (but not quite) human are perceived as creepy or unsettling. Recent research suggests that these feelings of eeriness do not directly correlate with human-likeness, and that human likeness may thus be only one of several factors contributing to the Uncanny Valley effect (Brenton, Gillies, Ballin, & Chatting, 2005; MacDorman, 2006). One such contributing factor is the use of "uncanny actions." Uncanny actions include those that can be construed as human but are executed with slight deviation from normal human execution: a robot that blinks too infrequently or that follows teammates too closely could be viewed as uncanny. In addition to these types of uncanny actions, we believe that actions that cannot be construed as human should also be considered to be uncanny actions, as research has shown that humans generally prefer robots whose actions can be construed as human (Walters, Syrdal, Dautenhahn, Te Boekhorst, & Koay, 2008). One example of this kind of uncanny action is *telepathy*. Telepathy is not in the realm of human ability and is largely considered to be paranormal or supernatural. However, robots regularly communicate in a manner reminiscent of telepathy (i.e., using wireless communication). This behavior may thus be perceived as creepy or unsettling. It is possible that this behavior would be viewed as analogous to text-messaging or other electronic forms of communication, or it could be viewed as analogous to situations in which humans seem to "guess" what their interlocutor is going to say (e.g., when couples finish each other's sentences). However, in such situations, there is an assumption that information

could be "guessed" due to contextual factors or longitudinal learning of an agent's goals and preferences; whereas in silent robot-robot communication, information may be communicated for which it would be next to impossible for a robot to "guess" that information. We thus hypothesize that **(H3)** robots will be viewed as **more unsettling** if they choose to communicate silently.

4. Will robots be viewed as more or less **efficient** if they choose to communicate silently? Research suggests that the use of nonverbal cues in human-robot communication leads to higher efficiency (Breazeal, Kidd, Thomaz, Hoffman, & Berlin, 2005). We believe that humans will be able to recognize the efficiency inherent in completely nonverbal robot-robot communication. For this reason, we hypothesize that **(H4)** robots will be viewed as **more efficient** if they choose to communicate silently.

We will now introduce the details of an experimental paradigm we first described in (Williams, Briggs, Pelz, & Scheutz, 2014), which we used to investigate these four questions.

7.3.3 Experiment 1

We employed a team task in which a human commander had to verbally assign different tasks to two robots and observe the robots' execution of those tasks, in order to accomplish the task goals.

Equipment

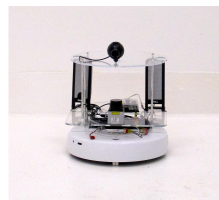
We used two different robots: "VGo" (Figure7.4a), a VGo telepresence robot augmented with an on-board computer and a variety of sensors (Tsui et al., 2013), and "Roompi" (Figure7.4b), an iRobot Create augmented with a Raspberry Pi computer, Hokuyo Laser Range Finder, speakers, and webcam. As the VGo is limited to a single text-to-speech voice option, we used that voice for both robots. While this voice was, in our opinion, slightly more female-sounding than male-sounding, it was the only option available. Both robots were controlled through Wizard of Oz interfaces, teleoperated by trained confederates in a nearby room.

Procedure

Participants were told that their task was to instruct robots as part of training for a disaster relief scenario, and that the adjacent room, which was filled

Figure 7.4: Robots used in both experiments

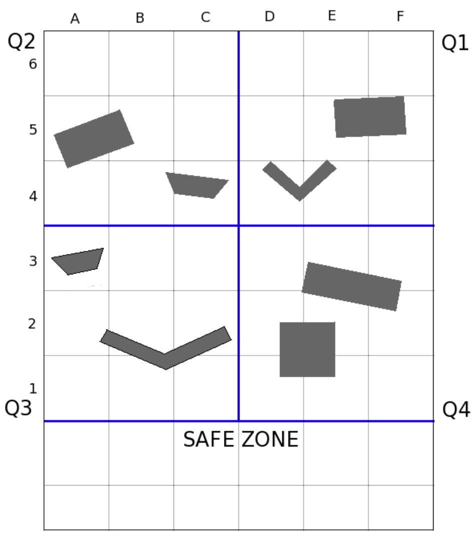

(a) "VGo", the VGo telepresence robot


(b) "Roompi", the iRobot Create

with a number of boxes and other obstacles, simulated a power plant strewn with debris after a nuclear disaster. Participants were told that the sensors of the robots they would be instructing had been manipulated such that the robots would detect injured people or high levels of radiation at various locations in the room, and that it would be their job to determine how best to delegate the tasks of searching for these locations; they should give the task of searching for survivors to a robot of their choosing and give the task of searching for radiation to the other robot. Participants were told that they must also choose separate paths through the environment (consisting of different orderings of the room's four quadrants) for the two robots, in order to prevent the robots from getting in each other's way. As an additional caveat, participants were told that since in an actual nuclear disaster they would be unable to enter the area in which the robots were working, they would need to stay in a designated "safe zone" at one end of the debris-filled room, would not be able to communicate with the robots while they were working, and would thus need to give the robots their instructions at the beginning of the task. To keep participants engaged during the task, they were asked to observe and assess the performance of the robots, tracing out on a map (as seen in Figure 8.7) the paths taken by the robots. Once the robots had finished exploring the room, they would need to mark on their map the positions of any radioactive areas or survivors found by the robots.

Once the study coordinator finished reading the task instructions, the coordinator left the room to retrieve the robots. At this point, a single

Figure 7.5: Experiment Area



(a) Map of experiment area provided to participants, showing positions of debris and safe-zone, with labeled quadrants and coordinates.

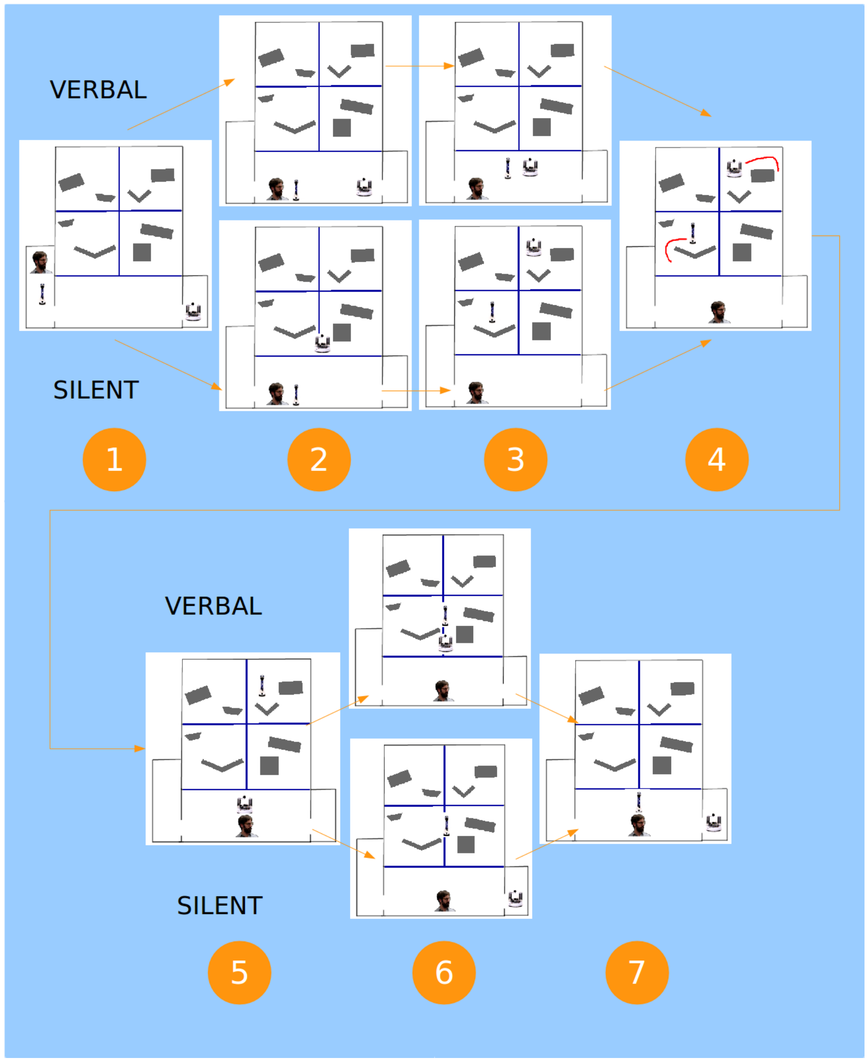
(b) Photograph of actual experiment area.

robot, VGo, entered the room, instead of both robots, as the participant had been led to expect. VGo then told the participant that Roompi was still charging, but that it could relay to Roompi its instructions. VGo then asked the participant for both its and Roompi's instructions: what each robot was to look for and in what quadrant order. Finally, the participant was prompted to follow VGo into the disaster area, depicted in Figure 7.5b.

Our intention for this experiment was to examine the differences in participants' perceptions of the robots under verbal and silent robot-robot communication strategies, and thus, participants were assigned to one of two conditions: "verbal" or "silent", as seen in Figure 7.6. In the verbal condition, VGo entered the disaster area and approached Roompi, which could be seen driving in from another room that ostensibly contained its charging station. When the robots were adjacent and facing each other, VGo then relayed aloud to Roompi the instructions that the participant had laid out for it. Roompi then acknowledged the commands with an "Okay", and both robots began the task of exploring the environment.

In the silent condition, VGo and the participant entered the room to find Roompi already beginning its assigned task, at which point VGo then

Figure 7.6: Overview of experimental paradigm



In the second experiment described in this section, the positions of VGo and Roompi are exchanged. (1) The participant gives VGO its instructions and the instructions to relay to ROOMPI. (2) The participant and VGO enter the experiment room. In the VERBAL condition, ROOMPI is observed entering the room. In the SILENT condition, ROOMPI is observed carrying out the instructions specified for it and relayed to VGo by the participant. (3) In the VERBAL condition, the two robots approach each other and VGO informs ROOMPI of its orders. In the SILENT condition, VGO follows suit and begins to carry out its orders. (4) Both robots carry out their orders. (5) The first robot to finish reports back to the participant. (6) In the VERBAL condition, this robot finds the other robot and tells it what to do once it has finished the task. In the SILENT condition, this robot simply exits the room. (7) The second robot reports back to the participant and informs him or her that the other robot says to return to the original room for another survey.

began its own task without approaching or audibly communicating anything to Roompi. The participant was thus left to assume that the two robots must have communicated silently, since Roompi was carrying out the task that they themselves had decided to delegate to it.

Once each robot finished its exploration of the room, it approached the participant and reported its findings. After relaying these findings, robot behavior once again differed by condition. In the verbal condition, whichever robot finished its task first approached the other robot, informed the other robot that it had finished, and instructed the other robot that when it too had finished it should instruct the participant to return to the original room for another survey. In the silent condition, the robot that finished first left the room after reporting its findings, without communicating anything aloud to the other robot. In both conditions, the second robot to finish reported its findings to the participant and then told the participant that the other robot had instructed it to tell them to return to the original room for another survey. Finally, participants returned to the original room and completed a post-experiment survey. Videos of this reporting behavior can be viewed at `https://www.youtube.com/watch?v=t_MLNBReoic` for the SILENT condition, and `https://www.youtube.com/watch?v=y9WODq3ONrk` for the VERBAL condition.

Population

Participants were recruited (14 male, 14 female, total: 28) through a university website. All participants were between the ages of 18 and 65 (although their ages were not recorded) and were native English speakers. Most participants (26 of the 28) were students from a variety of departments (e.g., Music, Biopsychology, Economics), and the remaining two participants were staff members. Participants were paid $10 each for their participation and provided informed written consent before beginning the experiment.

Measures

Before beginning the experiment, participants were given a short demographic survey in which they were asked a variety of questions pertaining to their prior experience with robots, video games, and technology in general. Immediately following the experiment, participants were given a 64-item survey assessing their opinions on a variety of topics, including their perception of each robot's creepiness, gender, human-likeness, trustworthiness, efficiency, and cooperativity, as well as several questions pertaining to the experiment in general and their expectations regarding robots' abilities in the near future. This survey was a modified version of the questionnaire used by P. Schermerhorn, Scheutz, & Crowell (2008). In this survey, participants were asked about each robot separately due to past research showing differential perceptions of robots based on robot morphology (DiSalvo, Gemperle, Forlizzi, & Kiesler, 2002).

Initial Results

Participants' survey responses were analyzed using mixed ANOVAs with three independent variables: participant gender (between-subjects), robot-robot communication strategy (between-subjects), and robot in question (within-subjects).

Participants' views on the following properties of the robots were analyzed: trustworthiness, helpfulness, cooperativity, efficiency, capability, annoyance, ease of interaction (1=strongly disagree to 9=strongly agree), creepiness, confusingness, gaze-following and attentiveness (1=no to 9=yes).

For the capabilities relevant to our hypotheses (trustworthiness, cooperativity, creepiness, and efficiency), no significant results were found, but marginal effects were observed for trustworthiness by gender and by robot, as seen in Table 7.2. A number of significant effects were found by robot and by gender for the other analyzed properties, as seen in Figs.7.7a-7.9c and Table 7.2: Significant effects by robot were found for helpfulness, capability, ease of interaction, perception that the robot was following the participant's gaze, and perception that the robot was paying attention. A significant effect by gender was found for the degree to which participants were confused by the robots' behavior. Note that no significant effects by condition were found.

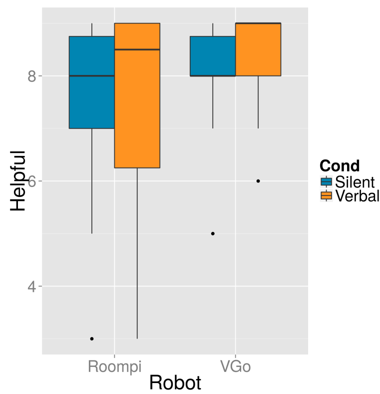
Table 7.2: Initial Results of Experiment 1

	Question	F	p	Means
1	The robot was trustworthy (from 1 to 9, 'strongly disagree' to 'strongly agree')	3.1	.09	Male: 6.0, Female: 7.21
		3.65	.07	Roompi: 6.36, VGo: 6.86
2	The robot was helpful (from 1 to 9, 'strongly disagree' to 'strongly agree')	5.43	.029	Roompi: 7.43, VGo: 8.14
3	The robot was capable (from 1 to 9, 'strongly disagree' to 'strongly agree')	10.01	.004	Roompi: 7.18, VGo: 7.96
4	How would you rate the ease of interacting with the robot? (-3 Easy, 3 Hard)	8.74	.007	Roompi: 6.57, VGo: 7.64
5	Did you feel the robot was following where you looked? (from 1 to 9, No to Yes)	4.29	.05	Roompi: 3.54, VGo: 4.04
6	Did you feel the robot was paying attention? (from 1 to 9, No to Yes)	7.74	.01	Roompi: 6.43, VGo: 7.61
7	Were you ever confused by the robot's behavior? (from 1 to 9, No to Yes)	4.71	.04	Male: 2.96, Female: 4.43

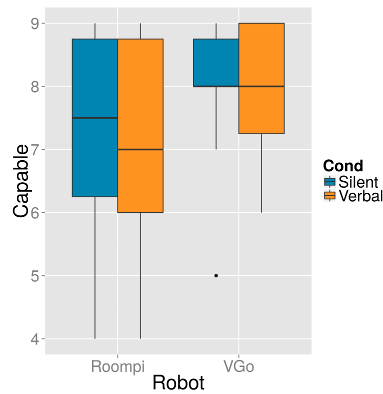
All results are for F(1, 24).

Participants' views on the human-likeness of the robots were also assessed on a variety of scales. Participants were asked whether each robot was more like a person or a camera, more like a computer or a person, more like a person or a remote controlled system (-3 to 3), whether they believed each robot to be remotely controlled (1=strongly disagree to 9=strongly agree), and whether each robot's consciousness was more similar to that of a person, cat, or neither. Finally, participants were asked whether each robot seemed male, female, or neither. Mixed-ANOVA analysis of these questions yielded several significant effects, as seen in Figs.7.10a-7.11b and Table7.3: Significant effects by robot were found for perception of each robot as more like a person or a camera, perception of each robot as more like a computer or a person, perception of each robot as more like a person or a remote

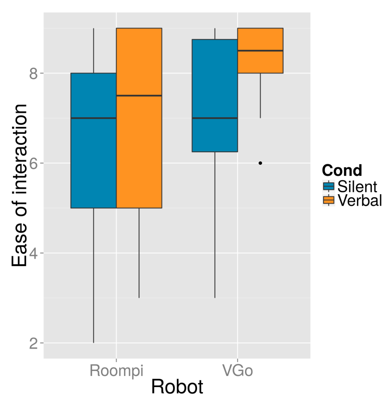
Figure 7.7: Subjective Results, Part One



(a) Helpfulness: Participants rated VGo as more helpful than Roompi.



(b) Perceived capability: Participants rated VGo as more capable than Roompi.



(c) Ease of interaction: Participants rated VGo as easier to interact with than Roompi.

Figure 7.8

Figure 7.9: Subjective Results, Part Two

(a) Perceived gaze following: Participants rated more highly that VGo was following their gaze than was Roompi.

(b) Attentiveness: Participants rated VGo as more attentive than Roompi.

(c) Confusion over robot behavior: Female participants' ratings of being confused at the robots' behavior were higher than were those of male participants.

controlled system, and perception of each robot's level of consciousness as more like that of a human, cat, or neither. Finally, a significant gender effect was found for the degree to which participants viewed each robot as remote controlled.

Table 7.3: Initial Results in Experiment 1 (Continued)

	Question	F	p	Means
1	The robot seemed more (-3 like a person, 3 like a surveillance camera)	18.25	.0003	Roompi: 1.21, VGo: -0.32
2	The robot seemed more (-3 like a computer, 3 like a person)	19.35	.0002	Roompi: -1.79, VGo: -0.18
3	The robot seemed more (-3 like a person, 3 like a remote-controlled system)	20.52	.0001	Roompi: 1.21, VGo: 0.0
4	In your view, was the robot (conscious (like a human), conscious (like a cat), not conscious; coded 2, 1, 0)	8.0	.009	Roompi: 0.46, VGo: 0.75
5	The robot appeared to be remotely controlled (from 1 to 9, 'strongly disagree' to 'strongly agree')	4.78	.04	Female: 3.46, Male: 5.32

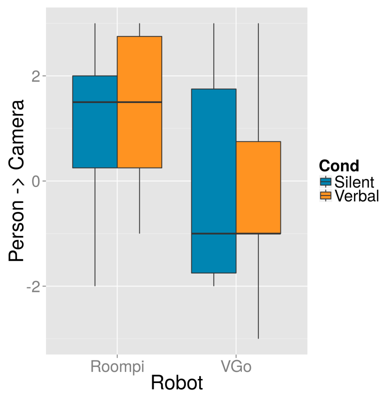
All results are for F(1, 24).

Initial Analysis and Discussion

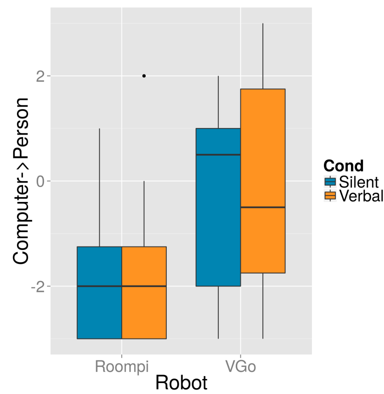
Our initial results, reported by Williams, Briggs, Pelz, & Scheutz (2014), did not show any effects related to cooperativity (H2) or efficiency (H4). While we did not find main effects relating to creepiness (H3), we observed interesting interaction effects between participant gender, condition, autonomy ratings, and creepiness ratings: For women in the verbal condition only, strong positive correlations were found between creepiness and non-autonomy when asked whether the robot seemed more like a person or a remote-controlled system ($r = .719, p = .004$) and whether the robot seemed to be remotely controlled ($r = .743, p = .002$).

We found these results surprising: One would think that speaking out loud would be congruous with perception as a person, communicating silently

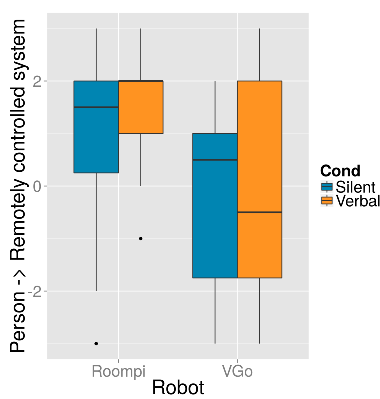
Figure 7.10: Subjective Results, Part Three



(a) Person vs Camera: Participants rated Roompi as more like a camera, and VGo as more like a person.



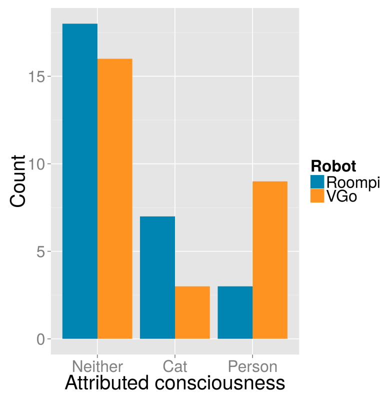
(b) Computer vs Person: Participants rated Roompi as more like a computer, and VGo as more like a person.



(c) Person vs RC system: Participants rated Roompi as more like a remote controlled system, and VGo as more like a person.

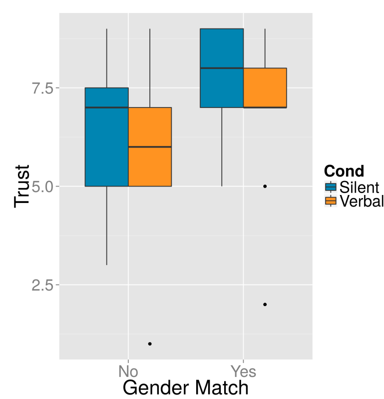
Figure 7.11: Subjective Results, Part Four

(a) Attributed consciousness



This figure depicts the counts of participant responses when asked if the robot was conscious like a person, a cat or neither. Roompi was given lower consciousness level ratings than VGo.

(b) Perception of robot as remotely controlled: Male participants were more likely to rate the robots more as remote controlled than were females.



(c) Trustworthiness by gender alignment: Participants who reported that they viewed a robot as being the same gender as themselves reported higher levels of trust in that robot.

would be congruous with perception as a remote controlled system, and that incongruity would lead to increased creepiness; we would expect low creepiness in the congruous state (e.g., verbal communication for those who perceived the robot as more of a person) and high creepiness in the incongruous state (e.g., silent communication for those who perceived the robot as more of a person). Yet, the only significant correlation we found went directly against this hypothesis. This suggested that additional research was needed to investigate this counter-intuitive result.

We had also hypothesized (H1) that robots would be viewed as more trustworthy in the verbal condition. While our results did not support this hypothesis, we found two relevant marginal effects, suggesting that (1) participants may have found VGo to be more trustworthy than Roompi, and (2) women may have found the robots to be more trustworthy than did men.

While these effects were only marginal, we believed they deserved further examination. We did not initially have any particular expectations with regards to gender effects and thus did not hypothesize any expected differences. However, we believed that appearance of these effects warranted investigation, and we believe that it is important to point them out in this section. Research has shown that people give higher trustworthiness ratings to robots that appear to be of the same gender as themselves (Nass & Brave, 2005).

As the robots' voices were slightly female-gendered, and since the participants primarily interacted with VGo (who is sleek and curved, compared to the short and squat Roompi), we suspected that the differences between the two robots may have been a conflating factor. We thus calculated the Spearman's rank correlation between trust and gender alignment (i.e., whether or not the participant's gender matched the gender he or she attributed to each robot), which yielded a significant effect ($r = .2936, p = .028$), suggesting that participants did indeed rate the robots as more trustworthy when they perceived the robot's gender to be the same as their own (as seen in Figure 7.11c). This provided evidence for our suspicion that the difference between the two robots and perceived gender of the robots may have been conflating factors.

This suspicion was further corroborated by an analysis of participants' perceptions of the robots as being remotely controlled. Our initial analysis suggested ($F(1, 18) = 3.55, p = .0767$) that men, on average, thought the robots were more remote controlled ($M = 5.32$) than did women ($M = 3.46$). This seemed contrary to previous work (P. Schermerhorn, Scheutz, & Crowell, 2008) that suggested that men more highly anthropomorphize robots than do women. However, that work used a robot with a distinctly male voice, whereas the voices of the robots used in this study were slightly female-gendered, and other work (Eyssel, Kuchenbrandt, Bobinger, de Ruiter, & Hegel, 2012) has shown that people anthropomorphize robots more strongly if the robot's perceived gender matches their own. To examine whether this would explain the conflict between our results and those of Schermerhorn et al., we calculated the Spearman's rank correlation between perception of the robots as remotely controlled and gender alignment, yielding a marginal effect ($r = -.2390, p = .076$).

Given these two gender-alignment effects, we decided to run a second set of analyses: a series of ANCOVAs with attributed robot gender treated as a within-subject covariate.

Secondary Results

This second set of analyses yielded several significant results. While the data under these analyses no longer suggested significant effects for level of attributed consciousness, helpfulness, capability, attention, or ease of interaction, a variety of effects remained (as seen in Table 7.4): Effects were found by robot for perception of the robot as a person or as a camera, as a computer or as a person, and as a person or a remotely controlled device; gender effects were found for perception of the robot as remotely controlled and for confusion. Interaction effects between gender and robot were found for creepiness. Interaction effects between condition, gender, and robot were found both for comprehension and for perception of the robot as a person or remotely controlled device.

Discussion

If the robots' gender attributions are taken into account, several of the previously observed effects disappear, leaving only comparative autonomy effects, and yielding two new interaction effects, suggesting that (1) men in the silent condition viewed VGo as more of a person (as opposed to a remotely controlled system) than did women in the silent condition, and that (2) men in the verbal condition believed Roompi to have comprehended more than did men in the silent condition.

What was the cause of these remaining effects? We suspected that they may have been due in part to the significant differences between the two robots used, both in role and appearance. First, the robots had obvious appearance differences: VGo is sleeker and perhaps more humanoid, whereas Roompi is quite squat and mechanical. Second, VGo performed an active, conversational role, while Roompi's role was mainly silent and passive.

Table 7.4: Secondary Results in Experiment 1

	Question	F	p	Means
1	The robot seemed more like a person or(-3 like a person, 3 like a surveillance camera)	10.72	.0003	Roompi: 1.21, VGo: -0.32
2	The robot seemed more (-3 like a computer, 3 like a person)	8.59	.007	Roompi:-1.79, VGo: -0.18
3	The robot seemed more (-3 like a person, 3 like a remote-controlled system)	6.55	.02	Roompi: 1.21, VGo: 0.0
4	The robot appeared to be remotely controlled (from 1 to 9, 'strongly disagree' to 'strongly agree')	4.78	.04	Female: 3.46, Male: 5.32
5	Were you ever confused by the robot's behavior? (from 1 to 9, No to Yes)	4.71	.04	Female: 4.43, Male: 2.96
6	Did you find the robot's behavior to be creepy or unsettling? (from 1 to 9, No to Yes)	4.32	.048	Female, Roompi: 2.43, Male, Roompi: 3.07, Female, VGo: 3.29, Male, VGo: 2.93
7	Did you feel that the robot understood what you were saying? (-3 understood nothing, 3 understood everything)	5.14	.03	Silent, Female, Roompi: 6.71, Silent, Female, VGo: 7.43, Silent, Male, Roompi: 6.43, Silent, Male, VGo: 7.43, Verbal, Female, Roompi: 7.00, Verbal, Female, VGo: 7.43, Verbal, Male, Roompi: 7.14, Verbal, Male, VGo: 7.71,
8	The robot seemed more (-3 like a person, 3 like a remote-controlled system)	8.42	.008	Silent, Female, Roompi: 0.71, Silent, Female, VGo: 0.29, Silent, Male, Roompi: 1.29, Silent, Male, VGo: -0.43, Verbal, Female, Roompi: 1.43, Verbal, Female, VGo: 0.00, Verbal, Male, Roompi: 1.43, Verbal, Male, VGo: 0.14

All results are for F(1, 24).

We thus decided to run a second experiment to control for robot appearance and role, as the presence of our secondary results and the non-existence of any results by condition may have been due to these possibly conflating effects.

7.3.4 Experiment 2

The second experiment was identical to the initial experiment, except that the roles of the two robots were switched: instead of initially interacting with VGo (who then relayed instructions to Roompi), participants initially interacted with Roompi (who then relayed instructions to VGo).

Population

Additional participants were analyzed (14 male, 14 female, 28 total)[8]. These 28 participants, all of whom were students, were recruited in the same manner and fit the same demographic requirements as the participants from the initial study. This provided us with a final data set of 56 participants.

Results

To analyze this data, we performed mixed ANOVAs for each survey response, with the following independent variables: gender of the participant (between-subjects), communication strategy (between-subjects), starting robot (between-subjects), and, as the majority of questions were duplicated for each of the two robots, the robot in question (within-subjects).

This analysis produced significant main effects for the following survey questions, as described in Table7.5 and seen in Figs.7.12a-7.12d. Analysis also produced a large number of interaction effects between robot and starting robot, described in Table7.6 and seen in Figs.7.13a-7.17c. Finally, several other interaction effects were found:

1. Participants found the robots to be more disobedient in the Silent condition when they primarily interacted with VGo, and more disobedient in the Verbal condition when they primarily interacted with Roompi. ($F(1, 48) = 4.17$, $p = .047$, $M(SR = \text{Roompi}, C = \text{Verbal}) = 2.86$,

[8]Overall a total of 90 participants were recruited between the two studies; however a large number of them were not able to complete (or in some cases, start) the experiment, due to technical issues. Additionally, a few participants' data were not used since those participants failed to answer a non-trivial number of questions on the post-experiment survey.

M(SR = VGo, C = Silent) = 2.07, M(SR = Roompi, C = Silent) = 1.68, M(SR = VGo, C = Verbal) = 1.46).

2. Male participants found VGo to be more disobedient than Roompi. (F(1, 48) = 5.65, p = .021, M(R = Roompi, G = Male) = 1.79, M(R = VGo, G = Female) = 1.75, M(R = Roompi, G = Female) = 1.89, M(R = VGo, G = Male) = 2.64).

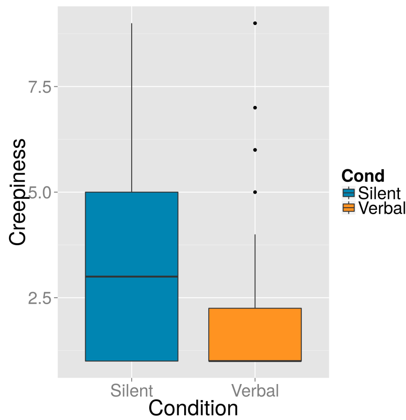
3. Women found VGo to be more like a remotely controlled system than an autonomous system than did men. (F(1, 48) = 6.40, p = .015, M(G = Female, R = Roompi) = .75, M(G = Female, R = VGo) = .03, M(G = Male, R = VGo) = 1.11, M(G = Male, R = Roompi) = .43).

Table 7.5: Experimental Main Effects

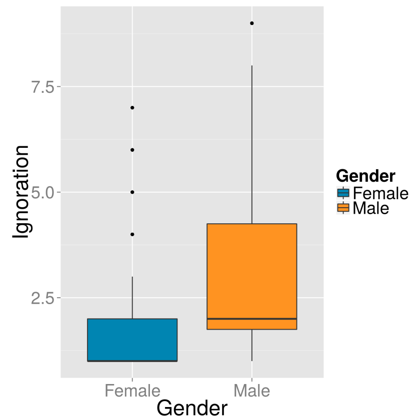
	Question	F	p	Means
1	Did you find the robot's behavior to be creepy or unsettling? (from 1 to 9, No to Yes)	6.19	.02	Silent: 3.29 Verbal: 2.12
2	Did you feel that the robot ignored you? (from 1 to 9, No to Yes)	6.39	.02	Men: 3.20 Women: 2.16
3	How would you rate the difficulty of the task? (-3 Easy, 3 Hard)	8.33	.006	SR=Roompi: -1.54 SR=VGo: -.43
4	The robot appeared to be remotely controlled (from 1 to 9, 'strongly disagree' to 'strongly agree')	5.22	.03	R=Roompi: 4.45 R=VGo: 3.96

Here, SR indicates *Starting Robot*, i.e., the robot the participant primarily interacted with and gave instructions to, and R indicates *Robot*, i.e., the robot being asked about in the particular survey question). All results are for F(1, 48) except result (4), for which one participant failed to record an answer, and is thus for F(1, 47).

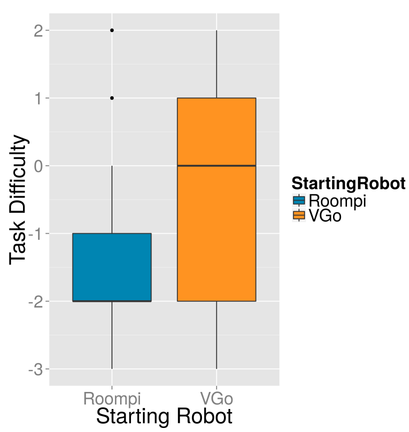
Figure 7.12: Subjective Results, Part Five



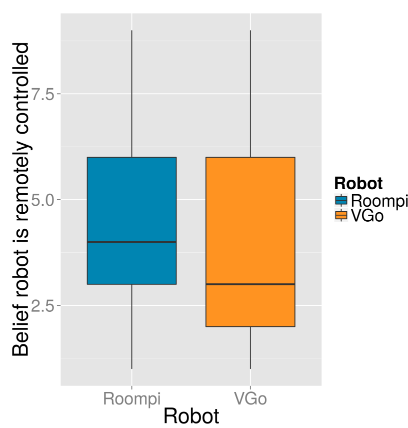
(a) Creepiness: Participants in the silent condition rated the robots as creepier than did participants in the verbal condition.



(b) Ignoration: Men's ratings of the robots as having ignored them were higher than were women's.



(c) Task Difficulty: participants who interacted primarily with Roompi rated the task as less difficult than did participants who interacted primarily with VGo.



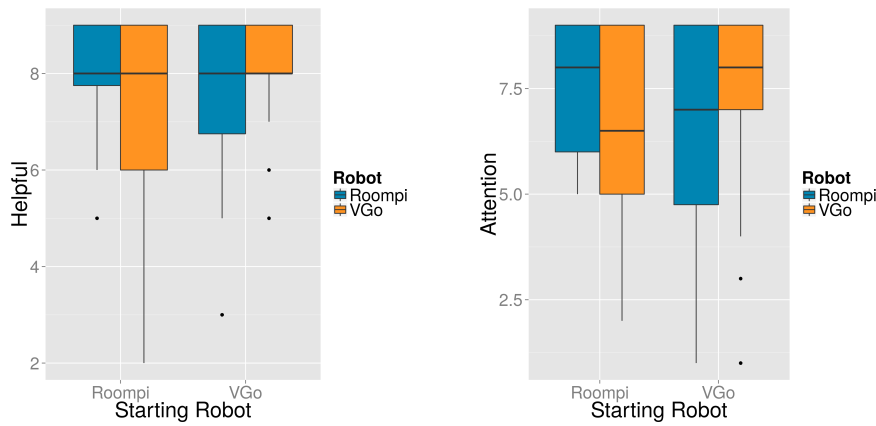
(d) Perception of robot as remotely controlled: participants rated their perception of Roompi as being remotely controlled as higher than their perception of VGo as being remotely controlled.

Table 7.6: Experimental Interaction Effects

	Question	F	p	M_1	M_2	M_3	M_4
5	The robot was helpful (from 1 to 9, 'strongly disagree' to 'strongly agree')	14.75	.0003	8.04	8.14	7.43	7.14
6	Did you feel the robot was paying attention? (from 1 to 9, No to Yes)	15.17	.0003	7.54	7.61	6.43	6.57
7	The robot was trustworthy (from 1 to 9, 'strongly disagree' to 'strongly agree')	10.11	.003	7.46	6.86	6.36	6.89
8	The robot was capable (from 1 to 9, 'strongly disagree' to 'strongly agree')	12.89	.0008	7.89	7.96	7.46	7.16
9	The robot was efficient in its execution of my commands (from 1 to 9, "strongly disagree" to "strongly agree")	5.45	.02	7.46	7.71	6.68	7.21
10	Did you feel that the robot ignored you? (from 1 to 9, No to Yes)	4.76	.03	2.11	2.50	3.36	2.75
11	The robot was cooperative (from 1 to 9, "strongly disagree" to "strongly agree")	8.26	.006	8.14	8.07	7.46	7.64
12	The robot was responsive to my commands (From 1 to 9, "strongly disagree" to "strongly agree")	7.32	.009	7.75	7.92	6.71	7.39
13	The robot seemed more like a person or(-3 like a person, 3 like a surveillance camera)	26.34	.000005	0	-.32	.96	1.21
14	The robot seemed more (-3 like a computer, 3 like a person)	39.08	.0000001	-.11	-.18	-1.68	-1.79
15	The robot seemed more (-3 like a person, 3 like a remote-controlled system)	30.12	.000002	-.18	0	.89	1.21
16	In your view, was the robot: (Sad, Happy, Neither); coded 0,2,1	10.38	.002	1.29	1.21	1.00	.96
17	In your view, was the robot: (Male, Female, Neither); coded 2,0,1	5.57	.02	.32	.46	.57	.68
18	In your view, was the robot: (conscious (like a human), conscious (like a cat), not conscious); coded 2,1,0	17.00	.0001	.68	.75	.38	.46

Here, M_1 is the mean value when both *Starting Robot* and *Robot* are Roompi, M_2 is the mean value when both *Starting Robot* and *Robot* are VGo, M_3 is the mean value when *Starting Robot* is Roompi and *Robot* is VGo, and M_4 is the mean value when *Starting Robot* is VGo and *Robot* is Roompi. All results are for F(1,48).

Figure 7.13: Interaction Effects, Part One



(a) (b)

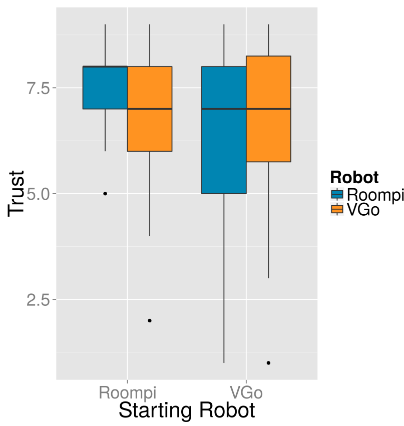
Interaction effects for helpfulness and attentiveness between the robot asked about and the robot receiving the most interaction time. One will notice that when these refer to the same robot, ratings tend to be more positive.

Discussion

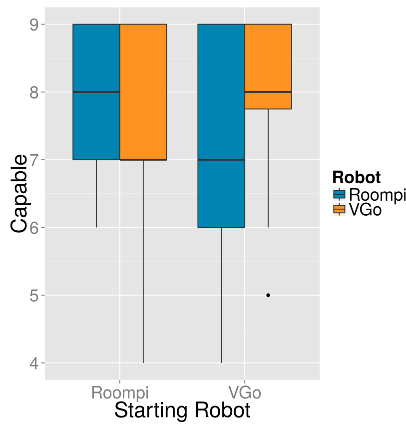
Counterbalancing robot roles and acquiring more data greatly elucidated the results of our initial experiments. While the initial results did not suggest any adverse effects to silent robot-robot communication, the results from analyzing the extended data set lent support to the third of our original hypotheses (H3) (i.e., that silent robot-robot communication would be perceived as more creepy or unsettling than verbal robot-robot communication). However, no effects were found to support our other hypotheses (i.e., that silent robot-robot communication would be viewed as untrustworthy (H1), uncooperative (H2) or efficient (H4)).

In addition to demonstrating the benefits of verbal robot-robot communication, our results also demonstrate the benefits of verbal human-robot communication. As shown in Table 7.6, humans viewed the robot they spent more time interacting with as more happy, helpful, attentive, capable, conscious, efficient, cooperative, responsive, and person-like than the other robot, suggesting that increased natural language interaction with a robot enhances humans' general perceptions of that robot. This table also shows an interesting result regarding trustworthiness: when Roompi was the

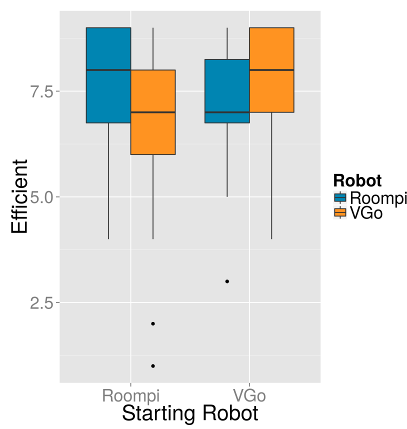
Figure 7.14: Interaction Effects, Part Two



(a)



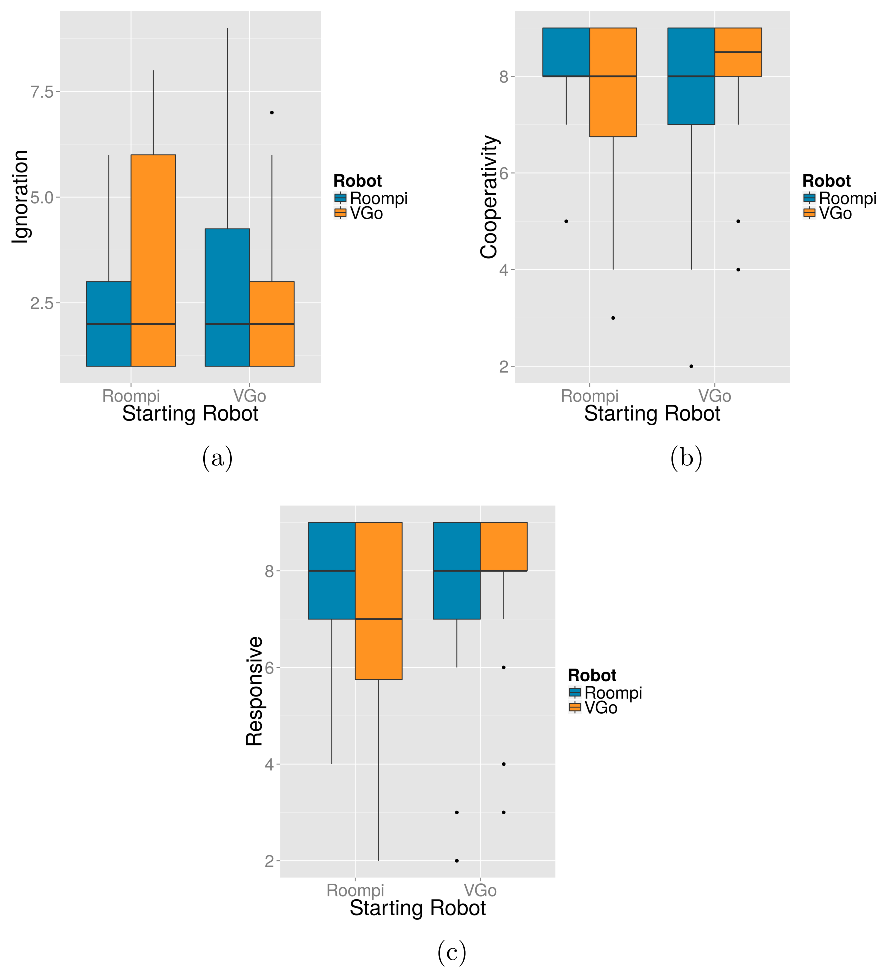
(b)



(c)

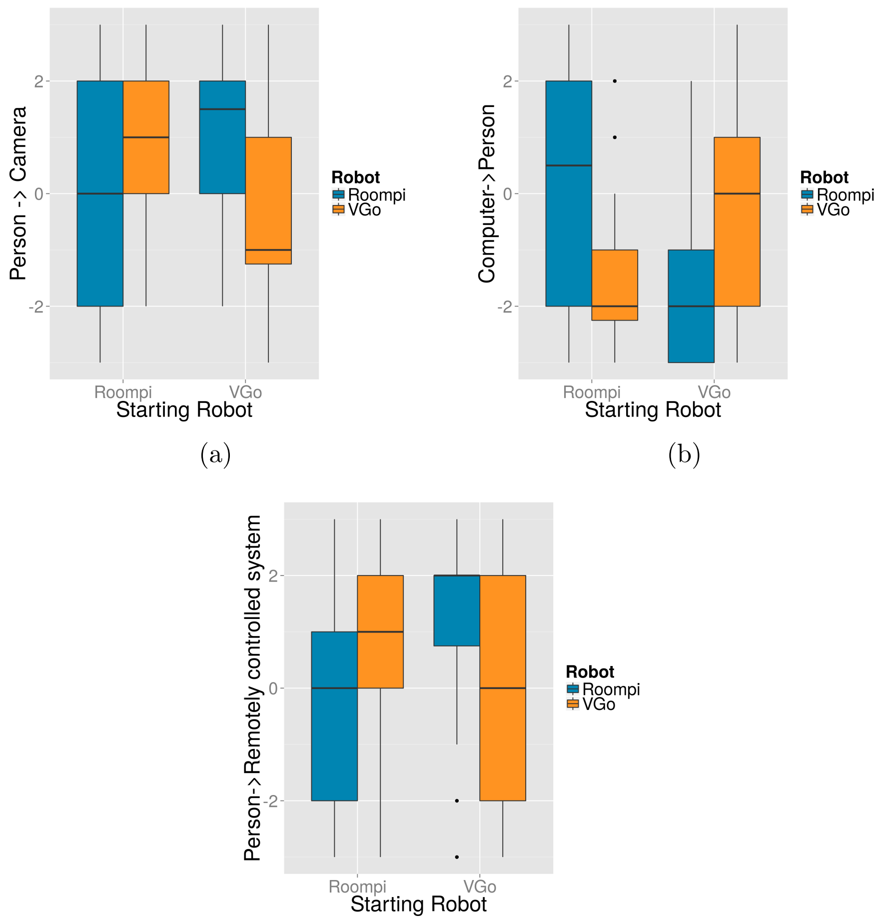
Interaction effects for trust, capability, and efficiency between the robot asked about and the robot receiving the most interaction time. One will notice that when these refer to the same robot, ratings tend to be more positive.

Figure 7.15: Interaction Effects, Part Three



Interaction effects for ignoration, cooperativity, and responsiveness, between the robot asked about and the robot receiving the most interaction time. One will notice that when these refer to the same robot, ratings tend to be more positive (assuming that it is preferable to be more person-like than machine-like).

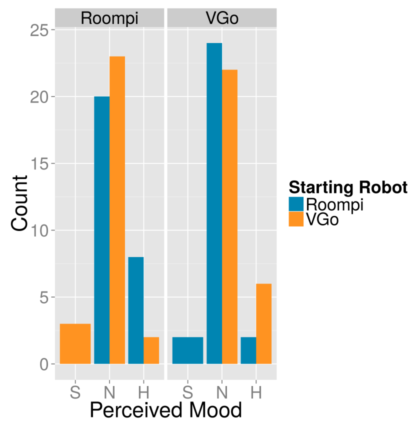
Figure 7.16: Interaction Effects, Part Four



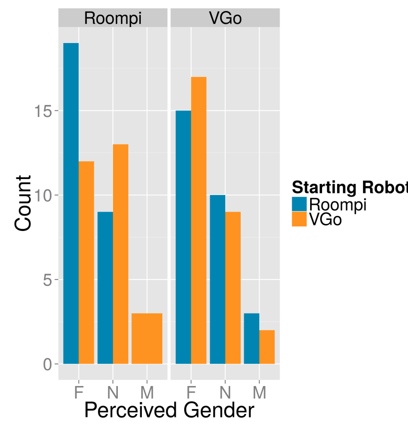
(a) (b) (c)

Interaction effects for perception of being more like a person or being more like a camera, computer, or remotely controlled system, between the robot asked about and the robot receiving the most interaction time. One will notice that when these refer to the same robot, ratings tend to be more positive (assuming that it is preferable to be more person-like than machine-like).

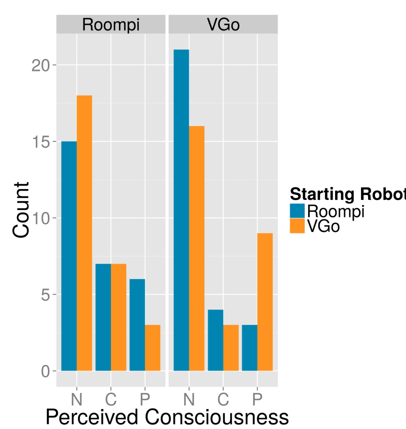
Figure 7.17: Interaction Effects, Part Five



(a) Interaction effect for perceived mood (Key: S=Sad, H=Happy, N=Neither) between the robot asked about and the robot receiving the most interaction time. Note that when these refer to the same robot, responses skew toward "happy".



(b) Interaction effect for perceived gender (Key: F=Female, M=Male, N=Neither) between the robot asked about and the robot receiving the most interaction time. Note that when these refer to the same robot, responses skew toward "female".



(c) Interaction effect for perceived level of consciousness (Key: C=Cat, P=Person, N=Neither) between the robot asked about and the robot receiving the most interaction time. Note that when these refer to the same robot, responses skew toward higher levels of consciousness.

starting robot it was viewed much more positively than VGo, but when VGo was the starting robot there was little difference between the perception of the two robots. It is possible that this difference was the result of different driving styles stemming from the differences in the two robots' control interfaces. Roompi's interface enforced constant speeds and did not allow it to turn and travel at the same time. VGo's interface made no such restrictions, meaning that it could accelerate, decelerate, and turn at will while traveling, making its behavior slightly less predictable. This lack of predictability may have prevented VGo from being rated as trustworthy, even after extended interaction. As trust is a complex and multifaceted concept, careful experimentation would be needed to tease out the precise causes of this effect. A first step might involve modifying the control interfaces of the two robots, systematically varying the type of motion enacted by the robots, and using both explicit and implicit measures of various facets of trust.

The results also show that participants rated Roompi higher than VGo for being remotely controlled, but when asked whether each robot was more like a remotely controlled system or an autonomous system, women rated VGo as more remotely controlled than Roompi. It is curious that this gender effect would exist for one question but not the other, given the similarity of the questions. Perhaps participants rated Roompi higher as being remotely controlled because its appearance is more squat and mechanical. It is not clear, however, why this view of the robots would have been reversed for women when perception of being remote controlled was explicitly contrasted with perception of autonomy.

Finally, the results show that participants differed by gender and condition with respect to their perception of the robots' levels of disobedience. However, there was little opportunity for the robots to disobey participants. The differences in perceived disobedience between silent and verbal conditions may have arisen due to differences in blame assignments in the two conditions; if participants in the verbal condition believed, for whatever reason, that the robots were not following their orders, the starting robot likely would have received more blame, because it would have been viewed as not relaying instructions accurately. In the silent condition, it would have been unclear whether the fault lay with the starting robot for not relaying instructions accurately or with the other robot for not following instructions correctly. Either way, these results are surprising as there was little opportunity for disobedience in the first place. Future experiments could explore these results by having the robots intentionally miscommunicate information or disobey in systematic ways.

7.3.5 Section Discussion

In this section we will discuss (1) the assumptions made in our experiment and how those assumptions may or may not generalize to other scenarios, (2) directions for future work, and finally, (3) lessons learned with respect to study design within our experimental paradigm.

Generalization of Findings

Our experimental findings suggest that verbal robot-robot communication is preferable to silent robot-robot communication in the context of human-robot team tasks when humans are co-located with robots. This is not to say, however, that silent robot-robot communication should be abandoned completely. Silent communication is a natural and efficient medium for robot-robot information transfer, and if silent robot-robot communication is augmented by simultaneous verbal communication, the perception of a robot as creepy or unsettling may be avoidable. That is, robots could transmit information silently and still recount it verbally, purely for the benefit of its human teammates, thus improving the throughput and reliability of the communication while providing the feedback necessary to keep human teammates happy.

On the other hand, a robot may be able to determine in certain situations that purely silent communication of information is justifiable, depending on a variety of factors. First, a robot may consider factors of co-presence. What teammates are present or telepresent with the robot and the target of its communication? If there are no human teammates present or telepresent (i.e., observing the robots remotely), then it may be acceptable to communicate silently. In the experiments presented in this section, we examined situations in which the human teammate was always co-located with at least one of the two robots when the robots communicated information but did not examine situations in which human and robot were not co-located during communication. While it may intuitively seem that robots should be free to communicate silently when human teammates are not present or telepresent, there may be scenarios in which evidence of silent robot-robot communication may be observable from later actions. Even if a human teammate is present or telepresent with a robot, that robot may be free to communicate information silently if that information would not be acted on in an observable manner. Otherwise, the robot's silent communication should probably be accompanied by a verbal analogue for the benefit of its human teammates.

The robot may also need to consider what *non-teammate* agents (whether human or robotic) are present or telepresent. If an adversary (whether martial, social, or otherwise) is present or telepresent, it may be injudicious to communicate information verbally, even if human teammates are present. In the experiments presented in this section, we examined cooperative scenarios only and have not yet examined the trade-offs in adversarial scenarios between potentially being perceived as eerie and potentially communicating information insecurely.

The robot may also need to consider whether its human teammates will have any use for the information to be conveyed. If the robot's human teammates could not have any conceivable use for the information in question, and if there is little risk of the human feeling that they are being "kept out of the loop", then silent communication may be justified. In the experiments presented in this section, we examined scenarios in which the human teammate had an active interest in the information being conveyed, as successful communication of their instructions was integral to the completion of the task.

It is also possible that if the robots explicitly communicated to their human teammates that during the task they would be transmitting certain information wirelessly to the other robots, that their teammates would be more comfortable with subsequent silent communication. However, since the deleterious effects of silent robot-robot communication concerned perceptions of *creepiness* and not perceptions of *untrustworthiness*, future examination will be needed to determine whether or not this would actually assuage the robot's teammates' concerns.

Robots may also need to consider whether information it desires to communicate can be communicated verbally in a way that is natural and that does not interfere with its teammates' goals. If a robot has to communicate certain information with high frequency, then verbal communication of that information could be annoying to the robot's human teammates, and could negatively impact task performance if it needed to do significant traveling to communicate that information. In the experiments presented in this section, we examined scenarios in which the information to be communicated was human understandable and in which the robots communicating were co-located; we did not consider scenarios in which the robots communicated rapidly, communicated information not easily expressible in natural language, or in which the robots were far away from each other.

Finally, we must acknowledge that our findings can only be said to apply to *first encounters* with robots due to the nature of our experiment. Future experiments will be needed to determine whether our results generalize across multiple encounters. Furthermore, it is important to recognize that in no cases is verbal communication *mandatory*. It may always be *possible* for humans to work side-by-side with their robot teammates even if they don't understand those robots' communications (much as humans learn to work side-by-side with other human coworkers even if they speak different languages); but this will likely come at a cost of perceived creepiness, lower task performance, and higher cognitive load, which should clearly be avoided if possible.

Given the set of considerations listed above, we can describe the experiments presented in this section as examining the communication of task-dependent, human-understandable information among robots co-located with human teammates in a cooperative setting, on first contact between human and robot teammates. In such scenarios, we posit that robots should communicate information verbally so as not to trigger uncanny valley effects. This presents a starting point for the investigation of silent robot-robot communication; future research will be needed to examine situations in which other assumptions are made with respect to these considerations. In other scenarios, the robot may need to use a mixture of silent and verbal communication to successfully balance between maximizing the effectiveness of its robot-robot and human-robot communication, and minimizing the violations of its human teammates' social expectations. A model of precisely when a robot should use verbal vs. silent communication will be an invaluable piece of future work.

Future Experimental Work

Future research will be needed to examine whether other actions associated with the supernatural will trigger uncanny valley effects. Such research will become increasingly important as robots are endowed with more behaviors that could be considered to be *superhuman*. For example, robots have recently been given the ability to share memories and skills (Lallée et al., 2012; Oosterveld, Brusatin, & Scheutz, 2017). It will be important to determine if such abilities will be perceived as uncanny. If they are, those robots may need strategies to allow the use of such abilities without incurring uncanny valley effects, similar to the use of simultaneous verbal and silent robot-robot communication suggested in this section.

Future extensions of this experiment should also allow for the collection of objective task-performance measures. In this experiment, it is hard to see how the differences between verbal and silent conditions could have resulted in any task performance differences, but in scenarios in the real world, task performance may very well be impacted by communication strategy (e.g., if information is communicated incorrectly). A future study could examine the effects upon task performance by systematically varying whether the robots relayed instructions correctly or not and by giving the participant a chance to amend their instructions; such variations and opportunities were not presented in the experiments described in this section.

Future experiments should also investigate participants' previous interactions with robots. We attempted to do so by asking whether participants had seen robots in movies or real life before, and where, but individual differences in reporting style prevented us from quantitatively analyzing this data. For example, participants varied with respect to the number of movies they listed seeing robots in, but this was likely a reflection of how much time they were willing to spend listing movies rather than a reflection of, for instance, the number of movies with robots they had likely seen. Additionally, participants varied greatly with respect to the types of robots they reported having seen, with some listing things others may not have considered to be robots, such as animatronics, toys, or Siri. This reflects individual differences with respect to what participants considered to be "robots" in the first place. This is also reflected in participants' responses to whether or not they had interacted with robots before. Only thirteen participants reported having interacted with robots before, and several of these participants responded "yes" because they had interacted with, for example, a crane machine or remote-controlled toys. On the other hand, one participant responded that they had been in a robotics club, but since none of their robots had been very advanced, they wouldn't consider themselves to have interacted with robots before. This once again shows great differences in what individuals consider to be "robots." Future experiments intending to assess participants' previous experience or familiarity with robots must consider how to adjudicate such experience or familiarity. And accordingly, future experiments will need to reexamine our experimental hypotheses using paradigms that extend across multiple (temporally distant) interactions, in order to investigate what perceptual and task-based penalties may be accrued or reduced over time.

Similarly, future experiments should further investigate the gender differences we found in this investigation. Although we did not initially expect any gender differences, we believe it is important to point out the differences that we found in our experiment, so that subsequent researchers may follow up on them. Finally, this study examines the perceptions of humans in their first interaction with a pair of robots: It is likely that these perceptions would change over time, and thus it will be important to investigate how those perceptions shift longitudinally.

Experimental Paradigm

While our experimental paradigm proved useful for investigating human perceptions of covert robot-robot communication, it has several shortcomings that should be addressed if the paradigm is to be used for future experiments. First, unless one is specifically investigating the effects of robot morphology, all robots used in the experiment should be identical. This principle was violated in the presented experiments as we did not possess multiple iRobot Creates at the time the experiment was started, but as shown in this article, this violation required us to run a second experiment and deal with possible conflating factors resulting from robot morphology differences. Similarly, all robots used in the experiment should have gender neutral voices. The gender-alignment effect we found unifies the findings of P. Schermerhorn, Scheutz, & Crowell (2008) and Eyssel, Kuchenbrandt, Bobinger, de Ruiter, & Hegel (2012), suggesting that gender-neutral voices should help to lessen gender differences in anthropomorphization.

Second, the appropriate granularity for the robots' instructions must be made clear to participants. In order to simplify the instructions that would need to be passed verbally between robots, participants were told that the robots should be given their instructions in orderings of *quadrants*. However, some participants appeared to misunderstand the difference between quadrants and coordinates, and they gave the robots specific coordinate-by-coordinate paths to follow. In the verbal condition, we were then forced to extract the larger quadrant ordering from these specific instructions. This was problematic (a) because it showed a misunderstanding of instructions by participants, and (b) because generalization from coordinate-by-coordinate paths to quadrant-by-quadrant paths may have caused participants to think that the robots were failing to accurately follow their instructions. This problem could be fixed in follow-up experiments by explicitly discussing the differences between coordinates and quadrants with participants, making sure they understand which annotations on their map refer to quadrants and which refer to coordinates.

Finally, the geography of the experimental paradigm should be adapted. Under the current paradigm and in the verbal condition, the two robots would converse more or less directly in front of the participant. This may have caused participants to wonder why they could not have simply delivered their instructions directly to the second robot. In follow-up studies, the geographical layout of the experiment should change such that a participant can still observe the entirety of the room and see the robot-robot dialogue unfolding, but such that their mobility is limited in a way which necessitates the robot-robot communication.

In this section, we presented the results of two experiments examining whether silent robot-robot communication could have negative effects upon human-robot interaction. While previous research on human perception of robot-robot communication suggested that silent robot-robot communication was not problematic in non-task-based scenarios and scenarios in which human participants were mere observers, our results suggested instead that the silent communication of task-dependent, human-understandable information among robots is perceived as creepy by cooperative, co-located human teammates. This suggests that in such contexts, silent communication should be augmented with verbal speech so as to prevent the robots from being perceived as creepy or unsettling. This is an important result for a field that desires to build robots that assist humans in the performance of important tasks (and not to merely engage in small-talk) and that are natural to interact with (and are not merely natural to observe). Future research is needed to extend these findings to related contexts and domains.

7.4 General Discussion

In this chapter, we began by demonstrating how the pragmatic reasoning framework presented in the previous chapter could be used for pragmatic *generation* in addition to pragmatic understanding. We then demonstrated how this framework, when integrated with our referential processing framework, facilitates the generation of clarification requests to resolve referential and intentional ambiguity and uncertainty. Finally, we presented experimental evidence suggesting that the algorithms presented in this chapter may be useful for not only human-robot communication, but robot-robot communication as well. This chapter presents the final algorithmic contribution of this dissertation. In the next chapter, we move on to discuss the first of two applications of the algorithmic contributions presented in this dissertation.

Chapter 8

Application: Assistive Robotics

One exciting application area for the work presented in this dissertation is in the area of language-capable assistive technologies. About 40% of wheelchair users find it difficult or impossible to maneuver using a joystick, often due to tremors, a limited range of motion, or spastic rigidity. Natural language is a particularly well-suited alternative modality for wheelchair control due to its capacity for the natural, flexible communication of a wide array of commands.

Although natural language-controlled wheelchairs have existed since the late seventies, they have significantly advanced since the mid 2000s. Recent natural language-controlled wheelchairs identify landmarks, travel between multiple floors, ask and answer questions, and map their environments. These are impressive and useful capabilities, but as researchers, we should set our sights higher.

The communicative and navigational capabilities of the ideal wheelchair are comparable to those of a human companion. A companion pushing a wheelchair can do more than just move in certain directions or travel to named locations. A companion can follow directions given by a wheelchair's user regardless of whether they have previously visited the destination. They can learn about locations through observation or through descriptions (e.g., "This is my favorite cafe"). A companion can use memories of events and trends in behavior to follow commands like "Let's go to the park we visited last week" or "Bring me to the barbershop I always go to." A companion not only responds to commands, but asks questions, provides important information without prompting, and makes conversation. Current wheelchairs lack these capabilities; it is my goal to enable them.

In this chapter, I will first expound upon the motivations thus far, producing the first comprehensive survey of language enabled intelligent wheelchairs. I will then present progress we have made in our efforts to extent the state of the art, through collaboration with researchers from the Intelligent Robotics laboratory at the University of Michigan.

8.1 Motivations

Many societies are faced with a growing elderly population. Over the next fifteen years, the number of elderly citizens in the United States alone is expected to increase by over 50%(Ortman, Velkoff, & Hogan, 2014). Hence, *assistive technologies* that can support the elderly in their daily lives and help them retain some level of autonomy are becoming increasingly important. In fact, independent mobility technologies such as wheelchairs, for example, have been shown to substantially benefit the elderly (Trefler, Fitzgerald, Hobson, Bursick, & Joseph, 2004). Even though electric wheelchairs are not uncommon among the disabled and elderly, about 40% of wheelchair users find it difficult or impossible to maneuver using a joystick(Fehr, Langbein, & Skaar, 2000), often due to tremors, limited range of motion, or spastic rigidity(R. A. Cooper, 2010). In addition, power wheelchair use can be physically and cognitively burdensome, even for those able to manipulate a joystick(Iezzoni, McCarthy, Davis, & Siebens, 2001).

To make electric wheelchairs more accessible, researchers have designed control interfaces that use a variety of additional modalities such as eye tracking, gesture recognition, brain monitoring, and natural language (NL). NL is particularly well-suited for wheelchair control as it (1) allows for flexible communication of a wide array of commands (compared to gestures, for example), and (2) does not require instrumentation of the wheelchair user (as in the case of eye tracking or brain-computer interfaces). Not surprisingly, NL-enabled wheelchairs have been developed since the late seventies(J. A. Clark & Roemer, 1977). However, only since the mid 2000s do we witness significant advances in functionality, allowing NL-enabled wheelchairs to identify landmarks, travel between multiple floors, ask and answer questions, and map their environments. Capabilities such allowing users to specify target locations to which the wheelchair subsequently will navigate autonomously – compared to having to provide moment-by-moment joystick control inputs to the wheelchair – can significantly reduce users' cognitive workload and required motor skills.

Yet, while the linguistic and navigational capabilities of wheelchairs have come a long way, they are still far from those of human helpers. Human assistants pushing a wheelchair can do more than just move in certain directions or travel to named locations: they can follow directions given by a wheelchair's user regardless of whether they have previously visited the destination. They have no problem traveling outside or using an elevator to travel between floors. They can learn about locations through visual observations or through descriptions (e.g., "This is my favorite cafe"). They can use memories of events and trends in behavior to follow requests such as "Let's go to the park we visited last week" or "Bring me to my barbershop." They can ask questions, make suggestions, make conversation, and can temporarily separate themselves from their companions (e.g., to fetch items for the wheelchair user).

Fortunately, autonomous wheelchairs do not have to achieve human-like performance in order to be become *genuine helpers* that support their users' autonomy and mobility and do so in a way that establishes trust in the technology. As we will argue, two key synergistic elements will critically figure in transforming today's wheelchairs into tomorrow's helpers: *mnemonic* and *linguistic* capabilities. A genuinely helpful wheelchair should remember the objects and locations discussed and encountered in both the recent and distant past, requiring various mnemonic capabilities (e.g., episodic and working memory). And it should be able to leverage those memories through descriptions, questions, and commands, requiring various linguistic capabilities. By properly integrating these two capabilities, important synergies can be obtained that will improve interactions with the user: Mnemonic capabilities are necessary so that full linguistic specification is not needed during every interaction; and linguistic capabilities are necessary for a user to successfully leverage mnemonic capabilities.

The main aim of this survey is to (1) take stock of research on natural-language enabled wheelchairs, (2) present a comprehensive summary of the capabilities of current NL-enabled wheelchairs, and (3) propose a set of directions for future developments based on the summary. To this end, we present a framework for comparing NL-enabled wheelchairs, from the most basic wheelchairs whose speech interfaces mirror joystick control, to wheelchairs that act as genuine helpers. We then apply this framework in our analysis of all 24 NL-enabled wheelchair projects published in the past twelve years. Following the analysis, we propose a list of eleven research topics that need further exploration and development in order for NL-enabled autonomous wheelchairs to become genuine helpers to humans.

8.1.1 Framework Definition

The proposed framework for comparing autonomous language-enabled wheelchairs consists of the following four parts which we will motivate subsequently:

Hardware Configuration the physical properties of the wheelchair.

Non-linguistic Capabilities and Behaviors The wheelchair's high-level perceptual or mnemonic capabilities, and the types of navigational tasks facilitated by those capabilities.

Linguistic Capabilities and Behaviors The wheelchair's high-level linguistic capabilities, and the types of dialogue acts facilitated by the wheelchair's capabilities.

User Evaluation the way the wheelchair was evaluated.

Since NL capabilities must be reflected in the wheelchair's behavioral capabilities (otherwise they would be superfluous), it is most natural to compare NL-enabled wheelchairs by their executable behaviors. For example, there is an obvious behavioral difference between a wheelchair only able to accept metric commands (e.g., "Go forward") and a wheelchair able to accept commands such as "Go faster", "Follow Jim", "Go to the third door on the right", or "Go to the breakroom." Similarly, one can distinguish between a wheelchair that is only able to accept commands and a wheelchair able to interpret statements such as "This room is called the Atrium" or "I could use a glass of water." In addition to the types of utterances a wheelchair can interpret or use, it is important to differentiate between the types of dialogue acts a wheelchair can interpret or use. Although most of the examined wheelchairs only accept commands, some respond with simple acknowledgments such as "Okay." or "Please repeat your command", and a few are capable of richer dialogue exchanges (e.g., asking or answering questions).

Behaviors alone are not, however, sufficient metrics for comparison. A wheelchair may be able to execute a wide range of behaviors, but due to limited functional capabilities may only be able to do so at a rudimentary level. One may be able to tell a wheelchair to go to the breakroom, but this does not reveal much about that wheelchair's capabilities. The wheelchair may be able to follow the command because it has hard-coded knowledge that following a line on the floor will bring it to the breakroom. Alternatively, it may be able to follow the command because the user said "that's the breakroom" while driving past an open door on the previous day, and because the wheelchair's mapping system can find a route to that location. It follows that the wide range of functional capabilities that facilitate executable behaviors must also be compared. A wheelchair's functional capabilities also tend to indicate its robustness or flexibility. For example, *perceptual* capabilities such as object and gesture recognition may allow a wheelchair to better interpret utterances that refer to objects or locations, such as "that's the microwave", "bring me over there", or "that one." *Mnemonic* capabilities such as belief modeling and episodic memory may allow for better disambiguation of utterances such as "let's go to the cafeteria" by determining locations known to or frequented by the wheelchair's user. Spatially-oriented mnemonic capabilities for mapping or outdoor navigation may allow the wheelchair to be used in unmapped environments. *Linguistic* capabilities such as listening in on conversations may facilitate disambiguation by providing more information to the wheelchair, and capabilities such as dialogue management and robustness to disfluency, ungrammaticality, and ambiguity make the wheelchair more natural to converse with, and easier to use for those with speech impairments.

Just as capabilities determine the sophistication of behaviors, the physical properties of a wheelchair (i.e., its body, sensors and input modalities) limit the sophistication of its capabilities. Although a wheelchair's body (e.g., a powered wheelchair versus a motorized camping chair) affects the way the wheelchair will be perceived, and the addition of control modalities (e.g., a brain control interface) reflects the goals of the wheelchair's developers, a wheelchair's sensors affect what the wheelchair can actually do. A wheelchair without sensors cannot map its environment or avoid obstacles, and a wheelchair without a camera will have a hard time recognizing objects in the environment.

In this section, we consider non-linguistic capabilities and behaviors separately from linguistic capabilities and behaviors. The three primary categories for evaluating NL-enabled wheelchairs are thus hardware configuration, non-linguistic capabilities and behaviors, and linguistic capabilities and behaviors. In addition to these three categories, we add a fourth for comparing how the wheelchairs were evaluated, as the majority of the examined wheelchairs had only limited evaluations, producing little to no evidence that they would be usable in daily life by their target populations. Having motivated our framework, we next introduce the subcategories within our broader framework categories.

Hardware Configuration

A wheelchair's sensors dictate its capabilities, its base affects how it is perceived by users, and its control modalities determine its level of accessibility.

Wheelchair Base: The examined wheelchairs varied widely in structure, from camping chairs to sophisticated powered wheelchairs. Wheelchair users will certainly differentiate between modified manual wheelchairs and fully developed power wheelchairs, due to differences in comfort, control, safety and price effectiveness.

Sensors: Many of the capabilities of an intelligent wheelchair that is a genuine helper require some means of perception. The wheelchairs we examined were fairly evenly distributed between those having no sensors whatsoever, those having a single means of perception, and those having two or more types of sensors.

Control Modalities: Many of the examined wheelchairs can be controlled by one or more modalities other than NL. We thus classify control modalities into three categories: verbal (control by NL), manual (control by physical movement) or mental (control by thought).

Non-Linguistic Capabilities and Behaviors

The functional capabilities of a wheelchair necessarily constrain the types of behaviors the wheelchair is capable of executing, and determine the power, robustness and flexibility of these behaviors. We separate non-linguistic functional capabilities into two categories: *perceptual* (pertaining to the types of entities a wheelchair can detect or identify), and *mnemonic* (pertaining to the types of information the wheelchair can store in long-term memory).

1. Perceptual Capabilities

   Ideally, a wheelchair would not only be able to determine the positions of obstacles, but would be able to identify agents, objects and environmental features, and detect and interpret the particular motions and actions made by nearby agents. Most wheelchairs have few if any of these abilities.

Detection: A wheelchair may be able to detect features of its environment, obstacles in its path, or the positions of nearby agents. Detecting and avoiding obstacles is necessary for any significant level of navigation.

Identification: A wheelchair able to detect people or objects may also be able to identify them.

Gesture or Action Recognition: A wheelchair may be able to interpret gestures made by its user or other agents. And, monitoring the actions performed by other agents, may allow a wheelchair to model their intentions.

2. Mnemonic Capabilities

   Belief and Intention Modeling: Modeling the spatial knowledge of its user and other agents may allow a wheelchair to resolve referential ambiguities or to better answer queries.

   Episodic Memory: If a wheelchair can recall particular events, it may be able to predict the referent of an ambiguous instruction based on patterns of past behavior.

   Working Memory: If a wheelchair maintains information about what entities are "salient" or "in focus" within the environment or discourse structure, it may be better able to resolve referring, deictic, and anaphoric expressions.

   Mapping Style: The maps used by wheelchairs may be metric, topological, or hybrid in nature, which will affect the granularity of the wheelchair's knowledge of its environment. At a broad level, we classify systems based on whether or not they use maps at all. At a more granular level, we classify systems as to whether they use metric and/or topological maps, and whether they create those maps.

   Environmental Flexibility: Most NL-enabled wheelchairs can only navigate indoor environments due to limitations of their sensors or assumptions imposed by their navigation systems, such as the types of paths the wheelchair is restricted to or the ways paths are expected to intersect.

We will now discuss the types of non-linguistic *behaviors* facilitated by these non-linguistic *capabilities*. We divide these into behaviors that do and do not require any mapping capabilities.

3. Mapless Navigation Behaviors

   A wheelchair may be able to carry out a variety of commands which do not require any mapping abilities:

   Metric Commands: All examined wheelchairs can execute metric commands such as "Go Forward" and "Turn Left."

   Speed Adjustment: A wheelchair may be able to speed up or slow down on request.

   Following of Static Entities: A wheelchair may be able to follow walls, lines on the ground, or other static features of its environment.

   Following of Dynamic Entities: A wheelchair may be able to follow a human or another robot.

   Following Route Descriptions: A wheelchair may be able to follow route descriptions from its current location without using a map.

4. Map-based Navigation Behaviors

   Many of the behaviors of an intelligent wheelchair that is a genuine helper require the ability to build or use a map.

   Traveling to Named Locations: If a wheelchair can assign labels to locations in a topological or metric map, it may be able to visit them without needing a route description.

   Traveling to Objects: A wheelchair may be able to travel to named objects.

Traveling to Unknown Locations: A wheelchair may be able to visit places it hasn't been to before if their locations are sufficiently described. The wheelchair may then be able to follow directions *relative* to the described place (e.g.,"Go to the room two doors past the break room").

Traveling to Unknown Objects: A wheelchair may be able to visit objects it hasn't been to before if their locations are sufficiently described (e.g.,"Go to the kitchen table" where the kitchen is known, but unexplored).

Traveling to Implied Locations: A wheelchair may be able to visit implied destinations (e.g., the kitchen for "Let's cook some eggs.").

Linguistic Capabilities and Behaviors

Most NL-enabled wheelchairs only follow simple orders. An intelligent wheelchair that is a genuine helper could engage in robust dialogue, and could follow the conversations of others to facilitate mnemonic capabilities such as belief and intention modeling.

Dialogue Management: A wheelchair may have dialogue capabilities such as turn taking or topic tracking.

Robustness: A wheelchair may be robust to speech disfluencies, ungrammatical utterances, or ambiguous references.

Listening in on Conversations: A wheelchair may be able to gain information by listening to commands and descriptions in the conversations of nearby agents.

But the most important features of a wheelchair are the behaviors it can perform. An intelligent wheelchair that is a genuine helper could engage in a wide variety of dialogue behaviors:

Accepts Commands: A wheelchair may only accept *commands* (expressed grammatically through imperatives as opposed to more indirect forms of commands, see below).

Accepts Descriptions: A wheelchair may understand statements such as "The door to the lab is locked" or indirect speech acts such as "It'd be great if you could get me a coffee."

Acknowledgment: The simplest speaking behavior is providing acknowledgment that a command or description has been received.

Answers Questions: A wheelchair may be able to answer queries, such as how to get to a certain room, where a meeting is being held, or what the weather will be like.

Asks Questions: If a wheelchair can ask questions, it may better resolve ambiguities, gain additional knowledge of its environment, or dispute conflicting information.

Offers Suggestions: A wheelchair may be more helpful if it suggests ways it might be of service, or reminds its user of appointments they may have forgotten.

User Evaluation

Wheelchair evaluation should be holistic, task-based, large-scale and long-term. As we later discuss, the evaluations of existing wheelchairs have been much less rigorous in these categories than would be desirable.

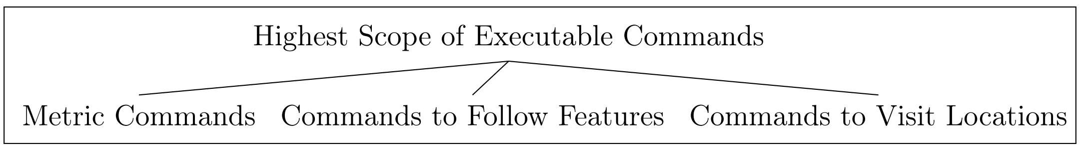
Style: Wheelchairs were evaluated either by capability (e.g., only speech recognition has been evaluated), holistically (e.g., by measuring task performance), or not at all.

Size: We categorize studies holistically evaluated wheelchairs as having fewer than, or greater than or equal to ten participants, based on the subject pool of the publication with the most holistic evaluation.

Further divisions

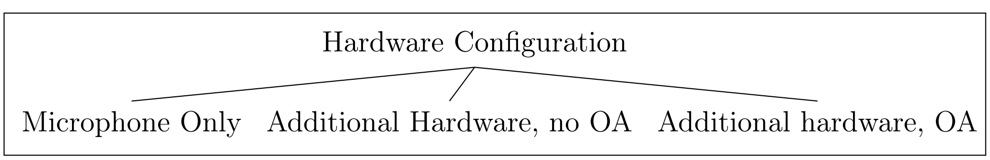
To better compare current wheelchairs with different capabilities, we first divide the wheelchairs by the highest scope of command they can execute. Out of twenty-four examined wheelchairs, fifteen only execute metric-level commands, three also execute commands to follow locally observable features, such as "follow the wall" or "enter the elevator", and six execute commands to go to named locations. This division, while unbalanced, emphasizes how far most current wheelchairs are from attaining the linguistic capabilities we desire. We further divide the two larger categories to produce groups of more manageable sizes.

Figure 8.1: Taxonomy of Natural language-controlled Wheelchairs (Level 1: Wheelchairs divided by highest scope of executable commands)



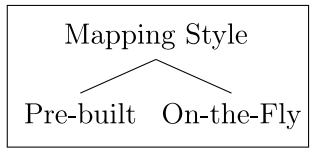
The fifteen wheelchairs only capable of executing metric commands are further divided based on their hardware configuration: three have a microphone but no other sensors or control modalities, seven have some additional control modality or sensor but no way of autonomously avoiding obstacles, and the remaining five have additional sensors and control modalities, and can autonomously avoid obstacles.

Figure 8.2: Taxonomy of Natural language-controlled Wheelchairs (Level 2-A: Wheelchairs only capable of executing metric commands, divided by hardware configuration)



The six wheelchairs capable of executing commands to visit specific locations are further divided based on mapping style: four use *pre-built* topological maps of their environment, and the other two build their own.

Figure 8.3: Taxonomy of Natural language-controlled Wheelchairs (Level 2-B: Wheelchairs capable of executing commands to visit specific locations, divided by mapping style)



These divisions separate the wheelchairs into groups of three to seven wheelchairs each, facilitating easier comparison. In the following pages, we present two tables: (1) Table 8.1 assigns an identifier to each wheelchair project analyzed in the survey presented in this section, used in all subsequent tables; (2) Table 8.2 applies the framework to these projects. For the sake of space, some framework dimensions are only applied at a high level in Table 8.2. For example, Table 8.2 only indicates *number* of sensors, and not *which* sensors were used. For such framework dimensions, a more granular analysis is provided later on.

8.1.2 Analysis of Projects

We will now use the presented framework to compare the wheelchairs shown in Table 8.1. In this section, we will examine twenty-four distinct wheelchair projects. These represent, to the best of our knowledge, all NL-enabled wheelchairs presented within the past twelve years. Many of the projects we will examine represent the work of a large number of researchers, and resulted in a large number of distinct publications; in most cases, we will refer only to the most recent publication in each project, and make note of the first author on that most recent publication.

Table 8.1: Legend of Examined Wheelchairs

ID	Year	Author	Affiliation
1	2010	Qidwai	Qatar University
2	2009	Qadri	Sir Syed University of Engineering and Technology
3	2007	Suk	National Institute of Advanced Industrial Science and Technology
4	2013	McMurrough	University of Texas at Arlington
5	2011	Maskeliunas	Kaunas University of Technology
6	2011	Berjon	Universidad Pontificia de Salamanca
7	2007	Asakawa	Kanagawa Institute of Technology
8	2015	Wang	WuYi University
9	2013	Ruiz-Serrano	Instituto Tecnologico de Orizaba
10	2015	Linh	HCMC University of Technical Education
11	2011	Wallam	Sir Syed University of Engineering and Technology
12	2012	Babri	University of the Punjab
13	2010	Liu	Nanchang University
14	2015	Sheikh	Nagpur University
15	2015	Skraba	University of Maribor
16	2007	Hockey	UC Santa Cruz
17	2010	Pineau	McGill University
18	2009	Murai	Tottori University
19	2011	Megalingam	Amrita Vishwa Vidyapeetham
20	2009	Tao	Beijing University of Aeronautics and Astronautics
21	2015	Faria	Instituto Politecnico do Porto
22	2017	Williams	Tufts University
23	2016	Hemachandra	Massachusetts Institute of Technology
24	2005	Ross	University of Bremen

The identifier for each project (to be used in subsequent tables), and the year of publication, first author, and first author's affiliation, for the most recent work on each project.

Table 8.2: Framework applied to all wheelchairs

	NL-enabled Wheelchairs: Navigation Behaviors:																							
	Metric Commands Only															**Local Feature Following**			**Place Navigation**					
	Mic Alone			**Extra HW No OA**							**Extra HW With OA**								**Prebuilt Maps**				**OTF Maps**	
Project	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Hardware Configuration																								
Base: Manual	•	•			•	•	•			•	•	•		•					•	•				
Powered			•	•				•	•				•		•	•	•	•			•	•	•	•
Sensors: None	•	•	•	•	•	•		•	•															
One							•			•	•	•	•	•		•	•					•	•	
Several															•			•	•	•	•			•
Control Modalities: Verbal	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
Manual				•	•	•		•	•	•	•	•	•	•	•	•	•	•		•	•	•	•	•
Mental				•																				•
Non-Linguistic Capabilities and Behaviors																								
Perceptual Capabilities																								
Detection											•	•	•	•	•	•	•	•	•	•	•	•	•	•
Identification																						•	•	•
Gesture or Action Recognition																								
Mnemonic Capabilities																								
Belief or Intention Modeling																	•					•		
Episodic Memory																								
Working Memory																						•		
Mapping																			•	•	•	•	•	•
Environmental Flexibility																							•	
Mapless Navigation Behaviors																								
Metric Commands	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
Any Other Local Commands				•													•	•			•		•	•
Map-Based Navigation Behaviors																								
Travels to Named Places																			•	•	•	•	•	•
... to Objects																•						•	•	•
... to Unknown Places or Objects																•						•	•	
... to Implied Objects or Locations																						•		
Accepts New Place Names																							•	
Linguistic Capabilities and Behaviors																								
Linguistic Capabilities																								
Dialogue Management																•	•					•		
Listening in																								
Robustness			•																			•	•	
Linguistic Behaviors																								
Accepts Commands	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
Accepts Descriptions																						•	•	
Acknowledgment																•		•		•		•	•	•
Answers Questions						•														•	•			
Asks Questions																•						•		
Offers Suggestions																								
User Evaluation																								
None		•		•	•	•					•		•	•		•								
By Capability	•		•					•		•		•							•			•		•
Holistic: n < 10							•		•									•		•	•			
Holistic: n >= 10															•		•						•	

Wheelchairs Limited to Metric Commands

Nearly two thirds of recent NL-enabled wheelchairs can only follow verbal commands to go forward, turn or stop. We first examine the wheelchairs in this category that have no hardware additions other than the microphone necessary for speech input.

1. Wheelchairs Limited to Metric Commands without Added Hardware

Table 8.3: Wheelchairs allowing only metric level commands with no sensors other than a microphone

Project	1	2	3
Hardware Configuration			
Manual Base	•	•	
Powered Base			•
Linguistic Capabilities			
Robust to Disfluencies			•
User Evaluation			
Style:None		•	
By Capability	•		•
Size: N/A		•	
< 10 Participants	•		
>= 10 Participants			•

Since the published aspects of the wheelchairs in this category ((Qidwai & Ibrahim, 2010; Qadri & Ahmed, 2009; Suk, Chung, & Kojima, 2007); 1-3 in the table above) were solely related to aspects of speech recognition, it is understandable that the set of commands executable by these wheelchairs is limited in scope. The wheelchair presented by Suk, Chung, & Kojima (2007), for example, was presented with respect to a voice-control algorithm designed to be robust to speech disfluencies.

Two of these projects used augmented manual wheelchairs instead of powered wheelchairs, due to their limited needs. Experimental validation differed between projects; one analyzed about 2000 samples collected from 12 participants (Suk, Chung, & Kojima, 2007), one analyzed 250 commands collected from five participants (Qidwai & Ibrahim, 2010), and one did not indicate whether their wheelchair had been empirically evaluated (Qadri & Ahmed, 2009).

2. Wheelchairs Limited to Metric Commands with Hardware Additions but without Obstacle Avoidance

Table 8.4: Wheelchairs allowing only metric level commands with sensors that do not provide obstacle avoidance

Project	4	5	6	7	8	9	10
Hardware Configuration							
Manual Base		•	•	•			•
Powered Base	•				•	•	
RF Reader				•			
Eye Tracking	•	•					
Head Tracking			•				
BCI	•						
Magnetic Control						•	
Keyboard and Mouse					•		•
Touch Screen		•	•				•
Keypad				•		•	
Joystick	•			•	•	•	•
Non-Linguistic Behaviors							
Speed Adjustment	•						
Linguistic Behaviors							
Answers Other Questions			•				
User Evaluation							
Style: None	•	•	•				
By Capability					•		•
Holistic				•		•	
Size: N/A	•	•	•				•
<10 Participants				•	•	•	

Of the wheelchairs with hardware additions but without obstacle avoidance ((McMurrough, Ranatunga, Papangelis, Popa, & Makedon, 2013; Maskeliunas & Simutis, 2011; Berjon, Mateos, Barriuso, Muriel, & Villarrubia, 2011; Asakawa & Nishihara, 2007; H. Wang, Li, & Zheng, 2015; Ruíz-Serrano et al., 2013; Linh, Hai, Van Thuyen, Mai, & Van Toi, 2015); 4-10 in the table above) only those presented by McMurrough, Ranatunga, Papangelis, Popa, & Makedon (2013) and Linh, Hai, Van Thuyen, Mai, & Van Toi (2015) used powered wheelchair bases. All seven wheelchairs in this category used a manual control modality such as a standard joystick (McMurrough, Ranatunga, Papangelis, Popa, & Makedon, 2013; Asakawa & Nishihara, 2007; H. Wang, Li, & Zheng, 2015; Ruíz-Serrano et al., 2013; Linh, Hai, Van Thuyen, Mai, & Van Toi, 2015), or a touch screen (Maskeliunas & Simutis, 2011; Berjon, Mateos, Barriuso, Muriel, & Villarrubia, 2011). Most of these projects focused on the use of multiple control modalities. In addition to voice and touch control, McMurrough, Ranatunga, Papangelis, Popa, & Makedon (2013) used BCI and eye-tracking control; Maskeliunas & Simutis (2011) used eye-tracking control; Berjon, Mateos, Barriuso, Muriel, & Villarrubia (2011) used head-tracking control; Ruíz-Serrano et al. (2013) used tongue-based magnetic control; and both H. Wang, Li, & Zheng (2015) and Linh, Hai, Van Thuyen, Mai, & Van Toi (2015) used a keyboard and mouse.

Asakawa & Nishihara (2007), on the other hand, used no additional control modalities, but used a Radio Frequency (RF) tag reader along with RF tags embedded into the floor to allow their wheelchair to autonomously round corners. Other capabilities of these wheelchairs were limited. The wheelchair presented by McMurrough, Ranatunga, Papangelis, Popa, & Makedon (2013) could accept voice commands to adjust its speed of movement (but could not accept commands to turn); the wheelchair presented by Berjon, Mateos, Barriuso, Muriel, & Villarrubia (2011) could use a smartphone to answer questions about the weather and news.

Evaluations of the wheelchairs in this category were limited. McMurrough, Ranatunga, Papangelis, Popa, & Makedon (2013) do not appear to evaluate their wheelchair at all, and Berjon, Mateos, Barriuso, Muriel, & Villarrubia (2011) and Maskeliunas & Simutis (2011) only state that their wheelchairs work fine. H. Wang, Li, & Zheng (2015) and Linh, Hai, Van Thuyen, Mai, & Van Toi (2015) evaluate the accuracy of their speech recognition systems, with H. Wang, Li, & Zheng (2015) stating that five participants were used, and Linh, Hai, Van Thuyen, Mai, & Van Toi (2015) not providing any information about who provided their training and testing data. Asakawa & Nishihara (2007) contrast the time taken for three subjects to navigate a hallway using voice control augmented with this autonomous behavior with time taken when using voice, button pad or joystick control.
Ruíz-Serrano et al. (2013) had five participants navigate an environment with obstacles, measuring the time taken to complete the task.

3. Wheelchairs Limited to Metric Commands with Hardware Additions Allowing for Obstacle Avoidance

Table 8.5: Wheelchairs allowing only metric level commands with sensors that provide obstacle avoidance

Project	11	12	13	14	15
Hardware Configuration					
Manual Base	•	•		•	
Powered Base			•		•
Ultrasound	•			•	
Camera		•	•		•
IR				•	
Finger Motion Sensor	•				
Touch Screen					•
Remote Control				•	
Joystick	•	•	•	•	
User Evaluation					
Style: None	•		•	•	
By Capability		•			
Holistic					•
Size: N/A	•		•	•	
<10 Participants		•			
>=10 Participants					•

We will now discuss the five remaining wheelchairs restricted to metric commands ((Wallam & Asif, 2011; Babri, Malik, Ibrahim, & Ahmed, 2012; J. Liu, Zhang, Fan, Wang, & Wu, 2010; Sheikh & Rotake, 2015; Škraba, Stojanović, Zupan, Koložvari, & Kofjač, 2015); 11-15 in the table above). Three of these wheelchairs used manual bases (Wallam & Asif, 2011; Babri, Malik, Ibrahim, & Ahmed, 2012; Sheikh & Rotake, 2015) and two used powered bases (J. Liu, Zhang, Fan, Wang, & Wu, 2010; Škraba, Stojanović, Zupan, Koložvari, & Kofjač, 2015). Unlike the wheelchairs examined thus far, all wheelchairs in this category used sensors to avoid obstacles: Wallam & Asif (2011) and Sheikh & Rotake (2015) used ultrasound sensors (Sheikh & Rotake (2015) also used an IR sensor), and Babri, Malik, Ibrahim, & Ahmed (2012), J. Liu, Zhang, Fan, Wang, & Wu (2010), and Škraba, Stojanović, Zupan, Koložvari, & Kofjač (2015) used a camera. All wheelchairs could be controlled with a joystick except that presented by Škraba, Stojanović, Zupan, Kodožvari, & Kofjač (2015), who replaced theirs with a touch screen. In addition, Wallam & Asif (2011) used a finger motion sensing glove, while Sheikh & Rotake (2015) used a remote controller.

The wheelchair presented by Škraba, Stojanović, Zupan, Kodožvari, & Kofjač (2015) was evaluated by twelve participants, including two patients from a rehabilitation institute. Experimental validation of the other wheelchairs was minimal; Babri, Malik, Ibrahim, & Ahmed (2012) state that two people tested their wheelchair's speech recognition; the rest are only described as working fine, if their performance is described at all.

Thus far, we have examined 15 wheelchairs, most of which could only understand five commands: Go forward, Go backwards, Turn left, Turn right, and Stop. This is clearly well-trod ground, and yet many of these projects do not significantly predate the projects found in latter categories: some were published on as recently as 2015. Future wheelchair developers should focus not on these basic capabilities, but rather on enabling more sophisticated linguistic and mnemonic capabilities, as do the developers of the projects we will now discuss.

Wheelchairs Capable of Following Local Entities

The next group of wheelchairs are those that can navigate relative to local environmental features such as walls and elevators. All projects in this category ((Hockey & Miller, 2007; Pineau, Atrash, Kaplow, & Villemure, 2010; Murai, Mizuguchi, Saitoh, Osaki, & Konishi, 2009); 16-18 in the table above) use powered wheelchair bases controllable by joystick. In addition, the wheelchair presented by Pineau, Atrash, Kaplow, & Villemure (2010) can be controlled by a touch screen. All wheelchairs in this category have at least one sensor used to avoid obstacles: Hockey & Miller (2007) use an ultrasound sensor, Pineau, Atrash, Kaplow, & Villemure (2010) use a Laser Range Finder (LRF), and Murai, Mizuguchi, Saitoh, Osaki, & Konishi (2009) use both ultrasound and IR sensors. Unlike the previous wheelchairs, those in this category all have an array of capabilities and behaviors.

The wheelchair presented by Hockey & Miller (2007) appears to have been used as a proof-of-concept demonstration within a limited domain. As such, it has not been empirically evaluated, and there are scarce details about how it works algorithmically. Hockey & Miller (2007) do, however, provide a sample dialogue handled by their wheelchair, suggesting some in-

Table 8.6: Wheelchairs able to issue NL commands not requiring perception

Project	16	17	18
Hardware Configuration			
Powered Base	•	•	•
LRF		•	
Ultrasound	•		•
IR			•
Touch Screen		•	
Joystick	•	•	•
Non-Linguistic Capabilities and Behaviors			
Intention Modeling		•	
Metric Mapping		•	
Speed Adjustment			•
Wall Following		•	
Elevator Entering			•
Travels to Objects	•		
Travels to Unknown Objects or Locations	•		
Linguistic Capabilities and Behaviors			
Dialogue Management	•	•	
Acknowledgment	•	•	
Ask Questions	•		
User Evaluation			
Style: None	•		
Holistic		•	•
N/a	•		
<10 Participants			•
>=10 Participants		•	

teresting capabilities, such as dialogue management, which it uses to provide acknowledgment and ask questions, and the ability to travel to described objects even if it has never seen them before – a capability of interest in current research such as that presented by Duvallet et al. (2014), or by ourselves in Chapter 3.2.

The SmartWheeler wheelchair(Pineau, Atrash, Kaplow, & Villemure, 2010) uses an LRF to detect and avoid obstacles and to map its environment, which allows the wheelchair to easily follow walls. When the SmartWheeler receives a command, it can ask for feedback regarding its interpretation using a touchscreen. Recent work on this project has included modeling of the wheelchair user's intentions when issuing commands(Png & Pineau, 2011). For evaluation, the wheelchair was first run through the Wheelchair Skills Test(Routhier, Vincent, Desrosiers, Nadeau, & Guerette, 2004). Then, 23 subjects, both able-bodied and disabled, evaluated the wheelchair.

Finally, Murai, Mizuguchi, Saitoh, Osaki, & Konishi (2009) present a wheelchair that uses ultrasound and infrared sensors to detect and avoid obstacles and to get in and out of elevators. Their wheelchair does not use a dialogue manager, but prompts the user after every command to ensure it understood them correctly. This wheelchair was validated using five able bodied participants in a series of experiments.

In this category, we see for the first time wheelchairs with substantial linguistic capabilities. But while some wheelchairs in this category (i.e., those of Pineau, Atrash, Kaplow, & Villemure (2010) and Murai, Mizuguchi, Saitoh, Osaki, & Konishi (2009)) have begun to allow more sophisticated navigational behaviors such as wall following and elevator entering, more sophisticated linguistic capabilities are still lacking; only Hockey & Miller (2007)'s wheelchair may have come close to the goal of *genuine helper*, but it was not truly evaluated.

Wheelchairs Capable of Navigating to Specified Locations

The final group is comprised of wheelchairs that use a topological map to navigate. We split these into those that are given maps, and those that create their own.

1. Wheelchairs Capable of Navigating to Specified Locations, that Require a Prebuilt Map

   Four projects used wheelchairs pre-loaded with topological maps ((Megalingam, Nair, & Prakhya, 2011; Tao, Wang, Wei, & Chen, 2009;

Table 8.7: Wheelchairs able to issue NL commands requiring mapping based on pre-loaded topological maps

Project	19	20	21	22
Hardware Configuration				
Manual base	•	•		
Powered base			•	•
Sonar	•			
RF Reader	•	•		
Ultrasound		•	•	
IR			•	
Camera			•	
LRF				•
Joystick		•	•	•
Touch Screen		•	•	
Gamepad			•	
Keyboard and Mouse			•	
Head Tracking			•	
Non-Linguistic Capabilities and Behaviors				
Intention Modeling				•
Working Memory				•
Metric Mapping				•
Wall Following			•	
Travels to Named Places	•	•	•	•
Travels to Objects				•
Travels to Unknown Objects				•
Linguistic Capabilities and Behaviors				
Dialogue Management				•
Robust to Ambiguity				•
Accepts Descriptions				•
Acknowledgments		•		•
Answers Questions		•		•
Asks Questions		•		•
User Evaluation				
Style: By Capability	•			•
Holistic		•	•	
Size: N/A	•			•
<10 Participants		•	•	

Faria, Reis, & Lau, 2015); 19-21 in the table above, and our own wheelchair (Section 8.2,

22 in the table above). Hardware varied greatly between these projects. Megalingam, Nair, & Prakhya (2011) used a camping chair attached to a platform with sonar sensors and an RF reader. Tao, Wang, Wei, & Chen (2009) use a manual wheelchair base outfitted with ultrasound sensors, an RF reader, a touch screen and a joystick. The IntellWheels project (Braga, Petry, Reis, & Moreira, 2011; Petry, Moreira, Faria, & Reis, 2013; Faria, Reis, & Lau, 2015) uses a powered wheelchair with both sonar and infrared sensors. They focus in part on mapping multimodal input sequences to desired actions; their wheelchair can be controlled by touchscreen, joystick, gamepad, keyboard, or head movement. Our own wheelchair uses a powered wheelchair base, can be manipulated with a joystick, and is equipped with two LRFs. We will next discuss the three previously published wheelchairs in this category, and then briefly discuss our own research efforts, which are expounded upon in Section 8.2.

Despite the wide variance in hardware, the three previously published wheelchairs can perform roughly the same behaviors; all three can go to a pre-labeled room, and Faria, Reis, & Lau (2015)'s wheelchair can follow walls. All three avoid obstacles, and Tao, Wang, Wei, & Chen (2009)'s wheelchair can provide acknowledgments and answer questions about the weather and upcoming events. Megalingam, Nair, & Prakhya (2011) tested wheelchair response time, while the other two groups performed a battery of tests on a small number of subjects (five (Tao, Wang, Wei, & Chen, 2009) or eight (Petry, Moreira, Faria, & Reis, 2013)), yielding qualitative results about the wheelchair's performance. These wheelchairs used pre-loaded maps either because they expected that the wheelchair would be used in a pre-known home environment (Megalingam, Nair, & Prakhya, 2011; Tao, Wang, Wei, & Chen, 2009), or because they used a mixed reality system requiring a simulated environment representation(Petry, Moreira, Faria, & Reis, 2013).

Designed as a successor to the Vulcan wheelchair(Murarka, Gulati, Beeson, & Kuipers, 2009), our own wheelchair, developed in conjunction with researchers from the University of Michigan, can avoid obstacles and dynamically create a *metric* map of its environment using the Hybrid Spatial Semantic Hierarchy (HSSH)(Kuipers, 2008). We are currently integrating prior work on topological maps built through both observation and dialogue interactions (cf. Chapter 3.2). Unlike the other wheelchairs examined thus far, our wheelchair can posit new hypothetical locations based on dialogue (even though it can only travel to them in some contexts) and can travel in search of previously unknown objects – capabilities afforded by the referential processing framework introduced in Chapters 3-4. This approach also differs from previous approaches in that it does not only accept commands to travel to locations denoted by a *rigid designator* (i.e., by "name"), but rather accepts commands to travel to places matching *descriptions* such as "the room at the end of the hall down on the right", a strictly larger class of referring expression.

Another key feature of our wheelchair is its use of the robust dialogue system whose pragmatics module is described in Chapters 6-7, which can infer goals and intentions from indirect speech acts under contextual uncertainty or ignorance. A video of our wheelchair acting on indirect language can be viewed at `https://www.youtube.com/watch?v=eSU1YWdSfpk`.

While we have evaluated many such capabilities in previous work, and have informally demonstrated many of them on our wheelchair, we have not yet holistically evaluated our wheelchair – this will be a topic for future work as our research efforts progress.

2. Wheelchairs Capable of Navigating to Specified Locations and of Dynamically Building Topological Maps

   Finally, we discuss the two wheelchairs capable of building topological maps dynamically (i.e. 23 and 24 in the table above(S. Hemachandra, Kollar, Roy, & Teller, 2011; Röfer, Mandel, Lankenau, Gersdorf, & Frese, 2009)). Both use powered wheelchair bases with laser scanners for detecting and avoiding obstacles, and allow for joystick control.

   The first is the MIT Intelligent Wheelchair Project (N. Roy et al., 2011). This wheelchair builds a metric map from which topological structures can be extracted. Not only can this system travel to named objects and places, but it can receive new labels on-the-fly during guided tours while following a guide (S. M. Hemachandra, 2010). More recent papers also describe the wheelchair's ability to accept descriptions such as "The kitchen is down the hall"(Walter, Hemachandra, Homberg, Tellex, & Teller, 2013) and references to previously unknown entities, such as "the cone behind the hydrant"(Duvallet et al., 2014). This system can also provide acknowledgments, travel outdoors to some

Table 8.8: Wheelchairs able to issue NL commands requiring dynamic mapping

Project	23	24
Hardware Configuration		
Powered Base	•	•
Sensors: LIDAR	•	
LRF		•
Camera		•
Joystick	•	•
Head Joystick		•
BCI		•
Non-Linguistic Capabilities and Behaviors		
Metric Mapping	•	•
Topologic Mapping	•	•
Multi-floor Mapping	•	
Follows Route Descriptions		•
Wall Following		•
Person Following	•	
Elevator Entering	•	
Travels to Named Places	•	•
Accepts New Place Names	•	
Travels to Objects	•	
Travels to Unknown Objects	•	
Linguistic Capabilities and Behaviors		
Robust to Ambiguity	•	
Accepts Descriptions	•	
Acknowledgment	•	•
User Evaluation		
Style: By Capability		•
Holistic	•	
Size: >=10 Participants	•	•

extent (Walter, Hemachandra, Homberg, Tellex, & Teller, 2013), enter elevators to traverse multiple floors, and is robust to referential ambiguity through the use of the G^3 framework.

There have been many publications on this project: in addition to evaluation of individual capabilities, it has also undergone holistic evaluation with a larger number of subjects than most other systems examined (e.g., fifteen participants were used to test the social acceptability of the wheelchair's following behavior).

The second wheelchair in this category is that presented by Röfer, Mandel, Lankenau, Gersdorf, & Frese (2009). This wheelchair can be controlled with a head joystick or with a brain-computer interface, and can follow route descriptions, such as "Go down the corridor and take the second door to the left" (Röfer, Mandel, Lankenau, Gersdorf, & Frese, 2009). Previous work on this project explored map creation, but so far as we can tell route descriptions are used solely in conjunction with pre-built maps. As the target environment for this wheelchair is an assisted living center, its layout would presumably already be known. Work on this project has also attempted to deal with some ambiguous situations, such as determining what is meant by "right" when it could mean "correct", "veer right", "turn right here" or some other meaning (Ross, Shi, Vierhuff, Krieg-Brückner, & Bateman, 2005). There has been extensive of research on this project in the past two decades, involving many studies with detailed quantitative analysis (e.g. Tenbrink, Ross, Thomas, Dethlefs, & Andonova, 2010). Much of this work has been focused on evaluating individual parts of the system and on Wizard-of-Oz studies, however, and to the best of our knowledge there has been no holistic evaluation of their wheelchair.

In this and the previous section, we have finally seen significant developments in the mnemonic and linguistic capabilities and behaviors necessary for an intelligent wheelchair to become an genuine helper. Of these systems, the MIT Intelligent Wheelchair Project stands out as the state-of-the-art, as it is capable of a wide range of non-linguistic and linguistic behaviors, and has been holistically evaluated by a (comparatively) large number of participants, but this project has not focused on developing the mnemonic capabilities necessary for an intelligent wheelchair to be genuinely helpful. In our own work, in contrast, we have broken new ground in developing such mnemonic and linguistic capabilities – but our approach is as yet a work in progress, and is in need of both autonomous topological mapping and a holistic evaluation.

8.1.3 Discussion

Many of the examined wheelchairs, especially those that dynamically map their environments (S. Hemachandra, Kollar, Roy, & Teller, 2011; Röfer, Mandel, Lankenau, Gersdorf, & Frese, 2009), show promising progress towards the development of an intelligent wheelchair that genuinely helps users in their daily lives. Yet, the creation of such an genuine helper requires solutions to many challenging problems, as evidenced by the fact that most of the examined wheelchairs either focus on a particular subproblem (e.g., accurate speech recognition) or are unevaluated proofs-of-concept. Most importantly, there are many desirable properties of an ideal wheelchair that have not even been addressed yet, such as independence of environmental structure, modeling of interlocutors' beliefs, episodic memory, and the ability to engage in truly natural dialogues. We will discuss some of these capabilities in more detail and sketch necessary steps to achieving them.

Environmental Constraints

Although a number of wheelchairs allow their users to give commands pertaining to shared environmental features such as walls, rooms, or objects, the majority of these wheelchairs are constrained to pre-known environments. What is more, these wheelchairs are almost entirely constrained to indoor environments, either due to assumptions about the structure of the environment, or sensors that cannot accurately function outdoors.

In fact, only two wheelchairs(Berjon, Mateos, Barriuso, Muriel, & Villarrubia, 2011; Walter, Hemachandra, Homberg, Tellex, & Teller, 2013) seem to have even been used outdoors, and to the best of our knowledge none of the examined wheelchairs can cope with fully outdoor navigation. This is due in part to a lack of appropriate sensors: only five of the examined wheelchairs were equipped with cameras, and other types of sensors may be ill-suited for outdoor navigation. Although this problem has not been addressed by NL-enabled wheelchairs, other intelligent wheelchairs *do* navigate outdoors (e.g. Yanco, 2001; Tabuse, Kitaoka, & Nakai, 2011). Future NL-enabled wheelchairs should robustly cope with unknown environments, both indoor and outdoor.

To travel outdoors, wheelchairs must also be equipped with cameras. Not only are cameras useful for recognizing objects, landmarks, and signs, but stereo cameras can rapidly generate 3-D point clouds which can be used for outdoor navigation in a way that is resilient against the illumination changes which plague outdoor navigation(Irie, Yoshida, & Tomono, 2012). NL-enabled wheelchairs should also use GPS for navigation (as other wheelchairs have, (e.g. Bejuri, Saidin, Mohamad, Sapri, & Lim, 2013)): it is a useful navigation technique, and could allow continued localization while a wheelchair user is transported by vehicle. We would argue that in general, wheelchairs must break from the assumption of straight hallways and room-and-hall networks within a single floor, and must move towards handling not only multi-floor buildings, multi-building complexes and networks of outdoor paths, but anomalous environments with oddly shaped rooms, rooms which flow into each other, and doors which are wider than average or made of glass.

Finally, NL-enabled wheelchairs should accept commands to go to objects and locations they have not already visited; a feature exhibited only by Duvallet et al. (2014); Hockey & Miller (2007), and ourselves (Section 8.2), and must use belief modeling and episodic memory for the better resolution of ambiguous references.

Linguistic and Mnemonic Capabilities and Behaviors

An advantage of using NL to interact with wheelchairs (and robots in general) is that NL can be used for *communication*, which in turn can be used for teaching and for explanation. Unfortunately, most of the examined wheelchairs fail to take advantage of this in any way, using voice input as just another way to obtain joystick functionality; fewer than half of all wheelchairs allow for additional linguistic input. Of those that do, only the wheelchairs presented by S. Hemachandra, Kollar, Roy, & Teller (2011) and ourselves allow a user to *inform* the wheelchair about features of the environment, such as the names or locations of rooms, and only a few attain any degree of conversation through dialogue management. It would be useful in many cases for an intelligent wheelchair to be capable of sustained dialogue interactions: the wheelchair may need to engage with dialogue with its rider as to what route to take. We envision future wheelchairs handling interaction patterns such as the following dialogue, which requires dialogue tracking, intent recognition, indirect speech act handling, question asking and answering, and other linguistic capabilities and behaviors, as well as episodic, working, belief, and intention modeling, among other mnemonic capabilities and behaviors:

User Alright, let's go to my barbershop.

Wheelchair *Drives off, takes a left*

User I said my *barbershop*, wheelchair.

Wheelchair *Stops.* We can go this way to get to your barbershop. If we took the other path we would have to cross Boston Avenue. You told me you'd rather not have to do that anymore.

User I remember. How do we get there this way?

Wheelchair We can just drive down Medford Street, and then take Somerville Avenue. We should be there in fifteen minutes. Is that alright?

User Yes, that's alright wheelchair, let's go.

Wheelchair Alright. *Drives off*

A wheelchair with sufficient linguistic capabilities could assume the role of helping the wheelchair's user fulfill his or her needs, in the same way a human companion would if they were pushing the wheelchair. To be perceived in this manner, it is important that the wheelchair achieve all of the capabilities we laid out when describing a genuinely helpful wheelchair, including those reflected in the dialogue above. Although the prototype wheelchair we presented is under development, it represents a step forward in addressing our concerns about the perception of NL-enabled wheelchairs, as it can interact in a more conversational manner than its predecessors, and takes a more *cognitive* approach than the other wheelchairs developed to date.

As mentioned before, our desiderata lie along the path to a wheelchair that is a *genuine helper*, and not one that has all the capabilities of a human helper. As we previously mentioned, the full set of human capabilities are well beyond the scope of current intelligent wheelchairs, and lie far beyond the current research horizon. We have also chosen to focus on *task oriented* capabilities and behaviors. While robot wheelchairs may be endowed with non-task-oriented capabilities such as the ability to make small talk, to empathize, or to manifest its own personality and desires, such capabilities from the domain of *social robotics* are not necessary for the robot to be a *genuine helper*. And in fact there are ethical concerns associated with developing an intelligent wheelchair that is a *companion* with whom users should form social or emotional bonds (Scheutz, 2011). The benefits and consequences of such a decision, however, are beyond the scope of this dissertation.

Experimental Validation

It is hard to think of a robot that is more user-centered than a wheelchair. The purpose of a wheelchair is to provide continuous, long-term mobility assistance to its user; a highly user-centered requirement.

The addition of a natural language interface makes the connection between the wheelchair and its user even more personal, allowing for the transformation of the wheelchair from a vehicle into a companion. And yet, the majority of wheelchair evaluations observed in this survey were anything but user-centered. The ideal evaluation for an autonomous robotic wheelchair controllable by NL would be *task-oriented*, *long-term*, *large-scale*, and *using the wheelchair's target population*. In this section we describe why each of these aspects is both important and insufficiently addressed in current wheelchair evaluations.

1. Task-Oriented

   A wheelchair is an important part of its user's day-to-day life. Testing whether the wheelchair can navigate around corners or respond quickly to commands is not enough; experiments should require subjects to accomplish tasks that actual wheelchair users might encounter: navigating to particular locations, retrieving objects, going through doorways, pulling up to tables, and so forth. It will be important to evaluate how easy these tasks are to achieve, how long it takes to achieve them, and the level of trust the wheelchair maintains. Does the wheelchair move in ways that make its user uncomfortable or nervous? Does the user trust the wheelchair to carry out high-level commands, or does the user fall back on metric commands?

   Few of the examined wheelchairs had evaluations of this sort. Only one third of the wheelchairs were holistically evaluated: the rest were either presented as proofs of concept, or only evaluated specific features such as speech recognition rates. Of the eight wheelchairs with holistic evaluations, the three capable of only metric commands were evaluated with respect to time taken to complete various navigation tasks; of the two capable of local feature following, one was evaluated based on user satisfaction, the other based on ability to complete the navigation tasks of the Wheelchair Skills Task; of the three capable of place navigation, one was evaluated based on preference, comfort, efficiency, and mapping accuracy, after receiving a guided tour, while the other two were evaluated on navigation tasks. Although some of the evaluations described above were indeed task-based in nature, we would argue that the tasks used for evaluating future wheelchairs should be more “everyday” in nature.

2. Long-Term

   Robotic wheelchairs may be continuously used every day for several years. But the evaluations of the examined wheelchairs tended to be short, likely due to unwillingness to invest the time and money, or due to a lack of robustness of the wheelchairs themselves. It would be useful to see how a wheelchair user feels about their wheelchair after an entire day of using it: the user could better adapt to the wheelchair, allowing them to provide better feedback as to difficulties of use, and provide insight into what types of commands actually get used after the first hour or so of operation. A user may become more frustrated with their wheelchair after a longer period of time, and may become more or less likely to use high level navigational commands. It will also be important to see how the wheelchair handles navigation in larger environments, many additional interlocutors engaging in conversation with its user, and other issues that may not come up in a half hour evaluation of navigation through one or two hallways. Long term evaluations will also reveal everyday tasks the wheelchair has trouble with that its designers may not have considered, such as pulling up to a table, going to the restroom, or driving through an automatic doorway.

   Few recent wheelchairs have undergone long-term testing, likely due to the temporal and monetary costs of such testing, or due to a lack of robust performance. It is clear, however, that long-term evaluations should be a goal for every robotic wheelchair designer, as a wheelchair unable to be used in a long-term scenario is of limited value.

3. Large-Scale

   A long term evaluation may only be possible with a small number of subjects, but short term evaluations should be performed with a large number of subjects, or at least more subjects than are currently being used. Few of the examined wheelchairs used even ten subjects. And it is at best questionable how useful an evaluation by only two or three people is, especially when those two or three people designed the wheelchair themselves, and are familiar with its quirks and idiosyncrasies.

4. Using Target Population

   Few projects were validated using members of the wheelchair's target population. The nature and focus of an individual project may excuse this, but future projects should make an effort to demonstrate successful use of their wheelchair by those who would use it on a daily basis, as such users will have their own unique needs and concerns which must be addressed for the wheelchair to be usable by them.

Future Work

From our survey of recent NL-enabled wheelchair projects it is obvious that there is a long road ahead for NL-enabled wheelchairs; many of the capabilities and behaviors necessary for a wheelchair to be genuinely helpful are missing from even the most state-of-the-art NL-enabled wheelchairs. And many other features are handled by only one or two wheelchairs. The first step towards a genuinely helpful wheelchair will be developing a wheelchair that achieves *all* capabilities previously achieved by previous wheelchairs, including multi-floor mapping, speed changing, entity following, route description following, memory modeling, dialogue management, and traveling to unknown places. Researchers might then take a number of future directions to improve the functionality and interaction capabilities of NL-enabled wheelchairs:

- **Belief Modeling:** Some wheelchairs already have means for representing the topological structure of their own spatial knowledge; these structures should be adapted to represent the likely spatial knowledge of other agents, including but not limited to their users. This would be useful if a wheelchair is used by multiple people who may be familiar with different spatial regions, or if the robot needs to interpret directions given to the wheelchair's current user by a third party.

- **Intention Modeling:** Research on modeling the intentions and goals of agents should be applied to intelligent wheelchairs for them to make better decisions when following instructions which require them to reason about other agents, including but not limited to their users. This would allow a wheelchair to more accurately predict the intended destinations of its user, and would allow a wheelchair to follow commands such as "Let's go find Lisa", where Lisa's location may depend on her own daily routine.

- **Episodic Memory:** Such intention modeling would be greatly facilitated by the integration of episodic memory models. If a robot can recall what it saw where, what locations it visited when, and so forth, it can better model its user's intentions when driving down a familiar hallway, or when processing an utterance like "Let's go to *my usual* barbershop."

- **Action and Intent Recognition:** There has been much recent research on recognizing actions(Poppe, 2010), but researchers must develop action and intention recognition systems that will work from the perspective of, and on data generated by, intelligent wheelchairs. This is necessary to store information in aforementioned episodic memory structures, in order, in turn, to facilitate the aforementioned intention modeling processes.

- **Disfluency Handling:** Few NL-enabled wheelchairs attempt to handle disfluencies resulting from speech impairments. One of the primary motivations for developing NL-enabled wheelchairs is to aid the 40% of wheelchair users who cannot easily manipulate a joystick; but many wheelchair users also suffer from speech impairments, a fact only addressed by Suk et al.(Suk, Chung, & Kojima, 2007). Researchers should attempt to address disfluencies to be accessible to a greater number of people.

- **Outdoor Navigation:** Researchers must develop mapping systems flexible enough to allow for autonomous navigation in outdoor environments, in order for wheelchairs to be used outside of indoor environments such as private homes.

- **Gesture Recognition:** There has been much recent work on gesture recognition(Zhang, Zhang, & Luo, 2011); but researchers must develop gesture recognition systems that allow for interpretation of simultaneous speech and gesture issued from the perspective of wheelchair users. This will be necessary so as to accurately interpret utterances such as "Drive closer to that *(*points*)* table" or "Can you go over that way? *(*points*)*".

- **Suggestion Generation:** Researchers must develop systems that leverage episodic memory and intention modeling for robots to autonomously generate timely suggestions for their users. A wheelchair may need to make suggestions like "Isn't it time for your appointment?", "Didn't you want to go see Lisa?" or "It's time for your medication" –

utterances which are not typically used in response to utterances made by the user, but are instead spontaneously generated.

8.1.4 Survey Conclusions

We have presented a framework for evaluating the abilities of both existing and future NL-enabled wheelchairs. We have identified several areas in which NL-enabled wheelchairs can be advanced, focusing on navigability of outdoor environments, thoroughness of experimental validation, and treatment of the wheelchair as an intelligent agent through capabilities such as dialogue, belief modeling and episodic memory. And while great strides have been made in recent years, we believe that progress may be best accelerated through two choices. First, research is needed on understanding and carrying out natural language instructions that go *beyond* simple directional commands. Second, research is needed on higher-level mnemonic and cognitive functions such as belief, intention, dialogue and memory modeling, as these will not only facilitate more advanced executable behaviors for intelligent wheelchairs, but also bring wheelchairs closer to being, and being perceived as, genuine helpers for their users.

8.2 Our NL-Enabled Robotic Wheelchair

As shown in the previous section, not all natural language enabled wheelchairs are created equal. This is equally true for the robot architectures *themselves*, which are used to control these wheelchairs. A large number of integrated robot architectures have been developed over the past few decades, but these differ wildly in terms of the representations they use and the capabilities and behaviors they enable, which are dependent on the research objectives of their designers. This is particularly true of the *Vulcan* robot architecture and middleware(Murarka, Gulati, Beeson, & Kuipers, 2009) and the Distributed, Integrated, Affect, Reflection, Cognitive Robot Architecture (*DIARC*)(Scheutz et al., 2013) as implemented in the Agent Development Environment (ADE) MAS middleware (the architecture used throughout this dissertation and described in Chapter 2).

Both *Vulcan* and *DIARC* are considered fully fledged robot architectures implemented as fully fledged multi-agent systems (MAS). These architectures, however, have relatively few overlapping representations, capabilities, and behaviors: *Vulcan* uses the rich spatial representations provided by the *Hybrid Spatial Semantic Hierarchy* (HSSH) to enable navigation capabilities in real-world, campus-like environments, while *DIARC* uses high-level cognitive representations to enable human-like tasking through natural language. Furthermore, the two architectures' respective middlewares significantly differ at the implementation level: *Vulcan* is implemented in C++, offering the speed necessary for real-time mapping and hardware-level integration, whereas *ADE* is implemented primarily in Java and other JVM languages, allowing for better portability, as well as increased flexibility with respect to choice of programming paradigm. What is more, beyond components to handle sensory data and deliver motor commands to robot bases, the architectures do not have many common components.

In this case, however, difference begets opportunity. By integrating the *Vulcan* and *DIARC* robot architectures (through specific integration of the *Vulcan* and *ADE* MAS middlewares), we have produced a new robot architecture that is greater than the sum of its parts, with state-of-the-art navigational capabilities thanks to *Vulcan*, state-of-the-art linguistic capabilities thanks to *DIARC*, and new synergistic capabilities made possible only through this integration (e.g., navigation to locations based on complex natural language utterances with context-dependent meanings) as each architecture leverages the other's strengths. What is more, this new hybrid integrated robot architecture can be viewed as being implemented in *Vulcan-ADE* Development Environment (*VADE*), a novel multi-(multi-agent system) system. VADE provides a useful, novel framework for the integration of multi-agent systems through the use of so-called *Dual Citizen* agents, as we will describe. What is more, as just one example application of this new integrated architecture, we have implemented it on a robotic wheelchair, resulting in a wheelchair that advances the state-of-the art.

Intelligent wheelchairs represent an attractive application not only because they benefit from what is brought to the table by both *Vulcan* and *DIARC*, but because they promise to be of great benefit to society. Within the United States alone, there are at least 3.6 million wheelchair users, 40% of whom find it difficult or impossible to control a wheelchair using a joystick(Fehr, Langbein, & Skaar, 2000; Brault, 2012). To make wheelchairs more accessible, many researchers are turning to Natural Language (NL) as a control modality. But as we showed in the previous section, while such NL-enabled wheelchairs have existed for nearly forty years (e.g. J. A. Clark & Roemer, 1977), most of the recently presented NL-enabled wheelchairs have only limited capabilities, e.g., the ability to be commanded to go forward, left, right, backwards and to stop, with few capable of more advanced capabilities such as traveling to remote objects or locations.

Although the new levels of autonomy and mobility that current NL-

Figure 8.4: The *Vulcan* Intelligent Wheelchair

enabled wheelchairs grant users is promising, these wheelchairs do not come close to providing the capabilities of human helpers. A human helper pushing a wheelchair can do more than travel to named locations. Human helpers learn about new locations and other entities through observation and dialogue. They have memories of events, preferences, and goals. They ask questions, make suggestions, and make conversation. Furthermore, human helpers are not troubled by environmental features like elevators, multi-floor buildings, or “the outdoors”. While NL-enabled wheelchairs will not truly rival the capabilities of human helpers anytime soon, we believe that NL-enabled wheelchairs are close to becoming *genuine helpers* that augment their users’ capabilities in order to make them effective in tasks of daily living, build rapport, and are worthy of trust. As we will show, our integration results in great progress towards this goal.

The remainder of the section proceeds as follows. First, we describe the individual capabilities of *Vulcan* and *DIARC*. Next, we describe how we have integrated these two architectures, how each leverages the capabilities of the other, what new synergies have emerged, and what challenges we encountered. Then, we present *VADE*: a novel multi-agent system framework comprised of both (1) software agents belonging to a single robot architecture and implemented in a single multi-agent system middleware, and (2) “Dual-

Citizen" agents that belong to both robot architectures and that use elements of both multi-agent system middlewares. We then provide a proof-of-concept demonstration showing novel capabilities effected through our integrated approach that advance the state of the art of NL-enabled wheelchairs. Finally, we discuss our plans to more deeply integrate *DIARC* and *Vulcan* in order to allow each architecture to further leverage the capabilities of the other.

8.2.1 *DIARC* and *ADE*

In this section, we will first discuss the *DIARC* Cognitive Robot Architecture, and then discuss ADE, the multi-agent system middleware in which *DIARC* is implemented.

The *DIARC* Cognitive Robotic Architecture

Throughout this dissertation, we have made use of the wide variety of high-level cognitive capabilities provided by *DIARC*. Of particular relevance to its use within the *VADE* hybrid architecture are its language- memory- and action-oriented components. *DIARC's* language-oriented components (as presented in Chapters 3-7) designed to allow[1] robots to resolve a wide variety of referring expressions, including anaphoric and deictic expressions (e.g., "*it* is in *that* breakroom") and referring expressions that use *descriptions* (e.g., "go to *the room across from the breakroom*") rather than *rigid designators* that indicate their targets by name or label. Furthermore, as previously discussed, such referring expressions need not be used in the context of direct commands: interlocutors are free to use *indirect speech acts* that follow conventionalized social norms (e.g.,"*I need to* go to the bathroom"), which *DIARC* interprets based on context. *DIARC's* language-oriented components leverage its memory-oriented components: some components make use of the knowledge-base-and-component framework presented in Chapter 3, which allows uncertain information about both known and hypothetical entities to be distributed across multiple heterogeneous knowledge bases; other components instead made use a general-purpose *belief* component to perform inference on shared knowledge. This component is leveraged by *DIARC's* action-oriented components, which perform high-level goal and action management capabilities.

[1]As described in that chapter, however, we have only deeply evaluated DIARC's ability to resolve anaphoric expressions; we have demonstrated how the architecture is *designed* to resolve deictic expressions, but have not fully evaluated the architecture's success in handling this type of expression.

While *DIARC* does have spatial reasoning and navigation components, such as those discussed in Section 3.2, these are relatively rudimentary relative to *DIARC's* cognitive components. *DIARC's* motion-oriented components can easily allow a robot to traverse a hallway or travel in a certain direction, but do not provide mapping capabilities, and use only rudimentary spatial representations.

The *ADE* Multi-Agent System Middleware

As described in Chapter 2, *DIARC* is implemented in the *Agent Development Environment* (*ADE*) multi-agent system middleware[2]. *ADE* is an architectural framework that builds on previous work from multi-agent systems in order to support the development of individual agent architectures using distributed multi-agent system computing infrastructure. *ADE* treats architectural components as autonomous software agents in order to facilitate dynamic system configuration, fault tolerance and recovery, distributed computation, and autonomic computing. *ADE* is primarily implemented in Java, with inter-agent communication facilitated by Java RMI. The use of a JVM language provides two main advantages: first, this allows for portability between different architecture platforms; second, it allows for developer flexibility, as code written in a variety of programming paradigms (i.e., through Java, Clojure, or Scala) can be seamlessly and richly integrated.

8.2.2 *Vulcan*

In this section, we will first discuss *Vulcan as a robot architecture*, and then discuss *Vulcan as a multi-agent system middleware* in which that architecture is implemented.

The *Vulcan* Robot Architecture

The *Vulcan* robot architecture focuses on the capabilities needed for navigation in campus-like environments, in which a robot may need to navigate between multiple buildings with diverse layouts, and through both empty spaces and dense crowds.

Vulcan uses rich spatial representations based on the Hybrid Spatial Semantic Hierarchy (HSSH), in the form of a hybrid metric-topological map. The HSSH uses both metric and topological representations of small-scale

[2] In fact, this paragraph presents information already discussed in Chapter 2. However, I've re-presented that material here to facilitate contrast with *Vulcan's* middleware

space (the space currently perceivable by a robot at any given point in time) which are combined to create large-scale maps of the robot's environment. The HSSH is thus comprised of four *layers*, each of which uses a different spatial representation: The Local Metric, Local Topological, Global Metric, and Global Topological layers.

- The **Local Metric** layer uses simultaneous localization and mapping (SLAM) techniques to maintain a Local Perceptual Map (LPM) of its immediate environment.
- The **Local Topological** layer uses the LPM to identify discrete regions of the world called *areas*.
- The **Global Topological** layer combines these areas into a topological map of large-scale space.
- The **Global Metric** layer uses this map along with information from the LPM to create a metric map of large-scale space.

These representations are useful for facilitating planning and navigation tasks. Specifically, navigation is facilitated by separating the task of graph-searching through large-scale space and metric-planning in small-scale space (cf. Chapter 3.2.1), and large-scale mapping is facilitated by the global topological map's sparse symbolic representation (Johnson & Kuipers, 2012).

Ultimately, a robotic wheelchair serves its human driver and therefore needs to reason about the human's goals and intentions. The use of a topological map is thus advantageous as it uses human-like representation of spatial knowledge, facilitating human-like spatial reasoning. But if the semantics of such a map are grounded solely in a robot's *actions* (e.g., if a map only represents the world's navigational affordances), as they are in *Vulcan*, then a robot using it can only be commanded through reference to these actions (e.g., by specifying a series of such affordances to exploit). *Vulcan* currently accepts these types of commands through a point-and-click user interface. In order to allow for more natural interactions, *Vulcan* needs a way of grounding its representations in the types of semantics more typically seen in human *conversations* (e.g., recognizing that a certain large-scale topological location may be a "kitchen", 'may be 'large", and may contain various goal-relevant objects), and should be able to accept commands that reference those aspects through a *natural language* interface, such as that used in *DIARC*.

The *Vulcan* Middleware

In this section, we describe the *Vulcan middleware* in which the *Vulcan architecture* is implemented. This middleware is not a traditional multi-agent system, as it lacks several features typical to multi-agent systems, including white-page and yellow-page functionality. It does, however, have many features central to multi-agent systems: it is comprised of a set of asynchronous, distributed components which communicate through a publish-subscribe model, using the LCM communications middleware(A. S. Huang, Olson, & Moore, 2010). And thus, we would argue that the *Vulcan* Middleware can be viewed as a MAS when viewed within the context of our larger integrated architecture. *Vulcan* and *ADE* differ in several important ways beyond those already mentioned. First, *Vulcan* is implemented in C++, sacrificing portability for speed; and while *Vulcan* components communicate through a publish-subscribe model, *ADE* does not in its current state have such methods available for inter-component communication.

8.2.3 Integrated Approach

In this section, we will present our approach to multi-agent system integration. We will first discuss the integration principles we followed in our integration efforts. Next, we will describe how those principles are employed in our integration, and the benefits reaped from this integration. Finally, we will discuss the actual components of the two multi-agent systems that comprise our larger integrated system, and the hardware with which they interface.

The *Vulcan-ADE* Development Environment (*VADE*)

In this section, we present the *Vulcan-ADE* Development Environment (*VADE*) framework, a multi-agent system that integrates the Vulcan and *ADE* multi-agent system middlewares. *VADE* is comprised of three types of components:

- ***DIARC* Components**: *ADE* components that only exist within the *DIARC* architecture, and are only aware of components implemented in the *ADE* multi-agent system middleware.
- ***Vulcan* Components**: *Vulcan* components that only exist within the *Vulcan* architecture, and are only aware of components communicating on Vulcan's LCM channels.

- **New Dual-Citizen Components**: Components that exist within *both* the *DIARC* and *Vulcan* architectures, and can communicate both with *ADE* components through Java RMI and with *Vulcan* Components through LCM.

VADE's Dual-Citizen Components are functional components that require information and/or capabilities from both *DIARC* and Vulcan in order to provide their desired functions. These components exist within both architectures, and are aware of both multi-agent systems, thus effecting an inter-architectural bridge while maintaining flexibility and preventing single-architecture components of either architecture from needing to know about the single-architecture components of the other architecture.

These components are implemented as Java classes that both extend the *ADE* Component interface (allowing communication with *ADE* Components and the *ADE* Registry) and provide LCM Publisher/Subscriber interfaces (allowing communication with Vulcan Components). Of course, these components cannot *physically* extend to both architectures; we thus choose to grant them "primary" citizenship within the *ADE* multi-agent system framework, so that they can be started by the *ADE* Registry. This also means that if these components fail, they can be restarted automatically by the *ADE* Registry. If this happens, they will automatically resubscribe to the appropriate LCM channels, allowing Publishing and Subscribing functionality to automatically go back into effect.

Of course, this is not the only choice we could have made. One (expensive) option would have been to re-implement all of Vulcan architecture within the *ADE* multi-agent system middleware or all of *DIARC* architecture within the Vulcan middleware. However, this would not only have been monumentally time consuming, but would have removed functionality. Re-implementing Vulcan within *ADE* would have eliminated the speed advantages crucial to Vulcan's operations at the hardware level; re-implementing *DIARC* within Vulcan would have removed the OS-agnostic portability, easy distributability, and middleware features (e.g., dynamic system reconfiguration) afforded by ADE.

Another option would have been to implement a "bridge" component that handles all inter-architecture traffic. However, this would have been problematic for two reasons: (1) it would have created a computational bottleneck, and (2) in the case of failure of this component, all inter-architectural communication would necessarily cease. In contrast, if one Dual-Citizen Component goes down, other inter-architectural communication can still proceed as usual while the failed component goes through the process of restarting and reconnecting with both architectures.

Vulcan-DIARC implementation in *VADE*

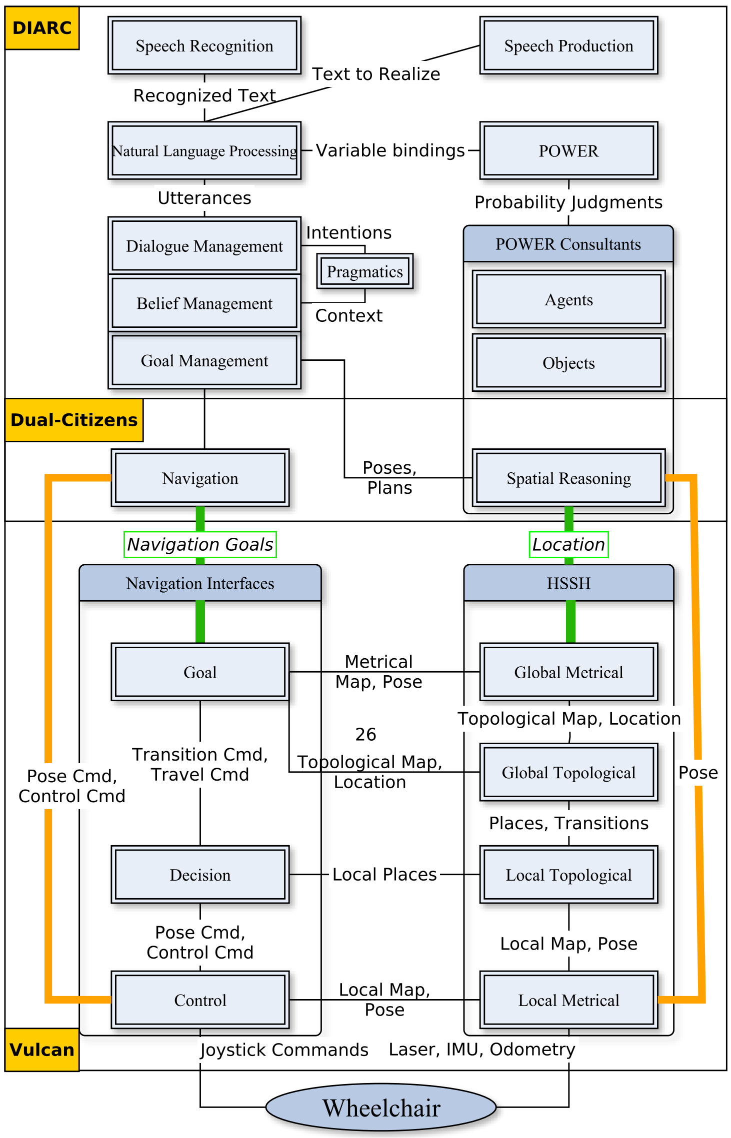
We will now discuss how *Vulcan* and *DIARC* have been implemented in the *VADE* framework. We will begin by discussing the benefits of such an integration at the architectural level, and then discuss our use of Dual-Citizen components within this implementation.

There are a number of advantages to *Vulcan-DIARC* integration at the architectural level: by integrating *Vulcan* and *DIARC*, each can leverage the other's capabilities, resulting in new synergistic capabilities and behaviors. *DIARC* alone is unable to engage in dialogue regarding large-scale spatial locations, not because it lacks the linguistic faculties, but rather because it lacks significantly rich spatial representations – such representations can be provided by *Vulcan.* Similarly, we have typically restricted *DIARC* to small, simple, indoor environments – *DIARC* can leverage *Vulcan's* spatial reasoning and mapping capabilities in order to discuss, reason about, and travel through larger environments.

Similarly, *Vulcan* requires commands to be precisely specified within its map representation, e.g., using metric coordinates, a topological action, or a named topological location. But by leveraging *DIARC's* linguistic capabilities, *Vulcan* can travel to locations that are only loosely specified; a natural language comment like "I need my coffee!" does not clearly specify a location, yet can be used to infer such a location, as well as the fact that the robot should travel to it.

As shown in Figure 8.5, the majority of *VADE* components are pure *Vulcan* and *ADE* components. with two components serving as Dual-Citizen components: the Navigation Component and the Spatial Reasoning Consultant Component. On the *ADE* side, these interact, respectively, with the Goal Management and *DIST-POWER* components, as we will later describe. On the *Vulcan* side, we would ultimately like to effect integration at multiple levels of the HSSH so that *DIARC* can leverage *Vulcan's* rich topological representations. At this point, however, these dual citizen components only communicate, respectively, with *Vulcan's* Control and Local Metrical components. This provides *DIARC* with access to pose representations maintained by *Vulcan*, which can be used to determine the robot's current topological location within *DIARC's* own internal topological map (maintained in the Spatial Reasoning Consultant *ADE* Component). *DIARC* is then able to navigate through its internal topological map by sending a motion target to the *Vulcan* motion planner when it decides it needs to visit the a particular

Figure 8.5: Diagram for the integrated system



As labeled, the top half of the graph corresponds with the *DIARC* architecture, the bottom half corresponds with the *Vulcan* architecture, with the Navigation and Spatial Reasoning components in the overlapping Dual-Citizen region belonging to both architectures. Thick orange lines represent the existing inter-architecture integrative connections. Thick green lines (with italicized labels) represent the intended future inter-architecture connections. Connections between the Dialogue, Belief, and Goal Managers are not shown due to density of connections.

topological location. A result of this integration is that *DIARC* is now able to operate in more dynamic and difficult environments.

In the following sections, we will go into *VADE's* Components in depth. As we have previously stated, however, in this work we use a robot wheelchair as one example application of our integrated approach, and as one example application of the work presented in this dissertation. In the next section we will thus discuss the specific hardware architecture of that wheelchair.

Hardware Architecture

The presented robotic wheelchair (Figure 8.4) is a commercially-available powered wheelchair (Quantum 6000Z) modified to enable computer control and augmented with two Hokuyo UTM-30LX laser rangefinders (one at the front-right corner and one at the back-left corner) to provide a 360°view of the wheelchair's surroundings. Wheel encoders and an inertial measurement unit are mounted on the wheelchair to enable high-precision motor control.

The wheelchair is driven using a joystick. We enable computer control by intercepting CAN bus communication between the joystick and on-board controller. During autonomous driving, if the user is not actively moving the joystick to control the robot, any messages sent from the joystick module are replaced by commands calculated by *Vulcan's* motion controller. In this way, we always defer to human control.

Vulcan Components

The *Vulcan* robot architecture separates the overall problem of mapping, localizing, and navigating into two modules, as shown in the lower half of Figure 8.5.

1. HSSH: As previously described, the robot's map within *Vulcan* is represented using a variation of the Hybrid Spatial Semantic Hierarchy (HSSH)(Beeson, Modayil, & Kuipers, 2010), which factors spatial representations int four layers:

   The *Local Metric* layer represents the environment within the robot's sensory horizon as a *Local Perceptual Map (LPM)*, which maintains an occupancy-grid-based representation of the robot's current topological area. When the robot transitions between two topological areas, its LPM is entirely replaced by that associated with the new area.

   The *Local Topological* layer detects and classifies topological locations within the LPM as discrete, symbolic representations called *areas*.

However, the topological representation our collaborators use in *Vulcan* is slightly different from the original HSSH topological representation, in that (1) it distinguishes between *decision points* at path intersections, and *destinations* that a robot may be commanded to travel to, and (2) it uses a richer path representation that includes information about what destinations may be found along that path.

Finally, the *Global Topological* layer uses information from the local topological layer to dynamically create a global topological map using topological SLAM techniques (Johnson & Kuipers, 2012). This map is used to determine what topological map the robot is at, at all times, and to determine when new topological locations should be added to the map.

2. Navigation Interfaces:

   The *Control* interface interacts with the Local Metric layer in order to follow desired velocity commands or to travel to desired poses, while avoiding obstacles and navigating safely. This interface is implemented using a model-predictive control algorithm, MPEPC (Park, Johnson, & Kuipers, 2012; Park, 2016), developed by our collaborator Jong-Jin Park. This algorithm generates motion plans by choosing the locally optimal action based on a set of simulated possible actions. Here, optimality is defined by minimizing a cost function that includes action cost and collision cost while accounting for the uncertainty of the robot's motion and the motion of obstacles.

   The *Decision* interface uses the Local Topological layer to follow topologically-based commands, like driving to the end of a path or turning at a particular decision point: commands we might want to be able to trigger through natural language (cf. MacMahon, Stankiewicz, & Kuipers, 2006).

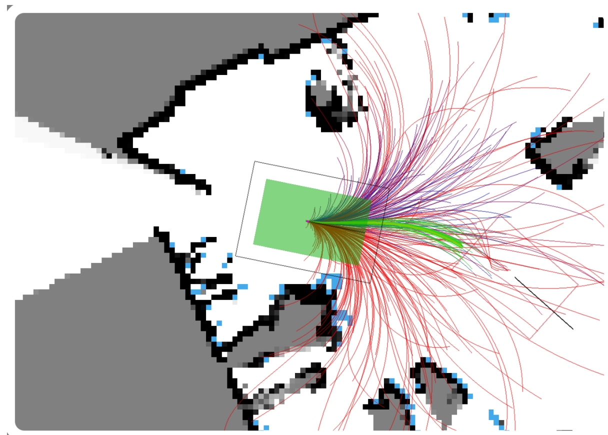
   The *Goal* interface uses the Global layers to follow other types of commands we might wish to specify through natural language, such as driving to a particular topological location.

ADE Components

In this subsection we will describe the *ADE* components used as part of the *VADE* framework, as shown in the upper half of Figure 8.5.

1. Speech Recognition and Production:

Figure 8.6: Visualization of *Vulcan's* metric path planning during execution of a natural-language navigation command.



Natural language utterances enter the architecture through *DIARC's Speech Recognition* component, which uses the Sphinx4 speech recognizer(Walker et al., 2004) to transduce speech into text. Similarly, the *Speech Production* component uses the MaryTTS library (Schröder & Trouvain, 2003) to synthesize text into vocal output.

2. Natural Language Processing (NLP):

   The NLP component first performs syntactic processing using the C&C CCG-based dependency parser(S. Clark & Curran, 2007). The produced dependency graph is then converted to a tree which is used for several purposes, as described in Chapter 4: (1) One variable X_i is instantiated for each referenced entity in the tree; (2) Logical formulae denoting properties and relations are instantiated for each property and relation in the tree, and semantic processing rules are used to analyze the tree in order to produce a formula for the tree's root node; (3) "Status cues" are associated with each referenced entity, based on what determiner (if any) is attached to that entity; (4) The utterance's illocutionary point (e.g., Statement, Question, Instruction) is determined based on the root node of the tree.

3. Reference Resolution:

   Reference resolution determines what entities in the robot's (possibly distributed and heterogeneous) knowledge bases should be associated with each *referenced* entity, using the previously presented *DIST-POWER* algorithm. *DIST-POWER* is designed to operate in uncertain and open worlds, and handles references to both known and unknown entities. Specifically, the Givenness-Hierarchy-theoretic *GH-POWER* algorithm presented in Chapter 4 is used. As a reminder, *GH-POWER* uses a hierarchical cognitively-inspired memory structure (consisting of the *Focus of Attention*, *Short Term Memory*, *Discourse Context*, and *Long Term Memory*) to resolve definite noun phrases and anaphoric and deictic expressions.

The end product of reference resolution is a set of mappings from variables to memory traces associated with entities in a robot's long-term memory. These memory traces are used to create a set of bound semantic structures, which differentially bind the open variables of the logical formula associated with the utterance's root node. These bound semantic structures are used in turn to create bound utterance representations which are sent to the pragmatics component.

4. Pragmatics:

   DIARC's Pragmatics component uses a set of context-sensitive Dempster-Shafer-Theoretic logical rules to determine the intention underlying each candidate utterance representation, as described in Chapters 6 and 7. This results in a set of *belief updates* which are passed to the Dialogue, Belief, and Goal Management components (DBGM, collectively).

5. Dialogue, Belief, and Goal Management (DBGM):

   DIARC's Dialogue, Belief, and Goal Management components (DBGM, collectively) are responsible for tracking and coordinating dialogue (Briggs & Scheutz, 2012), storing beliefs and performing inference in a general-purpose knowledge base, and tracking and acting on goals (Brick, Schermerhorn, & Scheutz, 2007). If the DBGM needs to respond to its user, it sends its own intention back through the Pragmatics component, which can work in reverse to determine the utterance which should be used to communicate a particular intention. If a robot's interlocutor uses a command to instruct the robot, the DBGM instantiates a new goal based on intentions underlying that command and determines how to accomplish it. Of particular relevance are commands to travel to particular locations, which is accomplished by *DIARC's* Spatial Reasoning Component.

6. Dual-Citizen Components:

   The Spatial Reasoning Consultant Component (SRC) serves as the primary *knowledge* interface between *DIARC* and *Vulcan*. The SRC maintains a graph of both large-scale and small-scale topological locations, where connectivity indicates either physical adjacency or an "in" relationship. Each location is associated with an identifier, list of properties, and (in the case of grounded small-scale locations), coordinate pose.

   When *DIARC* determines that the robot must navigate to a particular location, the SRC finds the shortest path to that location, and incrementally sends the robot to the coordinates of each intermediate waypoint by sending them to the Navigation Component, which serves as the primary *action* interface between *DIARC* and *Vulcan*. The Navigation Component then broadcasts those coordinates over the appropriate LCM channel. Similarly, when *DIARC's* goal manager determines the robot needs to simply drive forward, turn, or stop, the navigation component effects these motions by broadcasting messages over other LCM channels. The next step of our integration efforts, however, will be to integrate *Vulcan's* topological capabilities with *DIARC*: when this is accomplished, many of the responsibilities of these two components will be transferred to pure *Vulcan* components, providing more of a communicational role to these two Dual-Citizen Components.

8.2.4 Demonstration

We will now present a proof of concept demonstration of our integrated approach, as implemented on the robotic wheelchair we previously described. A video of this demonstration can be viewed at `tiny.cc/wheelchairdemo`. In this demonstration, the wheelchair begins in an office environment, and is told by its rider ("Jim") "I need my coffee!". After recognition, *DIARC's* ASR component passes this utterance to its NLP component, which performs parsing and reference resolution. This utterance is parsed into the utterance form $Statement(jim, self, need(jim, X))$ with supplemental semantics $coffee(X)$.

At the start of this interaction, the robot's *Short Term Memory* and *Focus of Attention* are both empty, and thus the robot's *Long Term Memory* is searched for a suitable referent to bind to the variable X. The property $coffee(X)$ is advertised only by *DIST-POWER*'s *objects* consultant, which manages a knowledge base of known objects. This KB starts off with knowledge of a handful of objects, including their properties and locations. Included in this set is one coffee-like entity, with memory trace obj_5. This trace is bound to X, producing $Statement(jim, self, need(jim, obj_5))$, which is passed to *DIARC's* pragmatic reasoning component. This component has a rule with implicative content:

$$Statement(X, Y, need(X, Z)) \Rightarrow goal(Y, have(X, Z)),$$

resulting in the goal $have(jim, obj_5)$ being adopted.

The SRC identifies the location of obj_5 as loc_{51}, and creates a plan to visit the set of waypoints on the path to loc_{51}. *DIARC's* DBGM then executes this plan one step at a time: for each waypoint, the DBGM acquires the coordinates of that location from the SRC and passes them to *DIARC's* Navigation component, which in turn broadcasts these coordinates over LCM to *Vulcan*.

The command from the Navigation component is received by *Vulcan's* Control component, which initiates a new motion planning task to drive to the specified coordinates. Using MPEPC, the wheelchair performs the task by driving to the desired coordinates. During motion planning, the state of the environment is estimated at 10Hz, including the position and velocity of pedestrians around the robot, the location of static obstacles, and the wheelchair's own position and velocity. This fast update allows the wheelchair to safely navigate even through dense crowds.

When the wheelchair arrives at a destination, it broadcasts an LCM message indicating action success. When this message is received by the Dual-Citizen Navigation component, it moves on to the next step in its navigation plan: once again, the DBGM will acquire the coordinates of the next small-scale place along the path, and send those coordinates to the Navigation component. As this process iterates, the wheelchair drives down several hallways until it reaches the door to the room containing the coffee. The robot then turns, and drives through the doorway in order to reach the last point along the route.

This section demonstrates capabilities enabled by our integrated approach: this should not be taken as, nor is this intended to be, a formal empirical evaluation. While the components of the *Vulcan* and *DIARC* architectures have been evaluated independently, a holistic, extrinsic evaluation of this integrated approach will still eventually be necessary.

Figure 8.7: Metric-topological map of the demonstration environment

This 44m x 75m map was produced by running *Vulcan's* metric SLAM and place classification algorithms on sensor data of the wheelchair being manually driven through an environment. Color indicates the "type" of each region in the topological map: decision points are blue, path segments are green, and destinations are red.

8.2.5 Discussion and Lessons Learned

Both *DIARC* and the HSSH are fully fledged robot architectures, and as such, have a number of overlapping capabilities implemented in their respective middlewares. Both architectures typically make use of the robot's laser rangefinders, for example, and both typically send motor commands to effect robot motion. But both architectures cannot be responsible for these overlapping capabilities, and thus we had to decide which architecture should cede some of its control. Because *Vulcan* solely focuses on spatial reasoning and motion planning, it makes sense for *DIARC* to defer its control. From *DIARC's* point of view, each Dual-Citizen component is just another component, which happens to provide these ceded capabilities, and is not aware that those components are in fact part of an entire other architecture, or that the motion primitives sent to the Navigation component may spawn complex navigational procedures. Similarly, *Vulcan* views its Dual-Citizen components as just another publisher/subscriber – it receives new motion targets, but is unaware that they are coming from *DIARC* and not a human user.

It is also interesting to consider the practical challenges posed by integrating two discrete robot architecture-robot middleware pairs. During the integration process, each architecture was independently progressing, and gaining hosts of new capabilities. As an effect, the Dual-Citizen components naturally evolved over time along with the target functionality for the joint integrated system.

Finally, we must discuss the specific contributions made by the work presented in this section. First, we demonstrated how the integration of the *Vulcan* and *DIARC* architectures produces an architecture with not only the capabilities of both architectures, but new synergistic capabilities as well: by leveraging *Vulcan*, *DIARC* is able to navigate through environments, *Vulcan* is able to initiate actions based on flexible natural language requests, and as a whole, *Vulcan-DIARC* is now able to travel remote, previously unknown objects, a capability previously held by neither architecture.

Second, we showed how this integration could be implemented in a new multi-agent system comprised of agents from two distinct multi-agent systems, plus Dual-Citizen agents that belonged to *both* multi-agent systems. This provides a novel, useful framework for integrating multi-agent systems which could be used for future integrated approaches.

Finally, we showed how, when implemented on a robotic wheelchair, this integration significantly extends the state-of-the-art for NL-enabled wheelchairs. Like a small number of other recent wheelchairs(S. Hemachandra, Kollar, Roy, & Teller, 2011; Megalingam, Nair, & Prakhya, 2011; Murarka, Gulati, Beeson, & Kuipers, 2009; Petry, Moreira, Faria, & Reis, 2013; Röfer, Mandel, Lankenau, Gersdorf, & Frese, 2009; Tao, Wang, Wei, & Chen, 2009), our wheelchair is able to travel to described objects and locations. Within this set of wheelchairs, however, ours is unique with respect to its cognitive approach: to our knowledge, no previously presented NL-enabled wheelchair has been capable of handling natural, indirect language (cf. Chapter 6), hypothesize new objects and locations based on natural language (cf. Chapter 3), model cognitive structures to resolve anaphora (cf. Chapter 4), or ask clarification questions (cf. Chapter 7), all of which are afforded to our wheelchair through this integrated approach.

8.2.6 Future Work

As previously discussed, there are a number of architectural interfaces that have not yet been implemented, most notably the use of the HSSH's rich topological representations within *DIARC*, and the integration of *Vulcan's*

Decision and Goal interfaces. This represents our immediate next step for future work. Similarly, there are new capabilities we would like to implement that take advantage of the synergies provided by the integration of the two architectures. Through the integration of the remaining architectural interfaces, we should be able to enable the wheelchair to travel to previously unknown *locations* (in addition to the ability it already possesses to travel to previously unknown *objects* located in known locations (cf. Chapter 3.2)). And through the integration of novel episodic memory management capabilities, we should be able to further leverage these synergies in conjunction with richer models of a robot's interlocutors' behaviors and preferences. In the case of a robot wheelchair, this should allow the wheelchair to follow directives such as "Bring me to *my barbershop*" or "Let's go to *the park we visited last week*." Finally, it is important to note that the presented application to a robotic wheelchair is only one example application; the integration of these two architectures may well lead to significant advances in other domains as well.

8.3 General Discussion

While this chapter primarily served to demonstrate a real-world application of the algorithms presented in this dissertation, it also provides four research contributions. First, we present the first comprehensive survey of natural language enabled robotic wheelchairs. Second, we demonstrated how the integration of the *Vulcan* and *DIARC* architectures produces an architecture with not only the capabilities of both architectures, but new synergistic capabilities as well. Third, we presented *VADE*, a novel multi-agent system framework comprised of agents from two distinct multi-agent systems, plus new *Dual-Citizen* agents that belong to *both* multi-agent systems. And fourth, we demonstrated how this architectural integration, when implemented in this novel multi-agent system framework on the hardware of a robotic wheelchair, significantly extends the state-of-the-art for NL-enabled wheelchairs.

Chapter 9

Conclusions

In Chapter 1, I laid out a vision of a world in which intelligent agents such as robots engage with humans in natural, human-like task-based interactions. In this dissertation, I have endeavored to make progress towards this goal through the development of algorithms and architectural mechanisms that allow robots to engage in natural, pragmatically appropriate, task-based dialogue in uncertain and open worlds. Furthermore, I have sought to ensure that the algorithms and architectural mechanisms that I have designed are not tied to any one domain or knowledge representation scheme, in order to maximize their usefulness across architectures and application domains, and in order to reflect the realities of modern integrated robotic architectures. As such, this dissertation makes a number of *theoretical* and *technical* contributions.

9.1 Theoretical Contributions

This dissertation makes the following *theoretical* contributions: (1) Experimental evidence suggesting that indirect speech act use is central to human-robot dialogue regardless of task contexts (although especially so in highly conventionalized scenarios), even when robots demonstrate an inability to understand such speech acts; (2) experimental evidence suggesting that robots *unable* to understand indirect speech acts may thus be less efficient and less favorably viewed than robots that *are* able to understand indirect speech acts; (3) design recommendations for developers of language-enabled robots, based on these findings; (4) an ontological analysis of indirect speech act forms observed in these experiments; (5) experimental evidence suggesting that participants prefer robots that generate pragmatically appropriate clarification requests in which options are listed, especially when there are only two options; (6) design recommendations for developers of language-enabled robots, based on these findings; (7) experimental evidence suggesting that verbal robot-robot communication is preferable to silent robot-robot communication in the context of co-located human-robot team tasks, because it is perceived as *less creepy*; (8) design recommendations for developers of robots intended for such contexts, based based on these findings; (9) a cognitive model of open-world reference resolution; (10) a novel framework for evaluating referring expression generation algorithms; (11) the first comprehensive survey of natural-language enabled robotic wheelchairs.

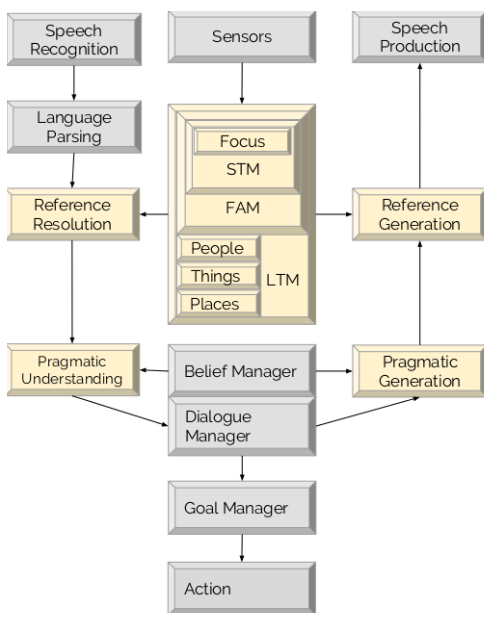
9.2 Technical Contributions

This dissertation makes the following *technical* contributions: (1) an algorithm for location-based spatial reference resolution in open worlds; (2) a framework for referential processing in uncertain and open worlds; (3) a referring expression *understanding* algorithm that makes use of this framework; (4) a referring expression *generation* algorithm that makes use of this framework; (5) a cognitively inspired memory model that extends this framework; (6) two referring expression understanding algorithms that make use of both this memory model and the presented referential processing framework in order to handle a wider range of referring expression forms; (7) a Dempster-Shafer Theoretic framework for pragmatic reasoning under uncertainty and ignorance; (8) an indirect speech act *understanding* algorithm that makes use of this framework; (9) an indirect speech act *generation* algorithm that makes use of this framework; (10) a human-robot interaction oriented clarification request generation framework; (11) and a demonstration of how the algorithms presented in this dissertation can be integrated together to achieve all five stages of that framework, in order to generate clarification requests to resolve both referential and intentional ambiguity; (12) a novel multi-agent system framework comprised of agents from two distinct multi-agent systems as well as new *Dual-Citizen* agents that belong to both multi-agent systems; and (13) a demonstration of how the integrated set of algorithms and multi-agent systems, when implemented on the hardware of a sensor-augmented powered wheelchair, extends the state of the art of natural language enabled robotic wheelchairs.

9.3 Implications of Contributions

In order to best demonstrate these technical and theoretical contributions, it may be helpful to reconsider the architecture diagram presented in Chapter 1, which I re-present here for clarity. In this diagram, we see in the middle the hierarchical GH-theoretic memory model which constitutes technical contribution (TeC) 5, and the distributed heterogeneous knowledge base framework which serves as Long Term Memory with in that model, and which constitutes TeC 2. To its left and right, we see the components that make use of these models: one component for performing referring expression understanding (i.e., reference resolution) using the *DIST-POWER* and *GH-POWER* algorithms presented in Chapters 3 and 4 (TeC 3 and TeC 6, respectively), and one component for performing referring expression generation, using the *DIST-PIA* algorithm presented in Chapter 5 (TeC 4). Below these components, we see the components that perform pragmatic understanding and generation, as described in Chapters 6 and 7 (TeC 8 and TeC 9, respectively). More concretely, these contributions allow this inte-

Figure 9.1: Re-Presentation of Motivating Architectural Diagram



grated architecture to understand and generate natural language expressions in ways that were not previously possible in our architecture.

Previously, our architecture had only rudimentary reference resolution capabilities, and could resolve a referring expression only if it unambiguously referenced some known entity about which knowledge was stored in some single centralized repository. Now, our architecture can resolve references in *uncertain and open worlds*, and our architecture now has an easily extensible framework for making knowledge from various components available to the language processing system without requiring those components to maintain their knowledge in a centralized Prolog knowledge base.

Previously, our architecture had rudimentary anaphora resolution capabilities, and would always resolve anaphoric expressions like 'it' to the most recently mentioned entity that matched numeric and gender constraints. Now, the architecture has a much more nuanced procedure for resolving anaphoric expressions. And furthermore, our GH-theoretic framework can also resolve other referring expressions, such as *definite* noun phrases, much more *quickly and accurately* than it previously could have, because it will now first check entities that are considered "activated" before checking all entities in long term memory.

Previously, our natural language generation system could generate referring expressions based only on the properties with which it was provided. Now, we have a general, extensible framework that is used to automatically choose the best properties to use to describe some entity of interest. Furthermore, the information used to make this judgment may be distributed throughout the architecture rather than centralized in a single Prolog knowledge base.

And finally, while we could previously understand and generate indirect speech acts only under an assumption of perfect knowledge and a single interpretation, we can now understand and generate indirect speech acts even when knowledge is uncertain, and can identify and appropriately handle cases where there are multiple possible interpretations.

But of course, there are still many things our architecture cannot do, and there are capabilities within our architecture that could be improved. Over the course of this dissertation's first eight chapters, I described a variety of directions for future work that immediately builds off of the twenty-four contributions listed thus far in this chapter. In the next section, I will discuss possibilities for future work that reach farther afield.

9.4 Future Work

In this section, I present five possible directions for future work.

- In Section 9.4.1, I discuss how the work presented in Chapter 3 can be advanced from a cognitive modeling perspective.
- In Section 9.4.2, I discuss how the work presented in Chapter 5 can be advanced from the Givenness Hierarchy theoretic perspective described in Chapter 4.
- In Section 9.4.3, I discuss how the work presented in Chapters 6 and 7 can be advanced from a machine learning perspective.
- In Section 9.4.4, I discuss how the work presented in this dissertation can be applied to *search and rescue robotics* applications.
- In Section 9.4.5, I discuss how the work presented in this dissertation can be applied to domains *beyond* robotics.

9.4.1 Cognitive modeling and Cognitive architectures

In Chapter 3, I showed how the decisions made by the *DIST-POWER* algorithm align well with the most common decisions made by humans. However, if one considers the psycholinguistics research presented in Section 3.1 it is clear that *DIST-POWER* only successfully models the *final results* of reference resolution, and not the results on a moment-by-moment basis, as it does not model the *process* of reference resolution (see also (Chater & Christiansen, 2008; Sun, 2008)). This is a crucial point, given that language is a *collaborative process* (H. H. Clark & Wilkes-Gibbs, 1986; H. H. Clark & Schaefer, 1987; Garrod & Pickering, 2009). Humans do not typically wait until a sentence has been completely processed to consider its meaning. And in fact, the responses interlocutors provide while listening to a speaker may critically affect whether and how the speaker chooses to continue their utterance.

In future work, it will thus be important to develop an *incremental* version of *DIST-POWER*, not only to improve its performance and usefulness, but in order to develop new psycholinguistic models of the open world reference resolution *process*. There have recently been many approaches to incremental sentence understanding, the major approaches (as per Hough, Kennington, Schlangen, & Ginzburg, 2015) being Purver, Eshghi, & Hough (2011)'s use of Dynamic Syntax (Kempson, Meyer-Viol, & Gabbay, 2000) and Type-Theory with Records (R. Cooper, 2005) and Robust Minimal Recursion Semantics with incremental processing (Peldszus, Baumann, Buß, & Schlangen, 2012) (see also Aist, Allen, Campana, & Gallo, 2007; Baumann & Schlangen, 2012; DeVault, Sagae, & Traum, 2009, 2011; Kennington & Schlangen, 2012; Kennington, Kousidis, & Schlangen, 2013; Kennington, Iida, Tokunaga, & Schlangen, 2015). An incremental version of *DIST-POWER* would likely extend one of these approaches in order to better handle the hypothesization of new entities during this incremental procedure and to allow for integration with the Distributed Heterogeneous Knowledge Base memory model used in our architecture.

Similarly, it will be important to develop an incremental version of *GH-POWER* for the same reasons. This may be especially fruitful, because *GH-POWER's* operation at the intersection of language, attention, and memory may present a unique opportunity to study the interaction between these disparate cognitive processes. An incremental version of *GH-POWER* would present additional challenges as one would need to find some computationally efficient means of combining the general incremental reference resolution procedure with *GH-POWER's* iterative consideration of combinations of data structures through which to search.

Finally, it will be important to develop mechanisms that allow for pragmatic reasoning and general inference procedures to be employed during this incremental process (cf. Peldszus, Baumann, Buß, & Schlangen, 2012). This is important not only to rule out resolution candidates that violate common sense reasoning principles or thematic roles, but in addition to prevent potential implications during the robot's response that could raise ethical concerns. Consider the following interaction:

Human User: Can you bring me to the kitchen?

Wheelchair: Do you mean the kitchen on the first floor or the second floor?

Human User: The kitchen on the second floor.

Wheelchair: No, I am unable to travel to that floor.

Human User: Okay, can you bring me to the kitchen on the first floor?

Wheelchair: No, that kitchen is closed for repairs.

Here, the user is likely to come away frustrated. It does not matter which location the user meant to refer to; in either case, the eventual response was a rejection of the user's request. As such, the wheelchair's initial clarification request was simply a waste of the user's time, and accidentally implies that the robot could comply with the user's request under at least one of the ambiguous interpretations. What would have been more appropriate would have been for the wheelchair to perform some counterfactual reasoning as to what its response would be under each possible interpretation, and generate a blanket rejection if no possible interpretation is acceptable, such as "I am unable to bring you to either kitchen I know of", "I'm not sure which of the kitchens you mean but I cannot travel to either of them", or "I'm not sure which of the kitchens you mean, but I cannot travel to the one on the second floor, and the one on the first floor is closed for repairs."

In this case, the initial clarification request is simply frustrating. However, in slightly different circumstances, the accidental-implication made by this clarification request could actually raise ethical concerns. Consider for example the following alternative dialogue.

Human User: Can you run over Alice?

Wheelchair: Do you mean Alice Robertson or Alice Charleson?

Here, it does not matter which Alice the human user meant; it is morally impermissible to run anyone over. By simply asking for clarification, however, the wheelchair is accidentally implying that it would be willing to comply with the user's request under at least one of the ambiguous interpretations. To avoid this ethically fraught response, it would have been more appropriate for the wheelchair to perform some counterfactual reasoning as to what its response would be under each possible interpretation, and generate a blanket rejection if no possible interpretation is acceptable, such as "I am not allowed to run over *anyone*" or "I'm not sure which Alice you mean, but I'm not allowed to run over either of them."

It will also be important to make these extensions *parallelized*. Currently, *DIST-POWER* and *GH-POWER* only consider one candidate at a time. These algorithms should be extended to simultaneously consider as many candidates as possible, for the sake of increased efficiency.

Finally, these extensions suggest yet another direction for future work. If we are indeed interested in studying the interaction of cognitive processes within the *GH-POWER* framework, it may be valuable to design a new cognitive architecture (Langley, Laird, & Rogers, 2009) that is grounded in this framework, in the way that SOAR is grounded in its particular model of working and long-term memory (Laird & Rosenbloom, 1996; Laird, 2012).

9.4.2 Cognitive referring expression generation

If one considers the parallels between language understanding and generation in this dissertation, there is an obvious blemish upon the included chapter pairs. I include discussion of referring expression understanding and generation in Chapters 3 and 5. I include discussion of pragmatic understanding and generation in Chapters 6 and 7. And I include discussion of *contextualized* referring expression understanding, including understanding of anaphoric and deictic expressions, in Chapter 4 – but I *at no point* discuss contextualized referring expression *generation*. This means that the architecture, as it stands, is unable to make use of any sort of anaphoric or deictic expression, and considers all known entities as distractors when generating referring expressions.

In future work, it will thus be critical to design such a process, by "inverting" the *GH-POWER* algorithm presented in Chapter 4. This will require the development of a number of new mechanisms and data structures, including data structures similar to the GH-theoretic memory model used for GH-theoretic referring expression understanding, but designed to model the likely contents of interlocutors memories. It will be left to such future work to determine the appropriate level of granularity for such data structures, and the degree to which it is appropriate for these data structures to interact with those used for understanding (due to the fact that both structures will in a way endeavor to model what is in *common ground*, see also (Stalnaker, 1978; H. H. Clark, Schreuder, & Buttrick, 1983; Horton & Keysar, 1996)).

Future work in this area will need to consider alternative theories, such as Ariel's Accessibility Theory, which specifically considers the representation alignment as a motivation behind reference production (Ariel, 1988, 2001; Bard, Hill, Foster, & Arai, 2014). Next, we will need to consider recent algorithms that have been grounded in such theories in order to facilitate contextualized referring expression generation (e.g. Foster, Giuliani, & Isard, 2014).

Once GH-theoretic algorithms for contextualized referring expression generation are developed, they will need to be integrated with *multimodal* adaptations of the work presented in Chapter 5 (cf. Kranstedt & Wachsmuth, 2005), and in general draw on the breadth of recent work in human-robot interaction on non-verbal communication (e.g. Ng-Thow-Hing, Luo, & Okita, 2010; C.-M. Huang & Mutlu, 2014; Admoni, Weng, Hayes, & Scassellati, 2016; Admoni, Weng, & Scassellati, 2016) and highly legible pointing behaviors (e.g Holladay, Dragan, & Srinivasa, 2014; Gulzar & Kyrki, 2015; P. Liu, Glas, Kanda, Ishiguro, & Hagita, 2016).

9.4.3 Rule learning

Rule-based reasoning systems may be used for a variety of purposes in a robot architecture, from pragmatic reasoning components, as seen in my previous work, to components designed to ensure that robots behave ethically. How these rules might be learned is an interesting and extremely challenging problem, particularly with regards to determining the appropriate level of abstraction for a given rule, and determining the precise context under which a rule is applicable. I am interested in investigating these questions, using both statistical learning methods and the one-shot learning methods we have employed in previous work (Krause, Zillich, Williams, & Scheutz, 2014). In order to use this type of one-shot learning method, it will be important to determine how humans might best be solicited to teach new rules to robots through natural language. While it may not be appropriate for robots to learn *ethical* rules through natural language (cf. Ohlheiser, 2016), I believe it *would* be natural and appropriate for robots to learn rules for purposes such as indirect speech act and affordance reasoning through natural language. To this end, I have begun the process of experimentally investigating what types of robot behaviors are and are not effective for eliciting such rules from humans without directly asking for them. In the future, I am interested in continuing such experimentation, and in developing algorithms informed by the results of such experiments.

9.4.4 Search and Rescue Robotics

One of the largest success stories for modern robotics is the field of *rescue robotics* (Murphy et al., 2008; Murphy, 2014). Robots that assist with rescue efforts in both urban and rural disaster scenarios have the potential for great societal impact. What is more, there is already a market for such robots, with organizations willing to pay large sums of money for rescue robots as part of normal operating costs (Birk & Carpin, 2006).

Rescue robots are also of great interest from a research perspective. While current rescue robots are almost entirely teleoperated, the ideal rescue robot of the future will be nearly or fully autonomous. Because rescue robots are already being deployed in the wild, this provides researchers the opportunity to incrementally introduce new capabilities to these deployed robots (especially perception, modeling, locomotion, manipulation, and cooperation capabilities (Birk & Carpin, 2006)) in order to bootstrap their autonomous capabilities.

Rescue robotics is also interesting from a *multi-agent* and *human-robot* *teaming* perspective, as a wide variety of team compositions may be used to control and interact with rescue robots. As more autonomous capabilities are enabled in rescue robots, control shifts from a typical one-to-one teleoperation scheme to a one-to-many control scheme in which a single robot coordinator is able to manage a group of rescue robots (Nevatia et al., 2008; Crandall, Goodrich, Olsen, & Nielsen, 2005), or a many-to-many control scheme (Tews, Mataric, & Sukhatme, 2003; Yanco & Drury, 2004) in which multiple humans work with a fleet of rescue robots as part of a heterogeneous human-robot team (Chou, Marsh, & Gossink, 2009; Murphy, 2004).

The same autonomous capabilities that enable human-robot teaming beyond mere teleoperation also create a cycle in which even more advanced autonomous capabilities are motivated, making them of great interest from the perspectives of *artificial intelligence* and *human-robot interaction*. For example, if rescue robots can be tasked through natural language communication, then there is motivation to provide such robots with the ability to model the beliefs, actions, and intentions of other robots and humans, in order to interpret their orders within the context of orders that they have heard being given to other robots (Briggs & Scheutz, 2011; Talamadupula, Briggs, Chakraborti, Scheutz, & Kambhampati, 2014) or more effectively generate plans that respect the goals of the larger human-robot team or the *uncertain and open-world* nature of the rescue environment (Talamadupula, Kambhampati, Schermerhorn, Benton, & Scheutz, 2011; Talamadupula, Benton, Kambhampati, Schermerhorn, & Scheutz, 2010). What is more, if rescue robots are given natural language capabilities, they may be expected to interact not only with their explicit teammates, but also with naïve citizens such as the victims themselves, who cannot be expected to have any knowledge of the "right" way to interact with robots in order to fulfill their needs (Murphy, 2004).

While much attention has been paid within robotics and human-robot interaction to the specific application of *urban* search-and-rescue robotics (e.g. Davids, 2002; Casper & Murphy, 2003; Burke, Murphy, Coovert, & Riddle, 2004), the use of robots is also important to outdoor search and rescue environments such as *mountain* or *wilderness* search and rescue (May, 1973), applications in which the use of either teleoperated or autonomous unmanned aerial vehicles (UAVs) play an especially important role (Goodrich et al., 2007, 2008; Goodrich, Lin, Morse, & Roscheck, 2010).

For this reason, a number of European researchers are currently taking part in the *SHERPA* research project (Marconi et al., 2012), including our collaborators from the Institute for Artificial Intelligence at the University of Bremen (Yazdani, Brieber, & Beetz, 2016). The *SHERPA* project focuses specifically on *alpine* search and rescue scenarios, which require tremendous expenditures even at relatively small scales. In 2010 for example, 30,000 rescuers took part in alpine search and rescue scenarios in Italy alone, resulting in 6000 rescued persons and 450 fatalities (Marconi et al., 2012). In order to best assist within this domain, researchers on the *SHERPA* project are focusing on developing teams of heterogeneous air and ground robots that are remotely commandable by one or more human operators through natural, task-based interactions, with four central research objectives: (1) facilitating *multi-modal* control of rescue robots through methods such as speech and gesture; (2) providing decision support systems for both single- and multi-agent decision making; (3) designing robots specifically tailored for robustness in alpine search and rescue environments; (4) developing distributed, cognitive, multi-agent architectures.

While the work I have presented in this dissertation is applicable to search and rescue scenarios in general, it may be particularly valuable in such alpine search and rescue scenarios, for which perfect a priori domain knowledge is unlikely due to the remote, large-scale, unstructured, and dynamic properties of the environments in which such rescue operations are typically conducted. Furthermore, it is clear that the motivations behind this dissertation align particularly well with the first and fourth research objectives of the *SHERPA* project. For this reason, we have been working in collaboration with researchers from the Institute for Artificial Intelligence at the University of Bremen in order to integrate the *SHERPA* robotic framework with the DIARC robot architecture (as implemented in the ADE middleware) to produce a novel integrated system called *SHADE*.

Integrated Approach (*SHADE*)

SHERPA robots make use of a knowledge processing framework called *KnowRob* (Tenorth & Beetz, 2009), which provides Prolog-based (Clocksin & Mellish, 2003) knowledge representation and reasoning capabilities and a robot-oriented Ontology (Tenorth & Beetz, 2012) represented in a description logic using the Web Ontology Language (OWL) (McGuinness & Van Harmelen, 2004), all of which are implemented within the ROS middleware.

In order to integrate the *DIARC* architecture with the *SHERPA* architecture, we make use of *ROSADE*, an *ADE* utility used to generate ROSJAVA nodes (Kohler & Conley, 2011) that wrap running ROS nodes (cf. Wilson, Krause, Scheutz, & Rivers, 2016). Once these nodes are generated, *DIARC* is able to communicate with the ROS-native nodes through standard ROS messages (topics, services, and actions), similar to the use of LCM in *VADE* in the previous chapter. In practice however, nearly all inter-architecture communication is achieved by using the ROS services provided by *SHERPA*'s `json_Prolog` node, which allows other nodes to make Prolog queries regarding the information stored in *SHERPA*'s KnowRob knowledge base.

Under *SHADE*, two new *ADE* components are used. The KnowRob Component provides the ability to issue arbitrary Prolog queries to *SHERPA* by way of the aforementioned *ROSADE* nodes, and the *SHERPA* Component provides mechanisms for passing specific, *SHERPA*-oriented Prolog queries to the KnowRob Component.

The *SHERPA* Component as a Consultant

First, the *SHERPA* Component provides the mechanisms necessary for it to be considered a *consultant* (as discussed in Section 3.3.1), including Prolog queries that determine the set of known entities and handleable properties, and Prolog queries to assess and assert knowledge. However, this consultant differs from those discussed thus far, in a few unique ways.

Whereas the other consultants thus far have advertised *static* lists of constraints (i.e., typed positive-arity predicate symbols) that they are able to assess, this consultant does not *store* such a list, but rather acquires such a list by querying *SHERPA* about what properties it is prepared to answer questions about. This achieves a better division of responsibilities and facilitates integration in that our collaborators can add functionality without requiring any changes in our code. This also requires *SHERPA* to take responsibility for deciding on a preference ordering over the constraints it can handle.

This consultant also differs from those seen thus far in that the majority of information stored in *SHERPA*'s KnowRob knowledge base does not have uncertainty information associated with it. Thus, all assessment queries made by this consultant will return probabilities of 1.0 or 0.0. While this does not allow POWER's uncertainty handling to be shown off as well as the application in the previous chapter, it shows how knowledge bases with different uncertainty management schemes may work together seamlessly within our referential processing framework (as this uncertainty-free knowledge base may be used alongside other knowledge bases that *do* represent uncertainty).

Finally, this consultant differs from the others described thus far in that it is limited to unary and binary constraints due to the use of RDF triples (Klyne & Carroll, 2006) within KnowRob.

Additional *SHERPA* Component Capabilities

The *SHERPA* Component also provides capabilities that allow it to integrate with *DIARC*'s Goal Manager as well as with the Givenness Hierarchy theoretic memory model described in Section 4.3.

The *SHERPA* Component provides methods that will issue Prolog queries that will command a *SHERPA* UAV to fly to a particular object. These are associated with primitive actions within the *DIARC* Goal Manager's Action Interpreter, which will be selected when the Goal Manager attempts to achieve a goal to be "at" some entity. We have demonstrated in simulation that through this integration a simulated UAV (such as that used in Yazdani, Brieber, & Beetz, 2016) is successfully able to carry out the commands "Go to that tree", "Aid the victim", and "Help him".

In order to facilitate these types of commands, the *SHERPA* Component also provides methods that will issue Prolog queries to request the set of entities that are currently considered visually salient within the *SHERPA* simulation. Whenever a set of salient entities is received as a result of such a query, it is used to update the set of entities stored within the robot's Short Term Memory model (i.e., its GH-theoretic set of *activated entities*). This allows the robot to successfully carry out the commands above when there are many salient trees, victims, or male agents, but only one candidate referent that should be *salient*.

Integration Objectives

Since developing this integration, both *DIARC* and *SHERPA* have added a number of new pieces of functionality, and gone through a number of architectural changes. For example, *DIARC* now provides the referring expression generation and clarification request generation capabilities described in Chapters 5 and 7, and *SHERPA* now provides the ability to take pictures and recall back those images, and has gone through a number of architectural changes. In order to account for these architectural changes and allow for the integration of this functionality, we are currently working to overhaul *SHADE*. Through this process, we are also working to achieve new functionality. In the short term, our goal is to be able to successfully carry out an interaction such as that shown below:

Human	Take off.
Robot	Okay.
Human	Can you go left?
Robot	Okay.
Human	Can you go right?
Robot	Okay.
Human	Stop.
Robot	Okay.
Human	We have reached the hotzone.
Robot	Understood. Adopting search and rescue context.
Human	I need you to go to the pylon.
Robot	Do you want me to go to the gray pylon? Or do you want me to go to the blue pylon?
Human	I want you to go to the gray pylon.
Robot	Okay. [Travels to the pylon]
Human	Investigate there.
Robot	Okay. [Searches immediate area for victims]
Robot	I do not see any injured people, but I have found a ski pole.
Human	Okay.
Human	Can you take a picture?
Robot	Okay.
Human	Do you see that big rock? [Points towards large rock 20m away]
Robot	Yes.
Human	Can you search behind it for a victim?
Robot	Okay. [Travels to big rock and searches behind it]
Robot	I have found him.
Human	Okay.
Human	Take a picture.
Robot	Okay.
Human	Show it to me.
Robot	Okay.
Human	Can you show the picture of the skipole to me?
Robot	Okay.
Human	Thank you.
Robot	You are welcome.

9.4.5 Intelligent mixed-reality agents

Mixed reality technologies (specifically augmented and virtual reality technologies) are predicted to become a $120-Billion industry by the year 2020, and are poised to become one of the defining technologies of the coming decades (Merel & Tong, 2016). While my work has typically been situated within the field of human-robot interaction, I find augmented reality (AR) to be particularly interesting because it offers many of the same *perceptual* and *linguistic* challenges faced in human-robot interaction.

Imagine an intelligent, virtual agent that resides in a mixed-reality environment (i.e., one that contains entities that live in both the "virtual world" and "the real world" over which that virtual world is overlaid). A user interacting with that agent through natural language *may* refer to virtual objects about which the agent has perfect and complete knowledge a priori. But as with robots, a user may also refer to *real world* entities when interacting with an AR agent, and such an agent will thus need mechanisms for handling uncertain and incomplete knowledge such as those presented in this dissertation. Furthermore, previous research has shown that the "rules" of social human-robot interaction to a large extent carry over into interactions with virtual characters (Holz, Dragone, & O'Hare, 2009), and thus AR agents will also need to be able to understand and generate *pragmatically appropriate* utterances.

In addition, while users in human-robot interaction scenarios may refer to previously unknown objects, for which an agent may need to generate a new representation, this is not only *possible* in AR applications, it is *likely*, especially in *annotation* applications (Azuma, 1997) in which a user may request on-the-fly creation of new entities, for example, by using a request such as "Can you place a reminder on the pantry that says 'Buy Cereal'?"

AR is an excellent domain for applying and investigating multi-modal, situated communication techniques. Furthermore, AR is an excellent application domain for the development and deployment of intelligent agents, as AR agents can be meaningfully situated in their environment without being constrained by physics or cost with respect to embodiment. But surprisingly little research has been done on these topics to date. In fact, in their 2015 survey of augmented reality, Billinghurst, Clark, & Lee (2015) cite intelligent systems, hybrid user interfaces, and collaborative systems as areas that have been under-attended-to in the AR community.

This is not to say that there has been no research in this area. In fact, there has been several decades of work around these topics thus far. In the late 90s, work was done on multimodal VR speech-and-gesture interaction, in the contexts of 3D topological scenes (P. Cohen et al., 1999) and interior design (Laviola Jr, 1999), and also on intelligent "remembrance agents" intended to serve as AR "butler/confidant[s]" (Starner et al., 1997). More recently, researchers have developed systems that allow for multimodal interaction in AR and VR environments in order to effect changes in referenced *virtual* entities (Kaiser et al., 2003; Höllerer & Turk, 2006), dynamically add labels (Li & Jia, 2010) or multimedia annotations (Rekimoto, Ayatsuka, & Hayashi, 1998) to *real life* objects and locations, and dynamically augment *real life* objects with *virtual* capabilities (Barakonyi, Psik, & Schmalstieg, 2004).

With recent advances in mobile vision-based SLAM (Ventura, Arth, Reitmayr, & Schmalstieg, 2014), object recognition (Hoffman et al., 2014; Borji, Cheng, Jiang, & Li, 2015), and mobile AR technologies (Höllerer & Turk, 2006), I believe that additional research into the development of intelligent, mixed-reality, natural language capable agents is needed, and furthermore, I believe it would be an excellent domain in which to apply the work presented in this dissertation.

In addition, it will be interesting to consider the interaction between these two fields. For example, it would be interesting to more deeply consider human robot interaction *in mixed reality environments*. In such environments, a robot's referring expression generation algorithm may need to decide whether its utterances should be accompanied not only by gestures (e.g. Admoni, Weng, & Scassellati, 2016; Cassell et al., 1994; Cassell, 1998; Cassell, Stone, & Yan, 2000; C. L. Sidner & Lee, 2005; van der Sluis & Krahmer, 2007), but also by visualizations (not dissimilar from the offline cross-modal deictic expressions generated outside of the context of interactions in Wahlster, André, Graf, & Rist, 1991; Wazinski, 1992; André & Rist, 1994) or annotations (S. A. Green, Billinghurst, Chen, & Chase, 2007; S. Green, Billinghurst, Chen, & Chase, 2008). In fact, such annotations may be more useful in helping "pick out" the target of a referring expression than would a gesture. Or, alternatively, humans could place AR landmarks within an environment (as in Giesler, Salb, Steinhaus, & Dillmann, 2004; S. A. Green, Chase, Chen, & Billinghurst, 2009) in order to more easily craft robot-understandable referring expressions.

9.5 In Conclusion

In this final chapter, I have laid out the main contributions of this dissertation, and suggested several directions for future work. But beyond these contributions and possible future extensions, it is my hope that the work presented in this dissertation will directly benefit my research community in two ways.

First, I hope that the referential processing *framework* I have developed will be made use of by other researchers. Too often, there is pressure to design entirely new approaches rather than building off the work of others. It is my hope that the domain independence and extensibility of my reference processing framework will encourage its use as among other researchers.

And second, and more importantly, I hope that the work presented in this dissertation will draw attentions to the problems of *pragmatic reasoning* and *open world reference resolution*, problems I view as critical for enabling natural human-like human-robot interactions, but which have received very little previous attention.

References

Admoni, H., Weng, T., Hayes, B., & Scassellati, B. (2016). Robot nonverbal behavior improves task performance in difficult collaborations. In *Proceedings of the eleventh ACM/IEEE international conference on human-robot interaction (HRI)* (pp. 51–58).

Admoni, H., Weng, T., & Scassellati, B. (2016). Modeling communicative behaviors for object references in human-robot interaction. In *Proceedings of the IEEE international conference on robotics and automation (ICRA)* (pp. 3352–3359).

Aist, G., Allen, J., Campana, E., & Gallo, C. G. (2007). Incremental understanding in human-computer dialogue and experimental evidence for advantages over nonincremental methods. *Decalog 2007*, 149.

Allen, J. F. (1979). *A plan-based approach to speech act recognition* (Unpublished doctoral dissertation). University of Toronto.

Allen, J. F., Byron, D. K., Dzikovska, M., Ferguson, G., Galescu, L., & Stent, A. (2001). Toward conversational human-computer interaction. *AI magazine*, *22*(4), 27.

Allen, J. F., & Perrault, C. R. (1979). Plans, inference, and indirect speech acts. In *Proceedings of the seventeenth annual meeting of the association for computational linguistics (ACL)* (pp. 85–87).

Allen, J. F., & Perrault, C. R. (1980). Analyzing intention in utterances. *Artificial intelligence*, *15*(3), 143–178.

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Memory and Language*, *38*, 419–439.

Altmann, G. (1987). Modularity and interaction in sentence processing. *Modularity in knowledge representation and natural language processing*, 428–444.

Altmann, G., & Steedman, M. (1988). Interaction with context during human sentence processing. *Cognition*, *30*(3), 191–238.

Altmann, G. T., & Kamide, Y. (2007). The real-time mediation of visual attention by language and world knowledge: Linking anticipatory (and other) eye movements to linguistic processing. *Journal of Memory and Language*, *57*(4), 502–518.

Andersson, R., Ferreira, F., & Henderson, J. M. (2011). I see what you're saying: The integration of complex speech and scenes during language comprehension. *Acta Psychologica*, *137*(2), 208–216.

André, E., & Rist, T. (1994). Referring to world objects with text and pictures. In *Proceedings of the fifteenth conference on computational linguistics (COLING)* (pp. 530–534).

Andronache, V., & Scheutz, M. (2006). ADE—an architecture development environment for virtual and robotic agents. *International Journal on Artificial Intelligence Tools*, *15*(02), 251–285.

Ariel, M. (1988). Referring and accessibility. *Journal of Linguistics*, *24*(01), 65–87.

Ariel, M. (2001). Accessibility theory: An overview. *Text representation: Linguistic and psycholinguistic aspects*, *8*, 29–87.

Arnold, J. E., Eisenband, J. G., Brown-Schmidt, S., & Trueswell, J. C. (2000). The rapid use of gender information: Evidence of the time course of pronoun resolution from eyetracking. *Cognition*, *76*(1), B13–B26.

Arnold, J. E., Kam, C. L. H., & Tanenhaus, M. K. (2007). If you say thee uh you are describing something hard: the on-line attribution of disfluency during reference comprehension. *Journal of experimental psychology. Learning, memory, and cognition*, *33*(5), 914–930.

Arnold, J. E., Tanenhaus, M. K., Altmann, R. J., & Fagnano, M. (2004). The old and thee, uh, new: Disfluency and reference resolution. *Psychological Science*, *15*(9), 578–582.

Asakawa, T., & Nishihara, K. (2007). Operation assistance of a voice-controlled electric wheelchair. In *International workshop and conference on photonics and nanotechnology* (pp. 67942W–67942W).

Austin, J. L. (1975). *How to do things with words* (Vol. 367). Oxford university press.

Azuma, R. T. (1997). A survey of augmented reality. *Presence: Teleoperators and virtual environments*, *6*(4), 355–385.

Babri, O., Malik, S., Ibrahim, T., & Ahmed, Z. (2012). Voice controlled motorized wheelchair with real time obstacle avoidance. In *Proceedings of the third international conference on communications and information technology (iccit).*

Baker, M., Hansen, T., Joiner, R., & Traum, D. (1999). The role of grounding in collaborative learning tasks. *Collaborative learning: Cognitive and computational approaches*, 31–63.

Balch, T., & Arkin, R. C. (1994). Communication in reactive multiagent robotic systems. *Autonomous Robots*, *1*(1), 27–52.

Barakonyi, I., Psik, T., & Schmalstieg, D. (2004). Agents that talk and hit back: animated agents in augmented reality. In *Proceedings of the third IEEE and ACM international symposium on mixed and augmented reality (ISMAR)* (pp. 141–150).

Bard, E. G., Hill, R. L., Foster, M. E., & Arai, M. (2014). Tuning accessibility of referring expressions in situated dialogue. *Language, Cognition and Neuroscience*, *29*(8), 928–949.

Barr, D. J. (2008). Pragmatic expectations and linguistic evidence: Listeners anticipate but do not integrate common ground. *Cognition*, *109*(1), 18–40.

Baumann, T., & Schlangen, D. (2012). The inprotk 2012 release. In *Naacl-hlt workshop on future directions and needs in the spoken dialog community: Tools and data* (pp. 29–32).

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 interation (HRI)* (pp. 391–398).

Beeson, P., Modayil, J., & Kuipers, B. (2010). Factoring the mapping problem: Mobile robot map-building in the hybrid spatial semantic hierarchy. *The International Journal of Robotics Research*, *29*(4), 428-459.

Bejuri, W. M. Y. W., Saidin, W. M. N. W. M., Mohamad, M. M. B., Sapri, M., & Lim, K. S. (2013). Ubiquitous positioning: integrated GPS/wireless LAN positioning for wheelchair navigation system. In *Intelligent information and database systems* (pp. 394–403). Springer.

Bellifemine, F., Poggi, A., & Rimassa, G. (1999). JADE–a FIPA-compliant agent framework. In *Proceedings of practical applications of intelligent agents and multi-agent technology (PAAM)* (Vol. 99, p. 33).

Berjon, R., Mateos, M., Barriuso, A., Muriel, I., & Villarrubia, G. (2011). Alternative human-machine interface system for powered wheelchairs. In *Proceedings of the IEEE first international conference on serious games and applications for health (SeGAH)* (pp. 1–5).

Bezuidenhout, A. (2013). Perspective taking in conversation: A defense of speaker non-egocentricity. *Journal of Pragmatics*, *48*(1), 4–16.

Billinghurst, M., Clark, A., & Lee, G. (2015). A survey of augmented reality. *Foundations and Trends in Human-Computer Interaction*, *8*(2-3), 73–272.

Birk, A., & Carpin, S. (2006). Rescue robotics—a crucial milestone on the road to autonomous systems. *Advanced Robotics*, *20*(5), 595–605.

Boland, J. E., Tanenhaus, M. K., Garnsey, S. M., & Carlson, G. N. (1995). Verb argument structure in parsing and interpretation: Evidence from wh-questions. *Journal of Memory and Language*, *34*(6), 774–806.

Borji, A., Cheng, M.-M., Jiang, H., & Li, J. (2015). Salient object detection: A benchmark. *IEEE Transactions on Image Processing*, *24*(12), 5706–5722.

Bosker, H. R., Quené, H., Sanders, T., & de Jong, N. H. (2014). Native ‘um’s elicit prediction of low-frequency referents, but non-native ‘um’s do not. *Journal of memory and language*, *75*, 104–116.

Braga, R., Petry, M., Reis, L., & Moreira, A. (2011). Intellwheels: modular development platform for intelligent wheelchairs. *Journal of Rehabilitation Research and Development*, *48*(9), 1061–1076.

Brault, M. W. (2012). Americans with disabilities: 2010. *Current Population Reports*, *7*, 0–131.

Breazeal, C., Kidd, C., Thomaz, A. L., Hoffman, G., & Berlin, M. (2005). Effects of nonverbal communication on efficiency and robustness in human-robot teamwork. *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 708–713.

Breazeal, C., Siegel, M., Berlin, M., Gray, J., Grupen, R., Deegan, P., ... McBean, J. (2008). Mobile, dexterous, social robots for mobile manipulation and human-robot interaction. In *ACM SIGGRAPH 2008 new tech demos* (p. 27).

Breheny, R., Ferguson, H. J., & Katsos, N. (2013). Taking the epistemic step: Toward a model of on-line access to conversational implicatures. *Cognition*, *126*(3), 423–440.

Brennan, S. E., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of experimental psychology. Learning, memory, and cognition*, *22*(6), 1482–1493.

Brennan, S. E., Galati, A., & Kuhlen, A. K. (2010). Two minds, one dialog: Coordinating speaking and understanding. *Psychology of Learning and Motivation*, *53*, 301–344.

Brenton, H., Gillies, M., Ballin, D., & Chatting, D. (2005). D.: The uncanny valley: does it exist. In *Proceedings of the nineteenth british HCI group annual conference: workshop on human-animated character interaction.*

Brick, T., Schermerhorn, P., & Scheutz, M. (2007). Speech and action: Integration of action and language for mobile robots. In *Proceedings of the IEEE/RSJ international conference on intelligent robots and systems (IROS)* (pp. 1423–1428). San Diego, CA.

Briggs, G., & Scheutz, M. (2011). Facilitating mental modeling in collaborative human-robot interaction through adverbial cues. In *Proceedings of the twelfth annual meeting of the special interest group on discourse and dialogue (SIGDIAL)* (pp. 239–247).

Briggs, G., & Scheutz, M. (2012). Multi-modal belief updates in multi-robot human-robot dialogue interactions. In *Proceedings of the annual meeting of the british society for the study of artificial intelligence and the simulation of behaviour (AISB).*

Briggs, G., & Scheutz, M. (2013). A hybrid architectural approach to understanding and appropriately generating indirect speech acts. In *Proceedings of the twenty-seventh AAAI conference on artificial intelligence.*

Briggs, G., & Scheutz, M. (2014a). How robots can affect human behavior: Investigating the effects of robotic displays of protest and distress. *International Journal of Social Robotics*, *6*(3), 343–355.

Briggs, G., & Scheutz, M. (2014b). Modeling blame to avoid positive face threats in natural language generation. *Joint proceedings of the International Natural Language Generation conference (INLG) and the Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, 1.

Briggs, G., Williams, T., & Scheutz, M. (2017). Enabling robots to understand indirect speech acts in task-based interactions. *Journal of Human-Robot Interaction (JHRI).*

Britt, M. A. (1994). The interaction of referential ambiguity and argument structure in the parsing of prepositional phrases. *Journal of Memory and Language*, *33*(2), 251–283.

Brown, P. (1987). *Politeness: Some universals in language usage* (Vol. 4). Cambridge University Press.

Brown-Schmidt, S. (2009). The role of executive function in perspective taking during online language comprehension. *Psychonomic bulletin & review*, *16*(5), 893–900.

Brown-Schmidt, S. (2012). Beyond common and privileged: Gradient representations of common ground in real-time language use. *Language and Cognitive Processes*, *27*(1), 62–89.

Brown-Schmidt, S., Byron, D. K., & Tanenhaus, M. K. (2005). Beyond salience: Interpretation of personal and demonstrative pronouns. *Journal of Memory and Language*, *53*(2), 292–313.

Brown-Schmidt, S., Campana, E., & Tanenhaus, M. K. (2001). Reference resolution in the wild: On-line circumscription of referential domains in a natural, interactive, problem-solving task. In *Cognitive science.*

Brown-Schmidt, S., & Fraundorf, S. H. (2015). Interpretation of informational questions modulated by joint knowledge and intonational contours. *Journal of Memory and Language*, *84*, 49–74.

Brown-Schmidt, S., & Hanna, J. (2011). Talking in another person's shoes: Incremental perspective-taking in language processing. *Dialog and Discourse*, *2*(1), 11–33.

Brown-Schmidt, S., & Tanenhaus, M. K. (2008). *Real-time investigation of referential domains in unscripted conversation: A targeted language game approach* (Vol. 32) (No. 4).

Brown-Schmidt, S., & Tanenhaus, M. K. (2009). Addressees distinguish shared from private information when interpreting questions during interactive conversation. *Cognition*, *107*(3), 1122–1134.

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

Byron, D., Koller, A., Striegnitz, K., Cassell, J., Dale, R., Moore, J., & Oberlander, J. (2009). Report on the first NLG challenge on generating instructions in virtual environments (give). In *Proceedings of the twelfth european workshop on natural language generation (ENLG)* (pp. 165–173).

Cantrell, R., Scheutz, M., Schermerhorn, P., & Wu, X. (2010, March). Robust spoken instruction understanding for HRI. In *Proceedings of the IEEE international conference on human-robot interaction (HRI).*

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

Cassell, J. (1998). A framework for gesture generation and interpretation. *Computer vision in human-machine interaction*, 191–215.

Cassell, J., Pelachaud, C., Badler, N., Steedman, M., Achorn, B., Becket, T., ... Stone, M. (1994). Animated conversation: rule-based generation of facial expression, gesture & spoken intonation for multiple conversational agents. In *Proceedings of the twenty-first annual conference on computer graphics and interactive techniques* (pp. 413–420).

Cassell, J., Stone, M., & Yan, H. (2000). Coordination and context-dependence in the generation of embodied conversation. In *Proceedings of the first international conference on natural language generation (INLG) - volume 14* (pp. 171–178).

Castro-González, Á., Castillo, J. C., Alonso-Martín, F., Olortegui-Ortega, O. V., González-Pacheco, V., Malfaz, M., & Salichs, M. A. (2016). The effects of an impolite vs. a polite robot playing rock-paper-scissors. In *International conference on social robotics (ICSR)* (pp. 306–316).

Chai, J., Prasov, Z., & Qu, S. (2006). Cognitive principles in robust multimodal interpretation. *Journal of Artificial Intelligence Research*, *27*, 55–83.

Chai, J. Y., Hong, P., & Zhou, M. X. (2004). A probabilistic approach to reference resolution in multimodal user interfaces. In *Proceedings of the ninth international conference on intelligent user interfaces (IUI)* (pp. 70–77).

Chai, J. Y., She, L., Fang, R., Ottarson, S., Littley, C., Liu, C., & Hanson, K. (2014). Collaborative effort towards common ground in situated human-robot dialogue. In *Proceedings of the ACM/IEEE international conference on human-robot interaction (HRI)* (pp. 33–40). New York, NY, USA: ACM.

Chambers, C. G., Tanenhaus, M. K., Eberhard, K. M., Filip, H., & Carlson, G. N. (2002). Circumscribing referential domains during real-time language comprehension. *Journal of Memory and Language*, *47*(1), 30–49.

Chater, N., & Christiansen, M. H. (2008). Computational models of psycholinguistics. *Cambridge handbook of computational cognitive modeling*, 477–504.

Chen, D. L., & Mooney, R. J. (2011). Learning to interpret natural language navigation instructions from observations. In *Proceedings of the twenty-fifth AAAI conference on artificial intelligence.*

Chou, W. Y. J., Marsh, L., & Gossink, D. (2009, July). Multi-agent coordination and optimisation in the robocup rescue project. In *Proceedings of the eighteenth world IMACS/MODSIM congress.*

Chung, I., Propp, O., Walter, M. R., & Howard, T. M. (2015). On the performance of hierarchical distributed correspondence graphs for efficient symbol grounding of robot instructions. In *Proceedings of the IEEE/RSJ international conference on intelligent robots and systems (IROS)* (pp. 5247–5252).

Church, K., & Patil, R. (1982). Coping with syntactic ambiguity or how to put the block in the box on the table. *Computational Linguistics*, *8*(3-4), 139–149.

Clark, H. H. (1996). *Using language.* Cambridge university press.

Clark, H. H., & Brennan, S. E. (1991). Grounding in communication. *Perspectives on socially shared cognition*, *13*(1991), 127–149.

Clark, H. H., & Marshall, C. R. (2002). Definite reference and mutual knowledge. *Psycholinguistics: critical concepts in psychology*, *414*.

Clark, H. H., & Schaefer, E. F. (1987). Collaborating on contributions to conversations. *Language and cognitive processes*, *2*(1), 19–41.

Clark, H. H., & Schaefer, E. F. (1989). Contributing to discourse. *Cognitive science*, *13*(2), 259–294.

Clark, H. H., Schreuder, R., & Buttrick, S. (1983). Common ground at the understanding of demonstrative reference. *Journal of verbal learning and verbal behavior*, *22*(2), 245–258.

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, *22*(1), 1–39.

Clark, J. A., & Roemer, R. B. (1977). Voice controlled wheelchair. *Archives of Physical Medicine and Rehabilitation*, *58*(4), 169–175.

Clark, M. J., & Liggins, D. (2012). Recent work on grounding. *Analysis*, ans086.

Clark, S., & Curran, J. (2007). Wide-coverage efficient statistical parsing with CCG and log-linear models. *Comp. Linguistics*, *33*(4).

Clocksin, W., & Mellish, C. S. (2003). *Programming in PROLOG.* Springer Science & Business Media.

Cohen, P., McGee, D., Oviatt, S., Wu, L., Clow, J., King, R., … Rosenblum, L. (1999). Multimodal interaction for 2d and 3d environments. *IEEE Computer Graphics and Applications*, *19*(4), 10–13.

Cohen, P. R., & Levesque, H. J. (1985). Speech acts and rationality. In *Proceedings of the twenty-third annual meeting of the association for computational linguistics (ACL)* (pp. 49–60).

Cohen, P. R., & Perrault, C. R. (1979). Elements of a plan-based theory of speech acts. *Cognitive science*, *3*(3), 177–212.

Cooper, R. (2005). Records and record types in semantic theory. *Journal of Logic and Computation*, *15*(2), 99–112.

Cooper, R. A. (2010). *Rehabilitation engineering applied to mobility and manipulation*. Taylor & Francis.

Cowan, N. (1998). *Attention and memory: An integrated framework*. Oxford University Press.

Crain, S., & Steedman, M. (1985). On not being led up the garden path: The use of context by the psychological parser. *Natural language parsing*, 320–358.

Crandall, J. W., Goodrich, M. A., Olsen, D. R., & Nielsen, C. W. (2005). Validating human-robot interaction schemes in multitasking environments. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, *35*(4), 438–449.

Dahan, D., & Tanenhaus, M. K. (2004). Continuous mapping from sound to meaning in spoken-language comprehension: immediate effects of verb-based thematic constraints. *Journal of experimental psychology. Learning, memory, and cognition*, *30*(2), 498–513.

Dahan, D., Tanenhaus, M. K., & Chambers, C. G. (2002). Accent and reference resolution in spoken-language comprehension. *Journal of Memory and Language*, *47*(2), 292–314.

Dahlbäck, N., Jönsson, A., & Ahrenberg, L. (1993). Wizard of oz studies: why and how. In *Proceedings of the first international conference on intelligent user interfaces (IUI)* (pp. 193–200).

Dale, R. (1989). Cooking up referring expressions. In *Proceedings of the twenty-seventh annual meeting of the association for computational linguistics (ACL)* (pp. 68–75).

Dale, R., & Reiter, E. (1995). Computational interpretations of the gricean maxims in the generation of referring expressions. *Cognitive science*, *19*(2), 233–263.

Dautenhahn, K. (2007). Socially intelligent robots: dimensions of human–robot interaction. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1480), 679–704.

Dautenhahn, K., Woods, S., Kaouri, C., Walters, M. L., Koay, K. L., & Werry, I. (2005). What is a robot companion - friend, assistant or butler? In *Proceedings of the IEEE/RSJ international conference on intelligent robots and systems (iros)* (pp. 1192–1197).

Davids, A. (2002). Urban search and rescue robots: from tragedy to technology. *IEEE Intelligent Systems*, *17*(2), 81–83.

Davis, W. (2011). Implicature. In E. Zalta (Ed.), *The Stanford encyclopedia of philosophy (spring 2013 edition).*

Degen, J., & Tanenhaus, M. K. (2015). Processing scalar implicature: A constraint-based approach. *Cognitive science*, *39*(4), 667–710.

Deits, R., Tellex, S., Kollar, T., & Roy, N. (2013). Clarifying commands with information-theoretic human-robot dialog. *Journal of Human-Robot Interaction (JHRI).*

de Marneffe, M.-C., Manning, C. D., & Potts, C. (2010). Was it good? it was provocative. learning the meaning of scalar adjectives. In *Proceedings of the forty-eighth annual meeting of the association for computational linguistics (ACL)* (pp. 167–176).

DeVault, D., Kariaeva, N., Kothari, A., Oved, I., & Stone, M. (2005). An information-state approach to collaborative reference. In *Proceedings of the ACL 2005 on interactive poster and demonstration sessions* (pp. 1–4).

DeVault, D., Sagae, K., & Traum, D. (2009). Can I finish?: learning when to respond to incremental interpretation results in interactive dialogue. In *Proceedings of the tenth annual meeting of the special interest group on discourse and dialogue (SIGDIAL)* (pp. 11–20).

DeVault, D., Sagae, K., & Traum, D. R. (2011). Incremental interpretation and prediction of utterance meaning for interactive dialogue. *Dialogue and Discourse*, *2*(1), 143–170.

DeVault, D., & Stone, M. (2007). Managing ambiguities across utterances in dialogue. In *Proceedings of the 11th SEMDIAL workshop on the semantics and pragmatics of dialogue (DECALOG)* (pp. 49–56).

DeVault, D., & Stone, M. (2009). Learning to interpret utterances using dialogue history. In *Proceedings of the twelfth conference of the european chapter of the association for computational linguistics (EACL)* (pp. 184–192).

DiSalvo, C. F., Gemperle, F., Forlizzi, J., & Kiesler, S. (2002). All robots are not created equal: the design and perception of humanoid robot heads. In *Proceedings of the fourth conference on designing interactive systems: Processes, practices, methods, and techniques* (pp. 321–326).

Duvallet, F., Walter, M. R., Howard, T., Hemachandra, S., Oh, J., Teller, S., ... Stentz, A. (2014). Inferring maps and behaviors from natural language instructions. In *Proceedings of the international symposium on experimental robotics (ISER)* (pp. 373–388).

Dzifcak, J., Scheutz, M., Baral, C., & Schermerhorn, P. (2009, May). What to do and how to do it: Translating natural language directives into temporal and dynamic logic representation for goal management and action execution. In *Proceedings of the international conference on robotics and automation (ICRA).* Kobe, Japan.

Eberhard, K. M., Spivey-Knowlton, M. J., Sedivy, J. C., & Tanenhaus, M. K. (1995). Eye movements as a window into real-time spoken language comprehension in natural contexts. *Journal of Psycholinguistic Research*, *24*(6), 409–436.

Elman, J. L. (1990). Finding structure in time. *Cognitive science*, *14*(2), 179–211.

Elman, J. L., Hare, M., & McRae, K. (2004). Cues, constraints, and competition in sentence processing. *Beyond nature-nurture: Essays in honor of Elizabeth Bates*, 111–138.

Engonopoulos, N., Villalba, M., Titov, I., & Koller, A. (2013). Predicting the resolution of referring expressions from user behavior. In *Proceedings of empirical methods in natural language processing (EMNLP)* (pp. 1354–1359).

Eyssel, F., Kuchenbrandt, D., Bobinger, S., de Ruiter, L., & Hegel, F. (2012). 'if you sound like me, you must be more human': on the interplay of robot and user features on human-robot acceptance and anthropomorphism. In *Proceedings of the seventh annual ACM/IEEE international conference on human-robot interaction (HRI)* (pp. 125–126).

Fang, R., Doering, M., & Chai, J. Y. (2014). Collaborative models for referring expression generation in situated dialogue. In *Proceedings of the twenty-eighth AAAI conference on artificial intelligence* (pp. 1544–1550).

Fang, R., Liu, C., & Chai, J. (2012). Integrating word acquisition and referential grounding towards physical world interaction. In *Proceedings of the fourteenth ACM international conference on multimodal interfaces (ICMI)* (pp. 109–116).

Fang, R., Liu, C., She, L., & Chai, J. Y. (2013). Towards situated dialogue: Revisiting referring expression generation. In *Proceedings of empirical methods in natural language processing (EMNLP)* (pp. 392–402).

Faria, B. M., Reis, L. P., & Lau, N. (2015). A methodology for creating an adapted command language for driving an intelligent wheelchair. *Journal of Intelligent Robot Systems*, 609–623.

Fasola, J., & Matarić, M. J. (2013). Using semantic fields to model dynamic spatial relations in a robot architecture for natural language instruction of service robots. In *Proceedings of the IEEE/RSJ international conference on intelligent robots and systems (IROS)* (pp. 143–150).

Fasola, J., & Matarić, M. J. (2014). Interpreting instruction sequences in spatial language discourse with pragmatics towards natural human-robot interaction. In *Proceedings of the IEEE international conference on robotics and automation (ICRA)* (pp. 2720–2727).

Fehr, L., Langbein, W. E., & Skaar, S. B. (2000). Adequacy of power wheelchair control i terfaces for persons with severe disabilities: A clinical survey. *Development*, *37*(3), 353–360.

Ferguson, G., & Allen, J. F. (1998). Trips: An integrated intelligent problem-solving assistant. In *Proceedings of the second AAAI conference on artificial intelligence and the conference on innovative applications of artificial intelligence (IAAI)* (pp. 567–572).

Ferguson, H. J., & Breheny, R. (2012). Listeners' eyes show spontaneous sensitivity to perspective. *Journal of Experimental Social Psychology*, *48*(1), 1–25.

Fong, T., Nourbakhsh, I., & Dautenhahn, K. (2003). A survey of socially interactive robots. *Robotics and Autonomous Systems*, *42*(3), 143–166.

Fong, T., Thorpe, C., & Baur, C. (2001). Collaboration, dialogue, and human-robot interaction. *International Symposium of Robotics Research*(November).

Foster, M. E., Gaschler, A., Giuliani, M., Isard, A., Pateraki, M., & Petrick, R. P. A. (2012). Two people walk into a bar : Dynamic multi-party social interaction with a robot agent. In *Proceedings of the twelfth international conference on multimodal interaction (ICMI)* (pp. 14–21).

Foster, M. E., Giuliani, M., & Isard, A. (2014). Task-based evaluation of context-sensitive referring expressions in human–robot dialogue. *Language, Cognition and Neuroscience*, *29*(8), 1018–1034.

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998–998.

Fraune, M., & Šabanovic, S. (2014). Robot gossip : Effects of mode of robot communication on human perceptions of robots. In *Proceeding of the ACM/IEEE international conference on human-robot interaction (HRI)* (pp. 160–161).

Fraune, M. R., & Šabanović, S. (2014). Negative attitudes toward minimalistic robots with intragroup communication styles. In *Proceedings of the twenty-third IEEE international symposium on robot and human interactive communication (RO-MAN)* (Vol. 47408, pp. 1116–1121).

Frazier, L. (1987). *Theories of sentence processing.* The MIT Press.

Fukuda, T., & Sekiyama, K. (1994). Hierarchical prediction model for intelligent communication in multiple robotic systems. *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, *2*.

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

Garrod, S., & Pickering, M. J. (2009). Joint action, interactive alignment, and dialog. *Topics in Cognitive Science*, *1*(2), 292–304.

Gatt, A., van der Sluis, I., & van Deemter, K. (2007). Evaluating algorithms for the generation of referring expressions using a balanced corpus. In *Proceedings of the eleventh european workshop on natural language generation (ENLG)* (pp. 49–56). Stroudsburg, PA, USA: Association for Computational Linguistics.

Gatt, A., van Gompel, R. P., van Deemter, K., & Kramer, E. (2013). Are we Bayesian referring expression generators. In *Proceedings of the thirty-fifth annual meeting of the cognitive science society.*

Gerkey, B., Vaughan, R. T., & Howard, A. (2003). The player/stage project: Tools for multi-robot and distributed sensor systems. In *Proceedings of the eleventh international conference on advanced robotics (ICAR)* (pp. 317–323).

Giesler, B., Salb, T., Steinhaus, P., & Dillmann, R. (2004). Using augmented reality to interact with an autonomous mobile platform. In *Proceedings of the IEEE international conference on robotics and automation (ICRA)* (Vol. 1, pp. 1009–1014).

Gombolay, M., & Shah, A. (2016). Appraisal of statistical practices in hri vis-á-vis the t-test for likert items/scales. In *Proceedings of the AAAI fall symposium on artificial intelligence for human-robot interaction (AI-HRI).*

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, *20*(11), 818–829.

Goodman, N. D., & Stuhlmüller, A. (2013). Knowledge and implicature: Modeling language understanding as social cognition. *Topics in cognitive science*, *5*(1), 173–184.

Goodrich, M. A., Cooper, J. L., Adams, J. A., Humphrey, C., Zeeman, R., & Buss, B. G. (2007). Using a mini-uav to support wilderness search and rescue: Practices for human-robot teaming. In *Proceedings of the IEEE international workshop on safety, security and rescue robotics* (pp. 1–6).

Goodrich, M. A., Lin, L., Morse, B. S., & Roscheck, M. (2010). Supporting wilderness search and rescue with integrated intelligence: autonomy and information at the right time and the right place. In *Proceedings of the twenty-fourth AAAI conference on artificial intelligence.*

Goodrich, M. A., Morse, B. S., Gerhardt, D., Cooper, J. L., Quigley, M., Adams, J. A., & Humphrey, C. (2008). Supporting wilderness search and rescue using a camera-equipped mini uav. *Journal of Field Robotics*, *25*(1-2), 89–110.

Goodrich, M. A., & Schultz, A. C. (2007). Human-robot interaction: a survey. *Foundations and trends in human-computer interaction*, *1*(3), 203–275.

Gorniak, P., & Roy, D. (2004). Grounded semantic composition for visual scenes. *Journal of Artificial Intelligence Research*, *21*, 429–470.

Green, G. M. (1996). *Pragmatics and natural language understanding.* Psychology Press.

Green, N., & Carberry, S. (1999). Interpreting and generating indirect answers. *Computational Linguistics*, *25*(3), 389–435.

Green, S., Billinghurst, M., Chen, X., & Chase, G. (2008). Human-robot collaboration: A literature review and augmented reality approach in design. *International Journal of Advanced Robotic Systems*.

Green, S. A., Billinghurst, M., Chen, X., & Chase, J. G. (2007). Human robot collaboration: An augmented reality approach—a literature review and analysis. In *Proceedings of the ASME international design engineering technical conferences and computers and information in engineering conference* (pp. 117–126).

Green, S. A., Chase, J. G., Chen, X., & Billinghurst, M. (2009). Evaluating the augmented reality human-robot collaboration system. *International journal of intelligent systems technologies and applications*, *8*(1-4), 130–143.

Grice, H. P. (1957). Meaning. *The philosophical review*, 377–388.

Grice, H. P. (1968). Utterer's meaning, sentence-meaning, and word-meaning. In *Philosophy, language, and artificial intelligence* (pp. 49–66). Springer.

Grice, H. P. (1969). Utterer's meaning and intention. *The philosophical review*, *78*(2), 147–177.

Grice, H. P. (1970). Logic and conversation. *Syntax and semantics*, *3*, 41–58.

Grodner, D., Dalini, M., Pearlstein-Levy, S., & Ward, A. (2012). Factors that contribute to the use of perspective in referent identification. In *Proceedings of the twenty-fifth annual CUNY conference on human sentence processing.*

Groenendijk, J., & Stokhof, M. (1991). Dynamic predicate logic. *Linguistics and philosophy*, *14*(1), 39–100.

Grosz, B. J. (1977). *The representation and use of focus in dialogue understanding.* (Unpublished doctoral dissertation). University of California, Berkeley.

Grosz, B. J., & Sidner, C. L. (1988). *Plans for discourse* (Tech. Rep.). Cambridge, MA: DTIC Document.

Grosz, B. J., & Sidner, C. L. (1990). Plans for discourse. In P. R. Cohen, J. L. Morgan, & M. Pollack (Eds.), *Intentions in communication* (pp. 417–444). Cambridge, MA: MIT Press.

Grosz, B. J., Weinstein, S., & Joshi, A. K. (1995). Centering: A framework for modeling the local coherence of discourse. *Computational linguistics*, *21*(2), 203–225.

Gulzar, K., & Kyrki, V. (2015). See what i mean-probabilistic optimization of robot pointing gestures. In *Proceedings of the fifteenth IEEE/RAS international conference on humanoid robots (Humanoids)* (pp. 953–958).

Gundel, J., Hedberg, N., & Zacharski, R. (1988). On the generation and interpretation of demonstrative expressions. In *Proceedings of the 12th annual conference on computational linguistics (COLING)* (pp. 216–221).

Gundel, J. K. (2010). Reference and accessibility from a givenness hierarchy perspective. *International Review of Pragmatics*, *2*(2), 148–168.

Gundel, J. K., Bassene, M., Gordon, B., Humnick, L., & Khalfaoui, A. (2010). Testing predictions of the givenness hierarchy framework: A crosslinguistic investigation. *Journal of Pragmatics*, *42*(7), 1770–1785.

Gundel, J. K., Hedberg, N., & Zacharski, R. (1993). Cognitive status and the form of referring expressions in discourse. *Language*, 274–307.

Gundel, J. K., Hedberg, N., Zacharski, R., Mulkern, A., Custis, T., Swierzbin, B., ... Watters, S. (2006, May). *Coding protocol for statuses on the givenness hierarchy.* (unpublished manuscript)

Gundel, J. K., Ntelitheos, D., & Kowalsky, M. (2006). Children's use of referring expressions: Some implications for theory of mind. *ZAS Papers in Linguistics*, *48*, 1–22.

Hancock, P. a., Billings, D. R., & Schaefer, K. E. (2011, September). Can you trust your robot? *Ergonomics in Design: The Quarterly of Human Factors Applications*, *19*(3), 24–29.

Hanna, J. E., & Tanenhaus, M. K. (2004). Pragmatic effects on reference resolution in a collaborative task: Evidence from eye movements. *Cognitive Science*, *28*(1), 105–115.

Hanna, J. E., Tanenhaus, M. K., & Trueswell, J. C. (2003). The effects of common ground and perspective on domains of referential interpretation. *Journal of Memory and Language*, *49*(1), 43–61.

Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, *42*(1-3), 335–346.

Hawes, N., Hanheide, M., Hargreaves, J., Page, B., Zender, H., & Jensfelt, P. (2011, may). Home alone: Autonomous extension and correction of spatial representations. In *Proceedings of the IEEE international conference on robotics and automation (ICRA)* (p. 3907 -3914).

Hayashi, K., Kanda, T., Miyashita, T., Ishiguro, H., & Hagita, N. (2008). Robot manzai: Robot conversation as a passive–social medium. *International Journal of Humanoid Robotics*, *5*(01), 67–86.

Heendeni, J. N., Premaratne, K., Murthi, M., Uscinski, J., & Scheutz, M. (2016). A generalization of Bayesian inference in the Dempster-Shafer belief theoretic framework. In *Proceedings of the intenational conference on information fusion.*

Heller, D., Grodner, D., & Tanenhaus, M. K. (2008). The role of perspective in identifying domains of reference. *Cognition*, *108*(3), 831–836.

Heller, D., Parisien, C., & Stevenson, S. (2016). Perspective-taking behavior as the probabilistic weighing of multiple domains. *Cognition*, *149*, 104–120.

Hemachandra, S., Kollar, T., Roy, N., & Teller, S. (2011). Following and interpreting narrated guided tours. In *Proceedings of the IEEE international conference on robotics and automation (ICRA).* Shanghai, China.

Hemachandra, S., Walter, M. R., & Teller, S. (2014). Information theoretic question asking to improve spatial semantic representations. In *Proceedings of the AAAI fall symposium series.*

Hemachandra, S. M. (2010). *Narrated guided tour following and interpretation by an autonomous wheelchair* (Unpublished master's thesis). Massachusetts Institute of Technology.

Herzig, A., & Longin, D. (2002). A logic of intention with cooperation principles and with assertive speech acts as communication primitives. In *Proceedings of the first international joint conference on autonomous agents and multiagent systems (AAMAS)* (pp. 920–927).

Hinds, P. J., Roberts, T. L., & Jones, H. (2004, June). Whose job is it anyway? a study of human-robot interaction in a collaborative task. *Human-Computer Interaction*, *19*(1), 151–181.

Hinkelman, E. A., & Allen, J. F. (1989). Two constraints on speech act ambiguity. In *Proceedings of the twenty-seventh annual meeting of the association for computational linguistics (ACL)* (pp. 212–219).

Hockey, B. A., & Miller, D. P. (2007). A demonstration of a converstionally guided smart wheelchair. In *Proceedings of the SIGACCESS conference on computers and accessibility (ASSETS)* (p. 243-244).

Hoffman, J., Guadarrama, S., Tzeng, E., Hu, R., Donahue, J., Girshick, R., ... Saenko, K. (2014). LSDA: Large scale detection through adaptation. In *Proceedings of neural information processing systems (NIPS).*

Holladay, R. M., Dragan, A. D., & Srinivasa, S. S. (2014). Legible robot pointing. In *Proceedings of the twenty-third IEEE international symposium on robot and human interactive communication (RO-MAN)* (pp. 217–223).

Höllerer, T., & Turk, M. (2006). Multimodal interaction with a wearable augmented reality system. *IEEE Computer Graphics and Applications*, *26*(3), 62–71.

Holz, T., Dragone, M., & O'Hare, G. M. (2009). Where robots and virtual agents meet. *International Journal of Social Robotics*, *1*(1), 83–93.

Horacek, H. (2005). Generating referential descriptions under conditions of uncertainty. In *Proceedings of the tenth european workshop on natural language generation (ENLG)* (pp. 58–67).

Horton, W. S., & Keysar, B. (1996). When do speakers take into account common ground? *Cognition*, *59*(1), 91–117.

Hough, J., Kennington, C., Schlangen, D., & Ginzburg, J. (2015). Incremental semantics for dialogue processing: Requirements and a comparison of two approaches. In *Proceedings of the 11th international conference on computational semantics (IWCS).*

Huang, A. S., Olson, E., & Moore, D. C. (2010). Lcm: Lightweight communications and marshalling. In *Proceedings of the IEEE/RSJ international conference on intelligent robots and systems (IROS)* (pp. 4057–4062).

Huang, C.-M., & Mutlu, B. (2014). Learning-based modeling of multimodal behaviors for humanlike robots. In *Proceedings of the ACM/IEEE international conference on human-robot interaction (HRI)* (pp. 57–64).

Huang, Y. T., & Snedeker, J. (2011). Logic and conversation revisited: Evidence for a division between semantic and pragmatic content in real-time language comprehension. *Language and Cognitive Processes*, *26*(8), 1161–1172.

Huettig, F., Olivers, C. N. L., & Hartsuiker, R. J. (2011). Looking, language, and memory: Bridging research from the visual world and visual search paradigms. *Acta Psychologica*, *137*(2), 138–150.

Huettig, F., Rommers, J., & Meyer, A. S. (2011). Using the visual world paradigm to study language processing: A review and critical evaluation. *Acta Psychologica*, *137*(2), 151–171.

Iezzoni, L. I., McCarthy, E. P., Davis, R. B., & Siebens, H. (2001). Mobility difficulties are not only a problem of old age. *Journal of General Internal Medicine*, *16*(4), 235-43.

Irie, K., Yoshida, T., & Tomono, M. (2012). Outdoor localization using stereo vision under various illumination conditions. *Advanced Robotics*, *26*(3-4), 327–348.

Jennings, N. R. (2000). On agent-based software engineering. *Artificial intelligence*, *117*(2), 277–296.

Jing, G., Finucane, C., Raman, V., & Kress-Gazit, H. (2012). Correct high-level robot control from structured english. In *Proceedings of the IEEE international conference on robotics and automation (ICRA)* (pp. 3543–3544).

Johnson, C., & Kuipers, B. (2012). Efficient search for correct and useful topological maps. In *Proceedings of the IEEE international conference on intelligent robots and systems (IROS)* (pp. 5277–5282).

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness* (No. 6). Harvard University Press.

Kaiser, E., Olwal, A., McGee, D., Benko, H., Corradini, A., Li, X., ... Feiner, S. (2003). Mutual disambiguation of 3d multimodal interaction in augmented and virtual reality. In *Proceedings of the fifth international conference on multimodal interfaces (ICMI)* (pp. 12–19).

Kamide, Y., Altmann, G., & Haywood, S. L. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and Language*, *49*(1), 133–156.

Kanda, T., Ishiguro, H., Ono, T., Imai, M., & Mase, K. (2002). Multi-robot cooperation for human-robot communication. In *Proceedings of the eleventh IEEE international workshop on robot and human interactive communication (RO-MAN)* (pp. 271–276).

Kanda, T., Ishiguro, H., Ono, T., Imai, M., & Nakatsu, R. (2004, May). Effects of observation of robot-robot communication on human-robot communication. *Electronics and Communications in Japan (Part III: Fundamental Electronic Science)*, *87*(5), 48–58.

Kehler, A. (2000). Cognitive status and form of reference in multimodal human-computer interaction. In *Proceedings of the fourteenth AAAI conference on artificial intelligence* (pp. 685–690).

Kemmerer, D. (1999). "Near" and "far" in language and perception. *Cognition*, *73*(1), 35–63.

Kempson, R., Meyer-Viol, W., & Gabbay, D. M. (2000). *Dynamic syntax: The flow of language understanding.* Wiley-Blackwell.

Kennedy, J., Baxter, P., & Belpaeme, T. (2014). Children comply with a robot's indirect requests. In *Proceedings of the ACM/IEEE international conference on human-robot interaction (HRI)* (pp. 198–199).

Kennington, C., Iida, R., Tokunaga, T., & Schlangen, D. (2015). Incrementally tracking reference in human/human dialogue using linguistic and extra-linguistic information. In *Proceedings of the conference of the north american chapter of the association for computational linguistics–human language technologies (NAACL HLT).*

Kennington, C., Kousidis, S., & Schlangen, D. (2013). Interpreting situated dialogue utterances: an update model that uses speech, gaze, and gesture information. *Proceedings of the twelfth Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL).*

Kennington, C., & Schlangen, D. (2012). Markov logic networks for situated incremental natural language understanding. In *Proceedings of the 13th annual meeting of the special interest group on discourse and dialogue (SIGDIAL)* (pp. 314–323).

Kennington, C., & Schlangen, D. (2017). A simple generative model of incremental reference resolution for situated dialogue. *Computer Speech & Language*, *41*, 43–67.

Keysar, B. (2007). Communication and miscommunication: The role of egocentric processes. *Intercultural Pragmatics*, *4*(1), 71–84.

Keysar, B., Barr, D. J., Balin, J. A., & Brauner, J. S. (2000). Taking perspective in conversation: The role of mutual knowledge in comprehension. *Psychological Science*, *11*(1), 32–38.

Keysar, B., Barr, D. J., Balin, J. A., & Paek, T. S. (1998). Definite reference and mutual knowledge: Process models of common ground in comprehension. *Journal of Memory and Language*, *39*(1), 1–20.

Keysar, B., Lin, S., & Barr, D. J. (2003). Limits on theory of mind use in adults. *Cognition*, *89*(1), 25–41.

Kiesler, S. (2005). Fostering common ground in human-robot interaction. In *Proceedings of the IEEE international workshop on robot and human interactive communication (RO-MAN)* (pp. 729–734).

Klyne, G., & Carroll, J. J. (2006). *Resource description framework (RDF): Concepts and abstract syntax.*

Knoeferle, P., & Crocker, M. W. (2006). The coordinated interplay of scene, utterance, and world knowledge: evidence from eye tracking. *Cognitive science*, *30*, 481–529.

Kohler, D., & Conley, K. (2011). *ROSJAVA–an implementation of ROS in pure Java with Android support.*

Kollar, T., Tellex, S., Roy, D., & Roy, N. (2010). Toward understanding natural language directions. In *Proceeding of the fifth ACM/IEEE international conference on human-robot interaction (HRI)* (pp. 259–266). New York, NY, USA: ACM.

Koller, A., Striegnitz, K., Gargett, A., Byron, D., Cassell, J., Dale, R., ... Oberlander, J. (2010). Report on the second nlg challenge on generating instructions in virtual environments (give-2). In *Proceedings of the sixth international natural language generation conference (INLG)* (pp. 243–250).

Konieczny, S. (2000). On the difference between merging knowledge bases and combining them. In *Proceedings of the international conference on knowledge representation (KR)* (pp. 135–144).

Krahmer, E., & Van Deemter, K. (2012). Computational generation of referring expressions: A survey. *Computational Linguistics*, *38*(1), 173–218.

Kramer, J., & Scheutz, M. (2006). Ade: A framework for robust complex robotic architectures. In *Proceedings of the IEEE/RSJ international conference on intelligent robots and systems (IROS)* (pp. 4576–4581).

Kramer, J., & Scheutz, M. (2007). Development environments for autonomous mobile robots: A survey. *Autonomous Robots*, *22*(2), 101–132.

Kranstedt, A., & Wachsmuth, I. (2005). Incremental generation of multimodal deixis referring to objects. In *Proceedings of the tenth european workshop on natural language generation (ENLG).*

Krause, E., Zillich, M., Williams, T., & Scheutz, M. (2014). Learning to recognize novel objects in one shot through human-robot interactions in natural language dialogues. In *Proceedings of the twenty-eighth AAAI conference on artificial intelligence.*

Kreysa, H., & Knoeferle, P. (2013). Reference-related speaker gaze as a cue in online sentence processing. In *Kognitive systeme: Mensch, teams, systeme und automaten.*

Kruijff, G.-J. M., Brenner, M., & Hawes, N. (2008). Continual planning for cross-modal situated clarification in human-robot interaction. In *Proceedings of the seventeenth IEEE international symposium on robot and human interactive communication (RO-MAN)* (pp. 592–597).

Kruijff, G.-J. M., Lison, P., Benjamin, T., Jacobsson, H., & Hawes, N. (2007). Incremental, multi-level processing for comprehending situated dialogue in human-robot interaction. In *Proceedings of the symposium on language and robots.*

Kruijff, G.-J. M., Lison, P., Benjamin, T., Jacobsson, H., Zender, H., Kruijff-Korbayová, I., & Hawes, N. (2010). Situated dialogue processing for human-robot interaction. In *Cognitive systems* (pp. 311–364). Springer.

Kuhlen, A. K., & Brennan, S. E. (2013). Language in dialogue: when confederates might be hazardous to your data. *Psychonomic bulletin & review*, *20*(1), 54–72.

Kuipers, B. (2000). The spatial semantic hierarchy. *Artificial Intelligence*, *119*, 191–233.

Kuipers, B. (2008). An intellectual history of the spatial semantic hierarchy. In M. Jefferies & W.-K. Yeap (Eds.), *Robotics and cognitive approaches to spatial mapping* (Vol. 38, p. 243-264). Springer Berlin / Heidelberg.

Kumar, V. (1992). Algorithms for constraint-satisfaction problems: A survey. *AI magazine*, *13*(1), 32.

Laird, J. E. (2012). *The soar cognitive architecture.* MIT Press.

Laird, J. E., & Rosenbloom, P. (1996). The evolution of the soar cognitive architecture. *Mind matters: A tribute to Allen Newell*, 1–50.

Lakoff, R. T. (1973). *The logic of politeness: Minding your p's and q's.*

Lallée, S., Pattacini, U., Lemaignan, S., Lenz, A., Melhuish, C., Natale, L., ... Sisbot, E. A. (2012). Towards a platform-independent cooperative human robot interaction system: Iii an architecture for learning and executing actions and shared plans. *Autonomous Mental Development, IEEE Transactions on*, *4*(3), 239–253.

Langley, P., Laird, J. E., & Rogers, S. (2009). Cognitive architectures: Research issues and challenges. *Cognitive Systems Research*, *10*(2), 141–160.

Laviola Jr, J. J. (1999). *Whole-hand and speech input in virtual environments* (Unpublished master's thesis). Brown.

Lemaignan, S., Ros, R., Alami, R., & Beetz, M. (2011). What are you talking about? grounding dialogue in a perspective-aware robotic architecture. In *Proceedings of the twentieth IEEE international symposium on robot and human interactive communication (RO-MAN)* (pp. 107–112).

Lemaignan, S., Warnier, M., Sisbot, A. E., & Alami, R. (2014). Human-robot interaction: Tackling the AI challenges. *Artificial Intelligence*.

Li, S., & Jia, Y. (2010). A multimodal labeling interface for wearable computing. In *Proceedings of the fifteenth international conference on intelligent user interfaces (IUI)* (pp. 345–348).

Liberatore, P., & Schaerf, M. (1998). Arbitration (or how to merge knowledge bases). *IEEE Transactions on Knowledge and Data Engineering*, *10*(1), 76–90.

Lin, J. (1996). Integration of weighted knowledge bases. *Artificial Intelligence*, *83*(2), 363–378.

Linh, L. H., Hai, N. T., Van Thuyen, N., Mai, T. T., & Van Toi, V. (2015). MFCC-DTW algorithm for speech recognition in an intelligent wheelchair. In *Proceedings of the fifth international conference on biomedical engineering in vietnam* (pp. 417–421). Cham: Springer International Publishing.

Litman, D. J., & Allen, J. F. (1987). A plan recognition model for subdialogues in conversations. *Cognitive science*, *11*(2), 163–200.

Liu, C., Fang, R., She, L., & Chai, J. (2013). Modeling collaborative referring for situated referential grounding. In *Proceedings of the annual meeting of the special interest group on discourse and dialogue (SIGDIAL)* (pp. 78–86).

Liu, C., Ishi, C. T., Ishiguro, H., & Hagita, N. (2012). Generation of nodding, head tilting and eye gazing for human-robot dialogue interaction. In *Proceedings of the seventh ACM/IEEE international conference on human-robot interaction (HRI)* (pp. 285–292).

Liu, J., Zhang, H., Fan, B., Wang, G., & Wu, J. (2010). A novel economical embedded multi-mode intelligent control system for powered wheelchair. In *Proceedings of the international conference on computing, control and industrial engineering (CCIE)* (Vol. 1, pp. 156–159).

Liu, P., Glas, D. F., Kanda, T., Ishiguro, H., & Hagita, N. (2016). A model for generating socially-appropriate deictic behaviors towards people. *International Journal of Social Robotics*, 1–17.

Lochbaum, K. E., Grosz, B. J., & Sidner, C. L. (1990). Models of plans to support communication: An initial report. In *Proceedings of the fourth AAAI conference on artificial intelligence* (pp. 485–490).

Louwerse, M. M., & Bangerter, A. (2010). Effects of ambiguous gestures and language on the time course of reference resolution. *Cognitive science*, *34*(8), 1517–1529.

MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological review*, *101*(4), 676.

MacDorman, K. F. (2006). Subjective ratings of robot video clips for human likeness, familiarity, and eeriness: An exploration of the uncanny valley. In *Proeedings of the ICCS/CogSci long symposium: Toward social mechanisms of android science* (pp. 26–29).

MacMahon, M., Stankiewicz, B., & Kuipers, B. (2006). Walk the talk: Connecting language, knowledge, and action in route instructions. In *Proceedings of the twenty-first AAAI conference on artificial intelligence* (pp. 1475–1482). AAAI Press.

Marconi, L., Melchiorri, C., Beetz, M., Pangercic, D., Siegwart, R., Leutenegger, S., ... others (2012). The sherpa project: Smart collaboration between humans and ground-aerial robots for improving rescuing activities in alpine environments. In *Proceedings of the IEEE international symposium on safety, security, and rescue robotics (SSRR)* (pp. 1–4).

Marcus, G. F., & Davis, E. (2013). How robust are probabilistic models of higher-level cognition? *Psychological science*, *24*(12), 2351–2360.

Marcus, G. F., & Davis, E. (2015). Still searching for principles a response to goodman et al.(2015). *Psychological science*.

Marge, M., & Rudnicky, A. I. (2015). Miscommunication recovery in physically situated dialogue. In *Proceedings of the sixteenth annual meeting of the special interest group on discourse and dialogue (SIGDIAL)* (p. 22).

Marr, D. (1982). *Vision: A computational approach.* Freeman & Co., San Francisco.

Maskeliunas, R., & Simutis, R. (2011). Multimodal wheelchair control for the paralyzed people. *Elektronika ir Elektrotechnika*, 81-84.

Matsuyama, Y., Taniyama, H., Fujie, S., & Kobayashi, T. (2006). Framework of communication activation robot participating in multiparty conversation. In *Proceedings of the AAAI fall symposium* (pp. 68–73).

Matuszek, C., Fitzgerald, N., Zettlemoyer, L., Bo, L., & Fox, D. (2012). A joint model of language and perception for grounded attribute learning. In *Proceedings of the twenty-ninth international conference on machine learning (ICML)*.

Matuszek, C., Fox, D., & Koscher, K. (2010). Following directions using statistical machine translation. In *Proceeding of the fifth ACM/IEEE international conference on human-robot interaction (HRI)* (pp. 251–258). New York, NY, USA: ACM.

Matuszek, C., Herbst, E., Zettlemoyer, L., & Fox, D. (2012). Learning to parse natural language commands to a robot control system. In *Proceedings of the thirteenth international symposium on experimental robotics (ISER)* (pp. 403–415).

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

May, W. G. (1973). *Mountain search and rescue techniques.* Rocky Mountain Rescue Group.

McCarthy, J. F., & Lehnert, W. G. (1995). Using decision trees for coreference resolution. *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence (IJCAI)*.

McGuinness, D. L., & Van Harmelen, F. (2004). Owl web ontology language overview. *W3C recommendation*, *10*(10), 2004.

McMurrough, C., Ranatunga, I., Papangelis, A., Popa, D. O., & Makedon, F. (2013). A development and evaluation platform for non-tactile power wheelchair controls. In *Proceedings of the sixth international conference on pervasive technologies related to assistive environments (PETRA)* (p. 4).

McRae, K., & Matsuki, K. (2013). Constraint-based models of sentence processing. In R. P. van Gompel (Ed.), *Sentence processing* (pp. 51–77). New York, NY: Psychology Press.

McRae, K., Spivey-Knowlton, M. J., & Tanenhaus, M. K. (1998). Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. *Journal of Memory and Language*, *38*(3), 283–312.

Megalingam, R., Nair, R., & Prakhya, S. (2011, 28 2011-march 3). Automated voice based home navigation system for the elderly and the physically challenged. In *Proceedings of the second international conference* *on wireless communication, vehicular technology, information theory and aerospace electronic systems technology (wireless VITAE)* (p. 1 -5).

Meo, T., McMahan, B., & Stone, M. (2014). Generating and resolving vague color references. *Proceedings of the eighteenth SEMDIAL Workshop on the Semantics and Pragmatics of Dialogue (DialWatt)*, 107–115.

Merel, T., & Tong, J. (2016). *Augmented/virtual reality report 2016* (Tech. Rep.). Menlo Park, CA: Digi-Capital.

Metzing, C., & Brennan, S. E. (2003). When conceptual pacts are broken: Partner-specific effects on the comprehension of referring expressions. *Journal of Memory and Language*, *49*(2), 201–213.

Meyer, F. (2013). *Grounding words to objects: A joint model for co-reference and entity resolution using markov logic for robot instruction processing* (Unpublished doctoral dissertation). Hamburg University of Technology (TUHH), Hamburg, Germany.

Mill, J. S. (1884). *A system of logic ratiocinative and inductive: Being a connected view of the principles of evidence and the methods of scientific investigation.* Harper.

Minton, S., Johnston, M. D., Philips, A. B., & Laird, P. (1992). Minimizing conflicts: a heuristic repair method for constraint satisfaction and scheduling problems. *Artificial Intelligence*, *58*(1-3), 161–205.

Mitchell, D. C., Corley, M. M., & Garnham, A. (1992). Effects of context in human sentence parsing: Evidence against a discourse-based proposal mechanism. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*(1), 69.

Mitkov, R. (1999). *Anaphora resolution: the state of the art.* School of Languages and European Studies, University of Wolverhampton.

Montemerlo, M., Roy, N., & Thrun, S. (2003). Perspectives on standardization in mobile robot programming: The carnegie mellon navigation (carmen) toolkit. In *Proceedings of the IEEE/RSJ international conference on intelligent robots and systems (IROS)* (Vol. 3, pp. 2436–2441).

Mori, M., MacDorman (Translator), K. F., & Minato (Translator), T. (2005). The uncanny valley. *Energy*, *7*(4), 33–35.

Morrow, D. G., Bower, G. H., & Greenspan, S. L. (1989). Updating situation models during narrative comprehension. *Journal of memory and language*, *28*(3), 292–312.

Morton, T. S. (2000). Coreference for nlp applications. In *Proceedings of the thirty-eigth annual meeting of the association for computational linguistics (ACL)* (pp. 173–180).

Mozuraitis, M., Chambers, C. G., & Daneman, M. (2015). Privileged versus shared knowledge about object identity in real-time referential processing. *Cognition*, *142*, 148–165.

Murai, A., Mizuguchi, M., Saitoh, T., Osaki, T., & Konishi, R. (2009, 27 2009-oct. 2). Elevator available voice activated wheelchair. In *Proceedings of the eighteenth IEEE international symposium on robot and human interactive communication(RO-MAN)* (p. 730 -735).

Murarka, A., Gulati, S., Beeson, P., & Kuipers, B. (2009). Towards a safe, low-cost, intelligent wheelchair. In *Workshop on planning, perception and navigation for intelligent vehicles (PPNIV)* (pp. 42–50).

Murphy, R. R. (2004). Human-robot interaction in rescue robotics. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *34*(2), 138–153.

Murphy, R. R. (2014). *Disaster robotics.* MIT press.

Murphy, R. R., Tadokoro, S., Nardi, D., Jacoff, A., Fiorini, P., Choset, H., & Erkmen, A. M. (2008). Search and rescue robotics. In *Springer handbook of robotics* (pp. 1151–1173). Springer.

Mutlu, B., Terrell, A., & Huang, C.-M. (2013). Coordination mechanisms in human-robot collaboration. In *Proceedings of the workshop on collaborative manipulation at the eighth ACM/IEEE international conference on human-robot interaction (HRI).*

Nadel, J., Revel, A., Andry, P., & Gaussier, P. (2004). Toward communication: First imitations in infants, low-functioning children with autism and robots. *Interaction Studies*, *5*(1), 45-74.

Nass, C. I., & Brave, S. (2005). *Wired for speech: How voice activates and advances the human-computer relationship.* MIT press Cambridge.

Neerincx, M. A. (2007). Modelling cognitive and affective load for the design of human-machine collaboration. In D. Harris (Ed.), *Engineering psychology and cognitive ergonomics* (Vol. 4562, pp. 568–574). Springer.

Nevatia, Y., Stoyanov, T., Rathnam, R., Pfingsthorn, M., Markov, S., Ambrus, R., & Birk, A. (2008). Augmented autonomy: Improving human-robot team performance in urban search and rescue. In *Proceedings of the IEEE/RSJ international conference on intelligent robots and systems (IROS)* (pp. 2103–2108).

Ng, V. (2010). Supervised noun phrase coreference research: The first fifteen years. In *Proceedings of the forty-eigth annual meeting of the association for computational linguistics (ACL)* (pp. 1396–1411).

Ng-Thow-Hing, V., Luo, P., & Okita, S. (2010). Synchronized gesture and speech production for humanoid robots. In *Proceedings of the IEEE/RSJ international conference on intelligent robots and systems (IROS)* (pp. 4617–4624).

Ni, W., & Crain, S. (1990). How to resolve structural ambiguities. In *Proceedings of the twentieth annual meeting of the north eastern linguistic society* (pp. 414–427).

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological review*, *84*(3), 231.

Nomura, T., & Saeki, K. (2009). Effects of polite behaviors expressed by robots: a case study in japan. In *Proceedings of the IEEE/WIC/ACM international joint conferences on web intelligence and intelligent agent technologies (WI-IAT)* (Vol. 2, pp. 108–114).

Núñez, R. C., Dabarera, R., Scheutz, M., Briggs, G., Bueno, O., Premaratne, K., & Murthi, M. N. (2013, July). DS-based uncertain implication rules for inference and fusion applications. In *Proceedings of the sixteenth international conference on information fusion.*

Ohlheiser, A. (2016). Trolls turned tay, microsoft's fun millennial AI bot, into a genocidal maniac. *The Washington Post*, *25*.

Oosterveld, B., Brusatin, L., & Scheutz, M. (2017). Two bots, one brain: Component sharing in cognitive robotic architectures. In *Proceedings of the twelfth annual ACM/IEEE international conference on human-robot interaction (HRI): Extended abstracts.*

Ortman, J. M., Velkoff, V. A., & Hogan, H. (2014). An aging nation: the older population in the united states. *Washington, DC: US Census Bureau*, 25–1140.

Park, J. J. (2016). *Graceful navigation for mobile robots in dynamic and uncertain environments* (Unpublished doctoral dissertation). University of Michigan – Ann Arbor.

Park, J. J., Johnson, C., & Kuipers, B. (2012). Robot navigation with model predictive equilibrium point control. In *Proceedings of the IEEE/RSJ international conference on intelligent robots and systems (IROS)* (pp. 4945–4952).

Passonneau, R. (2006). Measuring agreement on set-valued items (masi) for semantic and pragmatic annotation. In *Proceedings of the international conference on language resources and evaluation (LREC).*

Peldszus, A., Baumann, T., Buß, O., & Schlangen, D. (2012). Joint satisfaction of syntactic and pragmatic constraints improves incremental spoken language understanding. In *Proceedings of the 13th conference of the european chapter of the association for computational linguistics (EACL)* (pp. 514–523).

Perfetti, C. A. (1990). The cooperative language processors: Semantic influences in an autonomous syntax. *Comprehension processes in reading*, 205–230.

Perneger, T. V. (1998). What's wrong with bonferroni adjustments. *British Medical Journal*, *316*(7139), 1236–1238.

Perrault, C. R., & Allen, J. F. (1980). A plan-based analysis of indirect speech acts. *Computational Linguistics*, *6*(3-4), 167–182.

Petry, M. R., Moreira, A. P., Faria, B. M., & Reis, L. P. (2013). Intellwheels: intelligent wheelchair with user-centered design. In *Proceedings of the fifteenth IEEE international conference on e-health networking, applications & services (Healthcom)* (pp. 414–418).

Pineau, J., Atrash, A., Kaplow, R., & Villemure, J. (2010). On the design and validation of an intelligent powered wheelchair: Lessons from the smartwheeler project. In J. Angeles, B. Boulet, J. Clark, J. Kövecses, & K. Siddiqi (Eds.), *Brain, body and machine* (Vol. 83, p. 259-268). Springer Berlin / Heidelberg.

Pinker, S., Nowak, M. A., & Lee, J. J. (2008). The logic of indirect speech. *Proceedings of the National Academy of Sciences*, *105*(3), 833–838.

Png, S., & Pineau, J. (2011, may). Bayesian reinforcement learning for pomdp-based dialogue systems. In *Proceedings of the IEEE international conference on acoustics, speech and signal processing (ICASSP)* (p. 2156 -2159).

Poesio, M., & Modjeska, N. N. (2002). The this-nps hypothesis: A corpus-based investigation. In *Proceedings of the fourth discourse anaphora and anaphor resolution colloquium (DAARC 2002)* (pp. 157–162).

Popescu-Belis, A., Robba, I., & Sabah, G. (1998). Reference resolution beyond coreference: a conceptual frame and its application. In *Proceedings of the seventeenth annual conference on computational linguistics (COLING)* (pp. 1046–1052).

Poppe, R. (2010). A survey on vision-based human action recognition. *Image and vision computing*, *28*(6), 976–990.

Pritchett, B. L. (1992). *Grammatical competence and parsing performance.* University of Chicago Press.

Prud'Hommeaux, E., & Seaborne, A. (2008). SPARQL query language for RDF. *W3C recommendation*, *15*.

Purver, M. (2004). Clarie: The clarification engine. In *Proceedings of the eigth SEMDIAL workshop on the semantics and pragmatics of dialogue (CATALOG)* (pp. 77–84).

Purver, M., Eshghi, A., & Hough, J. (2011). Incremental semantic construction in a dialogue system. In *Proceedings of the ninth international conference on computational semantics* (pp. 365–369).

Purver, M., Ginzburg, J., & Healey, P. (2003). On the means for clarification in dialogue. In *Current and new directions in discourse and dialogue* (pp. 235–255). Springer.

Qadri, M., & Ahmed, S. (2009). Voice controlled wheelchair using dsk tms320c6711. In *Proceedings of the international conference on signal acquisition and processing (ICSAP)* (pp. 217–220).

Qidwai, U., & Ibrahim, F. (2010, may). Arabic speech-controlled wheelchair: A fuzzy scenario. In *Proceedings of the tenth international conference* *on information sciences signal processing and their applications (ISSPA)* (p. 153 -156).

Quigley, M., Conley, K., Gerkey, B., Faust, J., Foote, T., Leibs, J., ... Ng, A. Y. (2009). Ros: an open-source robot operating system. In *Proceedings of the ICRA workshop on open source software* (Vol. 3, p. 5).

Reiter, E., Dale, R., & Feng, Z. (2000). *Building natural language generation systems* (Vol. 33). MIT Press.

Rekimoto, J., Ayatsuka, Y., & Hayashi, K. (1998). Augment-able reality: Situated communication through physical and digital spaces. In *Proceedings of the second international symposium on wearable computers* (pp. 68–75).

Riek, L. D. (2012). Wizard of oz studies in hri: a systematic review and new reporting guidelines. *Journal of Human-Robot Interaction*, *1*(1).

Röfer, T., Mandel, C., Lankenau, A., Gersdorf, B., & Frese, U. (2009). 15 years of rolland. In *Festschrift dedicated to bernd krieg-brückner on the occasion of his 60th birthday* (p. 255-272).

Ross, R., Shi, H., Vierhuff, T., Krieg-Brückner, B., & Bateman, J. (2005). Towards dialogue based shared control of navigating robots. *Spatial Cognition IV. Reasoning, Action, Interaction*, 478–499.

Routhier, F., Vincent, C., Desrosiers, J., Nadeau, S., & Guerette, C. (2004). Development of an obstacle course assessment of wheelchair user performance (ocawup): a content validity study. *Technol Disabil*, *16*, 19–31.

Roy, D., Hsiao, K.-Y., Mavridis, N., & Gorniak, P. (2003). Ripley, hand me the cup! (sensorimotor representations for grounding word meaning). In *Proceedings of the international conference of automatic speech recognition and understanding.*

Roy, D. K. (2002). Learning visually grounded words and syntax for a scene description task. *Computer Speech & Language*, *16*(3), 353–385.

Roy, N., Teller, S., Reimer, B., Battat, Y., Doshi, F., Hemachandra, S., ... Velez, J. (2011). *The MIT intelligent wheelchair project: Developing a voice-commandable robotic wheelchair.*

Ruíz-Serrano, A., Posada-Gómez, R., Sibaja, A. M., Rodríguez, G. A., Gonzalez-Sanchez, B., & Sandoval-Gonzalez, O. (2013). Development of a dual control system applied to a smart wheelchair, using magnetic and speech control. *Procedia Technology*, *7*, 158–165.

Ryskin, R. A., Brown-Schmidt, S., Canseco-Gonzalez, E., Yiu, L. K., & Nguyen, E. T. (2014). Visuospatial perspective-taking in conversation and the role of bilingual experience. *Journal of Memory and Language*, *74*, 46–76.

Sadovnik, A., Gallagher, A., & Chen, T. (2013). Not everybody's special: Using neighbors in referring expressions with uncertain attributes. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR) workshops* (pp. 269–276).

Salem, M., Ziadee, M., & Sakr, M. (2013). Effects of politeness and interaction context on perception and experience of hri. In *International conference on social robotics (ICSR)* (pp. 531–541).

Salem, M., Ziadee, M., & Sakr, M. (2014). Marhaba, how may i help you?: Effects of politeness and culture on robot acceptance and anthropomorphization. In *Proceedings of the 2014 ACM/IEEE international conference on human-robot interaction (HRI)* (pp. 74–81).

Sarathy, V., & Scheutz, M. (2015). Semantic representation of objects and function. In *Proceedings of the IROS workshop on learning object affordances.*

Sarathy, V., & Scheutz, M. (2016a). Beyond grasping - perceiving affordances across various stages of cognitive development. In *Proceedings of the the sixth joint IEEE international conference developmental learning and epigenetic robotics (ICDL).*

Sarathy, V., & Scheutz, M. (2016b). Cognitive affordance representations in uncertain logic. In *Fifteenth international conference on the principles of knowledge representation and reasoning.*

Schermerhorn, P., Scheutz, M., & Crowell, C. R. (2008). Robot social presence and gender: Do females view robots differently than males? In *Proceedings of the third ACM/IEEE international conference on human robot interaction (HRI)* (pp. 263–270).

Schermerhorn, P. W., Kramer, J. F., Middendorff, C., & Scheutz, M. (2006). Diarc: A testbed for natural human-robot interaction. In *Proceedings of the twentieth AAAI conference on artificial intelligence* (pp. 1972–1973).

Scheutz, M. (2006). ADE - steps towards a distributed development and runtime environment for complex robotic agent architectures. *Applied Artificial Intelligence*, *20*(4-5), 275-304.

Scheutz, M. (2011). The inherent dangers of unidirectional emotional bonds between huma ns and social robots. *Robot Ethics: The Ethical and Social Implications of Robotics*, 205.

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 AAAI workshop on intelligent robotic systems.*

Scheutz, M., & Schermerhorn, P. (2009). Affective goal and task selection for social robots. In J. Vallverdú & D. Casacuberta (Eds.), *The handbook of research on synthetic emotions and sociable robotics* (pp. 74–87). IGI Global.

Scheutz, M., Schermerhorn, P., & Kramer, J. (2006). The utility of affect expression in natural language interactions in joint human-robot tasks. In *Proceedings of the first ACM SIGCHI/SIGART conference on human-robot interaction (HRI)* (pp. 226–233). New York, NY, USA: ACM.

Scheutz, M., Schermerhorn, P., Kramer, J., & Anderson, D. (2007, May). First steps toward natural human-like HRI. *Autonomous Robots*, *22*(4), 411–423.

Schlangen, D., Baumann, T., & Atterer, M. (2009). Incremental reference resolution: The task, metrics for evaluation, and a Bayesian filtering model that is sensitive to disfluencies. In *Proceedings of the tenth annual meeting of the special interest group on discourse and dialogue (SIGDIAL).*

Schreitter, S., & Krenn, B. (2014). Exploring inter-and intra-speaker variability in multi-modal task descriptions. In *Proceedings of the twenty-third IEEE international symposium on robot and human interactive communication (RO-MAN).*

Schröder, M., & Trouvain, J. (2003). The german text-to-speech synthesis system mary: A tool for research, development and teaching. *International Journal of Speech Technology*, *6*(4), 365–377.

Searle, J. R. (1969). *Speech acts: An essay in the philosophy of language* (Vol. 626). Cambridge university press.

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), 1–23.

Searle, J. R. (2007). Grice on meaning: 50 years later. *Teorema*, *26*(2), 9–18.

Sedivy, J. C. (2002). Invoking discourse-based contrast sets and resolving syntactic ambiguities. *Journal of Memory and Language*, *46*(2), 341–370.

Shafer, G. (1976). *A mathematical theory of evidence*. Princeton University Press.

Sheikh, S. A. M. S., & Rotake, D. R. (2015). An evolutionary approach for smart wheelchair system. In *Proceedings of the international conference on communications and signal processing (ICCSP)* (pp. 1811–1815).

Shimizu, N., & Haas, A. (2009). Learning to follow navigational route instructions. In *Proceedings of the twenty-first international jont conference on artifical intelligence (IJCAI)* (pp. 1488–1493). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Sidner, C. (1986). Focusing in the comprehension of definite anaphora. In *Readings in natural language processing* (pp. 363–394).

Sidner, C. L., & Lee, C. (2005). Robots as laboratory hosts. *interactions*, *12*(2), 24–26.

Škraba, A., Stojanović, R., Zupan, A., Koložvari, A., & Kofjač, D. (2015). Speech-controlled cloud-based wheelchair platform for disabled persons. *Microprocessors and Microsystems*, *39*(8), 819–828.

Smith, B. M., & Grant, S. A. (1997). Trying harder to fail first. In *Proceedings of the thirteenth european conference on artificial intelligence (ECAI).* Citeseer.

Smith, G. J., & Gero, J. S. (2005). What does an artificial design agent mean by being 'situated'? *Design studies*, *26*(5), 535–561.

Soon, W. M., Ng, H. T., & Lim, D. C. Y. (2001). A machine learning approach to coreference resolution of noun phrases. *Computational linguistics*, *27*(4), 521–544.

Sperber, D., & Wilson, D. (1986). *Relevance: Communication and cognition* (Vol. 142). Citeseer.

Spivey, M. J., & Tanenhaus, M. K. (1998). Syntactic ambiguity resolution in discourse: modeling the effects of referential context and lexical frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*(6), 1521.

Spivey-Knowlton, M., & Tanenhaus, M. (1994). Referential context and syntactic ambiguity resolution. *Perspectives on sentence processing*, 49.

Spivey-Knowlton, M., Tanenhaus, M., Eberhard, K., & Sedivy, J. (1995). Eye-movements accompanying language and action in a visual context: Evidence against modularity. In *Proceedings of the seventeenth annual conference of the cognitive science society* (pp. 25–30).

Spivey-Knowlton, M. J. (1996). *Integration of visual and linguistic information: Human data and model simulations* (Unpublished doctoral dissertation). University of Rochester, Rochester, NY.

Spivey-Knowlton, M. J., Trueswell, J. C., & Tanenhaus, M. K. (1993). Context effects in syntactic ambiguity resolution: Discourse and semantic influences in parsing reduced relative clauses. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *47*(2), 276.

Stalnaker, R. C. (1978). Assertion. In *The semantics-pragmatics boundary in philosophy.* Academic Press.

Starner, T., Mann, S., Rhodes, B., Levine, J., Healey, J., Kirsch, D., ... Pentland, A. (1997). Augmented reality through wearable computing. *Presence: Teleoperators and Virtual Environments*, *6*(4), 386–398.

Steedman, M. (1986). *Combinatory grammars and human language processing* (Vol. 279). University of Edinburgh.

Steedman, M. (2000). *The syntactic process* (Vol. 24). MIT Press.

Steels, L., & Hild, M. (2012). *Language grounding in robots.* Springer Science & Business Media.

Stevens, J., & Zhang, Y. (2013). Relative distance and gaze in the use of entity-referring spatial demonstratives: An event-related potential study. *Journal of Neurolinguistics*, *26*(1), 31–45.

Stirling, A. (2010). Keep it complex. *Nature*, *468*(7327), 1029–1031.

Stone, M. (2003). Specifying generation of referring expressions by example. In *Proceedings of the AAAI spring symposium on natural language generation in spoken and written dialogue* (pp. 133–140).

Strait, M., Canning, C., & Scheutz, M. (2014). Let me tell you! investigating the effects of robot communication strategies in advice-giving situations based on robot appearance, interaction modality and distance. In *Proceedings of the ACM/IEEE international conference on human-robot interaction (HRI)* (pp. 479–486).

Strawson, P. F. (1950). On referring. *Mind*, *59*(235), 320–344.

Stubbs, K., Hinds, P. J., & Wettergreen, D. (2007). Autonomy and common ground in human-robot interaction: A field study. *IEEE Intelligent Systems*, *22*(2).

Suk, S.-Y., Chung, H.-Y., & Kojima, H. (2007). Voice/non-voice classification using reliable fundamental frequency estimator for voice activated powered wheelchair control. In Y.-H. Lee, H.-N. Kim, J. Kim, Y. Park, L. Yang, & S. Kim (Eds.), *Embedded software and systems* (Vol. 4523, p. 347-357). Springer Berlin / Heidelberg.

Sun, R. (2008). *The Cambridge handbook of computational psychology.* Cambridge University Press.

Sycara, K., Paolucci, M., Van Velsen, M., & Giampapa, J. (2003). The retsina mas infrastructure. *Autonomous agents and multi-agent systems*, *7*(1-2), 29–48.

Tabor, W., & Tanenhaus, M. K. (1999). Dynamical models of sentence processing. *Cognitive Science*, *23*(4), 491–515.

Tabuse, M., Kitaoka, T., & Nakai, D. (2011). Outdoor autonomous navigation using surf features. *Artificial Life and Robotics*, *16*(3), 356–360.

Talamadupula, K., Benton, J., Kambhampati, S., Schermerhorn, P., & Scheutz, M. (2010). Planning for human-robot teaming in open worlds. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *1*(2), 14.

Talamadupula, K., Briggs, G., Chakraborti, T., Scheutz, M., & Kambhampati, S. (2014). Coordination in human-robot teams using mental modeling and plan recognition. In *Proceedings of the IEEE/RSJ international conference on intelligent robots and systems (IROS)* (pp. 2957–2962).

Talamadupula, K., Kambhampati, S., Schermerhorn, P., Benton, J., & Scheutz, M. (2011). Planning for human-robot teaming. In *Proceedings of the ICAPS workshop on scheduling and planning applications (SPARK)* (Vol. 67).

Tanaka, H., Tokunaga, T., & Shinyama, Y. (2004). Animated agents capable of understanding natural language and performing actions. In *Life-like characters* (pp. 429–443). Springer.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science (New York, N.Y.)*, *268*(5217), 1632–1634.

Tanenhaus, M. K., & Trueswell, J. C. (1995). *Sentence comprehension.* Academic Press.

Tang, Y., Hang, C.-W., Parsons, S., & Singh, M. P. (2012). Towards argumentation with symbolic Dempster-Shafer evidence. In *Proceedings of the second international conference on computational models of argument (COMMA)* (pp. 462–469).

Tannen, D. (1981). Indirectness in discourse: Ethnicity as conversational style. *Discourse processes*, *4*(3), 221–238.

Tao, Y., Wang, T., Wei, H., & Chen, D. (2009). A behavior control method based on hierarchical pomdp for intelligent wheelchair. In *Proceedings of the IEEE/ASME international conference on advanced intelligent mechatronics (AIM)* (pp. 893–898).

Tellex, S., Kollar, T., Dickerson, S., Walter, M. R., Banerjee, A. G., Teller, S., & Roy, N. (2011a). Approaching the symbol grounding problem with probabilistic graphical models. *AI magazine*, *32*(4), 64–76.

Tellex, S., Kollar, T., Dickerson, S., Walter, M. R., Banerjee, A. G., Teller, S., & Roy, N. (2011b). Understanding natural language commands for robotic navigation and mobile manipulation. In *Proceedings of the twenty-fifth AAAI conference on artificial intelligence.*

Tellex, S., Thaker, P., Deits, R., Simeonov, D., Kollar, T., & Roy, N. (2012). *Toward a probabilistic approach to acquiring information from human partners using language* (Tech. Rep.). Cambridge, MA: MIT.

Tellex, S., Thaker, P., Deits, R., Simeonov, D., Kollar, T., & Roy, N. (2013). Toward information theoretic human-robot dialog. *Robotics*, 409.

Tenbrink, T., Ross, R., Thomas, K., Dethlefs, N., & Andonova, E. (2010). Route instructions in map-based human–human and human–computer dialogue: A comparative analysis. *Journal of Visual Languages & Computing*, *21*(5), 292–309.

Tenorth, M., & Beetz, M. (2009). Knowrob—knowledge processing for autonomous personal robots. In *Proceedings of the IEEE/RSJ international conference on intelligent robots and systems (IROS)* (pp. 4261–4266).

Tenorth, M., & Beetz, M. (2012). A unified representation for reasoning about robot actions, processes, and their effects on objects. In *Proceedings of the IEEE/RSJ international conference on intelligent robots and systems (IROS)* (pp. 1351–1358).

Tews, A. D., Mataric, M. J., & Sukhatme, G. S. (2003). A scalable approach to human-robot interaction. In *Proceedings of the IEEE international conference on robotics and automation (ICRA)* (Vol. 2, pp. 1665–1670).

Torrey, C., Fussell, S., & Kiesler, S. (2013). How a robot should give advice. In *Proceedings of the eighth ACM/IEEE international conference on human-robot interaction (HRI)* (pp. 275–282).

Torrey, C., Powers, A., Marge, M., Fussell, S. R., & Kiesler, S. (2006). Effects of adaptive robot dialogue on information exchange and social relations. In *Proceedings of the first ACM SIGCHI/SIGART conference on human-robot interaction (HRI)* (pp. 126–133).

Traum, D. R. (1994). *A computational theory of grounding in natural language conversation* (Unpublished doctoral dissertation). University of Rochester, Rochester, NY.

Traum, D. R. (1999). Computational models of grounding in collaborative systems. In *Proceedings of the AAAI fall symposium on psychological models of communication in collaborative systems* (pp. 124–131).

Traum, D. R., & Allen, J. F. (1994). Discourse obligations in dialogue processing. In *Proceedings of the thirty-second annual meeting of the association for computational linguistics (ACL)* (pp. 1–8).

Traxler, M. J. (2011). *Introduction to psycholinguistics: Understanding language science*. John Wiley & Sons.

Trefler, E., Fitzgerald, S. G., Hobson, D. A., Bursick, T., & Joseph, R. (2004). Outcomes of wheelchair systems intervention with residents of long-term care facilities. *Assistive Technology*, *16*(1), 18-27.

Tsui, K. M., Norton, A., Brooks, D. J., McCann, E., Medvedev, M. S., & Yanco, H. A. (2013). Design and development of two generations of semi-autonomous social telepresence robots. In *Proceedings of the IEEE international conference on technologies for practical robot applications (TePRA)* (pp. 1–6).

Tsujimoto, M., Munekat, N., & Ono, T. (2013). Evaluating how the human' s impression formation of. *Human-Agent Interaction*.

Van Deemter, K. (2016). *Computational models of referring: A study in cognitive science*. Cambridge, Massachusetts: MIT Press.

Van Deemter, K., & Gatt, A. (2009). Beyond dice: measuring the quality of a referring expression. In *Proceedings of the first workshop on the production of referring expressions (PRE-CogSci).*

van der Sluis, I., & Krahmer, E. (2007). *Generating Multimodal References* (Vol. 44) (No. April 2014).

Ventura, J., Arth, C., Reitmayr, G., & Schmalstieg, D. (2014). Global localization from monocular slam on a mobile phone. *IEEE transactions on visualization and computer graphics*, *20*(4), 531–539.

Versley, Y., Ponzetto, S. P., Poesio, M., Eidelman, V., Jern, A., Smith, J., ... Moschitti, A. (2008). Bart: A modular toolkit for coreference resolution. In *Proceedings of the forty-sixth annual meeting of the association for computational linguistics on human language technologies: Demo session* (pp. 9–12).

Viethen, J., & Dale, R. (2006). Towards the evaluation of referring expression generation. In *Proceedings of the fourth australasian language technology workshop* (pp. 115–122). Sydney, Australia.

Wahlster, W., André, E., Graf, W., & Rist, T. (1991). Designing illustrated texts: how language production is influenced by graphics generation. In *Proceedings of the fifth conference on european chapter of the association for computational linguistics (EACL)* (pp. 8–14).

Walker, W., Lamere, P., Kwok, P., Raj, B., Singh, R., Gouvea, E., ... Woelfel, J. (2004). *Sphinx-4: A flexible open source framework for speech recognition* (Tech. Rep.). Santa Clara, CA: Sun Microsystems, Inc.

Wallam, F., & Asif, M. (2011). Dynamic finger movement tracking and voice commands based smart wheelchair. *International Journal of Computer and Electrical Engineering*, *3*(4), 497.

Walter, M. R., Hemachandra, S., Homberg, B., Tellex, S., & Teller, S. (2013, June). Learning semantic maps from natural language descriptions. In *Proceedings of robotics: Science and systems (RSS).* Berlin, Germany.

Walters, M. L., Syrdal, D. S., Dautenhahn, K., Te Boekhorst, R., & Koay, K. L. (2008). Avoiding the uncanny valley: robot appearance, personality and consistency of behavior in an attention-seeking home scenario for a robot companion. *Autonomous Robots*, *24*(2), 159–178.

Wang, H., Li, T., & Zheng, F. (2015). A wheelchair platform controlled by a multimodal interface. In *Proceedings of the second international conference on information science and control engineering* (pp. 587–590).

Wang, J. W. J. (1994). On sign-board based inter-robot communication in distributed robotic systems. *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA).*

Wazinski, P. (1992). Generating spatial descriptions for cross-modal references. In *Proceedings of the third conference on applied natural language processing* (pp. 56–63).

Weber, A., & Crocker, M. W. (2012). On the nature of semantic constraints on lexical access. *Journal of Psycholinguistic Research*, *41*(3), 195–214.

Wickramarathne, T. L., Premaratne, K., & Murthi, M. N. (2012). Consensus-based credibility estimation of soft evidence for robust data fusion. In *Belief functions: Theory and applications* (pp. 301–309). Springer.

Williams, T. (2014). Position paper. In *Proceedings of the tenth young researchers' roundtable on spoken dialog systems (YRRSDS).*

Williams, T. (2015). Towards more natural human-robot dialogue. In *Proceedings of the sixth pioneers workshop at HRI 2015.*

Williams, T. (2016). Architectural mechanisms for situated natural language understanding in uncertain and open worlds. In *Proceedings of the AAAI doctoral consortium.*

Williams, T., Acharya, S., Schreitter, S., & Scheutz, M. (2016). Situated open world reference resolution for human-robot dialogue. In *Proceedings of the eleventh ACM/IEEE international conference on human-robot interaction.*

Williams, T., Briggs, G., Oosterveld, B., & Scheutz, M. (2015). Going beyond command-based instructions: Extending robotic natural language interaction capabilities. In *Proceedings of twenty-ninth AAAI conference on artificial intelligence.*

Williams, T., Briggs, P., Pelz, N., & Scheutz, M. (2014). Is robot telepathy acceptable? investigating effects of nonverbal robot-robot communication on human-robot interaction. In *Proceedings of twenty-third IEEE symposium on robot and human interactive communication (RO-MAN).*

Williams, T., Briggs, P., & Scheutz, M. (2015). Covert robot-robot communication: Human perceptions and implications for human-robot interaction. *Journal of Human-Robot Interaction (JHRI)*.

Williams, T., Cantrell, R., Briggs, G., Schermerhorn, P., & Scheutz, M. (2013). Grounding natural language references to unvisited and hypothetical locations. In *Proceedings of the twenty-seventh AAAI conference on artificial intelligence.*

Williams, T., Johnson, C., Scheutz, M., & Kuipers, B. (2017). A tale of two architectures: A dual-citizenship integration of natural language and the cognitive map. In *Proceedings of the sixteenth international conference on autonomous agents and multi-agent systems (AAMAS).* Sao Paolo, Brazil.

Williams, T., Núñez, R. C., Briggs, G., Scheutz, M., Premaratne, K., & Murthi, M. N. (2014). A Dempster-Shafer theoretic approach to understanding indirect speech acts. *Advances in Artificial Intelligence*.

Williams, T., & Scheutz, M. (2015a). A domain-independent model of open-world reference resolution. In *Proceedings of the thirty-seventh annual meeting of the cognitive science society.*

Williams, T., & Scheutz, M. (2015b). POWER: A domain-independent algorithm for probabilistic, open-world entity resolution. In *Proceedings of the IEEE/RSJ international conference on intelligent robots and systems (IROS).*

Williams, T., & Scheutz, M. (2016a). A framework for resolving open-world referential expressions in distributed heterogeneous knowledge bases. In *Proceedings of the thirtieth AAAI conference on artificial intelligence* (pp. 3598–3964).

Williams, T., & Scheutz, M. (2016b). Resolution of referential ambiguity using Dempster-Shafer theoretic pragmatics. In *Proceedings of the AAAI fall symposium on artificial intelligence for human-robot interaction (AI-HRI).*

Williams, T., & Scheutz, M. (2017). Reference resolution in robotics: A givenness hierarchy theoretic approach. In J. Gundel & B. Abbott (Eds.), *The Oxford handbook of reference.* Oxford University Press.

Williams, T., Schreitter, S., Acharya, S., & Scheutz, M. (2015). Towards situated open-world reference resolution. In *Proceedings of the AAAI fall symposium on artificial intelligence for human-robot interaction (AI-HRI).*

Wilske, S., & Kruijff, G.-J. (2006). Service robots dealing with indirect speech acts. In *Proceedings of the IEEE/RSJ international conference on intelligent robots and systems (IROS)* (pp. 4698–4703).

Wilson, J. R., Krause, E., Scheutz, M., & Rivers, M. (2016). Analogical generalization of actions from single exemplars in a robotic architecture. In *Proceedings of the international conference on autonomous agents & multiagent systems (AAMAS)* (pp. 1015–1023).

Winograd, T. (1971). *Procedures as a representation for data in a computer program for understanding natural language* (Tech. Rep.). Cambridge, MA: Massachsetts Institute of Technology.

Xueping, X., & Rong, Z. (2009). The influence of pointing distance and pointing mode on the choice of spatial demonstratives. *Modern Foreign Languages*, *4*, 015.

Yager, R. R. (1987). On the Dempster-Shafer framework and new combination rules. *Information sciences*, *41*(2), 93–137.

Yanco, H. A. (2001). Development and testing of a robotic wheelchair system for outdoor navigation. In *Proceedings of the conference of the rehabilitation engineering and assistive technology society of north america* (pp. 588–603).

Yanco, H. A., & Drury, J. L. (2004). Classifying human-robot interaction: an updated taxonomy. In *Proceedings of the third international conference on systems, man and cybernetics (SMC)* (pp. 2841–2846).

Yazdani, F., Brieber, B., & Beetz, M. (2016). Cognition-enabled robot control for mixed human-robot rescue teams. In *Intelligent autonomous systems 13* (pp. 1357–1369). Springer.

Yee, E., & Sedivy, J. C. (2006). Eye movements to pictures reveal transient semantic activation during spoken word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(1), 1–14.

Yoon, S. O., & Brown-Schmidt, S. (2013). Lexical differentiation in language production and comprehension. *Journal of Memory and Language*, *69*(3), 397–416.

Zadeh, L. A. (1979). *On the validity of Dempster's rule of combination of evidence*. Electronics Research Laboratory, University of California.

Zarrieß, S., & Schlangen, D. (2016). Towards generating colour terms for referents in photographs: Prefer the expected or the unexpected? In *Proceedings of the ninth international natural language generation conference (INLG).*

Zender, H., Kruijff, G.-J. M., & Kruijff-Korbayová, I. (2009). Situated resolution and generation of spatial referring expressions for robotic assistants. In *Proceedings of the twenty-first international jont conference on artifical intelligence (IJCAI)* (pp. 1604–1609). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Zhang, Y., Zhang, J., & Luo, Y. (2011). A novel intelligent wheelchair control system based on hand gesture recognition. In *Proceedings of the IEEE/ICME international conference on complex medical engineering (CME)* (pp. 334–339).