TOPICS
TOPICS IN COGNITIVE SCIENCE

Topics in Cognitive Science 6 (2014) 513–533
Copyright © 2014 Cognitive Science Society, Inc. All rights reserved.
ISSN:1756-8757 print / 1756-8765 online
DOI: 10.1111/tops.12101

Strategies for Human-Driven Robot Comprehension of Spatial Descriptions by Older Adults in a Robot Fetch Task

Laura Carlson,[a] Marjorie Skubic,[b] Jared Miller,[a] Zhiyu Huo,[b] Tatiana Alexenko[b]

[a]*Department of Psychology, University of Notre Dame*
[b]*Electrical and Computer Engineering Department, University of Missouri-Columbia*

Received 30 March 2012; received in revised form 26 March 2013; accepted 1 April 2014

Abstract

This contribution presents a corpus of spatial descriptions and describes the development of a human-driven spatial language robot system for their comprehension. The domain of application is an eldercare setting in which an assistive robot is asked to “fetch” an object for an elderly resident based on a natural language spatial description given by the resident. In Part One, we describe a corpus of naturally occurring descriptions elicited from a group of older adults within a virtual 3D home that simulates the eldercare setting. We contrast descriptions elicited when participants offered descriptions to a human versus robot avatar, and under instructions to tell the addressee how to find the target versus where the target is. We summarize the key features of the spatial descriptions, including their dynamic versus static nature and the perspective adopted by the speaker. In Part Two, we discuss critical cognitive and perceptual processing capabilities necessary for the robot to establish a common ground with the human user and perform the “fetch” task. Based on the collected corpus, we focus here on resolving the perspective ambiguity and recognizing furniture items used as landmarks in the descriptions. Taken together, the work presented here offers the key building blocks of a robust system that takes as input natural spatial language descriptions and produces commands that drive the robot to successfully fetch objects within our eldercare scenario.

Keywords: Human–robot interaction; Spatial language; Spatial descriptions; Fetch task; Task instructions; Assistive robotics; Eldercare

Correspondence should be sent to Laura Carlson, Department of Psychology, University of Notre Dame, Notre Dame, IN 46556. E-mail: lcarlson@nd.edu

1. Introduction

Human comprehension of spatial language is a complex activity. Consider (1):

(1) "*Your eyeglasses are behind the radio on the table in the bedroom.*"

First, the spatial term "*behind*" is qualitative, circumscribing a vague region of space, rather than a precise metric location (Landau & Jackendoff, 1993; Talmy, 1983). Second, this region depends upon the interpretation of the spatial term *behind*, which could be based on the perspective of the speaker or addressee, the orientation of the room, or the sides of the objects (radio, table) (Levelt, 1996; Levinson, 1996). Third, the description is typically comprehended within a conversational context that includes the speaker's assumptions about the addressee's capabilities and knowledge (Clark, Schreuder, & Buttrick, 1983) and the establishment of a common ground (Clark, 1996). Despite these complexities, human interpretation of such instructions normally proceeds naturally and fluently.

In sharp contrast, the comprehension of spatial descriptions is particularly problematic for robots. First, robots "think" and move in terms of quantitative rather than qualitative relations, relying on mathematical expressions and numbers. Second, given that spatial terms are ambiguous, strategies are required for determining perspective, and allowing for reinterpretation when necessary. Third, assumptions about common ground and the capabilities of the addressee depend upon whether the addressee is a human or robot (Tenbrink, Fischer, & Moratz, 2002). Previous work has proposed a qualitative spatial representation for robot navigation (Gribble, Browning, Hewett, Remolina, & Kuipers, 1998; Kuipers, 2000). Others have proposed the use of directional commands for mobile robots, for example, *turn left, go forward*, or *go past <the landmark>* (Muller et al., 2000; Tellex & Roy, 2006; Levit & Roy, 2007; see Klippel & Montello, 2007 for more on the interpretation of directions). However, with such work confined to a 2D ground plane, further development is needed to achieve human-like comprehension of natural 3D spatial descriptions. While it is possible to train a speaker to restrict robot directives to a set of constrained commands (i.e., make the user adapt to the robot), our intent instead is to explore how the robot can adapt to the human user, with all of the ambiguities and complexities inherent in natural language.

We addressed these complexities by collecting a corpus of spatial descriptions elicited within a 3D virtual setting in the context of a fetch task. In Part One, we describe the corpus, looking systematically at how various measures, including word count, the inclusion of different categories of words, perspective, and the dynamic or static nature of the description, vary as a function of the addressee and the instruction to speakers. We conclude Part One by indicating next steps for the corpus, which include assessing the effectiveness of the spatial descriptions (Carlson, Skubic, Miller, Huo, & Alexenko, 2013), and comparing performance of robot simulations with human performance in the fetch task by contrasting path metrics, including length and number of pauses (Skubic, Huo, Alexenko, Carlson, & Miller, 2013). In Part Two, we discuss how the corpus in Part One provides the basis for the development of robot strategies designed to comprehend these descriptions. We then provide an overview of the system that we are building that contains the components of natural language processing (NLP), navigation instruction representation, identification of perspective, and robot behavior including perceptual recognition of objects in the environment (Skubic et al., 2013). Then for illustration, we present an in-depth description of two of these processes that are essential for establishing common ground between speaker and addressee: identification of perspective, and recognition of furniture items used as landmarks.

2. Part 1. The spatial description corpus

Our corpus of spatial descriptions was collected within an eldercare scenario in which a participant navigates through a virtual 3D house environment to find a target, and then provides spatial descriptions that specify the target's location to an avatar in the context of a fetch task. The house consisted of a long hallway with a living room on the left and a bedroom on the right that was modeled after lab space at the University of Missouri. Fig. 1, Panel A shows a survey view of the space; Panel B shows portions of the virtual rooms; and Panel C shows portions of the rooms within the physical lab space. This correspondence between virtual and physical space enables future comparisons across types of environment. The selection of this scenario is motivated by the fact that older adults identify fetch tasks in which the robot retrieves a desired object as one of the top five tasks for assistive devices (Beer et al., 2012). Moreover, older adults report a strong preference for being able to speak naturally to assistive devices, rather than other types of interfaces (Scopelliti, Giuliani, & Fornara, 2005).

We focused on three critical aspects of the scenario. First, research has shown an increased reliance on landmarks during wayfinding by older adults (Davis, Therrien, & West, 2009). In our virtual house, there were several types of landmarks that could be used to specify the location of the targets, which were positioned on tables in the virtual scene: house units such as walls or rooms; furniture units such as a couch, bed, or table; and object units that were co-located on the tables, such as a lamp or wallet on the table next to the target. Our interest was in identifying which type of landmarks older adults would typically include. Another interest was whether descriptions relied on adopting an intrinsic reference frame for a furniture object. For example, consider the couch shown in Fig. 1. The phrase "*the table in front of the couch*" (see the full description in Table 1) is potentially ambiguous; it might refer to the circular table, with *front* defined by the viewing perspective when entering the living room, or the rectangular table, with *front* defined as the front side of the couch. The robot strategies must recognize the furniture objects and also note their orientations, which will then be used during ambiguity resolution.

Second, within the virtual environment, participants offered spatial descriptions to either a human avatar named Brian (see Fig. 2, Panel A) or a robot avatar modeled after our real-life robot (see Fig. 2, Panel B). The participant and the addressee always faced each other, so that their perspectives were offset 180 degrees. This enabled us to code the spatial descriptions as being consistent with the participant's perspective or the addressee's perspective, based on the use of the terms *left* and *right* to guide the addressee to the target. When neither term was used (e.g., *"The wallet is on the table in the bedroom"*), the per-

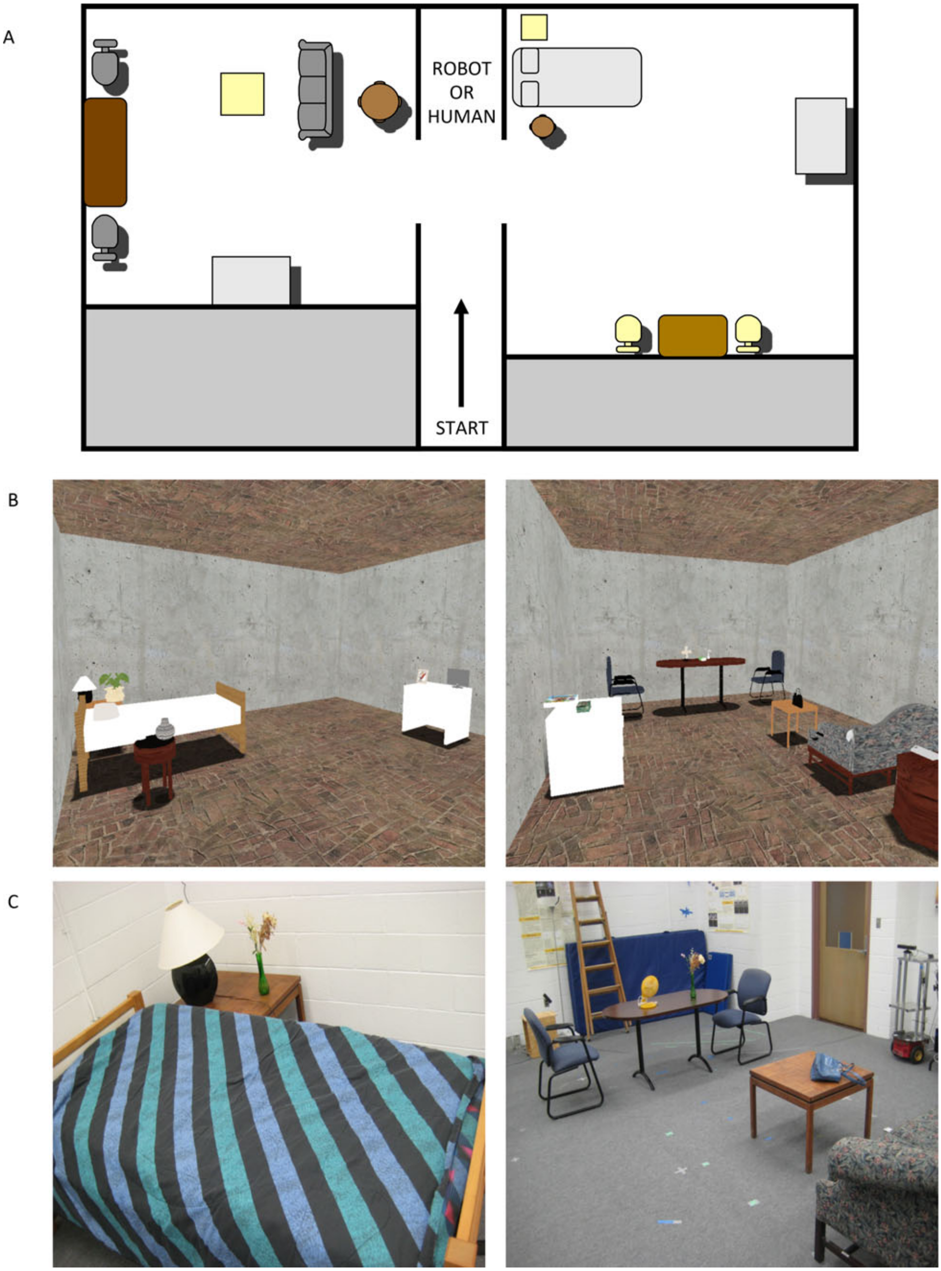

Fig. 1. Panel A shows a survey perspective of the house environment that contains a living room on the left, a hallway in the middle, and a bedroom on the right. A trial within the virtual environment started at the location in the hallway indicated by "start" with the perspective indicated by the arrow, facing the robot or human avatar addressee. Panel B shows part of the living room and bedroom at the eye-level height adopted in the experiment. Panel C shows the corresponding rooms in Skubic's lab at Missouri. The similarity across the virtual and physical spaces will enable us to test in the future for any differences based on virtual versus actual environments and virtual versus actual addressees (robot and human).

Table 1
Sample descriptions for the target "letter" from the corpus, presented as a function of addressee and instruction. Each description is provided by a different participant, and each condition shows two descriptions

	Brian	Robot
How	Brian go to your right the room on your right and as you enter take another right turn and on the table in front of you you'll see a letter Oh, turn right, and then right again, and then it will be on the table	Turn left into that room and the letter is on the table near the door Okay come four steps forward. Turn my left into the room and then take a sharp right toward the table in the corner and the letter is right in right at the edge of the table. . .the forward edge
Where	The letter is in the living room to your right on that table Brian if you go in to the room on the left and make an immediate right to the table in front of the couch you will find the letter	It is in the room on the right The letter is in the living room on the six-sided table

spective was coded as indeterminate. When speakers and addressees are misaligned, speakers generally prefer the addressee's perspective (Mainwaring, Tversky, Ohgishi, & Schiano, 2003; Schober, 1993), with this preference also present for robot addressees (Tenbrink et al., 2002). However, DiSalvo, Gemperle, Forlizzi, and Kiesler (2002) showed that humans define their interactions with avatars based on the degree of anthropomorphism, which suggests potential variation in the preference for an addressee perspective across the two types of avatars. Moreover, Scopelliti et al. (2005) found that older adults showed negative emotional responses to robots. Such attitudes may be reflected in a preference to use a speaker's perspective when speaking to a robot, given that this minimizes the speaker's cognitive load with an increased burden for the addressee (Schober, 1993).

Third, we manipulated the instructions, using these two prompts:

Where prompt: Tell (Brian/the robot) where the <target> is
How prompt: Tell (Brian/the robot) how to find the <target>

Plumert, Carswell, DeVet, and Ihrig, (1995) found that *where* prompts elicited descriptions with hierarchical sequences ordered from small to big units (e.g., "*behind the lamp, on the table, in the bedroom*"), whereas *how* prompts elicited sequences ordered from big to small (e.g., "*in the bedroom, on the table, behind the lamp*"). Because our environment was not as hierarchically structured as Plumert et al.'s, we characterized descriptions instead along a dynamic versus static dimension. Dynamic descriptions actively moved the listener through the environment to the destination, and resembled step-by-step directions (e.g., "*Go forward, turn right, look to the left. . .*"). In contrast, static descriptions described the location of the object in a stable, stationary manner independent of the

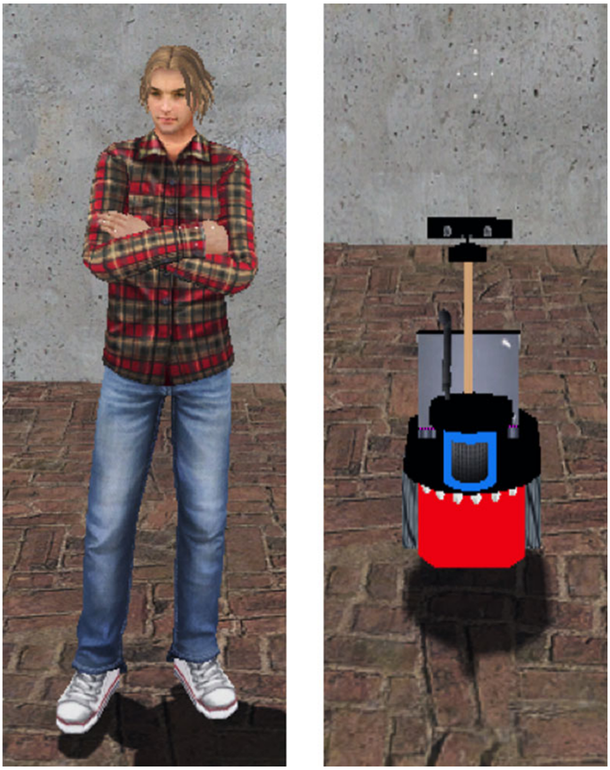

Fig. 2. Addressees in the fetch task: Brian and the robot. All participants were told that they were facing the robot to ensure that they knew the robot's front side.

addressee (e.g., "*The cell phone is on the table by the bed in the bedroom*"). These types of descriptions differ with respect to phrase structure, parts of speech, and types of spatial relations and word choice. The natural language processing strategies for the robot must be flexible enough to accommodate both types, but it would be helpful to understand the contexts in which each type is preferred. For example, in the natural language processing component of our system, separate parsing procedures are being developed for dynamic and static descriptions. Therefore, knowing whether older adults prefer to use static descriptions with the robot as opposed to dynamic descriptions enables the system to make an initial selection about which parsing procedure to try first. We are also investigating consistent features in the language structure that can be used to classify the description as static versus dynamic, to further aid in the parsing.

2.1. Method

2.1.1. Subjects

Sixty-four older adults participated, with a mean age of 76 (range 64–96), recruited from local senior centers in South Bend, IN, and Columbia, MO, and compensated with

$10 for their participation. All gave informed consent and were treated in accordance with APA ethical guidelines. The data were collected at the local senior centers and in the lab at Notre Dame for those who preferred to travel on-site. Participants were prescreened for cognitive impairment with the Mini-Cog Assessment Instrument for Dementia (Borson et al., 2000). Two older adults failed this screening and were replaced in the final sample. Participants also completed three assessments at the end of the experiment to characterize their general health and cognitive functioning. First, they filled out a general health measure that included questions such as date of birth, years of education, and basic health questions about audition and vision and level of satisfaction with their health and physical condition. All participants reported good basic health and overall satisfaction scores (M = 3.88, range of 2–4 on a scale from 1 (dissatisfied) to 4 (satisfied). Second, they completed the Mill Hill vocabulary test (Raven, Court, & Raven, 1977), which provides an indicator of general verbal intelligence. Mean scores were 20.7 (range of 9–31) out of a possible score of 33, indicating unimpaired performance. Third, participants completed the Digit-Symbol test (Wechsler, 1997), which measures visual-motor speed and complexity and motor coordination. Mean scores were 61.2 (range of 24–93) out of a possible score of 93, indicating unimpaired performance.

2.1.2. Stimuli and Design

A virtual house environment (see Fig. 1) was created using 3dsMax Design 2010© (AutoDesk WorldWide Headquarters: Autodesk, Inc., San Rafael, CA) and Google SketchUp© (Trimble Navigation Limited Sunnyvale, CA) and rendered using Half-Life 2 (Valve, Bellevue, WA) gaming software. Each room contained four tables, two chairs, and a couch or bed. Within each room, two potential reference objects and a target were placed on top of each table; objects are listed in the Appendix. Eight versions of the house were created, each containing a single target. A different house version was used for each trial to prevent a preview of targets for future trials. The design was a 2 (addressee: Brian vs. robot avatar) X 2 (instruction: how or where) between subjects factorial design with 16 participants assigned to each of the four conditions.

2.1.3. Procedure

The session began with a brief video tour (42 s) of the house that showed the structure of the rooms. After this, participants explored the virtual environment by telling the experimenter when and where to go. The experimenter controlled the navigation, because pilot testing showed that older adults felt uncomfortable and unfamiliar with this technology, consistent with Ezer (2008). On average it took about 3–5 commands for participants to feel comfortable with this procedure. Participants with the robot as addressee were also shown labeled pictures of the front and back of the robot so that they understood they were facing the front of the robot at the start of each trial.

Each participant completed eight experimental trials. On each trial, participants were shown a picture of the to-be-found target on a gray background. The experimenter named the target to ensure full identification. Participants started in the hall and told the experimenter how to navigate to find the target. Participants were allowed to search as long as necessary to discover the target; the procedure was not timed. No trials were excluded due to an inability to locate the target. Participants were then returned to the starting location, received their assigned "where" or "how" prompt, and described the location of the target to the addressee; these descriptions were recorded. There were no constraints on the content or format of the descriptions. The order of the trials was randomized across all participants. Finally, after completing all trials, participants drew a map of the house that was coded to verify that participants had an accurate representation of the environment in terms of the layout of the rooms.

2.2. Results and discussion

2.2.1. Details of the corpus

The corpus consisted of transcriptions of 512 spatial descriptions (64 participants X 8 trials). Table 1 provides sample descriptions broken down by addressee and instruction. On average, *how* descriptions contained more words (M = 27.0 words/description) than *where* descriptions (M = 19.4), $F(1, 60) = 10.7$, $\eta_p^2 = .151$, $p < .05$. In addition, descriptions given to Brian (M = 25.5 words/description) contained more words than descriptions given to the robot (M = 20.9), $F(1, 60) = 4.0$, $\eta_p^2 = .062$, $p = .05$. There was no interaction, $F < 1$.

We classified the 11,854 words used in the descriptions into major categories, including spatial terms (count = 2689), landmark types (count = 1206: house units = 624, furniture units = 537, and object units = 45), and hedges (count = 80) that modified the directional heading, for example, "*immediately to the right*." We present histograms of the words occurring within these categories in Fig. 3. Table 2 provides counts for each category as a function of addressee and instruction.

2.2.1.1. Spatial terms: The predominant spatial terms were *on, to, left, right, in, and into*. There were more spatial terms in *how* descriptions (M = 4.1) than in *where* descriptions (M = 2.3), $F(1, 60) = 18.5$, $\eta_p^2 = .236$, $p < .05$, consistent with *how* descriptions specifying a path to a location rather than the location per se. There was no difference in the spatial terms given to Brian (M = 3.5) or the robot (M = 2.9), ($F(1,60) = 2.48$, $\eta_p^2 = .040$, $p = .12$), and no interaction ($F < 1$).

2.2.1.2. Landmark types: Both furniture and house units were popular, consistent with Cassenti, Kelley, Avery, and Yagoda (2011), who found improved performance for robot directives that included both spatial terms and larger objects. Strikingly, object units were rarely included. *Where* descriptions contained on average more house units (M = 1.4) than *how* descriptions (M = 1.1), $F(1,60) = 5.0$, $\eta_p^2 = .077$, consistent with *where* descriptions specifying a location rather than a path. There were no differences as a function of addressee or instruction for the number of object units or furniture units (all Fs < 1.4, ps > .24).

2.2.1.3. Hedges: Hedges were not used very frequently (count of 80 out of 512 utterances); however, the two most popular hedges (e.g., "*right* as you enter," "*immediately*

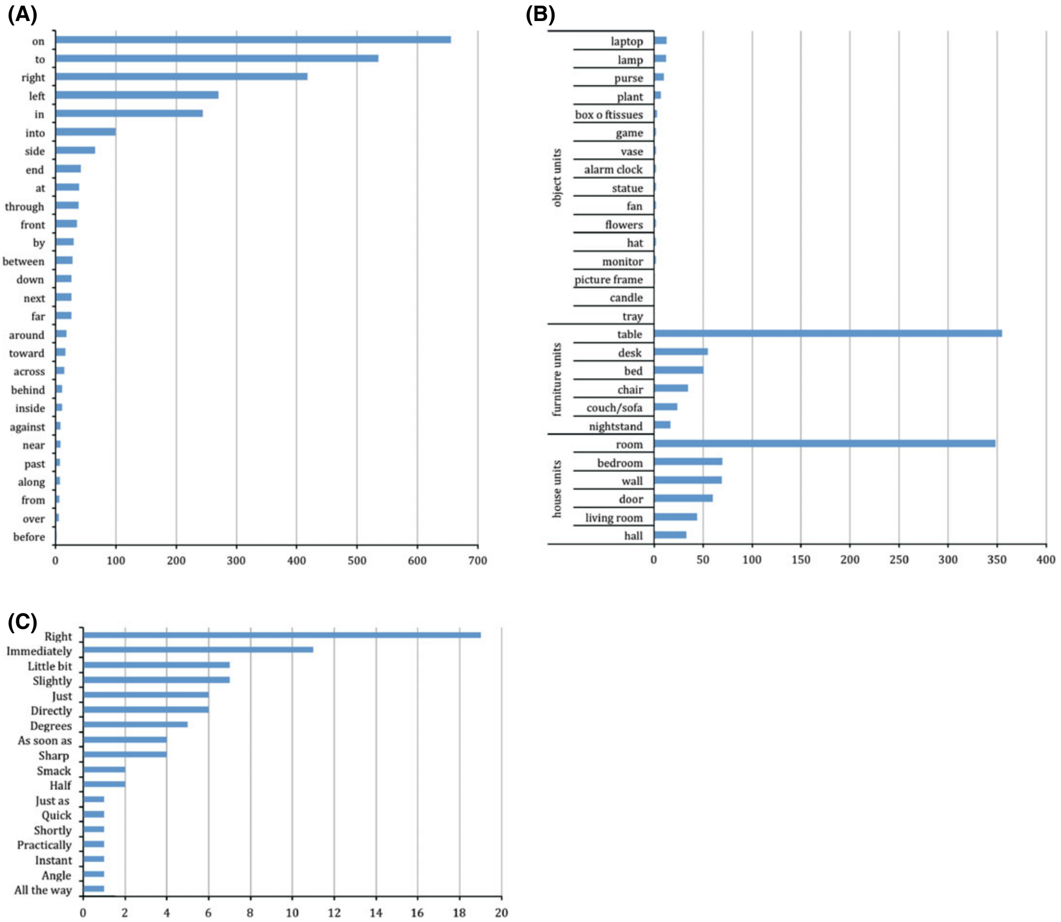

Fig. 3. Histograms for different categories of terms used in the spatial descriptions. Panel A shows spatial terms; Panel B shows object units, furniture units, and house units; Panel C shows hedges.

Table 2
Mean word counts per description for spatial terms, house units, furniture units, object units and hedges as a function of addressee and instruction

	Robot		Brian	
	How	Where	How	Where
Spatial terms	3.65 (.36)	2.18 (.39)	4.47 (.36)	2.59 (.44)
House units	1.21 (.16)	1.18 (.19)	.88 (.11)	1.59 (.12)
Furniture	1.04 (.09)	.97 (.09)	1.05 (.14)	1.13 (.07)
Object units	.04 (.02)	.09 (.04)	.13 (.05)	.09 (.03)
Hedges	.23 (.08)	.02 (.01)	.20 (.06)	.17 (.04)

on your left") tended to occur at the point of discovering the target, with significantly more hedges for *how* instructions (M = .22) than for *where* instructions (M = .09), F (1,60) = 5.4, η_p^2 = .083. Interestingly, there was a marginal interaction between addressee and instruction, F(1,60) = 3.05, η_p^2 = .048, p = .09. With *how* descriptions that emphasized a path, speakers used hedges at the same rate for both Brian (.20) and the robot (.23). However, with *where* descriptions, participants continued to use hedges at that same rate with Brian (.17) but stopped using hedges for the robot (.02). This is an indication that older adults may alter their descriptions when speaking to the robot, due perhaps either to different perceived capabilities or different degrees of willingness to accommodate to the robot. This is an issue that we are exploring in further studies.

2.2.2. Dynamic versus static

Two raters coded the descriptions as dynamic or static, with an interrater reliability of 99%. The mean percentage of dynamic descriptions broken down by addressee and instruction is shown in Fig. 4. A 2 (instruction) X 2 (addressee) between subjects ANOVA revealed a significant effect of instruction, F(1, 60) = 48.5, n_p^2 = .447. For *how* descriptions, the overwhelming preference was to use a dynamic description, almost all of the time (M = 95.3%); however, for *where* descriptions, there was more variation, with fewer dynamic descriptions (M = 35.8%) than static descriptions (64.2%). Moreover, within the *where* instructions, the number of dynamic descriptions was below chance for the robot (M = 28.6; t(15) = 1.98, p < .066) but not for Brian (M = 43.1%; t < 1). This means that older adults were not as willing to offer step-by-step instructions for the robot when the task instructions did not require it. This is consistent with the lack of accommodation in the use of hedges observed for robots in this condition.

2.2.3. Perspective

Of the 512 descriptions in the corpus, 231 were coded as adopting the addressee perspective, 165 as the speaker perspective, and 116 coded as indeterminate. Fig. 5 shows the breakdown of each perspective as a function of addressee and instruction. The addressee perspective used more frequently for Brian (M = 63.7%) than for the robot

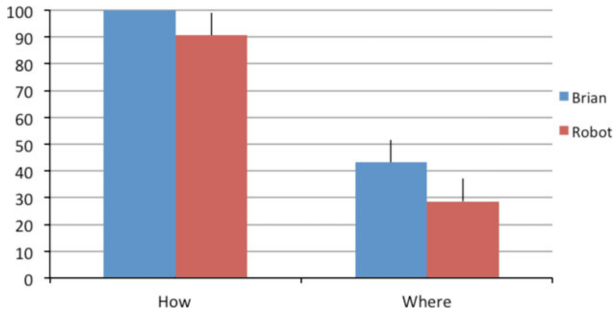

Fig. 4. Percentage of dynamic descriptions as a function of addressee and instruction. Error bars correspond to the standard error of the mean.

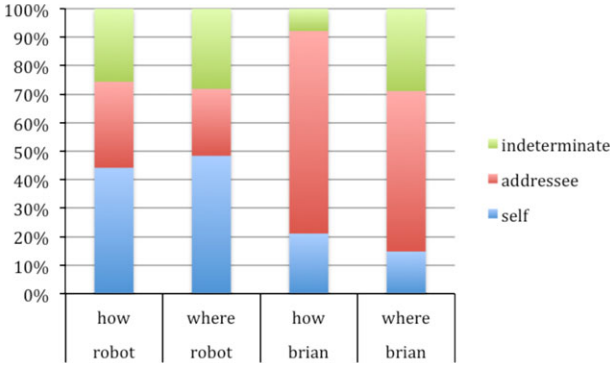

Fig. 5. Percentages of addressee, self and indeterminate perspectives, broken down by addressee type and instruction.

(M = 27.0%), $F(1,60) = 11.5$, $n_p^2 = .161$, $p < .05$; indeed, for the robot, the speaker's perspective was preferred (M = 46.0%).

2.3. Summary of the behavioral results

Generally, descriptions contained a combination of spatial terms and house and furniture landmarks but very few object landmarks. Moreover, there were key differences as a function of addressee and instruction. When talking to the robot, participants preferred to use fewer words and to adopt a speaker's perspective, whereas when talking to Brian, participants used more words and preferred an addressee perspective. We are planning on examining the extent to which these differences as a function of addressee are based on differences in appearance between Brian and the robot or differences in the inferences that speakers make about the capabilities of the addressee.

When describing *how* to find the target, participants consistently used dynamic descriptions that contained more spatial terms and fewer house units and hedges, regardless of addressee. However, when describing *where* the target was, participants used fewer spatial terms and more house units. They were also more likely to use dynamic descriptions for Brian but static descriptions for the robot and to avoid the use of hedges.

2.4. Next steps for the corpus

It is beyond the scope of this article but worth noting that we have also begun assessing the effectiveness of the spatial descriptions that are in our corpus, examining whether the descriptions are accurate, as reflected in whether a human could find the appropriate target given the spatial description. One interesting finding is that in general the *where* descriptions are associated with more successful performance than the *how* descriptions (Carlson et al., 2013). We are also contrasting the performance of the human and the robot in terms of the efficiencies of their paths to the target, using path metrics such as path length and number of pauses (Skubic et al., 2013).

3. Part 2. Establishing common ground with the addressee

The corpus was collected with the objective of capturing typical language used by seniors for the "fetch" task. In particular, the motivation was to see what type of language was used when giving directives to either a robot or another person with the intent of exploring how well these natural directives can be translated into robot commands. As such, the corpus is a central component of a system that we are building for the robot that includes the components of natural language processing (NLP), navigation instruction representation, and robot behavior (for an overview of the system, see Skubic et al., 2013). With the development of this system, the robot will be able to receive natural spatial descriptions and navigate to find the target in a fetch task. Indeed, we have begun to conduct simulations with the robot using templates derived from the natural spatial descriptions from the corpus, in order to assess how well the robot finds the target, as compared to humans who receive the same templates (Skubic et al., 2013).

For the purposes of the current article, we focus on two critical processes within our system: (a) reasoning about the perspective used in the directives, and (b) recognition of furniture items used as spatial references. These processes emerge directly from our analysis of the corpus in Part One. With respect to perspective, there were systematic differences in the perspective adopted by speakers as a function of the addressee. Of relevance here is the finding that while both perspectives were used when speakers talked to robots, they preferred their own perspective (as opposed to the addressee perspective that was preferred for talking to Brian). With respect to landmarks, there was a strong preference for using furniture objects as landmarks rather than the smaller objects on the tables next to the targets. Therefore, the robot will need to identify and differentiate the different furniture items, including the many different tables (four per room). Addressing these challenges to facilitate the interpretation of natural language directives will advance our goal of creating a human–robot interface that establishes a common ground with the user. Our objective is to provide a robot that adapts to the user needs by giving the robot advanced perceptual and reasoning capabilities modeled after those of a human.

Our physical robot is built on a Pioneer 3DX base and uses the Robot Operating System (ROS) (Quigley et al., 2009). The robot's sensors consist of a laser range-finder for obstacle avoidance; the Microsoft Kinect provides perceptual capabilities for recognizing furniture and other items in the scene. The Kinect is positioned at a height of 1 m; both color (RGB) and depth images are used. The sensing capabilities of the Kinect constrain the useable distance and viewing cone for furniture recognition.

The robot system includes the *linguistic processes* for part-of-speech (POS) tagging, chunking, and meaning extraction; *cognitive processes* for reasoning about one's position in the environment, correctly interpreting perspective, and using POS tags to classify words into meaningful categories (e.g., house units and furniture units); and *perceptual processes* for recognizing objects and successfully navigating within the environment. Here, we focus on the development of the cognitive process of determining perspective and the perceptual process of recognizing furniture objects. The natural language processing component is further discussed in Skubic et al. (2013).

3.1. Determining perspective

As shown in Fig. 5, the corpus contains descriptions from an addressee perspective, from a speaker perspective, and indeterminate descriptions for which perspective could not be derived. Although the corpus revealed some preferences for a given perspective in certain contexts, these are not completely predictable. Therefore, the robot strategies will need to reliably determine perspective for each description. Other approaches to this problem include Trafton et al. (2005) cognitive architecture that helps the robot reason about perspective; Berlin, Gray, Thomaz, and Breazeal (2006) use of the teacher's perspective in robot learning by demonstration; and Matuszek, Fox, and Koscher's (2010) reliance on environmental structures to reason about perspective. Our approach makes use of prior knowledge of the environment, landmarks, and context as a means of establishing common ground with the elderly user.

Perspective can be derived within our environment by the use of *left* and *right* as directional commands given that speakers and addressees were offset by 180 degrees and given that the starting position for each trial was the middle hallway in the house (see Fig. 1, Panel A). Our approach requires that two conditions be met. First, the robot requires an approximate map of the environment that minimally includes the entrances to the rooms and their names, for example, living room, bedroom. Ideally, the map should also include a list of possible furniture items in each room. The map does not have to include all of the furniture, although including some of the larger, fixed items (e.g., bed) will improve the efficiency of the fetch. Providing the robot with such a map serves a dual purpose of determining the perspective taken by the speaker as well as speeding up and simplifying furniture recognition (see below). Second, the robot needs to know its own location and orientation on the map, at least in relation to the entrances to the rooms.

For determining perspective, during parsing there is a search within the description for the appearance of a room name and the spatial terms *left* or *right,* with the requirement that they be in the same noun phrase. When found, the robot compares this information with the map and its starting position and orientation. For example, if the speaker told the robot to "*turn right into the bedroom*" and the combination of the map and the robot's position and orientation indicate that the bedroom is on the robot's left, the robot can deduce that the speaker adopted his or her own perspective, and that the robot should indeed turn left to go into the bedroom. The observed differences in the corpus for preferences to use the speaker perspective with the robot, but the addressee perspective with Brian can be incorporated into the strategies as probabilities that factor into the initial commitment that the robot makes in determining perspective. Additionally, the corpus can be used to identify phrases that indicate perspective, and these phrases can be used as features of a particular perspective. For example, if the speaker tells the addressee to "turn around," this is an indication that the speaker wants to align the perspectives so that his or her own perspective can be used.

Obviously, this room-name approach will fail for descriptions that do not contain the name of the room, such as "*go into the room on the right.*" In this case, the furniture items that are included in the description are compared against a list of items typically included or known to be included in each room as a way of guessing which room was intended. This models human reasoning about room purpose. However, this, too, can fail if the furniture items in the description do not clearly indicate the room, or if multiple rooms contain the same furniture items (e.g., table). In this case, it is important that the robot recognize that ambiguity still exists. At this point, the robot will ask for clarification. Since such discourse is time consuming, however, we use it as a last resort if the reasoning steps fail.

3.2. *Furniture recognition*

Given our intent to study the robot fetch task in the physical world, it is important to confront the perceptual challenges placed on the robot for accomplishing the task. This is an important step toward language-based human robot interaction (HRI), as grounding of language is related to human perception (Roy, 2005). We focus here on the recognition of the furniture units, given the high frequency of occurrence of this type of landmark within the corpus across all addressees and all instructions. Furniture could be included in a map; however, some items may be moved, so we do not want to rely on precise, mapped locations. Instead, we need a strategy that recognizes not only the furniture item but also its orientation, given that some descriptions may assume an intrinsic front or back for the furniture (e.g., *in front of the couch*).

In related work, others have proposed language-based HRI approaches that require landmark recognition but have not included recognition strategies (e.g., Chen & Mooney, 2011). Moreover, there is previous work on object recognition using the Kinect system that we adopt here. Lai, Bo, Ren, and Fox (2011) use color and depth images to recognize small objects. Janoch et al. (2011) use the histogram of oriented gradients and size to recognize a variety of objects, including furniture. However, much of this work focuses on recognition only and is not necessarily concerned with execution speed, which is important for timely human–robot interaction. Moreover, the previous work does not generate a model of furniture with spatial information such as position and orientation, and relies on a comprehensive training dataset such that only those specific objects are recognized. In contrast, our approach considers execution speed; determines the orientation of the furniture item along with its recognition, using both depth and color images from the Kinect; and allows type classification for objects outside the training set.

3.2.1. *Furniture recognition methods*

Large objects in the scene are first segmented based on the point cloud generated from the depth image by clustering; the corresponding color information of the points is then extracted from the background. Many furniture items found in the home have a primary horizontal plane, for example, chairs, beds, couches, and tables. To eliminate the effect from the clutter on top of furniture samples, we use the main (horizontal) plane to help to build the furniture model. The main plane is identified using the RANSAC algorithm (Golovinskiy, Kim, & Funkhouser, 2009). Tests indicate it is possible to find the horizontal plane even when surface clutter occupies half of the area. Seven features are used in the furniture classifier:

Furniture size (area of the main plane)

Main plane height (average height of all points in the plane)

Main plane texture (local binary pattern operator (Ojala, Pietikäinen, & Harwood, 1996))

Furniture type (chair-like or table-like, computed based on shape)

Main plane red color proportion, normalized

Main plane green color proportion, normalized

Main plane blue color proportion, normalized

All features are normalized and have an equal weighting. The furniture classification process has two steps. In Step 1, the first four features are used as inputs into a system of fuzzy rules to classify the general type of furniture item, based on the class with the highest membership value. In Step 2, furniture items are further separated by color, with the last three features used with a support vector machine to make the final decision of instance. Thus, a furniture sample that is not contained in the training dataset can still have a type classification based on general shape, which assists the human–robot interaction process.

The confidence of the recognition result is computed from two components: intrinsic confidence, which is determined by the scores of type classification and instance recognition, and extrinsic confidence, which depends on the robot's position with respect to the object. There are three factors in extrinsic confidence: distance, viewing direction, and viewing completeness, that is, based on whether the entire item is in view. The confidence of the recognition result is the mean of these two kinds of confidence. For large furniture items, such as the couch and the bed, the robot is seldom able to view the entire item due to the viewing cone of the Kinect. Therefore, the viewing completeness measurement for these items is relaxed to prevent them from being ignored by the robot due to a low recognition confidence.

Orientation of the furniture item is determined differently for symmetric versus asymmetric objects. For symmetric, table-like objects with no intrinsic front, the closest visible edge is assigned as the front. For rectangular tables, the closest long edge is assigned as the front, based on a previous human subject experiment (Blisard & Skubic, 2005). For chair-like objects, orientation is based on the direction of the chair back relative to the main plane, as shown in Fig. 6.

3.2.2. *Experiments and results*

To test this approach, nine furniture items were selected to represent the items from the virtual house (see Fig. 7). Color and depth images were taken around each object in eight directions and eight distances from 1 to 3 m. Of these 64 images, 48 were used for training and 16 for testing. As a further test, eight images at 1.5 m were collected for a subset of six items on cluttered tables to better represent an unstructured home environment. The furniture recognition process runs in about 9 ms on an Intel core i7 CPU at 1.6 GHz, making the process feasible for real time human–robot interaction. The results

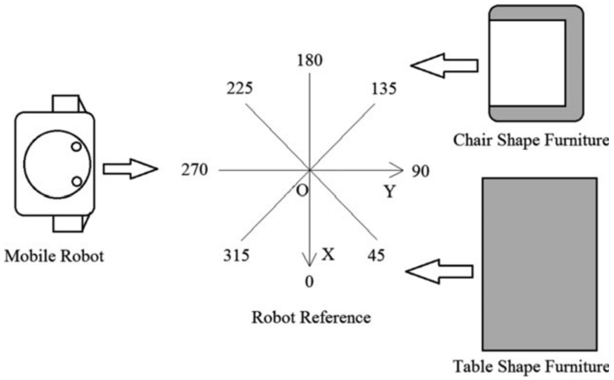

Fig. 6. Chair-shaped objects have an intrinsic front, as shown by the arrow, independent of the robot's relative position. For table-shaped objects, the front is determined by the robot's relative position and viewing perspective. The robot reference axes show the viewing angles used for the test results in Table 4.

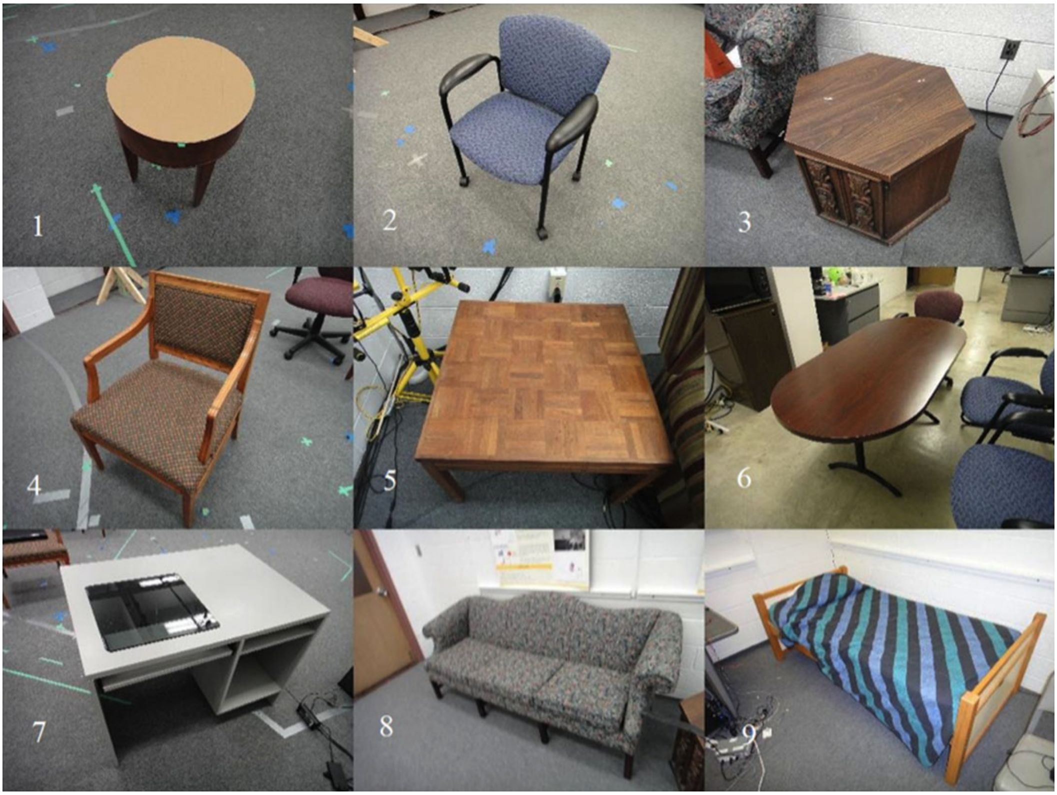

Fig. 7. The nine furniture items tested: (1) round table, (2) blue chair, (3) hexagon table, (4) wood chair, (5) coffee table, (6) dinner table, (7) desk, (8) couch, and (9) bed.

are shown in Table 3 for the test images after training. The recognition results for most of the smaller furniture items are excellent. The larger furniture items of the couch and bed present some challenges, in part due to their size and the sensing limitations of the Kinect. As the distance increases from the Kinect, the resolution of the depth data decreases, resulting in increased uncertainty and reduced recognition. At the same time, the Kinect's viewing cone of 60 degrees may prevent the complete view of larger items at a close distance.

Furniture orientation was also tested using the data from the uncluttered furniture recognition test. Results are shown in Table 4 for the eight directions tested, as error values between the orientation angle detected and the ground truth, in degrees. Objects 1 and 3 (small round table and hexagon table) are excluded from this test due to their general round shape. For the table shaped items that are symmetrical (coffee table, dining table, desk, and bed), an orientation of less than 180 degrees is computed. The results show that orientation is easier to compute for some viewing angles. For the symmetrical objects, including the bed, orientation angle can be determined with low error rates from a range

Table 3
Recognition results for furniture items

Furniture Sample	Without Clutter (%)	With Clutter (%)
1. Round table	100	100
2. Blue chair	100	N/A
3. Hexagon table	100	100
4. Wood chair	87.5	N/A
5. Coffee table	100	87.5
6. Dinner table	100	100
7. Desk	100	100
8. Couch	67.5	N/A
9. Bed	75	N/A

Table 4
Error results of the orientation test for eight directions (in degrees)

Orientation								
Furniture Sample	0	45	90	135	180	225	270	315
1. Round table	×	×	×	×	×	×	×	×
2. Blue chair	47	28	112	25	32	**4**	**1**	**6**
3. Hexagon table	×	×	×	×	×	×	×	×
4. Wood chair	10	35	47	37	12	**2**	**4**	**7**
5. Coffee table	**1**	**0**	**2**	**2**	×	×	×	×
6. Dinner table	**1**	**3**	**1**	**3**	×	×	×	×
7. Desk	**5**	**5**	**1**	**5**	×	×	×	×
8. Couch	48	172	21	51	15	**5**	**5**	**1**
9. Bed	**6**	**2**	**5**	**9**	×	×	×	×

Note: Low error values are shown in bold.

of viewing angles. For the chair-like objects, including the couch, much better results are obtained when viewing from the intrinsic front. This is not surprising, as the shape is not always visible from the back. Thus, both distance and viewing angle can affect recognition and orientation results. Our strategy in the fetch task is to allow the robot to move closer or approach from a different angle, if necessary, to provide a more confident recognition.

4. Conclusions

In this contribution we describe a corpus of spatial descriptions offered by older adults for finding a target within a virtual house environment in the context of a fetch task. We uncovered systematic differences in the word choice, selection of particular landmarks such as furniture items, perspective adopted, and structure, as a function of the addressee and instruction. We then describe the development of a critical cognitive process (determining perspective) and an essential perceptual process (recognizing furniture units) that are informed by the corpus and that in conjunction with the linguistic strategies will enable natural descriptions to be converted into commands that will direct the robot to the target in question.

More generally, the key features of our approach that are informed by the corpus include sensitivity to the addressee and the differential assumptions speakers make about its capabilities; differential willingness to accommodate to the addressee; the task context within which the descriptions are offered (specifying how to find an object vs. specifying where the object is); the likely perspective and the likely structure of the description as a function of addressee and instruction; and the reliance on certain objects in the house (house units and furniture units but not object units) as landmarks.

Our goal is to establish a common ground with the elderly user by making the robot adapt to the user's needs as much as possible. Clark (1996) argues that common ground in language is achieved as a joint activity through the interaction. This is illustrated in the HCRC Map Task corpus (Anderson et al., 1991) for human–human communication and proposed as language games for better robot communication (Steels, 2001). However, the results of our study show that seniors may want a more streamlined communication with a task-oriented robot and do not necessarily want to speak to robots the same way they speak to other people. A discourse with interaction and/or learning can be added if necessary to further establish a common ground between the fetch robot and the elderly user.

In conclusion, the work presented here is part of a larger project that has the goal of developing an intelligent system that uses natural spatial descriptions to direct a robot to a target object's location in a fetch task. To accomplish this task, important capabilities will be required for cognitive, linguistic, and perceptual processing by the robot. This article presents the first step toward developing such a system that adapts to the elderly user's language preferences instead of requiring a specialized language for the robot.

Acknowledgments

This research was supported with funding by the National Science Foundation Grant IIS-1017097. We thank Xiao Ou Li for constructing the virtual house environment and Erin Gibson, Oscar Gonzalez, Diane Garritson, Ashley Herrmann, Mary Johnson, Kevin Kimberly, Kathleen Loftus, Angelique Laboy-Capparopa, Elliott Mitchem, and Joe Wernke for help collecting and coding the data.

References

Anderson, A. H., Bader, M., Bard, E. G., Boyle, E. H., Doherty, G. M., Garrod, S. C., Isard, S. D., Kowtko, J. C., McAllister, J. M., Miller, J., Sotillo, C. F., Thompson, H. S., & Weinert, R. (1991). The HCRC Map Task Corpus. *Language and Speech*, *34*(4): 351–366.

Beer, J. M., Smarr, C., Chen, T. L., Prakash, A., Mitzner, T. L., Kemp, C. C., & Rogers, W. A. (2012). The domesticated robot: design guidelines for assisting older adults to age in place. In *Human-Robot Interaction (HRI), 2012–7th ACM/IEEE International Conference* on (pp. 335–342). IEEE.

Berlin, M., Gray, J., Thomaz, A. L., & Breazeal, C. (2006). Perspective taking: An organizing principle for learning in human-robot interaction. In *Proceedings of the National Conference on Artificial Intelligence* (Vol. 21, No. 2, p. 1444). Menlo Park, CA: AAAI Press.

Blisard S, Skubic M. (2005). Modeling spatial referencing language for human-robot interaction. In *Proceedings, 2005 IEEE International Workshop on Robots and Human Interactive Communication,* Nashville, TN, Aug 13–15, 2005, pp. 698–703.

Borson, S., Scanlan, J., Brush, M., Vitaliano, P., & Dokmak, A. (2000). The Mini-Cog: a cognitive "vital signs" measure for dementia screening in multi-lingual elderly. *International Journal of Geriatric Psychiatry*, *15*(11), 1021–1027.

Carlson, L. A., Skubic, M., Miller, J., Huo, Z., & Alexenko, T. (2013). Assessing the effectiveness of older adults' spatial descriptions in a fetch task. Paper submitted to the 35th Annual Conference of the Cognitive Science Society. Berlin, Germany, July, 2013.

Cassenti, D. N., Kelley, T. D., Avery, E., & Yagoda, R. E. (2011). Location label speech options improve robot operator performance. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *55*, 439–443.

Chen, D. L., and Mooney, R. J. 2011. Learning to interpret natural language navigation instructions from observations. In Association for the Advancement of Artificial Intelligence (AAAI), pp. 128–135, Cambridge, MA.

Clark, H. H. (1996). *Using language*. Cambridge, UK: Cambridge University Press.

Clark, H. H., Schreuder, R., & Buttrick, S. (1983). Common ground and understanding demonstrative reference. *Journal of Verbal Learning and Verbal Behavior*, *22*, 245–258.

Davis, R. L., Therrien, B. A., & West, B. T. (2009). Working memory, cues, and wayfinding in older women. *Journal of Applied Gerontology*, *28*(6), 743–767.

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 4th conference on Designing interactive systems: processes, practices, methods, and techniques* (pp. 321-326). ACM.

Ezer, N. (2008). Is a robot an appliance, teammate, or friend? Age-related differences in expectations of and attitudes toward personal home-based robots. Dissertation, Georgia Institute of Technology.

Golovinskiy, A., Kim, V. G., & Funkhouser, T. (2009). Shape-based recognition of 3D point clouds in urban environments. In *Computer Vision, 2009 IEEE 12th International Conference on* (pp. 2154-2161). IEEE.

Gribble, W., Browning, R., Hewett, M., Remolina, E., & Kuipers, B. (1998). Integrating vision and spatial reasoning for assistive navigation, in assistive technology and artificial intelligence. In V. Mittal, H. Yanco, J. Aronis, & R. Simpson (Eds.), *Lecture Notes in Computer Science* (pp. 179–193). Berlin: Springer-Verlag, 1998.

Janoch, A., Karayev, S., Jia, Y., Barron, J. T., Fritz, M., Saenko, K., & Darrell, T. (2011). *A category-level 3-D object dataset: Putting the kinect to work*. Barcelona, Spain: ICCV Workshop on Consumer Depth Cameras in Computer Vision.

Klippel, A., & Montello, D. R. (2007). Linguistic and nonlinguistic turn direction concepts. In S. Winter, B. Kuipers, M. Duckham, & L. Kulik (Eds.), Spatial information theory. In *Proceedings, 9th International Conference, COSIT 2007,* Melbourne, Australia, September 19-23, 2007 (pp. 354–372). Berlin: Springer.

Kuipers, B. (2000). The spatial semantic hierarchy. *Artificial Intelligence*, *119*, 191–233.

Lai, K., Bo, L., Ren, X., & Fox, D. (2011). Sparse distance learning for object recognition combining rgb and depth information. In *Robotics and Automation (ICRA), 2011 IEEE International Conference on* (pp. 4007-4013). IEEE.

Landau, B., & Jackendoff, R. (1993). What and where in spatial language and spatial cognition. *The Behavioral and Brain Sciences*, *16*, 217–265.

Levelt, W. J. M. (1996). Perspective taking and ellipsis in spatial descriptions. In P. Bloom, M. A. Peterson, L. Nadel, & M. Garrett (Eds.), *Language and space* (pp. 77–108). Cambridge, MA: MIT Press.

Levinson, S. G. (1996). Frames of reference and Molyneux's question: Crosslinguistic evidence. In P. Bloom, M. A. Peterson, L. Nadel, & M. Garrett (Eds.), *Language and space* (pp. 109–169). Cambridge, MA: MIT Press.

Levit, M., & Roy, D. (2007). Interpretation of spatial language in a map navigation task. *IEEE Transactions on Systems, Man, and Cybernetics. Part B, Cybernetics*, *37*(3), 667–679.

Mainwaring, S., Tversky, B., Ohgishi, M., & Schiano, D. (2003). Descriptions of simple spatial scenes in English and Japenese. *Spatial Cognition and Computation*, *3*, 3–42.

Matuszek, C., Fox, D., & Koscher, K. (2010). Following directions using statistical machine translation. In *Proceedings of the 5th ACM/IEEE international conference on Human-robot interaction* (pp. 251-258). IEEE Press.

Muller, R., Rofer, T., Landkenau, A., Musto, A., Stein, K., & Eisenkolb, A. (2000). Coarse Qualitative Descriptions in Robot Navigation, in Spatial Cognition II. In C. Freksa, W. Braner, C. Habel, & K. Wender (Eds.), *Lecture Notes in Artificial Intelligence* (pp. 265–276). Berlin: Springer-Verlag.

Ojala, T., Pietikäinen, M., & Harwood, D. (1996). A comparative study of texture measures with classification based on feature distributions. *Pattern Recognition*, *29*, 51–59.

Plumert, J. M., Carswell, C., DeVet, K., & Ihrig, D. (1995). Content and organization of communi- cation about object location. *Journal of Memory and Language*, *34*, 477–498.

Quigley, M., Conley, K., Gerkey, B., Faust, J., Foote, T., & Leibs, J., & Ng, A. Y. (2009). ROS: an open-source Robot Operating System. In *ICRA workshop on open source software* (Vol. 3, No. 3.2).

Raven, J. C., Court, J. H., & Raven, J. (1977). *Manual for raven's progressive matrices and vocabulary scales*. London: Lewis.

Roy, D. (2005). Grounding words in perception and action: computational insights. *Trends in Cognitive Sciences*, *9*(8), 389–396.

Schober, M. F. (1993). Spatial perspective-taking in conversation. *Cognition*, *47*(1), 1–24.

Scopelliti, M., Giuliani, M. V., & Fornara, F. (2005). Robots in a domestic setting: A psychological approach. *Universal Access in the Information Society*, *4*, 146–155.

Skubic, M., Huo, Z., Alexenko, T., Carlson, L., & Miller, J. (2013). Testing an assistive fetch robot with spatial language from older and younger adults. In review for 2013 IEEE RO-MAN: The 22nd IEEE International Symposium on Robot and Human Interactive Communication, (ROMAN13, August 26-29, 2013, Gyeongju, Korea).

Steels, L. (2001). Language games for autonomous robots. *Intelligent Systems, IEEE*, *16*(5), 16–22.

Talmy, L. (1983). *How language structures space*. In H. L. Pick Jr, & L. P. Acredolo (Eds.), *Spatial orientation: Theory, research and application* (pp. 225–282). New York: Plenum.

Tellex, S., & Roy, D. (2006). Spatial routines for a simulated speech-controlled vehicle. In *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction* (pp. 156-163). ACM.

Tenbrink, T., Fischer, K., & Moratz, R. (2002). Spatial strategies in human-robot communication. *KI*, *16*(4), 19–23.

Trafton, J. G., Cassimatis, N. L., Bugajska, M. D., Brock, D. P., Mintz, F. E., & Schultz, A. C. (2005). Enabling effective human-robot interaction using perspective-taking in robots. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, *35*(4), 460–470.

Wechsler, D. (1997). *WAIS-III: Administration and scoring manual: Wechsler adult intelligence scale*. Psychological Corporation.

Appendix

Target Object	Reference Object 1	Reference Object 2
Notepad	Video game	Kleenex
Book	Flower	Fan
Letter	Laptop	Tray
Mug	Purse	Hat
Cell phone	Vase	Candle
Wallet	Lamp	Plant
Keys	Monitor	Frame
Glasses case	Alarm clock	Statue