Group-based Spatial Reference in Linguistic Human-Robot Interaction

Thora Tenbrink (tenbrink@informatik.uni-bremen.de)
Faculty of Linguistics and Literary Sciences, Universität Bremen
Bibliothekstr. 1, 28359 Bremen, Germany

Reinhard Moratz (moratz@informatik.uni-bremen.de)
Department of Mathematics and Informatics, Universität Bremen
Bibliothekstr. 1, 28359 Bremen, Germany

Abstract

To enable a robot to interpret natural language instructions with regard to a specific object in a configuration, the robot needs to be equipped with a repertory of representations that human users might employ for reference. In this paper, we focus on the ways in which objects are referred to in relation to other similar objects, by analysing the results of an experiment involving a robot and human users unfamiliar with it. The robot is designed to understand reference via basic intrinsic and relative reference systems. Results showed that our implemented computational model could handle standard instances of spatial reference, but needed refinement to be able to interpret specific variations that were employed by users based on contextual factors analysed in this paper.

Introduction

Service robotics is an area in which technological progress leads to rapid development and continuous innovation. One central aim is to enable future service robots to be instructed intuitively by laypersons, such as elderly or handicapped people, rather than needing to be programmed by experts, as is still the case with stationary industry robots utilised in factories.

In a typical service robotics scenario, a robot is instructed by a human user to act upon a specific object. To this aim, both participants need to negotiate their internal conceptual representations linguistically to identify the referent. Between humans, such communication is fairly straightforward, as humans do not differ fundamentally with regard to perception and conception; furthermore, humans have a broad range of options at their disposal for negotiation of a referent. For instance, the object can be referred to via its class name, its function, or via details distinguishing it from competing referents in the shared visual field of the interactants. In a scenario such as a bakery where the distinctions between the objects are hard to specify, such as diverse kinds of rolls, humans naturally employ pointing gestures.

Currently, modern robots have only limited access to either of these options. Visual details cannot be recognised and distinguished easily [Moratz et al., 2001], while (coverbal) gestures involve complex cognitive processes [McNeill, 2000] that are currently being researched and made accessible for automatic processing in major research projects, for instance, in Bielefeld [Wachsmuth and Sowa, 2002]. Moreover, capturing gestures via data gloves may in many cases not be feasible. Our research therefore concentrates on a different option, namely, the spatial differentiation between objects that the robot cannot distinguish sensorially or conceptually.

Spatial location descriptions can serve to distinguish objects within a group of similar objects that can be recognised both by the robot and the human instructor. Relevant candidates for this aim are projective spatial expressions, which have been researched extensively in human-human interaction (e.g., [Vorwerg et al., 1997], [Zimmer et al., 1998]). However, the concept of *group-based reference*, which is central to our aims, has largely been neglected in the literature. Most scenarios investigated experimentally involve either reference to one of several *different* objects, or the specification of the location of one object, rather than the identification of one of several similar objects in a group. Moreover, there is no way of predicting whether human users in a robotic scenario adhere to the same preferences and principles of spatial reference that have been identified in human-human settings.

In this paper, we present an experiment involving a robot and human users unfamiliar with it, and analyse the users' choices of group-based reference in different kinds of spatial configurations. The results are used to improve the existing system that is designed to interpret intrinsic and relative reference systems, focussing on instructions that involve spatial reference relative to the rest of a group of objects (rather than relative to the robot's intrinsic features or to a different object, which are also typical kinds of instructions that could be handled by the system).

A brief classification of reference systems

In the following, we present a brief classification of reference systems that serves as the basis for our implemented computational model. In localising reference objects in space, humans have - broadly speaking - three kinds of reference systems at their disposal, which (in Levinson's [Levinson, 1996] terminology) may be called *intrinsic*, *relative*, and *absolute*. In intrinsic reference systems, objects are located by referring to the intrinsic properties of another entity, such as the speaker's front in *The ball is in front of me*. Relative reference systems depend on the presence of a further entity (the so-called *relatum*), as in *The ball is in front of the table*. Absolute reference systems depend on the earth's cardinal directions, such as *north* or *south*. Additionally, speakers may variously employ either their own or their listener's *point of view* (also called *origin*) - or, which in some situations may also be useful, the perspective of a third entity (as in, *Viewed from the church's entrance, there is a bookshop on the right*).

This classification of reference systems has proved useful in order to identify one entity's position in relation to one or two different entities. We augment it by assuming that speakers can also specify the position of one object in a group of similar ones by specifying its relation to the rest of the group ([Moratz et al., 2001]). This is also the case if the group consists only of two similar objects. For reference of this kind projective adjectives are employed, which "can be used only as restrictive modifiers to select an object (the figure) in contrast to another object (called here 'ground') of the same category." [Eschenbach, 2003].

Speakers can combine any of these reference systems, they can modify them by providing additional information of various kinds, and they can also refer to a target object on the basis of relative distance, which is also a qualitative kind of spatial reference.

Results of psycholinguistic experiments, e.g. [Zimmer et al., 1998], point to the fact that the employment of specific projective terms is dependent on the spatial relationship between the entities. For example, canonical expressions are used for prototypical relations: "in front of" is only used without modification if a (target) object is located in a certain restrictive area with regard to another object. For non- prototypical relations, composites are used. [Eschenbach, 2003] formally defines the semantics of projective terms, taking into account contextual, functional, and geometric components. Such findings serve as a suitable benchmark for generating expectations with regard to the employment of projective terms. Furthermore, our previous findings [Moratz et al., 2001] already allow us to narrow down the scope of possible kinds of reference, since users consistently use the robot's perspective as origin, and they (mostly) refrain from employing absolute reference systems. We apply these expectations to human-robot interaction scenarios and aim at equipping our robot with the respective cognitive abilities. The next section outlines the status at the time of the experiment.

A robot designed to model human spatial representations

The following components interact in the system: the speech component, the semantic component, the spatial reasoning component, and the sensing and action component. The *speech component* uses the Nuance speech recognition module and generates feature-value structures. The *semantic component* produces underdetermined propositional representations of the spatial domain. The semantic component also implements the computational model of projective relations described in more detail below. It maps the spatial reference expressions of the given command to the relational description delivered by the perception component. The *spatial reasoning component* plans routes through the physical environment. To follow an instruction, the goal representation constructed by the semantic component is mapped onto the perceived spatial situation. The *sensing and action component* consists of two subcomponents: perception and behavior execution. The *perception subcomponent* uses a laser range finder. The measurements of the laser range finder are segmented and objects are classified. The spatial arrangement of these objects is delivered to the semantic component as a qualitative relational description. The *behavior execution subcomponent* manages the control of the mobile robot (Pioneer 1).

The *computational model* of projective relations, implemented in the semantic component, can be characterised as follows. To model reference systems that take the robot's point of view as origin, all objects are represented in an arrangement resembling a plan view (a scene from above). The reference axis is a directed line through the center of the object used as relatum (see figure 1), which may be the robot itself, the group of objects, or other salient objects.

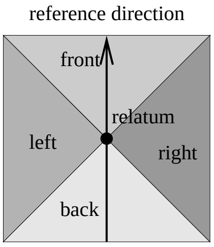

Figure 1: Relatum and reference direction

The partitioning into sectors of equal size is a sensible model for the directions "links" (left), "rechts" (right), "vor" (front) and "hinter" (back) relative to the relatum. However, this representation only applies if the robot serves as both relatum and origin. If a salient object or the group is employed as the relatum, front and back are exchanged, relative to the reference direction [Herrmann, 1990]. The result is a qualitative distinction, as suggested, for instance, by Hernandez (1994) . An example for this configuration is shown in figure 2. In this variant of relative localisation, the "in front of" sector is directed towards the robot [1].

In cases with a group of similar objects, the centroid of the group serves as virtual relatum. Here the reference direction is given by the directed straight line from the

[1] The assumption of a partition with unique correlations was called into question by our experimental results (see below).

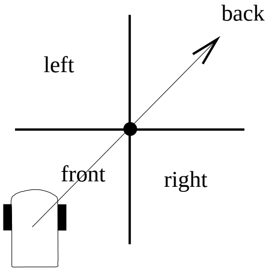

Figure 2: Relative reference model

robot center to the group centroid. [2] The object closest to the group centroid can be referred to as the “middle object” (see figure 3).

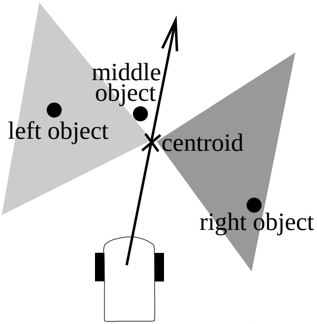

Figure 3: Group based references

Experimental Study

The people who are familiar with our robot system as described in the previous section usually have no problems achieving reference to a target object on the basis of their knowledge about the implemented features. For instance, reference to the “left object” in a configuration involving three similar objects in a row does not cause complications. In this sense, we have already achieved smooth and effective communication using natural language instructions. Related achievements with regard to spatially situated human- robotic systems have been mastered successfully, for instance, by major research projects in Saarbrücken (e.g., [Stopp et al., 1994]) and Bielefeld (e.g., [Knoll et al., 1997]), focussing on different aspects of human robot interaction scenarios.

However, to achieve *natural* communication between (service) robots and lay users it is necessary to determine what kind of language users employ in a realistic human-robot interaction situation. Based on a method developed by [Fischer, 2003], we therefore confront naive human users with a robot they are unfamiliar with in order to identify their intuitive strategies. The experiment is designed to influence the test persons as little as possible in order to trigger strategies of spatial reference that seem suitable and natural to the users.

Procedure

We asked 25 test persons to make the robot described above move to particular locations pointed at by the experimenter. They perceived a scene in which two or three objects (identical cardboard boxes, and in some cases a barrel of similar size) were placed on the floor together with the robot. The objects were arranged in five controlled variations, the first of which is depicted in figure 4. In the second variant, the middle box was exchanged for a barrel. The other variants constituted further configurations of the objects. Thus, different kinds of instructions were triggered, involving, for instance, reference to the barrel as a salient object, or reference to the three boxes as a group. When the instruction could not be interpreted by the robot, the only response the users received was “Ich verstehe nicht” (I don’t understand). When the instruction was interpreted, the robot simply started to move. This procedure was carried through for all five object configurations in four conditions for different groups of test persons. The four conditions differed as to whether the test persons started using spoken or written instructions, and whether the robot initially gave a scene description of what it perceived, or said nothing in the beginning. The effects resulting from these different conditions are analysed, for instance, in [Moratz and Tenbrink, 2003]. In the following, we focus on the instructions involving group-based reference that were uttered spontaneously by our test persons.

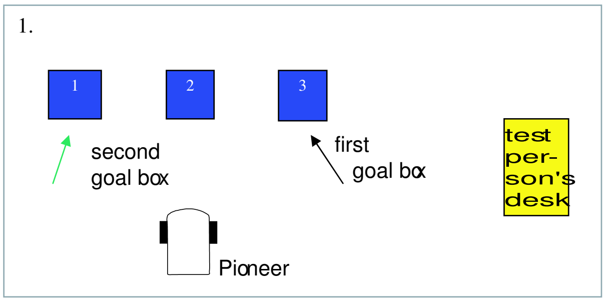

Figure 4: Configuration 1.

Results

In general, it took some time until our test persons experienced success. Until then, most of them tried out diverse kinds of instructions which the robot (being designed to interpret instructions referring directly to the goal object) could not interpret, such as incremental directions reminiscent of remote control: “move forward”, “turn around”, etc. (replicating previous results, see [Moratz et al., 2001]). The number of unsuccessful attempts varied greatly between users, and the reasons for changing instructional strategies also differed (cf. [Moratz and Tenbrink, 2003]). Most of our subjects experienced their first success by employing a group- based instruction such as “Gehe zum linken Karton” (Go to the

[2] Our experimental results motivated the additional implementation of an alternative model identifying a different reference direction, see below.

left box)[3], although there were also other possibilities, such as intrinsic reference using the robot's front as relatum, etc., as described above. Following this, they tried to employ the same kind of instruction for further reference, i.e., they used the same syntactic construction and employed group-based reference in the following configurations.

In two cases, the ensuing instructions did not conform to our expectations. In one configuration (named configuration 3), there was only one box that the robot could see at all, which was located diagonally to the left in front of the robot. Another box was placed behind the robot. There were basically two kinds of reference that we expected to be employed here: one is to simply refer to the box by its class name without reference to its location in space, as the robot could not perceive any other box. As the human users could not be sure about that, it would be equally sensible to refer to its location in relation to the robot, as in "Gehe zum Karton vorne links" (Go to the box in front left). It is hard to conceive how a group-based reference system could be employed in a situation like the present one, where the boxes were located in considerable distance to each other and, above all, at different sides of the robot.

However, several of our test persons did use references that did not seem to be suitable for the present configuration after they had been successful with group- based reference, using projective adjectives, in previous instructions. The following protocol of one test person's written instructions (who is not a rare exception but rather one of several who acted similarly) illustrates this:

Configuration 1 (as depicted in figure 4)

USER: *rechts (right)*
ROBOT: *Ich verstehe nicht. (I don't understand)*
USER: *right*
ROBOT: *Ich verstehe nicht.*
USER: *start*
ROBOT: *Ich verstehe nicht.*
USER: *fahre zur rechten Kiste (drive to the right box) (*success*: the robot moves to the correct box)*

Configuration 2 (same as 1, but middle box is exchanged for a barrel)

USER: *fahre zur linken Kiste (drive to the left box) (*success*)*

Configuration 3 (one box diagonally ahead of the robot, one behind it)

USER: *fahre zur linken Kiste*
ROBOT: *Ich verstehe nicht.*
USER: *fahre zur linken Kiste*
ROBOT: *Ich verstehe nicht.*
USER: *fahre zur vorderen Kiste (drive to the front box)*
ROBOT: *Ich verstehe nicht.*
USER: *fahre zur Kiste (*success*)*

This user employed group-based reference three times before she tried stating the class name without reference to the object's location in space, which was expected and, indeed, often employed by other test persons.

According to our results, previously experienced success involving group-based reference encourages users to employ a similar instruction format in configuration 3, even though it does not seem very suitable there. Users with a different dialogue history did not employ group-based reference in this configuration. But - conversely - those users who experienced their first success by simply stating the class name in configuration 3 tried to achieve reference in a similar way also in other configurations where such a description is not sufficient, as there were - in contrast to configuration 3 - other boxes present that could be perceived by the robot.

The second case in which group-based reference caused problems was the fourth configuration, which is depicted in figure 5. Here, the target object could be referred to - in relation to the other objects - via the fact that it was situated in the middle between them, which would yield an utterance such as "go to the middle object"; i.e., referring to the target box not by its class name, but by a higher-order category that includes the barrel as well. However, this construction was disregarded. Overwhelmingly, users spontaneously uttered a construction like "go to the middle box". In a strict sense, this is not correct; accordingly the robot could not interpret it, as it only perceived two boxes (and a barrel).

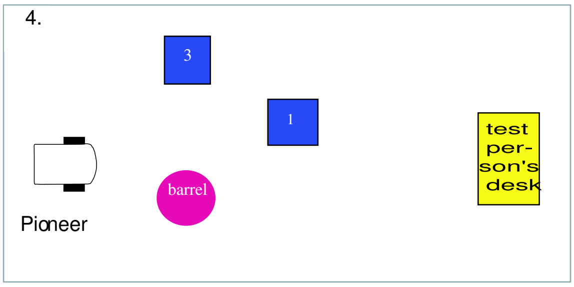

Figure 5: Object configuration 4.

Many users then attempted a different group-based instruction in which the target box was referred to as "der hintere Karton" (the rear box). This kind of reference conflicted with the implemented computational model, triggering considerations of modification (see below). However, they could equally well have referred to it as "der rechte Karton" (the right box), as the target box

[3] Translations are approximations and have to be treated with caution. In the mapping of spatial reference systems to linguistic surface, there is no one-to-one correspondence between English and German.

was situated to the right of the only other box that was present. This was done by some users, but seldom as first instruction in this configuration.

Ensuing modifications of the computational model

Following the insights we gained through our experiment, we improve the range of our robot's ability to interpret natural language instructions on diverse levels. Here, we outline the modifications to the implemented computational model that enable the system to cover more of the empirically observed variations of group-based reference.

Reference to a single perceived box by a projective adjective, which should, formally, only be employed if there are other objects of the same category present, [Eschenbach, 2003], can be interpreted by assuming that the projective term does not point to the relation to the rest of the group (if there is no such group), but to the robot itself. This yields an interpretation close to the implemented model of intrinsic reference.

The acceptability area of the projective terms needs to be enlarged as users do not necessarily aim at identifying an exact reference direction. Note that specifying a reference direction more precisely would indeed only be necessary if the task involved stating the location of an object rather than identifying one out of several objects; or if the setting involved many objects rather than just a few. In contrast, in our scenario, "the left object" in configuration 3 can only be the one diagonally ahead, as there is no way in which this description could apply to the object behind the robot. The improved model (see figure 6) also reflects better a finding by [Franklin et al., 1995], namely that positions on the right diagonal can be accepted both as "to the right" and as "in front" of a reference point.

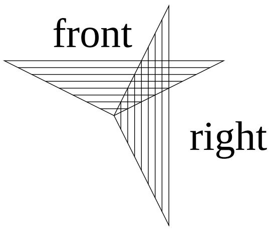

Figure 6: Enlarged acceptance areas (from $90°$ to $120°$).

Reference to "the middle box" in a configuration where there are only two boxes are captured by allowing for a different object to be included into the group.

As noted above, the expression "der hintere Karton" (the rear box) also conflicted with the implemented computational model. In the given configuration, the group of similar objects consisted of only two boxes. Thus, the centroid of the group was located between the two boxes, yielding a clear *left* and a clear *right* box, but no obvious *rear* one. Still, humans have no difficulties identifying the rear box in the given configuration. In our view, there are several possible reasons for this that need to be accounted for in the computational model. First, users could - just as in the expression "der mittlere Karton" (the middle box) - include the barrel into the group without making this explicit by using a superordinate term instead of the class name. This can be captured by including the different object into the group, as above. Second, the users could consider the robot's view direction as salient, which is plausible especially in a configuration like no. 4 where the robot is located directly opposite of the test person's desk. Then, the reference direction would not by determined by the group's centroid, but rather by robot's symmetry axis. The relatum would be still the group centroid. This would yield a clear rear sector that includes the target object.

Further experiments will be necessary to ascertain whether these adaptations are capable of interpreting the full range of variability of spatial reference employed naturally by human users.

Conclusion and Future Perspectives

In this paper, we have presented the results of an experiment involving human users and a robot they are unfamiliar with, focusing on instructions naturally employed by the users that involve group-based reference. The system's functionality is successful whenever users employ expected, or typical, options; however, users often resort to more remote choices in dependence on several discoursal and configurational factors that need to be accounted for by the system. We have outlined some of these factors based on our experiment and improved the computational model to ensure a broader range and flexibility of interpretation of spatial reference.

In the future, we will carry out further rounds of experiments in order to analyse more closely the users' choices of group-based vs other kinds of spatial reference, and contextual factors influencing such choices. The results will be used incrementally for further refinement of the computational model. Gradually, the scenario will become broader to include spatial reference on a larger scale and involve more complicated configurations of diverse kinds of objects.

Psycholinguistic experiments employing statistical measures could be useful at this point to highlight the degree to which diverse factors influence the users' choices, in order to predict more precisely (quantitatively) what could happen in a given scenario. Our (qualitative) method, which we consider different from but compatible with such studies, enables us to determine the range of what to expect in human-robot interaction scenarios. The implementation of a flexible computational model can then allow for the interpretation of even remote possibilities of reference, ensuring the applicability to other, more complicated settings also on a larger scale.

Further considerations regarding the setting will improve the communication between human users and robots in various ways. Having proved (in this and our previous study, see [Moratz et al., 2001]) that test persons are reluctant to employ instructions referring directly to the goal object, we will test how users react when provided with an example instruction such as “go to the left box”. Based on the findings of the present study, we are especially interested in how far users will react by employing projective adjectives (usually confined to group-based reference) even in situations where there is no group. - We will also experiment with context dependent robot output to investigate how interpretable instructions can be triggered unobtrusively.

Regarding the complexity of finding an appropriate and unambiguous reference direction, we will investigate specifically the influence of the robot's orientation in relation to the group of objects and to the test person. For instance, we expect the establishment of a reference system to be decisively simplified if the robot actively orients its front towards the group centroid in relation to the group's principal axis, e.g., by moving parallel to the group's principal axis and turning around to face the center of the group.

Acknowledgment

The authors would like to thank John Bateman and Kerstin Fischer for interesting and helpful discussions related to the topic of the paper, and Tilman Vierhuff for constructing the parser. Furthermore, we thank Soeren Schwertfeger, Thomas Berndt, Andi Winterboer, Daniil Liberman, and Andreas Meyer for programming the robot and for supporting the experiments. We acknowledge the active participation of the student project group Dialog, the support by Verena Mangold and Jessica Renziehausen, and, last but not least, the financial support by the DFG for the research project SFB/TR8 Spatial Cognition.

References

[Eschenbach, 2003] Eschenbach, C. (2003). Contextual, functional, and geometric features and projective terms. In Carlson, L. and van der Zee, E., editors, *Functional and Spatial Features in Language and Space*. Oxford University Press, Oxford. *In print*.

[Fischer, 2003] Fischer, K. (2003). Linguistic Methods for Investigating Concepts in Use. In Stolz, T. and Kolbe, K., editors, *Methodologie in der Linguistik*. Lang, Frankfurt a.M.

[Franklin et al., 1995] Franklin, N., Henkel, L., and Zangas, T. (1995). Parsing surrounding space into regions. *Memory and Cognition*, 23:397–407.

[Hernández, 1994] Hernández, D. (1994). *Qualitative representation of spatial knowledge*. Lecture Notes in Artificial Intelligence. Springer Verlag, Berlin, Heidelberg, New York.

[Herrmann, 1990] Herrmann, T. (1990). Vor, hinter, rechts und links: das 6h-modell. psychologische studien zum sprachlichen lokalisieren. *Zeitschrift für Literaturwissenschaft und Linguistik*, 78:117–140.

[Knoll et al., 1997] Knoll, A., Hildebrandt, B., and Zhang, J. (1997). Instructing cooperating assembly robots through situated dialogues in natural language. In *Proc. IEEE Conference on Robotics and Automation*, Albuquerque, New Mexico.

[Levinson, 1996] Levinson, S. C. (1996). Frames of Reference and Molyneux”s Question: Crosslinguistic Evidence. In Bloom, P., Peterson, M., Nadel, L., and Garrett, M., editors, *Language and Space*, pages 109–169. MIT Press, Cambridge, MA.

[McNeill, 2000] McNeill, D. (2000). *Language and Gesture*. Cambridge University Press, Cambridge, UK.

[Moratz et al., 2001] Moratz, R., Fischer, K., and Tenbrink, T. (2001). Cognitive Modeling of Spatial Reference for Human-Robot Interaction. *International Journal on Artificial Intelligence Tools*, 10(4):589–611.

[Moratz and Tenbrink, 2003] Moratz, R. and Tenbrink, T. (2003). Instruction modes for joint spatial reference between naive users and a mobile robot. In *Proceedings of RISSP IEEE International Conference on Robotics, Intelligent Systems and Signal Processing, Special Session on New Methods in Human Robot Interaction, October 8-13, 2003, Changsha, China*.

[Stopp et al., 1994] Stopp, E., Gapp, K., Herzog, G., Laengle, T., and Lueth, T. (1994). Utilizing spatial relations for natural language access to an autonomous mobile robot. In *KI-94: Proceedings of the Eighteenth German Conference on Artificial Intelligence*, Berlin, Heidelberg. Springer.

[Vorwerg et al., 1997] Vorwerg, C., Socher, G., Fuhr, T., Sagerer, G., and Rickheit, G. (1997). Projective relations for 3d space: Computational model, application, and psychological evaluation. In *AAAI'97*, pages 159 – 164.

[Wachsmuth and Sowa, 2002] Wachsmuth, I. and Sowa, T., editors (2002). *Gesture and Sign Languages in Human-Computer Interaction, International Gesture Workshop, GW 2001, London, UK, April 18-20, 2001, Revised Papers*, volume 2298 of *Lecture Notes in Computer Science*. Springer.

[Zimmer et al., 1998] Zimmer, H. D., Speiser, H. R., Baus, J., Blocher, A., and Stopp, E. (1998). The Use of Locative Expressions in Dependence of the Spatial Relation between Target and Reference Object in Two-Dimensional Layouts. In Freksa, C., Habel, C., and Wender, K. F., editors, *Spatial Cognition*, Lecture Notes in Artificial Intelligence, pages 223–240. Springer-Verlag, Berlin.