Assessing the effectiveness of older adults' spatial descriptions in a fetch task

Laura A. Carlson[1] (lcarlson@nd.edu)
Marjorie Skubic[2] (skubicm@missouri.edu)
Jared Miller[1] (jmille39@nd.edu)
Zhiyu Huo[2] (zhiyuhuo@mail.missouri.edu)
Tatiana Alexenko[3] (ta7cf@mail.missouri.edu)
[1]Department of Psychology, 119-D Haggar Hall, University of Notre Dame, Notre Dame, IN 46556 USA
[2]Electrical and Computer Engineering Department, University of Missouri, Columbia, MO
[3]Computer Science Department, University of Missouri, Columbia, MO

Abstract

The current paper examines spatial descriptions provided by older adults in the context of a fetch task in a virtual house environment that mimics an eldercare setting. Sixty-four older adults provided directions for how to find a target or where to find a target to a robot or human (named Brian) avatar. There were systematic differences in the form and structure of the descriptions based on the communicative task. Specifically, *how* descriptions were longer, contained more detail, and were dynamically structured as compared to *where* descriptions. However, *where* descriptions were found to be more effective in conveying the target location, as assessed with a subsequent target selection task. Implications for the development of robot algorithms for the comprehension of naturalistic spatial language across these two communicative tasks are discussed.

Keywords: Human-robot interaction (HRI); spatial language; dynamic and static; how and where; effectiveness; fetch task; assistive robotics; eldercare.

Introduction

An emerging line of research in human-robot interaction involves the development of assistive devices for use in eldercare settings, either as social companions (e.g., Heerink et al., 2008; Kidd, Taggart, & Turkle, 2006; Libin & Cohen-Mansfield, 2004; Shibata, Kawaguchi, & Wada, 2011; Wada et al., 2003) or as task-oriented robots assisting with navigation (Montemerlo et al., 2002), managing medication (Tiwari et al., 2011)], and providing reminders (Pollack et al., 2002). Older adults also report wanting help with tasks such as cleaning, heavy lifting, and fetching objects (Beers et al., 2012). They also prefer to speak naturally to these assistive devices, rather than use a more constrained interface (Scopelliti, Guiliani, & Fornara, 2005).

To accommodate these preferences, recently we gathered a corpus of spatial descriptions from older adults who interacted with an avatar within a virtual house setting in the context of a fetch task. Our primary goal in this project is to develop robot algorithms for the online comprehension of these natural language spatial descriptions and to test these algorithms in an analogous physical environment with a physical robot. In working toward this goal, on the basis of the corpus, we have identified key components that need to be developed for the robot including speech recognition for older adults (Alexenko et al., 2013), parsing the natural language descriptions and coding them into chunks that can be converted into robot commands, recognizing key furniture items within a cluttered environment that are included in the descriptions, and identifying spatial relations within the horizontal plane (e.g., behind the couch) and the vertical plane (e.g., on top of the table) (Skubic et al, 2012).

Given that the robot algorithms are driven by the properties of the spatial descriptions, in the current paper we examine how the communicative task of the speaker impacts the features of the descriptions, and present data that reflect the effectiveness of the descriptions.

Spatial Directions and Spatial Descriptions

A fetch task is one in which a speaker specifies the location of a desired target for an addressee whose goal is to retrieve the target. There are two ways in which the location can be indicated by the speaker. The speaker could provide directions that tell *how* to get to the target location or the speaker could provide descriptions that specify information about *where* a given target location is. Research has shown systematic differences in the type and structure of the language that is used for each of these communicative tasks. For example, Plumert et al. (1995) found that written directions on how to find a target in a hierarchically organized doll-house environment were more likely to provide more detailed messages and contain more spatial units that tended to be organized in a descending sequence (floor → room → reference object. e.g., *The keys are on the first floor in the living room on the table*.) as compared to written descriptions of where to find a target that were less detailed and organized in an ascending sequence (reference object → room → floor, e.g., *The keys are on the table in the living room on the first floor*).

This distinction between *how* and *where* has also been characterized as *dynamic* and *static* (Wahlster. 1995, Fasola and Mataric 2012) spatial language, respectively, with dynamic stepping the addressee through the environment in a point by point fashion and *static* offering spatial information that does not embed the addressee in the environment. Dynamic spatial directions are also inherently sequential, while static descriptions are not. Nevertheless, static descriptions are often overlooked or treated the same

as dynamic directions by other researchers (Tellex et al, 2011), perhaps due to the focus on two-dimensional route instructions or the assumption that dynamic descriptions are better or more prevalent (Kollar et al., 2010, MacMahon et al., 2006, Vogel and Jurafsky 2010, Shimizu and Haas 2009).

In the current work, we assess two questions related to this how/where distinction: First, we ask whether there are consistent differences in the type and form of the spoken spatial language that is produced by older adults in response to *how* and *where* instructions that might echo Plumert et al's (1995) findings with written spatial language. We focus on the type of language included in the descriptions, and the amount of detail, and ignore the hierarchical sequencing that Plumert et al. (1995) measured, because our environment consists of a single floor, as intended for mimicking an eldercare setting. Second, we ask whether these differences are associated with differences in the relative effectiveness of the descriptions.

Corpus of descriptions from older adults in a virtual fetch task

Our corpus consists of 512 spatial descriptions collected from 64 older adults (mean age = 76 years) who specified the location of 8 targets embedded in the virtual house environment shown in Figure 1. Targets were placed in the living room (on the left in Figure 1) and bedroom (on the right in Figure 1) on tables that also contained two other objects that could potentially serve as reference objects. On each trial, older adults explored the virtual house with the assistance of an experimenter and found a designated target. They were then positioned in the central hallway (marked by "Start" in Figure 1), and provided a description

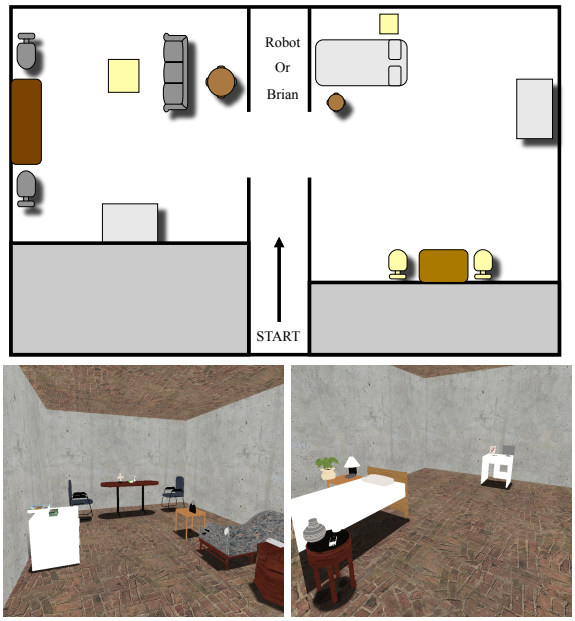

Figure 1: Overview and screen shots of the virtual house

of the target location to either a robot or human avatar (named Brian) who faced them (as indicated by the arrow in Figure 1), such that their perspectives were misaligned by 180 degrees. Previous research has shown a preference for speakers to use an addressee perspective when perspectives between speaker and addressee are misaligned (Mainwaring et al., 2003; Schober, 1993), with such preference also observed for robot addressees (Tenbrink et al., 2002). However, given that older adults have shown negative emotional responses to robots (Scopelliti et al, 2005), we included the addressee manipulation to assess whether older adults in particular would be more likely to adopt their own perspective rather than the perspective of the robot. These perspective results are presented in Carlson et al. (2013). A second manipulation was related to the task instructions. Specifically, older adults were instructed to either provide directions for **how** to find the target or to provide descriptions of **where** the target was located. Both the addressee manipulation (robot or Brian) and the task instruction manipulation (how or where) were between subject manipulations, with the consequence that 16 older adults each provided 8 descriptions (128) for each of the 4 addressee X instruction combination (128 X 4 = 512 descriptions in total).

A full report of the older adult corpus can be found in Carlson et al. (2013). We focus here on the *how* versus *where* differences. Figure 2 provides the task instructions (adapted from Plumert et al., 1995), and examples from the corpus.

	How	**Where**
Instructions to participants	Please tell the *Brian/the robot* how to find the target.	Please tell *Brian/the robot* where the target is.
Addressee: Brian Target 1 = Book	Brian turn to your right. Go to the room on your right and go straight ahead and the book is right there before you.	Brian the book is in the room to your right and its on a table at the far side of the room.
Target 2 = Cell phone	Brian please take 1 or 2 steps forward and enter the room on your left. Upon entering the room take about 3 steps forward and you'll see a small brown table with a black top. The cell phone that you are looking for is on that table.	The cell phone is in the bedroom on the bedside table.
Addressee: Robot Target 1 = Book	Enter the room on your right and continue straight until you hit a brown covered table and on table of the table is the book.	The book is in the living/dining room on the dining room table.
Target 2 = Cell phone	Enter the room on your immediate left and by the little nightstand next to the bed there it is on top of it. It's a round night table and the cell phone is in the center.	The cell phone was on the left side on a round table near the bed.

Figure 2: Instructions and sample descriptions by how/where and addressee

As shown in Table 1, *how* descriptions contained more words overall per description, and included more spatial terms (such as "on", "to" and "right") and more hedges (such as "immediately" and "slightly"). In contrast, *where* descriptions contained more house units (such "room", "door" and "wall"). Descriptions often contained large furniture items in the rooms (such as "bed" and "couch"), and rarely contained reference objects that were collocated

on the tables (such as "lamp"), with the incidence of these categories not varying across *how* and *where* descriptions. Finally, *how* descriptions were more likely to have a dynamic form than the *where* descriptions.

	How	Where
Words per description	M = 27.0, SE = 1.8	M = 19.4, SE = 1.5
Spatial terms per description	M = 4.1, SE = .26	M = 2.3, SE = .29
House units per description	M = 1.1, SE = .10	M = 1.4, SE = .12
Hedges per description	M = .22, SE = .05	M = .09, SE = .02
% dynamic descriptions	M = 95.3, SE = 3.3	M = 35.8, SE = 7.9

Table 1: Significant differences between how and where older adult descriptions

These results are consistent with Plumert et al. (1995) who also found that *how* descriptions contained more spatial units and were more detailed than *where* descriptions. What remains unclear is whether these differences are associated with any differences in effectiveness. That is, these descriptions were all collected by older adult speakers in the context of a fetch task in which accurately specifying the location of the target is critical for the success of the task. We ask next whether *how* or *where* descriptions are more effective in identifying the location of the target.

Differences in how vs. where effectiveness

To assess effectiveness, we randomly selected from the corpus of older adult descriptions two from each speaker's set of 8 descriptions (with half of the speakers addressing Brian and half the robot), with the constraint that the location of each target was specified an equal number of times across the set of descriptions that we were assessing. These descriptions were then provided to sixty-four younger adults to assess effectiveness. Their task was to listen to a description without the target, navigate through the house in accordance with the description, and then guess the identity of the target. 4 targets were placed on tables in the living room and 4 targets were placed on tables in the bedroom. Each table contained a target and two distractor objects. Each participant performed two trials (one in the living room and one in the bedroom). Before the trials began, the younger adults were shown a video tour of the house that did not include the targets. This was to familiarize them with the house environment and the relative locations of the rooms and their contents. On each trial, participants started in the hallway of the house, standing in the location and at the orientation of the avatar, as specified by the label "robot or avatar" shown in Figure 1, and facing the original speaker's location (which is marked in Figure 1 with the label "Start" and with an orientation specified by the arrow). They were therefore facing the position of the participants from which the descriptions were gathered. A participant was given a description from the corpus with the target item removed, and they navigated through the house until they thought they found the target, and then named it. The key dependent measure was their accuracy in selecting the target.

As shown in Figure 3, we examined two indicators of this accuracy: selection of the correct target, and selection of the correct table on which the target appeared. This latter measure is important because two potential reference objects appeared on the tables next to the targets, and often the descriptions did not provide enough information to identify which object on the table was the target (see example descriptions in Figure 2). The infrequent use of the reference objects that appeared next to the target is consistent with Plumert et al. (1995) who found that such reference objects were only consistently used when the target was located on the reference object as opposed to beside it.

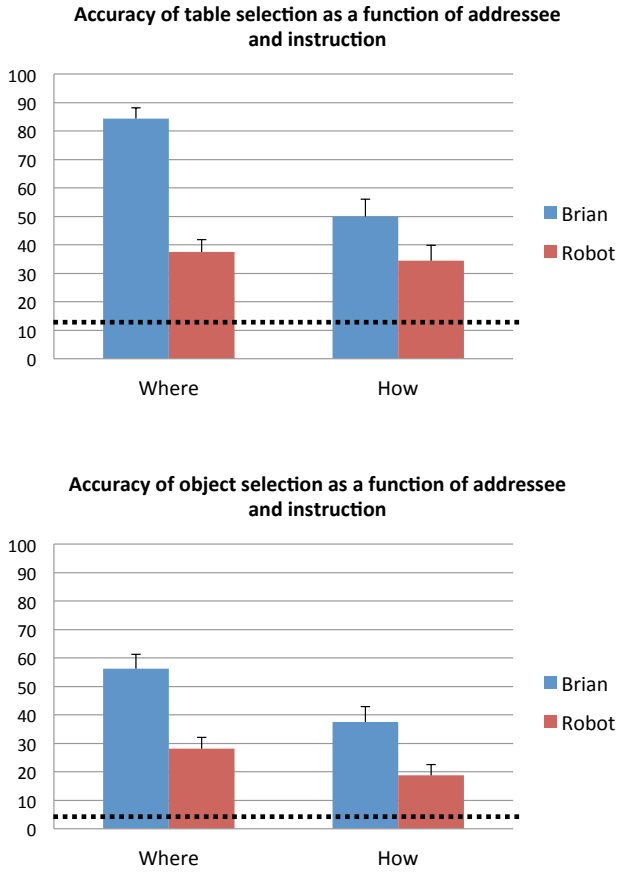

Figure 3: Selection accuracy for correct table (top) and correct target (bottom) as a function of how/where and addressee. Dotted line indicates chance selection.

With respect to selection of the correct table, performance in all conditions was significantly above chance performance of 12.5, based on 8 possible tables in the environment. In addition, significantly better performance was observed for "where" descriptions (M = 61%) than for "how" descriptions (M = 42%), $F(1,60) = 4.51$, $p < .05$. In addition, there was a significant effect of addressee, with more accurate performance for descriptions provided to Brian (M = 67%) than to the robot (M = 36%), $F(1,60) =$

12.55, $p < .01$. The interaction between instruction and addressee was marginal, $F(1,60) = 3.13$, $p = .08$.

With respect to the selection of the correct object, performance in all conditions was significantly above chance performance of 4.2 (based on 24 possible targets in the environment (3 on each of 8 tables)). For this analysis, there was only a main effect of addressee: $F(1,60) = 7.10$, $p < .05$; the effect of instruction and the interaction were not significant ($Fs < 1.6$, $ps < .21$).

For the object selection measure, we also assessed how likely it was that participants selected the correct object, given that they selected the correct table. Chance performance in this case is 33%, given that there are three objects (target and two reference objects on each table).

Figure 4 shows that in all conditions, accuracy was significantly above chance. We expect that this is because the target objects were generally smaller than the reference objects on the tables. Clark, Schreuder and Buttrick (1983) argue that when a reference is under-determined by a speaker, the addressee will select an object from a group of

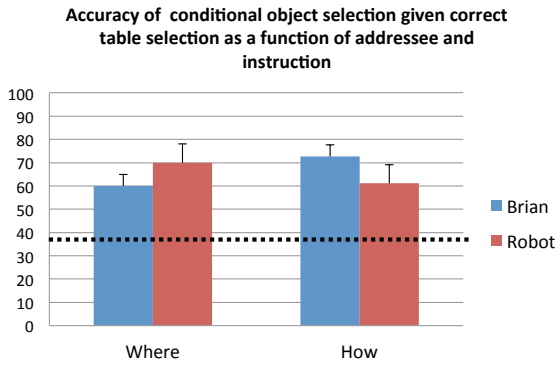

Figure 4: Selection accuracy for the correct object, conditional on correct table selection. Dotted line indicates chance selection.

objects that offers the most contrast from the others along a given dimension. For example, imagine a speaker tells an addressee to pick up a ball and refers to a collection of three balls (a golf ball, a squash ball and a basketball) that are placed on a table in front of them. Clark et al. argue that it is likely that the addressee will select the basketball because it is the most unique item in the set, standing out in terms of size.

Finally, we also examined whether there were differences in accuracy for the individual targets. Given that each of the targets appeared in a given location (and location was not counterbalanced across targets), this serves an indicator as to whether any of the target locations were particularly difficult to describe and find. Figure 5 shows accuracy as a function of the targets, both as indicated by the correct selection of the table and correct selection of the object.

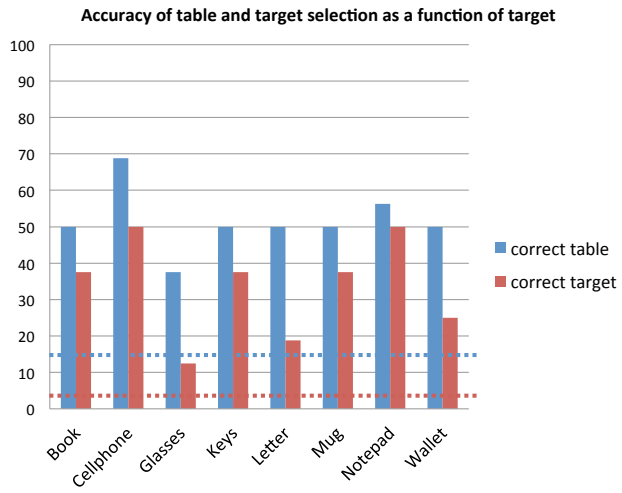

Figure 5: Selection accuracy for table and target selection as a function of target. Dotted lines indicate chance selection.

Chi-square analyses revealed no significant differences among targets, for either the correct table accuracy or the correct target accuracy. This indicates that there were not any targets or locations that were particularly difficult to describe and/or find. This is likely due to the simple layout and relatively impoverished contents of the rooms in the virtual house.

Overall, the results for the assessment of the older adults spatial descriptions indicate that the *where* descriptions allowed participants to more easily select the target and its table than the *how* descriptions. We suspect that the differences in accuracy for table selection as a function of addressee that were observed are likely due to other properties of the descriptions, such as the perspective adopted by the speaker. For a full report of the older adult corpus, see Carlson et al. (2013).

Conclusions

Together, the detailed analysis of the corpus and the results of the experiment assessing the effectiveness of the descriptions point to an interesting contrast. On the one hand, the corpus analysis reveals that *how* descriptions are longer, offer more detail, include more spatial terms, and are dynamic, as compared to the *where* descriptions that are shorter, include more references to house structures, and are often static. On the other hand, these same *how* descriptions are not as effective in communicating the location of the target, as assessed by the accuracy for selecting the target and its table. We are currently comparing the effectiveness of these older descriptions with the effectiveness of a corpus of descriptions collected from younger adults within the same virtual environment. Moreover, we are also recording the paths that participants take to the target in response to these descriptions, with the idea that the paths may offer an additional online measure of effectiveness. Metrics we are

examining include path length, navigation speed, number of pauses, and changes in heading.

These results also have several interesting implications for the development of robot algorithms in this task. For example, it may be beneficial for the robot algorithm to initially classify a description as one that is conveying directions or one that is conveying location, given that the form and content of the descriptions vary as a function of communicative task. In a natural setting, of course, the speaker may not be explicit about whether he or she is providing directions or specifying location (that is, the speaker is not assigned a *how* or *where* task per se, as in our current work). This classification would need to be based on the properties of the descriptions themselves.

In addition, the robot algorithms will need to take into account the differential effectiveness of the two types of descriptions. The "how" descriptions may provide a more explicit approach to allow direct translation into robot commands; however, varying viewing perspectives will complicate the interpretation. To follow the directions of the "how" descriptions, a robot does not rely as much on perception, which may improve the efficiency of the fetch navigation in some static environments but not necessarily the effectiveness. In contrast, the "where" descriptions provide more hints using reference structures and objects so that the robot can navigate to the target using perception. The "how" descriptions may be easier to interpret but have a lower probability to navigate the robot to the specified target, especially given a dynamic environment in which reference furniture items have been moved. The "where" descriptions require the challenge of translating them into navigation commands but may provide more reliable fetch results, even in the case of moved reference items.

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

Beer, J., Smarr, C.A., Chen, T.L., Prakash, A., Mitzner, T.L., Kemp, C.C., and Rogers, W.A. (2012). "The Domesticated Robot: Design Guidelines for Assisting Older Adults to Age in Place," *Proc., IEEE Conference on Human Robot Interaction - Session: Living and Working with Service Robots*, Boston, MA, pp. 335-342.

Carlson, L. A., Skubic, M., Miller, J., Huo, Z., & Alexenko, T. (2013). Strategies for human-driven robot comprehension of spatial descriptions by older adults in a robot fetch task. Revised manuscript submitted for publication in special issue of *Topics*.

Clark, H. H., Schreuder, R., & Buttrick, S. (1983). Common ground and the understanding of demonstrative reference. *Journal of Verbal Learning and Verbal Behavior, 22*, 245-58.

Fasola, Juan, and Maja J. Matarić. "Using Spatial Language to Guide and Instruct Robots in Household Environments." (2012), AAAI Technical Report, FS-12-07.

Heerink, M., Kröse, B., Wielinga, B., and Evers, V. (2008). "Enjoyment, Intention to Use and Actual Use of a Conversational Robot by Elderly People," *Proc., Intl. Conf. on Human-Robot Interaction*, pp. 113-119.

Kidd, C., Taggart, C.W., and Turkle, S. (2006). "A Sociable Robot to Encourage Social Interaction among the Elderly," *Proc., IEEE Intl Conf on Robotics and Automation*, pp. 3972-3976.

Kollar, T., Tellex, S., Roy, D., and Roy, N. (2010). "Toward Understanding Natural Language Directions," *Proc., 5th ACM/IEEE International Conference on Human-Robot Interaction*, pp. 259.

Libin, A. and Cohen-Mansfield, J. (2004). "Therapeutic Robocat for Nursing Home Residents with Dementia: Preliminary Inquiry," *American Journal of Alzheimer's Disease and Other Dementias*, 19(2):111-116.

MacMahon, M., Stankiewicz, B., and Kuipers, B. (2006). "Walk the Talk: Connecting Language, Knowledge, and Action," *Route Instructions*, pp 1475-1482.

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

Montemerlo, M., Pineau, J., Roy, N., Thrun S., and Verma, V. (2002). "Experiences with a Mobile Robotic Guide for the Elderly," *Proc., AAAI-02*, pp. 587-592.

Plumert, J. M., Carswell, C., DeVet, K., and Ihrig, D. (1995). "The Content and Organization of Communication about Object Locations," *Journal of Memory & Language*, 34:477-498.

Pollack, M.E., Brown, L., Colbry, D., Orosz, C., Peintner, B., Ramakrishnan, S., Engberg, S., Matthews, J.T., Dunbar-Jacob, J., and McCarthy, C.E. (2002) "Pearl: A Mobile Robotic Assistant for the Elderly," *Proc., AAAI Workshop on Automation as Eldercare*.

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

Scopelliti, M., Giuliani, V., and Fornara, F. (2005). "Robots in a Domestic Setting: A Psychological Approach," *Universal Access in the Information Society, 4,* 146-155.
Scopellite, Guiliani, & Fornara, 2005

Shibata, T., Kawaguchi, Y., and Wada, K. (2011). "Investigation on People Living with Seal Robot at Home - Analysis of Owners' Gender Differences and Pet Ownership Experience," *International Journal of Social Robotics, 4(1)*:53-63.

Shimizu, N. and Haas, A. (2009). "Learning to Follow Navigational Route Instructions," *Proc., Intl. Joint Conf. on Artificial Intelligence*, pp.1488-1493.

Skubic, M., Alexenko, T., Huo, Z., Carlson, L., and Miller, J. (2012). "Investigating Spatial Language for Robot Fetch Commands," *Grounding Language for Physical Systems, AAAI Technical Report, WS-12-07*, pp. 39-45.

Tellex, S., Kollar, T., Dickerson, S., Walter, M., Banerjee, A., Teller, S. and Roy, N. (2011). "Understanding Natural Language Commands for Robotic Navigation and Mobile Manipulation," *Proc., Conf. on Artificial Intelligence* (AAAI).

Tenbrink, T., Fischer, K., & Moratz, R. (2002). Spatial strategies of human-robot communication." KI #4 [Online]. Available: http://citeseerx.ist.psu.edu/ viewdoc /summary?doi=10.1.1.58.2151

Tiwari, P., Warren, J., Day, K.J., and Datta, C. (2011). "Comprehensive Support for Self-Management of Medications by a Networked Robot for the Elderly" *Proc., Health Informatics New Zealand 10th Annual Conference and Exhibition*, Aukland.

Vogel, A. and Jurafsky, D. (2010). "Learning to Follow Navigational Directions," *Proc., 48th Annual Meeting of the Association for Computational Linguistics*, pp. 806-814.

Wada, K., Shibata, T., Saito, T. and Tanie, K. (2003). "Psychological and social effects of robot assisted activity to elderly people who stay at a health service facility for the aged," in *Proc*., *Intl. Conf. on Robotics and Automation*, pp. 3996–4001.

Wahlster, W., Statiques Et Dynamiques, and Gerd Herzog. "Coping with Static and Dynamic Spatial Relations." (1995).