TEAM3 Challenge: Tasks for Multi-Human and Multi-Robot Collaboration with Voice and Gestures

Michael J. Munje
College of Computing
Georgia Institute of Technology
Atlanta, GA, USA
michaelmunje@gatech.edu

Lylybell K. Teran
Data Science Institute
Columbia University
New York, NY, USA
klt2162@columbia.edu

Bradon Thymes
Dept. of Computer Science
Cornell University
Ithaca, NY, USA
bmt63@cornell.edu

Joseph P. Salisbury
Open Innovation Center
Riverside Research
Lexington, MA, USA
jsalisbury@riversideresearch.org

ABSTRACT

Intuitive human-robot collaboration requires adaptive modalities for humans and robots to communicate and learn from each other. For diverse teams of humans and robots to naturally collaborate on novel tasks, robots must be able to model roles for themselves and other team members, anticipate how team members may perceive their actions, and communicate back to team members to continuously promote inclusive team cohesion toward achieving a shared goal. Here, we describe a set of tasks for studying mixed multi-human and multi-robot teams with heterogenous roles to achieve joint goals through both voice and gestural interactions. Based around the cooperative game TEAM3, we specify a series of dyadic and triadic human-robot collaboration tasks that require both verbal and nonverbal communication to effectively accomplish. Task materials are inexpensive and provide methods for studying a diverse set of challenges associated with human-robot communication, learning, and perspective-taking.

CCS CONCEPTS

• **Computing methodologies** → **Cognitive robotics**; *Spatial and physical reasoning*; Reasoning about belief and knowledge; • **Human-centered computing** → **Interface design prototyping**; *Natural language interfaces*; Collaborative interaction.

KEYWORDS

non-dyadic HRI, interaction techniques in groups with robots, human-robot collaboration, robot-assisted tower construction, natural language interaction, interaction design and prototyping

ACM Reference format:

Michael J. Munje, Lylybell K. Teran, Bradon Thymes, Joseph P. Salisbury. 2023. TEAM3 Challenge: Tasks for Multi-Human and Multi-Robot Collaboration with Voice and Gestures. In *HRI '23 Companion, March 13-16, 2023, Stockholm, Sweden.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3568294.3580049

HRI '23 Companion, March 13–16, 2023, Stockholm, Sweden

ACM ISBN 978-1-4503-9970-8/23/03.
https://doi.org/10.1145/3568294.3580049

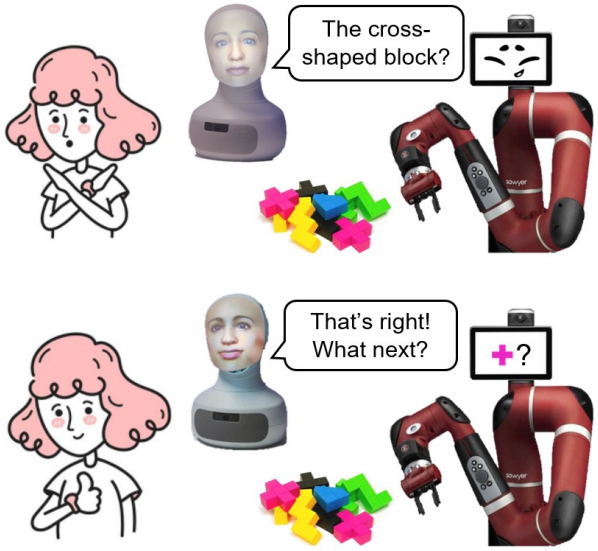

Figure 1: In the game TEAM3, three players with distinct roles work together to build a structure. Each role requires different skills that make TEAM3 an interesting challenge for triadic multi-human and multi-robot collaboration.

1 INTRODUCTION

The potential for multi-human multi-robot teams to form and collaborate on novel open-ended tasks offers both great opportunities and challenges [1]. For example, effective disaster response requires a diverse set of human expertise that could be supported by a variety of robot collaborators performing tasks such as search, rubble removal, structural inspection, medical support, evacuation, and logistics [2]. When time is critical, robots will be expected to adaptively respond to evolving team composition, changing priorities, and task requirements. To address this, robots need to have a representation of the perceived mental model of humans collaborators – an artificial theory of mind [3]–[5]. To determine the mental models of teammates through natural interactions, a language-capable robot must communicate effectively to establish common ground for collaboration [6]. For these interactions to be inclusive, robots must be able to adapt their approaches for communication and task allocation to meet the individualized needs of their teammates [7]–[11].

To develop and evaluate approaches for robust natural communication, we sought a collaborative task that requires teamwork with distinct roles that currently available robots could perform if they were provided sufficient communication and learning skills. Furthermore, we sought an inexpensive and easy to implement task to enable scalable collection of video data sets for machine learning. Here, we describe how materials from the collaborative tower building game TEAM3 [12] may be used to study communication and learning strategies in teams. Briefly, TEAM3 is played in groups of three players, with each player taking on one of three roles: the *Architect*, who knows what tower the team must build but can't speak; the *Builder*, who needs to build the tower but can't see; and the *Supervisor* in the middle who must facilitate communication between the two (Fig. 1).

Using TEAM3, human and human-robot collaboration can be studied using variously constrained communication (e.g., speech, gesture) and sensing (e.g., vision, listening, touch) modalities. We specify a set of dyadic and triadic interactions using TEAM3 for incrementally developing capabilities that would enable robots to participate in TEAM3. Finally, we describe a prototype implementation of a TEAM3 interaction using the robot Sawyer.

2 RELATED WORK

A fundamental aspect of communication is the ability to refer to objects [13]. Referring expression generation (REG) is a core component of many natural language generation systems [14]–[16]. When the robots' task domain can be defined, such as in warehouse logistics, natural language referring expressions can be interpreted and mapped to task specifications [17]. To perform new tasks in more open and dynamic environments, robots will need to be able to learn relevant information on the fly from humans, as in interactive task learning [18], [19]. For robots to communicate effectively during task learning and execution, research has been devoted to more challenging aspects of open-world REG [20]. This includes strategies for using context such as spatial relationships [21]–[23] and affect [24], perspective taking [25]–[27], and autonomously handling misinterpretations and ambiguities [28]–[30]. Interactive visual grounding of referring expressions (INGRESS) achieved object manipulation using natural language with perspective correction and interactive disambiguation unconstrained by object categories or expressions [31]. Language-conditioned learning (LANCON-LEARN) [32] uses an attention-based approach to learn new manipulation skills based on reasoning about relationships between skills and task objectives through natural language and interaction. ProgPrompt leverages a programmatic large language model prompt structure to enable plan generation across situated environments, robot capabilities, and tasks [33].

In addition to speech, the ability to recognize and produce nonverbal communication, such as gestures and eye gaze, can be used to reference objects and facilitate understanding between humans and robots. For example, robotic co-verbal gesture can increase listener attention and memory recall [34]. Implicit nonverbal communication can positively impact understandability of the robot, efficiency of task performance, and robustness of errors that arise from miscommunication [35]. Eye gaze can enable implicit communication, such as turn-taking cues, between humans and robots [36]–[40]. To build a large dataset of human gestures, charades can be used as a playful method to allow a robot to optimize its own gesture production and recognition abilities [41]. Personalized learning of gestures can also be achieved with a dialogue interaction [42].

For robots to learn new tasks from a human, robots must be able to develop a generalizable understanding of the task specification intended by the human and then be able to plan a sequence of actions to achieve task goals in the world. For training and evaluation of task learning and autonomous planning, block-based tower building provides a sequential task that is well-suited for HRI research. For example, tower building has been used to study goal inference based on past experiences and present context [43], flexible human-aware task planning [44], preferred interaction style with robots [45], the impact of a robot's allocation behavior on interpersonal dynamics and collaboration in groups [46], the influence of robot emotion expressions [47], and human linguistic forms to refer to objects that cannot be seen at the time of reference [48]. A multimodal reinforcement learning framework leveraging gestures and speech to communicate intention was utilized to collaboratively build a Jenga tower [49].

3 TASK GOALS

Our task design goal was to identify a collaborative task that could be used to: 1) facilitate the collection of a corpus of data on a collaborative construction task with both verbal and nonverbal object referring; 2) evaluate the performance of robots in a task that requires them to perform a specific role in a team through varying combinations of verbal and nonverbal communication and task performance. To better facilitate crowdsourced video collection of humans performing a task, we sought a game that was inexpensive. By utilizing an existing game, we could start to leverage video that was already available online from the gaming enthusiast community. Importantly, we wanted a game with rules that were specifically designed for cooperation, as opposed to competition. This was to avoid the need to modify instructions, which could lead to confusion and not allow us to leverage existing video available online.

To make the task feasible for current robots, we sought a game with pieces large enough to easily manipulate with a robotic arm with a gripper end-effector. The game should require object manipulation, including changes in object orientation, without requiring excessive fine motor control, dexterity, or balance. This was to help limit the need for extensively precise robotic control (e.g., assembling a house of cards) and maintain focus on higher level manipulations (e.g., "place object X upside-down on the left half of object Y…no no, not that far left"). Games that required tossing and catching were also avoided so precisely timed actions would not be necessary. Finally, we sought a game that required more than two players, to investigate non-dyadic interactions, without requiring more than 6 players, to avoid the challenges associated with recruiting large groups of people. From these requirements, the cooperative game TEAM3 was identified.

4 TASK AND TASK VARIANTS

4.1 Task Description

Each TEAM3 set includes two sets of five pieces – a pink X-shaped piece, a black T-shaped piece, a green S-shaped piece, a blue L-shaped piece, and a yellow W-shaped piece. Each piece's shape is composed of a set of 1×1-inch cubes, ranging from 3-5 cubic units. TEAM3 includes a set of "blueprint" cards that depict towers composed of a subset of pieces (for example, see Fig. 2 goal structure). Standard blueprints have only "flat" towers, where pieces stack so all cubic units are aligned in two dimensions. TEAM3 also includes three role cards – *Architect, Supervisor, and Builder*. Players select role cards before each round.

At the start of each round, the *Architect* draws a blueprint card that is kept hidden from the other players. The *Builder* closes their eyes and the game pieces are shuffled around on a table so they are within easy reach of the *Builder*. The *Architect* starts a 3-minute timer and begins the round by gesturing to the *Supervisor*, who will speak to the *Builder*. During gameplay, the *Architect* can't speak or make any verbal sound, and must only communicate in gestures. Gestures may include hand signals, facial expressions, clapping, etc. However, the *Architect* is not permitted to point at any construction piece on the table. The *Supervisor* interprets *Architect* gestures and may provide the *Builder* with verbal instruction. The *Builder* builds the tower with their eyes closed, following the *Supervisor*'s instructions.

4.2 Team Variations

For variations of TEAM3 where a robot may take on one or more TEAM3 role, we define each variation as a set T containing players P_j, where player P is an element of the set {H, R}, with H = a human player and R = a robot player, and $\mathcal{J}$ is the set of jobs (i.e., TEAM3 roles) that player has. More precisely, $\mathcal{J}$ is a nonempty subset of {A, S, B} where A = *Architect*, S = *Supervisor*, and B = *Builder*. Thus, $T = \{H_A, H_S, H_B\}$ represents standard TEAM3, with three humans having exactly one job each (n.b., set notation omitted around elements of $\mathcal{J}$ for brevity). If we prevent any two players in T from having the same job and require all jobs be assigned to someone, then there are two sets with only one player, twelve dyadic team sets, and eight triadic team sets.

4.3 Dyadic Teams

For dyadic teams, there are three job set pairings (Table 1):

Architect-Supervisor and Builder: In this pairing, the first player knows the target structure and must communicate that information to the second player to build. While the traditional *Architect* could only use nonverbal gestures, the *Architect-Supervisor* can communicate verbally, as the traditional *Supervisor* has this ability, and we suggest abilities be additive in this case. When a robot is *Architect-Supervisor*, it must guide a human who can't see to build the tower using verbal communication. When a robot is *Builder*, it builds the tower following human speech and without optical sensors.

Table 1: Dyadic ("Two-Player") Teams

P1	P2	Team Variants
A+S	B	$\{H_{A,S}, H_B\}$, $\{H_{A,S}, R_B\}$, $\{R_{A,S}, H_B\}$, $\{R_{A,S}, R_B\}$
A	S+B	$\{H_A, H_{S,B}\}$, $\{H_A, R_{S,B}\}$, $\{R_A, H_{S,B}\}$, $\{R_A, R_{S,B}\}$
S	A+B	$\{H_S, H_{A,B}\}$, $\{H_S, R_{A,B}\}$, $\{R_S, H_{A,B}\}$, $\{R_S, R_{A,B}\}$

Architect and Supervisor-Builder: In this pairing, the first player knows the target structure and must communicate to the second player through gestures. If abilities are additive, the second player can speak and see while building the tower, although a potential variation could allow the *Supervisor-Builder* to see the gestures of the *Architect*, but not be able to see the tower they were building. When a robot is *Architect*, it must guide the *Builder* to build the tower using only gestures. When a robot is *Supervisor-Builder*, it builds the tower following *Architect* gestures, with optical sensing potentially limited to viewing the *Architect* but not the blocks.

Supervisor and Architect-Builder: Here, the *Architect-Builder* knows the goal structure but cannot see the game pieces. Restricting the *Architect-Builder* from communicating verbally to the *Supervisor* could be interesting to consider.

4.4 Triadic Teams

Triadic teams can be categorized by the number of robots involved (Table 2). Considering a robot in each role:

Robot as Architect: The robot provides gestural feedback to the *Supervisor* on how to guide the *Builder*. The robot *Architect* will need to adjust gestural feedback based on both how the *Supervisor* is interpreting its gestures, as well as the state of the *Builder*.

Robot as Supervisor: The robot interprets *Architect* gestures to build an internal model of what the tower should be, and then provides the *Builder* with instructions for how to achieve that. The robot *Supervisor* will need to adapt its instructions to *Builder* actions if the *Builder's* actions are not producing a tower that matches the robot's internal model of what the tower should be.

Robot as Builder: Robot's actions are driven by *Architect* speech. However, as the robot *Builder* acts, if it maintains an internal model of the tower's state, it will be able to avoid collisions and understand commands such as placing a block atop another.

Table 2: Triadic ("Three-Player") Teams

# Robots	Team Variants
0	$\{H_A, H_S, H_B\}$
1	$\{H_A, H_S, R_B\}$, $\{H_A, R_S, H_B\}$, $\{R_A, H_S, H_B\}$
2	$\{H_A, R_S, R_B\}$, $\{R_A, R_S, H_B\}$, $\{R_A, H_S, R_B\}$
3	$\{R_A, R_S, R_B\}$

5 PROTOTYPE DEVELOPMENT

5.1 Motivation

Having identified TEAM3 as a collaborative task useful for studying effective verbal and nonverbal referring expression behavior, we propose TEAM3 as a challenge task for human/multi-robot and multi-human/robot interactions. To demonstrate human-level TEAM3 play, robots must 1) learn how to play TEAM3 from human instruction and then 2) be effective in any of the three TEAM3 roles. While our longer-term goal is to be able to explain TEAM3 to a robot and then evaluate its ability to achieve human performance levels, we first sought to establish a robot could perform in a TEAM3 role with a domain-specific set of actions. We focused on demonstrating a robot in the *Builder* role, using a Sawyer robot (Rethink Robotics) with a microphone and speakers to enable voice communication.

5.2 Interaction Design

When humans perform the *Builder* role, they benefit from the ability to search for and identify specific blocks from touch alone. While we could provide a series of verbal commands to the robot to be able to locate and grasp specific blocks on the table, we hypothesized the subsequent manipulation of blocks via voice commands to a specific orientation to build the tower would be time-intensive and frustrating. Rather than have the robot continuously manipulate objects live, we utilized Sawyer's face display to show a representation of the robot's internal model of the goal structure (Fig. 2). Using this approach, we propose an interaction flow where: 1) the goal structure is specified using voice commands on the display screen; 2) once the goal structure is complete, the robot's end effector is guided by voice commands to locations of blocks on the table; 3) as each block is reached and grasped, knowledge of the orientation of the block is provided to the robot by rotating a 3D block model on the display screen to fit its actual orientation in the gripper via voice commands; 4) once the orientation of the block in the end effector is determined, the robot can execute a motion plan to reorient it into its previously determined goal position in the target structure.

5.3 Implementation

A catkin workspace with Sawyer-specific dependencies [50] was setup on a development workstation running Ubuntu 20.04 LTS and ROS Noetic [51]. Demonstration code was developed in Python 3.2. For speech generation, Amazon Polly was used [52]. The child-like voice "Kevin" was selected based on a survey administered to eight people, which included three other voice choices with names removed. A set of computer speakers were affixed to the back of Sawyer's facial display to play voice clips. Similar to [53], we used the Python library SpeechRecognition [54] with PyAudio [55] and Google Speech Recognition. To provide Sawyer with animated facial expressions, we adapted code from [56]. A graphical interface for constructing TEAM3 towers using voice commands was developed using Turtle graphics [57].

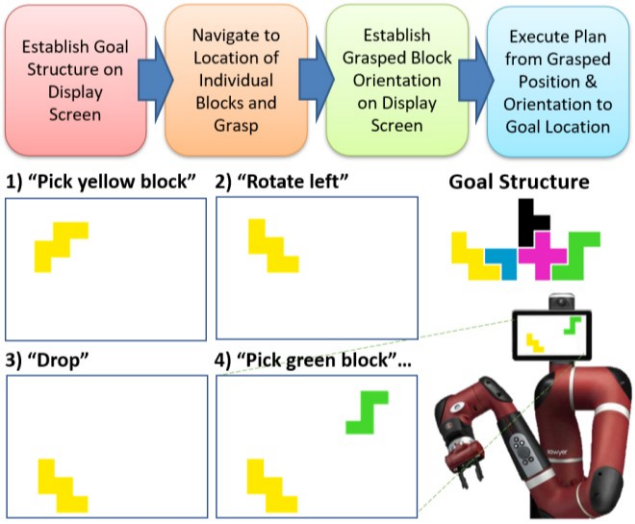

Figure 2: Speech-based TEAM3-specific interaction flow with robot as *Builder*.

5.4 Progress, Challenges, and Next Steps

Using the face display interface, we can build TEAM3 towers using a domain-specific set of voice commands to select blocks, set their position and rotation, and confirm when the tower is complete and ready to physically construct. Prior to implementing the physical block reorientation workflow, we examined Sawyer stacking blocks from set positions that didn't require reorientation. Sawyer was successful at building some TEAM3 towers, although blocks tended to shift upon release. As we improve placement consistency, we intend to develop the interaction workflow further in a simulation environment. To demonstrate a $\{H_A, R_S, R_B\}$ interaction with gesture recognition, we intend to incorporate a Furhat robot as *Supervisor*, as in Fig. 1.

6 CONCLUSION

We propose using TEAM3 as a tower construction task to collect verbal and gestural referring forms and study triadic interactions. TEAM3 requires minimal training, setup, and materials while demanding communication and coordination that is challenging even for humans. TEAM3 may be useful as a benchmarking task for both interactive task learning and intuitive human-robot communication for object manipulation and multi-task planning. TEAM3 variations may be utilized to develop foundational skills needed to achieve human levels of play without prior domain understanding. TEAM3 may also be useful for evaluating approaches for robots to adaptively communicate shared goals with teammates. TEAM3 requires this to be done across multiple teammates, each with personalized needs and responsibilities. Robots capable of adaptive multimodal communication will help promote more intuitive and inclusive human-robot teaming.

ACKNOWLEDGMENTS

Research support provided by the GEM Fellowship Program and Riverside Research.

REFERENCES

[1] E. Schneiders, "Non-Dyadic Human-Robot Interaction: Concepts and Interaction Techniques," in *Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction*, Sapporo, Hokkaido, Japan, Mar. 2022, pp. 1176–1178.

[2] R. R. Murphy, *Disaster robotics*. MIT press, 2014.

[3] N. C. Rabinowitz, F. Perbet, H. F. Song, C. Zhang, S. M. A. Eslami, and M. Botvinick, "Machine Theory of Mind," *ArXiv180207740 Cs*, Mar. 2018, Accessed: Feb. 07, 2022. [Online]. Available: http://arxiv.org/abs/1802.07740

[4] J. Williams, S. M. Fiore, and F. Jentsch, "Supporting Artificial Social Intelligence With Theory of Mind," *Front. Artif. Intell.*, vol. 5, p. 750763, Feb. 2022, doi: 10.3389/frai.2022.750763.

[5] T. Chakraborti, S. Sreedharan, and S. Kambhampati, "Human-Aware Planning Revisited: A Tale of Three Models," *IJCAI-ECAI XAIICAPS XAIP Workshop*, p. 10, 2018.

[6] P. Ramaraj, "Robots that Help Humans Build Better Mental Models of Robots," in *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, New York, NY, USA, Mar. 2021, pp. 595–597. doi: 10.1145/3434074.3446365.

[7] S. Janarthanam and O. Lemon, "Learning to Adapt to Unknown Users: Referring Expression Generation in Spoken Dialogue Systems," in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Uppsala, Sweden, Jul. 2010, pp. 69–78. Accessed: Nov. 18, 2022. [Online]. Available: https://aclanthology.org/P10-1008

[8] S. Rossi, F. Ferland, and A. Tapus, "User profiling and behavioral adaptation for HRI: A survey," *Pattern Recognit. Lett.*, vol. 99, pp. 3–12, Nov. 2017, doi: 10.1016/j.patrec.2017.06.002.

[9] T. Munzer, M. Toussaint, and M. Lopes, "Preference learning on the execution of collaborative human-robot tasks," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*, May 2017, pp. 879–885. doi: 10.1109/ICRA.2017.7989108.

[10] A. Cunha *et al.*, "Towards Collaborative Robots as Intelligent Co-workers in Human-Robot Joint Tasks: what to do and who does it?," in *ISR 2020; 52th International Symposium on Robotics*, Dec. 2020, pp. 1–8.

[11] N. Monaikul, B. Abbasi, Z. Rysbek, B. Di Eugenio, and M. Žefran, "Role Switching in Task-Oriented Multimodal Human-Robot Collaboration," in *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, Aug. 2020, pp. 1150–1156. doi: 10.1109/RO-MAN47096.2020.9223461.

[12] "TEAM3," *braingames*. https://www.publishing.brain-games.com/team3 (accessed Nov. 22, 2022).

[13] K. van Deemter, *Computational Models of Referring: A Study in Cognitive Science*. MIT Press, 2016.

[14] R. Dale and E. Reiter, "Computational interpretations of the Gricean maxims in the generation of referring expressions," *Cogn. Sci.*, vol. 19, no. 2, pp. 233–263, Apr. 1995, doi: 10.1016/0364-0213(95)90018-7.

[15] A. Gatt and E. Krahmer, "Survey of the State of the Art in Natural Language Generation: Core tasks, applications and evaluation," *J. Artif. Intell. Res.*, vol. 61, pp. 65–170, Jan. 2018, doi: 10.1613/jair.5477.

[16] T. C. Ferreira, D. Moussallem, Á. Kádár, S. Wubben, and E. Krahmer, "NeuralREG: An end-to-end approach to referring expression generation." arXiv, May 21, 2018. Accessed: Nov. 17, 2022. [Online]. Available: http://arxiv.org/abs/1805.08093

[17] P. Detzner, T. Kirks, and J. Jost, "A Novel Task Language for Natural Interaction in Human-Robot Systems for Warehouse Logistics," in *2019 14th International Conference on Computer Science & Education (ICCSE)*, Aug. 2019, pp. 725–730. doi: 10.1109/ICCSE.2019.8845336.

[18] P. Ramaraj, M. Klenk, and S. Mohan, "Understanding intentions in human teaching to design interactive task learning robots," in *RSS 2020 Workshop: AI & Its Alternatives in Assistive & Collaborative Robotics: Decoding Intent*, 2020.

[19] P. Ramaraj and J. E. Laird, "Establishing common ground for learning robots," in *RSS 2018: Workshop on Models and Representations for Natural Human-Robot Communication*, 2018.

[20] F. I. Doğan and I. Leite, "Open Challenges on Generating Referring Expressions for Human-Robot Interaction." arXiv, Apr. 19, 2021. Accessed: Nov. 18, 2022. [Online]. Available: http://arxiv.org/abs/2104.09193

[21] Z. Huo and M. Skubic, "Natural Spatial Description Generation for Human-Robot Interaction in Indoor Environments," in *2016 IEEE International Conference on Smart Computing (SMARTCOMP)*, May 2016, pp. 1–3. doi: 10.1109/SMARTCOMP.2016.7501708.

[22] R. Hu, M. Rohrbach, J. Andreas, T. Darrell, and K. Saenko, "Modeling Relationships in Referential Expressions with Compositional Modular Networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, Jul. 2017, pp. 4418–4427. doi: 10.1109/CVPR.2017.470.

[23] R. Fang, M. Doering, and J. Chai, "Collaborative Models for Referring Expression Generation in Situated Dialogue," *Proc. AAAI Conf. Artif. Intell.*, vol. 28, no. 1, Art. no. 1, Jun. 2014, doi: 10.1609/aaai.v28i1.8934.

[24] S. A. Akgun, M. Ghafurian, M. Crowley, and K. Dautenhahn, "Using Affect as a Communication Modality to Improve Human-Robot Communication in Robot-Assisted Search and Rescue Scenarios." arXiv, Aug. 19, 2022. Accessed: Nov. 20, 2022. [Online]. Available: http://arxiv.org/abs/2208.09580

[25] M. Liu, C. Xiao, and C. Chen, "Perspective-Corrected Spatial Referring Expression Generation for Human–Robot Interaction," *IEEE Trans. Syst. Man Cybern. Syst.*, vol. 52, no. 12, pp. 7654–7666, Dec. 2022, doi: 10.1109/TSMC.2022.3161588.

[26] M. Berlin, "Perspective Taking: An Organizing Principle for Learning in Human-Robot Interaction," p. 7.

[27] M. E. Foster, M. Giuliani, and A. Isard, "Task-based evaluation of context-sensitive referring expressions in human–robot dialogue," *Lang. Cogn. Neurosci.*, vol. 29, no. 8, pp. 1018–1034, Sep. 2014, doi: 10.1080/01690965.2013.855802.

[28] A. Tabrez, J. Kawell, and B. Hayes, "Asking the Right Questions: Facilitating Semantic Constraint Specification for Robot Skill Learning and Repair," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Sep. 2021, pp. 6217–6224. doi: 10.1109/IROS51168.2021.9636375.

[29] J. Hoelscher, D. Koert, J. Peters, and J. Pajarinen, "Utilizing Human Feedback in POMDP Execution and Specification," in *2018 IEEE-RAS 18th International Conference on Humanoid Robots (Humanoids)*, Nov. 2018, pp. 104–111. doi: 10.1109/HUMANOIDS.2018.8625022.

[30] F. I. Doğan, I. Torre, and I. Leite, "Asking Follow-Up Clarifications to Resolve Ambiguities in Human-Robot Conversation," in *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, Mar. 2022, pp. 461–469. doi: 10.1109/HRI53351.2022.9889368.

[31] M. Shridhar and D. Hsu, "Interactive Visual Grounding of Referring Expressions for Human-Robot Interaction." arXiv, Jun. 11, 2018. Accessed: Nov. 18, 2022. [Online]. Available: http://arxiv.org/abs/1806.03831

[32] A. Silva, N. Moorman, W. Silva, Z. Zaidi, N. Gopalan, and M. Gombolay, "LanCon-Learn: Learning With Language to Enable Generalization in Multi-Task Manipulation," *IEEE Robot. Autom. Lett.*, vol. 7, no. 2, pp. 1635–1642, Apr. 2022, doi: 10.1109/LRA.2021.3139667.

[33] I. Singh *et al.*, "ProgPrompt: Generating Situated Robot Task Plans using Large Language Models." arXiv, Sep. 22, 2022. doi: 10.48550/arXiv.2209.11302.

[34] P. Bremner, A. G. Pipe, C. Melhuish, M. Fraser, and S. Subramanian, "The effects of robot-performed co-verbal gesture on listener behaviour," in *2011 11th IEEE-RAS International Conference on Humanoid Robots*, Oct. 2011, pp. 458–465. doi: 10.1109/Humanoids.2011.6100810.

[35] C. Breazeal, C. D. Kidd, A. L. Thomaz, G. Hoffman, and M. Berlin, "Effects of nonverbal communication on efficiency and robustness in human-robot teamwork," in *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Aug. 2005, pp. 708–713. doi: 10.1109/IROS.2005.1545011.

[36] O. Palinko, F. Rea, G. Sandini, and A. Sciutti, "Eye gaze tracking for a humanoid robot," in *2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids)*, Nov. 2015, pp. 318–324. doi: 10.1109/HUMANOIDS.2015.7363561.

[37] N. F. Duarte, M. Raković, J. Marques, and J. Santos-Victor, "Action Alignment from Gaze Cues in Human-Human and Human-Robot Interaction," in *Computer Vision – ECCV 2018 Workshops*, vol. 11131, L. Leal-Taixé and S. Roth, Eds. Cham: Springer International Publishing, 2019, pp. 197–212. doi: 10.1007/978-3-030-11015-4_17.

[38] S. Lallée *et al.*, "Cooperative human robot interaction systems: IV. Communication of shared plans with Naïve humans using gaze and speech," in *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Nov. 2013, pp. 129–136. doi: 10.1109/IROS.2013.6696343.

[39] S. Gillet, M. T. Parreira, M. Vazquez, and I. Leite, "Learning Gaze Behaviors for Balancing Participation in Group Human-Robot Interactions," in *Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction*, 2022, p. 10.

[40] M. Moujahid, H. Hastie, and O. Lemon, "Multi-party Interaction with a Robot Receptionist," in *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, Mar. 2022, pp. 927–931. doi: 10.1109/HRI53351.2022.9889641.

[41] J. de Wit *et al.*, "Playing charades with a robot: collecting a large dataset of human gestures through HRI," in *Proceedings of the 14th ACM/IEEE International Conference on Human-Robot Interaction*, Daegu, Republic of Korea, Mar. 2019, pp. 634–635.

[42] H. Brock and R. Gomez, "Personalization of Human-Robot Gestural Communication through Voice Interaction Grounding," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Sep. 2021, pp. 846–853. doi: 10.1109/IROS51168.2021.9636105.

[43] V. Mohan and A. A. Bhat, "Joint Goal Human Robot collaboration-From Remembering to Inferring," *Procedia Comput. Sci.*, vol. 123, pp. 579–584, Jan. 2018, doi: 10.1016/j.procs.2018.01.089.

[44] S. Devin, A. Clodic, and R. Alami, "About Decisions During Human-Robot Shared Plan Achievement: Who Should Act and How?," in *Social Robotics*, Cham, 2017, pp. 453–463. doi: 10.1007/978-3-319-70022-9_45.

[45] R. Schulz, P. Kratzer, and M. Toussaint, "Preferred Interaction Styles for Human-Robot Collaboration Vary Over Tasks With Different Action Types," *Front. Neurorobotics*, vol. 12, 2018, Accessed: Nov. 20, 2022. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnbot.2018.00036

[46] M. F. Jung, D. Difranzo, S. Shen, B. Stoll, H. Claure, and A. Lawrence, "Robot-Assisted Tower Construction—A Method to Study the Impact of a Robot's Allocation Behavior on Interpersonal Dynamics and Collaboration in Groups," *ACM Trans. Hum.-Robot Interact.*, vol. 10, no. 1, pp. 1–23, Mar. 2021, doi: 10.1145/3394287.

[47] S. Zhou and L. Tian, "Would you help a sad robot? Influence of robots' emotional expressions on human-multi-robot collaboration," in *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, Aug. 2020, pp. 1243–1250. doi: 10.1109/RO-MAN47096.2020.9223524.

[48] Z. Han and T. Williams, "A Task Design for Studying Referring Behaviors for Linguistic HRI," in *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, Mar. 2022, pp. 783–786. doi: 10.1109/HRI53351.2022.9889512.

[49] Z. Cai, Z. Feng, L. Zhou, C. Ai, H. Shao, and X. Yang, "A Framework and Algorithm for Human-Robot Collaboration Based on Multimodal Reinforcement Learning," *Comput. Intell. Neurosci.*, vol. 2022, p. e2341898, Sep. 2022, doi: 10.1155/2022/2341898.

[50] "Sawyer Robot." Rethink Robotics GmbH, Sep. 03, 2022. Accessed: Nov. 14, 2022. [Online]. Available: https://github.com/RethinkRobotics/sawyer_robot

[51] "noetic - ROS Wiki." http://wiki.ros.org/noetic (accessed Nov. 09, 2022).

[52] "Text to Speech Software – Amazon Polly – Amazon Web Services," *Amazon Web Services, Inc.* https://aws.amazon.com/polly/ (accessed Nov. 09, 2022).

[53] R. Adamini *et al.*, "User-friendly human-robot interaction based on voice commands and visual systems," in *2021 24th International Conference on Mechatronics Technology (ICMT)*, Dec. 2021, pp. 1–5. doi: 10.1109/ICMT53429.2021.9687192.

[54] "SpeechRecognition: Library for performing speech recognition, with support for several engines and APIs, online and offline." [Online]. Available: https://github.com/Uberi/speech_recognition#readme

[55] "PyAudio: Cross-platform audio I/O with PortAudio." Accessed: Nov. 09, 2022. [Online]. Available: https://people.csail.mit.edu/hubert/pyaudio/

[56] "cwru-robotics/baxter_facial_animation." CWRU Robotics, Dec. 24, 2021. Accessed: Nov. 09, 2022. [Online]. Available: https://github.com/cwru-robotics/baxter_facial_animation

[57] "turtle — Turtle graphics — Python 3.11.0 documentation." https://docs.python.org/3/library/turtle.html (accessed Nov. 23, 2022).