Spatial References and Perspective in Natural Language Instructions for Collaborative Manipulation

Shen Li*, Rosario Scalise*, Henny Admoni, Stephanie Rosenthal, Siddhartha S. Srinivasa

***Abstract*— As humans and robots collaborate together on spatial tasks, they must communicate clearly about the objects they are referencing. Communication is clearer when language is unambiguous which implies the use of spatial references and explicit perspectives. In this work, we contribute two studies to understand how people instruct a partner to identify and pick up objects on a table. We investigate spatial features and perspectives in human spatial references and compare word usage when instructing robots vs. instructing other humans. We then focus our analysis on the clarity of instructions with respect to perspective taking and spatial references. We find that only about 42% of instructions contain perspective-independent spatial references. There is a strong correlation between participants' accuracy in executing instructions and the perspectives that the instructions are given in, as well between accuracy and the number of spatial relations that were required for the instruction. We conclude that sentence complexity (in terms of spatial relations and perspective taking) impacts understanding, and we provide suggestions for automatic generation of spatial references.**

I. Introduction

As people and robots collaborate more frequently on spatial tasks such as furniture assembly [1], warehouse automation [2], or meal serving [3], they need to communicate clearly about objects in their environment. In order to do this, people use a combination of visual features and spatial references. In the sentence "The red cup on the right", 'red' is a visual feature and 'right' is a spatial reference.

There is a long line of research in robotics related to communicating about spatial references like 'furthest to the right','near the back', and 'closest' for navigation task [4]–[10]. However, there are fewer studies involving the communication of spatial references for tabletop or assembly tasks [11]. A common theme in the space of tabletop manipulation tasks is clutter which we view as many potential objects to reason about. See Fig. 1

*Both authors contributed equally.

Robotics Institute, Carnegie Mellon University, Pittsburgh PA 15213 {shenli,robo,henny,siddh}@cmu.edu, srosenthal@sei.cmu.edu

This material is based upon work funded and supported by the Department of Defense under Contract No. FA8721-05-C-0003 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center. [Distribution Statement A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution. Carnegie Mellon is registered in the U.S. Patent and Trademark Office by Carnegie Mellon University. DM-0003432.

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.

A cluttered table introduces the problem of *object uniqueness* where if there are two objects which are identified in the same manner (e.g. the red cup among two red cups), we are left with an ambiguity. One possible solution to this is to utilize *spatial references* which allow the use of spatial properties to establish a grounding or certainty about the semantic relationship between two entities.

However, even with the use of spatial references, it is still possible to encounter additional ambiguity which originates from the reference frame. Humans often use perspective to resolve this ambiguity as in the example 'the red cup on your right'. Often times, in tabletop scenarios, the person giving instructions will be situated across the table from their partner and thus will have a different perspective. Therefore, robots that collaborate with humans in tabletop tasks have to both understand and generate *spatial language* and *perspective* when interacting with their human partners. We investigate these key components by collecting a corpus of natural language instructions and analyzing them with our goal of clear communication in mind.

We first conducted a study in which we asked participants to write instructions to either a robot or human partner sitting across the table to pick up an indicated block from the table as shown in Fig. 1. This task raises a perspective problem: does the participant use the partner's perspective or their own perspective, if any? Blocks were not always uniquely identifiable, and so the task required participants to describe spatial relationships between objects as well. We analyze the instructions from participants for 1) language differences between instructing a human versus a robot partner, 2) trends in language for visual and spatial references, and 3) the perspective(s) participants use when instructing their partners.

To investigate the effect of perspective, we conducted a second study in which we presented new participants with the instructions from the first study and asked them to select the indicated block. We utilized the correct selection of the indicated block as an objective measure of clarity. In order to establish which instructions contained ambiguities (lack of clarity), we first manually coded the instructions for whether the reference perspective was unknown or explicit (participant's, partner's, or neither) and whether there were multiple blocks that could be selected based on the instruction. An unknown perspective implies the instruction is dependent on perspective, but it is not explicitly stated.

Results from the first study show that participants explicitly take the partner's perspective more frequently when they

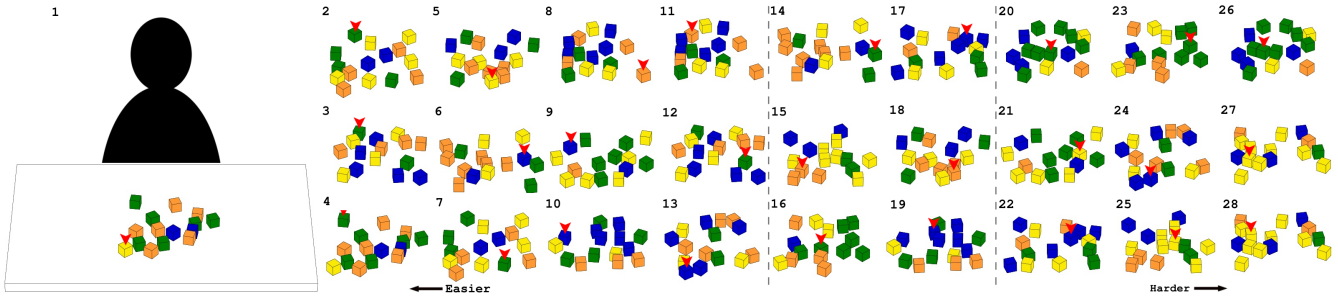


Fig. 1: Scenes used to elicit spatial references. Online participants were asked to write how they would instruct the silhouetted figure to pick up the block indicated with the red arrow. For each participant, the silhouette was either referred to as a robot or a human partner. The block configurations on the left were rated as the easiest to describe, while the configurations on the right were the most difficult.

believe they are instructing a person rather than a robot. Additionally, we find that people use color most frequently to refer to a block, while block density (e.g. the cluster of green blocks), block patterns (e.g. lines of red blocks), and even certain precise quantitative terms (e.g. 2nd block to the left) are also widely used. Finally, people spend more time writing the instructions and rate their tasks as more challenging when their instructions require the use of more spatial references.

From the second study, we find that 58% of our collected instructions contain perspective-dependent spatial references. Of this 58% more than half fail to explicitly specify the perspectives. This results in participants taking longer amounts of time to process the instructions and lower accuracies in discerning the intended block. The other 42% of instructions contained perspective-independent spatial references. These instructions demonstrated quicker completion times and higher correct block selection accuracies. We conclude that it is beneficial for instructions to avoid perspective-dependent spatial references when possible.

II. Related Work

A. Visual Search

Visual search, defined as the routine visual behavior to find one object in a visual world filled with other distracting items [12], is well aligned with our tabletop task. Wolfe divides visual search into 2 steps: processing easy information from all locations at the same time in parallel and focusing on the complex information from a few spatial locations. In the first step, people respond to visual stimuli from the scene, including object color, stereoscopic depth, line arrangement, curvature, intersection, and terminator [12]. In this work, we match these features to our collected instructions and analyze the frequency of each feature.

B. Spatial Reference

When similar objects are involved, referring to a group as a whole is an easy and natural way of specifying the target [13]. In spatial navigation system, there are three hierarchical levels: landmark, route, and survey knowledge of an environment. Landmarks are unique objects at fixed locations; routes correspond to fixed sequences of locations; survey knowledge abstracts and integrates knowledge from different experiences into a single model [14].

C. Perspective

When people collaborate together on spatial tasks, they often must take each other's perspectives when referring to objects in and features of the environment [15], [16]. In an analysis of 4000 utterances made by NASA astronauts training together for a mission, 25% of the utterances involved perspective taking [10].

Levelt separates perspectives into three categories: deictic perspective (referring to the participants' points of view, e.g. "on my left"), intrinsic perspective (referring to the objects' points of view, e.g. "in front of the car"), and absolute perspective (referring to the world frame, e.g. "north") [17]. Levinson merges addressee-centered and deictic perspectives into relative perspective (referring to landmark object) [18].

Most work on spatial references and perspective taking for robots assumes people always take robots' perspective when giving instructions for tabletop [19], [20] or navigation [21] tasks. When a person instructs a robot to perform a task with some ambiguity, the person prefers the robot to take the person's perspective [22]. In object identification tasks, people intuitively use their robot partners' perspectives [23]. Conversely, human-human collaboration literature reveals that solo people with imaginary human partners are uniform in taking their partners' perspectives while people with real human partners are not [24], indicating that there is no consensus on common perspectives. Hence, in our task where participants instruct partners sitting across the table, we analyze and rank different perspectives participants use.

D. Ambiguity

Instructions become obscure when the instruction givers are not explicit about the perspectives they are taking and the instructees have to make assumptions [17]. Moreover, the ambiguities occurred in target objects, landmarks, and spatial relationships between them [25] also make the instruction harder for people to understand. For example, in object identification tasks, applicability regions for spatial references are fairly large and not mutually exclusive, which makes instructions not necessarily precise [23].

In an experiment in which people were asked to write navigation instructions to another person, the other person

was only able to successfully follow 69% of the nearly-700 instructions while the others are ambiguous [6]. In a similar study, subjects were only able to navigate to the final destination 68% of the time [26]. We analyze the effects of general ambiguity and the ambiguity caused by unknown perspective on the easiness that people can understand the instruction.

E. Human Partner vs. Robot Partner

Robot is treated as a communication partner who needs more basic instructions than human interlocutor [25]. This is consistent with another study where half of the participants instruct robots by decomposing the action and describing paths to adapt to robots' assumed linguistic and perceptual abilities. [13]. Seniors want a streamlined communication with a task-oriented robot and do not want to speak to robots the same way they speak to people [27]. Therefore, we also investigate the difference between the way people speak to a robot and to a human partner in our tabletop task.

III. Study 1: Collecting Language Examples

To understand how people describe spatial relationships between similar objects on a tabletop, we collected a broad corpus of spatial references generated by 100 online participants. We analyzed this corpus for the types of words participants used and the word choice across differences in perceived difficulty of providing a spatial reference.

A. Study design

To collect spatial references that represents tasks that required perspective taking as well as object grounding, we created a set of stimulus images. Each image represents a configuration with 15 simplified block objects in different colors (orange, yellow, green, or blue) on a table. (Fig. 1). We first generated 14 images of configuration independently, each of which included different visual features and spatial relationships, such as a single block of one color, pairs of blocks closely placed, blocks separated from a cluster, and blocks within or near clusters of a single color. Then we placed red-arrow indicators above two different target blocks independently in each image and ended up with 14 pairs of configuration (28 images of configuration in total).

This stimulus design is chosen to elicit instructions that rely more on the visual and spatial arrangement of the blocks than their individual appearance for the purposes of human-robot interaction. In order to capture clear instructions for a potential partner, this task asked participants to instruct a hypothetical partner to pick up the indicated block as though that partner could not see the indication arrow. The partner (indicated by the silhouetted figure in the images) was seated across the table from the participant viewing the scene. This setup required participants to be clear about the target blocks and the perspectives where they were describing the blocks.

Prior work indicates that people communicate with robots differently from with other people [13], [25], [27]. Therefore, we varied whether participants were told that their partner (the silhouette) was human or robot. [1] Participants were randomly assigned to either the human or the robot condition, and this assignment was the same for every stimulus they saw. The stimuli were otherwise identical across conditions.

We analyze the results with respect to these hypotheses:

- H1 People use different words when talking to human and robot. Specifically, people are *more verbose*, *more polite*, and use *more partner-based perspective words* to human partners than robot partners.
- H2 The frequency of words used in all instructions correlates with the features used in visual search (*color, stereoscopic depth, line arrangement, curvature, intersection, and terminator* [12]).
- H3 Subjective ratings of sentence difficulty correlate with the number of spatial references required to indicate the target blocks.

B. Study Procedure

We deployed our study through Amazon's Mechanical Turk[2]. Each participant was randomly assigned a partner condition (human vs robot) and 14 trials. In each trial, participants were presented with an image, like the one on the left side of Fig. 1, which was randomly chosen from the two predefined configurations in each of the 14 pairs of configuration III-A. The participants then typed their instructions and rated the difficulty of describing that block on a 5-point scale. For each trial, we also collected the completion time. After completing 14 trials, participants were asked 1) if they followed any particular strategies when giving instructions, 2) how challenging the task was overall, and 3) for any additional comments they had about the task. Finally, we collected demographics such as age, gender, computer usage, handedness, primary language (English or not), and experience with robots.

C. Metrics

We analyze the collected corpus for language features. To analyze the differences on word choice between human-partner group and robot-partner group (H1), we computed

- *word count* - number of words for each instruction,
- *politeness* - presence of the word "please" in each instruction,
- *perspective* - whether the instruction explicitly refers to participant's perspective (egocentric), partner's perspective (addressee-centered), neither perspective[3], or unknown perspective (instruction implicitly refer to some perspectives) (see Table I for details).[4] [17], [18]

Word count and politeness were automatically extracted from the text. Perspective was manually coded by four raters

[1] We did not change the visual appearance of the silhouette

[2] www.mturk.com

[3] *Neither Perspective* sentences 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".

[4] Object-centered perspective is not considered because blocks are all the same except color

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).

Word Category	Description
Action	An action to perform
Object	An object in configuration
Color	Color of object
Ordering/Quantity	Ordering/Quantization of objects
Density	Concentration of objects (or lack of)
Pattern/Shape	A readily apparent formation
Orientation	The direction an object faces
Environmental	Reference to an object in the environment
Spatial Reference	Positional reference relating two things
Perspective	Explicitly indicates perspective

TABLE II: Word categories and their brief descriptions

who coded the same 10% of the data and iterated until high inter-rater reliability, measured by averaging the result of pairwise Cohen's κ tests. The average κ value for perspective was 0.85, indicating high inter-rater reliability. Once this reliability established, the four raters each processed one quarter of the remainder of the instructions.

To compare the features used in our collected instructions with visual search (H2), we classify words into categories adapted from visual search literature [12]. The categories are listed in Table II and presented in the order of *word frequency*, the number of instructions that contain words from the category divided by the size of the corpus.

To verify the correlation between perceived difficulty and the number of required spatial references (H3), we compare the subjective *difficulty rating* (Likert scale 1 (easy) to 5 (difficult)) to the following objective measures:

- *word count* - as computed for H1
- *spatial reference count* - as computed for H2
- *ordering and quantity word count* - as computed for H2
- *completion time* - the duration from when a participant loads a new stimulus to when the participant hits the submit button for his/her instruction.

D. Results

In the study, we recruited 120 participants and over-sampled 1680 instructions so that we could account for errors in data collection process and invalid responses. We remove 10 sentences (0.006%) that either do not refer to any blocks or are otherwise nonsensical. For consistent and organized analysis, we randomly select 1400 sentences from the remaining 1670 to ensure that each of the 28 configurations has exactly 50 instructions divided as evenly as possible between partner conditions. We analyze the 1400 sentences selected in this manner.

Visual Feature	Count	Frequency
Color	1301	0.929
Ordering/Quantity	498	0.356
Density	456	0.326
Pattern/Shape	60	0.043
Orientation	1	0.001

TABLE III: Visual feature frequencies and feature-included sentence counts over all 1400 sentences ranked from most to least frequent

1) Hypothesis H1: To evaluate the different words people used when speaking to a robot or human partner (H1), we analyze the overall *word count*, number of *politeness* words, and *perspective* used between the two partner conditions.

To analyze word count, we conduct an independent-samples t-test comparing number of words in the sentences for the two partner conditions. There is no significant difference in the mean sentence length by partner type (human: $M = 14.90, SD = 7.8$, robot: $M = 14.35, SD = 7.1$), $t(1398) = -1.389, p = 0.179$.

To analyze politeness, we conduct a Chi-squared test of independence between partner type (human or robot) and politeness word (present or absent). There is a significant relationship between the two variables, $\chi^2(1) = 6.685, N = 1400, p = 0.01$. Though use of politeness words is rare overall (only 4.6% of all the sentences contain "please"), politeness words are used significantly more often in human-partner condition (6.1%) than robot-partner condition (3.2%).

To analyze perspective, we conduct a Chi-squared test of independence between partner type (human or robot) and perspective used (participant's, partner's, neither, or unknown). There is a significant relationship between the two variables, $\chi^2(3) = 13.142, n = 1400, p = 0.004$. Post-hoc analysis with a Bonferroni correction identify that the partner perspective is used significantly more frequently in human-partner condition (28.1% of sentences) than in robot-partner condition (20.6% of sentences), $p = 0.001$. No other significant differences are found. This result is aligned with the idea that people adapt to the robot's assumed linguistic and perceptual abilities when talking to a robot. [13].

Thus, H1 is partially supported: there is no difference in sentence length between human and robot conditions, but people use "please" more often and take partner's perspective more frequently when they believe they are instructing another human than instructing a robot.

2) Hypothesis H2: To address our belief regarding the correlations between the visual features in our collected instructions and visual search (H2), we analyze how frequently sentences contain visual search features.

A summary of the results are in Table III.

First, a reference to color is used in nearly every sentence, since color is such a salient feature in our stimuli as well as in visual search. Next, although orientation is also strongly influential according to visual search literature, orientation is almost never referenced in our data. This is likely due to the fact that in our study, blocks have 4-way symmetry and are not oriented in any particular direction [12].

Without many other visual indicators, participants fre-

quently referred to "dense" regions of one particular color or to shapes or patterns they saw in the blocks such as a "line of red blocks". These references are observed in the literature with less consistency than color and orientation are [12].

Finally, although ordering/quantity does not fit the paradigm of visual search [12] as well as the previously mentioned features did, these words are closely related to the concepts of pattern/shape and density. "The third block in the line" and "The second block from the cluster" are examples respectively. We find high occurrence of ordering/quantity words especially in relation to other visual search terms.

In summary, we find that the observed frequency of many categories of words in our corpus, including color, density, shape, and ordering/quantity, closely matched what we expected based upon the visual search literature [12].

3) Hypothesis H3: We evaluate the effect of perceived difficulty on word choice in each instruction (H3) by investigating the correlations between subjective rating of difficulty, overall word count, number of spatial references, number of order/quantity words, and completion time. We excluded any trials on which the participant did not provide a subjective rating of difficulty and two outlier trials for which the response times were greater than 10 minutes, which ended up with 1353 sentences.

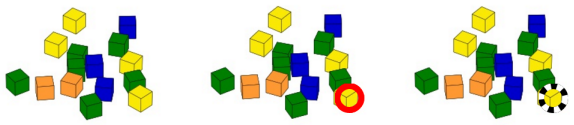
Because we use ordinal measures in this evaluation (e.g. subjective difficulty is rated on a 5-point scale), we conduct a Spearman's rank-order correlation to determine the relationship among the five metrics identified. There are statistically significant correlations across all pairs of metrics ($p < 0.01$ for all, which accounts for multiple comparisons).

Table IV details these correlations, and Fig. 2 visually displays some of them. Some of our key observations are:

1) As expected, there is a clear positive correlation (0.528) between word count and difficulty (Fig. 2a): easier scenes require fewer words to describe.
2) Also as expected, there is a clear positive correlation (0.508) between completion time and difficulty (Fig. 2b): harder scenes require more time.
3) Interestingly, easier rated tasks generally require fewer spatial references (Fig. 2c): more spatial references in a sentence imply a greater depth of search to find the correct answer.

These findings confirm that subjective ratings of sentence difficulty are strongly correlated with the number of spatial references required to indicate the target block.

We conclude that participants are more polite and use partner's perspective more frequently when instructing a human partner than a robot partner. Additionally, the words used in the instructions are in line with the words used by participants when helping partners perform visual search. Finally, there are strong correlations between subjective rating of difficulty with all of our objective measures. However, we are mostly interested in whether these collected instructions are clear enough for partners to understand. Our second study is aimed at analyzing the corpus from Study 1 for clarity.



(a) (b) (c)

Fig. 3: (a) In Study 2 (Sec. IV), we removed the red indication arrow. (b) As participants move their mouse over the image, a red circle will appear over the blocks to show them which block they could possibly select. (c) When they click on the block, a black and white checkered circle will appear around the selected block.

IV. Study 2: Evaluating Language for Clarity

To study the principles of clear spatial references in human robot collaboration, we need to validate the clarity of the instructions obtained in Study 1 (Sec. III). First, we manually coded the instructions in terms of two criteria (perspectives had already been coded in Study 1 (Sec. III)):

- *block ambiguity* - the number of blocks that people could possibly identify from the image based on the given instruction.
- *perspective* - whether there is an explicitly stated perspective provided in the instructions.

Subsequently, we ran a follow up study to empirically measure the clarity of the sentences. In this second study, participants were presented with the stimuli from Study 1 (Sec. III) (without red indication arrows) alongside the corresponding block descriptions from Study 1 (Sec. III), and were asked to click on the indicated blocks. We collected responses from ten participants for each instruction from Study 1 (Sec. III).

A. Coding Instructions for Clarity

We manually code each of the instructions from Study 1 (Sec. III) for perspective and general block ambiguity. The coding measures, inter-rater reliability scores, and preliminary findings are described next.

1) Perspective: As described in Sec. III-C and Table I, all sentences are labeled with perspective information. Among all the 1400 sentences, 454 (32.4%) sentences use unknown perspective, 339 (24.2%) sentences use partner perspective, 15 (1.07%) sentences use participant perspective, and 589 (42.1%) sentences use neither perspective.

2) Block Ambiguity: Block ambiguity is the number of blocks this sentence could possibly apply to. For our definition, no inferences are allowed when determining block ambiguity. Every detail which could possibly lead to ambiguity should be considered and expanded to different referred blocks. For example, the spatial relation "surrounded" could mean either partially or fully surrounded, which makes the sentence "the block that is surrounded by three blocks" potentially ambiguous. Unknown perspective could also lead to block ambiguity if different blocks are identified under the assumption of different perspectives.

We manually code each of the instructions from Study 1 (Sec. III) for "high" or "low" block ambiguity. If a sentence

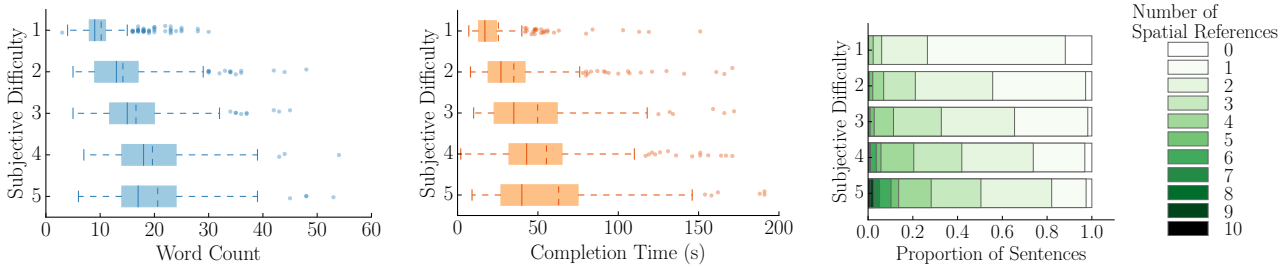


Fig. 2: The effect of subjective difficulty on ratings and measures of the sentences, such as (a) word count, (b) completion time, and (c) number of spatial references.

	Difficulty	Word Count	Spatial Reference	Order/Quantity Word	Completion Time
Difficulty	—	0.528	0.213	0.338	0.508
Word Count	0.528	—	0.416	0.425	0.682
Spatial Reference	0.213	0.416	—	0.082	0.262
Order/Quantity Word	0.338	0.425	0.082	—	0.350
Completion Time	0.508	0.682	0.262	0.350	—

TABLE IV: Spearman's rho correlations of sentence features and scene difficulty evaluations. All correlations are statistically significant with $p < 0.01$.

could refer to only one single block in the scene, it is rated as "low" ambiguity. Otherwise, it is rated as "high" ambiguity. We use the same process as in Sec. III-C to establish inter=rater reliability. On 10% of the data, the average Cohen's κ for the four raters is 0.68, indicating high rater agreement. Each rater subsequently code one quarter of the remaining data.

Among all the 1400 sentences coded, 895 (63.9%) sentences are not block ambiguous with only one block being referred to, while 492 (36.1%) sentences possibly refer to more than one block.

B. Online Study Design and Procedure

As mentioned above, the goal of the second study is to investigate the clarity of spatial instructions, which will guide us through the future research on robot-generated instructions. In this online study, new Amazon Mechanical Turk participants were shown 40 configurations random chosen from the pool of 28 configurations generated in Study 1 (Sec. III). Each configuration was presented alongside one of the corresponding instructions from Study 1 (Sec. III) corpus. We would make sure that the clarity of all the collected instructions in Study 1 (Sec. III) were evaluated here. Then the participants were asked to click on the block that best matched each instruction. As they moved their mouse over the image, a red circle appeared over the blocks to show them which block they would be selecting (Fig. 3b). When they clicked on the block, a black and white checkered circle would appear around the selected block (Fig. 3c). Continuing to move the mouse would present a red circle on those blocks which the participants could then click on to change their answer. Then we measured the participant's accuracy at selecting the indicated block.

We compute the following metrics for Study 2:

- *Final Answer* - whether a participant picks the correct block
- *Accuracy* - average over 10 participants of *final answer* for each instruction
- *Completion Time* - duration from moment when the page finishes loading to the moment when a participant clicks the next button to proceed.

Based on our ambiguity measures and the results from Study 2, we hypothesize that:

H4 *Block ambiguous* sentences will take participants in Study 2 **more time** and participants will be **less accurate** in discerning the referred block.

H5 Sentences with *unknown perspective* will take participants in Study 2 **more time** and they will be **less accurate** in discerning the referred block. Conversely, sentences with *neither perspective* will take **less time** and participants will be **more accurate** in discerning the referred block.

C. Results

We collect the responses from 356 participants and randomly select 10 responses for each of the 1400 sentences from Study 1 (Sec. III). We evaluate the participant performance in Study 2 on the set of sentences from Study 1 (Sec. III) by measuring their accuracy and completion time as described above. We also compare the objective accuracy measure to our manually-coded block ambiguity and perspective taking.

1) Hypothesis H4: First, we investigate block ambiguity by conducting an independent samples t-test measuring the effect of block ambiguity (low or high) on accuracy (Fig. 4a) and completion time (Fig. 4b). There are significant results for both accuracy ($t(1398) = 13.888, p < 0.005$) and completion time ($t(1398) = -5.983, p < 0.005$). Accuracy is lower and completion time is higher on sentences that contain ambiguous block references (H4). These results confirm that block ambiguous statements take longer amounts of time for

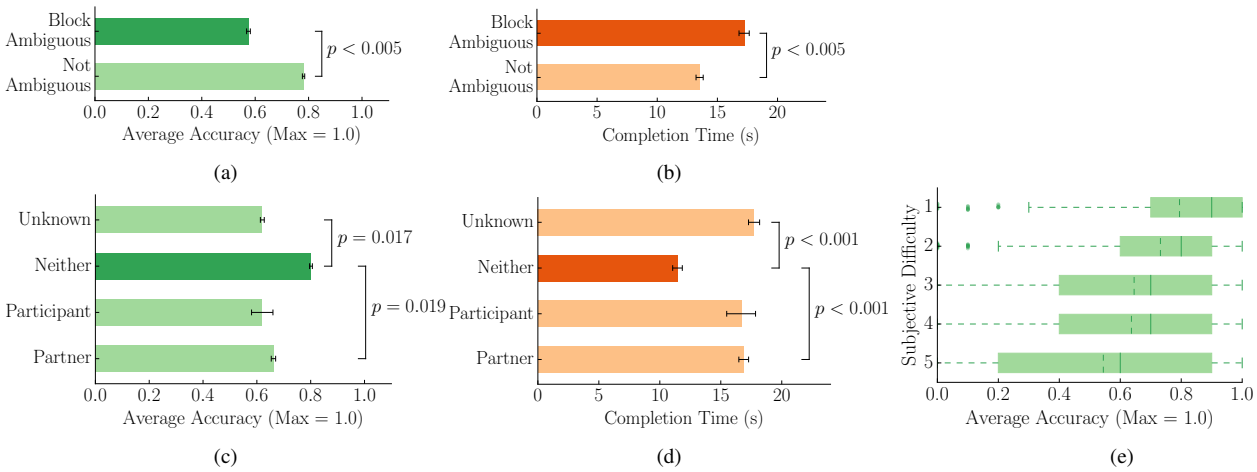


Fig. 4: The effect of block ambiguity on (a) average selection accuracy and (b) completion time. The effect of perspective on average selection (c) accuracy and (d) completion time. The effect of the subjective participant ratings of difficulty from Study 1 (Sec. III) on (e) average selection accuracy from Study 2 (Sec. IV).

participants to process and participants are less accurate in discerning the referred block.

2) Hypothesis H5: Next, we analyze perspective taking by conducting a one-way ANOVA measuring the effect of perspective type (participant, partner, neither, or unknown) on accuracy (Fig. 4c) and completion time (Fig. 4d). Perspective type has a significant effect for both accuracy ($F(3, 1396) = 43.655, p < 0.005$) and completion time ($F(3, 1396) = 34.607, p < 0.005$). Sentences that use neither perspective have higher accuracies ($M = 0.802, SD = 0.240$) than sentences that use partner ($M = 0.662, SD = .278, p = 0.019$) or unknown ($M = 0.619, SD = 0.307, p = 0.017$) perspective (H5). Similarly, average time is lower for sentences that use neither perspective ($M = 11.418s, SD = 10.56$) than partner ($M = 16.881, SD = 9.81, p < 0.001$) or unknown ($M = 17.756, SD = 12.03, p < 0.001$) perspective (H5). No other significant differences are found. These results confirm that neither perspective statements take shorter amounts of time for participants to process and participants are more accurate in discerning the referred block. At the same time, unknown perspective statements take participants longer time and participants are less accurate.

Additionally, we observe that participants in Study 2 have lower accuracy on sentences that participants in Study 1 (Sec. III) label as more difficult (Fig. 4e). This result is not surprising as participants who have trouble writing a clear sentence would likely rate the task as difficult.

We conclude that hypotheses 4 and 5 are both supported. Block ambiguity and unknown perspective are both correlated with higher completion times and lower accuracies. The type of perspective in the sentence has a significant effect on accuracy: when the instructions are written in neither perspective, participants in Study 2 have higher accuracy than any of the other perspectives.

V. Discussion

Keeping the goal of seemless human-robot collaboration in a tabletop manipulation setting in mind, we find the results from this first step forward quite encouraging. We created a corpus of natural language when specifying objects in a potentially ambiguous setting. We identified a cognitive process which plays a significant role in the formation of these specifying descriptions. We defined metrics to aid in scoring the optimality of a description. We designed an evaluation process based on these metrics. And finally, we performed an initial, yet broad, analysis on our corpus that was able to uncover a handful of insights. We will discuss a few of these insights in the following section.

In analyzing the corpus, we discovered that participants generally followed one of three approaches when writing instructions: (1) a *natural* approach where they used embedded clauses linked together by words indicating spatial relationships such as in the instruction "Pick up the yellow block directly to the left of the rightmost blue block.", (2) an *algorithmic* approach, which a majority of the users employed, where they partitioned their description in stages reflecting a visual search process such as in the instruction "There is an orange block on the right side of the table. Next to this orange block is a yellow block. Please pick up the yellow block touching the yellow block", (3) an *active language* approach where they provided instructions asking the partner to move their arms (usually) in a certain way so as to grasp the desired object such as in the instruction "Stand up, reach out over the table, and grab the yellow block that is touching the blue block closest to me.". In certain instructions, the participant would even offer active guidance (which is of course not an option in a one shot response written in a web form).

Among the three, the algorithmic approach is often the clearest but feels less natural. We believe that these observations about instruction approach types will lend themselves well to further investigation on user instruction preferences. For example, some users might prefer to give algorithmic descriptions which iteratively reduce ambiguity as needed, while other users might prefer to utilize active language where they guide the robots motions via iterative movement-

driven instruction.

Our findings suggest that sentence clarity suffers when there is either an ambiguity related to the number of blocks a sentence can specify or an ambiguity related to perspective. An interesting observation is the relationship between block ambiguity and perspective ambiguity. Because the process we used in coding the data did not exclude one from the others, it was highly possible that these two features were dependent although the Pearson correlation indicated the opposite ($r = -0.0287$). Perspective ambiguity will often results in block ambiguity, except in the case that there features in the instructions that are dominate enough to eliminate all the possible blocks aside from one. For example, in "It is the block all the way on the right side by itself", the perspective is unknown but only one block in the scene is identifiable since it is the only "by itself" candidate. In this case, we can reduce the instruction to "It is the block by itself". On the other hand, block ambiguity does not always imply unknown perspective. For example, in "pick up the closest green block", although the perspective is neither, not unknown, there are actually multiple possible blocks inferred from the instruction due to ambiguity in the landmark being referred to (e.g. closest to what?).

Further, descriptions requiring perspective always seem to include terms like 'right', 'left', 'above' and 'below'. We shall classify these as directional relative spatial references. If establishing perspective proves to be difficult in a scenario, and a sentence can avoid using directional relative spatial references, the robot should prefer to avoid these kinds of descriptions. That is, if the robot is able to generate a description using our definition of 'neither' perspective, it should prefer to do so over other descriptive strategies.

The intention of this work is to establish a baseline understanding of human preferences and behaviors when giving manipulation scenario instructions to a robot. We identify this one-shot language data analysis as a necessary step in laying the foundation for a truly interactive system which might take multiple rounds of input or ask questions to reduce uncertainty. Even without the element of active conversing, however, the results and insights we were able to extract are rather encouraging and have allowed us to establish effective grounding. We intend to gradually introduce interactivity in future works with varying approaches and modalities, and we believe the work we present in this paper provides an excellent initial benchmark.

REFERENCES

[1] R. A. Knepper, T. Layton, J. Romanishin, and D. Rus, "Ikeabot: An autonomous multi-robot coordinated furniture assembly system," in *International Conference on Robotics and Automation*. IEEE, 2013, pp. 855–862.

[2] H. R. Everett, D. W. Gage, G. A. Gilbreath, R. T. Laird, and R. P. Smurlo, "Real-world issues in warehouse navigation," in *Photonics for Industrial Applications*. International Society for Optics and Photonics, 1995, pp. 249–259.

[3] S. Ishii, S. Tanaka, and F. Hiramatsu, "Meal assistance robot for severely handicapped people," in *International Conference on Robotics and Automation*, vol. 2. IEEE, 1995, pp. 1308–1313.

[4] M. Skubic, D. Perzanowski, S. Blisard, A. Schultz, W. Adams, M. Bugajska, and D. Brock, "Spatial language for human-robot dialogs," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 34, no. 2, pp. 154–167, 2004.

[5] S. N. Blisard and M. Skubic, "Modeling spatial referencing language for human-robot interaction," in *IEEE International Workshop on Robot and Human Interactive Communication*, 2005, pp. 698–703.

[6] M. MacMahon, B. Stankiewicz, and B. Kuipers, "Walk the talk: Connecting language, knowledge, and action in route instructions," in *National Conference on Artificial Intelligence*, 2006.

[7] T. Kollar, S. Tellex, D. Roy, and N. Roy, "Toward understanding natural language directions," in *ACM/IEEE International Conference on Human-robot Interaction*. IEEE Press, 2010, pp. 259–266.

[8] S. Tellex, T. Kollar, S. Dickerson, M. Walter, A. Banerjee, S. Teller, and N. Roy, "Understanding natural language commands for robotic navigation and mobile manipulation," in *National Conference on Artificial Intelligence*, 2011.

[9] T. M. Howard, S. Tellex, and N. Roy, "A natural language planner interface for mobile manipulators," in *International Conference on Robotics and Automation*. IEEE, 2014, pp. 6652–6659.

[10] J. G. Trafton, N. L. Cassimatis, M. D. Bugajska, D. P. Brock, F. E. Mintz, and A. C. Schultz, "Enabling effective human-robot interaction using perspective-taking in robots," *IEEE Transactions on Systems, Man and Cybernetics, Part A (Systems and Humans)*, vol. 35, no. 4, pp. 460–470, 2005.

[11] Y. Bisk, D. Marcu, and W. Wong, "Towards a dataset for human computer communication via grounded language acquisition," in *AAAI Workshop on Symbiotic Cognitive Systems*, 2016.

[12] J. M. Wolfe, "Guided search 2.0 a revised model of visual search," *Psychonomic bulletin & review*, vol. 1, no. 2, pp. 202–238, 1994.

[13] R. Moratz, K. Fischer, and T. Tenbrink, "Cognitive modeling of spatial reference for human-robot interaction," *International Journal on Artificial Intelligence Tools*, vol. 10, no. 04, pp. 589–611, 2001.

[14] S. Werner, B. Krieg-Brückner, H. A. Mallot, K. Schweizer, and C. Freksa, "Spatial cognition: The role of landmark, route, and survey knowledge in human and robot navigation," in *Informatik als Innovationsmotor*. Springer, 1997, pp. 41–50.

[15] N. Franklin, B. Tversky, and V. Coon, "Switching points of view in spatial mental models," *Memory & Cognition*, vol. 20, no. 5, pp. 507–518, 1992.

[16] H. A. Taylor and B. Tversky, "Perspective in spatial descriptions," *Journal of memory and language*, vol. 35, no. 3, pp. 371–391, 1996.

[17] W. J. Levelt, "Perspective taking and ellipsis in spatial descriptions," *Language and space*, pp. 77–107, 1996.

[18] S. C. Levinson, "Frames of reference and molyneuxs question: Crosslinguistic evidence," *Language and space*, pp. 109–169, 1996.

[19] S. Guadarrama, L. Riano, D. Golland, D. Gouhring, Y. Jia, D. Klein, P. Abbeel, and T. Darrell, "Grounding spatial relations for human-robot interaction," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2013, pp. 1640–1647.

[20] D. K. Misra, J. Sung, K. Lee, and A. Saxena, "Tell me dave: Context-sensitive grounding of natural language to manipulation instructions," in *Robotics: Science and Systems*, 2014.

[21] K. Fischer, "The role of users? concepts of the robot in human-robot spatial instruction," in *Spatial Cognition V Reasoning, Action, Interaction*. Springer, 2006, pp. 76–89.

[22] J. G. Trafton, A. C. Schultz, M. Bugajska, and F. Mintz, "Perspective-taking with robots: experiments and models," in *IEEE International Workshop on Robot and Human Interactive Communication*. IEEE, 2005, pp. 580–584.

[23] R. Moratz and T. Tenbrink, "Spatial reference in linguistic human-robot interaction: Iterative, empirically supported development of a model of projective relations," *Spatial cognition and computation*, vol. 6, no. 1, pp. 63–107, 2006.

[24] M. F. Schober, "Spatial perspective taking in conversation," *Cognition*, vol. 47, pp. 1–24, 1993.

[25] K. Fischer and R. Moratz, "From communicative strategies to cognitive modelling," in *Workshop Epigenetic Robotics*, 2001.

[26] Y. Wei, E. Brunskill, T. Kollar, and N. Roy, "Where to go: Interpreting natural directions using global inference," in *International Conference on Robotics and Automation*. IEEE, 2009, pp. 3761–3767.

[27] L. Carlson, M. Skubic, J. Miller, Z. Huo, and T. Alexenko, "Strategies for human-driven robot comprehension of spatial descriptions by older adults in a robot fetch task," *Topics in cognitive science*, vol. 6, no. 3, pp. 513–533, 2014.