To appear in K.R. Coventry & P. Olivier (eds.), Spatial Language: Cognitive and Computational Aspects. Dordrecht: Kluwer, 2001.

Chapter 4

An Anthropomorphic Agent for the Use of Spatial Language

Tanja Jörding and Ipke Wachsmuth
Dresden University of Technology, Germany and University of Bielefeld, Germany

In this paper we describe 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 multi-agent 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 visualised in the scene. With reference to 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 a user can concentrate his/her imagination 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 expressions in natural language into quantitative commands in a graphical system. A particular aim is the processing of verbal spatial expressions. We developed an articulated anthropomorphic agent which is the visible instance of the VIA (see Figure 1). This agent can encourage the use of natural language and can be conceived of 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 communication by situated spatial instructions.

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. In order to facilitate smooth interaction, the flexible use of different perspectives must be possible. Some natural language systems (e.g., Retz-Schmidt, 1988; Olivier et al., 1994) consider deictic and intrinsic frames of view. An additional frame of reference, for example, addressee-centred, is considered by Schober (1995). Recently, Gapp presented an approach to the computation and empirical evaluation of the meanings of basic spatial relations in 3D space (Gapp, 1995). His main emphasis was on clarifying the dependencies between angle, distance and shape with respect to simple idealised objects.

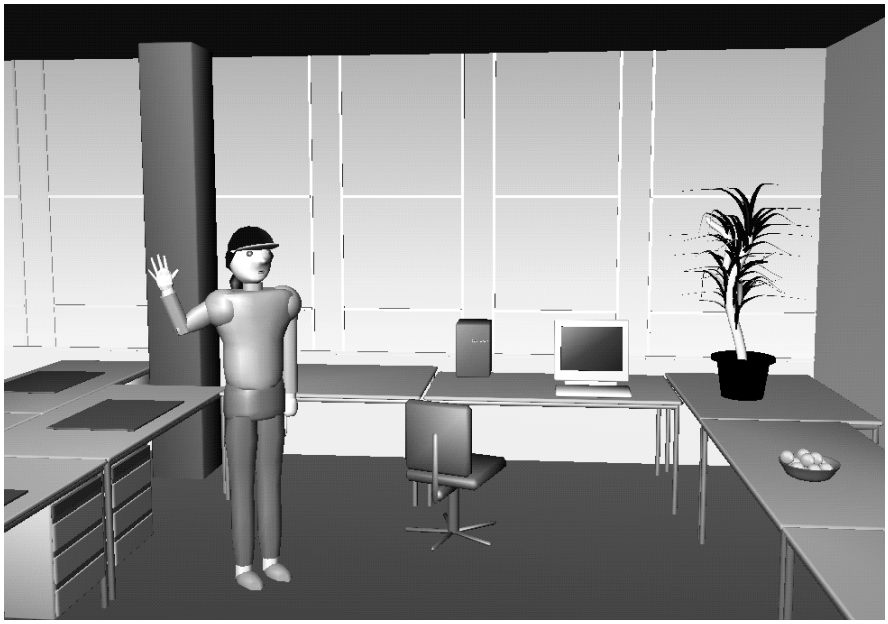

Figure 1. Example scene: the anthropomorphic agent in a virtual office room

In our project we consider a more complex setting with a visible listener, where spatial issues become more realistic. With a human-like figure and some kind of gestures there are better possibilities of simulating natural discourse, for example, 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 section we first describe 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. In the concluding section, we discuss our ideas and give an outlook on future work.

2. THE VIENA PROJECT

VIENA[1] ("Virtual Environment & Agents") is a project within the research focus theme "Artificial Intelligence and Computer Graphics" at the University of Bielefeld. The overall goal is to develop an intelligent form of communication with a virtual environment (Wachsmuth & Cao, 1995). 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 (see Figure 2), which altogether form a multimedia user interface, translate qualitative instructions from the human user into quantitative technical commands that update the visualisation scene model.

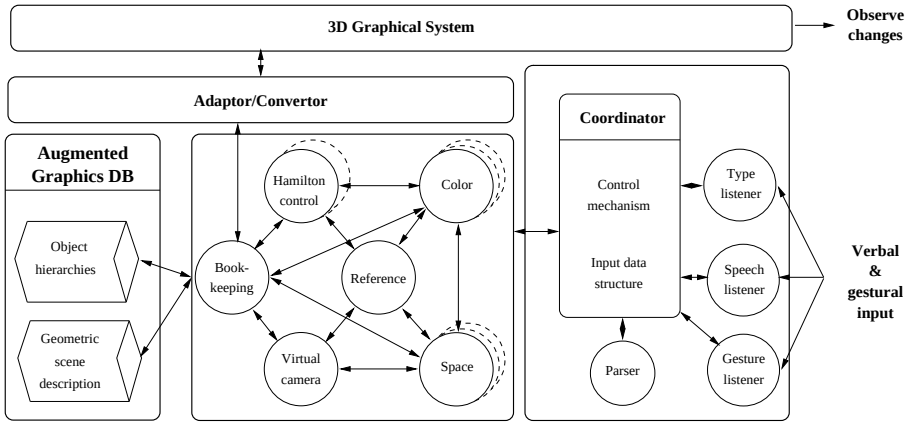

Figure 2. The architecture of the VIENA system (Lenzmann & Wachsmuth, 1997)

Instructions are issued via a multimodal input agency that combines input from different modalities. It consists of input listeners, a parser, and a co-ordinator. Three listener agents, the type listener, speech listener, and gesture listener, track and analyse sensor data from the keyboard, the microphone,

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

and the data glove, respectively. With the help of the parser, the co-ordinator analyses and integrates the inputs received from the listeners and generates an internal task description that is posted to the appropriate agents of Hamilton's agency. Hamilton is the anthropomorphic agent in the visual scene.

In mediating an instruction, invisible agents in the VIENA system track exact object locations and colourings, and they negotiate alternative ways of acting. For example, a space agent computes spatial transformations in the virtual environment such as translating, rotating, and scaling of scene objects. By inspecting or modifying RGB-vectors, a colour agent helps to identify an object by means of a colour description ("the red chair") or to change the appearance of objects (e.g., blue, lighter). A camera agent calculates transformations of the virtual camera to enable the navigation through the scene. To resolve ambiguous references in the qualitative instruction, a reference agent determines a ranked list of candidate reference scene objects. A Hamilton control agent realises the manipulation of the articulated figure. A bookkeeping agent is authorised to access and modify the augmented graphics database to supply current situation information to agents on request.

Some of these agents are realised as agencies, that means there are two or three instances of the same agent-type with a slightly varied functionality.

In the visual scene Hamilton, the anthropomorphic agent, can move around and change its appearance in the following ways;

- **Translation and rotation**
  The agent can move in the horizontal plane and turn around along its vertical axis. Gravity, and the collision resistance of solid bodies, 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 ensure that the user avoids loosing orientation in the virtual environment. After a short period of time the head turns back automatically, such that a special frame of reference 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 and the index finger stretched (see Figure 3). In comparison to other forms of body language, this gesture is easy to grasp

in communication. Before pointing, the agent turns around to view the object. After a certain idle time the arm also turns back automatically.

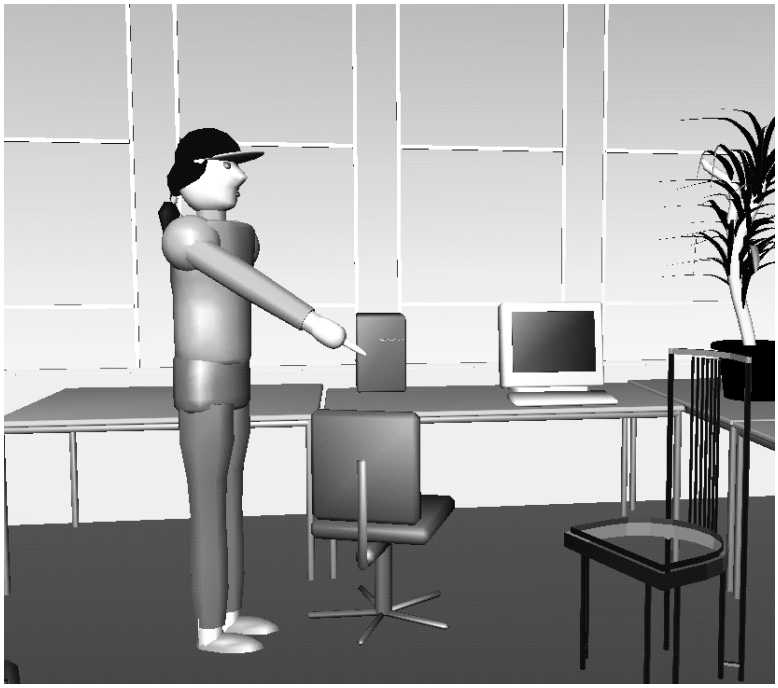

Figure 3. "Point to the designer-chair"

– **Changing the size**
  With 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. An adaptation to the individual size of the user could be included if the application demands for it.
– **"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 (e.g., looks at the user) and a waving arm is seen for some seconds (Figure 1). These actions can also be evoked in response to the user waving (we use a simple data glove for this).

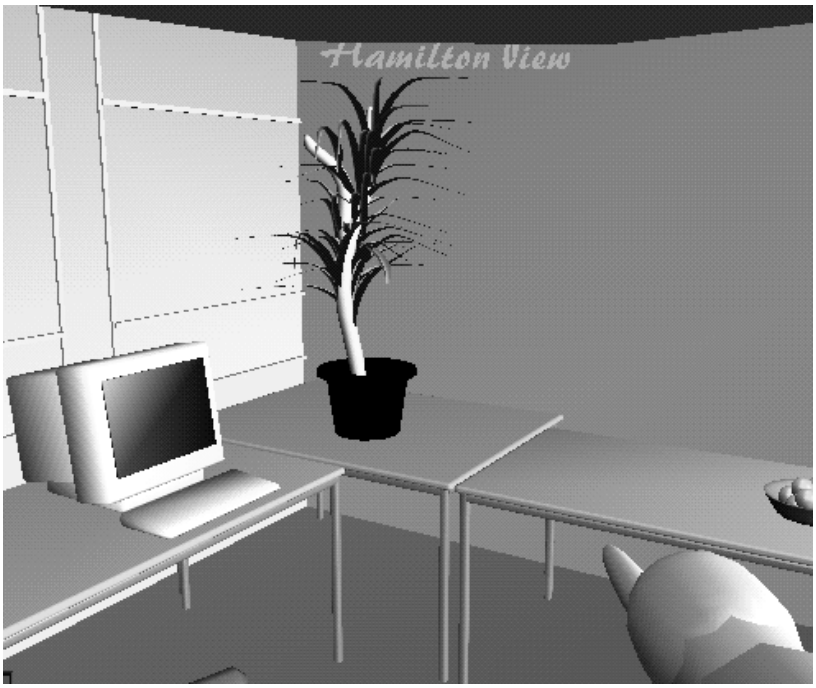

Figure 4. Looking "through the eyes" of Hamilton (involved view)

– **Changing the perspective**
  The system can also switch from an external view to a “situated” (or 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. 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 frame of reference, 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.
  – In the **involved view** the user adopts the same perspective and field of vision as the anthropomorphic agent, such that s/he can better immerse her/himself in the scene. Figure 4 shows the view “through the eyes” of Hamilton during a pointing gesture.

3. COMMUNICATING ABOUT SPACE

Sometimes in everyday life we become aware that communicating about space is very difficult in several respects, for example, when describing route directions or the furnishings of a room. 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 what the actual frame of reference is and what its position and orientation are (Graf & Herrmann, 1989). 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 subsection and discuss their application in an interactive graphics system. We consider two types of location references: *dimensional deixis* (directions, like "right", "left", "front", etc.) and *positional deixis* (positions, like "here" and "there").

3.1 Dimensional Deixis

Dimensional deixis is facilitated through "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, for example, 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.

3.1.1 Reference Systems

For an unambiguous description of spatial relations, a frame of reference must be given implicitly or explicitly, so that utterances can be understood by the listener. When determining the frame of reference, 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 frames of reference 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 can be established by its everyday use. For example, the front side of a desk is the side where people normally sit. Depending on the front side, the right and left axes can be structured in two ways, called facing or aligning modality (Hill, 1982).

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, that is, the knowledge about the actual situation, helps to understand ambiguous utterances.

3.1.2 Empirical findings

Selecting a frame of reference depends on various parameters, for example, geometric factors (angle and distance between Hamilton, user and reference object), visual factors (visibility), accessibility and contextual factors (social situation and application domain). To get an impression of which frame of reference would be chosen in an actual situation, we carried out a simple experiment with 62 subjects. Each subject had to stand at a distance of about 2 meters and an angle of 45 degrees in front of a desk on which a coffee-pot was located (see Figure 5).

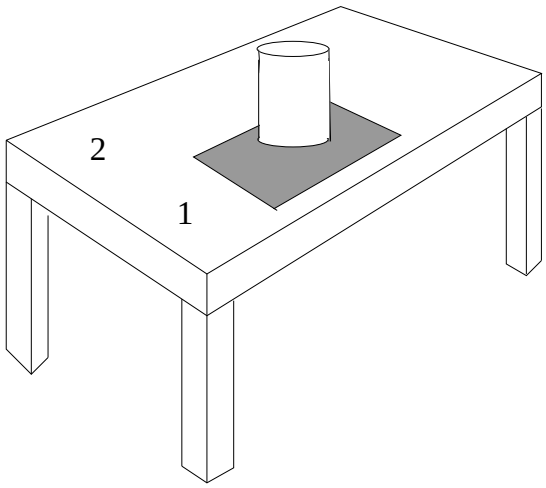

Figure 5. Setting for the experiment (see text)

Subjects then got the instruction *move the coffee-pot to the left* and had to move the coffee-pot to a new position. It turned out that about one-third (22) of the subjects placed the pot in position 1 (see Figure 5) and the other two-

thirds (40) of them placed the pot in position 2 on the table. First of all, this shows that there is a significant variation of preferences among subjects. Secondly, we are inclined to judge that the intrinsic left-right orientation of the table influenced those subjects choosing position 1 while the other subjects apparently chose position 2 from a deictic perspective. In any case we may conclude that there is no "best" solution but that the observed individual differences need to be taken into account.

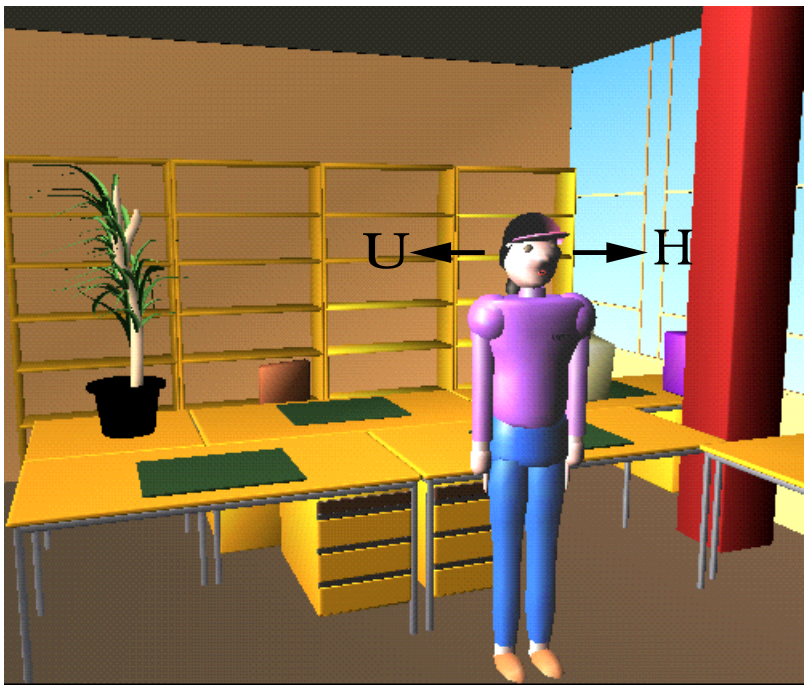

Figure 6. "Go left" U: user deictic view, H: Hamilton deictic view

3.1.3 Application in VIENA

In the VIENA system different frames of reference can be used. Hamilton can be directed to move from its own point of view or from the external viewpoint of the user. Depending on the frame of reference, translations in space are carried out differently, especially if the virtual agent and the user are face to face. Figure 6 shows the two possibilities for realising the instruction *go left* for this case.

When transforming objects the user sometimes has to choose (without becoming fully aware of it) between three frames of reference. Figure 7 shows possible realisations for the instruction *move the bowl to the right*. From the intrinsic point of view (imposed by intrinsic features of the desk),

Figure 7. "Move the bowl to the right"
I: desk intrinsic view, U: user deictic view, H: Hamilton deictic 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 realised when objects are moved to position U.

To get a deeper understanding of when to use a certain frame of reference, we evaluated relevant literature using similar experiments. The results are vague and sometimes contradictory. Miller and Johnson-Laird (1976) conclude that the intrinsic frame of reference is easier to use than the deictic one. Ehrich considered descriptions of the furnishings of a room and found that most people use the deictic point of view (Ehrich, 1992). Wunderlich postulates that in static situations the intrinsic frame of reference is more frequently used (Wunderlich, 1981). On the other hand, our study shows a significant preference for the deictic point of view, but also that the intrinsic one cannot be ignored, at least, not in our specific setting. This indicates that the selection of the actual frame of reference is highly influenced by contextual factors.

More recently, Schober explored how people choose spatial perspectives when they have an actual or an imaginary listener (Schober, 1995). He proposed 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-centred, addressee-centred, both-centred, object-centred, extrinsic, and neutral perspectives. The study's results showed the importance of interactive feedback from the listener, because in discourse the 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-centred and neutral descriptions that did not require taking one person's perspective. They may have wanted to minimise the effort expended by both themselves their partner. Schober stated:

> ...human conversational partners were highly egalitarian 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. (Schober, 1995) p.153

Considering these observations the choice of a frame of reference seems by 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. Consequently extracting general rules is obviously not very useful. Instead we need a flexible system, which takes into account all possible frames of reference.

3.1.4 Our Realisation

In the VIENA system, we consequently consider three instances of the space agent to calculate the transformation of objects. One instance (when appropriate) uses an object 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 realised in two instances, using the two deictic frames of reference mentioned.

Evaluating an instruction, the system first carries out a transformation from the deictic point of view. If this realisation does not match the expectation of the user, s/he can correct the system by stating *"wrong"*. The system then generates a solution where a different agent instance computes the transformation. Based on these different instances of agents, further work deals with adaptation to individual users' preferences (Lenzmann & Wachsmuth, 1997).

Another advantage of this realisation 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 can possibly find a solution. Further work is aimed at integrating simple hand gestures (issued by use of a data glove) to help resolve ambiguities. With reference to Figure 6, 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 frame of reference must be known, which can be complicated by different places and times 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 may overlap (Klein, 1978). In the literature (for example, Bühler, 1965), the following characterisation 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.

3.2.1 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 we describe possible interpretations of these adverbs. They are only suggestions by the system which can be corrected (e.g., 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" s/he should 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 reference to his/her own. If the user changes to the involved view, there is only one possible frame of reference. In response to the instruction mentioned above, the chair would be moved near the anthropomorphic agent or near the virtual camera, that is, toward 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, for example, 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. The positional *there* can also be complemented by a pointing gesture issued by the user from "outside", using a data glove. Thus describing regions or specifying objects in 3-dimensional space becomes 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 view 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 give a short hint with the head or the eyes. But in a graphics system, these kinds of gestures are so far difficult to understand and not supported by our current system.

3.2.2 Our Realisation

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 frames of reference. 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 realises the task *Go to the <object>*. In our current realisation, 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. In future work it is desirable to consider not only preceding gestures but also preceding interactions.

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 to the human-computer interaction, in particular, with respect to spatial language.

Introducing the visible agent, we illustrated its ability to move in the virtual room and to carry out pointing and waving gestures. Aside from psychological motivations (as an addressee, the human-like figure should encourage the use of natural language), our main interest is the improvement of situated spatial communication.

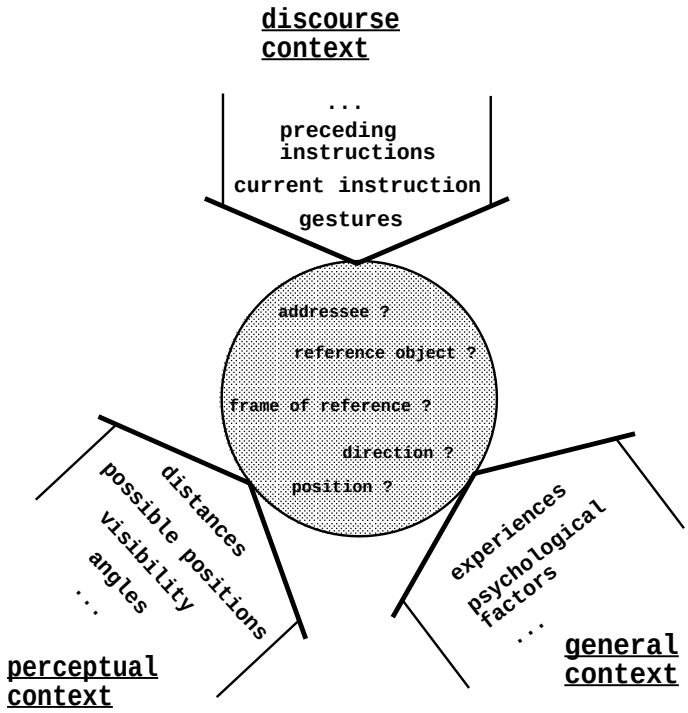

Figure 8. Contextual factors

Focusing on deictic expressions, we investigated dimensional adverbs and the use of different frames of reference which are dealt with by several instances of agents in our realisation. With these instances all possible expectations of the user (known to us by now) concerning the actual reference frame are taken into account and can be visualised by the system.

In addition, the use of positional adverbs (*here* and *there*) becomes possible through the perception of a human-like figure. Instructions from 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 realised that a human-like processing of spatial expressions requires a large amount of situational knowledge. People use pronouns to indicate the addressee, vague descriptions for the different reference objects and qualitative positions and directions. Therefore further work should be directed toward possible ways of taking the influence of contextual factors into account.

Giving a fuller account of contextual factors seems a promising area for future research. Figure 8 illustrates some of these factors. The centre of Figure 8 we show some spatial components that are frequently used ambiguously. The three arrows illustrate some contextual factors, divided into 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, K. (1965). *Sprachtheorie: Die Darstellungsfunktion der Sprache*. Stuttgart: Fischer.

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

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

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

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.

Hill, C. (1982). Up/down, front/back, left/right. A contrastive study of Hausa and English. In J. Weissenborn & W. Klein (Eds.), *Here and There. Cross-linguistic Studies on Deixis and Demonstration*. Amsterdam/Philadelphia: Benjamins.

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

Lenzmann, B. & Wachsmuth, I. (1997). Contract-Net-Based Learning in a User-Adaptive Interface Agency. In Weiss, G. (Ed.), *Distributed*

Artificial Intelligence Meets Machine Learning: Learning in Multi-Agent Environments (pp. 202-222). Berlin: Springer (LNAI 1221).

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

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

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, pp. 1405-1410). Seattle, WA: American Association for Artificial Intelligence.

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

Schober, M.F. (1995). How Addressees Affect Spatial Perspective Choice in Dialogue. In *Working Notes of the Representation and Processing of Spatial Expressions Workshop. International Joint Conference on Artificial Intelligence* (IJCAI-95), Montreal.

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

Wunderlich, D. (1981). Linguistic Strategies. In F. Coulmas (Ed.), *A Festschrift for Native Speaker* (pp. 279-296). The Hague: Mouton.