Spatial Referring Expressions in Child-Robot Interaction: Let's Be Ambiguous!

Christopher D. Wallbridge[1], Séverin Lemaignan[1], Emmanuel Senft[1], Charlotte Edmunds[1], Tony Belpaeme[1,2]

1 University of Plymouth
2 University of Ghent

*** christopher.wallbridge@plymouth.ac.uk**

Abstract

Establishing common ground when attempting to disambiguate spatial locations is difficult at the best of times, but is even more challenging between children and robots. Here, we present a study that examined how 94 children (aged 5-8) communicate spatial locations to other children, adults and robots in face-to-face interactions. While standard HRI implementations focus on non-ambiguous statements, we found this only comprised about 20% of children's task based utterances. Rather, they rely on brief, iterative, repair statements to communicate about spatial locations. Our observations offer strong experimental evidence to inform future dialogue systems for robots interacting with children.

1 Introduction

For children arriving in a new country, learning the language of their new home is an important part of their integration. Proficiency in the language of the host country is a vital condition for success at school. Even for children of migrants born in the host country, this may be an issue if the language used at school cannot be reinforced in the home. As tailored language classes are expensive and limited in time, we wish to explore if robot tutors can be used to complement language tutoring. This is encouraged by robots having been shown to be able to reduce anxiety in a second language learning when acting as a peer [1]. However there is still much to be considered when designing a robotic language tutor [5].

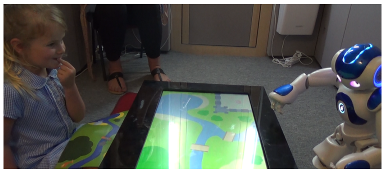

Figure 1. A child interacting with the robot in our study.

While most language tutoring systems focus on the learning of nouns and verbs, we wish to study the learning of spatial language instead: the vocabulary and grammatical constructions serving the communication of spatial relations. Spatial language is particularly challenging, as the semantics are often vague, context dependant and referent dependant. For example, in "the apple next to the bowl" the spatial referent "next" does not have boolean membership, but rather has a graded membership depending on the distance between objects and the size of the objects. A typical assumption in Natural Language Interaction Systems (NLIS) is that referring expressions (RE) are unambiguous descriptions of object locations and that a linguistic interaction between a user and a computer system follows a quite structured and clear interaction flow using unambiguous utterances [8]. This might be the case for spoken interfaces in banking systems or telephone ordering, but the literature in socio-linguistics and dialogue systems show that language is much more dynamic than NLIS typically allows for, and this is specifically prominent in spatial RE.

Socio-linguistics suggests that people do not tend to use fully specified RE. Instead, they reduce the cognitive load by under-specifying the description and then rely on a strategy of repair to correct misunderstanding if necessary [7]. Rather than this being a one-way communication, it is a fundamentally social process. The person being addressed is expected to be an active contributor to the process of reaching *common ground*. Each participant in the conversation will contribute until a *grounding criterion* is met [6], i.e. when each contributor to the communication believes that they have understood enough for their current purpose. Pickering and Garrod [11] describe this partial alignment of common ground as the natural way in which we communicate. Full common ground is only necessary when there is difficulty reaching alignment.

Dialogue management systems have to take into consideration these under specified statements. One assumption that often made in interaction between two agents is that what is said by one, is how the other understands it. However this is not always true, even in human-human interaction [10]. Instead, continuous communication can allow a system to re-evaluate its belief state of the current environment, and the belief state of other communicative agents. For spatial tasks they are able to use contextual language to help with the positioning of an item [2]. Instead of complex statements that try to pinpoint the exact location in one sentence, a series of much simpler statements is used.

By contrast, implementations of RE generation and understanding for use in robotics often follow Gricean Maxims [9], such as the Incremental Algorithm [8]. These algorithms focus on a single statement that eliminates ambiguity. While communicating clearly and unambiguously about spatial references is one solution to the problem of communicating about space, more recent systems also incorporate perspective taking [12], which may alleviate the need for precise but verbose REs. With perspective taking we do see a more interactive approach. But this process still relies on reaching full alignment by eliminating ambiguity.

Our present study provides real-world data of children establishing common ground in the natural course of playing a game. We observed them either interacting with other children, with adults or with a robot using a Wizard of Oz setup. The study provides opportunities for the children to use a large set of spatial language, perspective taking and establishing a common point of reference, whilst being easy to replicate.

2 Study Design

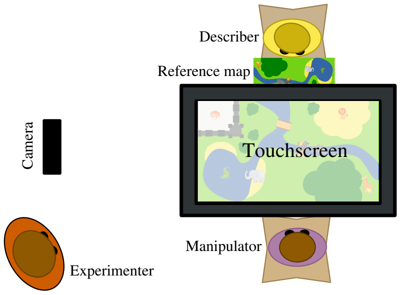

Figure 2.The experimental setup. A top down view showing the position of the manipulator and describer sitting opposite each other with the "Sandtray" screen in the middle. The experimenter is sitting to the side with a camera recording the participants.

We collected data from 94 children between the ages of 5 and 8. They were assigned to one of three conditions: child-child, child-adult or child-robot. For the child-child and child-adult conditions children from two different schools were used. They participated during the day at their school in a room for individual teaching. In the child-child condition two children from the same class participated together. In the child-adult condition a child participated with an experimenter. Those in the child-robot condition were recruited from register held by the Babylab at the University of Plymouth.

Following a sandbox paradigm [3], one child and a partner (child, robot or adult) are sitting on opposite sides of a large touchscreen (Fig. 2). The screen presents a background with different areas: a castle, a desert, two rivers with bridges, a lake, two beaches and many bushes or trees.

One agent, hereafter called the *describer*, has to guide the other agent, called the *manipulator*, to move items on the touchscreen to a desired location. The describer is provided with a reference map, which is kept hidden from the manipulator, with the desired position of eight items (Fig. 3).

While it has been shown that pointing can influence the words used [13], the task could be easily completed without words if gestures were allowed. As we were focused on the language being used, the describer was instructed not to use pointing gestures. If children attempted to use pointing they were reminded that this was not allowed.

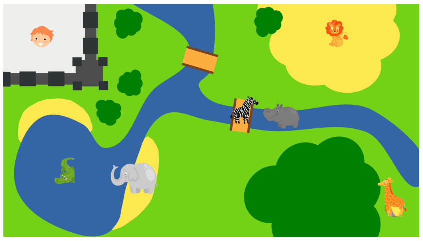

Figure 3.An example of the reference map given to a child to describe. The eight items (face, crocodile, elephant, zebra, hippo, lion, giraffe and ball) are shown in the desired location that they need to be moved to. The child describes the position on his map for an agent to manipulate into the correct position.

The touchscreen presents a background with different areas (Fig. 3). Eight movable items have to be moved to specified locations on the map. The reference maps were designed to elicit a number of different ways to describe the position of objects. Some objects were facing a particular direction, to encourage locutions like 'in front of' or 'behind'. Features, such as the bridges and bushes, were repeated so as to require disambiguation. Verbal disambiguation was also elicited by the relatively small size of the screen, which limits the effectiveness of joint gaze to identify the correct location for an object.

In the case of the child-child and child-adult conditions, after the first map was completed, the role of manipulator and describer would be swapped. In the case of the child-robot condition the child would be invited to describe the second map. The robot itself would appear to move objects around the touchscreen via the use of a Wizard of Oz control interface, held by an experimenter. The experimenter is able to move an object on their interface, the robot would then move its hand to point at the object and then move its hand to point at the target location, with the object moving with it.

3 Results

For statistical power reasons, we focused our current observation of results on the child-child interaction (Child-Child=60, Child-Adult=26, Child-Robot=8), while providing more qualitative observations of the other conditions in the discussion.

We observed an average of 7.12 (SD=7.50) repair statements used per round (one round consisted of one map with eight objects to be moved). The SD shows large inter-personal variations. There were comparatively few cases of repair statements requiring spatial perspective taking (M=0.56 per round). Despite being told not to use them, there was an average of 2.43 (SD=3.03) pointing gestures used per round.

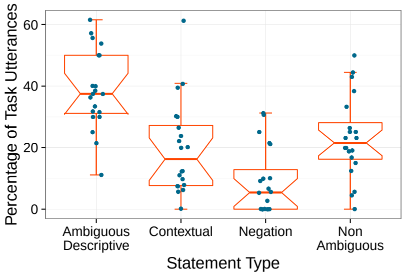

Figure 4. Break down of on-task statements. Ambiguous descriptive statements were a significantly higher proportion than the other statement types.

We took all the on-task statements from a sample of 10 child-child sessions, giving us data from 20 children. The statements were divided into the following categories: Ambiguous-Descriptive (statement refers to more than one location e.g.'the zebra is on a bridge'), Contextual (statement following from previous statements, that would make no sense to a third person entering the conversation e.g. 'the other one'), Negation(statement indicating that it is an incorrect location with no further description e.g. 'no'), Non-Ambiguous (statement that describes only one possible location e.g. 'the crocodile is in the big lake') and Pointing.

On average Ambiguous-Descriptive statements were used 38.6% of the time, Contextual in 13.1%, Negation in 9% and Non-Ambiguous in 23.2%. Using a Welch two-sample t-test we find that the Ambiguous-Descriptive statements are used significantly more than any other type of statements, and Cohen's d test shows a large effect size in each case (Contextual: $t(38) = 4.2$, $p < .001$, $d = 1.34$; Negation: $t(38) = 7.8$, $p < .001$, $d = 2.48$; Non-Ambiguous: $t(38) = 3.7$, $p < .001$, $d = 1.17$).

4 Discussion

Our observations show that interactions between children (and between children and robots) are highly dynamic, fast-paced and relying on the situatedness and embodiment of the conversation partners [4], very unlike the "walkie-talkie exchanges" typically used in Human-Robot Interaction. Between children, as soon as the manipulator has enough information to make a guess they will often start moving the objects, without waiting until enough information is given as to be non-ambiguous. This has two possible outcomes: either they guess right, or it causes the describer to generate a repair statement. It also appears that typically it is easier for the describer to let the manipulator start moving the objects – knowing that the position they described is ambiguous – so that they may then generate a short, easily understood, repair, reducing the cognitive load. In fact we see that the robot's inability to change course after it has started moving an object caused frustration to the child describing.

In the child-robot condition there appeared to be a reduction of the repair statements when the robot moved items incorrectly. This could be caused by many factors, such as the children feeling more nervous with the robot, the expectations they have of its abilities and the absence of some basic social cues, such as back channelling and lack of eye contact, all of which made the interaction laborious.

Pointing was still prevalent, despite it being disallowed and discouraged (even the experimenter was found pointing or indicating directions). Future work could look at a different methodology to encourage the combination of gestures and language.

5 Conclusion

Counter to many implementations that seek to eliminate ambiguity entirely, we find that children tend to use many ambiguous statements when describing the location of objects. As such *the robot, when being given RE, must expect ambiguous statements*. It should not wait for further information, but rather start acting on the information it has, as this will also assist in the process of description. This also means that the robot should be prepared to react quickly to repair statements by enabling it to diverge from its current action to take into account the new information.

This also means *the robot should be allowed to be ambiguous in its descriptions*. This may be beneficial to reduce processing requirements for the robot itself, but also may help reduce the cognitive load for its conversational partner. When doing so, the robot should monitor closely the reaction of its partner, and be prepared to provide timely repairs to lead the implicit, interactive disambiguation process.

Our next steps are to implement a more interactive robot to collect more data with children interacting with the robot. Using this data we will be able to build an effective framework for natural spatial communication between children and robots.

6 Acknowledgements

This work was supported by the EU H2020 L2TOR project (grant 688014), the EU H2020 Marie Sklodowska-Curie Actions project DoRoThy (grant 657227) and the EU FP7 DREAM project (grant 611391).

References

1. M. Alemi, A. Meghdari, and M. Ghazisaedy. The impact of social robotics on L2 learners' anxiety and attitude in English vocabulary acquisition. *International Journal of Social Robotics*, pages 1–13, 2015.

2. T. Baumann, M. Paetzel, P. Schlesinger, and W. Menzel. Using Affordances to Shape the Interaction in a Hybrid Spoken Dialog System. In *Proceedings of ESSV*, Bielefeld, Germany, Mar. 2013.

3. P. Baxter, R. Wood, and T. Belpaeme. A touchscreen-based'sandtray'to facilitate, mediate and contextualise human-robot social interaction. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*, pages 105–106. ACM, 2012.

4. T. Belpaeme, S. J. Cowley, and K. F. MacDorman. *Symbol grounding*, volume 21. John Benjamins Publishing, 2009.

5. T. Belpaeme, P. Vogt, R. van den Berghe, K. Bergmann, T. Göksun, M. de Haas, J. Kanero, J. Kennedy, A. C. Küntay, O. Oudgenoeg-Paz, et al. Guidelines for designing social robots as second language tutors. *International Journal of Social Robotics*, 2017.

6. H. H. Clark and E. F. Schaefer. Contributing to discourse. *Cognitive science*, 13(2):259–294, 1989.

7. H. H. Clark and D. Wilkes-Gibbs. Referring as a collaborative process. *Cognition*, 22(1):1–39, 1986.

8. R. Dale and E. Reiter. Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cognitive science*, 19(2):233–263, 1995.

9. H. P. Grice, P. Cole, J. Morgan, et al. Logic and conversation. *1975*, pages 41–58, 1975.

10. G.-J. M. Kruijff, M. Janíček, and P. Lison. Continual processing of situated dialogue in human-robot collaborative activities. In *RO-MAN, 2010 IEEE*, pages 594–599. IEEE, 2010.

11. M. J. Pickering and S. Garrod. Alignment as the basis for successful communication. *Research on Language & Computation*, 4(2):203–228, 2006.

12. R. Ros, S. Lemaignan, E. A. Sisbot, R. Alami, J. Steinwender, K. Hamann, and F. Warneken. Which one? grounding the referent based on efficient human-robot interaction. In *RO-MAN, 2010 IEEE*, pages 570–575. IEEE, 2010.

13. A. Sauppé and B. Mutlu. Robot deictics: How gesture and context shape referential communication. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*, pages 342–349. ACM, 2014.