Grounding Language for Physical Systems
AAAI Technical Report WS-12-07

Investigating Spatial Language for Robot Fetch Commands

Marjorie Skubic[1], Tatiana Alexenko[2], Zhiyu Huo[1] Laura Carlson and Jared Miller
[1]Electrical and Computer Engineering Dept.
[2]Computer Science Dept.
University of Missouri
SkubicM@missouri.edu

Dept. of Psychology
University of Notre Dame
LCarlson@nd.edu

Abstract

This paper outlines a study that investigates spatial language for use in human-robot communication. The scenario studied is a home setting in which the elderly resident has misplaced an object, such as eyeglasses, and the robot will help the resident find the object. We present results from phase I of the study in which we investigate spatial language generated to a human addressee or a robot addressee in a virtual environment and highlight differences between younger and older adults. Drawn from these results, a discussion is included of needed robot capabilities, such as an approach that addresses varying perspectives used and recognition of furniture items for use as spatial references.

Introduction

Recent studies have shown that one of the top five tasks noted by seniors for assistive robots is help with fetching objects, for example, retrieving missing eyeglasses (Beer et al., 2012). In addition, the most preferred domestic robot interface is natural language (Scopelliti et al., 2005). In this paper, we present an overview and initial results for a project designed to address the fetch task and study appropriate language that allows users to communicate naturally and effectively with a robot.

When people communicate with each other about spatially oriented tasks, they typically use relative spatial references rather than precise quantitative terms, e.g., *the eyeglasses are in the living room on the table in front of the couch* (Carlson and Hill, 2009). Here, we explore this type of spatial referencing language. A human subject experiment was performed, studying first college-age students and then adults over age 64 for comparison. The study was conducted in a virtual environment (VE), which provides a controlled setting and is easier for manipulating test conditions. In pilot work, the use of a VE was shown to have sufficient sensitivity to detect differences in test groups and also replicated key findings from work done in physical environments (Schober, 1995). Later studies are planned with robots in the physical world.

Human Subject Experiment

The first phase of human subject experiments has been completed with younger and older adults. We investigated the type of spatial language used naturally by participants when addressing either a human (called Brian) or a robot avatar. The VE included three rooms – a central hallway with a living room and a bedroom (Figure 1).

Fig. 1. *The virtual scene used for the experiments, showing the robot avatar in the hallway with the living room on the left and the bedroom on the right.*

Each participant begins with a brief video illustrating the room layouts. At this point, candidate reference objects are shown but no target objects are included in the scene. The participant is then asked to explore the scene, to look for a specified target object which is now included in the VE. Eight target objects are used for the study: a book, cell phone, eyeglasses case, keys, letter, mug, notepad, and wallet; each participant has eight trials, one for each target object. After locating the target object, the participant is brought back to the hallway, facing the avatar, and is asked to give instructions to the avatar on the location of the target object. Two test conditions were used to compare 64 younger to 64 older adults: (1) the addressee, either human or robot, and (2) the instruction given to the participant, either tell the addressee *where* to find the target object or *how* to find it. The descriptions given by the participants were recorded, transcribed, and coded for analysis as

follows: perspective taken (self or addressee), type of description: static vs. dynamic language, number of spatial phrases, reference object selected, type of spatial term, and use of spatial hedges (e.g., kind of near the middle of the room). Partial results are included here; additional results can be found in (Carlson et al, in review).

The issue of perspective taken (self or addressee) is especially pertinent to a robot interpreting spatial descriptions. Other work suggests a preference for the addressee perspective in human-human communications (Mainwaring et al., 2003) and human-robot communications (Tenbrink, Fischer & Moratz, 2002).

Figure 2 shows the results of the perspectives taken in the study; the how and where conditions are combined here to highlight the differences between younger and older adults. The younger adults preferred the addressee perspective when facing either the human (Brian) or robot avatar, consistent with previous work. The older adults preferred the addressee perspective when speaking to Brian; however, when speaking to the robot, they used a self perspective more often than the addressee perspective.

Figure 2 also shows the usage of ambiguous perspective, e.g., the keys are on the table in the bedroom. In this case, there are no cues to determine which perspective was used. Interestingly, younger adults almost never used this type of ambiguous language, whereas older adults often did. In general, older adults used significantly fewer words (24 vs. 28 per description), fewer spatial units, and fewer reference objects than younger adults, preferring instead to use a more concise description.

Figure 3 shows examples of descriptions, highlighting differences between older and younger adults for the *Robot, Where* test condition. Note the dynamic language in the last example (go to the room…). Although the *How* condition resulted in more dynamic descriptions compared to the *Where* condition, a dynamic structure was observed in both conditions (about half for the younger adults).

It is not clear why there are differences between younger and older adults in perspective, detail, or language structure. The study participants were explicitly shown the front side of the robot and told what the front was so it should have been clear that they were addressing the robot face to face. (Additional cues on the robot could clarify this for future studies.) We also looked for signs that the older adults might not have learned or remembered an accurate detailed map of the rooms; however, it does not appear that the elderly participants captured a less detailed mental model of the environment.

We were also interested in investigating spatial references used in the descriptions. Overall, there were very few references given to small objects, i.e., candidate reference objects purposely placed on horizontal surfaces around the rooms. For each target object, there were two candidate reference objects placed nearby on the same

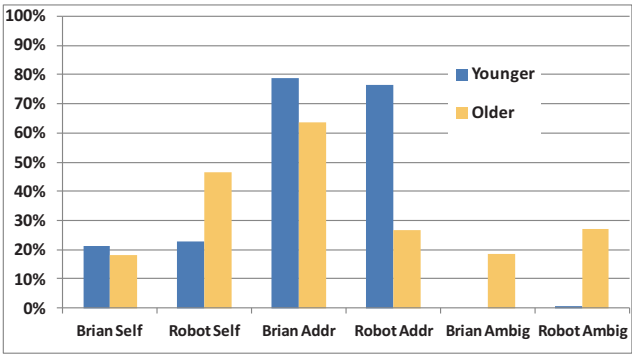

***Fig. 2**. Perspectives taken for younger and older adults*

Older Adults
- The wallet is in the bedroom on the bedside table
- The notepad is in the livingroom on the desk
- On the table in the livingroom in the back of the sofa

Younger Adults
- The wallet is in the room on your right around the bed and on the bedside table
- The notepad is in the room on your right. Walk in and its on the white...dresser to your left next to an empty box of kleenex.
- Go to the room on your right go past the couch… behind the couch there's an end table… the mug is on the end table

***Fig. 3.** Sample Descriptions for Robot, Where*

surface that could have been used in the description. However, these objects generally were not used. There were references made to larger units, both furniture items and house units, as shown in Figure 4. The most popular references were made to *room* and *table*. All target objects were located on some type of table, so this is not surprising. The many references to room instead of bedroom or living room show further ambiguity challenges in interpreting the spatial descriptions.

To perform this type of fetch task efficiently, the robot will need to be able to understand which room the speaker indicates, which is complicated by the varying perspectives used. The robot will also need the capability to recognize objects in the scene, especially furniture items and room units and understand spatial relationships between them. In the remaining paper, we discuss our approach for addressing these challenges.

Spatial Language Processing

Natural language, even when constrained to the domain of spatial descriptions in a limited and known environment can vary too greatly to be directly understood by the robot. The descriptions logged in the human subject experiment illustrate this. For this reason we are developing a process to convert spoken language into a limited set of robot commands that the robot can understand, i.e., a minimal set that will support the fetch task in a physical environment.

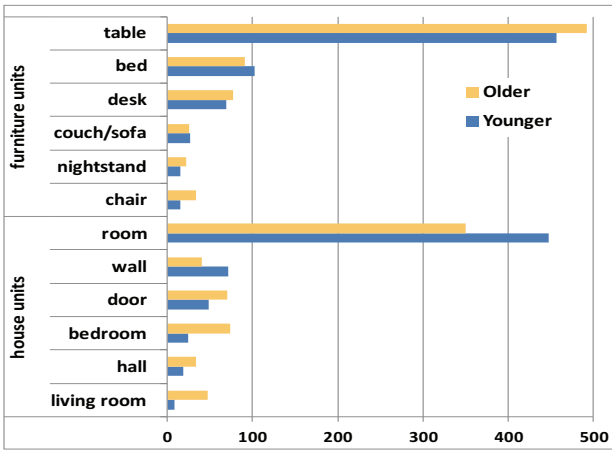

Fig. 4. *Total reference counts for furniture and room units.*

Our goal is not to solve the natural language processing problem in general, but rather to address the specific challenges in interpreting spatial descriptions. If ambiguities still exist, the robot can establish a dialog with the user for clarification; however, the goal is to allow the robot to reason about uncertain conditions first and ask questions only if necessary.

Overview

Figure 5 outlines the process used to convert spoken language into robot commands. After speech recognition, each word in the spatial description is tagged with an appropriate part-of-speech (POS). Next, the tagged description is chunked into a tree of phrases. The relevant Noun Phrases (NPs) are then extracted from the tree to adjust directional terms for accommodating the perspective taken by the user when giving the spatial description. The perspective adjustment process requires some additional inputs, namely prior knowledge of the environment and the current pose of the robot in the scene. The prior knowledge consists of a map of the rooms in the environment and a list of possible furniture items in each room. The adjustment process then returns the adjusted NPs and the join process combines them back into the tree of the entire description. The complete and perspective-adjusted description is then converted into robot commands. A hybrid map and path planning process similar to (Wei et al., 2009) could be added to support larger, more complex environments, although we have not considered this here.

Speech recognition is still a challenge and often results in errors. We assume that speech recognition accurately transcribes the speech into text. While this assumption is not realistic currently, it can be mitigated by training the robot to recognize its owner's voice and other methods. The capability of discourse between the robot and the user and using a constrained vocabulary are some methods to help mitigate the problems with speech recognition.

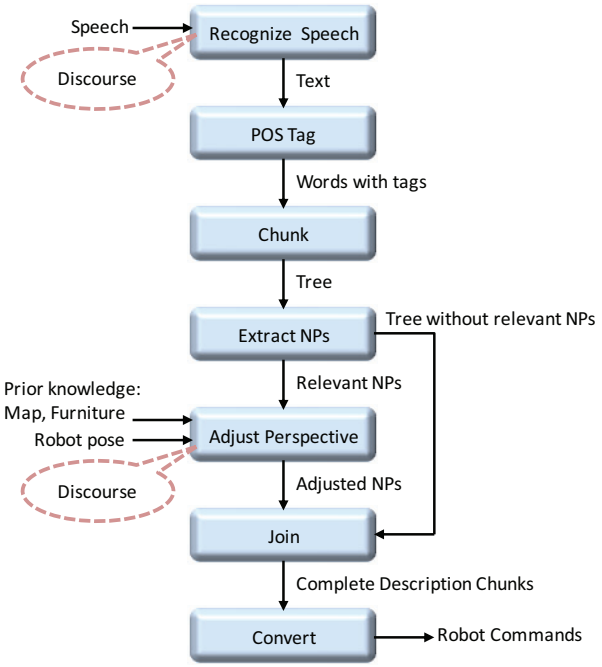

Fig. 5. *Language processing steps to translate spoken spatial descriptions into robot fetch commands*

Accurate POS tagging is also important to the overall process, because the proceeding steps explicitly depend on it. Chunking uses a set of grammatical rules, which are essentially regular expressions based on POS tags (Bird et al., 2009). Below is a simple grammar rule for a NP.

NP: {<DT|PP>?<JJ>*<NN>}
{<NN>+}

These rules define two common patterns of a NP. The first one consists of a determiner or possessive optionally followed by one or more adjectives and ending with a noun, i.e. "the blue chair". The second rule is consecutive nouns, i.e. "coffee table".

If POS tagging is done incorrectly, the chunking will also fail. For example, an issue observed with the default NLTK (Bird and Loper, 2004) POS tagger was that the word "bed" was tagged as a verb in every description where it was present, although it was clear to a human that it was used as a noun. In this situation the chunker would not recognize any phrase containing "bed" as an NP.

While it is possible to extend the rules for an NP to include phrases that end with a verb, this would result in some verb phrases being labeled as NPs by the chunker. A better solution is to improve the POS tagger by training it on a dataset. This dataset is simply a file of all of the descriptions collected from the experiment tagged by the default tagger which is then manually reviewed to fix the incorrect POS tags. Also, some rules are created for a bigram trained tagger to appropriately tag components of certain bigrams. If our dataset accurately depicts the spatial descriptions typically used, we would expect the trained

tagger to perform well because it will assign the correct tags from the training data and because of the constrained vocabulary. If a word not found in the training data is encountered, the nearest known synonym can be found using the WordNet database, as proposed in MacMahon et al. (2006). Further tests will explore this.

Conversion of the description to robot commands also depends on accurate POS tagging. The convert process is similar to the extract process, except only words with certain POS tags are extracted, and some words will actually be added to the command. For example, static (*Where*) descriptions do not necessarily contain action verbs such as "go" and "turn", so the convert process will add these verbs to form commands. Dynamic (*How*) descriptions are generally in a form very close to a robot command as they are given.

To simplify and optimize the extract and convert processes, specific custom POS tags will be used instead of general POS tags. Furniture items, rooms, directions, hedges, action verbs and target objects will each have their own specific tags in the training dataset. The chunker's grammar rules will also be based on these tags. The extract and convert processes then become very straightforward. These processes can simply pull out words that contain POS tags of interest. The convert process will add an action verb to form a command if it cannot find one.

Adjusting the Perspective

Figure 2 shows that the perspective taken by the speakers is not consistent in either group or in the addressee subgroups. Although older adults tend to take their own perspective more often when addressing the robot, there is no way to predict which perspective will be taken. Others have addressed this issue. Trafton et al. (2005) use a cognitive architecture to help the robot reason about perspective. Berlin et al. (2006) consider the teacher's perspective in robot learning by demonstration. Additionally people periodically confuse left and right, particularly when giving directions verbally as in this study. Here, we use prior knowledge of the environment to reason about perspective and correct human errors. Unlike Matuszek et al. (2010) where only environmental structures are considered, we use landmarks and context to reason about left and right.

Figure 6 shows a map of the virtual environment used in data collection. The Start point indicates the initial position of the speaker. This layout simplifies the situation significantly; however, more complex scenarios will be discussed. Also, given the initial position, determining the perspective is only relevant for rooms. However, the same approach can be applied if the robot, speaker, and target object are in the same room, provided the location and orientation of furniture items are included in the map.

Our approach requires that certain conditions are met; later we discuss some mitigation strategies if they cannot be met. The first condition is that the robot is given an approximate map of the environment. This map must at the very least include the entrances to the rooms and their names, e.g., 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 discussed in the next section. The second condition is that the robot knows its own location and orientation on the map, at least in relation to the entrances to the rooms.

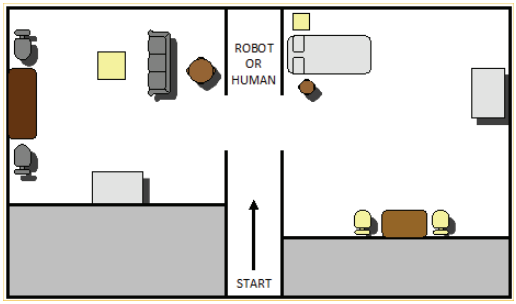

***Fig. 6**. A top-down layout of the virtual environment used in the human subject experiment.*

From Figure 5, the spatial description (which could be partial) is chunked into noun phrases. Proper chunking is critical; we are currently comparing and combining several approaches. Next we look for the appearance of a room name and "right" or "left" in each noun phrase. This is why chunking is important, because the directional term intended for a room or object must appear in the noun phrase about that room or object. If a noun phrase with a room name and a directional term is 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 it should turn left to go into the bedroom where the target object is. This will not work if there is a bedroom on the left and another on the right. This approach will also fail if the speaker leaves out the name of the room and just says "go into the room on the right".

To address these situations, additional steps are taken. The list of furniture items contained in each room is used to reason about the most likely room intended, by examining the furniture items mentioned. This also applies to the situation where the room is unnamed by the speaker (e.g., "room on the left"). For example a nightstand or a bed will generally be located in the bedroom and a couch

will generally be located in the living room. However, this, too, can fail if the furniture items in the description do not clearly indicate the room or multiple rooms contain the same furniture items. In this case, it is important that the robot recognize that ambiguity still exists. We propose discourse with the speaker in this situation. Since discourse is time consuming, however, we use it as a last resort if all of the reasoning steps fail.

Furniture Recognition for Spatial Referencing

Our intent is to study the robot fetch task in the physical world and include the perceptual challenges placed on the robot to accomplish the task. This is an important step towards language-based human robot interaction (HRI), as grounding of language is related to human perception (Roy 2005). The results of the human subject experiment indicate that furniture items are referenced often as landmarks in the spatial descriptions. Furniture placement could be included in a map; however, some items may be moved, so we do not want to rely on precise, mapped locations. As shown in Figure 3, the descriptions sometimes assume an intrinsic front or back of the furniture (e.g., …back of the sofa). Thus, to perform the fetch task in the physical world, it will be important for the robot to not only recognize furniture items but also capture their orientation. Our robot is built on a Pioneer 3DX base. The Microsoft Kinect provides the main sensing capabilities, positioned at a height of 1m; both color (RGB) and depth images are used.

Others have proposed language-based HRI approaches that require landmark recognition but have not included recognition strategies (e.g., Chen and Mooney, 2011). There is previous work on object recognition using the Kinect. Lai et al. (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. Much of the related work focuses on recognition only and is not necessarily concerned about execution speed. Speed is important for timely human-robot interaction. And, as noted above, detecting the orientation of the furniture item is important.

Furniture Recognition Methods

Large objects in the scene are first segmented based on the depth image; the corresponding color image segment is then used in the recognition process. Many furniture items found in the home have a primary horizontal plane, for example, chairs, beds, couches and tables. The main horizontal plane (the main plane) is identified using the RANSAC algorithm (Golovinskiy et al, 2009). Seven features are used in the furniture classifier:

1. Furniture size (area of the main plane)
2. Main plane height (average height of all points in the plane)
3. Main plane texture (local binary pattern operator (Ojala et al., 1996))
4. Furniture type (chair-like or table-like, computed based on shape)
5. Main plane red color proportion, normalized
6. Main plane green color proportion, normalized
7. 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 above are used as inputs into a system of fuzzy rules to recognize 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; the last three features are used with a support vector machine to make the final decision.

The confidence of the recognition result is determined from two aspects; the first is intrinsic confidence (determined by the features) which is the fuzzy membership value. The second is the 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, i.e., 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 dinner table, the couch and the bed, the robot is seldom able to view the entire item. 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.

Spatial referencing algorithms will require a known orientation of furniture items. The object front is computed differently for chair-like vs. table-like objects. The front of table shaped objects is based on the direction of the visible edge. The front of chair shaped objects is based on the direction of the chair back relative to the main plane, as shown in Figure 7.

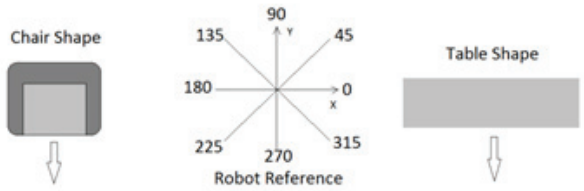

Fig. 7. *Chair shaped objects have an intrinsic front, as shown by the arrow, independent of the robot's relative position. For table shaped objects, the front is determined by the robot's relative position and viewing perspective.*

Experiments and Results

To test this approach, nine furniture items were selected, to represent items used in the VE for the human subject experiment (Figure 8). Color and depth images were taken

around each object in 8 directions and 8 distances from 1 to 3m. Of these 64 images, 48 were used for training and 16 for testing. As a further test, images were also collected with cluttered table tops to better represent an unstructured home environment. For each of these 6 items, 8 images were collected at 1.5 m. The results are shown in Table 1 for both tests. The furniture recognition process runs in about 9 ms on an Intel core i7 CPU at 1.6 GHz.

Fig. 8. *The nine furniture items tested*

Orientation was also tested using the same data as the uncluttered furniture recognition test. Results are shown in Table 2 for the 8 directions tested, as error values between orientation detected and the ground truth, in degrees. Objects 1 and 3 (small round table and hexagon table) are excluded from this test due to the general round shape. Other table shaped items (coffee table, dining table, desk, and bed) are symmetrical; thus, an orientation of less than 180 degrees is computed. The results show that orientation is easier to compute for some viewing angles. To improve results, the robot can be directed to move to a better viewing angle, based on the confidence level.

Table 1. *Recognition Results for furniture in Fig. 8*

Furniture Sample	Without Clutter	With Clutter
1	100%	100%
2	100%	N/A
3	100%	100%
4	87.5%	N/A
5	100%	87.5%
6	100%	100%
7	100%	100%
8	67.5	N/A
9	75%	N/A

Spatial Referencing

The spatial referencing language with respect to furniture items will build on the Histogram of Forces (HoF) (Matsakis and Wendling, 1999) to model spatial relationships between two objects. The HoF offers a mathematical framework that produces similar results to the Attention Vector Sum (Regier and Carlson, 2001) without training data and also supports any arbitrary shape and size of either object. In previous work, the HoF was used to generate spatial descriptions of a robot's environment based on range data and support a dialog with a human user about objects in a scene (Skubic et al., 2003; 2004). Initially, the work considered planar relationships projected onto the horizontal plane. Later, 3D descriptions were considered by also projecting object models onto a vertical plane (Luke et al., 2005). A similar approach will be used for the fetch task to support 3D descriptions.

Table 2. *Error results of the orientation test for 8 directions (in degrees). Low error values are shown in red*

	0	45	90	135	180	225	270	315
1	×	×	×	×	×	×	×	×
2	47	28	112	25	32	4	1	6
3	×	×	×	×	×	×	×	×
4	10	35	47	37	12	2	4	7
5	1	0	2	2	×	×	×	×
6	1	3	1	3	×	×	×	×
7	5	5	1	5	×	×	×	×
8	48	172	21	51	15	5	5	1
9	6	2	5	9	×	×	×	×

Conclusion

In this paper, we present an overview of results from a human subject experiment showing differences between older and younger adults in generating spatial descriptions for a robot fetch task. The results illustrate key challenges in determining the perspective used by the speaker, supporting both dynamic and static language structure with varying detail, and using furniture items in spatial references, possibly with frames intrinsic to the object.

We address these challenges by providing the robot with recognition and reasoning strategies that are similar to human strategies and thus, establish a common ground between the robot and the human user. The robot is similar to a human in how it recognizes furniture and its understanding of spatial relationships using HoF. Common ground is the map of the house and furniture content of the rooms; a human navigating through the environment creates a similar map in her head. Through custom POS tags, more common ground is created because the robot knows that left and right are directions, a bed is furniture, and keys are a target object. Reasoning about the purpose of the room using furniture content to determine which room is meant by the speaker is another form of common ground. Thus, providing a robot with these capabilities will help create a natural interface with the human user.

Acknowledgements

This work is funded by the National Science Foundation, grant# IIS-1017097. The authors thank Xiao Ou Li for constructing the virtual environment and numerous MU and ND students for help in data collection and coding.

References

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 Proc., ACM/IEEE Intl. Conf. on Human-Robot Interaction, 335-342, March, 2012, Boston, MA

Berlin, M., Gray, J., Thomaz, A.L., and Breazeal, C. 2006. Perspective Taking: An Organizing Principle for Learning in Human-Robot Interaction. In Proc., National Conf. on AI.

Bird, S., and Loper, E. 2004. NLTK: The natural language toolkit. In Proc., 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04).

Bird, S., Klein, E., Loper, E. 2009. *Natural Language Processing with Python* (1st ed.). O'Reilly Media, Inc.

Carlson, L.A. and Hill, P.L. 2009. Formulating spatial descriptions across various dialogue contexts. In K. Coventry, T. Tenbrink, & J. Bateman (Eds.), *Spatial Language and Dialogue,* 88-103. New York, NY: Oxford University Press Inc.

Carlson, L., Skubic, M., Miller, J., Huo, Z., and Alexenko, T. In Review. Investigating Spatial Language Usage in a Robot Fetch Task to Guide Development and Implement of Robot algorithms for Natural Human-Robot Interaction. *Topics in Cognitive Science*.

Chen, D.A., Mooney, R.J. 2011. Learning to Interpret Natural Language Navigation Instructions from Observations. In Proc., AAAI Conf. on Artificial Intelligence.

Golovinskiy, A., Kim, V.G., Funkhouser, T. 2009. Shape-based Recognition of 3D Point Clouds in Urban Environments. In Proc., Intl. Conf. on Computer Vision (ICCV), Sept.

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. ICCV Workshop on Consumer Depth Cameras in Computer Vision, Barcelona, Spain.

Lai, K., Liefeng B., Xiaofeng R., Fox, D. 2011. Sparse distance learning for object recognition combining RGB and depth information. In Proc., IEEE Intl. Conf. on Robotics and Automation, 4007-4013, Shanghai, China.

Luke, R., Blisard, S., Keller, J. & Skubic, M. 2005. Linguistic Descriptions of Three Dimensional Scenes Using SIFT Keypoints. In Proc., IEEE Intl. Workshop on Robot and Human Interactive Communication, Nashville, TN.

MacMahon, M., Stankiewicz, B., Kuipers, B. 2006. Walk the talk: connecting language, knowledge, and action in route instructions. In Proc., 21st National Conf. on AI – Vol. 2, 1475-1482. Boston, Mass.: AAAI Press.

Mainwaring, S., Tversky, B., Ohgishi, M. & Schiano, D. 2003. Descriptions of simple spatial scenes in English and Japanese. *Spatial Cognition and Computation,* 3: 3-42.

Ojala, T., Pietikäinen, M. and Harwood, D. 1996. A Comparative Study of Texture Measures with Classification Based on Feature Distributions. *Pattern Recognition*, 29: 51-59.

Matsakis, P. and Wendling, L. 1999. A new way to represent the relative position between areal objects. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 21(7): 634–643.

Matuszek, C. Fox, D., Koscher, K. 2010. Following Directions using statistical machine translation. In Proc., 5th ACM/IEEE Intl. Conf. on Human-Robot Interaction, 251-258, Osaka, Japan.

Regier, T. and Carlson, L.A. 2001. Grounding spatial language in perception: An empirical and computational investigation. *Journal of Experimental Psychology: General*, 130(2):273–298.

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

Schober, M.F. 1995. Speakers, addressees and frames of reference: Whose effort is minimized in conversations about locations? *Discourse Processes*, 20: 219-247.

Scopelliti, M., Giuliani, M., and Fornara, F. 2005. Robots in a domestic setting: a psychological approach. *Universal Access in the Information Society*, 4(2): 146-155.

Skubic, M., Matsakis, P., Chronis, G. and Keller, J. 2003. Generating multilevel linguistic spatial descriptions from range sensor readings using the histogram of forces, *Autonomous Robots*, 14(1): 51-69.

Skubic, M., Perzanowski, D., Blisard, S., Schultz, A., Adams, W., Bugajska, M., and Brock, D. 2004. Spatial language for human-robot dialogs, *IEEE Trans. on SMC, Part C,* 34(2): 154–167.

Tellex, S. Kollar, T., Dickerson, S., Walter, M.R., Banerjee, A.G., Teller, S. and Roy, N. 2011. Understanding Natural Language Commands for Robotic Navigation and Mobile Manipulation. In Proc., National Conf. on AI, San Francisco, CA.

Tenbrink, T., K. Fischer, and R. Moratz, 2002. *Spacial strategies of human-robot communication,* KI, no. 4 [Online]. Available: iteseer.ist.psu.edu/tenbrink02spacial.html.

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 Trans. on SMC, Part A*, 35(4): 460-470.

Wei, Y, Brunskill, E, Kollar, T, and Roy, N. 2009. Where to Go: Interpreting Natural Directions Using Global Inference. In Proc., IEEE Intl. Conf. on Robotics and Automation, Kobe, Japan.