# Evaluating Referring Form Selection Models in Partially-Known Environments

**Zhao Han** and **Polina Rygina** and **Tom Williams**
MIRRORLab
Department of Computer Science
Colorado School of Mines
zhaohan@mines.edu, prygina@mines.edu, twilliams@mines.edu

## Abstract

For autonomous agents such as robots to effectively communicate with humans, they must be able to refer to different entities in situated contexts. In service of this goal, researchers have recently attempted to model the selection of *referring forms* on the basis of cognitive status (informed by Givenness Hierarchy), and have shown promising results with over 80% accuracy. However, we argue that the task environments lack ecological validity, due to their use of a small number of objects that are constantly activated and easily uniquely identifiable. Accordingly, we present a novel building-construction task that we believe has increased ecological validity. We then show how training cognitive status informed referring form selection models on data collected within this novel task environment yields substantially different results from those found in previous work, providing key insights and directions for future work.

## 1 Introduction

One of the most studied dimensions of natural language pragmatics is *reference*: the process by which speakers pick out, or *refer*, to things of interest in the environment, and how hearers interpret, or *resolve* those references. The generation (or production) side of this problem has attracted sustained attention across a variety of communities, including philosophy of language, psycholinguistics, and artificial intelligence – so much so that referring has been called the "fruit fly" of language (Van Deemter, 2016).

The vast majority of research on referring, however, has been focused on problems like *Referring Expression Generation* (Krahmer and Van Deemter, 2012), in which the goal is to select the *properties* that will be used in a generated expression (e.g., choosing to highlight the redness, or the boxiness, of a red box, among other possible properties). In contrast, very little research has been done on the problem of computationally modeling *Referring Form Selection*, in which a speaker must select a more general *referring form*, such as "it", "that", or "the $\langle N' \rangle$"[1], despite its accepted status as an important initial step during language production (Kibrik, 2011).

While the computational modeling of referring expression generation has been heavily understudied, it has been a key question of interest in the linguistics community, with a number of competing theories making different predictions, including Accessibility Theory (Ariel, 2001) and Givenness Hierarchy Theory (Gundel et al., 1993). Such theories thus provide natural starting points for computational modeling work. Yet while these theories provide critical linguistic insights about the nature of referring form selection, they provide little direct input into the cognitive processes, mechanisms, or algorithms that govern this process.

Recently, this has begun to change, with researchers like Pal et al. (2020) seeking to directly computationally model the mechanics of these underlying theories of reference (in their case, *Givenness* or *Cognitive Status*), and then build higher-level computational models of referring form selection that leverage those more fundamental models (Pal et al., 2021). These recent works have provided promising results, with over 80% accuracy in predicting the referring forms used by interactants in human-human and human-robot interaction scenarios.

Yet despite the promise of these results, concerns may be raised about the task environments in which those results were produced. Specifically, we argue that the task environment used in that previous work was not ideally suited for training or evalu-

[1] In this work we implicitly focus on Standard American English; but the types and distribution of general referring forms we consider have been observed across a wide variety of languages beyond English, including Mandarin, Japanese, Spanish, Russian, Eegimaa, Kumyk, Ojibwe, Arabic, Irish, Norwegian, Persian, and Turkish (Hedberg, 2013).



*Proceedings of the 15th International Conference on Natural Language Generation*, pages 1 - 14
July 18-22, 2022 

ating Cognitive Status informed Referring Form Selection models.

In this paper, we thus measure the performance of these previously published models using a better collection of tasks, making three key contributions in the process: (1) we present a novel task context that we argue is well designed for the studying of referring form selection; (2) we assess the performance of Pal et al. (2021)'s Referring Form Selection model in this setting to obtain a better estimate of its true performance in realistic task contexts; and (3) we use these results to motivate arguments as to how underlying models of cognitive status must be adapted to enable better performance on Referring Form Selection tasks.

## 2 Related Work

We will now describe prior what work has been done on Referring Form Selection, including the Cognitive Status informed work of Pal et al. (2021). We will then provide our specific critiques of the task context in which that work was trained and evaluated.

Referring Form Selection models fall into two main categories (Arnold and Zerkle, 2019). *Rational* models seek to explain how speakers egocentrically decide whether or not to use pronouns, e.g. for reasons of ease of production (Aylett and Turk, 2004; Frank and Goodman, 2012). In contrast, *pragmatic* models seek to explain how speakers allocentrically decide to use pronouns on the basis of their status as activated or focused within a conversation (Grosz et al., 1995; Brennan et al., 1987; Ariel, 1991; Gundel et al., 1993). These pragmatic models share an assumption that referring form selection is grounded in a relationship between discourse context and mnemonic or attentional states. For example, Gundel et al. (1993) suggest that referring forms are selected based on which of a hierarchically nested set of *Cognitive Statuses* ({in focus $\subseteq$ activated $\subseteq$ familiar $\subseteq$ uniquely identifiable $\subseteq$ referential $\subseteq$ type identifiable}) can be assumed to hold for the target referent.

While these models have shown promise in predicting whether or not someone chooses to use a definite noun phrase or a more reduced form, neither class of model is terribly effective at predicting precisely which form a speaker will choose to use. Rational models, for example, predict much more frequent use of reduced forms than are actually seen in practice, and fail to predict differential usage of "equally short" referring forms (Arnold and Zerkle, 2019). To make matters work, models in both categories tend to focus on specific referential phenomena, rather than trying to comprehensively model the entire process of reference production (Arnold and Zerkle, 2019; Grüning and Kibrik, 2005); and indeed often do not really try to model cognitive mechanisms or psycholinguistic processes at all (Arnold, 2016). And, of critical importance to those studying *situated* interaction, the vast majority of this previous work, in both camps, has predominantly been assessed on corpora not collected in or encoding any features of situated domains.

Work in the Artificial Intelligence community on Referring Form Selection suffers from similar problems. Most such work (Poesio et al., 2004; McCoy and Strube, 1999; Ge et al., 1998; Kibrik et al., 2016; Kibrik, 2011; Callaway and Lester, 2002; Kibble and Power, 2004) falls under *multifactorial process modeling*, in which the process of referring is modeled as a classification problem performed on the basis of a variety of features (Kibrik, 2011; Van Deemter et al., 2012; Gatt et al., 2014). Like the linguistic models discussed above, these models often do not attempt to select between referring forms at a fine-grained level, instead choosing to predict pronoun use as a whole. And, like the linguistic models discussed above, these models are often trained and evaluated in purely textual domains, such as the Wall Street Journal corpus (Krasavina and Chiarcos, 2007), thus avoiding many of the nuanced challenges that arise in situated domains, which are highly ambiguous and open worlds, and in which agents must make decisions on the basis of features that can be readily and immediately assessed, which may well go beyond purely linguistic features, including features of the environment in which dialogue is situated.

Some recent research efforts have attempted to fix these problems. Pal et al. (2020), for example, have presented models for dynamic modeling of Cognitive Status (a construct underlying Givenness Hierarchy theoretic accounts of referring (Gundel et al., 1993)), and have then used these models as informative features for Referring Form Selection, with apparently good results (Pal et al., 2021). Pal et al.'s work is also notable in that it is trained on data collected in situated interaction contexts. However, even this work suffers from certain flaws that may raise similar questions about generalizability

to situated domains. Specifically, we argue that Pal et al.'s work was conducted in a task environment that may not have been well suited for training or evaluation of cognitive status informed models. Pal et al.'s model was trained and evaluated using videos collected by Bennett et al. (2017), in which humans give instructions to humans or robot interactants as to how to re-arrange a large-scale environment to match a pre-determined pattern.

This task domain may be ill-suited to studying Cognitive Status informed Referring Form Selection for several reasons. First, this domain contains a relatively small number of candidate referents, i.e., three towers of cans and four labeled boxes. This could result in an irregular situation in which the majority of task-relevant objects are constantly at least *activated* (which, in a Givenness Hierarchy theoretic account, would enable the use of referring forms such as *this*), and are likely to remain so regardless of dialogue context merely due to the small number of observed task relevant objects. Second, all task-relevant objects in this domain are easily uniquely discriminable. Each of the "towers" has a unique context, and each of the boxes is labeled with a unique letter. This means that speakers may be able to over-rely on proper nouns and simple single-property descriptions, and would not need to seriously consider their choice of referring expression. Third, all task-relevant objects in this domain are visible at all times. This is likely to exacerbate the challenges listed above. Moreover, it is likely to completely preclude the need for indefinite descriptions, which are often used when the speaker assumes that the listener does not already have knowledge of their target referent.

In this work, we seek to address these challenges. We begin by collecting a new corpus of sequential referring expressions in a task context that does not have these shortcomings. Then, we re-assess the performance of Pal et al. (2021)'s Referring Form Selection models on the data collected in this more ideally suited task domain.

## 3 Environment and Task Design

To collect a wider variety of referring forms in a situated context, we designed a dyadic interaction task (Shown in Figure 1) in which pairs of participants perform four tower constructions tasks in four visually separated quadrants of a larger task environment. The task environment is separated into four quadrants to create a partially-observable environment in which participants can readily observe their current quadrant, but not the other quadrants. Each quadrant is filled with block shapes, including triangles, cubes, cuboids, cylinders, arches, and half-circles. All blocks are distributed to the corners and intersections of a $3 \times 3$ grid.

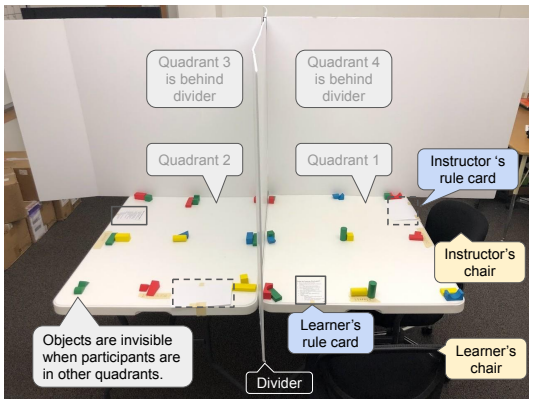


Figure 1: Two of four quadrants of the task environment. To promote a wide variety of referring forms, we placed objects in different quadrants with careful manipulation of target referent visibility (thus leading to course-grained variance in cognitive status) and by requiring repeated reference to task referents (thus leading to fine-grained variance in cognitive status).

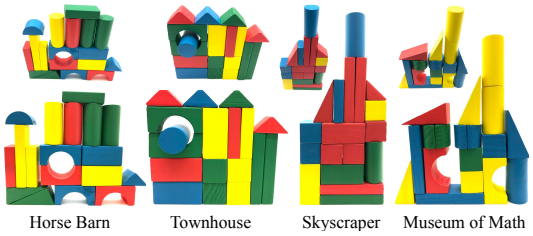


Figure 2: The buildings in the construction task. Two angles were provided to help participants to recognize constituent block shapes.

The described task environment is used as the setting for a series of four dyadic construction tasks, one in each of the four quadrants. Each task requires one participant (the Teacher) to instruct the other participant (the Learner) to construct a building based on a given image (Figure 2). The Learner, in turn, must work to construct the tower piece by piece as it is described to them, without speaking themselves, using only the resources available in their current quadrant unless the Teacher instructs them to seek a block in a different quadrant. Note that participants do not statically provide or listened to monolithic multi-minute monologues. In fact, the task is highly interactive, with teachers giving instructions, learners following instructions, and

then teachers providing corrections or proceeding. While learners were mostly silent while completing their tasks, this is perfectly reasonable given the particular domain we investigate in this work, i.e., deciding how to deliver multi-step task instructions.

Each building has 18 blocks, nine (50%) of which are placed in the quadrant where the building is being constructed, the other half of which are distributed in the other three quadrants. The large number of blocks in this task context ensure that, unlike in Pal et al.'s work, there are a large number of candidate referents that are not trivially distinguishable. The separation between quadrants ensures that, unlike in Pal et al.'s work, not all objects are visible at any given time. And, the distribution of blocks throughout the four quadrants ensures that the Teacher will need to refer to blocks that have not yet have been observed or which were observed in a previous construction task but which are no longer visible in the current quadrant, further diversifying the expected set of referring expressions used by Teachers.

## 4 Corpus Collection Procedure

The described environmental and task context were used to collect a new corpus of referring expressions, through the following IRB-approved procedure. Eleven pairs of participants were recruited from the campus of The Colorado School of Mines. Upon arrival, each pair of participants provided informed consent and were provided instructions about the structure of the tower construction task. Participants were then led to the task environment and seated in the first quadrant, where a photo of the target building was available to the participant assigned to be the Teacher, as seen in Figure 1. Participants were then videorecorded completing each of the four tower construction tasks in sequence. Each participant was paid $10 USD.

## 5 Corpus Annotation

The collected eleven-dyad corpus was comprised of eleven collections of four monologues each. These eleven collections averaged 27:32 minutes in length, with a minimum of 16:26 and maximum of 34:03. The average monologue length was 6:53. We first transcribed these recordings automatically online using the Dovetail qualitative analysis software[2]. The first two authors then manually verified and corrected these transcripts. The collected transcripts were then divided into a total of 1992 utterance clauses.

[2] https://dovetailapp.com/

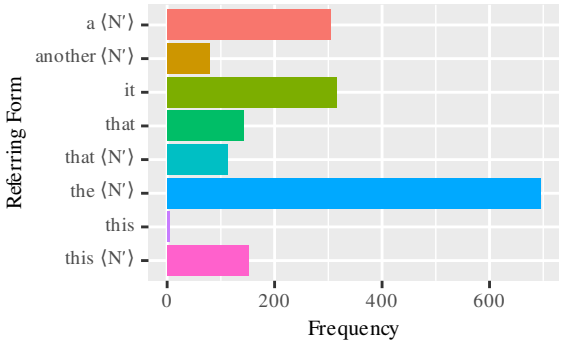


Figure 3: Distribution of a wide variety of referring forms. In addition to the six definitive nouns (right six columns), we also found participant frequently used indefinite nouns of "a $\langle N' \rangle$" and "another $\langle N' \rangle$" (left two columns).

After removing clauses that contained no referring forms (e.g., utterances made when switching quadrants and at study conclusion) or only plural referring forms (e.g., them, they), which we did not aim to model in this work, 1867 referring expressions remained, including from the corrective instructions. Each participant contributed an average of 169.7 referring forms, which is significantly more than the average of 18 (603/33) referring forms per participant in the situated interaction corpus (Bennett et al., 2017) used by Pal et al. (2021). The data does not have information that names or uniquely identifies individual people or offensive content. Below, we provide two sample utterance sequences from the collected corpus.

**Sample 1**

- *Alright. Do you see the red block over there?*
- *We need that but the blue.*
- *Awesome. Put that leg on that side to the left of the green cube, just like that.*
- *And you see the red thing I was talking about?*
- *Put that right on top of the blue thing. Perfect.*

**Sample 2**

- *And then the blue cylinder is going to go there.*
- *Okay. So for the next, we need this one.*
- *And you can go ahead and set that up right next to the triangle.*
- *And put it vertically on the inside.*

After translating each of the corpus' 44 monologues into a sequence of (non-plural) referring

forms, we categorized each into one of eight categories (See Figure 3), and annotated, at each reference point, key features of each candidate object in the environment. Critically, we ensured that the features used could all be assessed on the fly, to ensure they could actually be used in future robotics applications. In the following subsections, we detail both of these types of annotation.

### 5.1 Referring Forms

We categorized referring expressions into seven types of referring forms: *it*, *this*, *that*, *this* ⟨*N'*⟩, *that* ⟨*N'*⟩, *the* ⟨*N'*⟩ and ⟨*indefinite NP*⟩. While indefinite noun phrases took multiple forms (e.g., "a ⟨$N'$⟩", which accounted for 16.3% of all referring forms, and "another ⟨$N'$⟩", which accounted for 4.2% of referring forms (per Figure 3). Similarly, like (Pal et al., 2020), we take a descriptivist view (Frege, 1892; Russell, 2001; Nelson, 2002) and merge bare noun phrases together with definite noun phrases.

### 5.2 Object Features

Next, we discuss the features annotated for each object in the scene at each reference point. We used the same four simple features used with great success by Pal et al. (2021), both because they are easily assessable by autonomous agents like robots, and to facilitate direct comparison with Pal et al. (2021). Each of these four features is described in a subsection below.

#### 5.2.1 Cognitive Status

The first feature used was Cognitive Status, which was, unsurprisingly, the most informative feature used in Pal et al. (2021)'s Cognitive Status informed approach. To annotate the cognitive status of each object in the scene at each point of reference, we used the Cognitive Status model used by Pal et al. (2021), as defined in Pal et al. (2020). This approach uses a *Cognitive Status Engine* comprised of a set of *Cognitive Status Filters*, one for each object. Each Cognitive Status Filter is a Bayesian filter of the form:

$$p(S_o^t) = p(S_o^{t-1})p(L_o^t)p(S_o^t \mid S_o^{t-1}, L_o^t)$$

Here, $S$ is a cognitive status in $\{I, A, F\}$ (where $I$ is "In Focus", $A$ is "Activated", and $F$ is "Familiar", predicted from an object's cognitive status at the previous time point and the object's "linguistic status" at the previous timepoint $L \in \{N, M, T\}$ (where $N$ is "Not Mentioned", $M$ is "Mentioned", and $T$ is "Mentioned in a Topic Role". To compare directly with Pal et al. (2020) and Pal et al. (2021), we have made the same assumption that all objects are initially at lease familiar. This is a simplifying assumption that we will return to later.

Initially, using this model identically to how it was used by Pal et al. (2021) failed to predict any objects in the scene to be "In Focus" at any timepoint. To diagnose this problem, we created a *blended model* by linearly combining a non-probabilistic model $HL$ directly derived from linguistic rules (cp. (Pal et al., 2021)) with the probabilistic model $C$ trained by Pal et al. (2021): $C' = w_1 HL + w_2 C$, where $w_l = 0.1$, $w_c = 0.9$. Here, $HL$ is encoded as a $9 \times 3$ matrix ($S \times L$) where each row (column, as transposed below) represented a combination of a cognitive status and linguistic status at time $t-1$, and each column (row, as transposed below) represented a cognitive status at time $t$:

$$L^T = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \end{bmatrix}$$

This *blended model* was then used as the basis for each Cognitive Status Filter.

#### 5.2.2 Number of Distractors

Next, we considered the number of distractors, which are the number of objects that have a cognitive status at the same GH-theoretic tier or higher than the target. We believe the number of distractors affects how people determine referential choice, as evidenced by Ferreira et al. (2005).

#### 5.2.3 Physical distance

During our task design phase, we have intentionally placed blocks at a $3 \times 3$ grid, allowing us to classify each referring form in at least nine categories, a combination of {near (N), middle (M), far (F)} and {left (L), middle (M), right (R)}. The nine grid points are coded as shows in Table 1 (note that a participant sits below NM).

| | | |
|---|---|---|
| FL: 6 | FM: 5 | FR: 6 |
| ML: 4 | MM: 3 | MR: 4 |
| NL: 2 | NM: 1 | NR: 2 |
| | Instructor | |

Table 1: Codes for physical distance.

Additionally, each object can be in one of four distance-relevant task categories at any time: on-table (T), in-building (B), in-hand (H), and in-other-quadrant (O). Although B and H are specific to our

| Model | Removed Feature |
|---|---|
| M1 | N/A (full model) |
| M2 | Cognitive status |
| M3 | Number of distractors |
| M4 | Physical distance |
| M5 | Temporal distance |

Table 2: Five model types.

task scenario, they can be generalized: B can be seen as objects at the *task goal location*, and H can be generalized to *invisible locations*. Because T is a general term, we coded it the same as MM, i.e., 3. B and H do not have distance comparisons, we coded them as 0. As O is furthest, we coded it as 10. This was a simplifying assumption to best compare with prior work.

#### 5.2.4 Temporal Distance

Similar to Pal et al. (2021), we annotated recency of mention, i.e., temporal distance, for each object by indexing the previous occurrence of the object. TD is coded as 0 when an object is not yet mentioned in a monolog, 1 when the object is the last mentioned object, and $1/n$ where $n$ is the number of objects referred since the object was mentioned.

## 6 Computational Modeling

As we intended to interrogate the performance of previous published models, we use the same decision tree algorithm by Pal et al. (2021) for explanability and theory-building purposes. Specifically, we used the same decision tree implementation in Weka 3.8.6 (Eibe et al., 2016): REPTree (Reduced Error Pruning Tree) (Quinlan, 1987), an extention of the C4.5 algorithm. REPTree builds a decision tree using information gain and prunes the tree using reduced-error pruning (REP) with backfitting (Witten and Frank, 2002).

Similarly, we followed the same training procedure as Pal et al. (2021), training five distinct models (Table 2): a full model (**M1**), and four ablated models, removing either cognitive status (**M2**), distractors (**M3**), physical distance (**M4**), or temporal distance (**M5**). We initially set the maximum depth of the tree to six, the same as Pal et al.'s model, but the decision tree became complex and difficult to interpret/explain, we thus set the maximum allowed depth of the tree to five. Similar performance was observed at depth 5 vs. 6.

The performance of these five models (for unpruned and pruned trees) were evaluated using five-fold cross validation to further avoid over-fitting

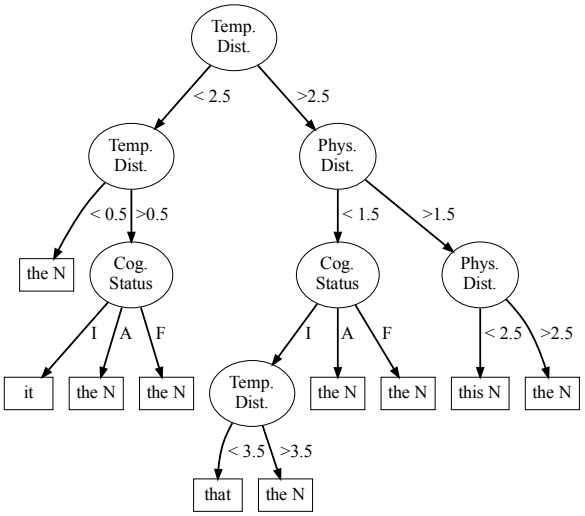


Figure 4: The decision tree visualization for the best-performing M1 (**six** main referring form categories informed by Givenness Hierarchy). {I,A,F}={In Focus, Activated, Familiar}.

(note that the tree pruning is also for this purpose). To quantify the models' performance, five common scoring metrics are used, as in (Pal et al., 2021): accuracy, root mean squared error (RMSE), precision, recall, and F1 score. The latter three are weighted by class size. Additionally, we used coverage (modeled as number of classes included in model predictions) and number of leaves to quantify model simplicity and explainability.

All data (which will be licensed under CC-BY 4.0) and code are attached to this submission.

### 6.1 Results

Table 3 shows the results for the full, unpruned trees; Table 4 shows the results for the pruned trees, which had lower coverage but are more readily interpretable. In both tables, the left and right sides show results with and without indefinite forms included. We consider these separately as Pal et al. (2021) did not consider indefinite noun phrases.

In this section we will more deeply interrogate the results of the pruned trees, as they are more readily interpretable. As seen in Table 4 left, we achieved 61%–66% accuracy for M1-M5. M1 and M3 (removing the number of distractors) are top-performing on all metrics. For M5 where the temporal distance feature was not used, the performance is slightly dropped to 61.72%. All models scored similarly in other metrics.

Fig. 4 shows a tree visualization for the M1 model. From the top node, it branches at the tem-

| | *Six GH informed referring forms* | | | | | *With two indefinite forms* | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **M1** | **M2** | **M3** | **M4** | **M5** | **M1'** | **M2'** | **M3'** | **M4'** | **M5'** |
| Accuracy | 66.01 | 63.41 | 65.87 | 61.58 | 61.08 | 59.50 | 59.00 | 59.67 | 51.02 | 57.17 |
| RMSE | 0.340 | 0.366 | 0.341 | 0.384 | 0.389 | 0.405 | 0.410 | 0.403 | 0.490 | 0.428 |
| Precision | 0.573 | 0.527 | 0.572 | 0.514 | 0.542 | 0.509 | 0.506 | 0.512 | 0.432 | 0.498 |
| Recall | 0.660 | 0.634 | 0.659 | 0.616 | 0.611 | 0.595 | 0.590 | 0.597 | 0.510 | 0.572 |
| F1 score | 0.597 | 0.571 | 0.596 | 0.546 | 0.560 | 0.544 | 0.539 | 0.545 | 0.454 | 0.522 |
| Coverage | 5 | 4 | 4 | 5 | 5 | 6 | 6 | 6 | 6 | 6 |
| Leaves | 35 | 31 | 34 | 29 | 16 | 35 | 31 | 30 | 30 | 23 |

Table 3: Evaluation metrics and results for unpruned trees.

| | *Six GH informed referring forms* | | | | | *With two indefinite forms* | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **M1** | **M2** | **M3** | **M4** | **M5** | **M1'** | **M2'** | **M3'** | **M4'** | **M5'** |
| Accuracy | **65.73** | 64.11 | **65.80** | 62.98 | 61.72 | **59.83** | 58.95 | **59.83** | 51.30 | 57.29 |
| RMSE | 0.343 | 0.359 | 0.342 | 0.370 | 0.383 | 0.402 | 0.411 | 0.402 | 0.487 | 0.427 |
| Precision | 0.552 | 0.543 | 0.552 | 0.509 | 0.521 | 0.493 | 0.487 | 0.493 | 0.435 | 0.476 |
| Recall | 0.657 | 0.641 | 0.658 | 0.630 | 0.617 | 0.598 | 0.589 | 0.598 | 0.513 | 0.573 |
| F1 score | 0.589 | 0.576 | 0.589 | 0.542 | 0.556 | 0.536 | 0.528 | 0.536 | 0.445 | 0.514 |
| Coverage | 4 | 3 | 4 | 2 | 3 | 4 | 4 | 4 | 3 | 4 |
| Leaves | 10 | 6 | 10 | 5 | 6 | 7 | 6 | 7 | 9 | 7 |

Table 4: Evaluation metrics and results.

poral distance (TD) at 2.5 (root) and 0.5 (depth 1). When $TD \in [1, 2]$ (left branch), the model looks at the cognitive status, where "it" is used if an object is in focus (I), "the $\langle N' \rangle$" is used otherwise. When $TD = 0$ (i.e., $TD < 0.5$), "the $\langle N' \rangle$" is used. When TD is far ($TD \geq 3$), i.e., the right side of the tree, physical distance (PD) is used to differentiate between "this $\langle N' \rangle$" and "the $\langle N' \rangle$" of $PD \geq 3$. When the objects are closer ($PD \leq 1$, i.e., the objects are in near middle (NM), in hand or in building), cognitive status and temporal distance plays a more important role. Specifically, "that" is used when the cognitive status is in focus and mentioned a few utterances ago ($TD \geq 4$). The number of distractors was not selected as a decision node.

For the eight-class referring form classification, the accuracy score dropped up to approximately 10% to 51.30%–59.83%. M3', without the number of distractors feature, performed as well as full model M1'.

Figure 5 shows the visualization of the M1' model. Because indefinite referring forms were added and they were used to refer to non-present objects, the physical distance feature determines when to use "a $\langle N' \rangle$", as seen in the rightmost traversal. Within the task environment, on the far side, physical distance separates the usage of "this $\langle N' \rangle$" and "the N" (the third-right most and the second-right most leaves); this is exactly the same as M1 model, as seen in the rightmost subtree in Figure 4. For the cognitive status, "the $\langle N' \rangle$" is used when it is lower than In Focus (I) and the object is in front of the instructor ($PD = 1$, i.e., $PD < 1.5$). For In Focus, if the object is less temporally distant ($TD >= 3$), "it" is used as expected; Otherwise, "the $\langle N' \rangle$" is used. When the cognitive status is Activated or Familiar, "the $\langle N' \rangle$" is used.

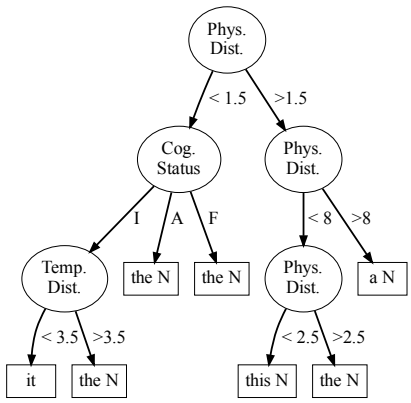


Figure 5: The decision tree visualization for the best-performing M1' model (**eight** referring form categories). Because indefinite nouns are included, physical distance became the root node.

The unpruned trees are available in Appendix 9. As there are many branches and leaves, we do not step through them here. To show the simplicity of the tree with maximum allowed depth 5, Appendix 9 shows the trees with maximum depth 6 and Appendix 9 shows those trees without maximum depth set.

Compared with Pal et al.'s model performance

2021 (72%-86%), the performance of those models trained with the new dataset, especially with the frequently-used indefinite nouns, yielded approximately 20% drop in performance.

## 7 Discussion

By designing a new task, we were able to collect a situated corpus with a wide variety of referring forms. The corpus also include two frequently used indefinite nouns that were not observed in previous corpora, thanks to careful manipulation of object visibility and the partially observable environment with four quadrants. Using this more ecologically valid task environment, we were able to show that the high performance of Pal et al. (2021)'s work may have been artificially inflated by the nature of their task environment.

Before continuing, we would like to state that this work is not a simple replication of Pal et al.'s paper. Our work contributes a novel situated building-construction task that is much improved over the task used by Pal et al. (2021). Moreover, we expand significantly beyond their work, dealing with more difficult issues such as significantly more objects, their visibility and cognitive status, and ambiguity. In the rest of this section we detail and further interrogate why we believe we observed this performance difference.

First, Pal et al. (2021)'s model was trained on a small dataset of referring forms, in which all have similar cognitive status (activated or in focus) due to the small set of 11 objects (compared to 72 objects in this work). Pal et al. (2021)'s task environment also involved very short dialogues, whereas our tower construction task took an average of half an hour to finish. The small dataset used by Pal et al. (2021) may have resulted in over-fitting.

Second, in Pal et al. (2021)'s task, all objects were either labeled or uniquely distinguishable. In contrast, our tower construction task had only a few shapes of blocks used across 72 blocks, significantly increasing *ambiguity*.

Third, indefinite nouns were not considered by Pal et al. (2021), who only used visible objects. As we see from Figure 3 (the left two bars), indefinite nouns were common in our task. In the previous modeling effort, the cognitive status filters (CSFs) assume all object are at least activated and do not attempt to reason about what is "not known of" to the interlocutor, as the assumption was that both interlocutor and autonomous agents such as robots know of the same objects in the scene. Future work should weaken this assumption to model Theory of Mind reasoning.

## 8 Limitations and Future Work

The observed performance gaps motivate possible improvements. During task design, we explicitly intended to collect a multimodel situated dataset, not only language but also gestures, which are particularly informative and suited for situated contexts. We plan to analyze our collected videos and extract gestures, which will likely serve as informative features, as deictic gestures will likely be used on objects' first reference to facilitate use of "this" and "that". In contrast, abstract gestures may be used when objects are in previous quadrants (Stogsdill et al., 2021).

As mentioned in Section 5.2.1, all objects were annotated as at least Familiar to best compare with Pal et al.'s work. Yet, this assumption is clearly violated, especially for objects not yet seen in the task. How to ascribe cognitive status to not-yet-seen objects is a challenging philosophical question, though. We plan to address this in future work.

Finally, to minimize differences between the model trained in this work and that trained by Pal et al. (2021), we excluded a feature that would likely have been informative: referent *visibility*. As discussed, non-visibility was coded as a physical distance of 10; in future work this should be treated as a separate feature.

## 9 Conclusion

We presented a new interaction-based task design to collect a new situated corpus to advance the computational modelling for referring form selection. Specifically, we adapted the modelling technique used by Pal et al. (2021) and reassess its performance on the new corpus. In future work, we plan to annotate the gestures used in our experiment and improve the computational modelling trained on the new multimodal dataset, moving beyond pure replication.

## Supplementary Materials

The data and decision tree code can be found at `https://osf.io/z3ths/`.

## Acknowledgements

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

## References

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

Mira Ariel. 2001. Accessibility theory: An overview. *Text representation: Linguistic and psycholinguistic aspects*, 8(8).

Jennifer E Arnold. 2016. Explicit and emergent mechanisms of information status. *Topics in Cog. Sci.*

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 *ACL*.

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

Frank Eibe, Mark A Hall, and Ian H Witten. 2016. The weka workbench. online appendix for data mining: practical machine learning tools and techniques. In *Morgan Kaufmann*. Elsevier Amsterdam, The Netherlands.

Victor Ferreira, L Slevc, and Erin Rogers. 2005. How do speakers avoid ambiguous linguistic expressions? *Cognition*, 96(3).

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

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*.

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

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

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

Nancy Hedberg. 2013. Applying the givenness hierarchy framework: Methodological issues. *International workshop on information structure of Austronesian languages*.

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*.

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*.

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

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

JR Quinlan. 1987. Simplifying decision trees. *International journal of man-machine studies*, 27(3):221–234.

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

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.

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*.

Ian H Witten and Eibe Frank. 2002. Data mining: practical machine learning tools and techniques with java implementations. *ACM SIGMOD Record*, 31(1):76–77.

## Appendix A: Unpruned Decision Trees

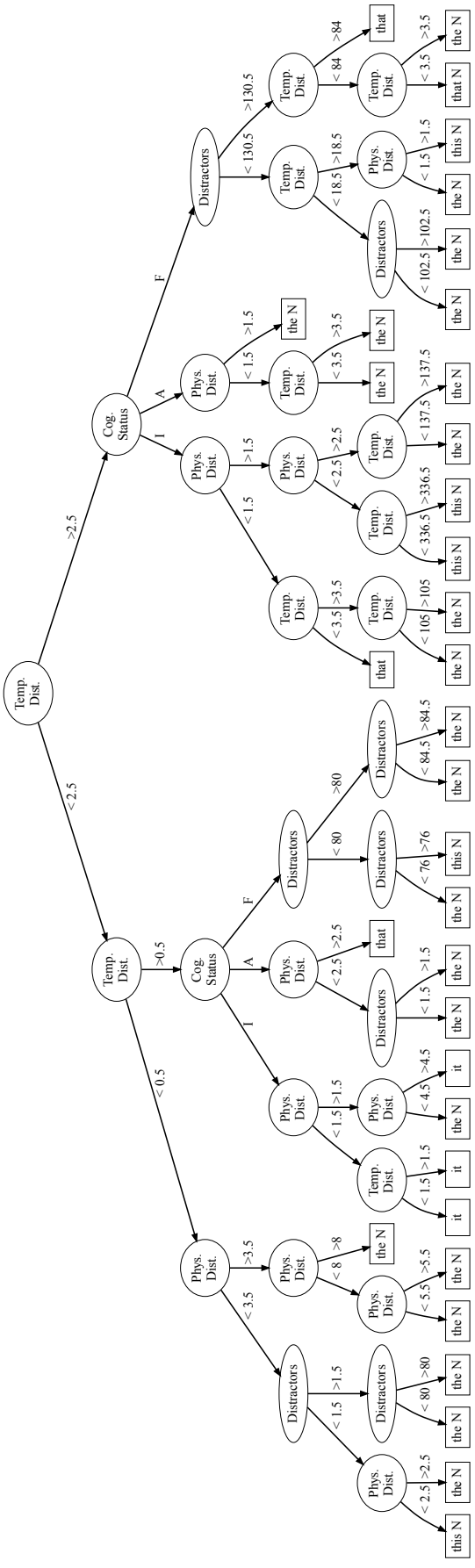


Figure 6: The unpruned decision tree (with six major referring forms).

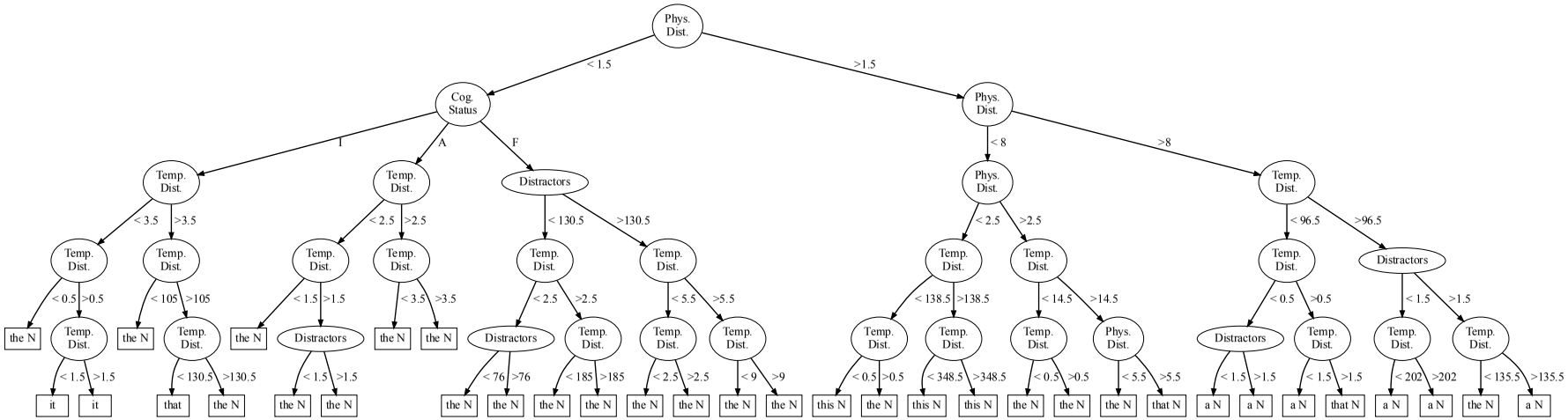


Figure 7: The unpruned decision tree (with six major referring forms and two indefinite referring forms).

## Appendix B: Pruned Decision Trees With Maximum Allowed Depth 6

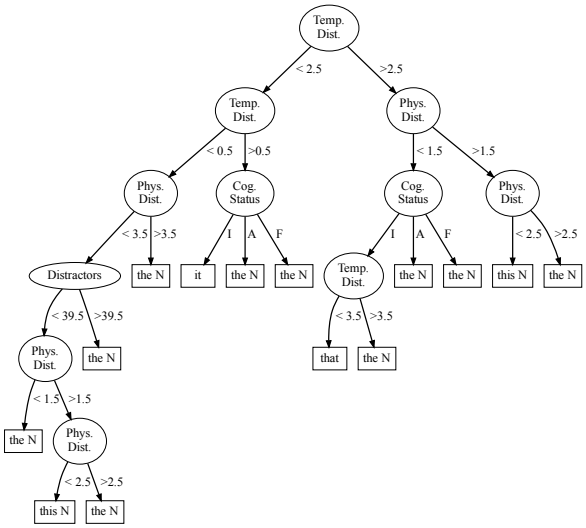


Figure 8: The decision tree visualization for M1 with maximum allowed depth *6* (**six** main referring form categories informed by Givenness Hierarchy).

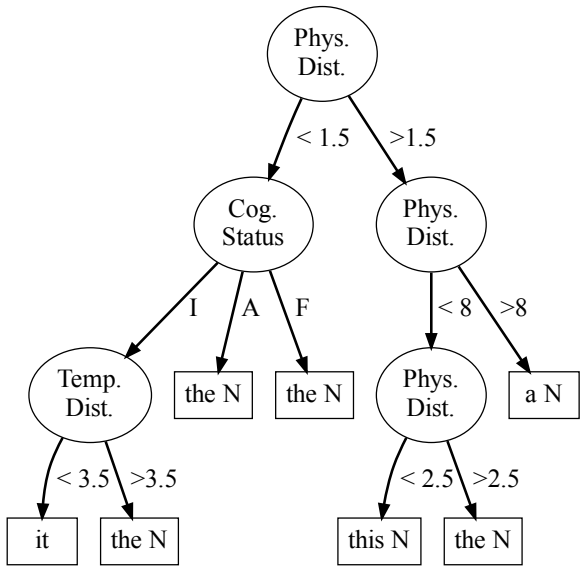


Figure 9: The decision tree visualization for M1 with maximum allowed depth *6* (**eight** main referring form categories). *Note that this is exactly the same as Figure 5.*

## Appendix C: Pruned Decision Trees With Maximum Allowed Depth Unset

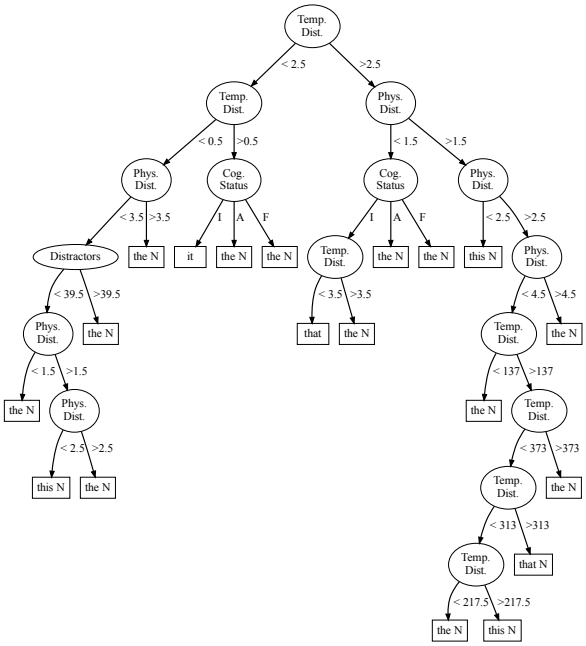


Figure 10: The decision tree visualization for M1 with maximum allowed depth *unset* (**six** main referring form categories informed by Givenness Hierarchy).

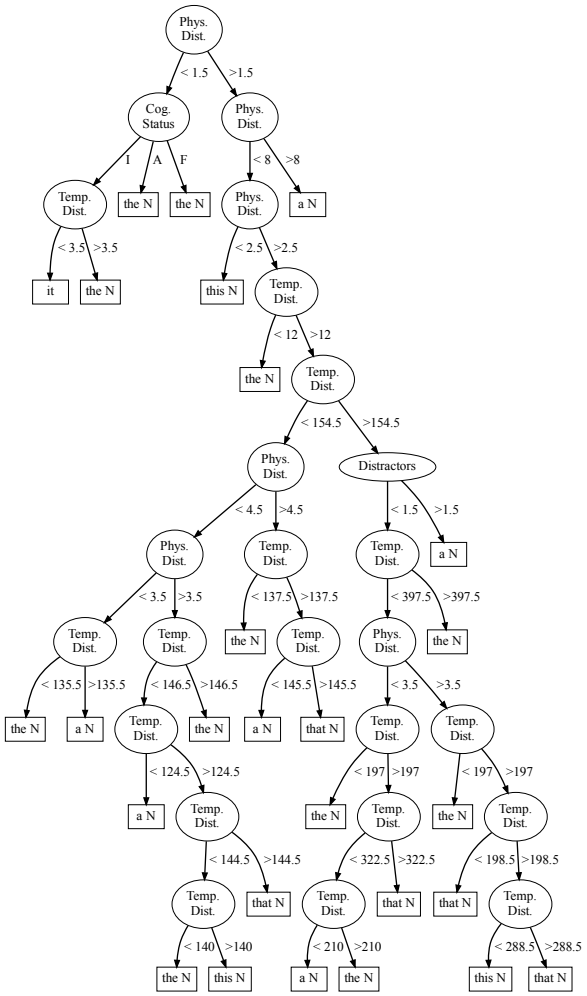


Figure 11: The decision tree visualization for M1 with maximum allowed depth *unset* (**eight** main referring form categories).