Exploring the Naturalness of Cognitive Status Informed Referring Form Selection Models

Gabriel Del Castillo* and **Grace Clark*** and **Zhao Han*** and **Tom Williams**
MIRRORLab
Department of Computer Science
Colorado School of Mines
gdelcastillo@mines.edu, geclark@mines.edu, zhaohan@mines.edu, twilliams@mines.edu

Abstract

Language-capable robots must be able to efficiently and naturally communicate about objects in the environment. A key part of communication is *Referring Form Selection* (RFS): the process of selecting a form like *it*, *that*, or *the N* to use when referring to an object. Recent cognitive status-informed computational RFS models have been evaluated in terms of goodness-of-fit to human data. But it is as yet unclear whether these models actually select referring forms that are any more *natural* than baseline alternatives, regardless of goodness-of-fit. Through a human subject study designed to assess this question, we show that even though cognitive status-informed referring selection models achieve good fit to human data, they do not (yet) produce concrete benefits in terms of naturality. On the other hand, our results show that human utterances also had high variability in perceived naturality, demonstrating the challenges of evaluating RFS naturality.

Keywords: Referring form selection (RFS), computational models, naturalness, Givenness Hierarchy, cognitive status

1 Introduction

Referring is a critical part of human communication, especially in situated, task-based interactions. Humans use a variety of referring forms during reference production, including both definite descriptions (e.g., *The red box on the table*) and concise referring forms (e.g., *this box*, *that*, or *it*). While more concise referring forms are less information-rich, they allow speakers to express their intentions more quickly, and allow their listeners in turn to more quickly and effectively infer those intentions (Gundel et al., 1993). The process of choosing what type of referring form to use, known as *Referring Form Selection*, is an important first step in the production of referring language (Krahmer and Van Deemter, 2012).

* The first three authors contributed equally to this work.

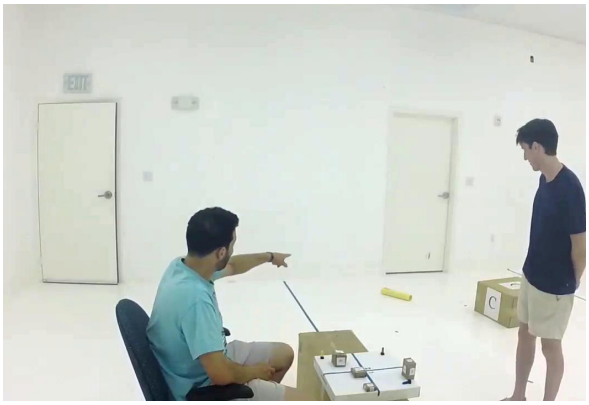

Figure 1: To investigate the naturalness of referring forms, we conducted a study where participants watched videos of human-human instruction tasks. After each video, participants were shown a referring utterance that could have followed as the next line in the dialogue, and were asked to assess the naturalness of that utterance.

Yet despite the wide variety of referring forms observed in human-human interaction, and the critical role of Referring Form Selection in language production, most research on generating referring language has focused solely on definite descriptions (Van Deemter, 2016; Krahmer and Van Deemter, 2012). While generating effective definite descriptions is a critical task, a speaker solely relying on this referring form would be an inefficient, unnatural, and annoying speaker. This discrepancy is critical not just for the psycholinguistics community, who seek to understand the cognitive dynamics of language production, but also for the Artificial Intelligence and Human-Robot Interaction communities, who seek to enable efficient, natural, and humanlike communication in task-based, situated domains (Tellex et al., 2013; Jackson and Williams, 2022; Cakmak and Thomaz, 2012; Williams et al., 2015; Gervits et al., 2021). As such, we argue that more attention to the problem of Referring Form Selection is needed across multiple areas of cognitive science.

Recently, a number of researchers have begun to give the Referring Form Selection problem additional consideration, using a variety of experimental and machine learning research techniques (Same and van Deemter, 2020; Pal et al., 2021; Chen et al., 2021; Han et al., 2022; Spevak et al., 2022). For example, Chen et al. (2021) examined the features learned by deep learning models of Referring Form Selection; Han et al. (2022) addressed the ecological validity of the contexts in which Referring Form Selection is studied; Spevak et al. (2022) studied how document planning of task instructions could be designed for optimal referring form selection; and Pal et al. (2021) and Han et al. (2022) studied how models of cognitive status could be used to enable cognitively informed models of Referring Form Selection.

This last set of work is of particular interest: Pal et al. (2021) and Han et al. (2022) leveraged the well-validated (Gundel et al., 2010) Givenness Hierarchy theory (Gundel et al., 1993), a linguistic theory that captures the relation between different referring forms and the *cognitive status* of referents in listeners' minds. For example, per this theory, when a speaker uses *this*, one can infer they assume their target referent to be *activated* in their listener's mind; when a speaker uses *it*, one can infer they assume their target referent to be *in focus*.

Previous work on *cognitive status*-informed models of Referring Form Selection have largely been evaluated in terms of fit to human data using objective metrics like accuracy and notably human evaluations of these computational model in live human-robot interactions by Han and Williams (2023). That is, previous researchers have only assessed whether the referring forms predicted by their models *match* the referring forms that people actually use in human-human interactions.

While assessing fit to human data supports these models *as cognitive models*, it obfuscates a key dimension of Referring Form Selection: when a human selects a Referring Form during Referring Form Selection, there is no one "correct" form for them to select. In many contexts, for example, *the N'* and *that N'* may be relatively equally appropriate. Even when an object is truly in focus, warranting the use of the extremely concise *it*, the use of *the-N* is not *wrong*; and in fact, in some such cases, the use of *the-N* may be advantageous as it is simply more natural sounding.

As such, while cognitive status-informed models of Referring Form Selection have been shown to achieve good fit to human data, (1) high goodness of fit may be an unnecessarily aggressive benchmark, and (2) it is unclear whether the referring forms selected by these models are actually any more natural than those that would be produced if simpler baseline models were used.

In this work, we thus compared the naturalness of referring forms selected by cognitive status-informed referring form selection models (specifically, that presented by Han et al. (2022)) to those that would be selected by a variety of baselines, including a random baseline, and a definite description baseline (in which a definite noun phrase *the N* is always used).

To do so, we conducted an experiment in which we modified a dataset of task-based referring expressions, systematically varied the referring forms shown to participants, allowing us to collect naturalness ratings for all possible referring forms that could have been used in those referring expressions. We then consider, for each of those referring expressions, what referring form each of the compared models would have predicted, and thus what the perceived naturalness would have been. Averaging these naturalness predictions for each model, we are able to compare the overall naturalness of the considered models.

As we will show, our results suggest that even though cognitive status-informed referring selection models achieve good fit to human data, they do not (yet) produce concrete benefits in terms of naturality. But our results also demonstrate the challenges of performing this type of evaluation, as even the utterances produced by humans had high variability in perceived naturality.

2 Related Work

Arnold and Zerkle (2019) argues that linguistic Referring Form Selection models generally fall into two categories: *rational* and *pragmatic*. *Rational* models (e.g. Aylett and Turk (2004); Frank and Goodman (2012)) could explain the use of pronouns from an egocentric perspective, i.e., in terms of their ease of use in conversations. *Pragmatic* models, on the other hand, could explain the use of pronouns from an allocentric perspective, i.e., in terms of the assumptions about interlocutors that lead to their use. These allocentric accounts are typically grounded in theoretical constructs like cognitive status (Grosz et al., 1995). Although

these pragmatic models vary in terms of the constructs they use to explain referring form choice (e.g., givenness (Gundel et al., 1993), and focus (Grosz et al., 1995; Brennan et al., 1987; Grosz and Sidner, 1986), these models are all centered around the assumption that referring form selection is based on the status a referent has in a conversation or in the mind of conversational participants.

While both of these models make important contributions to the literature, neither performs at exceptional levels when it comes to predicting which specific referring forms to use. As Arnold and Zerkle (2019) pointed out, rational models suggest using reduced forms vastly more often than seen in practice, and fail to predict referring forms that are equally short. Furthermore, Arnold and Zerkle (2019) and Grüning and Kibrik (2005) note that both kinds of models focus on individual events or factors, such as recency in conversation (Mann et al., 1989), instead of developing a fully comprehensive model for all of what reference production entails.

Artificial Intelligence (AI) researchers developing Referring Form Selection Models have the same problems (Ge et al., 1998; McCoy and Strube, 1999; Callaway and Lester, 2002; Poesio et al., 2004; Kibble and Power, 2004; Kibrik, 2011; Kibrik et al., 2016). AI RFS models can be broadly categorized as *multi-factorial process modeling*, where the prediction of referring forms is approached as a problem of classification based on various linguistic and contextual features. Much like previous models, those discussed by Kibrik (2011), Van Deemter et al. (2012) and Gatt et al. (2014) opt for predicting pronoun use in general, as opposed to predicting the use of specific referring forms. Additionally, models like those listed above tend to be trained using purely textual domains (e.g., Krasavina and Chiarcos (2007)'s) that are very different from situated domains. Situated domains are highly ambiguous, with large numbers of nearly identical objects, and require speakers to make run-time decisions based on linguistic features (like prosody) and non-linguistic features (like physical distance) that may be assessed using noisy sensors.

To fix these problems, Pal et al. (2020) presented dynamic models of cognitive status based on the Givenness Hierarchy (Gundel et al., 1993), which they used to produce cognitive status-informed RFS models. In addition to cognitive status, these models included situated features like physical distance, leading to promising results (Pal et al., 2021). Han et al. (2022) solved a number of external ecological issues in the task environment that referring form data was collected from, e.g., including repeated and non-present objects. This led to a wider variety of referring forms collected. However, both Pal et al. (2021) and Han et al. (2022) only evaluated their work in terms of model goodness-of-fit to human data[1]; that is, how well the model predicts the referring forms used by others. While goodness-of-fit is a valuable metric, it obscures the fact that when choosing a referring form, there are often multiple referring forms that might be appropriate. As such, focusing on goodness-of-fit simultaneously risks underselling the performance of a model and overselling the benefits of the model. Because of the ecological validity and the wide range of referring forms from Han et al. (2022)'s model, we used it in this work.

3 Hypothesis

Due to the incorporation of cognitive statuses of objects in listeners' mind, we believe that referring forms predicted by cognitive status-informed models will have higher perceived naturalness, compared to random selection and compared to use of definite descriptions alone.

4 Method

In order to evaluate the naturalness of different referring forms, we used a novel experimental design in which we (1) collected naturalness ratings for each of a large set of referring forms across a large set of referring contexts, and (2) used these ratings to determine the overall naturalness of several competing models, by seeing what the naturalness ratings *would have been* under the referring forms selected by those models across those referring contexts.

To collect naturalness ratings, we performed an online experiment in which participants watched a series of videos from a dataset of human-human task based interactions. After each video, the participant was shown, in text, the next utterance spoken

[1] In research performed in parallel, completed after, but formally published before this work, we also conducted in-person human evaluations of the naturalness of our referring form selection model (Han and Williams, 2023). While that work is beyond the scope of the present paper, readers may want to consult that paper, which reinforces and adds nuance to the result of this paper.

Figure 2: The task environment by Bennett et al. (2017) where a person instructs another person to re-configure objects to the layout of smaller models on the near table.

in the video, and asked to evaluate its naturalness.

We will now step through (1) the specific stimuli used in this experiment; (2) the experimental design that determined how these videos were shown to participants; (3) how we systematically varied the referring forms shown to participants and assessed the naturalness of those referring forms; (4) the overall experimental procedure; and (5) our participant demographics.

4.1 Stimuli

The videos we showed to participants were those collected by Bennett et al. (2017). This dataset contains videos from an experiment involving dyadic interactions in which one participant instructs another participant in how to rearrange a set of boxes and cans in order to match a desired configuration. The task environment is shown in Figure 2.

We selected ten videos from Bennett et al. (2017)'s dataset, and divided each into ten sub-videos, each of which ended immediately before the nth referring expression where $n \in \{1 \ldots 10\}$. That is, for each video, we constructed ten excerpts, the first of which started at the beginning of the task and ended immediately before the first referring expression, the second started at the beginning of the task and ended immediately before the second referring expression, and so forth. All videos were subtitled for clarity. Figure 3 left shows a video.

We selected videos that contained a wide range of referring forms. As shown in Table 2, the distribution of referring forms in the original dataset is extremely skewed, with *the* $\langle N \rangle$ and $\langle N \rangle$ taking 85%. In contrast, the first three referring forms in the chosen videos approximately take 30% each.

4.2 Experimental Design

Each participant watched ten videos, each of which was an excerpt from a different one of the ten videos (i.e., video 0, 1, 2, 3, 4, 5, 6, 7, 8, or 9), and each of which ended at a different cutpoint, (i.e., immediately before referring expression 0, 1, 2, 3, 4, 5, 6, 7, 8, or 9 in that video). The average length of the videos was 35.6 seconds. The sequence of videos watched by each participant was selected using a Graeco-Latin square design (Grant, 1948) to ensure that each participant saw ten different videos of ten different interactions of ten different lengths while controlling for ordering effects.

4.3 Manipulations and Measures

As mentioned above, at the end of each video, participants were shown, in text, the utterance that immediately followed where the video cut off. These utterances were manipulated to vary the referring form used in the expression, with the actual referring form from the video replaced by one of the following: {it, this, this-N', that, that-N', the-N', N'}. For example, if the original utterance was "Now push **box D** to the left", participants were shown "Now push **it** to the left", "Now push **this box D** to the left" and so forth. These referring forms were selected at random for each video according to a pre-determined schedule. After being shown this "next utterance", participants were asked to rate its naturalness on a 5-point Likert item, with 1 being very unnatural and 5 being very natural.

4.4 Procedure

Participants first completed an informed consent form, read their task instructions, and answered demographic questions. Next, to ensure participants could hear what was said in the videos and avoid bots automatically filling out the questionnaires, participants performed an audio/video check. If participants passed this check, they proceeded to watch ten videos based on their randomly assigned Graeco-Latin square row, after each of which they answered the naturalness questions. Finally, participants answered an attention check question (asking the color of the walls in the task environment). This experiment's design and procedure were approved by the authors' institution.

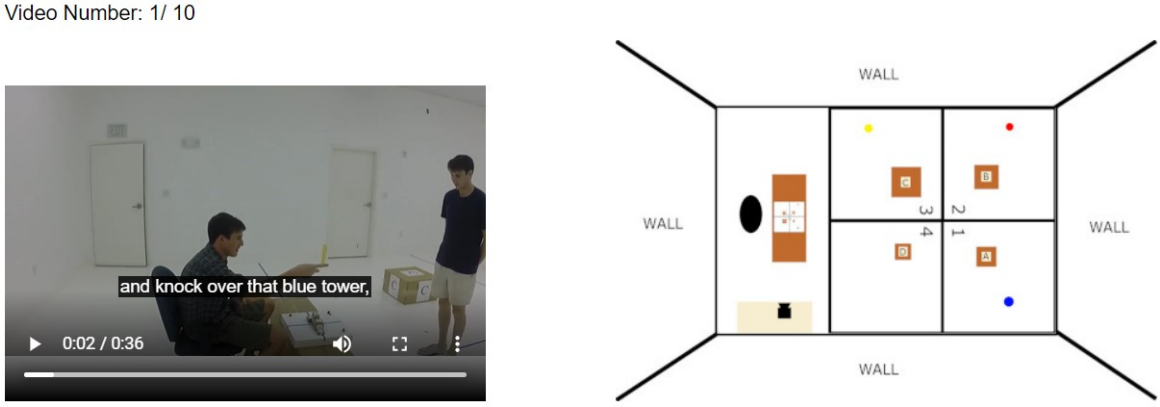

Figure 3: **Left**: The first video of ten videos that was playing and ended immediately before the first referring form. **Right**: The map used for participants to select the object being referred to (here, Box C). After the selection, the correct object would be similarly highlighted on the map.

4.5 Participants

361 participants were recruited globally from Prolific (Palan and Schitter, 2018) with a requirement that they must be fluent in English. 14 were excluded for failing the attention check, yielding 347 data points. The average age of the participants was 25.8 years (SD=7.06). The minimum age was 18, and the maximum was 65. 164 (47.3%) participants identified as female, and 175 (50.4%) identified as male. Four (1.2%) participants identified as non-binary and one (0.3%) identified as genderqueer. Three (0.9%) participants declined to identify their gender. Participants were asked to self-identify in terms of race and ethnicity. The categories with more than 5 participants are White/Caucasian (192, 55.3%), Black/African (72, 20.7%), and Latinx/Hispanic (31, 8.9%). All other racial or ethnic identities comprised less than 1.7% (6 participants). Each participant was paid USD $4.00 for their time.

5 Analysis

5.1 Data Matrix

All naturalness scores were aggregated into a 100×7 matrix, where each of the 100 rows represented a different video excerpt shown to participants, and each of the 7 columns represented a different referring form. For example, the first cell in the matrix contained the average naturalness scores for the utterance following cutpoint 0 in video 0, when the referring expression in that utterance was replaced with "it". The remainder of the first row contained the average naturalness scores for the other possible referring forms used following video 0 cutpoint 0. This data matrix was then used to evaluate a set of Referring Form Selection models, as described in the next section.

5.2 Models

To test our hypothesis, we compared five models using this data matrix. For each model, we considered each row in the data matrix, and identified which referring form the model would have predicted in the referring context encoded by that row. We then extracted the naturalness score from the column associated with that prediction. This produced a set of 100 naturalness scores for each model.

The five models we compared were (1) a *Random* baseline; (2) a *Definite Description* baseline; (3) a *Human* baseline; (4) Han et al. (2022)'s cognitive status-informed model where utterance-level temporal distance is used instead of object-level temporal distance; (5) A modified model trained with the physical distance being the furthest. This is in line with the data that the cognitive status model was trained on. Ideally, reference-level temporal distance would be utilized for both the CS model and the RF model, as it is more accurate on an object-mention-per-object-mention basis, but this would create conflict between the two models.

5.2.1 Random Model

The Random Model served as our first baseline. Under this model, a referring form was selected at random: for each row in our data matrix, the naturalness score from a random column was used.

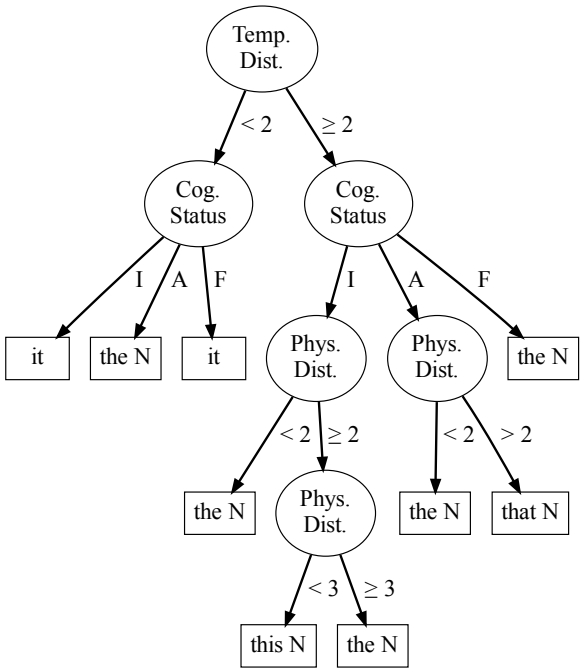

Figure 4: The visualization of the decision tree model used in the HanRFS condition, adapted from Han et al. (2022) with temporal distance at utterance level.

5.2.2 Definite Description Model

The Definite Description Model served as our second baseline. Under this model, *The N'* was always used: for each row in our data matrix, the naturalness score from the *The N'* column was extracted.

5.2.3 Human Model

The Human Model serves as our third baseline. Under this model, the referring form actually used by participants in the dataset was used. For example, if for a particular cutpoint in a particular video, the utterance following the cutpoint actually contained "that tower", then the naturalness score from the *That N'* column was extracted.

5.2.4 Cognitive Status-Informed Model (HanRFS)

For our fourth baseline, we used the cognitive status-informed model presented by Han et al. (2022) (under MIT licence). This model is a decision tree based machine learning model, shown in Figure 4, and uses three features: cognitive status, temporal distance, and physical distance.

Cognitive status was predicted using Pal et al. (2020)'s Bayesian cognitive status engine, which itself makes predictions based on an object's linguistic status and previously inferred cognitive status. That is, for a referring form to be predicted at cutpoint t, we fed each referring form from cutpoints $0 \ldots t-1$ (if any) to Pal's cognitive status engine. This produced a distribution over cognitive status that the target referent should have at time t. We then used the most likely cognitive status from this distribution as the feature passed to Han et al. (2022)'s decision tree.

Far Left Quad: 6	On Vertical Line between Far Quads: 5	Far Right Quad: 6
On Left Horizontal Line: 4	At Line Intersection: 3	On Right Horizontal Line: 4
Near Left Quad: 2	On Vertical Line between Near Quads: 1	Near Right Quad: 2

Table 1: Codes for physical distance.

Temporal distance was calculated as recency of mention: a target referent's temporal distance was calculated as the number of utterances since the utterance where the object was mentioned.

Physical distance was calculated in terms of qualitative distance-to-object. Han et al. (2022)'s original model was trained in a tabletop environment, and as such, they operationalized physical distance by assigning a set of distance scores 1-6 to each area in a 3×3 grid on the tabletop. We elected to do the same, breaking the task environment shown in the video into a 3×3 grid, and assigning a distance score 1-6 to each quadrant as shown in Table 1.

5.2.5 Modified Cognitive Status-Informed Model (HanRFS-RD)

Since the location of objects could have an impact on the choice of referring form and their perceived naturalness, we decided to include another baseline, HanRFS-RD (Remapped Distances), in which all physical distances were set to the furthest possible value (6), since all objects in Bennett et al. (2017)'s environment were further than any object used to train the decision tree model.

5.3 Data Analysis

To compare the predictions made by each of our five models, we used the Bayesian statistical framework (Wagenmakers et al., 2018), given its capability to quantify evidence both for and against a hypothesis, compared to the Frequentist approach. Specifically, we used JASP 0.17.1 (JASP Team, 2022) to run Bayesian statistical tests.

One important concept to understand in the

Table 2: Referring form distribution across conditions

	it	*the* $\langle N' \rangle$	$\langle N' \rangle$	*this*	*that*	*this* $\langle N' \rangle$	*that* $\langle N' \rangle$
Original	0.12	0.54	0.31	0.00	0.00	0.01	0.02
Random	0.13	0.16	0.08	0.14	0.13	0.16	0.21
The $\langle N' \rangle$	0.00	1.00	0.00	0.00	0.00	0.00	0.00
Human	0.28	0.29	0.29	0.01	0.02	0.04	0.07
HanRFS	0.30	0.32†	0.00†	0.00	0.00	0.27	0.10
HanRFS-RD	0.30	0.60†	0.00†	0.00	0.00	0.00	0.10

†The two models by Han et al. (2022) merged $\langle N \rangle$ with *the* $\langle N \rangle$

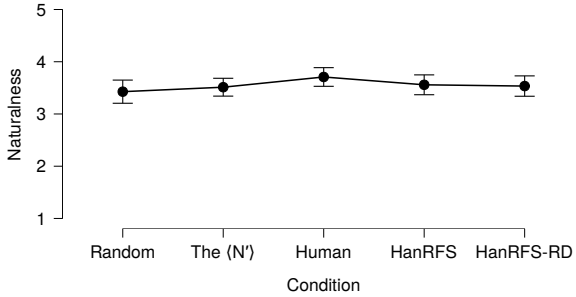

Figure 5: Mean naturalness ratings. Error bars show 95% credible intervals. Results favor no difference across conditions.

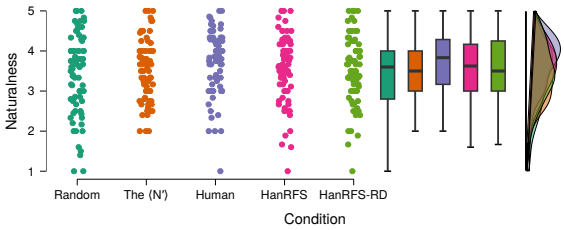

Figure 6: A raincloud plot (Allen et al., 2019) for naturalness ratings, combining a cloud of points with a box plot and a one-sided violin plot.

Bayesian approach is the Bayes factor (BF), defined as the 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 a favor of $\mathcal{H}_1$ that the data are five times more likely under $\mathcal{H}_1$ than under $\mathcal{H}_0$.

To help the decision-making process, we used the widely-accepted classification scheme (Lee and Wagenmakers, 2014). For evidence favoring $\mathcal{H}_1$, Bayes factor values are categorized as anecdotal ($BF_{10} \in (1, 3]$; inconclusive), weak ($BF_{10} \in (3, 10]$), moderate ($BF_{10} \in (10, 30]$), strong ($BF_{10} \in (30, 100]$), extreme ($BF_{10} \in (100, \infty)$). When evidence favors $\mathcal{H}_0$, these thresholds are inverted, and we can use BF_{01} for easier interpretation (Note the subscript is 01 rather than 10). For example, $BF_{10} = 1/5 = 0.2$ can be expressed as $BF_{01} = 5$.

6 Results

6.1 Model comparisons

Table 2 shows the distribution of the referring forms across the five conditions. The distribution of random referring forms, by definition, roughly follows a uniform distribution, with deviance due only to sampling noise. The $\langle N \rangle$ condition contains only itself. For the model condition, it predicts 30% of *it*, 31% of *the* $\langle N \rangle$, 27% of *this*, and 10% of *that* $\langle N \rangle$. On the other hand, for the fixed physical distance model, 30% of *it*, 60% of *the* $\langle N \rangle$, and 10% of *that* $\langle N \rangle$ was predicted, in line with the expected changes by increasing physical distance to its maximum value. Note that Han et al. (2022) took a descriptivist view (Frege, 1892; Russell, 2001; Nelson, 2002) and merged bare nouns ($\langle N \rangle$) with definite nouns (*the* $\langle N \rangle$).

6.2 Naturalness in Referring Form Selection

As seen from Figure 5, the mean naturalness scores are approximately the same in all five conditions and, surprisingly, the actual referring forms were only rated slightly higher: Random (M=3.427, SD=0.977), The $\langle N' \rangle$ (M=3.513, SD=0.752), HanRFS (M=3.559, SD=0.837), HanRFS-RD (M=3.534, SD=0.859), and Human (M=3.621, SD=0.814). Out of all of these, the Human baseline performed marginally better (M=3.621), followed by HanRFS (M=3.559), HanRFS-RD (M=3.534), The $\langle N' \rangle$ (M=3.513) and Random (M=3.427); however, due to the SD value being greater than 0.75 for all models, there is no statistically significant difference between them. Figure 6 shows the raw data points with boxplots and distribution estimation.

To see whether there is a difference, we conducted a Bayesian one-way analysis of variance (ANOVA) (Rouder et al., 2012) on the naturalness data. This analysis revealed strong evidence against the effect of different referring form selections and the actual referring forms ($BF_{01} = 28.358$), i.e., favoring $\mathcal{H}_0$. This means that the data are around 28.358 times more likely under models that did not include an effect than under those that did. Thus, the hypothesis is not supported: the referring forms predicted by the cognitive status-informed model were not perceived as more natural.

7 Discussion

7.1 Naturalness is in the Eye of the Beholder

We hypothesized that the cognitive status-informed RFS model's predictions would be more natural. However, results showed that all RFS models are equally natural, with extremely high variability in perceived naturality even for human-generated referring expressions. There are a variety of possible explanations for these observations.

First, humans may regularly generate unnatural sounding referring expressions. If so, human-level naturality may merely be a "low bar" that NLG research should seek to surpass.

Second, humans may vary dramatically in their perceptions of what is "natural". If so, human judgment may be a poor way to assess referring form naturality.

Third, our experimental paradigm may have been unsuccessful in measuring the naturality of referring forms on their own. Referring forms are always used in the context of a larger utterance, which itself may be viewed as natural or unnatural. To mitigate this concern, we specifically asked participants about the naturalness of the referring forms used in the utterance they were shown. However, it is possible that participants either did not follow these instructions, or were simply unable to adjudicate the naturalness of these forms without considering the broader context of their use. For example, in the utterance "And while you're there can you knock over **the blue tower**", participants rated the usage of **blue tower** ($\langle N \rangle$) as very natural. The separation of the naturalness of the referring form from its context of use is remarkably challenging because, without context, using concise referring forms becomes no longer useful.

Finally, our results may be due to the global population reflected in our sample. While participants were required to be fluent in English, most participants indicated that English was their second language. This may have led to significant variation in our naturalness ratings.

Takeaway 1: Future work needs to better separate the perceived naturalness of a referring form from its context, such as dialog. To confirm this, one may need to measure the naturalness of the context as a controlling factor, or may need to be particularly aggressive about reminding participants that they are rating only the referring form itself.

7.2 How Far is Far?

We included the modified HanRFS-RD model with remapped distances because of the differences between what is considered "close" and "far" in our analyzed dataset versus the dataset on which Han et al. (2022)'s model was trained on. This raises a larger question, however, of how to model referent distances in a task-agnostic way. Physical distance is clearly an important factor, and is known to play a role in differentiating referring forms like "this" vs "that", as well as differentiating the use of abstract versus precise deictic gestures (Stogsdill et al., 2021). Yet what is considered near versus far is highly task dependent, depending not only on the overall size of the space, but also on the physical affordances and explorability of the space. For example, Han et al. (2022)'s model was created in a space that was smaller than Bennett et al. (2017). But moreover, while in Bennett et al. (2017)'s experiment objects were out of immediate reach of the participants, in Han et al. (2022)'s experiment, objects were reachable without walking around, i.e., no farther than 60cm (2 feet) away, and in fact were touched and manipulated by participants. In other task environments, other features may also become relevant. In large-scale open-world environments, for example, many referents are non-visible (or may not even be known to exist) when they are referred to.

Takeaway 2: Future work needs to understand how referring form selection models can encode physical distance features in a way that is agnostic of, or relative to, the size of a task environment; should consider inclusion of a suite of distance features sensitive to different types of task environments; and should consider features related to but distinct from distance, like reachability, manipulability, and visibility.

8 Conclusions

To go beyond the focus on goodness-of-fit in cognitive status-informed computational referring form selection model evaluation, we conducted a human-subjects study to explore the naturalness ratings of the predictions. Surprisingly, results did not reveal an improvement in naturalness over random baselines, and in fact suggest that human perceptions of even *human-generated* referring forms are incredibly varied and not significantly different from those random baselines. Our results suggest several directions for future work, and new technical and

methodological considerations that must be made.

Supplementary Materials Availability Statement: All videos, data, and analysis scripts are available at https://osf.io/z2wyt/. All data was anonymized, replacing participants' names with automatically assigned numerical identifiers.

Acknowledgements

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

References

Micah Allen, Davide Poggiali, Kirstie Whitaker, Tom Rhys Marshall, and Rogier A Kievit. 2019. Raincloud plots: a multi-platform tool for robust data visualization. *Wellcome open research*, 4.

Jennifer E Arnold and Sandra A Zerkle. 2019. Why do people produce pronouns? pragmatic selection vs. rational models. *Language, Cognition and Neuroscience*, 34(9):1152–1175.

Matthew Aylett and Alice Turk. 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.

Maxwell Bennett, Tom Williams, Daria Thames, and Matthias Scheutz. 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)*, pages 6589–6594. IEEE.

Susan E Brennan, Marilyn W Friedman, and Carl Pollard. 1987. A centering approach to pronouns. In *25th Annual Meeting of the Association for Computational Linguistics*, pages 155–162.

Maya Cakmak and Andrea L Thomaz. 2012. Designing robot learners that ask good questions. In *2012 7th ACM/IEEE International Conference on Human-Robot Interaction*, pages 17–24. IEEE.

Charles B Callaway and James C Lester. 2002. Pronominalization in generated discourse and dialogue. In *ACL*.

Guanyi Chen, Fahime Same, and Kees van Deemter. 2021. What can neural referential form selectors learn? In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 154–166.

Michael C Frank and Noah D Goodman. 2012. Predicting pragmatic reasoning in language games. *Science*, 336(6084):998–998.

Gottlob Frege. 1892. Über sinn und bedeutung. *Zeitschrift für Philosophie und philosophische Kritik*, 100:25–50.

Albert Gatt, Emiel Krahmer, Kees Van Deemter, and Roger Van Gompel. 2014. Models and empirical data for the production of referring expressions. *Lang., Cognition and Neuroscience*, 29(8):899–911.

Niyu Ge, John Hale, and Eugene Charniak. 1998. A statistical approach to anaphora resolution. In *Workshop on Very Large Corpora*.

Felix Gervits, Gordon Briggs, Antonio Roque, Genki A Kadomatsu, Dean Thurston, Matthias Scheutz, and Matthew Marge. 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*, pages 150–158.

David A Grant. 1948. The latin square principle in the design and analysis of psychological experiments. *Psychological bulletin*, 45(5):427.

Barbara J Grosz, Aravind K Joshi, and Scott Weinstein. 1995. Centering: A framework for modelling the local coherence of discourse. *Computational Linguistics*.

Barbara J Grosz and Candace L Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational linguistics*, 12(3):175–204.

André Grüning and Andrej A Kibrik. 2005. Modelling referential choice in discourse: A cognitive calculative approach and a neural network approach. *Anaphora processing: Linguistic, cognitive and computational modelling*, 263:163.

Jeanette K Gundel, Mamadou Bassene, Bryan Gordon, Linda Humnick, and Amel Khalfaoui. 2010. Testing predictions of the givenness hierarchy framework: A crosslinguistic investigation. *Journal of Pragmatics*, 42(7):1770–1785.

Jeanette K Gundel, Nancy Hedberg, and Ron Zacharski. 1993. Cognitive status and the form of referring expressions in discourse. *Language*, pages 274–307.

Zhao Han, Polina Rygina, and Thomas Williams. 2022. Evaluating referring form selection models in partially-known environments. In *Proceedings of the 15th International Conference on Natural Language Generation*, pages 1–14.

Zhao Han and Tom Williams. 2023. Evaluating cognitive status-informed referring form selection for human-robot interactions. In *2023 Annual Meeting of the Cognitive Science Society (CogSci)*.

Ryan Blake Jackson and Tom Williams. 2022. Enabling morally sensitive robotic clarification requests. *ACM Transactions on Human-Robot Interaction (THRI)*, 11(2):1–18.

JASP Team. 2022. JASP (Version 0.16.4)[Computer software].

Rodger Kibble and Richard Power. 2004. Optimizing referential coherence in text generation. *Comp. Ling.*

Andrej A Kibrik. 2011. *Reference in discourse*. OUP.

Andrej A Kibrik, Mariya V Khudyakova, Grigory B Dobrov, Anastasia Linnik, and Dmitrij A Zalmanov. 2016. Referential choice: Predictability and its limits. *Frontiers in psychology*, 7:1429.

Emiel Krahmer and Kees Van Deemter. 2012. Computational generation of referring expressions: A survey. *Computational Linguistics*, 38(1):173–218.

Olga Krasavina and Christian Chiarcos. 2007. Pocospotsdam coreference scheme. In *Linguistic Annotation Workshop*.

Michael D Lee and Eric-Jan Wagenmakers. 2014. *Bayesian cognitive modeling: A practical course*. Cambridge university press.

William C Mann, Christian MIM Matthiessen, and Sandra A Thompson. 1989. Rhetorical structure theory and text analysis. *NASA STI/Recon Technical Report N*, 90:26733.

Kathleen F McCoy and Michael Strube. 1999. Generating anaphoric expressions: pronoun or definite description? In *The Relation of Discourse/Dialogue Structure and Reference*.

Michael Nelson. 2002. Descriptivism defended. *Noûs*, 36(3):408–435.

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*, volume 43.

Poulomi Pal, Lixiao Zhu, Andrea Golden-Lasher, Akshay Swaminathan, and Tom Williams. 2020. Givenness hierarchy theoretic cognitive status filtering. In *Annual Meeting of the Cognitive Science Society*.

Stefan Palan and Christian Schitter. 2018. Prolific.ac—a subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, 17:22–27.

Massimo Poesio, Rosemary Stevenson, Barbara Di Eugenio, and Janet Hitzeman. 2004. Centering: A parametric theory and its instantiations. *Computational linguistics*, 30(3):309–363.

Jeffrey N Rouder, Richard D Morey, Paul L Speckman, and Jordan M Province. 2012. Default bayes factors for anova designs. *Journal of mathematical psychology*, 56(5):356–374.

Bertrand Russell. 2001. *The problems of philosophy*. OUP Oxford.

Fahime Same and Kees van Deemter. 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*, pages 4575–4586.

Kevin Spevak, Zhao Han, Tom Williams, and Neil T Dantam. 2022. Givenness hierarchy informed optimal document planning for situated human-robot interaction. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*.

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*, pages 392–396.

Stefanie Tellex, Pratiksha Thakerll, Robin Deitsl, Dimitar Simeonovl, Thomas Kollar, and Nicholas Royl. 2013. Toward information theoretic human-robot dialog. *Robotics*, page 409.

Kees Van Deemter. 2016. *Computational models of referring: a study in cognitive science*. MIT Press.

Kees Van Deemter, Albert Gatt, Roger PG Van Gompel, and Emiel Krahmer. 2012. Toward a computational psycholinguistics of reference production. *Topics in cognitive science*, 4(2):166–183.

Eric-Jan Wagenmakers, Maarten Marsman, Tahira Jamil, Alexander Ly, Josine Verhagen, Jonathon Love, Ravi Selker, Quentin F Gronau, Martin Šmíra, Sacha Epskamp, et al. 2018. Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic bulletin & review*, 25(1):35–57.

Tom Williams, Gordon Briggs, Bradley Oosterveld, and Matthias Scheutz. 2015. Going beyond literal command-based instructions: Extending robotic natural language interaction capabilities. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.