An Anthropomorphic Agent for the Use of Spatial Language

Tanja Jörding, Ipke Wachsmuth

Knowledge-Based Systems Group
(Artificial Intelligence)
Faculty of Technology
University of Bielefeld
{tanja,ipke}@techfak.uni-bielefeld.de

Abstract. In this paper we describe the communication with a responsive virtual environment with the main emphasis on the processing of spatial expressions in natural language instructions. This work is part of the VIENA project in which we chose interior design as an example domain. A multiagent system acts as an intelligent mediator between the user and a graphics system. To make the communication about spatial relations more intuitive, we developed an anthropomorphic agent which is graphically visualized. Considering the human-like figure we explain the use of qualitative spatial expressions, like "right of" and "there".

1 Introduction

Interactive 3-dimensional graphics systems are more useful (e.g. in design), when users can concentrate on their imaginations and be free from technical considerations. Therefore it is important to improve interaction with the virtual environment by way of natural, intuitive communication forms.

In our work we consider a "virtual interface agent" (VIA) as an intelligent mediator in human-computer interaction, which translates qualitative natural language input to quantitative graphics commands. A particular aim is the processing of verbal spatial expressions. Therefore we developed an articulated anthropomorphic agent which is the visible instance of the VIA (fig.1). This agent can encourage the use of natural language and be conceived either as a second "person" or as a personification of the user. With the help of this agent we can place the user's eye in the virtual environment and allow the use of situated spatial instructions.

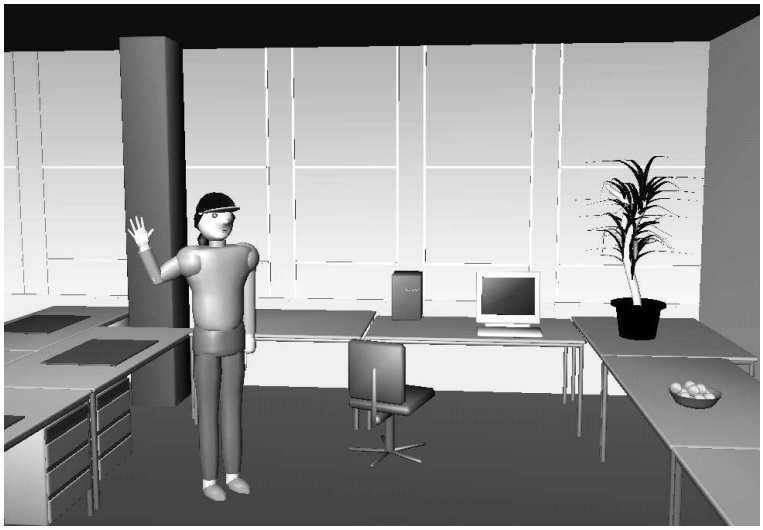

Figure 1: Example scene: the anthropomorphic agent in a virtual office room

In this paper we concentrate on the processing of spatial expressions. Because of the increasing interest in simulated human listeners, visible figures and robots that are controlled by natural language, the problem of varied perspectives has become very important. To get a smooth interaction, a flexible use of different perspectives must be possible. There are some natural language systems which deal with different perspectives , e.g. [Retz-Schmidt, 1988] [Olivier et al., 1994]. They consider deictic and intrinsic frames of view. An additional frame of reference (adressee-centered) is considered by [Schober, 1995]. Recently [Gapp, 1995] presented an approach to the computation and empirical evaluation of basic spatial relations' meanings in 3D space. His main emphasis is on clarifying the dependencies between angle, distance and shape with respect to simple idealized objects.

In our project we consider a more complex setting with a visible listener, where spatial issues becomes more realistic. With a human-like figure and some kind of gestures there are better possibilities to simulate a natural discourse, e.g. the use of indexical spatial expressions like 'here' and 'there'. On the other hand, compared to 'purified' settings, there are additional problems to be dealt with, like the selection of the actual frame of reference.

In the following we first explain the VIENA system in which the VIA is embedded. In Section 3 we describe the communication about space in the presence of an anthropomorphic agent, considering dimensional and positional adverbs. Concluding, we discuss our ideas and give an outlook on future work.

2 The VIENA Project

VIENA[1] ("Virtual Environment & Agents") is a project in the research focus theme "Artificial Intelligence and Computer Graphics" at the University of Bielefeld. The overall goal [Wachsmuth & Cao, 1995] is an intelligent communication with a virtual

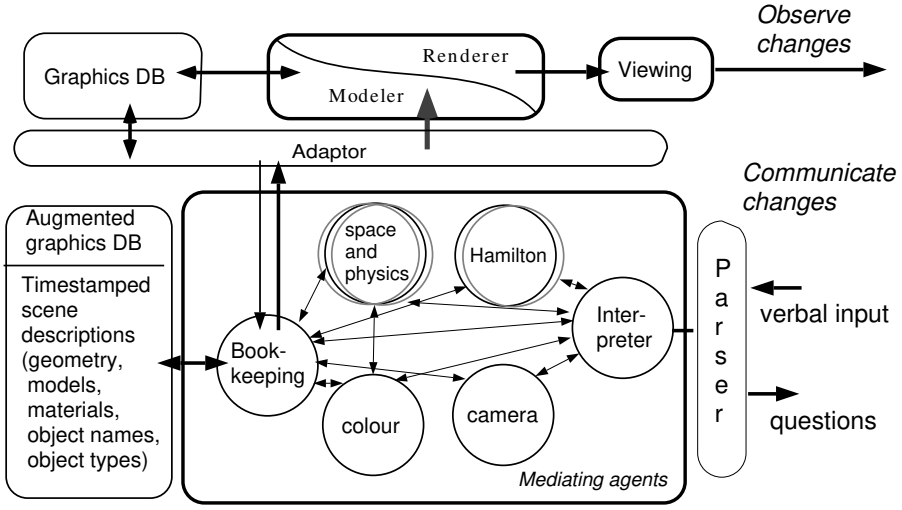

Figure 2: The architecture of the VIENA system (after [Wachsmuth & Cao,1995])

environment. As an example application we chose interactive design and exploration. Instead of manipulating scene details by mouse and menus, we communicate with the system by way of natural language. A set of agents (fig. 2), which altogether form a virtual interface agency, translate qualitative verbal instructions to quantitative commands that update the visualization scene model. The parser translates an instruction to a deep-level representation which outputs to the interpreter. There, instructions are distributed among specialized agents, e.g., a space-and-physics agent which computes physically grounded scene coordinates from qualitative spatial relations. A bookkeeping agent is authorized to access and modify the augmented data base to supply current situation information to agents on request. Some of these agents are realized as agencies, that means there are two or three instances of the same agent-type with a slightly varied functionality.

The anthropomorphic agent, named Hamilton in our system (fig. 2), can change its position and appearance in the following ways.

[1]Research in the VIENA Project is partly supported by the Ministry of Science and Research of the Federal State North-Rhine-Westphalia under grant no. IVA3-107 007 93.

- **Translation and rotation**
  The agent can move in the horizontal plane and turn around the vertical axis. Gravity and collision resistance of solid bodys are taken into account. In this way the user can deal with the agent by using experiences acquired in the physical world.

- **Looking**
  The head of the agent can rotate left, right, up and down. Rotations around the vertical axis are possible up to an angle of 45 degrees, rotations around the horizontal axis have a maximum of 20 degrees. Besides the optical aspects, these restrictions avoid that the user loses orientation in the virtual environment. After a short time period the head turns back autonomously, so a special reference frame for the head is not important in our current system.

- **Pointing gestures**
  Another objective of the agent is the improvement of communication by way of pointing gestures. Therefore we implemented a gesture with one arm extended

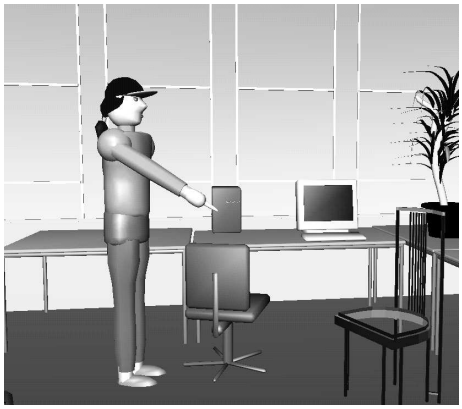

Figure 3: "Point to the designer-chair."

  and the index finger stretched (fig. 3). In comparison to other forms of body language, this gesture is easy to grasp in communication. Before pointing to an object, first the body of the agent rotates to the corresponding position to avoid a collision between arm and body. After a certain idle time the arm also turns back autonomously.

- **Changing the size**
  By the possibility of changing the body size of Hamilton, explorations become more flexible. Thus we have provided instructions that cause a shrinking or growing of the agent. Also an adaptation to the individual size of a user will soon be possible.

- **Hello**
  In addition to the pointing gesture, we also implemented a waving gesture. As an answer to the input "hello", the agent turns to the virtual camera, i.e. looks at the user, and a waving arm is seen for some seconds. These actions can also be evoked as an answer to a waving of the user (we use a simple data glove for this).

- **Changing the perspective**
  The system can also switch from an external view to a "situated" involved view by placing the virtual camera in the forehead of Hamilton. The virtual camera is positioned in such a way that a part of Hamilton's cap is visible without covering too much of the screen. To emphasize the change of perspective, the user can read "Hamilton View" in the upper part of the screen. Thus the user has two possibilities for exploring the virtual environment:

  – in the **external view** the agent is visible in the scene and can serve as an anchored allocentric reference frame, given by the three body axes: head/foot, front/back and right/left. The user can direct the agent to move in the scene and assess ergonomic features of the furnishings (e.g. size of a table) in comparison to the anthropometric features.

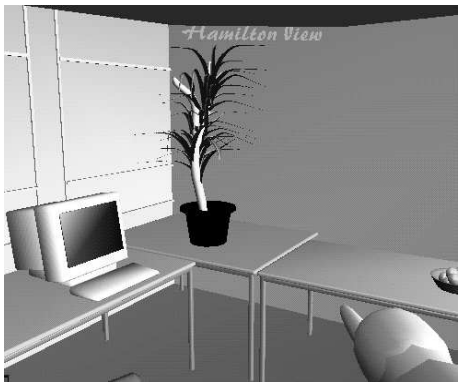

Figure 4: Looking "through the eye" of Hamilton (involved view)

  – in the **involved view** the user adopts the same perspective and field of vision as the anthropomorphic agent, so that s/he can be better immersed in the scene. Fig. 4 shows the view "through the eye" of Hamilton during a pointing gesture.

3 Communicating about Space

Communicating about space is very difficult in several aspects. One aspect is the structural difference between space and language: space is three-dimensional,

whereas language is based on the one-dimensionality of time [Friederici, 1989]. Another aspect is that speaker and listener often have different points of view. Thus it has to be clarified which actual reference frame has to be used and with which position and orientation [Graf & Herrmann, 1989]. Sometimes we get aware of these difficulties when describing route directions or the furnishings of a room. But most of the time we solve these problems by using contextual knowledge and gestures.

Since deictic references (like "here", "right", "in front of") play an important part in dialogues concerning space, we will focus on these topics in the following section and discuss their application in an interactive graphics system. We consider two types of locational references: *dimensional deixis* (directions, like "right", "left", "front", etc.) and *positional deixis* (positions, like "here" and "there").

3.1 Dimensional Deixis

Dimensional deixis is given by "up/down", "front/back" and "right/left". These terms normally indicate directions in three-dimensional space, depending on the position and the orientation of speaker and listener. Perception of the three dimensions is determined by biological and physical factors, e.g., by gravity and by the asymmetry of the human body. Using the left/right axis, confusion sometimes arises because the human body is nearly symmetrical in these directions.

Reference Systems

For an unambigious description of spatial relations, a reference frame must be given implicitly or explicitly, so that the utterances can be understood by the listener. When determining the reference frame, different coordinate systems can be chosen. The speaker can use deictic perspectives, where the origin is given by the position (1) of him- or herself, (2) of the listener or (3) of a third person. In each case, the axes are determined by the human perception of space [Levelt, 1986]. In addition to body position, there are also special reference frames for the eyes, the head and the upper part of the body, but most important are the position and orientation of the whole human body [Bühler, 1965].

As a further possibility, the speaker can make reference to a reference object (intrinsic perspective). Therefore the orientation of the reference object is given by its everyday use, e.g. the front side of a desk is the side where people normally sit.

In an utterance the speaker can explicitly mention the frame of reference ("from your position ..."). These expressions are used only if speakers are aware of possible ambiguities (e.g., if they are standing face to face so that changing of left and right happens frequently). But mostly situated knowledge, i.e. the knowledge about the actual situation helps to understand ambiguous utterances.

Application in VIENA

In the VIENA system different reference frames can be used. Hamilton can be directed to move from its (her? his?) own point of view or from the external viewpoint of the user. Depending on the reference frame taken on, translations in

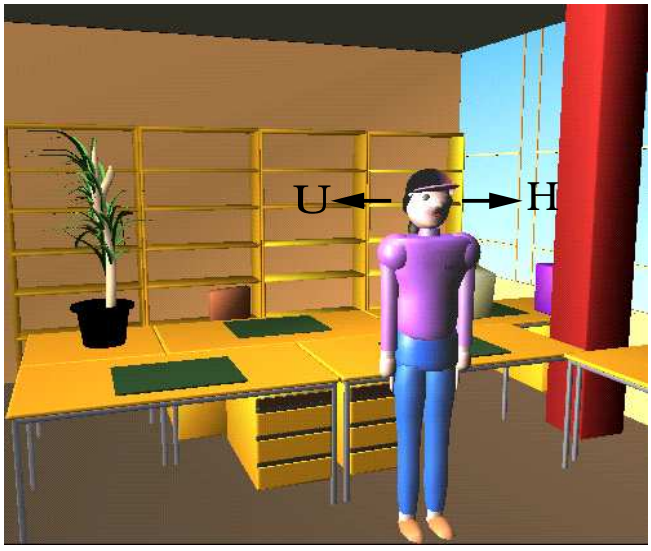

Figure 5: "Go left." U: user deictic view, H: Hamilton deictic view

space are carried out differently, especially if the virtual agent and the user are face to face. For this case, fig. 5 shows the two possibilities for realizing the instruction *go left*. When transforming objects, the user sometimes has to choose between three reference frames. Fig. 6 shows possible realizations for the instruction *move the bowl to the right*. From the intrinsic point of view, the bowl would be moved to position I, from the deictic view of Hamilton, it would be moved to position H, and the deictic view of the user is realized when objects are moved to position U.

Empirical findings

Selecting a reference frame depends on various aspects, e.g., geometric factors (angle and distance between Hamilton, user and reference object), optical factors (visibility and accessibility) and contextual factors (social situation and application domain). To get an impression of which reference frame would be chosen in an actual situation, we conducted an empirial study with 62 subjects. Each subject had to stand in a distance of about 2 meters and an angle of 45 degrees to a desk on which a coffeepot was located (fig. 7). Subjects then got the instruction *move the coffeepot to the left* and had to move the coffeepot to a new position. Results are shown in the table below.

Figure 6: "Move the bowl to the right." I: intrinsic view, U: user deictic view of, H: Hamilton deictic view

Position 1 (intrinsic)	Position 2 (deictic)
22 subjects (35.5 %)	40 subjects (64.5 %)

Results show that in our application domain both reference systems are relevant.

To get a deeper understanding when to use a certain reference frame, we evaluated relevant literature on similar experiments. The results are vague and sometimes contradictory. [Miller & Johnson-Laird, 1976] conclude that the intrinsic reference frame is easier to use than the deictic one. [Ehrich, 1992] considers descriptions of the furnishings of a room and finds out that most of the people use the deictic

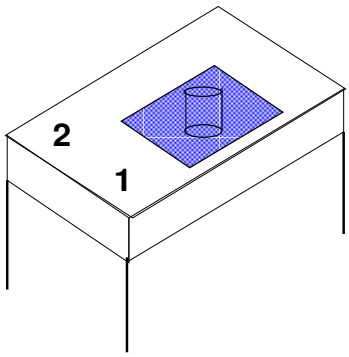

Figure 7: Setting for the empirical experiment

point of view. [Wunderlich, 1981] postulates that in static situations the intrinsic frame of reference is more frequently used. Our study, on the other hand, shows a significant preference for the deictic point of view, but also that the intrinsic one cannot be ignored, at least in our setting. This indicates that the selection of the actual reference frame is highly influenced by contextual factors.

More recently [Schober, 1995] explored how people choose spatial perspectives when they have an actual or an imaginary listener. He proposes that in human conversations more perspectives may be at work than most researchers have distinguished, especially when speaker and addressee don't share viewpoints. He looked at speaker-centered, addressee-centered, both-centered, object-centered, extrinsic and neutral perspectives. The study's results showed the importance of interactive feedback from the listener, because in discourse speakers' primary goal is to be understood. For this reason speakers sometimes took the addressee's perspective rather than their own (egocentric) perspective. Another interesting feature was the frequent use of both-centered and neutral descriptions that did not require to take one person's perspective. They may have wanted to minimize their own and their partner's effort. [Schober, 1995 p.153] says:

> "...human conversational partners were highly eglitarian in their perspective choices: (...) It is an open question what the optimal relationship for systems and users is. (...) I propose that rather than (or in addition to) trying to generate the perfect expression or create perfect interpretations, we might build in an architecture for accepting understanding and repairing misunderstandings."

Considering these empirical findings the choice of a reference frame seems by far no means decisive; all perspectives may be relevant. To make predictions, a variety of contextual factors would have to be considered. In addition, the individual perception of the situation and the individual use of language are very important. This means that extracting general rules is obviously not very useful. Instead we need a flexible system, which takes into account all possible frames of reference.

Our realization

In the VIENA system, we consequently realized three instances of the space agent to calculate the transformation of objects (fig. 2). One instance uses the intrinsic view for its calculation, the two other use the deictic view of the user and the deictic view of Hamilton. The Hamilton agent which calculates the movements of the anthropomorphic figure is realized in two instances, using the two deictic reference frames mentioned.

Evaluating an instruction, the system first carries out a transformation from the deictic point of view. If this realization does not meet the expectation of the user, s/he can correct the system (*"wrong"*). The system then generates a solution where a different agent instance has computed the transformation. Based on these

different instances of agents, further work deals with adaptation to individual users' preferences [Lenzmann & Wachsmuth, 1995].

Another advantage of this realization is the improved robustness of the system. If one instance of the agent cannot carry out the task (e.g., there is not enough unoccupied space at the goal position), another instance can be activated and possibly find a solution. In the near future we want to integrate simple hand gestures (issued by use of a data glove) to help resolve ambiguities. Referring to fig. 5, the instruction *go left*, combined with a hand gesture to the right, clearly indicates a movement from Hamilton's point of view.

3.2 Positional Deixis

In the English language the adverbs for positional deixis are *here* and *there*. They indicate positions in the 3-dimensional space depending on the position of the speaker. Because of their varied use in the language, interpretation of these adverbs is quite difficult. The relevant reference frame must be known, which can be complicated by different place and time of speaking and listening. The origin of the coordinate system can be moved by a pointing gesture or by a verbal expression, and one can also refer to abstract places. The regions can be expanded differently and can overlap [Klein, 1978]. In the literature, e.g. [Bühler, 1965], the following characterization is often found.

- *Here* is a region including the place of speaking.
- *There* is a region excluding the place of speaking. In the German language there are actually two meanings of *there* (*da* and *dort*) which refer to a shorter or wider distance between speaker and the indicated region.

Interpretation of positional adverbs in the VIENA System

In the VIENA system, communicating about space is restricted to simple instructions about the transformations of objects. When moving furniture in the virtual environment, the user refers to particular regions in the visible room. Regions outside of the visible room or abstract regions are not relevant. This limited discourse context makes the use of *here* and *there* possible. In the following section we describe possible interpretations of these adverbs. They are only suggestions of the system which can be corrected (i.e., negotiated) by the user in further interaction.

"Move the chair here!"

- From the deictic view, *here* usually refers to the user's own position. Because of the different perspectives the user can assume in the VIENA system, it has to be clarified where the user "feels" to be. In the external view, the user's position is formally given by the virtual camera which determines the

current field of vision. On the other hand, the user can take on the view of the anthropomorphic agent. Then s/he identifies the position of the agent with his/her own. If the user changes to the involved view, there is only one possible reference frame. In response to the instruction mentioned above, the chair would be moved near the anthropomorphic agent or near the virtual camera, that is, to the front of the screen.

- In addition, the region of *here* can be displaced by a pointing gesture. The region the user wants to indicate can be seen in the direction of the pointing arm [Ehrich, 1992].

"Move the desk there!"

- The verbal expression *there* is most often combined with different forms of gesture, e.g., facial expression, or pointing with arm or finger. In the VIENA system the anthropomorphic agent can carry out a pointing gesture with its right arm. A subsequent expression *there* can indicate a region in the direction of the pointing arm. At this time, gestures can only be produced by a verbal expression. In future more flexibility can be given by controlling the pointing arm with the mouse, so that describing regions in the 3-dimensional space could become easier.

- If no pointing gesture is issued and the agent is visible in the scene, *there* would refer to the position of the anthropomorphic agent (*there*, used by the speaker, has mostly the same meaning as *here* for the listener [Bühler, 1965].) In the following instructions, one can easily imagine this interpretation of *there*:

  "Hamilton, go left,
  a bit more,
  move the desk there."

- Another clue for locating the position *there* can be the line of vision of the user. When having the involved view, the user can move in the virtual room looking "through Hamilton's eyes".

  "Hamilton, go a bit backwards,
  look left,
  move the desk there."

  In particular when seeing a large part of the room, interpretation of such instructions is very vague. In real communication, the speaker would use a short sign with the head or the eyes. But in a graphics system, these kinds of gestures are too difficult to understand and not supported by our current system.

Our Realization

Computing instructions like *go there* or *come here* needs information from different agents. The bookkeeper has knowledge about preceding gestures, the Hamilton agent has the spatial knowledge to compute a goal position. There are two instances of the Hamilton agent which refer to different reference frames. After getting an instruction which contains *here* or *there*, one agent instance asks the bookkeeper for information. Looking up the database, the bookkeeper can decide if the adverb refers to a reference object because of a preceding gesture. Otherwise the position of the camera or Hamilton becomes relevant in the way mentioned above.

```
here (perspective, gesture)
 = reference_object, if gesture = TRUE;
   camera,           if gesture = FALSE and
                        perspective = deictic_view_of_user;
   hamilton,         if gesture = FALSE and
                        perspective = deictic_view_of_hamilton;


there (perspective, gesture)
 = reference_object, if gesture = TRUE;
   hamilton,         if gesture = FALSE and
                        perspective = deictic_view_of_user;
   indefinite,       if gesture = FALSE and
                        perspective = deictic_view_of_hamilton;
```

Getting a relevant position, Hamilton realizes the task *Go to the* < *object* >. In our current realization, the adverbs *here* and *there* are represented only as "zero objects", that is, as positions without an extended region. The actual goal position to which an object is moved is determined such that the object is placed as close as possible to the computed zero position, constrained by detected collisions, etc.

4 Discussion and Future Work

Spatial dialogues increasingly attract attention in different research areas. In this paper we presented an anthropomorphic agent for a graphics system to add comfort in the human-computer interaction, in particular, with respect to spatial language.

Introducing the visible agent, we illustrated its possibility to move in the virtual room and to carry out pointing and waving gestures. Besides psychological reasons (the human-like figure should encourage the use of natural language), our main interest is the improvement of situated spatial communication. Focusing on deictic expressions, we investigated dimensional adverbs and the use of different reference frames which are dealt with by several instances of agents in our realization.

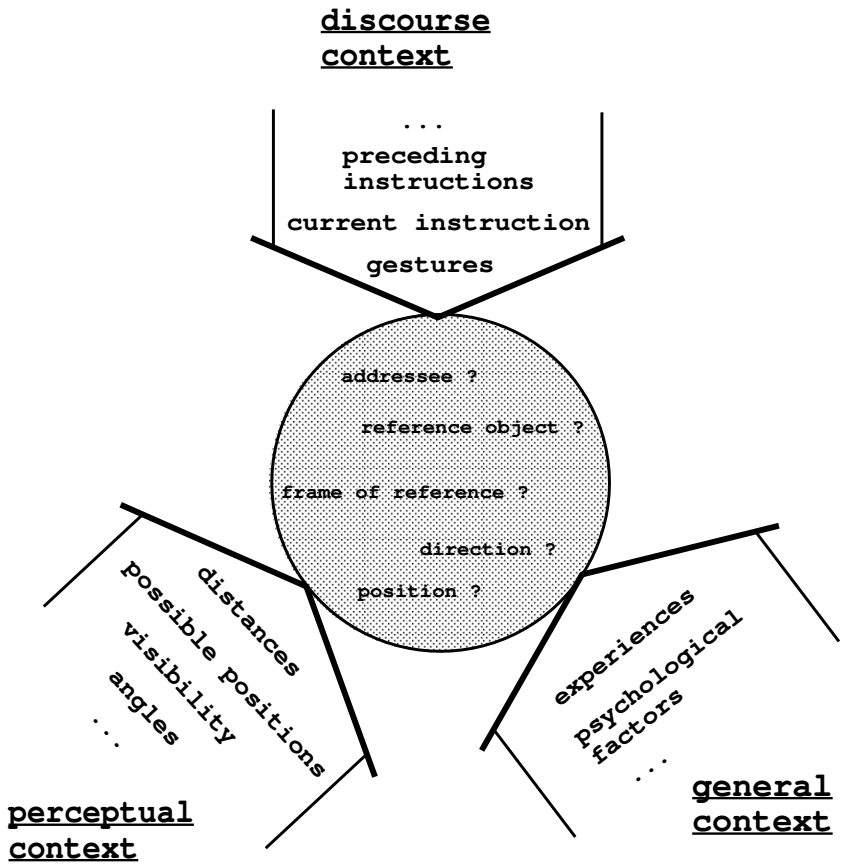

Figure 8: Contextual factors

With these instances all possible expectations of the user (as known to us by now) concerning the actual reference frame are taken into account and can be visualized by the system. In addition, the use of positional adverbs (*here* and *there*) becomes possible by the perception of a human-like figure. Instructions of the user can refer to the position and orientation of the anthropomorphic agent and can use pointing gestures of the agent to indicate positions or objects in 3D-space.

Considering the results presented in this paper, we realized that a human-like processing of spatial expressions requires a large amount of situational knowledge. People use pronouns to indicate the adressee, vague descriptions for the different reference objects and qualitative positions and directions. Therefore we are currently working on possibilities to regard the influence of contextual factors.

Giving a fuller account of contextual factors seems a promising area of future research. In the center of fig. 8 we show some spatial aspects which are frequently used ambiguously. In the three arrows there are some influence factors, divided in perceptual, discourse, and general context. Some of these may be important for the

processing of spatial expressions. In further work, this information might also be used to automatically adapt the system to individual user preferences.

References

[Bühler, 1965] Bühler, K. (1965). *Sprachtheorie: Die Darstellungsfunktion der Sprache.* Stuttgart: Fischer.

[Cao & Wachsmuth, 1993] Wachsmuth, I., Cao, Y. (1993). Situated Space Agent for 3-D Graphics Design. Extended abstract in *Virtual Reality Viena 1993, the Global VR-Focus in Europe - Abstracts (pp. 8-9).* Long version: KI-NRW (Applications of Artificial Intelligence in North-Rhine Westphalia) Report 93-13.

[Ehrich, 1992] Ehrich, V. (1992). *Hier und Jetzt, Studien zur lokalen und temporalen Deixis im Deutschen.* Linguistische Arbeiten. Tübingen: Niemeyer.

[Friederici, 1989] Friederici, A. D. (1989). Raumreferenz unter extremen perzeptuellen Bedingungen: Perzeption, Repräsentation und sprachliche Abbildung. In Ch. Habel, M. Herweg, K. Rehkämper (Hrsg.): *Raumkonzepte in Verstehensprozessen*, 17-33. Tübingen: Niemeyer.

[Gapp, 1995] Gapp, K.-P. (1995), An Empirically Validated Model for Computing Spatial Relations, In I. Wachsmuth et al. (Eds.) *KI-95: Advances in Artificial Intelligence*, 245-256, Springer-Verlag.

[Graf & Herrmann, 1989] Graf, R., Herrmann, T. (1989). *Zur sekundären Raumreferenz: Gegenüberobjekte bei nicht-kanonischer Betrachterposition.* Bericht Nr. 11, Arbeiten aus dem SFB 245 "Sprechen und Sprachverstehen im sozialen Kontext", Heidelberg/Mannheim.

[Herrman & Grabowski, 1987] Herrmann, T., Grabowski, J. (1987). *Sprechen: Psychologie der Sprachproduktion.* Spektrum Akademischer Verlag

[Klein, 1978] Klein, W. (1978). Wo ist hier? In *Linguistische Berichte 58*, 18-40.

[Lenzmann & Wachsmuth, 1995] Lenzmann, B. & Wachsmuth, I. (1995). A User-Adaptive Interface Agency for Interaction with a Virtual Environment. in *Working Notes of the Workshop: Adaptation and Learning in Multi-Agent-Systems*, 43-46, IJCAI 95 in Montreal.

[Levelt, 1986] Levelt, W. (1986). Zur sprachlichen Abbildung des Raumes: Deiktische und intrinsische Perspektive. In H.-G. Bosshardt (Hrsg.): *Perspektiven auf Sprache*, 187-211. Berlin: DeGruyter.

[Miller & Johnson-Laird, 1976] Miller, G.A., Johnson-Laird, P.N. (1976). *Language and Perception*. Cambridge University Press, 1976.

[Olivier et al., 1994] Olivier, P., Toshiyuki, M., Jun-ichi, T. (1994), Automatic Depiction of Spatial Descriptions, in *Proceedings of the 12th National Conference on Artificial Intelligence (AAAI-94)*, 1405-1410.

[Retz-Schmidt, 1988] Retz-Schmidt, G. (1988). Various Views on Spatial Prepositions. In *AI Magazine, 9(2)*, 95-105.

[Schober, 1995] Schober, M.F. (1995), How Adressees Affect Spatial Perspective Choice in Dialogue, in *Working Notes of the Workshop: Representation and Processing of Spatial Expressions*, IJCAI 95 in Montreal.

[Sichelschmidt, 1989] Sichelschmidt, L. (1989). Wo hier dort ist - primär- und sekundärdeiktische Raumreferenz. In Ch. Habel, M. Herweg, K. Rehkämper (Hrsg.): *Raumkonzepte in Verstehensprozessen*. Tübingen: Niemeyer.

[Wachsmuth & Cao, 1995] Wachsmuth, I., Cao, Y. (1995). Interactive Graphics Design with Situated Agents. In W. Strasser, F. Wahl (eds.): *Graphics and Robotics*, 73-85, Berlin, Heidelberg, New York: Springer.

[Wunderlich, 1981] Wunderlich, D. (1981). Linguistic Strategies. In F. Coulmas (ed.): *A Festschrift for Native Speaker*, 279-296, New York: Mouton.