Evaluating Cognitive Status-Informed Referring Form Selection for Human-Robot Interactions

Zhao Han (zhaohan@mines.edu)
Tom Williams (twilliams@mines.edu)
MIRRORLab, Department of Computer Science, Colorado School of Mines
Golden, CO 80401 USA

Abstract

Robots must be able to communicate naturally and efficiently, e.g., using concise referring forms like *it*, *that*, and *the* ⟨*N'*⟩. Recently researchers have started working on Referring Form Selection (RFS) machine learning algorithms but only evaluating them offline using traditional metrics like accuracy. In this work, we investigated how a cognitive status-informed RFS computational model might fare in actual human-robot interactions in a human-subjects study (N=36). Results showed improvements over a random baseline in task performance, naturalness, understandability, and mental workload. However, the model was not perceived to outperform a simple, naive, non-random baseline (constant use of indefinite noun phrases). We contribute several key research directions for further development of cognitive status-informed RFS models, the inclusion of multi-modality, and further development of testbeds.

Keywords: cognitive status; referential choice; anaphora generation; natural language generation (NLG), human-robot interaction (HRI)

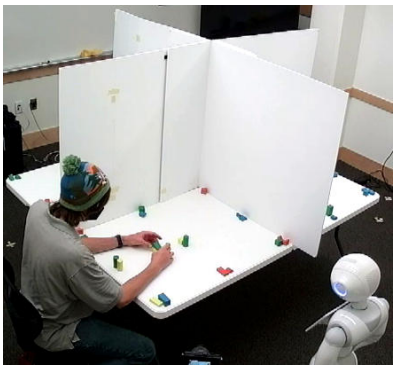

Figure 1: In the human-subjects study, a Pepper robot was teaching a participant to construct a building using instructions containing referring forms. In this work, we evaluated a cognitive status-informed referring form selection model.

Introduction

For language-capable robots to be genuinely helpful, they must be able to communicate naturally and efficiently. When generating descriptions of objects, locations, and people, robots must be able to use not only full definite descriptions (e.g., *the medkit*), but also more concise forms (e.g., or *that medkit*, *this*, or *it*). Humans regularly and strategically use them not only to express their intent more concisely, but also to allow their interlocutors to more quickly and effectively identify their target referents (Gundel et al., 1993).

However, these more concise referring forms are less well studied than the long, deeply nested, descriptive phrases that predominantly take center stage (Van Deemter, 2016; Krahmer & Van Deemter, 2012), particularly in the computational Natural Language Generation (NLG) community and the application domain of Human-Robot Interaction (HRI).

Instead, the focus was on higher-level decisions like what intent the robot should convey (Tellex et al., 2013; Jackson & Williams, 2022; Cakmak & Thomaz, 2012; Williams et al., 2015; Gervits et al., 2021), or lower-level decisions like the properties that should be included in definite descriptions, i.e., Referring Expression Generation (Tellex et al., 2014; Fang et al., 2015; Zender et al., 2009; Williams & Scheutz, 2017; Wallbridge et al., 2019; Doğan et al., 2019; Dogan & Leite, 2020; Sarthou et al., 2021).

Yet Referring Form Selection (RFS) is an important first step that robots must perform before considering including descriptive content (Krahmer & Van Deemter, 2012). As such, some researchers (Same & van Deemter, 2020; Pal et al., 2021; Chen et al., 2021; Han & Williams, 2022a; Spevak et al., 2022) have begun working on this important intermediate task of selecting a referring form among a set of candidate options. For example, Chen et al. (2021) used deep learning end-to-end methods, while Same & van Deemter (2020) and Pal et al. (2021) used feature-based machine learning computationally to model the mechanics of reference choices.

Pal et al. (2021)'s approach is of particular interest to this work due to explainability and its use of the Givenness Hierarchy theory with its constituent notion of Cognitive Status (Gundel et al., 1993). The theory suggests that different referring forms signal different cognitive statuses of objects that speakers assume in the mind of their interlocutors (Rosa & Arnold, 2011). For example, "this" signals that a speaker believes that their target referent is at least *activated* in the listener's mind. This theory has been validated across a wide range of languages with distinct origins (Gundel et al., 2010).

While Givenness Hierarchy theoretic approaches have achieved good performance in machine learning metrics like accuracy (Pal et al., 2021; Han & Williams, 2022a), they are *offline* comparisons to human reference choices. These models have not been evaluated in the context of live human-robot interactions with a physically situated robot. As such, it is unclear whether they will provide observable benefits in real-world collaborative human-robot interaction tasks. Addressing this key research gap is critical, both to understand whether and how these cognitive computational models improve task performance and subjective perceptions of robots.

In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society.*

In this work, we thus conducted a human-subjects study (N=36) evaluating a cognitive status-informed RFS model. Participants followed a robot's instructions to perform a series of building construction tasks. The referring forms used by the robot in its instructions were chosen using either the model, a random baseline, or a simple indefinite noun approach. These models were compared on the basis of objective task performance (i.e., instruction completion time), and participants' subjective perceptions of naturalness, understandability, and mental workload.

Related Work

Linguistic Models of RFS

Although the process of RFS has been understudied in the Computational Linguistics and HRI communities, it has been studied extensively by linguists and psycholinguists.

A number of competing theories (e.g., Gundel et al., 1993; Ariel, 2001) have been proposed to make different predictions. These models fall into two classes: rational and pragmatic selection models (Arnold & Zerkle, 2019). Rational models seek to explain how speakers egocentrically decide whether to use pronouns, e.g., for ease of production (Aylett & Turk, 2004; Jaeger & Levy, 2006; Frank & Goodman, 2012; Mahowald et al., 2013). Pragmatic models seek to explain how speakers allocentrically decide to use pronouns based on their cognitive status within a discourse or conversation, i.e., a mapping between cognitive representations and referring forms (Brown, 1983; Brennan et al., 1987; Ariel, 1991; Grosz et al., 1995; Gundel et al., 1993; Ariel, 2001).

One notable pragmatic model is the *Givenness Hierarchy* theory (Gundel et al., 1993). It suggests referring forms are selected based on a nested set of six tiers of *cognitive statuses*: {in focus $\subseteq$ activated $\subseteq$ familiar $\subseteq$ uniquely identifiable $\subseteq$ referential $\subseteq$ type identifiable}. A referring form choice relies on the cognitive status of the target referent in the mind of the listener. For example, if "it" is used by a speaker to refer to an entity, the listener can infer that the entity's cognitive status must be *in focus*. Similarly, when *that* $\langle N' \rangle$ is used, the listener can infer that the entity is at least *familiar*, but may also be *activated* or even *in focus*.

While these models are promising in predicting whether a speaker chooses to use a definite noun phrase or a more reduced form, neither category of models predicts exactly which referring form to be used. For example, rational models predict more for reduced forms, which are easier to produce, than the frequency that people use in reality.

Moreover, neither type of model attempts to comprehensively model reference production as a whole, but tends to focus on specific referential phenomena, e.g., the reduced forms (Arnold & Zerkle, 2019; Grüning & Kibrik, 2005). Similarly, these models do not attempt to model cognitive mechanisms or psycholinguistic processes (Arnold, 2016). Indeed, Grüning & Kibrik (2005) highlight that many linguists have narrowly focused on specific factors that may impact how referring forms are chosen, like linear (linguistic) distance (Givón, 1983), rhetorical distance (Fox, 1993; Mann et al., 1989), and narrative episodic structure (Tomlin, 1987; Marslen-Wilson et al., 1982). Finally, for human-robot interaction tasks, most research has used textual corpora without any situated features that exist in real-world task scenarios, such as the distance of the objects, which helps differentiate *this* and *that*.

Computational Models of RFS

These linguistic models serve as natural starting points for computational modeling. Yet while they provide critical linguistic insights into the nature of RFS, they offer little direct input into the cognitive processes, mechanisms, or algorithms that govern this process.

Work in the Artificial Intelligence or Computational Linguistics community has similar problems. Most relevant work from these fields (McCoy & Strube, 1999; Callaway & Lester, 2002; Poesio et al., 2004; Kibble & Power, 2004; Kibrik, 2011; Kibrik et al., 2016) falls under *multi-factorial process modeling*. They model the process of referring as a classification problem based on features (Kibrik, 2011; Van Deemter et al., 2012; Gatt et al., 2014). Like the linguistic or psycholinguistic models, these models do not predict specific referring forms but rather pronoun use as a whole.

Cognitive Status-Informed Computational Model Some recent research efforts have attempted to solve these problems, predicting referring forms at a fine-grained level. For example, Pal et al. (2020) proposed a computational model of cognitive status. Pal then leveraged this model along with a set of other features of target referents such as physical distance and temporal distance (recency of mention), as features for a decision tree-based model of Referring Form Selection (Pal et al., 2021), which achieved over 80% accuracy. The training and evaluation of this model were notably conducted on a corpus from a dyadic human-human situated task by Bennett et al. (2017).

Yet more recently, Han & Williams (2022a) found that this task lacks ecological validity for RFS modeling in several critical ways. First, the task domain only contains a small number of candidate referent targets, leading to irregular situations where most of the task-relevant objects are constantly at least *activated*. According to the Givenness Hierarchy, this results in a skewed use of referring forms such as "this". It is also likely the cognitive status of task-relevant objects will remain constant throughout the discourse due to the small number of task-relevant objects. Second, all task-relevant objects in this task are uniquely identifiable, with some even labeled with a unique letter. This means that all objects can be described by proper nouns and simple single-property descriptions, without the need to seriously consider the choice of referring expression. Finally, all objects in this task are visible at all times. This discourages the usage of indefinite nouns such as *a* $\langle N' \rangle$, which are often used when speakers assume that listeners do not already have knowledge of the target referent. These challenges were addressed by Han & Williams

(2022a), who developed a new task domain in which a wide variety of referring forms was collected and modeled while alleviating these problems. We thus used their comprehensive model in this work.

Hypotheses

Because the Givenness Hierarchy-based referential choice algorithm considers the cognitive status of the interlocutor, we expect positive effects on both task performance (H1) and subjective experience (H2–H4).

Hypothesis 1 (H1) – Increased Task Performance: It will take less time to finish tasks instructed using these referring forms.

Hypothesis 2 (H2) – Higher Perceived Naturalness: These referring forms are more natural.

Hypothesis 3 (H3) – Increased Understandability: These referring forms are more understandable.

Hypothesis 4 (H4) – Lower Workload: Following these referring forms requires less workload.

Method

We conducted a human-subject experiment in which participants were instructed by a robot to construct a series of buildings from wooden blocks. The study followed a within-subjects design, and the order was fully counterbalanced.

Apparatus and Materials

Robot Platform We used the SoftBank Pepper robot (Pandey & Gelin, 2018): a two-armed, 1.2m (3.9ft) tall humanoid robot with two speakers at the sides of its head. The voice speed was set at 90% to make its speech clearer.

Quadrants The task environment (Fig. 1) was constructed by adjoining two tables and erecting barriers from four pieces of foam board. This created a partially-observable environment to include uses of references like a $\langle N' \rangle$ to refer to non-present objects.

Blocks Nine distinct block shapes were used (Melissa & Doug, 2019). They include triangles (small and long), cubes, three types of cuboids, cylinders, arches, and half-circles. All blocks were randomly placed on a 3×3 grid within each quadrant. This leads to varying physical distance between blocks, which helps to differentiate referring forms whose use typically varies by distance, i.e., *this* vs. *that* (Dixon, 2003).

Buildings As shown in Fig. 2, participants constructed three buildings. Each building had 18 blocks to ensure candidate referents were not trivially distinguishable. Nine (50%) blocks were evenly distributed to other quadrants to include references introducing new objects.

Instruction Design

All the instructions given by the robot were based on instructions given by real humans. Specifically, one series of instructions was first selected from Han & Williams (2022b)'s publicly available dataset for each building. Next, some utterances were removed from these instruction sequences, such as corrective utterances (e.g., "Okay. I think that should be square.") and confirmative utterances (e.g., "Yeah." and "Perfect."). Finally, all referring expressions in the instructions were identified and replaced with new referring expressions.

Figure 2: The three buildings in the construction tasks. The number and variety of blocks were designed to lead to various cognitive statuses in instructions.

We divided a referring expression into two key parts: a referring form and a (possibly empty) set of propositional semantic content. The referring form is assumed to be one in $\{$*it, this, that, this* $\langle N' \rangle$*, that* $\langle N' \rangle$*, the* $\langle N' \rangle$*, a* $\langle N' \rangle\}$, which are associated with cognitive statuses by the Givenness Hierarchy: *In Focus* $\rightarrow$ *it, Activated* $\rightarrow$ $\{$*this, that, this* $\langle N' \rangle\}$, *Familiar* $\rightarrow$ *that* $\langle N' \rangle$, *Uniquely Identifiable* $\rightarrow$ *the* $\langle N' \rangle$, *Type Identifiable* $\rightarrow$ *a* $\langle N' \rangle$.

For referring expressions using referring forms that included a noun phrase $\langle N' \rangle$, we assumed that the propositional semantic content represented by this noun phrase always had the form $\{size\}$ $\{color\}$ $\{shape\}$, e.g., *"long yellow triangle"*. For the remaining *it, this, that, this* without a noun phrase $\langle N' \rangle$, the propositional semantic content was empty.

Accordingly, our three strategies comprising our three experimental conditions all followed the same strategy for selecting propositional content, but each used a different strategy for selecting referring forms:

1. **Random**: All referring form was randomly chosen from $\{$*it, this, that, this* $\langle N' \rangle$*, that* $\langle N' \rangle$*, the* $\langle N' \rangle$*, a* $\langle N' \rangle\}$.
2. **Indefinite** nouns. All referring forms were in the form of *a* $\langle N' \rangle$. Indefinite nouns are associated with the lowest (*type-identifiable*) tier and are thus always justifiable.
3. **Model**: All referring forms were predicted by Han & Williams (2022a)'s cognitive computational model, described in the end of the Related Work section.

When we started designing this experiment, we originally included a human condition where the robot directly repeated what the original human participant had said. However, we found that without also replicating that original speaker's gestures, the utterances did not make sense. We thus decided to leave out the condition. Modeling the impacts of gesture on cognitive status is a key direction for future work.

Procedure

Participants first completed an informed consent form and a demographics questionnaire, and then entered the first quadrant and sat on the left of that quadrant, as shown in Fig. 1.

The Pepper robot was positioned on the right side, 50*cm* away from the table edge, so it looked at the table with its vertical field of view covering all blocks. We disabled Pepper's autonomy mode so no built-in autonomous behavior could confound participants' perception during the experiment.

Participants followed Pepper's instructions to construct the three buildings sequentially and were asked to say "Ok" after they were ready for another instruction. To avoid wrong speech recognition for "Ok", the experimenter manually triggered the robot's next utterance right after hearing the word from earphones in another room. After a building was constructed, participants completed a survey to measure subjective experience and workload. Participants then entered the next quadrant.

The experiment was approved by the human subjects research committee at Colorado School of Mines in the US.

Measures

To test our hypotheses, we measured instruction completion time, workload, naturalness, and understandability.

To measure *instruction completion time*, we logged the time when the Pepper robot finished giving an instruction and when the robot spoke the next instruction. Intervals of more than a minute were removed after outlier analysis.

To measure *cognitive effort*, we used the widely-accepted NASA Task Load Index (Hart, 2006; NASA, 2019), including both its load survey and weighting survey components.

To measure *naturalness* and *understandability* of the referring forms, participants completed (after each within-subjects experimental block) 7-point Likert Items in which they were asked to indicate how natural and understandable Pepper's verbal references had been during the preceding block.

Participants

Thirty-six participants contributed valid data, while 38 participants were recruited from a university community in Colorado, USA. Two participants' data were excluded because part of their workload responses were not recorded.

The age ranges from 18 to 59 (M=26, SD=10.6). 18 (50%) were male, 16 (44.4%) were female, one (0.03%) was genderfluid, and one (0.03%) was gender nonconforming. They were mostly white (30, 83.3%) while six (16.3%) reported Asian identities. For whether they have experience with robots, 11 (30.6%) disagreed, two (0.06%) were neutral, and 23 (63.9%) agreed. Participants spent 40.41 minutes on average for the experiment and received $10 Amazon gift cards.

Data Analysis

We analyzed our data using the Bayesian statistical framework (Wagenmakers et al., 2018) in JASP 0.16.3. The Bayesian approach has one key benefit: It can quantify evidence both *for* and *against* hypotheses of interest using the *Bayes factor (BF)*, which is a ratio of the likelihood of given data being observed under each of two competing hypotheses, $\mathcal{H}_1$ and $\mathcal{H}_0$. For example, a Bayes Factor of BF_{10}=5 indicates that the data are five times more likely under $\mathcal{H}_1$ than under $\mathcal{H}_0$. To facilitate decision-making, we used the widely-accepted discrete classification scheme proposed by Lee & Wagenmakers (2014). For evidence favoring $\mathcal{H}_1$, a Bayes factor BF_{10} is deemed "anecdotal" when $BF \in (1,3]$, "moderate" when $BF \in (3,10]$, "strong" when $BF \in (10,30]$, "very strong" when $BF \in (30,100]$, and "extreme" when $BF \in (100,\infty]$. For data in favor of $\mathcal{H}_0$, these thresholds are inverted (1, 1/3, 1/10, 1/30, 1/100). In such cases, we also use BF_{01} ($1/BF_{10}$) rather than BF_{10} for easier interpretability.

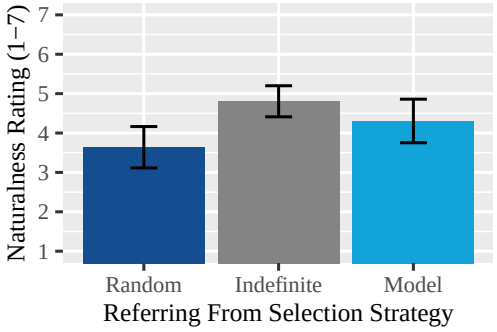

Figure 3: Mean naturalness ratings. Error bars show 95% CI. Results favored differences among all pairwise comparisons.

Results

Instruction Completion Time

As we planned to use Bayesian one-way repeated measures analysis of variance (RM-ANOVA) (Rouder et al., 2012), we assessed the normality of our data by visually inspecting the Q-Q (Quantile-Quantile) plots, a well-accepted practice among Bayesianists (Wagenmakers et al., 2018), revealing violation of normality and linearity. We thus log-transformed the data, which successfully addressed this violation.

We then ran a Bayesian one-way RM-ANOVA. Results showed moderate evidence in favor of $\mathcal{H}_0$ ($BF_{01} = 8.714$), meaning the data are around 8.7 times more likely under models that did not include an effect of condition than under those that did. Participants spent around the same time on each instruction ($M_{Random} = 9.623$, $M_{Indefinite} = 9.494$, $M_{Model} = 9.743$). Thus, **H1** was not supported: Participants did not spend less time on the task in the *Model* condition.

Naturalness

Visual inspection of Q-Q plots upheld the linearity and normality of Naturalness data across conditions. A Bayesian one-way RM-ANOVA showed extreme evidence ($BF_{10} = 37585.667$) favoring an effect of referring form selection strategy. Post-hoc Bayesian t-tests showed evidence favoring differences across all pairwise comparisons, as shown in Fig. 3. Specifically, these tests provide extreme evidence ($BF_{10} = 82216.093$) that utterances were perceived less natural in the *Random* condition (M=3.639, SD=1.552) than in the *Indefinite* condition (M=4.806, SD=1.167), strong evidence ($BF_{10} = 12.851$) suggesting that utterances were perceived less natural in the *Random* condition than in the

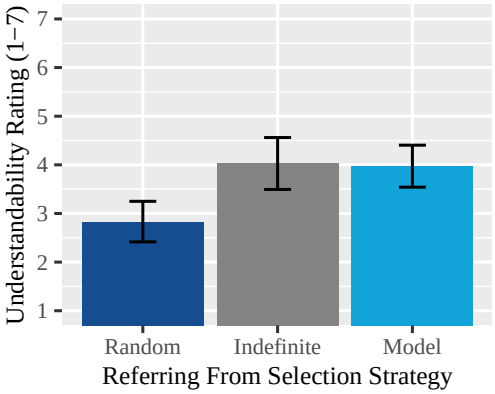

Figure 4: Mean understandability ratings. Error bars show 95% CI. Evidence was found favoring differences between the *Random* condition and the other conditions ($BF_{10} > 100$), but against a difference between the *Indefinite* condition and the *Model* condition ($BF_{01} = 5.464$).

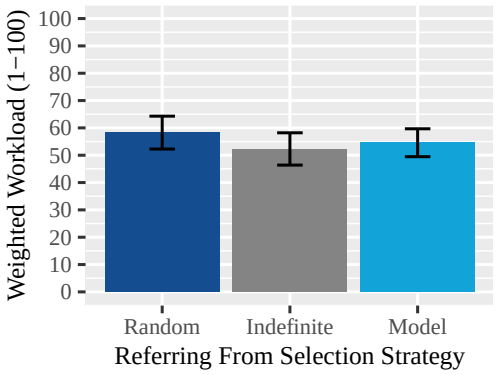

Figure 5: Mean weighted workload. Error bars show 95% CI. Results confirmed that there is no difference between the *Indefinite* condition and the *Model* condition ($BF_{01} = 3.639$).

Model condition (M=4.306, SD=1.636), and moderate evidence ($BF_{10} = 3.592$) that utterances were perceived less natural in the *Model* condition than in the *Indefinite* condition. Thus, **H2** was not supported. Although naturalness was rated higher in the *Model* condition than in the *Random* condition, participants did not perceive these cognitive status-informed referring forms as more natural than indefinite nouns.

Understandability

Visual inspection of Q-Q plots upheld the linearity and normality of Understandability data across conditions. A Bayesian one-way RM-ANOVA showed extreme evidence ($BF_{10} = 37617.862$) in favor of an effect on Understandability, as shown in Fig. 4. Post-hoc Bayesian t-tests provided extreme evidence ($BF_{10} = 170.949$) that utterances were perceived less understandable in the *Random* condition (M=2.833, SD=1.232) than in the *Indefinite* condition (M=4.028, SD=1.576), extreme evidence ($BF_{10} = 3014.702$) that utterances were perceived less understandable in the *Random* condition than in the *Model* condition (M=3.972, SD=1.276), but moderate evidence ($BF_{01} = 5.464$) *against* a difference between the *Indefinite* condition and the *Model* condition.

Thus, **H3** is not supported. In the *Model* condition, the referring forms were not perceived as more understandable.

Workload

According to the NASA Task Load Index manual (NASA, 2019), a weighted score was calculated for each participant. Visual inspection of Q-Q plots upheld the linearity and normality of Workload data across conditions.

A Bayesian one-way RM-ANOVA shows inconclusive anecdotal evidence ($BF_{10} = 1.047$) in favor of effect, as shown in Fig. 5. Pairwise comparisons by Bayesian posthoc t-tests showed moderate evidence ($BF_{01} = 3.639$) against a difference between the *Indefinite* condition (M=52.296, SD=17.465) and the *Model* condition (M=54.565, SD=15.099), and inconclusive evidence for the other two pairs. Specifically, there was probably no difference ($BF_{10} = 2.455$) between the *Random* condition (M=58.278, SD=17.774) and the *Indefinite* condition, but it is possible that utterances in the *Indefinite* condition induced lower levels of perceived cognitive load than the *Random condition*. Similarly, there was probably no difference ($BF_{01} = 1.916$) between the *Random* condition and the *Model* condition, but it is possible that the *Model* condition induced lower perceived cognitive load than the *Random condition*.

Thus, **H4** is not supported. The workload was rated the same in the *Model* condition as in the *Indefinite* condition.

Discussion

Surprisingly, results did not support any of our hypotheses regarding the cognitive status-informed model.

Hypothesis One: Task Performance

The first hypothesis was that it would take participants less time to finish their tasks when instructed with cognitively predicted, humanlike referring forms. Results suggested no difference across all conditions ($BF_{01} = 8.714$): Participants spent an average of approximately 9.5 seconds on each instruction. This might be due to the time it took participants to find and retrieve a referred block when it was in another quadrant. These results suggest the community needs to investigate better metrics to measure when an understanding of a referring form ends or is achieved. Such metrics would be challenging and need not only to be precise but also non-invasive because interruptions could interfere with cognitive status dynamics. *Future work needs to investigate the use of eye-tracking, neurophysiological measures, or through video coding, for detecting reference resolution timing while minimizing subjectiveness.*

Hypothesis Two: Naturalness

The second hypothesis was that humanlike referring forms would be more natural. Surprisingly, using cognitive status-

informed referring forms did not improve the naturalness. While the utterances were ranked more natural in the *Model* condition than in the *Random* condition, neither were rated as highly as in the *Indefinite* condition. While we explicitly asked participants about the naturalness of how robots *referred to* blocks rather than how the robot *spoke in general*, the overall naturalness of the utterances may still play a role. *Future work needs to investigate ways of encouraging participants to only attend to the referring forms to measure naturalness.*

Hypothesis Three: Understandability

Our third hypothesis was that humanlike referring forms would be more understandable. As shown in Fig. 4, both indefinite nouns and cognitive status-informed referring forms were rated equally highly.

This might be due to the limited ambiguity present in the experimental context. While the task context used in this work was originally developed to increase ambiguity by increasing the variety of different combinations of colors and shapes, this, unfortunately, leads to only at most two identical blocks in each building from our investigation (see Fig. 2). As such, an indefinite noun phrase to refer to the intended color and shape of the next block may be sufficient for an interactant to pick out the object, and thus is rated as understandable. *Future work should examine task contexts with more ambiguity, encompassing both reference comprehensiveness and ambiguity, so complex relations need to be described to disambiguate through a full noun phrase.*

Hypothesis Four: Mental Workload

Our fourth hypothesis was that humanlike referring forms would require less mental workload. Results instead showed no workload differences between cognitive status-informed referring form selection and indefinite noun phrases. One reason might be the same confluence of factors that led to the lack of differences in other dependent variables. For example, workload differences may not have been observable when assessed after each building was constructed. *While measuring workload at instruction level is challenging as cognitive status can be interrupted, future work could investigate designing a task with a constrained goal or time limit.*

General Discussion

One trend we noticed across our analyses was that indefinite nouns performed as well or better than cognitive status-informed referring form selection. We do not take this as evidence that robots should always use indefinite noun phrases, just as humans would not be advised to do so.

Rather, although we observed that the predicted referring forms in the *Model* condition were generally of high quality, the model occasionally made memorably poor predictions that may have singlehandedly ruined the naturalness ratings.

Specifically, there were some cases where the robot used "it" in contexts where it was not justified, leading to obvious difficulty for participants. Several participants visibly stared at Pepper for a few seconds, and some explicitly asked the experimenter whether the robot had "glitched".

It is possible that robots should simply avoid some overly restrictive referring forms unless with extremely high confidence. Future research could consider models that avoid using *it* and assess their performance.

Our results could also be due to the attempt to generate referring forms in isolation. Human speakers select referring form, content, word choice, and so forth as part of a single process. By filling new referring forms into the sentence structures selected by humans, it may have introduced incongruencies and irregularities that needlessly impaired the cognitive status-informed approach. Similarly, the model focused on verbal communication without gesturing. The predicted referring forms would need to appropriately trigger nonverbal cues for many reduced forms like *it* to be used successfully. Future work should investigate the evaluation of referring form selection models as part of a complete and multimodal natural language generation pipeline.

Finally, results raise key questions about the goal of humanlikeness in robot language generation. In fact, Han & Williams (2022a) pointed out that a model does not need to perfectly mimic human utterances to be successful, and there might be multiple referring forms that are equally appropriate. Moreover, humans may sometimes use concise referring forms for their own ease of production rather than to provide any benefit for listeners; if so, this suggests that robots can often be more prolix than human speakers. Finally, it is worth noting that many of the ethical and practical problems that plague neural Natural Language Generation models lie in their single-minded focus on fluency and humanlikeness. We would be well served to avoid overreliance on mimicry of humanlikeness as a guiding principle (Williams et al., 2020).

Conclusion

In this work, we conducted a human-subject study to investigate the objective and subjective performance of cognitive status-informed referring form selection models in a live collaborative HRI task. Results showed that these cognitive status-informed models have a long way to go in terms of performance in live human-robot interactions, with recent models outperformed by a naive indefinite noun phrase approach.

Rather than suggesting that robots can just use simple heuristic strategies like constant indefinite noun phrases going forward, we take the results as evidence that more work is needed to improve these cognitive status-informed models, as a nod towards the types of metrics we need to consider and the types of multimodal features needed to improve performance relative to those metrics, and as a reminder of the nuances of language and of the fragility of interactions with our new robotic teammates: even a single overly ambiguous pronoun may be enough to derail the overall interaction.

Supplementary Materials

All experiment materials, code, data, and analysis scripts are available at `https://osf.io/nzwvf/`.

Acknowledgments

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

References

Ariel, M. (1991). The function of accessibility in a theory of grammar. *Journal of pragmatics*, *16*(5), 443–463.

Ariel, M. (2001). Accessibility theory: An overview. *Text representation: Linguistic and psycholinguistic aspects*, *8*, 29–87.

Arnold, J. E. (2016). Explicit and emergent mechanisms of information status. *Topics in cognitive science*, *8*(4), 737–760.

Arnold, J. E., & Zerkle, S. A. (2019). Why do people produce pronouns? Pragmatic selection vs. rational models. *Language, Cognition and Neuroscience*, *34*(9), 1152–1175.

Aylett, M., & Turk, A. (2004). The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and speech*, *47*(1), 31–56.

Bennett, M., Williams, T., Thames, D., & Scheutz, M. (2017). Differences in interaction patterns and perception for teleoperated and autonomous humanoid robots. In *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)* (pp. 6589–6594).

Brennan, S. E., Friedman, M. W., & Pollard, C. (1987). A centering approach to pronouns. In *25th annual meeting of the association for computational linguistics* (pp. 155–162).

Brown, G. (1983). Prosodic structure and the given/new distinction. In *Prosody: Models and measurements* (pp. 67–77). Springer.

Cakmak, M., & Thomaz, A. L. (2012). Designing robot learners that ask good questions. In *2012 7th ACM/IEEE international conference on human-robot interaction* (pp. 17–24).

Callaway, C. B., & Lester, J. (2002). Pronominalization in generated discourse and dialogue. In *Proceedings of the 40th annual meeting of the association for computational linguistics* (pp. 88–95).

Chen, G., Same, F., & van Deemter, K. (2021). What can neural referential form selectors learn? In *Proceedings of the 14th international conference on natural language generation* (pp. 154–166).

Dixon, R. M. (2003). Demonstratives: A cross-linguistic typology. *Studies in Language. International Journal sponsored by the Foundation "Foundations of Language"*, *27*(1), 61–112.

Doğan, F. I., Kalkan, S., & Leite, I. (2019). Learning to generate unambiguous spatial referring expressions for real-world environments. In *2019 IEEE/RSJ international conference on intelligent robots and systems (IROS)* (pp. 4992–4999).

Dogan, F. I., & Leite, I. (2020). Open challenges on generating referring expressions for human-robot interaction. In *The 2nd workshop on nlg for hri at the international conference on natural language generation (inlg).*

Fang, R., Doering, M., & Chai, J. Y. (2015). Embodied collaborative referring expression generation in situated human-robot interaction. In *Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction* (pp. 271–278).

Fox, B. A. (1993). *Discourse structure and anaphora: Written and conversational english* (No. 48). Cambridge University Press.

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084).

Gatt, A., Krahmer, E., Van Deemter, K., & Van Gompel, R. (2014). Models and empirical data for the production of referring expressions. *Lang., Cognition and Neuroscience*, *29*(8), 899–911.

Gervits, F., Briggs, G., Roque, A., Kadomatsu, G. A., Thurston, D., Scheutz, M., & Marge, M. (2021). Decision-theoretic question generation for situated reference resolution: An empirical study and computational model. In *Proceedings of the 2021 international conference on multimodal interaction* (pp. 150–158).

Givón, T. (1983). Topic continuity in discourse: An introduction. *Topic continuity in discourse: A quantitative cross-language study*, *3*, 5–41.

Grosz, B. J., Joshi, A. K., & Weinstein, S. (1995). Centering: A framework for modelling the local coherence of discourse. *Computational Linguistics*.

Grüning, A., & Kibrik, A. A. (2005). Modeling referential choice in discourse: A cognitive calculative approach and a neural network approach. In *Anaphora processing: Linguistic, cognitive and computational modelling* (pp. 163–198). John Benjamins.

Gundel, J. K., Bassene, M., Gordon, B., Humnick, L., & Khalfaoui, A. (2010). Testing predictions of the givenness hierarchy framework: A crosslinguistic investigation. *Journal of Pragmatics*, *42*(7), 1770–1785.

Gundel, J. K., Hedberg, N., & Zacharski, R. (1993). Cognitive status and the form of referring expressions in discourse. *Language*, 274–307.

Han, Z., & Williams, T. (2022a). Evaluating referring form selection models in partially-known environments. In *Proceedings of the 15th international conference on natural language generation.* Association for Computational Linguistics.

Han, Z., & Williams, T. (2022b, Jul). *Hri025 [han2022inlg]: Evaluating referring form selection models in partially-known environments.* OSF. Retrieved from `osf.io/z3ths` doi: 10.17605/OSF.IO/Z3THS

Hart, S. G. (2006). Nasa-task load index (nasa-tlx); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 50, pp. 904–908).

Jackson, R. B., & Williams, T. (2022). Enabling morally sensitive robotic clarification requests. *ACM Transactions on Human-Robot Interaction (THRI)*, *11*(2), 1–18.

Jaeger, T., & Levy, R. (2006). Speakers optimize information density through syntactic reduction. *Advances in neural information processing systems*, *19*.

Kibble, R., & Power, R. (2004). Optimizing referential coherence in text generation. *Comp. Ling.*, *30*(4), 401–416.

Kibrik, A. A. (2011). *Reference in discourse*. Oxford University Press.

Kibrik, A. A., Khudyakova, M. V., Dobrov, G. B., Linnik, A., & Zalmanov, D. A. (2016). Referential choice: Predictability and its limits. *Frontiers in psychology*, *7*, 1429.

Krahmer, E., & Van Deemter, K. (2012). Computational generation of referring expressions: A survey. *Computational Linguistics*, *38*(1), 173–218.

Lee, M. D., & Wagenmakers, E.-J. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge university press.

Mahowald, K., Fedorenko, E., Piantadosi, S. T., & Gibson, E. (2013). Info/information theory: Speakers choose shorter words in predictive contexts. *Cognition*, *126*(2), 313–318.

Mann, W. C., Matthiessen, C. M., & Thompson, S. A. (1989). *Rhetorical structure theory and text analysis* (Tech. Rep.). University of Southern California.

Marslen-Wilson, W., Levy, E., & Tyler, L. K. (1982). Producing interpretable discourse: The establishment and maintenance of reference. *Speech, place, and action*, 339–378.

McCoy, K. F., & Strube, M. (1999). Generating anaphoric expressions: pronoun or definite description? In *The relation of discourse/dialogue structure and reference.*

Melissa & Doug. (2019). *100 piece wood blocks set.* `https://www.melissaanddoug.com/products/100-piece-wood-blocks-set`. (Accessed: 2022-08-29)

NASA. (2019). *NASA TLX paper & pencil version.* `https://humansystems.arc.nasa.gov/groups/tlx/tlxpaperpencil.php`. (Accessed: 2022-05-09)

Pal, P., Clark, G., & Williams, T. (2021). Givenness hierarchy theoretic referential choice in situated contexts. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 43).

Pal, P., Zhu, L., Golden-Lasher, A., Swaminathan, A., & Williams, T. (2020). Givenness hierarchy theoretic cognitive status filtering. In *Proceedings of the annual meeting of the cognitive science society.*

Pandey, A. K., & Gelin, R. (2018). A mass-produced sociable humanoid robot: Pepper: The first machine of its kind. *IEEE Robotics & Automation Magazine*, *25*(3), 40–48.

Poesio, M., Stevenson, R., Eugenio, B. D., & Hitzeman, J. (2004). Centering: A parametric theory and its instantiations. *Computational linguistics*, *30*(3), 309–363.

Rosa, E. C., & Arnold, J. E. (2011). The role of attention in choice of referring expressions. *Proceedings of PRE-Cogsci: Bridging the gap between computational, empirical and theoretical approaches to reference*, *20*.

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default bayes factors for anova designs. *Journal of mathematical psychology*, *56*(5), 356–374.

Same, F., & van Deemter, K. (2020). A linguistic perspective on reference: Choosing a feature set for generating referring expressions in context. In *Proceedings of the 28th international conference on computational linguistics* (pp. 4575–4586).

Sarthou, G., Buisan, G., Clodic, A., & Alami, R. (2021). Extending referring expression generation through shared knowledge about past human-robot collaborative activity. In *2021 IEEE/RSJ international conference on intelligent robots and systems (IROS)* (pp. 1879–1886).

Spevak, K., Han, Z., Williams, T., & Dantam, N. T. (2022). Givenness hierarchy informed optimal document planning for situated human-robot interaction. In *2022 IEEE/RSJ international conference on intelligent robots and systems (IROS).*

Tellex, S., Knepper, R., Li, A., Rus, D., & Roy, N. (2014). Asking for help using inverse semantics.

Tellex, S., Thakerll, P., Deitsl, R., Simeonovl, D., Kollar, T., & Royl, N. (2013). Toward information theoretic human-robot dialog. *Robotics*, 409.

Tomlin, R. S. (1987). Linguistic reflections of cognitive events. *Coherence and grounding in discourse*, *11*, 455–479.

Van Deemter, K. (2016). *Computational models of referring: a study in cognitive science*. MIT Press.

Van Deemter, K., Gatt, A., Van Gompel, R. P., & Krahmer, E. (2012). Toward a computational psycholinguistics of reference production. *Topics in cognitive science*, *4*(2), 166–183.

Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., ... others (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic bulletin & review*, *25*(1), 35–57.

Wallbridge, C. D., Lemaignan, S., Senft, E., & Belpaeme, T. (2019). Generating spatial referring expressions in a social robot: Dynamic vs. non-ambiguous. *Frontiers in Robotics and AI*, *6*, 67.

Williams, T., Briggs, G., Oosterveld, B., & Scheutz, M. (2015). Going beyond literal command-based instructions: Extending robotic natural language interaction capabilities. In *Twenty-ninth AAAI conference on artificial intelligence.*

Williams, T., & Scheutz, M. (2017). Referring expression generation under uncertainty: Algorithm and evaluation framework. In *Proceedings of the 10th international conference on natural language generation* (pp. 75–84).

Williams, T., Zhu, Q., Wen, R., & de Visser, E. J. (2020). The confucian matador: three defenses against the mechanical bull. In *Companion of the 2020 ACM/IEEE international conference on human-robot interaction* (pp. 25–33).

Zender, H., Kruijff, G.-J. M., & Kruijff-Korbayová, I. (2009). Situated resolution and generation of spatial referring expressions for robotic assistants. In *Twenty-first international joint conference on artificial intelligence.*