2017-01-04

Robot Perception Errors and Human Resolution Strategies in Situated Human-Robot Dialogue

Niels Schütte
Technological University Dublin, niels.schuette@gmail.com

Brian Mac Namee
University College Dublin, Ireland, brian.macnamee@ucd.ie

John D. Kelleher
Technological University Dublin, john.d.kelleher@tudublin.ie

Recommended Citation

Schuette, N., Mac Namee, B. & Kelleher, J. (2017) Robot Perception Errors and Human Resolution Strategies in Situated Human-Robot Dialogue, *Advanced Robotics,*January 4th, DOI : 10.1080/01691864.2016.1268973

This is an Accepted Manuscript of an article published by Taylor & Francis in

Advanced Robotics on the 4th January 2017

Available online: http://www.tandfonline.com/doi/full/10.1080/01691864.2016.1268973 .

Robot Perception Errors and Human Resolution Strategies in Situated Human-Robot Dialogue

Niels Schütte[a*] and John Kelleher[b] and Brian Mac Namee[c]

[a,b] *Dublin Institute of Technology, Kevin Street, Dublin 2, Ireland;* [c] *University College Dublin, Dublin 4, Ireland*

(*Received 00 Month 20XX; accepted 00 Month 20XX*)

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. We quantify the impact of perception errors on the dialogues, and investigate resolution attempts by users at a structural level and at the level of referring expressions.

Keywords: Dialogue Systems; Human-Robot Interaction; Perception Errors; Dialogue

1. 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]), misunderstandings 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.

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 relevance 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

*Corresponding author. Email: niels.schuette@gmail.com, [b] Email: john.d.kelleher@dit.ie [c]Email: brian.macnamee@ucd.ie

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: the **Toy Block** experiment. In the Toy Block 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 at its normal capacity. 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. In this paper we analyse two major aspects. In the first step, we investigate the effect the introduction of perception errors had on the dialogue, and the effectiveness of different information request options in resolving problems in the dialogues. In the second step, we identify instances in the dialogues, where perception errors lead to problems in the tasks, and investigate what strategies the participants used to resolve the arising problems. We focus on two sub-aspects of this problem:

- At a dialogue structure level we investigate what sequences of actions the participants performed in order to resolve the problems.
- At a referring expression level we investigate how participants modified the attributes in referring expressions after a reference had failed, and what influence the different information request options had on this choice.

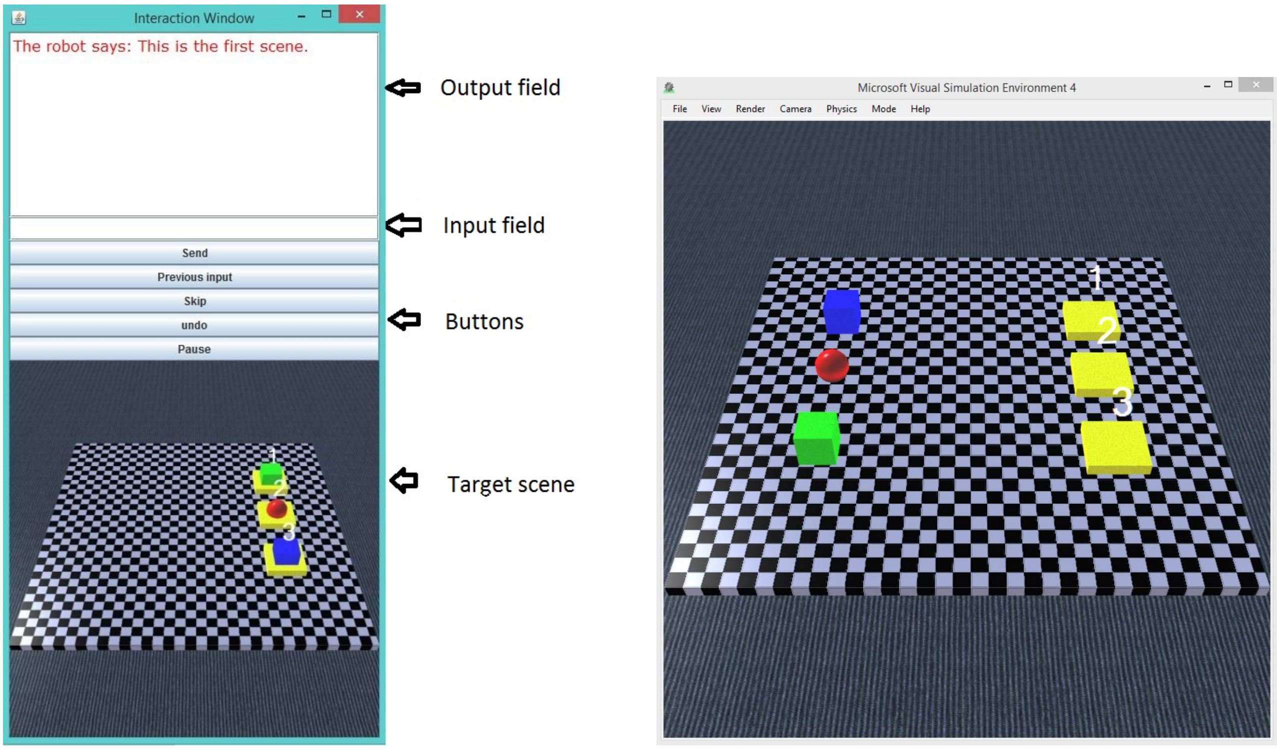
This paper is structured as follows: In Section 2 we describe the experiment system that was used to perform the experiment. In Section 3 we describe the setup of the experiment and the different tasks and perception problems the participants were faced with. In Section 4 we provide an overview of the data recorded during the experiment and the effect of perception errors. In Section 5 we describe the process of identifying perception based problems and the following resolution attempts in the recorded dialogues and report on the collected data. In Section 6 we describe structures identified in the recorded sequences, and in Section 7 we discuss the choice of attributes in referring expressions in unsuccessful references and in problem resolution attempts. In Section 8 we conclude the paper and discuss possible future work.

2. The Toy Block System

The Toy Block system enables users to interact through a dialogue system with a robot that can manipulate objects in a simulated world, similar to the *SHRDLU* system [46]. 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. The interface consists of two elements:

- The **simulation window** shows a real-time rendering of the simulation world.
- The **interaction window** provides access to a text based chat interface that the users use to interact with the simulated robot.



(a) The interaction window. (b) The simulation window.

Figure 1. The user interface.

When the user sends an instruction to the robot, it analyses the instruction and attempts to perform the requested 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.

2.1. *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 *colour* 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

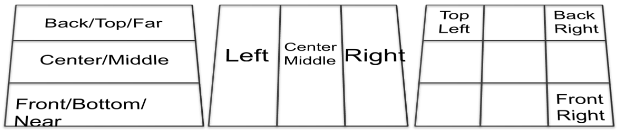


Figure 2. A partitioning of the scene for spatial references relative to the scene frame.

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, 39], 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).

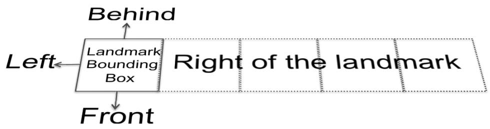


Figure 3. The definition of the spatial template for directional spatial terms: *right*, *left*, *front*, *behind*.

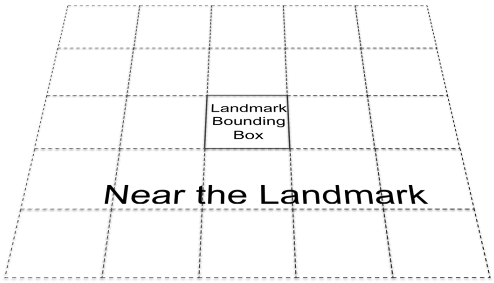


Figure 4. The spatial template for *near*.

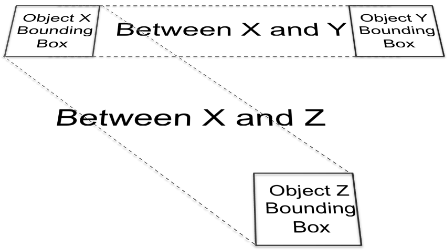


Figure 5. The spatial template for *between*.

3. The Toy Block experiment

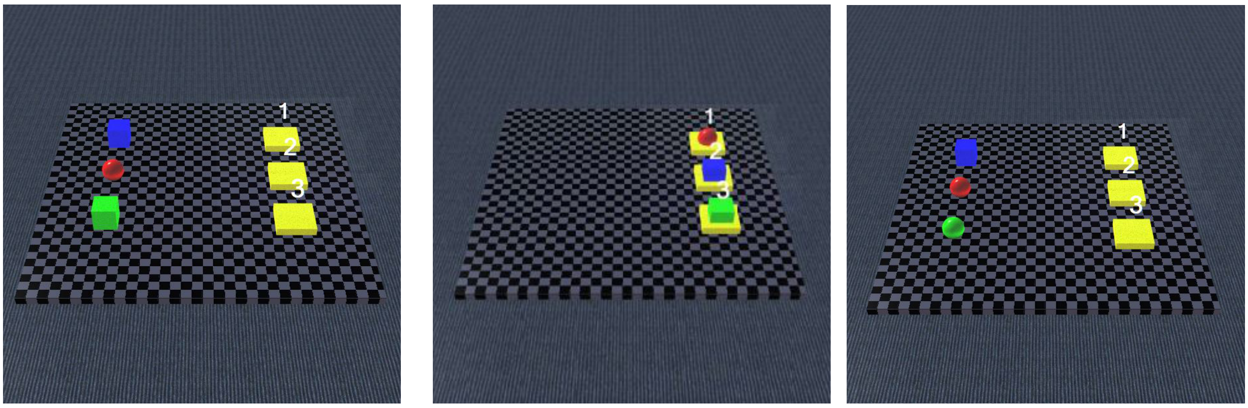
In each run of the experiment, the participants were presented with a set of 20 scenes. The scenes were presented in random order except for two simple introductory scenes which were intended as tutorial scenes, and that were always presented as the first and second scene. Each scene consisted of a **start 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].

3.1. *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, in order to encourage participants to use specific attributes or to use specific landmark-based expressions, we introduced distractor objects whose presence made referring expressions without these attributes ambiguous.

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



(a) The start scene. (b) The target scene. (c) The start scene as perceived by the robot.

Figure 6. A scene that is affected by a perception error.

incorrectly recognized the colour of an object; and the **wrong type** error, where the system misclassified the type of object. Errors could either directly 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.

The start and the target scene of a typical scene are shown in Figure 6(a) and Figure 6(b). For this scene an error was introduced into the robot's perception. The robot perceived the object that is shown as a green box in the bottom left of the scene (in Figure6(a)) as a green ball. Figure 6(c) shows the start scene as it appeared to the robot.

3.2. *Experiment phases*

The experiment consisted of five phases:

(1) **The 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) **The Error Phase**: Errors were introduced into the robot's perception. The purpose of this phase was to determine the effect of the perception errors on interactions.
(3) **The 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 that the scene description reflects the fact that the robot perceives the green box as a green ball due to a perception error. This phase represents a uni-directional language-based information option.
(4) **The Markup Phase**: Errors were introduced into the robot's perception. In this phase the participants were able to ask 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

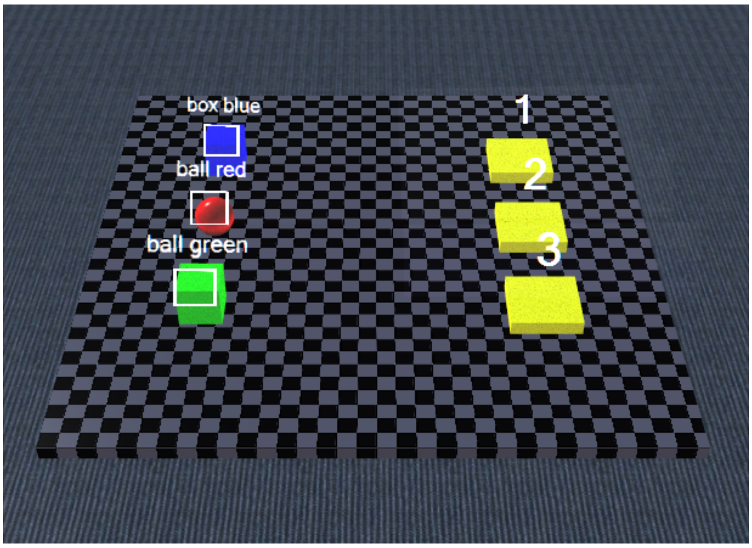


Figure 7. A scene with activated markup.

U:	Do you see a red ball?
S:	Yes, I see a red ball.
U:	Do you see two boxes?
S:	I see 1 box.
U:	Do you see a green box on the bottom left?
S:	Sorry, I don't see any green boxes that fit your description.
U:	Do you see a blue box?
S:	Yes, I see a blue box.

Figure 8. Some example questions for the Querying Phase.

markup. Again, the markup contains the information that the robot perceives the green box as a green ball. This phase represents a uni-directional visual information option.

(5) **The 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.

3.3. *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.

Table 1. Overview of the recorded data.

Phase	Number of participants	Scenes attempted	Total length
No Error Phase	10	200	04:16:08
Error Phase	17	338	09:03:36
Description Phase	11	220	08:08:03
Markup Phase	11	220	06:13:01
Querying Phase	11	220	06:12:38

Phase	Abandon rate	Reference problem rate	Average number of actions	SD	Average completion time (s)	SD
No Error Phase	5.05%	7.82%	5.18	2.83	108.64	94.12
Error Phase	19.16%	29.23%	7.57	5.74	140.28	146.89
Description Phase	12.08%	18.86%	7.3	5.08	201.47	256.41
Markup Phase	9.13%	16.37%	6.69	4.33	149.39	164.83
Querying Phase	9.72%	15.55%	6.2	3.69	150.14	133.1

Table 2. Measure values for the different phases.

4. Experimental Results

Table 1 contains an overview of the data recorded in the experiment, including the number of participants that took part in each phase and the total length of the dialogues recorded. Table 2 contains an overview of the dialogue measures that were recorded in each phase. The **abandon rate** is the most general indicator of problems in the interactions. It reports which proportion of the attempted the scenes the participants were not able to finish in each phase. We observe that all phases in which perception errors are present exhibit a higher abandon rate than the baseline condition. There is a marked increase between the No Error Phase and the Error Phase, which indicates that the presence of perception errors made the task more difficult. On the other hand, the phases in which the participants could request information about the robot's perception of the world (the Description Phase, the Markup Phase and the Querying Phase) show a clear improvement over the Error Phase. They, however, do not reduce the task abandon rate back to the level of the baseline condition. The **reference problem rate** denotes the proportion of the instructions that the participants made in each phase that the robot could not perform because it encountered a reference resolution problem (e.g. if the participant gave the instruction "Pick up the red ball", but the robot perceived two red balls). The reference problem rate in the No Error Phase represents a baseline for reference problems and is due to the general reference resolution capabilities of the system. The Error Phase shows a much higher reference problem rate than the No Error Phase. This indicates that the introduction of perception errors lead to an increase of reference resolution problems in the dialogues, which consequently lead to problems in the task as indicated by the scene abandon rate. We can observe that in the following phases in which the participants could request information about the robot's perception of the world the reference problem rate is much lower than in the Error Phase. This indicates that participants were able to use the information they received through the information requests to minimize reference problems.

The average number of actions denotes the number of actions a participant needed on average to complete a scene[1], while the average completion time denotes the average number of seconds participants needed to complete a scene. These figure indicate how much effort and time participants had to expend to complete a scene. The average number of actions displays a pattern similar to the reference problem rate. The lowest value is recorded for the No Error Phase. The Error Phase shows a strong increase while the remaining phases show values that lie in between these extremes.

The values for the average completion time however, are not consistent with this pattern. The highest value is recorded for the Description Phase, while the values for the Markup Phase and the

[1]We count every instruction the participant sent to the system as an action

Phase	Success Prop.	Success Count	Abandon Prop.	Abandon Count	Other Prop.	Other Count	Total Count
Error Phase	64.1%	82	29.7%	38	6.3%	8	128
Description Phase	80.5%	70	13.8%	12	5.7%	5	87
Markup Phase	69%	40	15.5%	9	15.5%	9	58
Querying Phase	74.3%	55	20.3%	15	5.4%	4	74

Table 3. The number of resolution sequences found in each phase and the percentage of their outcomes.

Querying Phase are similar to the one for the Error Phase. We believe that this may be explained by the fact that descriptions could get lengthy for more complex scenes and took more time for the system to present and for the participants to take in and process.

Since we have established that perception errors lead to problems in the dialogues, and that participants were able to resolve these problems through interaction with the system, we investigate in the following sections in more detail how they these problem resolutions were performed. In particular, we focus on *problem resolution sequences* as the sequences of actions the participants performed between the instance where a perception error occurred in the dialogue, and the instance where the problem was resolved.

5. Problem Resolution Sequences

We detected 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. the robot 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 in the action that triggered the resolution sequence the **initial reference**. We call the successful reference in the action that ended the resolution sequence the **final reference**. We call the sequence of actions from the initial to the final reference the **resolution sequence**. Figure 9 and Figure 10 contain two example resolution sequences that were extracted from the data.

Table 3 contains an overview of the number of resolution sequences that were recorded in each phase. *Success* denotes sequences that ended when the participant managed to successfully get the robot to pick up the object that they had intended it to pick up in the initial unsuccessful reference. *Abandon* denotes sequences that ended when the participant abandoned the scene when they felt they would not be able to successfully complete it. *Other* refers to sequences which ended when the participant created a scene that technically satisfied the completion conditions for the target scenes, but that avoided resolving the reference based problem.[2]

We analyse the resolution sequences under two aspects. In the first step we investigate the sequences in terms of the dialogue structures that are formed by the sequences of actions in order to discover how the participants typically approach the problem (Section 6). In the second step we investigate what attributes the participants used in the unsuccessful references and what attributes they used in successful references that concluded the resolution sequences in order to investigate the choice of attributes (Section 7). This is done in order to see how the participants modified their

[2]In a few scenes the participants could, due to oversights in the completion conditions, complete scenes using objects that were not affected by perception errors instead of the objects they were intended to use.

U: Pick up the red ball near the yellow ball
S: Sorry, I can't find any red balls that fit your description.
U: *(requests a description)*
S: There is a red ball to the left of a blue box on the bottom right. There is a red ball to the left of a yellow box on the top right. There is a place named place 1 on the left centre.
U: pick up the red ball near the yellow box.
S: Okay, I will pick it up.

Figure 9. A resolution sequence that involves a description request.

U: pick up the red ball near the yellow ball
S: Sorry, I can't find any red balls that fit your description.
U: pick up the right most red ball
S: Okay, I will pick it up

Figure 10. A resolution sequence that involves a no information request.

references, and to determine what effect the available information has on this choice.

6. Resolution Sequence Structures

We analysed the resolution sequences in terms of the actions the participants performed between encountering the problem and the end of the resolution sequence. In order to be able to recognize common structures between the sequences, we investigated them at a dialogue act level [40]. We annotated each action with one of the following dialogue act tags:

- **pickup:** The participant successfully instructs the system to pick up an object.
- **move:** The participant instructs the system to move an object the robot is holding to a given location.
- **put:** The participant instructs the system to put down an object the robot is holding.
- **description:** The participant requests a description of the scene (this was only possible in the Description Phase).
- **markupon:** The participant turns the markup information on (this was only possible in the Markup Phase).
- **markupoff:** The participant turns the markup off (this was also only possible in the Markup Phase).
- **query:** The participant makes a query (this was only possibly in the Querying Phase).
- **pause:** The user clicks on the pause button.

We furthermore introduced the following tags to represent the beginning and conclusions of dialogue act sequences.

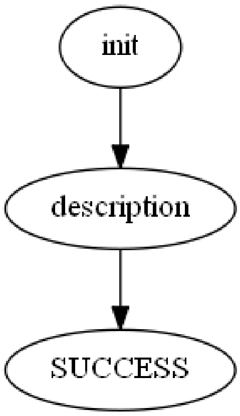


Figure 11. The dialogue act sequence for the sequence from Figure 9.

- **init:** This refers to the action that initializes the resolution sequence (i.e. an attempt to pick up an object that is affected by a perception error and that fails due to reference resolution problem).
- **success:** This denotes a pickup instruction that successfully completes the resolution sequence.
- **abandon:** The participant abandones the scene.
- **other:** The participant finishes the scene with an invalid solution (as discussed earlier).

In order to capture whether or not the robot could perform a requested instruction, we appended the following tags to each *move*, *pickup* or *put* instruction:

- **ok:** The system successfully performed the instruction.
- **not_ok:** The system was not able to successfully complete the instruction (e.g. because it was not able to resolve one of the references in the instruction).

For example *pickup_ok* represents a *pickup* instruction the system was able to perform, while *pickup_not_ok* represents a *pickup* instruction the system was not able to perform.

We then created for each resolution sequence a sequence of dialogue acts. For example, Figure 11 show the dialogue act sequence that was created for the resolution sequence shown in Figure 9. We selected the five most frequent resolution sequences in each phase and unified them into one graph to create a summary of the most frequent resolution strategies. Figure 12 to Figure 15 show the resulting graphs. Each graph can be read from the top to the bottom. The graphs begin at the *init* node which represents the initial unsuccessful instruction. Sequences of actions then form paths towards terminating nodes. Each arc in the paths is annotated with the number of times this particular arc was observed in the data set. Additionally, the thickness of arcs represents the relative frequency of each arc.

Figure 12 contains an overview of the five most frequent sequences in the **Error Phase**. In this phase the participants did not have the opportunity to request information about the robot's perception of the scene. The paths involving unsuccessful instructions (represented by the *pickup_not_ok* nodes) can therefore be interpreted as trial-and-error attempts by the participants to find a working description for the intended object. The rightmost path is interesting in that it represents sequences in which the participants picked up an object other than the one they had intended originally, put it down somewhere (either at a target location to fulfil some other part of the task, or they simply released it in the same place without moving it) and then resolved the problem eventually. In personal discussions with the participants after the experiment, some of them reported picking up objects that they did not need to move to complete the task in order to test out how the robot would interpret a given expression. This strategy therefore represents an

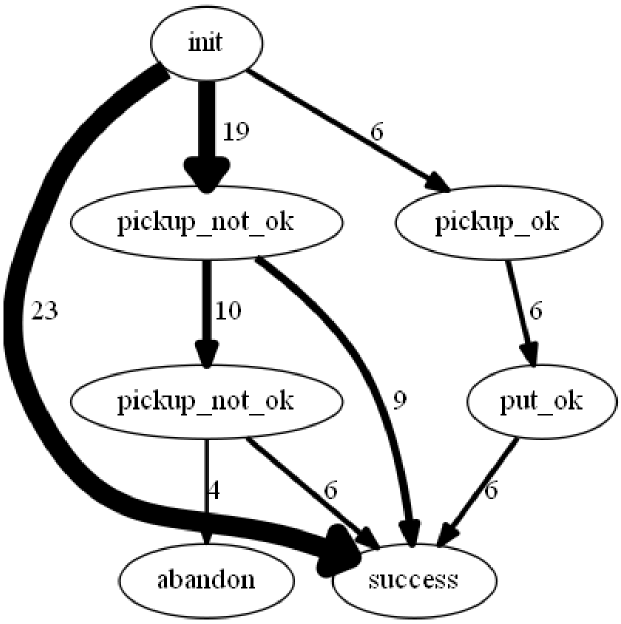


Figure 12. The sequence graph for the five most frequent sequences in the Error Phase.

approach towards querying the robot's understanding of the scene when no explicit information request options are available.

Figure 13 contains an overview of the five most frequent sequences in the **Description Phase**. We observe that the most prominent path leads from the *init* node to the *SUCCESS* node through a *description* node. This indicates that participants often resolved the problem by requesting a description, and were then able to pick up the intended object using the information obtained through the description. The other paths represent cases where the participants attempted to solve the problem by trial-and-error without requesting information.

Figure 14 contains an overview of the five most frequent sequences in the **Markup Phase**. The structures are similar to the graph for the Description Phase in that the most frequent path is the one that contains an information request in the form of a query, while the other paths represent trial-and-error attempts.

Figure 15 contains an overview of the five most frequent sequences in the **Querying Phase**. It is similar to the graphs from the Description Phase and the Markup Phase in that the most frequent path contains a query. It is noticeable however that, apart from the trial-and-error path, another path exists that represents series of multiple queries. This indicates that the participants frequently had to ask multiple queries to incrementally accumulate information before they were able to successfully resolve the problem.

6.1. *Discussion and Analysis*

Overall we find that in all phases the most frequent paths involve obtaining information. In the graphs for the Description Phase, the Markup Phase and the Querying Phase, information is chiefly obtained through explicit information requests (in the Querying Phase also through series' of queries). In the Error Phase information could not be requested explicitly. Participants therefore had to fall back on a trial-and-error strategy or the indirect pick-up based querying.

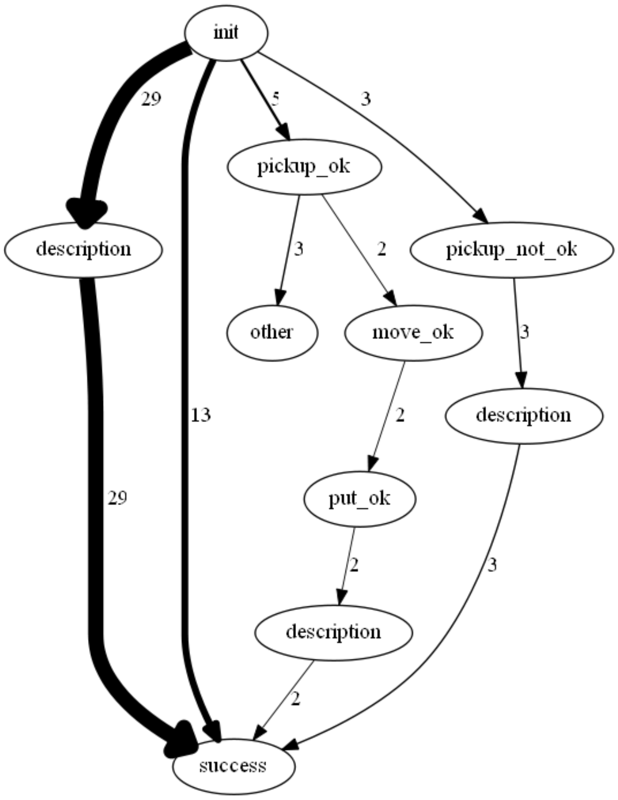


Figure 13. The sequence graph for the five most frequent sequences in the Description Phase.

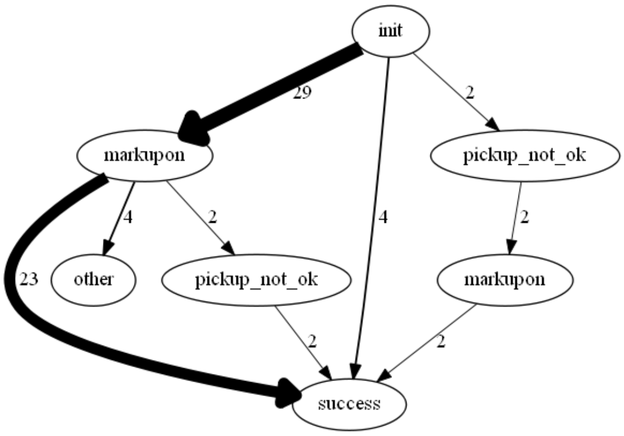


Figure 14. The sequence graph for the five most frequent sequences in the Markup Phase.

7. Attributes in Resolution Sequences

Apart from the sequences of actions the participants performed, we also investigated what attributes the participants used in their attempts to resolve the problems in the dialogues. We determined for each successful resolution sequence the set of attributes the participants used in the referring expression in their initial (unsuccessful) reference and in the final (successful) reference. For each expression we determined 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.

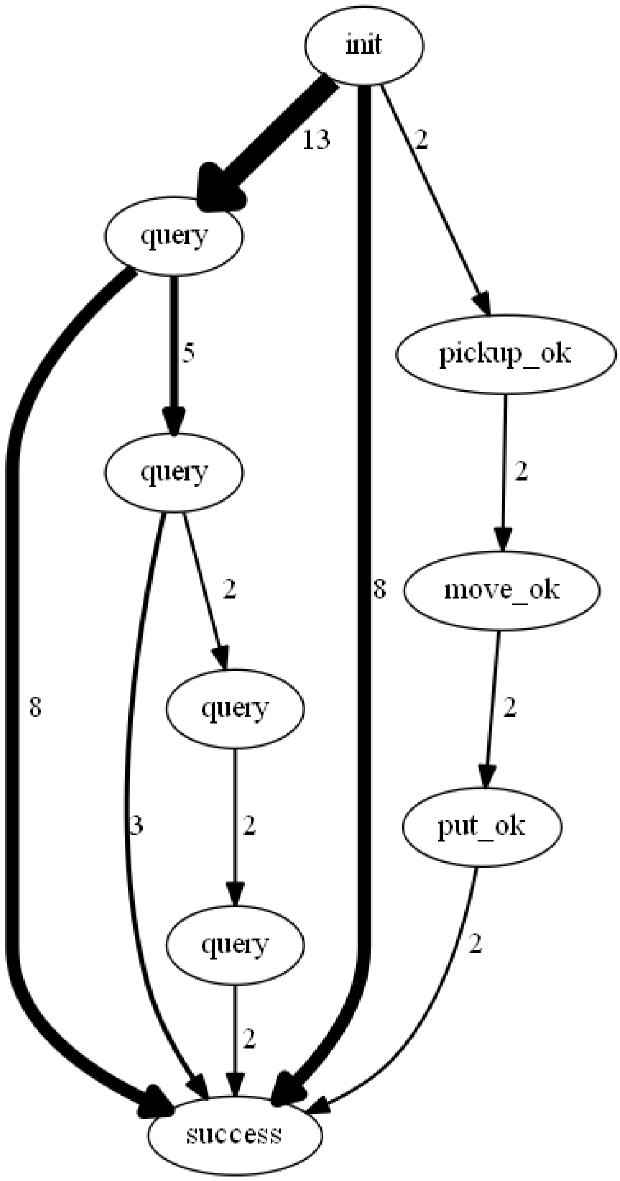


Figure 15. The sequence graph for the five most frequent sequences in the Querying Phase.

- **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 (and what type of) information was requested during each resolution sequence. The **uninformed set** contains sequences during which no information was requested (the sequence from Figure 10 is an example of such a sequence). The **description set** contains sequences during which a description was requested (the sequence from Figure 9 is an example), the **markup set** contains sequences during which markup was requested, and the **querying set** contains sequences during which at least one query was made.

Table 4 contains an overview of the attributes included in the initial references. Table 5 contains an overview of the attributes in the final references. For easier comparison, Table 6 contains the differences between the proportions in the initial and the final references.

7.1. *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. Overall, the proportion of expressions that contained a type attribute and the proportion of expressions that contained a colour attribute are lowest in the uninformed category. About half of the expressions in each category contain a landmark reference. It is remarkable that none of the observed initial expressions contained a directional description.

For the final references we observe that in all categories the proportion of expressions that contained a type attribute or a colour attribute is lower than in the initial references. We believe that this observation indicates that participants realized that the type attribute and the colour attribute were unreliable and therefore decided to use expressions that avoided values for these attributes (e.g. instead of referring to an object as "*the blue box*" they referred to it as "*the box*" (thereby removing the colour attribute from the expression) or "*the blue object*" (thereby removing the type attribute)). We also notice that the participants tended to include fewer landmark references in the final set. The drop-off is most pronounced for the uninformed category and for the querying category.

It is interesting that while the initial references did not include any directional descriptions, the final references frequently did. We believe that this is a consequence of the presence of perception errors. The type and colour of objects could be affected by perception errors, making it more difficult to produce successful references to the affected objects. Likewise, the type and colour of objects that were used as landmarks in referring expressions could be affected by perception errors. Direction based descriptions on the other hand did not rely on attributes that could be affected by 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 robust against perception errors.

We noted earlier that the drop-off in the use of the type and colour attribute was most strongly pronounced in the uninformed and the querying category, while it was not as strong in the description and the markup category. We believe that this may be explained by the type of information that these options provided. The description option and the markup option provided a complete description of how the scene appeared to the robot. Although the description of a scene did not explicitly state what the perception errors were that affected the robot's perception of the scene, the participants were able to compare the information provided by the system with their own perception of the world and figure out the divergence. They therefore had the option to align their own model of the world (temporarily for the purpose of producing a reference) to robot's flawed model of the world. This means that they did not necessarily have to abandon using unreliable attributes, but were able to use attribute values that were valid in the robot's understanding of the world.

In contrast to this, 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, and therefore had no model of the robot's understanding to align to. Instead they reached a successful reference by trial-and-error and were therefore 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

Table 4. Attributes included in the initial references.

	Type		Colour		Landmark Reference		Directional description	
Condition	**Proportion**	**Count**	**Proportion**	**Count**	**Proportion**	**Count**	**Proportion**	**Count**
Uninformed	96.34%	79	86.59%	71	45.12%	37	0.00%	0
Description	98.57%	69	87.14%	61	51.43%	36	0.00%	0
Markup	100.00%	40	92.50%	37	45.00%	18	0.00%	0
Querying	100.00%	55	94.55%	52	49.09%	27	0.00%	0

Table 5. Attributes included in the final references.

	Type		Colour		Landmark Reference		Directional description	
Condition	**Proportion**	**Count**	**Proportion**	**Count**	**Proportion**	**Count**	**Proportion**	**Count**
Uninformed	84.15%	69	58.54%	48	30.49%	25	17.07%	14
Description	98.57%	69	77.14%	54	34.29%	24	34.29%	24
Markup	95.00%	38	95.00%	38	37.50%	15	17.50%	7
Querying	94.55%	52	65.45%	36	32.73%	18	25.45%	14

Table 6. The differences between the proportions in the initial and the final references.

Condition	Type	Colour	Landmark reference	Directional description
Uninformed	-12.20	-28.05	-14.63	17.07
Description	0.00	-10.00	-17.14	34.29
Markup	-5.00	2.50	-7.50	17.50
Querying	-5.45	-29.09	-16.36	25.45

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. The extent to which this occurred is a matter for further research.

8. 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 showed that if errors are present in the perception of the robot, this leads to an increase of problems in the interactions and makes successfully completing the tasks more difficult. On the other hand, giving the human dialogue partners access to information about the robot's understanding of the scene allows them to reduce problems and complete the task more efficiently.

Furthermore we investigated references that failed due to perception errors and how the resulting problems were resolved by the participants. We found that participants resolved such problems either through a trial-and-error strategy, or, when information about the robot's perception was available, by requesting information and using it in their problem resolution attempts. We then investigated the choice of attributes in referring expressions that triggered perception based problems, the expressions that the participants produced to resolve the problems. In terms of the content of the referring expressions, we found that participants tended to include different attributes in the expressions in their final reference compared to their initial reference. The choice of which attributes to include or not to include appears to be 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, participants tend to align their referring expressions to the robot's understanding. The sequence shown 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. The sequence shown in Figure 10 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 approach to address this problem is to improve robot perception. Our work, however, indicates that another useful strategy could be to provide the human user access to the robot's perceptual model of the world. As our results show, users do request information about the robot's perception if it is available, and they are able to evaluate and relate it to their own understanding. They can then use it to either align their understanding of the world to the robot's understanding, or use it to avoid unreliable attributes and substitute them with reliable ones. An interesting implication of this is that there is potential for robot systems to use the adjustments that users make in their references to provide the robot vision system with information about possible weaknesses in its recognition mechanism. This information may be used by the robot to trigger a self-repair mechanism of its perception. We will explore this in future work.

References

[1] J. Aberdeen and L. Ferro. Dialogue patterns and misunderstandings. In *ISCA Tutorial and Research Workshop on Error Handling in Spoken Dialogue Systems*, 2003.

[2] A. Anderson, B. Bader, M., B. E., G. M. E., Doherty, S. Garrod, S. Isard, J. Kowtko, J. McAllister, J. Miller, C. Sotillo, H. S. Thompson, and R. Weinert. The HCRC map task corpus. *Language and Speech*, 34(4):351–366, 1992.

[3] J. Bateman, J. Hois, R. Ross, and T. Tenbrink. A linguistic ontology of space for natural langauge processing. *Artificial Intelligence*, 174(14):1027–1071, 2010.

[4] M. Brenner, N. Hawes, J. Kelleher, and J. Wyatt. Mediating between qualitative and quantitative representations for task-oriented human-robot interaction. In *In Proceedings of the 20th International Conference on Artificial Intelligence (IJCAI-07)*, pages 2072–2077, 2007.

[5] M. Burigo and P. Knoeferle. Visual attention during spatial language comprehension. *PLOS ONE*, 10(1), 2015.

[6] M. Burigo and S. Sacchi. Object orientation affects spatial language comprehension. *Cognitive Science*, 37(8):1471–1492, 2013.

[7] L. Carlson-Radvansky and G. Logan. The influence of reference frame selection on spatial template construction. *Journal of Memory and Langauge*, 37:411–437, 1997.

[8] H. H. Clark and E. F. Schaefer. Contributing to discourse. *Cognitive Science*, pages 259–294, 1989.

[9] F. Costello and J. Kelleher. Spatial prepositions in context: The semantics of *Near* in the presense of distractor objects. In *Proceedings of the 3rd ACL-Sigsem Workshop on Prepositions*, pages 1–8, 2006.

[10] K. Coventry, A. Cangelosi, R. Rajapakse, A. Bacon, S. Newstead, D. joyce, and L. Richards. Spatial prepositions and vague quantifiers: Implementing the funcitonal geometric framework. In *Spatial Cognition IV. Reasoning, Action, Interaction*. Springer Berling Heidelberg, 2005.

[11] K. Coventry and S. Garrod. *Saying, Seeing and Acting: The Psychological Semantics of Spatial Prepositions*. Psychology Press, 2004.

[12] S. Dobnik. *Teaching Mobile Robots to use Spatial Words*. PhD thesis, The Queen's College, University of Oxford, 2009.

[13] N. Hawes, M. Klenk, K. Lockwood, G. S. Horn, and J. D. Kelleher. Towards a cognitive system that can recognize spatial regions based on context. In *Proceedings of the 26th National Conference on Artificial Intelligence (AAAI'12), 2012*, 2012.

[14] M. Hoetjes, R. Koolen, M. Goudbeek, E. Krahmer, and M. Swerts. Reduction in gesture during the production of repeated references. *Journal of Memory and Language*, 79-80:1–17, 2015.

[15] H. Horacek. Generating referential descriptions under conditions of uncertainty. In *Proceedings of the 10th European Workshop on Natural Language Generation (ENLG)*, pages 58–67. Citeseer, 2005.

[16] J. Kelleher and F. Costello. Cognitive representations of projective prepositions. In *Proceedings of the Second ACL-Sigsem Workshop of The Linguistic Dimensions of Prepositions and their Use in Computational Linguistic Formalisms and Applications*, 2005.

[17] J. Kelleher and F. Costello. Applying computational models of spatial prepositions to visually situated dialog. *Computational Linguistics*, 35(2):271–306, June 2009.

[18] J. Kelleher, F. Costello, and J. van Genabith. Dynamically structuring, updating and interrelating representations of visual and linguistic discourse context. *Artificial Intelligence*, 167(1):62–102, 2005.

[19] J. Kelleher and R. Ross. Topology in compositie spatial terms. In D. Rapp, editor, *Proceedigns of Spatial Cognition 2010: Poster Presentations*, pages 46–50. SFB/TR8 Spatial Cognition, 2010.

[20] J. Kelleher, R. Ross, B. Mac Namee, and C. Sloan. Situating spatial templates for human-robot interaction. In *In Proceedings of the AAAI Symposium on Dialog with Robots*, 11th-13th Nov. 2010.

[21] J. Kelleher, R. Ross, C. Sloan, and B. Mac Namee. The effect of occlusion on the semantics of projective spatial terms: a case study in grounding language in perception. *Cognitive Processing*, 12(1):95–108, 2010.

[22] J. Kelleher and J. van Genabith. A computational model of the referential semantics of projective prepositiosn. In *Syntax and Semantics of Prepositions*, pages 211–228. Springer, 2006.

[23] J. D. Kelleher. Attention driven reference resolution in multimodal contexts. *Artificial Intelligence Review*, 25(1):21–35, 2006.

[24] C. Kennington and D. Schlangen. Simple learning and compositional applicaiton of perceptually grounded word meanings for incremental reference resolution. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th Internatonal Joint Conference on Natural Language Processing*, pages 292–301. Association for Computational Linguistics, Beijing, Ching, July 2015.

[25] L. Kunze, C. Burbridge, M. Albert, A. Tippur, J. Folkesson, P. Jensfelt, and N. Hawes. Combining top-down spatial reasoning and bottom-up object class recognition for scene understanding. In *In Intelligent Robots and Systems (iROS)*, pages 2910–2915. IEEE, 2014.

[26] D. J. Litman, M. Swerts, and J. Hirschberg. Characterizing and predicting corrections in spoken dialog systems. *Computational Linguistics*, 32:417–438, 2006.

[27] C. Liu, R. Fang, and J. Y. Chai. Towards mediating shared perceptual basis in situated dialogue. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 140–149. Association for Computational Linguistics, 2012.

[28] G. Logan and D. Sadler. A computational analysis of the apprehension of spatial relations. In P. Bloom, M. Peterson, L. Nadel, and M. Garret, editors, *Language and Space*, pages 493–530. MIT Press, 1996.

[29] R. López-Cózar, Z. Callejas, and D. Griol. Using knowledge of misunderstandings to increase the robustness of spoken dialogue systems. *Knowledge-Based Systems*, 23(5):471–485, July 2010.

[30] C. Manning, M. Surdeanu, J. Bauer, J. Finekl, S. Bethard, and D. McClosky. The stanford corenlp natural language processing toolkit. In *Proceedings of the 52nd Annual Meeting of the Association of Computational Linguistics: System Demonstrations*, pages 55–60, 2014.

[31] V. Mast and D. Wolter. A probabilistic framework for object descriptions in indoor route instructons. In T. Tenbrink, J. Stell, A. Galton, and Z. Wood, editors, *Spatial Information Theory*, pages 185–204. Springer, 2013.

[32] E. Ovchinnikova. *Integration of World Knowledge for Natural Language Understanding*. Atlantis Press, 2012.

[33] T. Regier and L. Carlson. Grounding spatial language in perception: An empirical and computational investigation. *Journal of experimental psychology: General*, 130(2):273–298, 2001.

[34] M. Richter, J. Lins, S. Schneegans, Y. Sandamirskaya, and G. Schoner. Autonomous neural dynamics to test hypotheses in a model of spatial language. In *36th Annual Conference of the Cognitive Science Society*, 2014.

[35] V. Rieser and D. Gkatzia. Generation for things unknown: Accounting for first-time users and hidden scenes. In *Proceedings of the 1st International Workshop on Data-to-text Generation*, 2015.

[36] R. Ross and J. Kelleher. Putting things "between" perspective. In *Proceedings of the 21st Irish Conference on Artificial Intelligence and Cognitive Sciance conference (AICS)*, September 2010.

[37] R. Ross and J. Kelleher. Using the situational context to resolve frame of reference ambiguity in route descriptions. In *Proceedings of the Second Workshop on Action, Perception and Langauge (APL'2)*, Uppsala, Sweden, November 2014.

[38] N. Schuette, J. Kelleher, and B. Mac Namee. A corpus based dialogue model for grounding in situated dialogue. In *Proceedings of the 1st Workshop on Machine Learning for Interactive Systems: Bridging the Gap Between Language, Motor Control and Vision (MLIS-2012).*, Montpellier, France, Aug. 2012.

[39] H. Schultheis, S. Bertel, and T. Barkowsky. Modelling mental spatial reasoning about cardinal directions. *Cognitive Science*, 38(8):1521–1561, November/December 2014.

[40] J. R. Searle. A taxonomy of illocutionary acts. *Language, Mind, and Knowledge*, 7:344–369, 1975.

[41] J. Shin, S. S. Narayanan, L. Gerber, A. Kazemzadeh, D. Byrd, and others. Analysis of user behavior under error conditions in spoken dialogs. In *INTERSPEECH*, 2002.

[42] K. Sjoo. *Functional Understanding of Space*. PhD thesis, KTH Computer Science and Communications, Sweden, 2011.

[43] M. Spranger and S. Pauw. Dealing with perceptual deviation - vague semantics for spatial language and quantification. In L. Steels and M. Hild, editors, *Language Grounding in Robots*, pages 173–192. New York: Springer, 2012.

[44] S. Tellex. *Natural language and spatial reasoning*. PhD thesis, Massachusetts Institute of Technology, 2010.

[45] M. A. Walker, D. J. Litman, C. A. Kamm, and A. Abella. PARADISE: a framework for evaluating spoken dialogue agents. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, pages 271–280. Association for Computational Linguistics, 1997.

[46] T. Winograd. *Understanding natural language*. New York: Academic Press, 1972.