PERSPECTIVE IN NATURAL LANGUAGE INSTRUCTIONS FOR COLLABORATIVE MANIPULATIONAND GESTALT PRINCIPLES

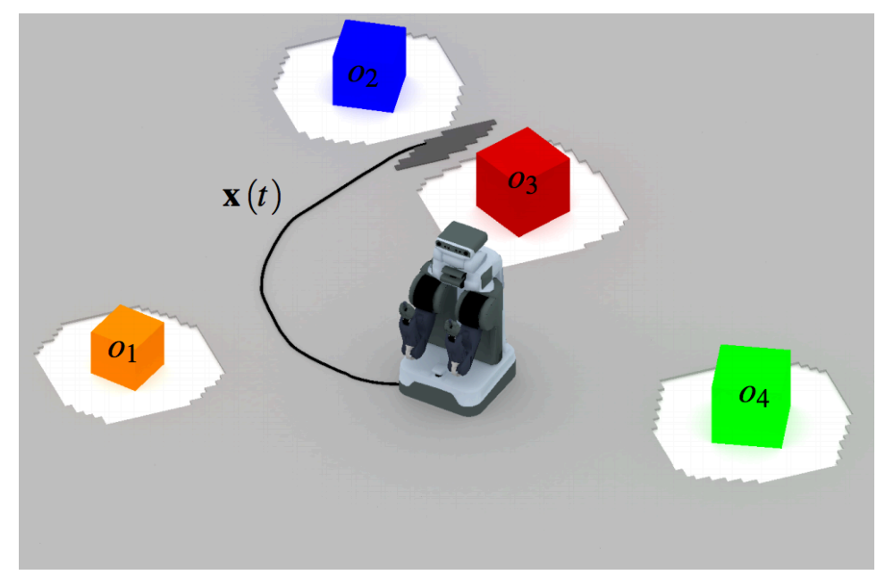
SHEN LI*, ROSARIO SCALISE*, HENNY ADMONI, STEPHANIE ROSENTHAL, SIDDHARTHA S. SRINIVASA

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 visual and spatial features in human spatial references and then focus on the clarity of instructions with respect to perspective taking. 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.

MOTIVATION OF PROBLEM

- Human-robot collaboration on spatial tasks, such as furniture assembly, warehouse automation, and meal serving
- Many research on spatial reference communication in **navigation** task
  - "move near the red box and the blue crate" [1]
- Less research on spatial reference communication in **tabletop** task
  - clutter



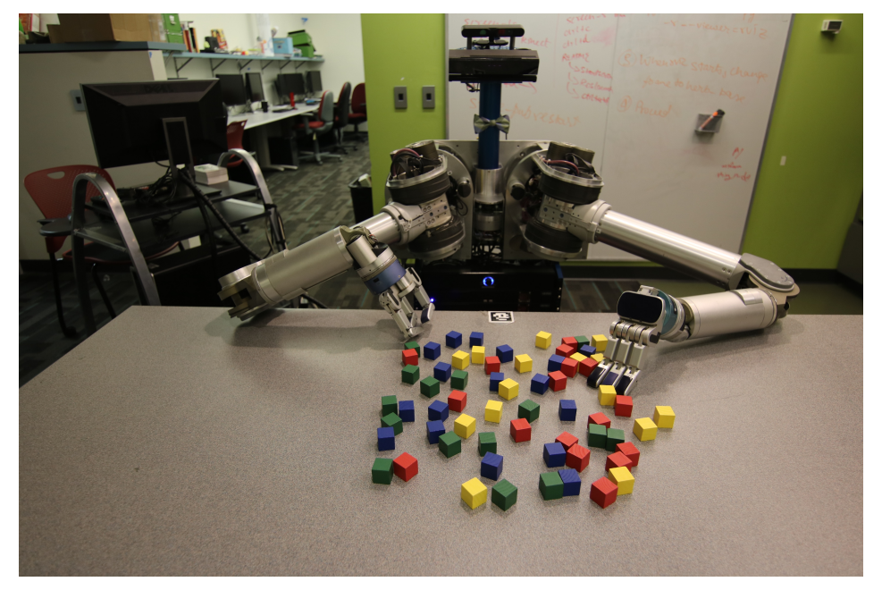
Left: Navigation task [1]

Right: Tabletop task [2]

- Object uniqueness problem
  - Visual features – red cup
  - Spatial references – on the left
    - Perspective – your
- Block task
  - elicit instructions that rely more on the visual and spatial features

Left: Cups in shelf

Right: Block task



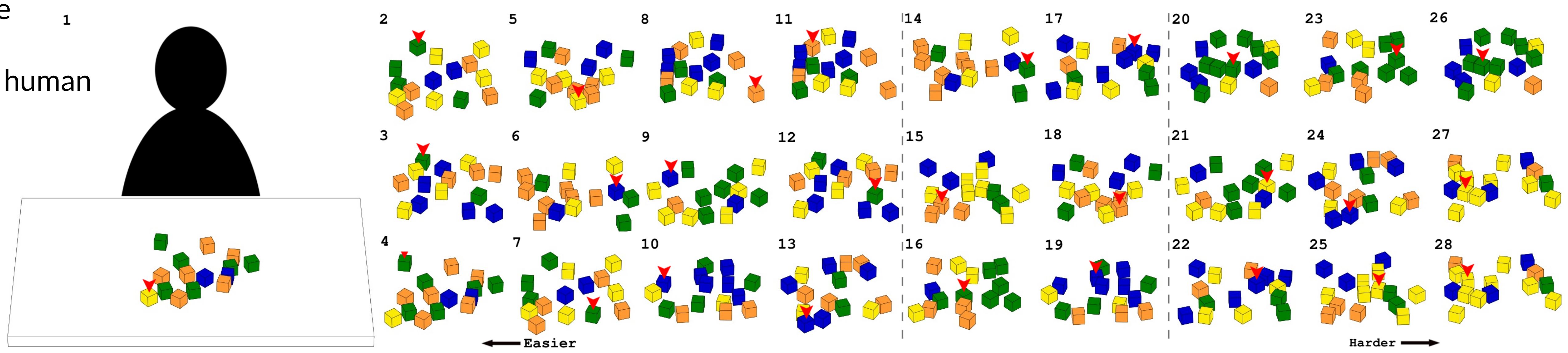
APPROACH

Study 1: Collecting Language Examples

- Amazon's Mechanical Turk
  - Instructions
  - Subjective Difficulty
  - Completion Time
- Within subject design
  - robot partner vs human partner

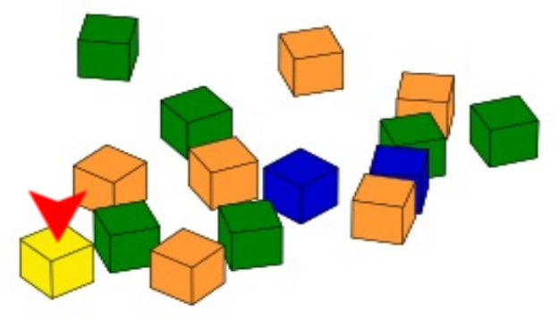
Study 2: Evaluating Language for Clarity

- Amazon's Mechanical Turk
  - Accuracy
  - Completion Time

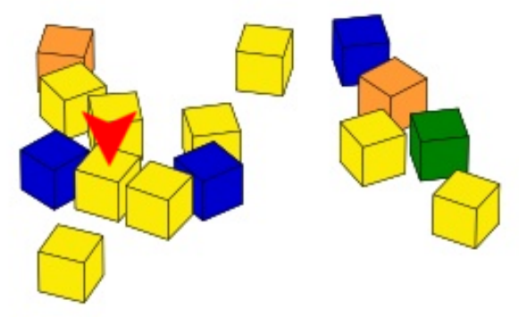


RESULTS

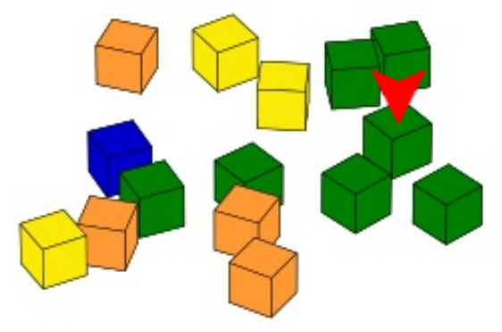
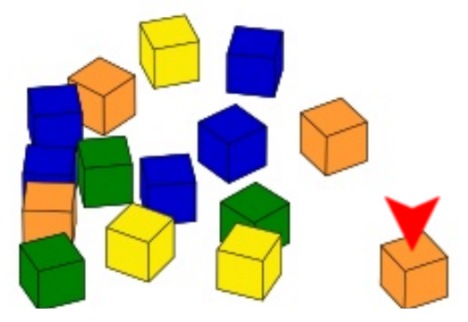
- **Robot Partner vs Human Partner**
  - Partner's perspective is used significantly more frequently in human-partner condition (28.1% of sentences) than in robot-partner condition (20.6% of sentences). P = 0.001
- **Visual feature**
  - Visual search
    - the routine visual behavior to find one object in a visual world filled with other distracting items [5]
    - Color, Stereoscopic depth (front and back), Line arrangement, Curvature, Intersection, Terminator
  - Visual feature in our task
    - Color
      - "green", "yellow", "blue", "orange"
      - Pick up the **yellow** block.



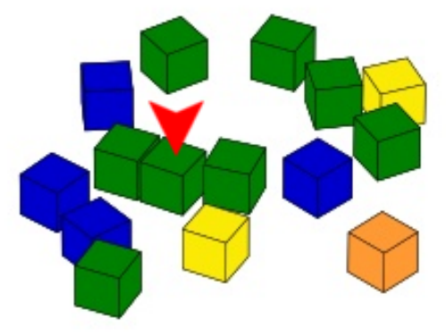
- Density – Terminator
  - edge, end, side, corner
  - Find the cluster of yellow blocks on your right. Pick up the yellow block that lays directly on the left side of the blue block, at a slight upward angle from the **edge** of the desk.



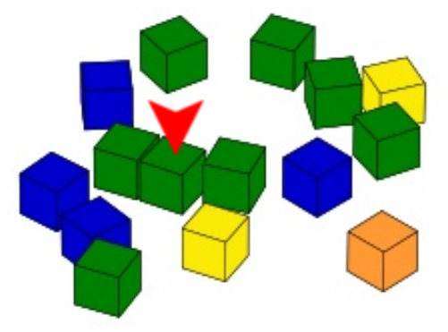
- Density - Intersection
  - cluster, pair, surround, sandwiched
  - pick up the middle green block from the **group** of 5
  - isolated, alone, apart, solitary
  - The **isolated** orange block on the corner

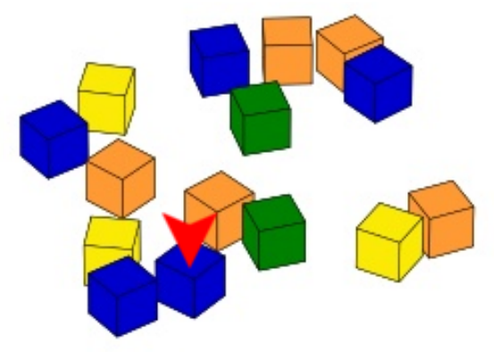
- Shape, pattern - Line
  - row, aligned, column, string, stack
  - Do you see the string of 3 green blocks in the center, near the yellow block near me? Grab the green block in the center of that **line** up.



- Shape, pattern - Curvature
  - diamond, rectangle, triangle, square
  - Look for a green block. Look for a green block that is very close to another green block. The green blocks should look like they combine to form a **rectangle**. Pick up the left most block of those two.



- Ordering, quantity
  - The **third** blue block from my left



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

Visual feature frequencies [3]

- **Data Coding**
  - Block ambiguity
    - 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.
  - Perspective

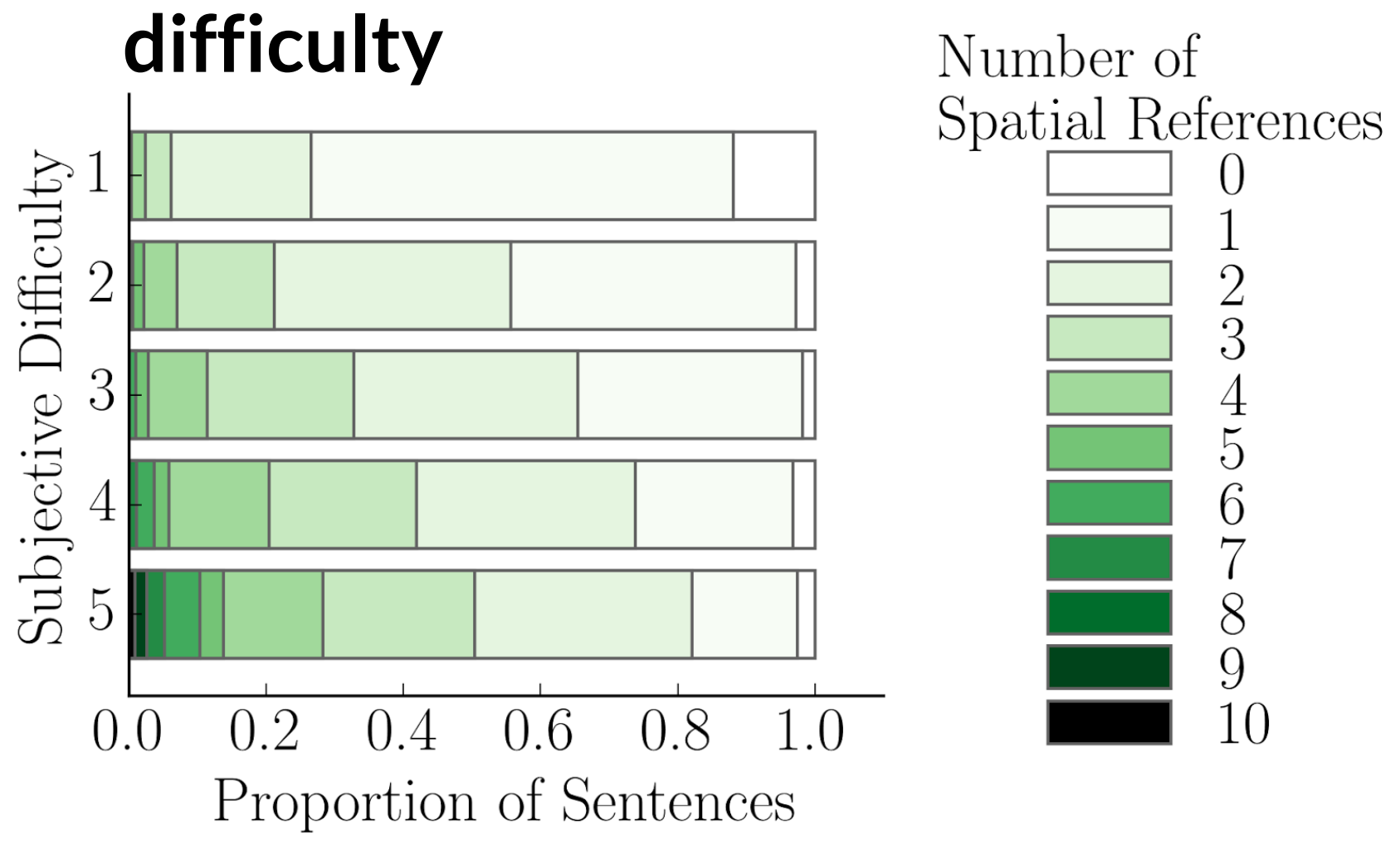
Type	P1	P2	Example
Participant Perspective	+	-	"the block that is to **my** rightest." "**my** left most blue block"
Partner Perspective	-	+	"the block on **your** left" "second from the right from **your** view'
Neither Perspective	-	-	"closest to you" "the top one in a triangle formation"
Unknown Perspective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

Possible perspectives (P1 = Participant; P2 = Partner) [8]

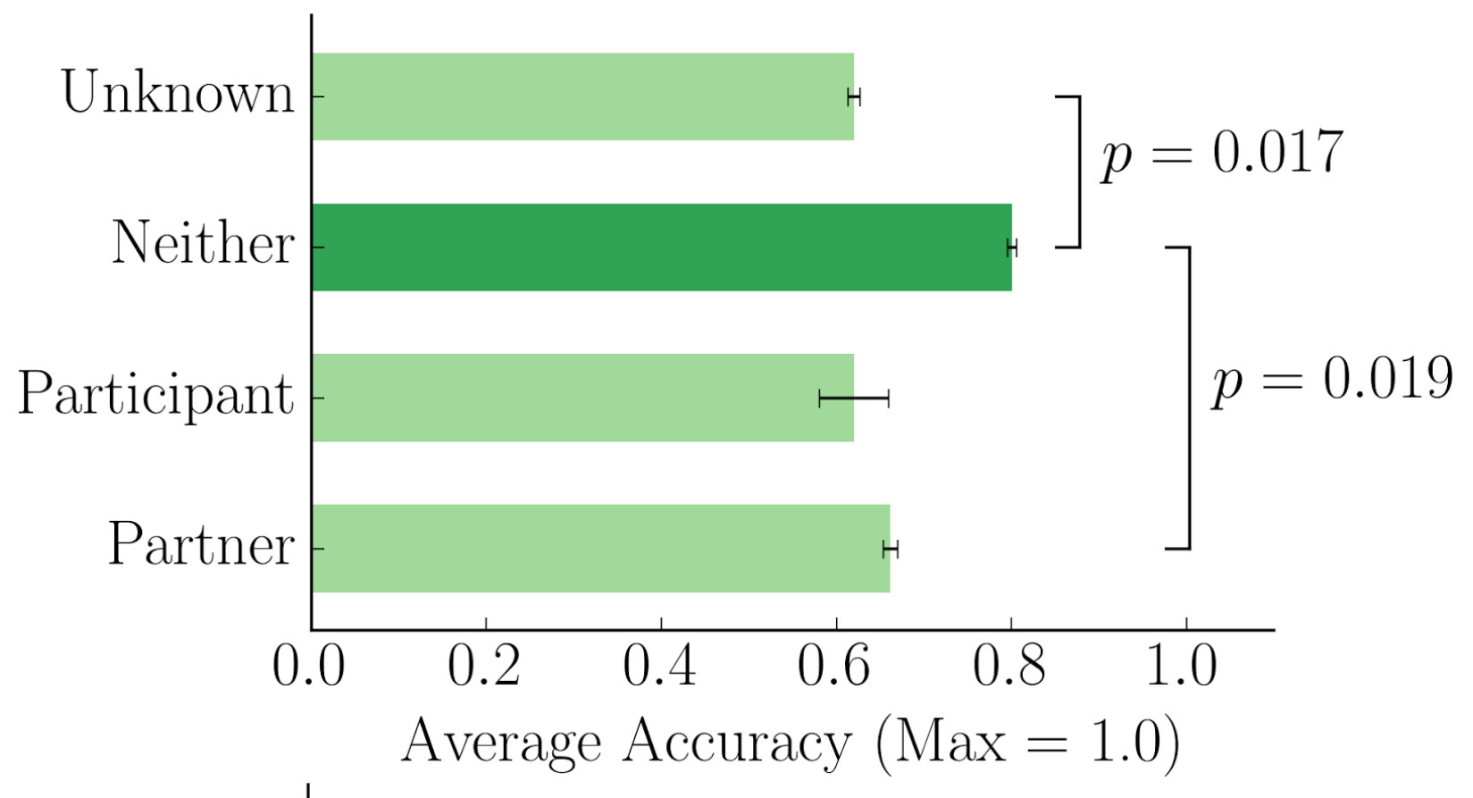
Perspective Type	Count	Percentage
Participant Perspective	15	1.07%
Partner Perspective	339	24.21%
Neither Perspective	592	42.29%
Unknown Perspective	454	32.43%
Total	1400	100%

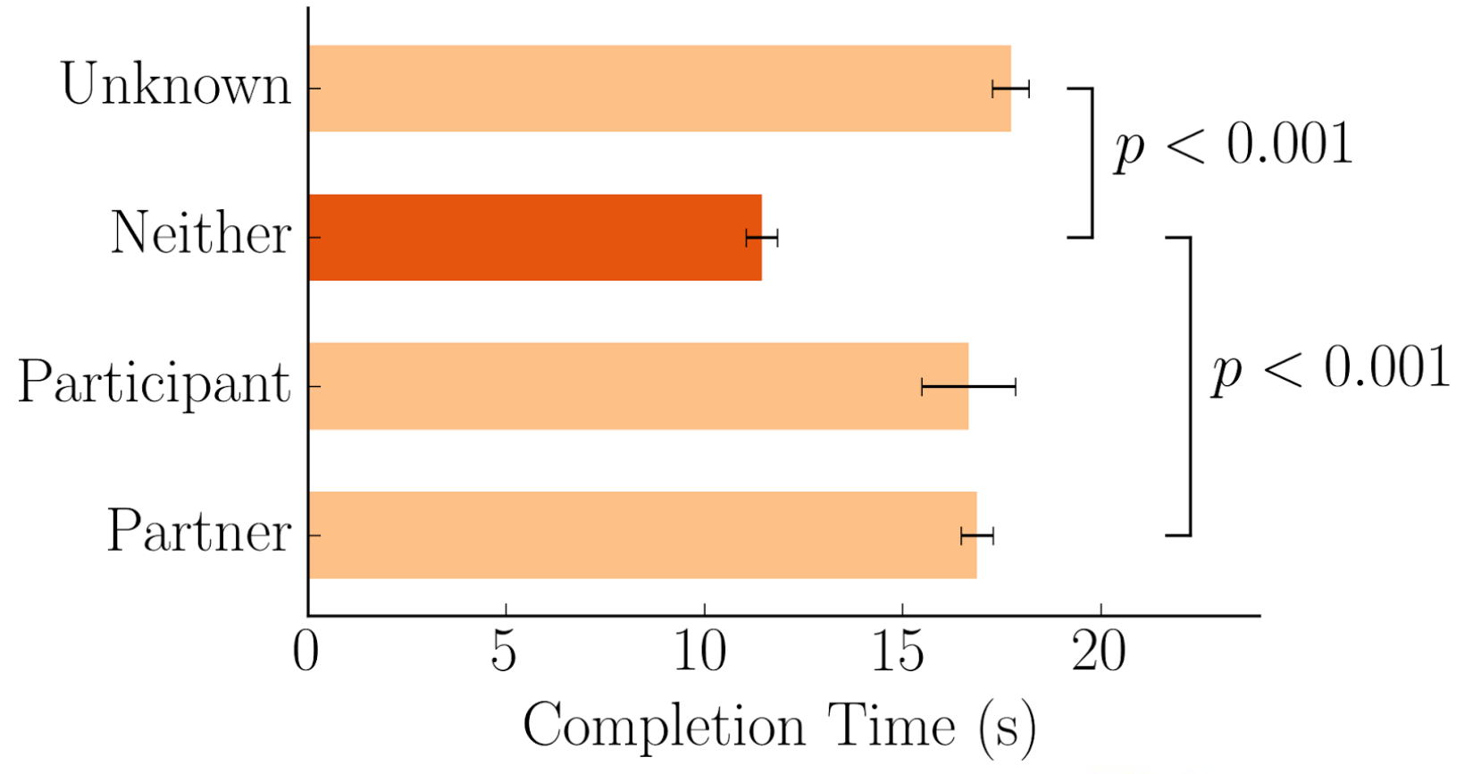
Number of each perspective (P1=Participant P2=Partner) [9]

- **Spatial references and subjective difficulty**



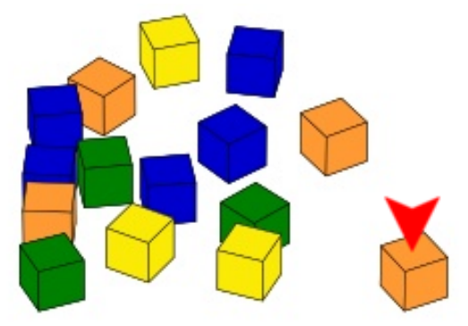
- **Perspective vs completion time and word count**





DISCUSSION

- **Prefer neither perspective**
  - Other factors, e.g. conciseness
- **Ambiguity and unknown perspective**
  - Unknown perspective usually implies ambiguity
    - Exceptions: dominant features
      - It is the block all the way on the right side **by itself**.
  - Ambiguity does not necessarily imply unknown perspective
    - Reasons: spatial reference ambiguity
      - Pick up the **nearest** orange block. **???**



- **Instructions**
  - **Natural approach**
  - **Algorithmic approach: To your right, there's an orange block next to a string of yellow blocks. Pick up the second yellow from the orange in this string.**
  - **Active language approach: Use your right arm and put it on the edge of the table straight ahead. Slide your hand sideways until you feel 3 blocks. Grab the middle one.**

REFERENCES

1. 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.
2. https://clumpaday.com/2014/08/page/2/
3. Li, S, Scalise, R, Admoni, H, Rosenthal, S, and Srinivasa, S (2016). Spatial References and Perspective in Natural Language Instructions for Collaborative Manipulation. In:Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN).
4. Li, S, Scalise, R, Admoni, H, Rosenthal, S, and Srinivasa, S (2016). Perspective in Natural Language Instructions for Collaborative Manipulation. In: Robotics: Science and Systems workshop on Model Learning for Human-Robot Communication.
5. Visual search: J. M. Wolfe, "Guided search 2.0 a revised model of visual search," Psychonomic bulletin & review, vol. 1, no. 2, pp. 202–238, 1994.