Challenges in Annotating Gesture-Based Cognitive Status in Human-Robot Collaboration Datasets

Logan Daigler
logandaigler@mines.org
Colorado School of Mines
Golden, Colorado, USA

Mark Higger
Colorado School of Mines
Golden, Colorado, USA
mhigger@mines.org

Terran Mott
Colorado School of Mines
Golden, Colorado, USA
terranmott@mines.org

Tom Williams
Colorado School of Mines
Golden, Colorado, USA
twilliams@mines.org

ABSTRACT

For robots to be effective at collaborating with humans, they must be able to effectively communicate about entities in open-world tasks. Existing research on natural language generation and referring expression generation has yet to address how gesture and cognitive status impact how humans or robots decide how to refer to entities, a process known as Referring Form Selection. To address these issues we present a novel experimental testbed that leverages the Givenness Hierarchy to produce an entity's cognitive status. We also discuss challenges in developing this testbed and how we surmounted them.

CCS CONCEPTS

• **Computer systems organization → Robotics**; • **Human-centered computing → Human computer interaction (HCI)**.

KEYWORDS

human-robot collaboration, open world, gesture, referring form selection

ACM Reference Format:
Logan Daigler, Mark Higger, Terran Mott, and Tom Williams. 2024. Challenges in Annotating Gesture-Based Cognitive Status in Human-Robot Collaboration Datasets. In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction (HRI '24 Companion), March 11–14, 2024, Boulder, CO, USA*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3610978.3640584

1 MOTIVATION

In the future, robots will have to offer physical assistance in care settings [19], in the home [8, 11], and in collaborative manufacturing [29]. To engage in natural and effective interactions in these domains, robots will have to be capable of conversing with human users about shared tasks using both verbal and non-verbal cues to refer to elements of the task. Crucially, these tasks may take place in open-world environments, and as such, robots will need to have dialogue and knowledge systems capable of discussing, interacting with, and referring to locations or entities that are brought up by humans, even if those entities or locations are not known prior. For example, while working on some task, a human might require a robot teammate to "go and grab a screwdriver on that table" when no screwdrivers have come up in the current context of the conversation. In this situation, a robot should be capable of understanding what a screwdriver is, know which table the speaker is referring to if there are multiple screwdrivers, and be able to clarify which screwdriver to grab, either through dialogue or through some non-verbal gesture.

Fortunately, there has been significant work done in the fields of Natural Language Generation (NLG) and Referring expression generation (REG), which has allowed robots to generate language and refer to entities and locations in open-world environments [31]. However, some aspects of situated natural language generation remain underexplored, including Referring Form Selection [15, 22], in which a robot selects how to refer to an object. Current systems tend to over-refer to objects in situations where it would have been more natural to use "it" or "that" to refer to an object that had already been introduced to the context of the current conversation.

To address this challenge, some recent work on robotic referring form selection [15, 26] has looked to linguistic models that explain how humans choose referring forms such as the Givenness Hierarchy (GH) [14]. Central to the Givenness Hierarchy is the notion of *cognitive status*. An entity's cognitive status is determined by the assumptions that a cooperative speaker can make regarding their addressee's knowledge of and attention to that entity. This informs whether the entity can be referred to with words like *it* and *that*, or if it should be described more fully [14]. While there has been some work to accurately model or predict the cognitive status of entities [25, 26], those previous efforts have not accounted for non-verbal cues like gesture, which are critical tools humans use to manipulate cognitive status.

So, in this paper, we present the design for a testbed that will allow us to collect data about the cognitive status of entities based on both utterances and gestures made during a task-based conversation. We reflect on the challenges of designing a testbed to label Cognitive Status in this way and describe the design decisions we made to mitigate these challenges. Finally, we describe how this testbed will be used in future work to work towards more natural human-robot collaboration.

2 RELATED WORK

2.1 The Givenness Hierarchy

The Givenness Hierarchy is a cognitive science framework that relates to the way humans refer to topics, concepts or objects during an interaction. The Givenness Hierarchy itself is a hierarchical mapping of *cognitive statuses* assigned to entities that determine how likely and in what manner they might be referred to in conversation [14]. For example, an object with the cognitive status "in-focus" may be referred to as "it," as in "Robot, can you hand it to me?" Specifically, the Givebness Hierarchy is comprised of six

HRI '24 Companion, March 11–14, 2024, Boulder, CO, USA

ACM ISBN 979-8-4007-0323-2/24/03.
https://doi.org/10.1145/3610978.3640584

GH Level	Description	Reference Example
In Focus	Objects that are "In-focus" are the focal point of the conversation and are likely to be topics of subsequent utterances, they are in short-term memory.	*Can you hand it to me?*
Activated	Objects that are "Activated" are in short-term memory and are present in the context of a conversation	*Can you hand me that?*
Familiar	Objects that are "Familiar" are objects that can be uniquely identified either because they have been recently mentioned or because they exist within long-term memory	*Could you get me that block?*
Uniquely Identifiable	A "uniquely Identifiable" object can be identified based on the nominal alone or can be identified by associating it with a previously activated referent	*Pick up the rightmost green block*
Referential	An object is "Referential" if it is subsequently mentioned in the conversation or it is clear that the speaker meant to refer to a specific object	*So there was this green block I saw in a previous quadrant. Perhaps that's the one you need?*
Type Identifiable	An object is "Type Identifiable" if its description is understood	*Pass me a green cube*

Table 1: The Givenness Hierarchy

hierarchically nested categories of cognitive status, shown in Table 1. In this way, it can play a pivotal role in how we can generate referring expressions for robotics [33].

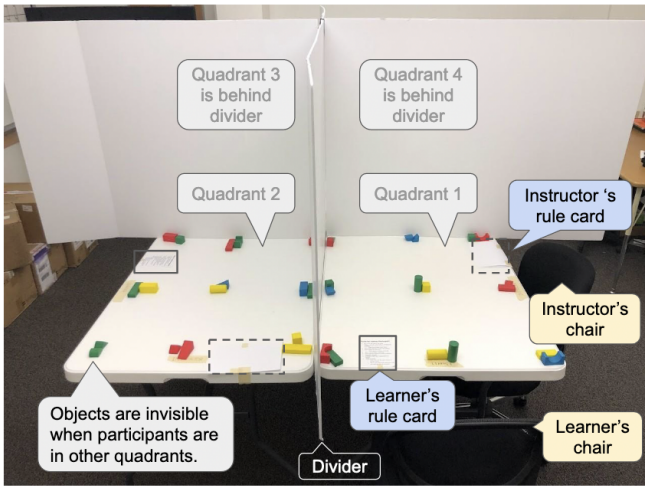

Figure 1: Setup for testing referring form selection from "Evaluating Referring Form Selection Models in Partially-Known Environments" [15]

2.2 The Givenness Hierarchy in HRI

HRI researchers have investigated how the Givenness Hierarchy can inform the design of effective, natural communication for robots doing collaborative tasks alongside humans. For example, Han et al. [15] present a methodological approach to studying how humans use the Givenness Hierarchy to communicate in collaborative tasks. The task environment used in Han et al. [15]'s work (shown in Figure 1) was partitioned into four quadrants, each containing a variety of colored blocks. Pairs of participants (an *instructor* and a *learner*) participated in a sequence of four building tasks in which one participant instructed the other in how to construct a specific "building" from the blocks. Each building task required some blocks that were unavailable in the current quadrant, such that the instructor was required to refer to blocks that were immediately visible, blocks that had been seen in previous quadrants, and blocks whose locations were as-yet-unknown. In this way, Han et al. [15] present a high-quality dataset of instances of open-world references to objects of various cognitive statuses within the Givenness Hierarchy.

2.3 Gesture & Reference in HRI

However, human references often use more than verbal language. In addition to verbal language, humans also use gestures to aid in generating clear, natural referring expressions [10, 21]. This type of non-verbal communication is critical for situated interaction [3, 4, 23]. Therefore, understanding how to interpret and generate gestures is a key component of multimodal human-robot collaboration [32]. Robots that can understand and use gesture are more likeable [20, 27] and more effective [1, 9, 12] collaborators.

Therefore, researchers have developed a variety of frameworks to categorize and represent gestures for human robot collaboration. Some frameworks focus on developing an exhaustive set of physical hand and arm gestures [2, 5], while other frameworks categorize gestures according to a small number of conversational roles [24]. HRI research shows that these frameworks can inform successful human-robot interaction [6, 18, 28]. However, these categorization schemes are limited in open-world environments, as open-world communication often gives rise to gestures that blur the boundaries between categories [7, 30]. Simple gestures in open worlds, such as pointing, can take on more complex meaning [7], including in open-world human-robot interactions [30].

3 METHOD

The goal of this work is to understand the ways that gestures inform, and are informed by, cognitive status, in the context of open-world

reference. Coming to this understanding may help us eventually develop ways for robots to both understand and generate natural, multimodal communication during collaborative tasks. To begin to explore this question, we leverage the dataset of references that resulted from Han et al. [15]'s open-world reference task, which contains a wide variety of open-world *gestures*.

We analyzed twelve videos from Han et al. [15]'s video dataset, a total of 337 minutes with an average of 28 minutes per video. From these videos, we identified 1067 gestures in total and an average of 89 gestures per video. In addition to identifying gestures, we required an understanding of how gestures indicate the cognitive status of referenced objects. Because the cognitive status an object has in the mind of an interactant cannot directly be observed, we needed a procedure whereby other humans could provide their intuitions as to the cognitive statuses objects might hold throughout the tasks. Therefore, the goal of our testbed was to create a platform for online experiments for identifying and labeling the cognitive status of objects during a collaborative task. In the following sections, we introduce a key set of challenges faced in developing such a testbed.

3.1 Experimental Design

In our testbed, participants are asked to watch a series of videos, complete a task, and respond to two questions following each video. Each video follows a *learner* and *instructor* as they build one of the structures in the four different quadrants and ends in a carefully selected utterance that includes both verbal and non-verbal cues.

In addition to watching videos of the block-moving task, participants are provided a virtual user interface that visualizes the 72 blocks' exact locations in each quadrant [16], shown in Figure 3. As participants watch people in video move blocks, they are asked to follow along with the interface by dragging virtual blocks into a "Pieces Moved" area. Through this task participants will have assumed the cognitive statuses of the learner and instructor.

3.2 Design Considerations

In order to develop the testbed that would enable this experimental design, we identified three key challenges during the design process.

3.2.1 Challenge 1 - How to prompt for an accurate Cognitive status?: While the Givenness Hierarchy describes how to model cognitive status and how to code cognitive status [13, 14, 17], actually applying these rules to a conversation when accounting for both verbal utterances and non-verbal gestures is a difficult task, as there has been little research on how gesture should be considered during manual or automatic coding. As such, there was no well-established way to prompt a person for the cognitive status of entities in their mind. Additionally, the nature of a virtual testbed means that the participant does not directly participate in the block-building experiment, and does not necessarily share the cognitive status of those participating in the previous experiment [15]. While we developed the testbed to mimic this experience as closely as possible (As described in Sections 3.3.2-3.3.3), this furthered the importance of the prompting questions. However, constructing questions to prompt participants for an entity's cognitive status is also a difficult task, and our questions required piloting to ensure they would elicit accurate responses.

3.2.2 Challenge 2 - Virtual Test-bed UI challenges: To carry out a virtual experiment that would allow us to accurately model Cognitive Status in the context of a conversation, we needed the test-bed to display videos to participants. The video medium allowed participants to watch the learner's and instructor's gestures, and therefore to account for these gestures in their assessment of each block's cognitive status. However, the videos provided by [15] introduced challenges as it was unclear in many situations which object was being referred to due to issues of color, the size of blocks, and the people in the prior experiment blocking the view of certain objects. Addressing these challenges required substantial forethought regarding the presentation of our experimental testbed, and required image processing of our videos to maximize clarity.

3.2.3 Challenge 3 - Mitigating cognitive load: In our preliminary testing of the virtual test-bed we quickly realized that participants could be cognitively overloaded while completing the experiment, potentially interfering with their ability to properly complete the experiment. This was a result of the high number of tasks our testbed requires of a participant: they must follow along with an unfamiliar video, watching and listening to the actions of the Instructor and Learner, while completing their own task of moving blocks around. The shape and color of the blocks were also difficult to make out from the desaturated security camera videos we used, further increasing difficulty. Addressing these challenges required additional thought as to the organization and presentation of our testbed, as well as further image processing.

3.3 Testbed Design

We will now describe how considering the challenges listed above informed the design of our experimental testbed.

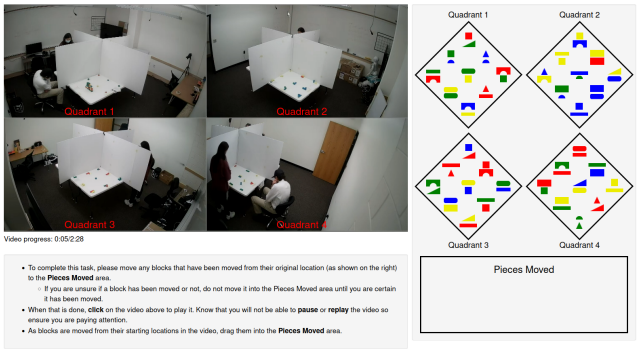

Figure 2: One step of the experiment; users are asked to move blocks to the location they see in the image

3.3.1 Developing the prompting questions: A key challenge in creating our test-bed was designing a way to accurately and intuitively prompt participants about the cognitive status of different blocks. Our goal for the prompting questions was to obtain the objects that were "Activated" and the 0-3 objects that may be "In-focus" given an utterance involving non-verbal gestures from [15].

We quickly realized that broad prompts were ineffective. For example, our first attempt at developing these questions was "Please

write down a description for the 7 blocks you think the learner is the most focused on." Then, the participant would be prompted with "Please click on the 7 blocks you just described in the order you described them." Our hope was that this would cause participants to write down objects starting with the "In-focus" objects, followed by all "Activated" objects. However, it became apparent that this style of cognitive status prompt was both unintuitive and inconsistent. Answers did not always match our manually coded cognitive status and did not follow the guidelines discussed in [13, 14, 17].

To improve our cognitive status prompting method we followed an iterative process using GH guidelines [13, 14] and came up with questions that led participants step-by-step through the cognitive status sorting and labeling process. Instead of asking participants a few vague questions, we developed a set of specific questions that were smaller in scope. For example, we chose to explicitly ask participants "If the instructor were to say "and move 'it' over here", please click on any objects (pick 0-3 objects) which the instructor could be referring to." In this way, we mitigated the challenge of designing intuitive cognitive status prompts by creating this modular series of precise questions that directly inform which objects hold certain cognitive statuses.

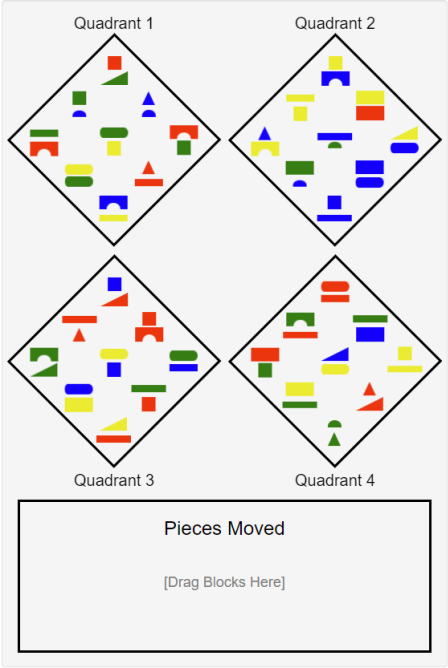

Figure 3: A section of the test-bed which allows participants to interact with the blocks seen in the video

3.3.2 Introduction of a virtual block-building experience: Our initial design had participants virtually build the structures seen in the videos, however, as discussed in 3.3.3, we found that this placed too much cognitive load on the participant. So, we simplified this task by including a "Pieces Moved" area that participants are required to move blocks into as they are moved in the video.

3.3.3 Improving the testbed to decrease cognitive load: We found we could decrease the cognitive load imposed by the testbed by modifying both tasks that participants had to undertake. To decrease the cognitive load of the block-moving task, we simplified the process so that all a participant had to do was move blocks into a "Pieces Moved" area to keep track of which blocks had been moved. This way, recalling these blocks would be easier when they are prompted to select blocks at the end of the utterance. We also aligned the position of the quadrants in the testbed's user interface to match the positions of the quadrants in the video. Then, to reduce the cognitive load of the video-watching task, we cropped the videos to include only information useful to the task, and increased the saturation of the videos to boost the colors of the blocks.

4 CONCLUSION

Multimodal communication, including both speech and gesture, is a key component of robots' ability to competently complete shared tasks with humans. In this paper, we present a novel method for collecting data about the cognitive status of entities during a task-based conversation and the challenges that came with developing such a testbed. We describe the development process for our testbed and how we overcame the challenges we encountered. This testbed will allow us to conduct future research on multimodal referring expression generation in human-robot collaboration.

In our future work, we hope to pilot our testbed online with a large number of participants. The data from those experiments can then be used to develop models of how robots should interpret, represent, and generate meaningful gestures while working with humans in shared environments. Our work can thus support the HRI community in working towards developing more natural referring expression generation models and algorithms in open-world task scenarios.

ACKNOWLEDGEMENTS

This work has been supported in part by the Office of Naval Research grant N00014-21-1-2418.

REFERENCES

[1] Henny Admoni, Thomas Weng, Bradley Hayes, and Brian Scassellati. 2016. Robot nonverbal behavior improves task performance in difficult collaborations. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (2016-03). 51–58. https://doi.org/10.1109/HRI.2016.7451733 ISSN: 2167-2148.

[2] Jens Allwood, Loredana Cerrato, Kristiina Jokinen, Costanza Navarretta, and Patrizia Paggio. 2007. The MUMIN coding scheme for the annotation of feedback, turn management and sequencing phenomena. *Language Resources and Evaluation* 41 (2007), 273–287.

[3] Cynthia Breazeal, Cory D Kidd, Andrea Lockerd Thomaz, Guy Hoffman, and Matt Berlin. 2005. Effects of nonverbal communication on efficiency and robustness in human-robot teamwork. In *2005 IEEE/RSJ international conference on intelligent robots and systems*. IEEE, 708–713.

[4] Rehj Cantrell, Paul Schermerhorn, and Matthias Scheutz. 2011. Learning actions from human-robot dialogues. In *2011 RO-MAN*. IEEE, 125–130.

[5] Nele Dael, Marcello Mortillaro, and Klaus R Scherer. 2012. The body action and posture coding system (BAP): Development and reliability. *Journal of Nonverbal Behavior* 36 (2012), 97–121.

[6] Jan de Wit, Paul Vogt, and Emiel Krahmer. 2022. The Design and Observed Effects of Robot-Performed Manual Gestures: A Systematic Review. *ACM Transactions on Human-Robot Interaction* (2022).

[7] Nick J Enfield, Sotaro Kita, and Jan Peter De Ruiter. 2007. Primary and secondary pragmatic functions of pointing gestures. *Journal of Pragmatics* 39, 10 (2007), 1722–1741.

[8] Jodi Forlizzi. 2007. How Robotic Products Become Social Products: An Ethnographic Study of Cleaning in the Home. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction* (Arlington, Virginia, USA) *(HRI '07)*. Association for Computing Machinery, New York, NY, USA, 129–136. https://doi.org/10.1145/1228716.1228734

[9] Brian Gleeson, Karon MacLean, Amir Haddadi, Elizabeth Croft, and Javier Alcazar. 2013. Gestures for industry intuitive human-robot communication from human observation. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 349–356.

[10] Susan Goldin-Meadow. 1999. The role of gesture in communication and thinking. *Trends in cognitive sciences* 3, 11 (1999), 419–429.

[11] Horst Michael Gross, Steffen Mueller, Christof Schroeter, Michael Volkhardt, Andrea Scheidig, Klaus Debes, Katja Richter, and Nicola Doering. 2015. Robot companion for domestic health assistance: Implementation, test and case study under everyday conditions in private apartments. *IEEE Int'l Conf. on Intelligent Robots and Systems* 2015-December (2015), 5992–5999. https://doi.org/10.1109/IROS.2015.7354230

[12] Stephanie Gross, Brigitte Krenn, and Matthias Scheutz. 2017. The reliability of non-verbal cues for situated reference resolution and their interplay with language: implications for human robot interaction. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. 189–196.

[13] Jeanette K. Gundel and Nancy Hedberg. 2015. 2. Reference and Cognitive Status: Scalar Inference and Typology. In *Information Structuring of Spoken Language from a Cross-linguistic Perspective*, M. M. Jocelyne Fernandez-Vest and Robert D. Van Valin (Eds.). De Gruyter, 33–54. https://doi.org/10.1515/9783110368758-003

[14] Jeanette K. Gundel, Nancy Hedberg, and Ron Zacharski. 1993. Cognitive Status and the Form of Referring Expressions in Discourse. *Language* 69, 2 (1993), 274–307. https://doi.org/10.2307/416535 Publisher: Linguistic Society of America.

[15] Zhao Han, Polina Rygina, and Tom Williams. 2022. Evaluating Referring Form Selection Models in Partially-Known Environments. In *The 15th ACL International Conference on Natural Language Generation (INLG)* (2022). https://zhaohanphd.com/publications/inlg22-evaluating-referring-form-selection-models-in-partially-known-environments/

[16] Zhao Han and Tom Williams. 2022. A Task Design for Studying Referring Behaviors for Linguistic HRI. In *2022 ACM/IEEE International Conference on Human-Robot Interaction (HRI), Late-Breaking Report (LBR)* (2022). https://zhaohanphd.com/publications/hri22lbr-a-task-design-for-studying-referring-behaviors-for-linguistic-hri/

[17] Nancy Hedberg. 2014. Applying the Givenness Hierarchy Framework: Methodological Issues. *Proceedings of Cross-Linguistic Perspectives not eh Information Structure of Austronesian Languages, Tokyo University of Foreign Studies* (Jan. 2014).

[18] Mark Higger, Polina Rygina, Logan Daigler, Lara Ferreira Bezerra, Zhao Han, and Tom Williams. 2023. Toward Open-World Human-Robot Interaction: What Types of Gestures Are Used in Task-Based Open-World Referential Communication?. In *The 27th Workshop on the Semantics and Pragmatics of Dialogue (SemDial)*.

[19] Khari Johnson. 2022. Hospital Robots Are Helping Combat a Wave of Nurse Burnout.

[20] Aelee Kim, Jooyun Han, Younbo Jung, and Kwanmin Lee. 2013. The effects of familiarity and robot gesture on user acceptance of information. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 159–160.

[21] Sotaro Kita. 2003. *Pointing: Where language, culture, and cognition meet*. Psychology Press, Chapter Pointing: A Foundational Building Block of Human Communication, 1 – 9.

[22] Emiel Krahmer and Kees van Deemter. 2012. Computational Generation of Referring Expressions: A Survey. *Computational Linguistics* 38, 1 (2012), 173–218. https://doi.org/10.1162/COLI_a_00088

[23] Nikolaos Mavridis. 2015. A review of verbal and non-verbal human–robot interactive communication. *Robotics and Autonomous Systems* 63 (2015), 22–35.

[24] David McNeill and Elena Levy. 1982. Conceptual representations in language activity and gesture. *Speech, place, and action* (1982), 271–295.

[25] Poulomi Pal, Grace Clark, and Tom Williams. 2021. Givenness Hierarchy Theoretic Referential Choice in Situated Contexts. In *Proceedings of the Annual Meeting of the Cognitive Science Society*.

[26] Poulomi Pal, Lixiao Zhu, Andrea Golden-Lasher, Akshay Swaminathan, and Tom Williams. 2020. Givenness Hierarchy Theoretic Cognitive Status Filtering. arXiv:2005.11267 [cs] http://arxiv.org/abs/2005.11267

[27] Maha Salem, Stefan Kopp, Ipke Wachsmuth, Katharina Rohlfing, and Frank Joublin. 2012. Generation and evaluation of communicative robot gesture. *International Journal of Social Robotics* 4 (2012), 201–217.

[28] Allison Sauppé and Bilge Mutlu. 2014. Robot deictics: How gesture and context shape referential communication. In *2014 9th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 342–349.

[29] Andrew Schoen, Nathan White, Curt Henrichs, Amanda Siebert-Evenstone, David Shaffer, and Bilge Mutlu. 2022. CoFrame: A System for Training Novice Cobot Programmers. In *Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction* (Sapporo, Hokkaido, Japan) *(HRI '22)*. IEEE Press, 185–194.

[30] Adam Stogsdill, Grace Clark, Aly Ranucci, Thao Phung, and Tom Williams. 2021. Is it pointless? modeling and evaluation of category transitions of spatial gestures. In *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*. 392–396.

[31] Stefanie Tellex, Nakul Gopalan, Hadas Kress-Gazit, and Cynthia Matuszek. 2020. Robots That Use Language. *Annual Review of Control, Robotics, and Autonomous Systems* 3, 1 (2020), 25–55. https://doi.org/10.1146/annurev-control-101119-071628 _eprint: https://doi.org/10.1146/annurev-control-101119-071628.

[32] Stefan Waldherr, Roseli Romero, and Sebastian Thrun. 2000. A gesture based interface for human-robot interaction. *Autonomous Robots* 9 (2000), 151–173.

[33] T Williams and M Scheutz. 2019. A givenness hierarchy theoretic approach. *The Oxford handbook of reference* (2019), 457.