Reformulation Strategies of Repeated References in the Context of Robot Perception Errors in Situated Dialogue

Niels Schütte
_Dublin Institute of Technology_, niels.schutte@student.dit.ie

John Kelleher
_Dublin Institute of Technology_, john.d.kelleher@dit.ie

Brian Mac Namee
_University College Dublin_, brian.macnamee@ucd.ie

Follow this and additional works at: [https://arrow.dit.ie/scschcomcon](https://arrow.dit.ie/scschcomcon)

Part of the Artificial Intelligence and Robotics Commons

Recommended Citation
Reformulation Strategies of Repeated References in the Context of Robot Perception Errors in Situated Dialogue

Niels Schütte$^1$ and John Kelleher$^2$ and Brian Mac Namee$^3$

Abstract—We performed an experiment in which human participants interacted through a natural language dialogue interface with a simulated robot to fulfil a series of object manipulation tasks. We introduced errors into the robot’s perception, and observed the resulting problems in the dialogues and their resolutions. We then introduced different methods for the user to request information about the robot’s understanding of the environment. In this work, we describe the effects that the robot’s perceptual errors and the information request options available to the participant had on the reformulation of the referring expressions the participants used when resolving an unsuccessful reference.

I. INTRODUCTION

Robots that interact with a human user through natural language in a spatial environment present a case of situated dialogue. The distinctive characteristic of a situated dialogue is that each participant has a specific perceptual perspective on a shared spatio-temporal context. Consequently, participants in a situated dialogue can not only make references that are evoking (i.e., denoting entities in the interlocutors’ conceptual knowledge) and anaphoric (i.e., denoting entities that have previously been mentioned in the dialogue), but can also make exophoric references (i.e., references denoting objects in the shared context of the dialogue). Therefore, in order to participate in a situated dialogue, robots must be able to perceive their environment and to communicate with the user about what they encounter in the world [18]. If the user’s perception of the world and the robot’s perception diverge (e.g., due to problems in the object recognition software used by the robot [25], or mismatches in the user’s and the robot’s understanding of spatial relations [4], [17]) misunderstanding may arise in the dialogue. In this paper, we investigate the effect of perception-based errors on human-robot dialogue, and how misunderstandings that arise from such errors are resolved. In particular, we analyse how human participants reformulate referring expressions in response to the robot failing to understand an initial reference.

Misunderstandings are frequent in human-human dialogue and humans use different strategies to establish a shared understanding or common ground [8]. The experiment reported in [14] is of relevant to our work because the experiment examined the adjustments made (in terms of gestures accompanying a reference) by a speaker in formulating repeated references in the context of negative feedback from the hearer to an initial reference. The differences between [14] and our work is that we focus on human-robot dialogues and on the adjustments made by speakers to the linguistic content (as distinct from the accompanying gestures) of repeated references. Furthermore, we are particularly interested in situations where the misunderstanding is caused by perceptual differences between the human and the robot. There are empirical studies that explore the effect of mismatched perception on dialogue; e.g., [2], [27], [38]. However, similar to [14], these studies target human-human dialogues.

Previously, the problem of misunderstandings in human-computer dialogue has mostly been addressed from the point of view of misunderstandings arising from difficulties in speech recognition or language understanding (e.g. [1], [26], [29], [41]). There has, however, been some prior research on problems arising from perceptual differences in natural language generation. For example, the problem of producing referring expressions when it is not certain that the other participant shares the same perception and understanding of the scene has been addressed by [15] and [35]. Another example of research investigating language misunderstanding based on perceptual errors is [43] which examines the effect of perceptual deviation on spatial language. However, [43] deals with robot-robot dialogues and the evolution of spatial term semantics in robot populations.

In this paper, we report on an experiment we recently completed and are currently in the process of evaluating: the Toy Block experiment. In the experiment, participants interacted with a simulated robot through a dialogue system and fulfilled a series of tasks by instructing the robot to manipulate a set of objects. The experiment consists of five phases. In the first phase, the robot performs as intended. In the second phase, artificial errors are introduced into the robot’s perception. In the third, fourth and fifth phases, the participants are offered different options to request information about the robot’s perception of the scene. The analysis we present in this paper focuses on: (1) how participants reformulated referring expressions in order to resolve misunderstandings due to perception differences with the robot; and, (2) what influence the information request options available to a participant had on their reformulation of referring expressions.

II. THE TOY BLOCK SYSTEM

The toy black system enables users to interact through a dialog system with a robot that can manipulate objects in a
simulated world. The world contains a set of objects that are intended to represent pieces from a toy building block set. The robot itself is abstract, and not physically represented in the simulation.

Users interact with the system through the user interface shown in Figure 1. This consists of two elements: (1) the simulation window shows a rendering of the simulation world that is updated in real time, and (2) the interaction window provides access to a text based chat interface that the users use to interact with the simulated robot. When the user sends an instruction to the robot, it analyses the instruction and attempts to perform the corresponding actions in the simulation world. If the robot cannot perform the instruction, it replies through the user interface and explains its problem.

The robot's perception is provided by a simulated computer vision system. In general its perception is correct, but sensor errors can be introduced. For example, it can be specified that the robot perceives entire objects or some of their properties incorrectly.

A. Natural Language Processing and Spatial Reasoning

The basic natural language processing pipeline of the toy block system involves: (1) parsing the user input (using the NLTK parser [30]); (2) analysing the resulting parse structure to populate a data frame designed to handle spatial descriptions (similar to the Spatial Description Clause structures in [44]); (3) grounding the referring expressions in the input against the context model of the robot; and (4), if the grounding succeeds executing the action. If the system is not able to perform an action, e.g. because it cannot find a unique referent for a referring expression, it generates an appropriate response. Referring expressions may involve simple attributes such as color and type. Referring expressions can also involve spatial descriptions such as relational referring expressions that describe the target object in relation to one or more landmark objects (e.g., “Pick up the red ball that is between the green box and the yellow box”), and directional descriptions that describe the general position of an object in the scene without reference to a landmark (e.g., “Pick up the ball on the right”). In the rest of this section we will focus on describing the computational models the toy block system uses to ground the semantics of spatial terms against the context model.

Psychological studies have identified a large number of phenomena that affect the semantics of spatial descriptions, including: the extent and shape of the spatial template associated with the spatial term [28]; the impact of the functional relationship between the objects and the goals of the agents [11]; the impact of other objects in the scene [9]; attentional factors [5], [33]; and perceptual phenomena, such as object occlusion [21], the speaker’s perspective on the landmark [20], [36], and the orientation of the objects [6]. This list can be extended further for composite directional spatial descriptions (e.g. “on the right of”, “at the front of”) where frame of reference and frame of reference ambiguity must be considered [7], [16], [37], [40], and the contribution of the topological term (e.g., at, on, in) to the overall semantics of the description is also a factor [19].

Given this array of factors it is not surprising that there is a spectrum of approaches to creating computational models of spatial language semantics, each motived by different considerations. For example, integrating world-knowledge [32] and/or linguistic ontological knowledge [3]; integrating spatial semantics into a compositional/attentional accounts of reference [23], [24], [31]; learning spatial semantics directly from sensor data using machine learning techniques [12], [34]; modelling the functional aspects of spatial semantics in terms of predicting the dynamics of objects in the scene [10], [42]; capturing the vagueness and gradation of spatial semantics [17], [22], [43]; and leveraging analogical reasoning mechanisms to enable agents to apply spatial semantics to new environments [13].

Compared to many of these previous models the approach to spatial semantics we took in this work is relatively simple. There are a number of simplifying factors within the design of the experiment that allowed this. For example, the objects in the world were simple shapes and all were of a similar size. As a result, the objects had no functional relationships between each other, nor did they have intrinsic frames of references associated with them. Also, all the objects appeared on a chequerboard patterned background, and the user’s view of the world was fixed.

In order to interpret spatial descriptions that located objects relative to the scene frame (e.g., “the ball on the right”) we simply partitioned the chequerboard up into a grid of 3 rows and 3 columns, and associated different spatial words with each of the regions. Figure 2 illustrates how the board the objects appeared on was split into regions. The user could refer to the area at the back of the board using terms such as back, top, or far, e.g. “Pick up the red ball at the back”). The regions denoted by other terms (such as middle, left, or near) are also shown. If the user input contained a description that combined spatial terms then the intersection of the regions was used. The image at the right in Figure 2 illustrates some of the possible labels for region intersections.

The system could also handle relative descriptions. If the description involved a projective spatial term (e.g. to right of X, to left of X, behind X, or in front of X) the system considered the spatial description to cover a region covering four times the bounding box of the landmark object X along the appropriate axis (see Figure 3). The use of 4 bounding boxes to define the region was chosen based on trial-and-error, and worked well for our experimental setup. The
region described by near X was also defined in terms of the bounding box of landmark object X—in this instance 2 bounding boxes in any direction (see Figure 4). Finally, the region described by between X and Y was taken to encompass the region along the axis going from one landmark’s centroid to the second landmark’s centroid (see Figure 5).

### III. THE TOY BLOCK EXPERIMENT

In each run of the experiment, the participants were presented with a set of 20 scenes. The scenes were presented in random order except for two simple introductory scenes which were intended as tutorial scenes, and that were always presented as the first and second scene. Each scene consisted of a start scene and a target scene. The start scene determined how the objects in the simulation world were arranged at the beginning of the scene. The target scene was presented to the participants as an image in the interaction window. The participants’ task was to interact with the robot to recreate the target scene in the simulation world. After a participant had successfully recreated the target scene, the system automatically advanced to the next scene.

All utterances by the participant and the system are transcribed and annotated with their semantic interpretation. The system also records task success and dialogue cost measures as described in [45]. Previously, we have reported the dialogue cost analysis of the Toy Block experiment pilot study in [39]. In this paper, however, we will focus on the analysis of the reference reformulation strategies adopted by participants in the experiment in situations where an initial reference failed.

#### A. The scenes

In total there were 20 scenes. The scenes were designed to encourage participants to use specific strategies and expressions to complete them. For example, we introduced distractor objects to encourage participants to include specific attributes or to use specific landmark-based expressions.
For 14 of the 20 scenes we designed perception errors that participants were likely to encounter when they attempted to solve the scenes. There were three types of errors: the **missing object** error, where the system failed to detect an object; the **wrong colour** error, where the system incorrectly recognized the colour of an object; and the **wrong type** error, where the system misclassified the type of object. Errors could either affect objects that participants were required to move to complete a scene, or they could affect objects that participants were likely to use as landmarks in relational referring expressions.

Figure 6 shows the start scene and the target scene of a typical scene. For this scene an error was introduced into the robot’s perception. The robot perceived the green box in the bottom left of the scene (in Figure 6a) as a green ball. Figure 6c shows the scene as it was perceived by the robot.

B. Experiment phases

The experiment consisted of five phases:

1) **No Error Phase**: The robot performed the instructions it was given to its best capabilities. The robot’s perception of the world were error-free. This phase represents a baseline condition for the performance of the system.

2) **Error Phase**: Errors were introduced into the robot’s perception. This purpose of this phase was to determine the effect of the perception errors on the user experience and the system performance.

3) **Description Phase**: Errors were introduced into the robot’s perception. Participants were able to ask the system to generate a description of the scene as it was perceived by the robot. The following is an example of a description the system produced for the scene in Figure 6:

   “There is a blue box on the top left. There is a red ball on the left. There is a green ball on the bottom left. There is a place named place 1 on the top right. There is a place named place 2 on the right. There is a place named place 3 on the bottom right.”

Note, the scene description reflects the fact that the robot perceives the green box as a green ball. This phase represents a uni-directional language-based information option.

4) **Markup Phase**: Errors were introduced into the robot’s perception. In this phase the participants were able to task the system to mark up the robot’s understanding of the scene in the simulation window. Figure 7 contains an example of the simulation window with active markup. This phase represents a uni-directional visual information option.

5) **Querying Phase**: Errors were introduced into the robot’s perception. In this phase, the participants were able to ask the system simple questions about whether or not the robot perceived an object of a given description. Figure 8 contains example questions and responses for the scene presented in Figure 6. This phase represents a dialogue based approach to requesting information.

C. Instructions

The participants were provided with an instruction manual that described the experiment, introduced the user interface and provided example interactions. In the phases in which errors were introduced, it mentioned the possibility that the robot’s perception might be problematic. After reading the instructions, the participants were shown a video recording of some example interactions with the system. This was done to prime the participants towards using language and concepts that were covered by the system. No time limit was set for experiment. There was no penalty or reward associated with the participants’ performance in the experiments.

IV. EXPERIMENTAL RESULTS

Table I contains an overview of the data recorded in the experiment, including the number of participants that took part in each phase. The key metric for these experiments is the **abandon rate**, which measures the frequency with which participants could not successfully complete a scene and so had to abandon it. We found that the presence of perception errors increases the abandon rate (and so decreases the task success). On the other hand, we found that introducing information request options increases the
task success. Overall, we found that the querying option and the markup option are similarly effective, and more effective than the description option. However, all options provide a clear improvement over the Error Phase.

<table>
<thead>
<tr>
<th>Phase</th>
<th>Number of participants</th>
<th>Recorded scenes</th>
<th>Abandon rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Error Phase</td>
<td>10</td>
<td>200</td>
<td>5.05%</td>
</tr>
<tr>
<td>Error Phase</td>
<td>17</td>
<td>338</td>
<td>19.16%</td>
</tr>
<tr>
<td>Description Phase</td>
<td>11</td>
<td>220</td>
<td>11.74%</td>
</tr>
<tr>
<td>Markup Phase</td>
<td>11</td>
<td>220</td>
<td>9.05%</td>
</tr>
<tr>
<td>Querying Phase</td>
<td>11</td>
<td>220</td>
<td>9.13%</td>
</tr>
</tbody>
</table>

TABLE I: Overview of the recorded data.

V. PROBLEM RESOLUTION SEQUENCES

We collected all instances where a participant instructed the robot to pick up an object that was affected by a perception error, causing the robot to encounter a perception problem (i.e., it was either not able to resolve the referring expression used in the instruction, or was not able to resolve it unambiguously). We then collected all subsequent actions until the participants either successfully managed to get the robot to pick up the intended object, or until they abandoned the current scene. We call the problematic reference that triggered the resolution sequence the **initial reference**. We call the successful reference that ended the resolution sequence the **final reference**. We call the sequence of actions between them a **resolution sequence**. Figure 9 contains two example resolution sequences that were extracted from the data. In total we extracted 247 resolution sequences that were successfully concluded.

We analyse the sequences to investigate how the participants modified their initially unsuccessful referring expressions to achieve a successful reference, and what influence the information request options had on this. To perform this analysis we first selected all resolution sequences that resulted in a successful resolution. We then determined for each referring expression that was contained in the initial reference of the sequence and the final reference of the sequence whether it contained the following attributes:

- **Type**: The expression contained a specific type attribute such as “ball” or “cube”. If the expression contained a general type such as “object” or “thing”, it was not counted.
- **Colour**: The expression contained a colour attribute such as “green” or “red”.
- **Landmark reference**: The expression contained a reference to a landmark object (such as in “the ball near the green box” or “the green box between the red ball and the yellow ball”).
- **Directional description**: The expression contained a directional expression that described the location of the target object in the world without reference to a landmark object (e.g., “the box on the left,” “the green box in the centre”).

We split the set of successful resolution sequences into four subsets depending on whether or not information was requested during the resolution sequence. The **uninformed set** contains sequences during which no information was requested (Sequence a in Figure 9 is an example of such a sequence). The **description set** contains sequences during which a description was requested (Sequence b from Figure 9 is an example), the **markup set** contains sequences during which markup was requested, and the **querying set** contains sequences during which a query was made.

Table II contains an overview of the attributes included in the initial references. Table III contains an overview of the attributes in the final references. For easier comparison, Table IV contains the differences between the proportions in the initial and the final references.

VI. DISCUSSION AND ANALYSIS

We find that almost all initial referring expressions contain a specific type attribute. Most of the initial referring expressions also contain a colour attribute. We do note that the proportion of expressions that contained a type attribute and
We believe that this may be explained by the type of information as strong in the description and the markup category. We uninformed and the querying category, while it was not and colour attribute was most strongly pronounced in the robust against perception errors.

We therefore believe that participants removed attributes that, in their experience, were potentially unreliable, and, to compensate for the loss of descriptive potential, instead substituted them with directional descriptions, which were to align their model of the scene to the robot’s flawed model of the world. This means they did not necessarily have to completely abandon unreliable attributes, but were able to use attribute values that were valid in the robot’s understanding of the world.

In the uninformed condition and the querying condition, the system did not provide an explicit description of how the robot perceived the world. In the querying condition, the participants could ask the system whether or not it perceived an object of a given description. In the uninformed condition, they did not request any information. Instead they reached a successful reference by trial-and-error and therefore were less likely to align their model of the scene to the robot’s model, and more likely to use more general terms and directional descriptions.

Another interesting observation is that the proportion of references that included a directional description is highest for the description condition. This may be related to the fact that the descriptions themselves also contained directional descriptions. It is therefore possible that the participants aligned their expressions to the descriptions provided by the system.

VII. Summary

We performed an experiment in which we artificially induced problems in a dialogue based human-robot interaction and observed how the problems were resolved. We investigated the choice of attributes in referring expressions before and after problems. We found that participants tended to include different attributes in the final expressions to their initial expressions, and that the choice of attributes is also
related to the type of information available about the robot’s perception. If information about the robot’s understanding of the scene is directly available, they tend to align their referring expressions to the robot’s understanding. Sequence a in Figure 9 is an example of this effect. The participant discovers through the description that the robot sees a yellow box as a yellow ball, and repeats the initial instruction modified to suit this understanding.

If this type of information is not available, participants tend to use strategies where they combine expressions that avoid unreliable attributes with robust directional descriptions. Sequence b in Figure 9 shows a case where the participant removed a landmark based description and replaced it with a directional description.

In conclusion, perception based errors may occur in human-robot dialogues. One obvious way to address this problem is to improve robot perception. Our work, however, indicates that another useful strategy to address this problem is to provide the human user access to the robot’s perceptual model of the world. As our results show, humans find it easy to align with the robot’s perception or to adjust their references so that the robot can understand them. In particular, in the uninformed condition humans adjusted their referring expressions to include descriptions that, in their experience, the robot could reliably interpret (in this experiment directional descriptions), and dropped the attributes from their descriptions that they believed the robot was not able to reliably handle (here, type and colour attributes). An interesting implication of this is that there is potential for robot systems to use the adjustments that humans make in their references to provide the robot with information regarding what errors in perception it is making. This information may be used by the robot to trigger a self-repair mechanism of its perception. We will explore this in future work. In other future work we are currently writing up the results of the experiments and plan to publish about them in more detail. In particular, we are going to compare and evaluate the performance of the different information request options in more detail and investigate their effect on the referring expressions and the structure of the resolution sequences.

**REFERENCES**


