A Task Design for Studying Referring Behaviors for Linguistic HRI

Zhao Han
Department of Computer Science
Colorado School of Mines
Golden, CO 80401, USA
Email: zhaohan@mines.edu

Tom Williams
Department of Computer Science
Colorado School of Mines
Golden, CO 80401, USA
Email: twilliams@mines.edu

***Abstract*—In many domains, robots must be able to communicate to humans through natural language. One of the core capabilities needed for task-based natural language communication is the ability to refer to objects, people, and locations. Existing work on robot referring expression generation has focused nearly exclusively on generation of definite descriptions to visible objects. But humans use many other linguistic forms to refer (e.g., pronouns) and commonly refer to objects that cannot be seen at time of reference. Critically, existing corpora used for modeling robot referring expression generation are insufficient for modeling this wider array of referring phenomena. To address this research gap, we present a novel interaction task in which an instructor teaches a learner in a series of construction tasks that require repeated reference to a mixture of present and non-present objects. We further explain how this task could be used in principled data collection efforts.**

***Index Terms*—linguistic HRI, data collection, referring form selection, dyadic interactions, human-robot interaction**

I. Introduction

To efficiently communicate with humans, especially during collaboration scenarios, language-capable robots must be capable of *referring* to objects. Referring has been referred to as the "Fruit Fly of Language" [1] due to the attention it has attracted in the Psycholinguistics and Natural Language Generation communities [2], [3], [4]. Similarly, significant work has been performed within the HRI community on *Referring Expression Generation* (REG), in which a speaker selects the properties to use to refer to an object, location, or person. However, across all these communities, the focus on REG has led to very little research (cp. [5]) on related and no-less-important aspects of referring, such as *Referring Form Selection*, in which the speaker makes the more fundamental decision of whether to use a definite description at all, or whether to instead use a pronominal or deictic expression such as "it" or "that". This divergence has also extended to the types of tasks used to collect data on natural language reference.

Previous research that has collected data with which to model referring language generation in HRI has typically been oriented around tabletop scenarios in which a set of candidate referents can all be seen at once throughout an environment [6], [7], [8], [9] (cp. [10]; and compare also to existing work on *open-world* reference resolution in HRI [11], [12]). However, these scenarios do not naturally promote the wider variety of referring forms observed in human language. Because all objects are visible and relatively equally centered in such scenarios, they all have similar degrees of *Cognitive Status* [13], and thus encourage similar forms of reference.

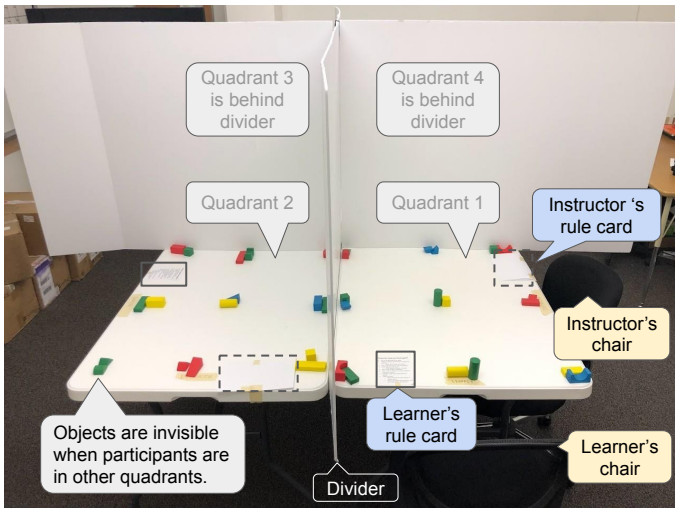

Fig. 1. Two of four quadrants of the task environment. Two rule cards are placed on each table, with Learner's on the left hand side and Instructor's (dashed box) on the right hand side (solid box). Objects are intentionally placed at the intersections of a 3×3 grid to encourage use of different referring forms whose use varies according to distance. Instructor teaches Learner to construct buildings whose constituents blocks are distributed across the visible and non-visible quadrants.

In this work, we present a novel instruction-oriented task design intended to facilitate a wider variety of referring forms. The dyadic interaction task (depicted in Fig. 1) encourages this wider variety of forms through careful manipulation of target referent visibility (thus leading to course-grained variance in cognitive status) and by requiring repeated reference to task referents (thus leading to fine-grained variance in cognitive status). Moreover, these same manipulations should, we believe, also facilitate the use of subtly different types of spatial gestures (cf. [14]). As such, our intention is to use this novel task to enable the collection of a new corpus of data in

This work has been supported in part by the Office of Naval Research under N00014-21-1-2418. We thank Ryan Blake Jackson and Terran Mott for their feedback on the experiment design.

which people use this wider range of referring forms (it, this, that, this-N, that-N, the-N, a-N) to refer to objects that are both physically present and not physically present, using both speech and gesture.

In the remainder of this paper, we will first provide additional motivation for our task interaction design goals. We will then present the task design and carefully explain the rationale behind its constituent elements. Finally, we recommend a procedure for collecting data using this task design.

II. Related Work

In this section, we consider the different tasks that HRI researchers have previously used to collect data for the purposes of modeling referring language use.

Many previous data collection efforts in this space have used blocks-based tabletop task designs. In one such human-robot tabletop scenario, Hsiao et al. [7] studied responsiveness of robots [7] to tabletop block-picking instructions. Matuszek et al. [8] collected a finger pointing data set for modeling object reference. And, similarly, Scalise et al. [9] collected a text corpus for modeling spatial relationship based chains of object references.

Other researchers have varied tabletop object configurations. Li et al. [15] considered a simple tabletop scenario with cluttered cubic blocks and asked participants to pick a single block; a paradigm also used by Weerakoon et al. [16] within a VR environment. Bisk et al. [17] use a similar paradigm in which blocks are stacked, in order to model more complex spatial relationships such as "mirroring" and "balancing". Dan et al. [18] also investigated 3D blocks world, but with a focus on manipulation with reference frames (e.g., absolute or relative).

These tasks have encouraged only some types of referring forms, as they do not, for example, tend to necessitate repeated reference (which would encourage referring forms like "it", "this", and "that", which are used when higher tiers of cognitive status can be assumed) or reference to objects not currently present (cf. our data coverage goal in Section III).

There are also Computer Vision datasets collected in visually similar contexts (e.g., [19], [20]); however, these datasets typically use static images rather than genuinely interactive task contexts, and focus primarily on referring expression accuracy rather than referring form selection [21].

All of the aforementioned tasks involve references to objects in contexts where all candidate referents are always visible. In contrast, we argue that a task design for collecting a wider array of referring forms must necessarily encourage references to both visible and non-visible.

III. Task Design Goals

Our task design goal is to facilitate the collection of a corpus of data in which a wide variety of linguistic referring forms and referring gestures arise. For the reasons previously discussed, we thus desire a task context in which the visibility of task-relevant objects, and the time since last reference to those objects, are carefully varied throughout the task.

We formulate eight hypotheses as to task design elements that should facilitate these design goals:

1) People will use *it* as long as there are immediately repeated references to the same object within a room.
2) People will use *this* and *that* as long as there are nearly-repeated references to objects at varying distances within a room.
3) People will use *this N*, *the N*, and *that N* as long as there are references to ambiguous objects at varying distances within a room
4) People will use *that N* and *the N* to refer to objects seen in previous rooms
5) People will use *a N* and *this N* to refer to objects not yet seen
6) People will use *deictic gesture* when objects are nearby, especially on their first reference
7) People will use *abstract gesture* when objects were in previous rooms
8) People will use *no gesture* when objects have not yet been seen or are repeatedly discussed

In the next section, we present our task design, with its design guided by these assumed hypotheses.

IV. Task Design

Our task is designed around a series of collaborative *tower construction* tasks (cf. [22]). These tasks are performed within a four-quadrant tabletop environment (Figure 1) where, in each quadrant, an instructor participant (Instructor) teaches a learner participant (Learner) to construct a building using wooden blocks. Across these four tasks, four buildings (Figure 2) are constructed from $18 \times 4 = 72$ blocks [23] of different colors and shapes, which are initially distributed across the four quadrants. The four buildings (inspired by the product photo of the building blocks from another brand [24]), are shown in Figure 2: a horse barn complex, a townhouse, a skyscraper, and a museum of math.

In each of the four construction tasks, half of the blocks are present in the current quadrant, while the other half are distributed to the other three quadrants, requiring either intentional reference towards objects acknowledged to not yet have been observed, and/or reference towards objects remembered as having been previously seen but which are no longer visible. As participants proceed throughout the four quadrants and perform the four tasks, the balance between these types of references necessarily change as more blocks become seen.

V. Materials

A. Quadrants

This task environment is constructed by adjoining two tables and erecting barriers from four pieces of foam board to create a partially-observable environment in which only one quadrant of objects can be seen at a time. The foams boards are used to make the barriers longer than the table, preventing participants from looking into other quadrants while they are seated. This

Fig. 2. Four buildings to be constructed, with 18 blocks for each. To help participants identify individual blocks, two angles were provided. For any building, half (9) of its blocks are in the current quadrant while the other half are distributed to the remaining three quadrants, allowing us to collect references to non-present objects.

is intended to encourage participants to use longer expressions, to facilitate collection of more data.

B. Blocks

A variety of block shapes are used, including triangles, cubes, cuboids, cylinders, arches, and half-circles. These blocks provide variety without introducing unnecessary complexity to object descriptions.

All blocks are randomly placed at the vertices of a 3×3 grid within each quadrant. This placement strategy leads to varying physical distance between blocks, encouraging different referring forms whose use typically varies by distance (i.e., "this" vs "that") [25]. This design is thus in accordance with hypotheses 1, 2, 3, and 6 as listed in Section III.

C. Buildings

In accordance with the remaining hypotheses (4, 5, 7, and 8), we further constrain block placement as follows: Half of the blocks needed for a given building are distributed to the quadrant in which that building is to be constructed (randomly distributed throughout placement locations in that quadrant), and the other half of the blocks need to be evenly distributed in the other three quadrants.

To meet this constraint, we made the following design decisions. Each building has an even number of 18 blocks. Nine of them are placed in the quadrant where the building is constructed, and the other nine are evenly distributed across the other three quadrants.

VI. Procedure

We recommend the following procedure to be used to encourage successful data collection within this task context. A pair of participants (Instructor and Learner) collaborate to construct the buildings in person. The experiment is recorded by four security cameras at the corners of the room for future transcription and annotation. The angles of the cameras are adjusted to face the middle of where the pair of participants sit, making sure to cover participants for gesture identification. To collect speaker utterances, a microphone is hung above each quadrant. This camera and microphone setup is inspired by the setup used by the STARS Laboratory[1].

[1]https://www.stars.msstate.edu/

Upon arrival, each participant should take one of the two seats as will and read an informed consent separately printed for either role, mainly stating the purpose and the rules of the study, as well as an audio release form. The seating and its resulting role is not preassigned to avoid implicit bias for who is better at instructing or constructing. After signing the forms, they are directed to the first quadrant.

At two table sides orthogonal to each other, the pair sit and *rule cards* are placed at table corners as reminders for participants to review while experimenters are preparing for video recording. The Learner rule card is visible and placed facing up. With two additional building photos at different angles (Figure 2), the Instructor rule card is initially flipped down and needed to be flipped orthogonally against the thick table edge, so the building photos are not visible to Learner, encouraging more speech and gestures from Instructor.

To solicit more data from Instructors, we recommend:

1) Instructors should not show cards to Learners (to encourage more speech);
2) Instructors should not touch any blocks (to encourage more references);
3) Instructors should remain seated but can ask Learners to find blocks in other quadrants (to encourage references to non-visible objects); and
4) Instructors should look for blocks in the current quadrant before asking Learners to visit other quadrants to seek out blocks (to encourage references to known objects).

For Learners, we similarly make three recommendations to encourage more Instructor speech and gestures:

1) Learners should not ask Instructors to see the building image;
2) Learners should not speak to Instructors unless absolutely necessary to proceed with the task; and
3) Learners should not look at or enter other quadrants unless asked by Instructor.

These rules are enforced by experimenters in another room, who watches the video stream and returns to reminding participants of rules if they are not followed.

Finally, when a building is constructed, its blocks must match in color, shape, and position, and Instructor asks the experimenter to check. If matched, both Instructor and Learner are asked to move to the next quadrant clockwise to construct another building, until all four buildings are built.

VII. Conclusion

In this paper, we contributed a building construction task designed to collect a wider array of referring forms than in previous tabletop reference tasks used in the HRI community. We discussed task design considerations, including how blocks are distributed, how buildings are chosen, and the procedure to maximize data abundance. In future work we intend to collect a corpus of referring forms using this task design, in order to enable more effective language-capable robots.

REFERENCES

[1] K. Van Deemter, *Computational models of referring: a study in cognitive science*. MIT Press, 2016.

[2] E. Reiter and R. Dale, "Building applied natural language generation systems," *Natural Language Engineering*, vol. 3, no. 1, pp. 57–87, 1997.

[3] T. Winograd, "Understanding natural language," *Cognitive psychology*, vol. 3, no. 1, pp. 1–191, 1972.

[4] E. Krahmer and K. Van Deemter, "Computational generation of referring expressions: A survey," *Computational Linguistics*, vol. 38, no. 1, pp. 173–218, 2012.

[5] P. Pal, G. Clark, and T. Williams, "Givenness hierarchy theoretic referential choice in situated contexts," in *Proceedings of the Annual Meeting of the Cognitive Science Society*, 2021.

[6] D. Roy, "Semiotic schemas: A framework for grounding language in action and perception," *Artificial Intelligence*, vol. 167, no. 1-2, pp. 170–205, 2005.

[7] K.-y. Hsiao, S. Vosoughi, S. Tellex, R. Kubat, and D. Roy, "Object schemas for responsive robotic language use," in *Proceedings of the 3rd ACM/IEEE international conference on Human robot interaction*, 2008, pp. 233–240.

[8] C. Matuszek, L. Bo, L. Zettlemoyer, and D. Fox, "Learning from unscripted deictic gesture and language for human-robot interactions," in *Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.

[9] R. Scalise, S. Li, H. Admoni, S. Rosenthal, and S. S. Srinivasa, "Natural language instructions for human–robot collaborative manipulation," *The International Journal of Robotics Research*, vol. 37, no. 6, pp. 558–565, 2018.

[10] K. Eberhard, H. Nicholson, S. Kübler, S. Gundersen, and M. Scheutz, "The indiana "cooperative remote search task"(crest) corpus," in *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, 2010.

[11] T. Williams and M. Scheutz, "Power: A domain-independent algorithm for probabilistic, open-world entity resolution," in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2015, pp. 1230–1235.

[12] ——, "A framework for resolving open-world referential expressions in distributed heterogeneous knowledge bases," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 30, no. 1, 2016.

[13] J. K. Gundel, N. Hedberg, and R. Zacharski, "Cognitive status and the form of referring expressions in discourse," *Language*, pp. 274–307, 1993.

[14] A. Stogsdill, G. Clark, A. Ranucci, T. Phung, and T. Williams, "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*, 2021, pp. 392–396.

[15] S. Li, R. Scalise, H. Admoni, S. Rosenthal, and S. S. Srinivasa, "Spatial references and perspective in natural language instructions for collaborative manipulation," in *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2016, pp. 44–51.

[16] D. Weerakoon, V. Subbaraju, N. Karumpulli, T. Tran, Q. Xu, U.-X. Tan, J. H. Lim, and A. Misra, "Gesture enhanced comprehension of ambiguous human-to-robot instructions," in *Proceedings of the 2020 International Conference on Multimodal Interaction*, 2020, pp. 251–259.

[17] Y. Bisk, K. J. Shih, Y. Choi, and D. Marcu, "Learning interpretable spatial operations in a rich 3d blocks world," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[18] S. Dan, P. Kordjamshidi, J. Bonn, A. Bhatia, J. Cai, M. Palmer, and D. Roth, "From spatial relations to spatial configurations," *arXiv preprint arXiv:2007.09557*, 2020.

[19] J. Mao, J. Huang, A. Toshev, O. Camburu, A. L. Yuille, and K. Murphy, "Generation and comprehension of unambiguous object descriptions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 11–20.

[20] L. Yu, P. Poirson, S. Yang, A. C. Berg, and T. L. Berg, "Modeling context in referring expressions," in *European Conference on Computer Vision*. Springer, 2016, pp. 69–85.

[21] F. I. Doğan, S. Kalkan, and I. Leite, "Learning to generate unambiguous spatial referring expressions for real-world environments," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 4992–4999.

[22] 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 Transactions on Human-Robot Interaction (THRI)*, vol. 10, no. 1, pp. 1–23, 2020.

[23] "100 piece wood blocks set," https://www.melissaanddoug.com/100-piece-wood-blocks-set/481.html, accessed: 2021-12-1.

[24] "Soft & quiet building blocks," https://www.lakeshorelearning.com/products/ca/p/LC1457/, accessed: 2021-12-1.

[25] R. M. Dixon, "Demonstratives: A cross-linguistic typology," *Studies in Language. International Journal sponsored by the Foundation "Foundations of Language"*, vol. 27, no. 1, pp. 61–112, 2003.