Perspective in Natural Language Instructions for Collaborative Manipulation

Shen Li*, Rosario Scalise*, Henny Admoni, Stephanie Rosenthal, Siddhartha S. Srinivasa
Robotics Institute
Carnegie Mellon University
Pittsburgh, Pennsylvania 15213
Email: {shenli,robo,henny}@cmu.edu, srosenthal@sei.cmu.edu, siddh@cmu.edu

I. Introduction

As humans and robots collaborate more together on spatial tasks such as furniture assembly [6], warehouse automation [3], or meal serving [5], they need to communicate clearly about the objects in their environment. Without an effective mechanism for conveying and understanding spatial language, robots will remain unable to collaborate with humans in a way that emulates human-human collaboration. Developing a strategy for robots to understand and generate spatial language is an inaugural step in allowing robots to collaborate with humans better.

There is a long line of successful research in robotics related to communicating about spatial relationships, including [1, 2, 4, 9, 10, 12, 13]. Humans often take perspectives when referring to objects in and features of the environment in spatial tasks [11]. For example, in an analysis of 4000 utterances made by NASA astronauts training together for a mission, 25% of the utterances involve perspective taking [13].

There are different types of perspectives humans take, such as deictic (referring to the participants' points of view), intrinsic (referring to the objects' points of view), or neither, and they often switch perspectives as they are speaking [7]. On the other hand, with many possible ways to refer to the same object in the environment, there is often ambiguity when instruction givers are not explicit about which perspective they are taking [7].

In this paper, we divide possible perspectives into four categories: participant's perspective, partner's perspective, neither perspective, and unknown perspective. We evaluate perspective on clarity of spatial language. We collect a dataset of participants' instructions for selecting a block located on a table. Then we study the relationship between perspective-taking and clarity by evaluating whether a partner can understand those instructions. Results show that sentences that do not use perspectives are clearer than sentences that use partner or unknown perspectives.

II. Data Collection

To generate a corpus of spatial language that requires perspective taking, we created 28 stimulus images containing simplified block objects (Fig. 1). These images showed 15 randomly-spaced, randomly-colored blocks (orange, yellow, green, or blue) on a table. This stimulus design was chosen to elicit descriptions that rely more on the spatial arrangement of the blocks and perspective taking than on their individual appearance. People on Amazon's Mechanical Turk[1] were hired to use natural language to instruct a partner to select an indicated block. We analyze 1400 responses from 120 participants.

Four raters manually coded the instructions for perspective. There are four possibilities (see Table I for details):

- *Participant Perspective* - the egocentric perspective where the instruction refers to the speakers themselves [14].
- *Partner Perspective* - the addressee-centered perspective where the instruction refers to the addressees [14].
- *Neither Perspective* - the instruction does not refer to any perspectives[2],
- *Unknown Perspective* - the instruction does refer to some perspectives but fails to state them explicitly

Type	P1	P2	Example
Participant Perspective	+	-	"the block that is to **my** rightest."
			"**my** left most blue block"
Partner Perspective	-	+	"the block on **your** left"
			"second from the right from **your** view"
Neither Perspective	-	-	"closest to you"
			"the top one in a triangle formation"
Unknown Perspective	?	?	"to the **left** of the yellow block"
			"the block that is on far **right**"

TABLE I: Possible perspectives. (P1=Participant P2=Partner).

The inter-rater reliability was established on 10% of the data and the average Cohen's κ value was 0.68, indicating high inter-rater reliability. The coding result is available in Table II.

III. Study Design and Procedure

Subsequently, we conducted a study to empirically measure the clarity of the instructions collected in section II. In this study, participants were presented with 40 stimuli alongside

[1] www.mturk.com

[2] *Neither Perspective* instructions only use perspective-independent directional information. For example, "closer to you" should be classified as neither perspective instead of partner perspective, because it contains a perspective-independent reference to a landmark, "you," but not perspective-dependent relationships such as "on my left" and "on your right".

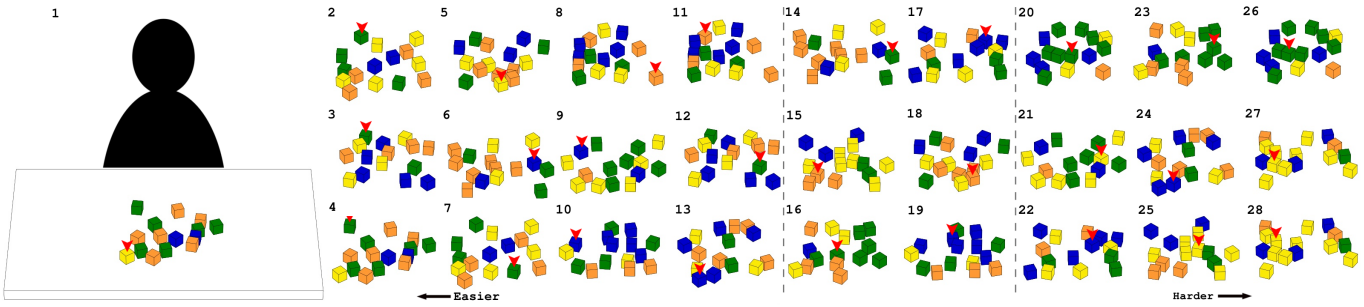

Fig. 1: The image on the left (labeled "1") is the full stimulus used to elicit spatial references. Online participants were asked to write how they would instruct the silhouetted figure to pick up the block indicated with the red arrow. The rest of the images are other possible configurations of blocks that were plugged into this full stimulus.

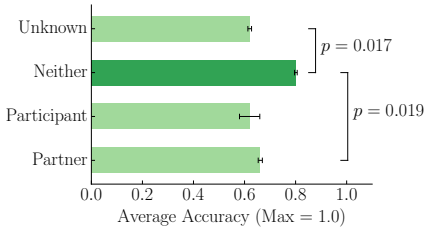

Fig. 2: The effect of perspective type on accuracy

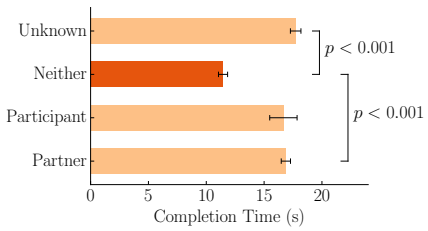

Fig. 3: The effect of perspective type on completion time

Perspective Type	Count	Percentage
Participant Perspective	15	1.07%
Partner Perspective	339	24.21%
Neither Perspective	592	42.29%
Unknown Perspective	454	32.43%
Total	1400	100%

TABLE II: Number of responses for each category of perspective

the corresponding block descriptions, and were asked to identify the indicated block. We collected 10 responses for each of the 1400 instructions from 356 participants.

We compute the following metrics for each response:

- *accuracy* - whether the identified block matches the indicated target block from the original instruction.
- *completion time* - how long the participant spends identifying the target block.

We hypothesize that instructions which do not refer to any perspectives will take participants *less time* and they will be *more accurate* in selecting the target block than instructions which use participant, partner, or unknown perspectives.

A one-way ANOVA measuring the effect of perspective type (participant, partner, ambiguous, or unknown) on accuracy and completion time reveals a significant effect for both accuracy ($F(3, 1396) = 43.655, p < 0.005$) (Fig. 2) and completion time ($F(3, 1396) = 34.607, p < 0.005$) (Fig. 3). Instructions that use neither perspective have higher accuracies than instructions that use partner ($p = 0.019$) or unknown ($p = 0.017$) perspectives. Similarly, completion time was lower for instructions that use neither perspective than instructions that used partner ($p < 0.001$) or unknown ($p < 0.001$) perspectives. No other significant differences are found.

These results confirm that instructions that use no perspective take less time to process and yield higher accuracy in discerning the referred block. Our hypothesis is supported.

IV. Discussion

According to our results, directional spatial references requiring perspectives will significantly increase sentence complexity and ambiguity. This suggests that for a collaborative robot, if it is able to generate a description using no perspectives, it should always prefer to do so over other sentence formulations. However, other factors, such as conciseness, must also be taken into consideration to produce clear instructions. Trade-offs between these approaches are our point of future work. More details about study and deeper analysis of data are contained in our recent publication [8].

Acknowledgments

This work was (partially) funded by the DARPA SIMPLEX program through ARO contract number 67904LSDRP, National Institute of Health R01 (#R01EB019335), National Science Foundation CPS (#1544797), the Office of Naval Research, and the Richard K. Mellon Foundation.

References

[1] Jacob Arkin and Thomas M Howard. Towards learning efficient models for natural language understanding of quantifiable spatial relationships.

[2] Yonatan Bisk, Daniel Marcu, and William Wong. Towards a dataset for human computer communication via grounded language acquisition. In *Proceedings of the AAAI Workshop on Symbiotic Cognitive Systems*, 2016.

[3] Hobart R Everett, Douglas W Gage, Gary A Gilbreath, Robin T Laird, and Richard P Smurlo. Real-world issues in warehouse navigation. In *Photonics for Industrial Applications*, pages 249–259. International Society for Optics and Photonics, 1995.

[4] Thomas M Howard, Stefanie Tellex, and Nicholas Roy. A natural language planner interface for mobile manipulators. In *International Conference on Robotics and Automation*, pages 6652–6659. IEEE, 2014.

[5] Sumio Ishii, Shinji Tanaka, and Fumiaki Hiramatsu. Meal assistance robot for severely handicapped people. In *IEEE International Conference on Robotics and Automation*, volume 2, pages 1308–1313. IEEE, 1995.

[6] Ross A Knepper, Todd Layton, John Romanishin, and Daniela Rus. Ikeabot: An autonomous multi-robot coordinated furniture assembly system. In *IEEE International Conference on Robotics and Automation*, pages 855–862. IEEE, 2013.

[7] Willem JM Levelt. Perspective taking and ellipsis in spatial descriptions. *Language and space*, pages 77–107, 1996.

[8] Shen Li, Rosario Scalise, Henny Admoni, Siddhartha S Srinivasa, and Stephanie Rosenthal. Spatial references and perspective in natural language instructions for collaborative manipulation. In *IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 2016.

[9] Matt MacMahon, Brian Stankiewicz, and Benjamin Kuipers. Walk the talk: Connecting language, knowledge, and action in route instructions. In *Proceedings of the National Conference on Artificial Intelligence*, 2006.

[10] Marjorie Skubic, Dennis Perzanowski, Samuel Blisard, Alan Schultz, William Adams, Magda Bugajska, and Derek Brock. Spatial language for human-robot dialogs. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 34(2):154–167, 2004.

[11] Holly A Taylor and Barbara Tversky. Perspective in spatial descriptions. *Journal of memory and language*, 35(3):371–391, 1996.

[12] Stefanie Tellex, Thomas Kollar, Steven Dickerson, Matthew Walter, Ashis Banerjee, Seth Teller, and Nicholas Roy. Understanding natural language commands for robotic navigation and mobile manipulation. In *AAAI Conference on Artificial Intelligence*, 2011.

[13] J Gregory Trafton, Nicholas L Cassimatis, Magdalena D Bugajska, Derek P Brock, Farilee E Mintz, and Alan C Schultz. 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, 2005.

[14] Jeremy M Wolfe. Guided search 2.0 a revised model of visual search. *Psychonomic bulletin & review*, 1(2): 202–238, 1994.