Projective relations for 3D space: Computational model, application, and psychological evaluation*

Constanze Vorwerg Gudrun Socher° Thomas Fuhr Gerhard Sagerer Gert Rickheit

Universität Bielefeld, SFB 360 "Situierte Künstliche Kommunikatoren", Postfach 100131, 33501 Bielefeld, Germany
°current address: California Institute of Technology, 136-93, Pasadena, CA 91125, USA
e-mail: {vorwerg,rickheit}@nov1.lili.uni-bielefeld.de, {gudrun,fuhr,sagerer}@techfak.uni-bielefeld.de

Abstract

We propose a 3D computational model for projective relations which is used in an integrated image and speech understanding system. The image and speech understanding system is being developed within a joint research project focusing on both technical and cognitive aspects of human-computer interaction. Psychological experiments have been carried out to evaluate our computational model as an approximation of the meaning of projective prepositions used by humans in spoken instructions. These experiments investigate the acceptance of the model by subjects as well as the regularities regarding human usage of projective relations. Results of the computational model, the overall system, and the psychological experiments are presented.

Introduction

In this paper we present a computational model for projective relations in 3D together with its application and results from psycholinguistic experiments. The experiments investigate the use of projective relations in spoken instructions and they form an empirical basis for the discussion and evaluation of the computational model. Projective relations are spatial relations that depend on a particular perspective or point of view (e.g. *left*; cf. Herskovits, 1986).

The computational model is motivated by the joint research project "Situated Artificial Communicators" at the University of Bielefeld which aims for the development of an integrated system with interacting visual, linguistic, sensory-motoric, and cognitive abilities. The system is supposed to understand and to carry out instructions which are given by a human. The system is equipped with a stereo camera and the main task is to relate the verbal instructions and the observed scene, i.e. the joint understanding of visual and speech input data. The underlying scenario is the cooperative assembly of toy airplanes using the Baufix®[1] construction kit.

Spatial expressions including projective relations are essential for communication in this scenario. The design of our computational spatial model is based on technical requirements according to the scenario as well as on results of psycholinguistic experiments and cognitive science. It is developed along the following guidelines:

- Localizations in 3D space are addressed in the construction scenario.
- The computational model must be able to cope with different frames of reference.
- Projective relations have vague and overlapping meanings (Hayward & Tarr, 1995; Herskovits, 1986).
- Object abstractions may be used for computational efficiency. These should preserve the extension and at least basic features of the shape of *both* the reference object (RO) and the intended object (IO). There is currently *no* strong psychological evidence that the objects' shape and extension do *not* influence the applicability of prepositions in certain configurations, although, e.g., Landau & Jackendoff (1993) are often interpreted that way. In contrast to other approaches (e.g. Gapp, 1994; Olivier & Tsujii, 1994), we do not abstract an object purely by its center of mass.

The computational model is outlined in the next section. The projective relations, which are implemented so far, are considered as models of the German projective prepositions *rechts* (*right*), *links* (*left*), *vor* (*in-front*), *hinter* (*behind*), *über* (*above*), and *unter* (*below*). The third section describes a prototype of a "situated artificial communicator" and the use of the spatial model within this system. The psychological investigations are addressed in the fourth section which is followed by a discussion focusing on the empirical results.

The 3D Spatial Model

In this section, we briefly describe our computation of the binary projective relations *right*, *left*, *in-front*, *behind*, *above*, and *below* between 3D objects in 3D space. A more detailed description of and discussion on other approaches can be found in (Fuhr et al., 1995).

Shape Abstraction and Space Partitioning

Instead of detailed geometric object models, we use surrounding boxes that are collinear to the objects' inertia axes

*This work has been supported by the German Research Foundation (DFG) in the project SFB 360.

[1]Wooden toy construction kit, see Fig. 2 for sample objects.

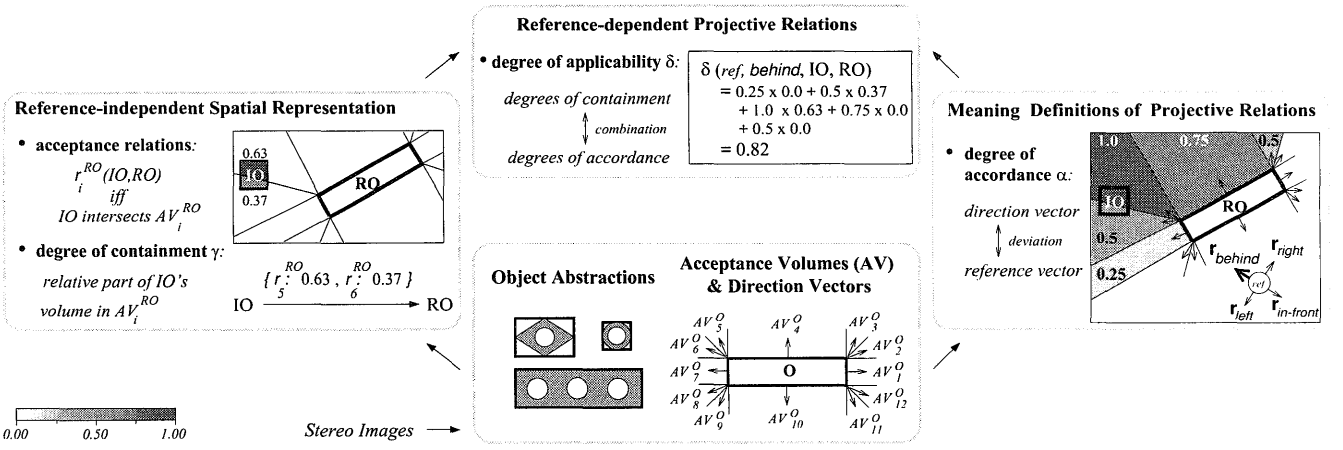


Fig. 1: Computation of scored projective relations for object pairs: The principles of the computation are demonstrated for 2D objects in 2D space. An object-specific partitioning of 2D space into 12 acceptance volumes is chosen. As an example the generation of the relation *behind* is shown for the objects IO and RO w.r.t. the reference frame *ref*.

(Abella & Kender, 1993) as abstractions. Although this is a rough shape abstraction, it is sufficient to investigate the influence of the objects' shape and extension to the use of projective prepositions. It must be pointed out that the principal computational structure of our approach is not bound to *this* sort of object abstractions. Other polyhedra and even object-specific polyhedra could be used as well. For simplicity, we currently use the same kinds of boxes for all kind of objects in the scenario. Throughout this section, we use "object" as a synonym for the object's box.

A finite number of acceptance volumes AV_i^O is associated with each object O. These are infinite open polyhedra bound to the sides, edges, and corners of the object. They partition the 3D space surrounding the object. A direction vector $\boldsymbol{d}(AV_i^O)$ corresponds to each acceptance volume. It roughly models the direction to which an acceptance volume extends in space. The object-specific partitioning is motivated by the assumption that the object itself may influence the way the surrounding space is perceived independently of specific reference frames[2].

Generation of Relations from Objects

The computation of relations from objects is a two-layered process. In the first layer, a *reference-independent spatial representation* is computed. Each acceptance volume induces a binary *acceptance relation* r_i^O that expresses whether an object P intersects with AV_i^O. Acceptance volumes are scored by calculating the corresponding *degree of containment*:

$$\gamma(P, r_i^O) = \frac{vol(P \cap AV_i^O)}{vol(P)}.$$

[2]The discrete partitioning is also motivated by the following requirement: The computational model shall allow us to infer possible image regions for the depiction of the IO with respect to a given RO under a given reference frame, that can be exploited by an object recognition component. However, we cannot discuss this issue in this paper (see Fuhr et al., 1995).

Thus, the relation between two objects P and O can be reference-*in*dependently expressed by a set of scored acceptance relation symbols with non-zero degree.

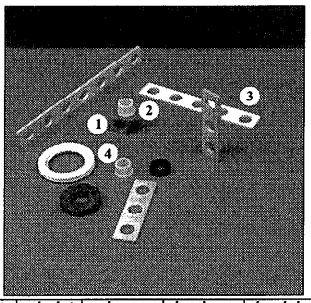
Furthermore, reference-*de*pendent *meaning definitions* of relations *rel* w.r.t. to certain ROs and a given reference frame $ref = \{\boldsymbol{r}_{right}, \boldsymbol{r}_{left}, \ldots, \boldsymbol{r}_{above}\}$ are also calculated in the first layer. The actual reference frame *ref* is assumed to be provided by other linguistic or cognitive components of our artificial communicator. *def*(*ref*,*rel*,RO) is given as the set of the symbols of all acceptance relations r_i^{RO} whose direction vector differs less than 90° from the corresponding reference vector $\boldsymbol{r}_{rel}$. The membership of an acceptance relation (symbol) to a meaning definition is scored by its *degree of accordance*:

$$\alpha(ref, rel, r_i^{RO}) = 1 - 2 \cdot \frac{\arccos(\boldsymbol{d}(AV_i^{RO}) \cdot \boldsymbol{r}_{rel})}{\pi}.$$

These two scored symbolic reference-independent and reference-dependent descriptions are the basis for the computation of reference-dependent relational expressions for IO-RO pairs in the second layer. The basic idea is, that the relation *rel* is applicable for an IO-RO pair w.r.t. a reference frame *ref* if at least one of the acceptance relations in *def*(*ref*,*rel*,RO) holds between IO and RO. Again the *degree of applicability* $\delta(ref, rel, \mathrm{IO}, \mathrm{RO})$ of *rel* varies gradually:

$$\delta(ref, rel, \mathrm{IO}, \mathrm{RO}) = \sum_{\substack{r_i^{RO} \in \\ def(ref, rel, RO)}} \alpha(ref, rel, r_i^{\mathrm{RO}}) \cdot \gamma(\mathrm{IO}, r_i^{\mathrm{RO}}).$$

Fig. 1 illustrates the steps of this computation. For easier visualization the steps are shown for 2D objects in 2D space. Furthermore, the table in Fig. 2 contains the relations and their applicability degrees computed for the scene shown in the image. The reference frame is assumed to be deictic and corresponding with the cameras' view of the scene. The table demonstrates that the results are very promising keeping in mind that they have been computed



IO	RO	left	right	above	below	behind	in-front
2	1	0.17	0.10	**0.51**	0	0.15	0.09
1	2	0.12	0.01	0	**0.86**	0.04	0.01
3	1	0	**0.60**	0.04	0.01	0.39	0
1	3	0.26	0	0.13	0.07	0	**0.64**
4	1	0	0.20	0.04	0.07	0	**0.75**
1	4	0.18	0.04	0.03	0.05	**0.74**	0
3	2	0	**0.52**	0	0.23	0.32	0
2	3	0.24	0	0.23	0	0	**0.65**
4	2	0.02	0.06	0	0.17	0	**0.79**
2	4	0.09	0.03	0.17	0	**0.78**	0
3	4	0	0.34	0	0	**0.65**	0
4	3	0.24	0	0.19	0	0	**0.66**

Fig. 2: Example of the projective relations and their degrees of applicability computed for the numbered objects from the scene shown in the image: The maximum applicability degree per RO-IO pair is highlighted in bold print. The chosen reference frame takes the position of the cameras as vantage point to allow for an easy verification of the results from the image of the scene.

from slightly erroneous 3D object poses reconstructed from *real* stereo image data.

Due to the fuzziness of our relation model several relations are usually applicable between an object pair. This points to another important issue: What are the relations most relevant to select? Obviously, nobody would specify the relation between an object pair using all projective prepositions. The table shows that in many cases a maximum decision yields good results. However, in ambiguous cases two relations might be chosen. We will return to this aspect later.

A Situated Artificial Communicator

Fig. 3 sketches our prototype of a "situated artificial communicator." It is a speech and image understanding system with the objective to mutually restrict the understanding processes through (intermediate) results from the respective other source of data. There are a number of other encouraging approaches towards the integration of vision and language (e.g. Nagel, 1988; Wahlster, 1989; Mc Kevitt, 1994). But most of this work concentrates either on the generation of scene descriptions from images or on the visualization of spoken phrases. In our system, the speech and image understanding modules extract qualitative descriptions from the speech and image data, respectively. A inference module, identifies the the intended object(s) and/or infers mutual restrictions for the understanding processes. The inference machine operates on the qualitative descriptions.

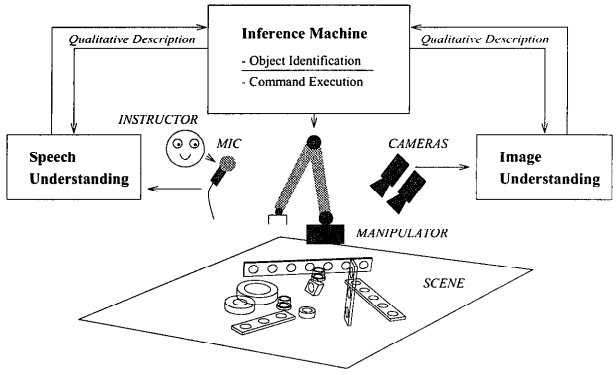


Fig. 3: Overview of the prototype of a *situated artificial communicator*: The speech and image understanding modules derive qualitative descriptions from acoustic and visual input data. The understanding processes work both data-driven and model-driven.

Object Identification

Objects in the scene are referred to by elliptic or complete noun groups, and possibly by projective terms. With the results from scene interpretation, utterances like "take the one behind the bar" can be sufficient to find the intended object. The object identification in our system is based on a common, qualitative description of the object's *type*, *color*, *size*, *shape*, and of *projective relations*. More details on the object recognition, color classification, and the size and shape assignment can be found in Socher et al. (1996). The instructions are given in real speech and the speech understanding is described in Socher & Naeve (1996).

All qualitative features are represented by vectorial random variables. For each symbol of a feature space (e.g. color or projective relations) the vectorial random variables contain a fuzzy degree of applicability of that symbol. The results of the computation of projective relations are used directly. The projective relation between two objects is thus specified by *all* computed projective relations and their degrees of applicability. This representation has the following advantages:

- Overlapping meanings and concurring hypotheses can be represented.
- The degree of applicability for each symbol of a feature space is captured. No irreversible, strict decision has to be made for the qualitative value of a feature in an early stage of the understanding processes.
- Object specifications in instructions can be of a wide variety. Using an inference machine, the understanding of object specifications is rather flexible according to which kind of features are named. Even partially false specifications can be handled.

The inference machine is based on a Bayesian network (Pearl, 1988). Nodes of the Bayesian network represent the features type, color, size, shape, and projective relation which result from speech understanding as well as the scene, the objects detected in the images, and the computed projective relations. The root node of this simple polytree network represents the identified object(s). The joint

probability for each object of being intended is estimated through propagation of the evidences in the network. If evidences are missing, restrictions are inferred. The conditional probability tables which are associated with the links in the Bayesian network are estimated based on confusion matrices obtained by the recognition procedures and by psycholinguistic experiments. Prior knowledge of the construction task is also incorporated.

Results

It is difficult to evaluate a complex system due to the system output is affected by various decisions, computations, and errors made at all processing steps.

For a first end-to-end evaluation of our system, we recorded spontaneous utterances produced by 10 naive users. Subjects had to refer to randomly selected objects in 11 different scenes. In each turn, an image of one of the scenes with one marked object was presented on a computer screen. The subject's utterance that is supposed to refer to the marked object was recorded. The first line in Table 1 shows the system results for 270 utterances. For this experiment the transcriptions of the spoken object references were used in order to avoid speech recognition errors here. Object identifications that contain the referred object are considered correct even when the system chose several objects that corresponded to the subject's utterance. The selection made by the system is unique for only 27% of the utterances.

input source	*correct*	*unique*	*false*	*nothing*
no projective rel.	195 (72%)	80 (27%)	16 (6%)	59 (22%)
projective rel.	66 (74%)	55 (62%)	13 (15%)	10 (11%)

Table 1: System results for 270 utterances containing no projective relations compared to 89 utterances of subjects that use projective relations to name the intended object.

In a second experiment, 6 subjects were explicitly asked to use projective relations when referring to objects in 6 different scenes. The second line in Table 1 shows the system output for these utterances. Here, the number of uniquely and correctly identified objects (62%) is much higher than without projective relations. Projective relations significantly improve the accuracy for exact localizations. For this experiment, judgment criteria were severer than for the first one. While in the first case, any object with the same type, color, size, or shape could be selected, here the system's task was to find exactly the intended object. Therefore, the number of false identifications is higher in the second experiment. Most of the false identifications are due to discrepancies between the computation of projective relations and the use of projective relations by the subjects.

Empirical Psycholinguistic Results

For an empirical evaluation of the computational model, two controlled psychological experiments were run. Subjects either generated spatial terms for describing the spatial relationship between object pairs or rated the applicability of spatial relations that were computed on the basis of the spatial model. Whereas most research in spatial cognition has focused on the canonical directions constituting the spatial framework using reference objects in canonical orientations (e.g., Carlson-Radvansky & Irwin, 1993; Hayward & Tarr, 1995), we are especially interested in the organization of visual space in non-canonical spatial configurations.

In both experiments, distance, orientation of the reference object, and position of the intended object (according to the computed acceptance volumes) systematically varied. To avoid conflicting perspectives, only reference and intended objects with no intrinsic parts were used. All objects were located on one horizontal plane; the use of the horizontal projective terms *rechts* (*right*), *links* (*left*), *hinter* (*behind*), and *vor* (*in-front*) was investigated. Stimuli in all experiments consisted of two objects, one of which was always in the center of the picture (the reference object: a bar) in one of four different orientations, and the other (the intended object: a cube), at one of 24 positions around the reference object and at one of three different distances (see Fig. 4). The four orientations used included two orientations collinear with the reference system (one horizontal, see Fig. 4a, and one sagittal, see Fig. 4b) and two rotated orientations where the inertia axes deviate by 45° (see Fig. 4c,d). Stimuli were displayed at a Silicon Graphics Workstation using a pair of Crystal Eyes Stereo Glasses.

Before running these two experiments, we had carried out a preliminary study in which subjects (N=36) had named these spatial relations in free speech in order to find out, among other things, what namings would be used. In 99.5% of the utterances, projective prepositions or adverbs were used. In German, two prepositions can be syntactically combined (e.g., *links vor*). Those combined namings were used in 61.5% of the utterances.

In **Experiment 1**, 40 German native speakers rated the applicability of projective prepositions that represented the relation judged maximally by the system for the configurations described above; two (equally judged) were combined when no single relation was judged best. To avoid answer bias and to have a reference system in which to see the results, a same number of distractors ("wrong answers") was introduced. Distractors had been constructed systematically by adding a second either neighboring or contrasting relation, leaving out the best or second best applicable relation, or substituting it by a neighboring or contrasting relation. The displayed relation naming could be rated from 0 to 1 by the subjects using a sliding scale.

Results show a generally high acceptance of the computed relations ($\mu = 0.90$, $\sigma^2 = 0.17$; cf. distractors: $\mu = 0.34$, $\sigma^2 = 0.39$). For some distractors in some positions, we find distractors rated the same as the system output (in the case of combination with the second applicable relation). A Multiple Analysis of Variance (ANOVA) reveals a significant influence of *orientation* and *distance* on the rating of system output with no interaction between the two factors. By post hoc mean comparisons, we find that the largest distance is rated slightly (but significantly) better and that a sagittal orientation of the reference object is

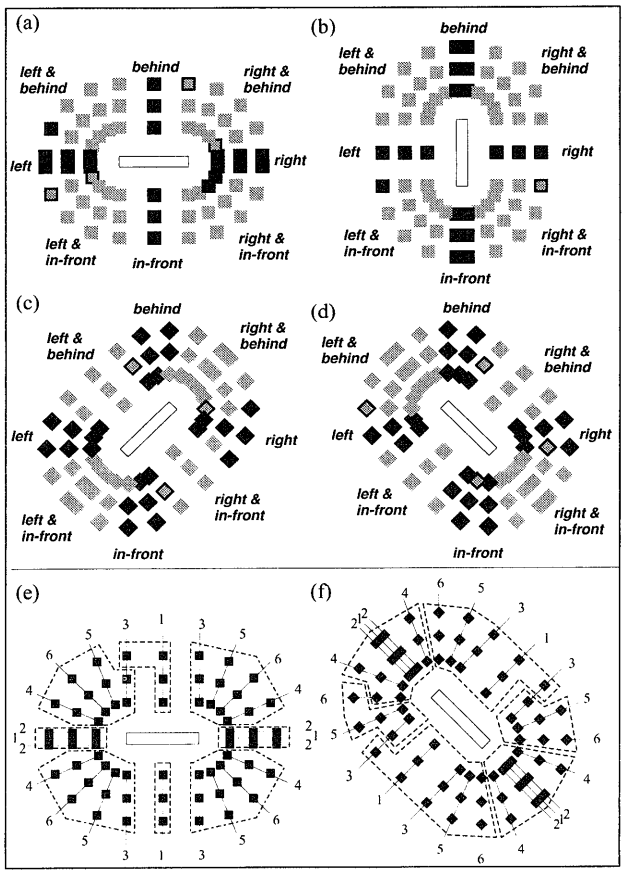


Fig. 4: (a) to (d) Preposition chosen most often in experiment 2 for each configuration used (*winner-takes-all*): A surrounding line indicates that two namings have been chosen equally often.
(e) and (f) Exemplary results of the Kohønen Maps for the collinear and the rotated conditions. The broken lines enclose those positions belonging to one cluster. Numbers indicate positions grouped together for an ANOVA.

rated better than a horizontal one, which again is rated better than the two rotated orientations. The ANOVA yielded a significant influence of *position* on the rating as well. To investigate this dependency in detail, we subdivided the 24 positions into 6 position groups according to the six positions contained in each quadrant around the reference object. In the collinear conditions, we find that the nearer the position groups are to the *left-right* and *in-front-behind* axes ($1/2 > 6 > 3/4 > 5$; see Fig. 4e) the better they are rated. In the rotated conditions some position groups are rated better than others as well ($5 > 6/3 > 1/4 > 2$; see Fig. 4f). Altogether, subjects' ratings show a significant correlation with the degrees of applicability computed by the system.

In **Experiment 2**, 20 German native speakers themselves named the direction relations, choosing from a given set of projective prepositions (on buttons) the one that fitted the displayed spatial configuration best. The given answer choice included *right*, *left*, *behind*, *in-front*, *below*, *above* and their respective combinations.

The results of experiment 2 show again that subjects tend

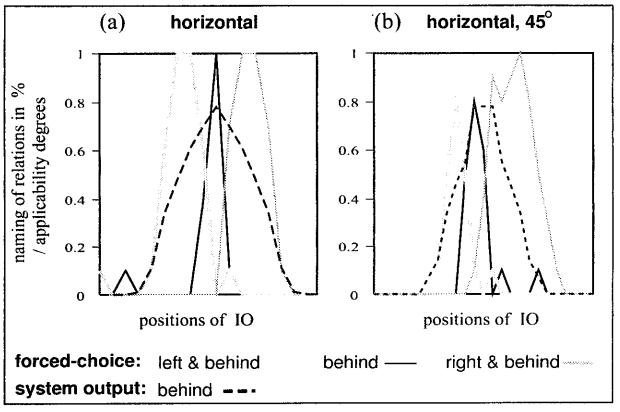


Fig. 5: Exemplary comparison of applicability degrees computed by the system and relative frequency of used prepositions in the forced-choice experiment (experiment 2): (a) horizontal orientation (see Fig. 4a) (b) horizontal orientation rotated by 45° (see Fig. 4c)

to use combined direction terms; single direction expressions are seemingly used only in unambiguous cases. As for the collinear reference objects (see Fig. 4a,b), the positions in which most subjects use single expressions are located centrally within the computed relation applicability areas. Regarding the rotated reference objects (see Fig. 4c,d), two combined direction terms are used significantly more often depending on the direction of the reference object (e.g., *left& in-front* and *right& behind* in the case of orientation (c); see Fig. 5b) indicating a stronger influence of the reference system on the partitioning of space. Otherwise (when surrounding space is partitioned exclusively by the reference object) we would expect a symmetric distribution of spatial terms.

A more detailed analysis shows the overlap of applicability regions when taking into account the distribution of different projective terms over the positions occupied by the intended objects. For the collinear orientations of the reference object, the peaks in frequency of the naming of a certain preposition correspond well with those in the computed applicability degree of the corresponding relation (see Fig 5a). Regarding the rotated reference objects, we find a small systematic shift of the empirical peak and a disproportion of certain combined prepositions in the same way as in the winner-takes-all-results (see Fig. 5b). A clustering of the data obtained by means of a Self-Organizing Kohønen Map reveals that only for the non-rotated reference objects is it possible to find eight clear-cut applicability areas of the four prepositions and their respective combinations. Whereas it is more difficult to distinguish applicability regions in the case of rotated reference objects. This result is supported by an Analysis of Variance using the information entropy obtained for each distance and ROs orientation. There is a significantly greater entropy (from what uncertainty follows) in the case of rotated reference objects as well as in the case of the nearest distance (with no interactions between the two factors).

Discussion

The generally high rating of the system's output shows the usability of the spatial computational model. In general, we find a correspondence between the central tendencies of the subjects' namings and the system's output, particularly good for the collinear reference objects.

Production data, especially the greater entropy found for the rotated reference objects (experiment 2) can account for the slightly smaller rating of these orientations in experiment 1. There is obviously more uncertainties about how to name projective relations in the case of rotated reference objects. This interpretation is supported by results of Kohønen Maps which show that subjects have no clear-cut applicability regions for the rotated reference objects.

In accordance with results reported in literature, our data provide strong evidence for the graded validity of projective relations and their overlap. The computational model takes this into account.

The empirical findings provide evidence for the assumption of the computational model that the reference object's extensions in space (Vorwerg, in prep.) as well as its rotation must be taken into account. For a rotated reference object, there is obviously a conflict between the object's axes and the reference system's axes, which seems to be resolved by a kind of trade-off resulting in using the nearest part (corner or edge) of the reference object as a reference point. Subjects' data can be interpreted as using the axes imposed by the reference system, but shifting those half-line axes (*left*, *right*, *in-front*, and *behind*) towards the nearest part of the reference object. The computational model's rotation of acceptance volumes may serve as an approximation for these cognitive processes in most positions. In some positions around rotated reference objects however, the assumed half-line axes cross acceptance volumes resulting in a discrepancy between computed projective relations and terms used by the subjects. These cases can account for the false identifications found in the testing of the artificial communicator system and could possibly be compensated for by using a non-linear and non-symmetric judgment function for the degree of accordance.

Subjects' frequent use of combined direction terms, which might in part be due to this special syntactic possibility in German, can easily be fitted into the computational model by changing the criterion according to which two relations are to be combined (e.g. take all relations which are assessed better than 0.3; in the case of two, combine them).

The fact that several positions exist where subjects rated combined and single terms equally indicates some flexibility in assigning spatial relations. Regarding the empirical evaluation of computational models, it must be taken into consideration that the uncertainty of subjects in naming a projective relation has some influence on the applicability rating of presented projective terms.

Conclusion

In this paper, we presented a computational model for projective relations that is used in a system for advanced human-computer interaction. In this context, psychological and cognitive aspects of human localization are of great interest. We carried out experiments investigating the use of projective relations by humans in our scenario. The experiments included an empirical evaluation of our system.

The experiments show a high acceptance of the computation of the projective relations by humans. They also show differences between the human production of projective relations and their computation by our model. Further investigations will concentrate on the use of the cognitive findings to improve the computational model. The emphasis will be placed on achieving a good trade-off between technical benefits (like the two-layered structuring and the existence of a reference-independent spatial representation) and cognitive adequacy of the model.

References

Abella, A. & Kender, J. (1993). Qualitatively Describing Objects Using Spatial Prepositions. In *Proc. of AAAI-93*, pp. 536–540.

Carlson-Radvansky, L. & Irwin, D. (1993). Frames of reference in vision and language: Where is above? *Cognition 46*, 223–244.

Fuhr, T., Socher, G., Scheering, C., & Sagerer, G. (1995). A three-dimensional spatial model for the interpretation of image data. In P. L. Olivier (Ed.), *IJCAI-95 Workshop on Representation and Processing of Spatial Expressions*, Montreal, pp. 93–102.

Gapp, K.-P. (1994). Basic Meanings of Spatial Relations: Computation and Evaluation in 3D space. In *Proc. of AAAI-94*, pp. 1393–1398.

Hayward, W. & Tarr, M. (1995). Spatial language and spatial representation. *Cognition 55*, 39–84.

Herskovits, A. (1986). *Language and spatial cognition*. Cambridge, Mass.: Cambridge University Press.

Landau, B. & Jackendoff, R. (1993). "What" and "where" in spatial language and spatial cognition. *Behavioral and Brain Sciences 16*, 217–265.

Mc Kevitt, P. (Ed.) (1994). *Special Issue on Integration of Natural Language and Vision Processing*, Volume 8 of *Artificial Intelligence Volume*. Kluwer Academic Publishers.

Nagel, H. (1988). From image sequences towards conceptual descriptions. *Image and Vision Computing 6*(2), 59–74.

Olivier, P. & Tsujii, J.-I. (1994). Quantitative Perceptual Representation of Prepositional Semantics. *Artificial Intelligence Review* Special Issue of Natural Language and Vision Processing *8*(2-3), 55–66.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo, CA: Morgan Kaufmann Publishers.

Socher, G. & Naeve, U. (1996). A Knowledge-based System Integrating Speech and Image Understanding – Manual Version 1.0. Report 95/15 – Situierte Künstliche Kommunikatoren, SFB 360, Universität Bielefeld.

Socher, G., Sagerer, G., Kummert, F., & Fuhr, T. (1996). Talking About 3D Scenes: Integration of Image and Speech Understanding in a Hybrid Distributed System. In *International Conference on Image Processing (ICIP-96)*, Lausanne, pp. 18A2.

Vorwerg, C. (in preparation). Categorization of spatial relations.

Wahlster, W. (1989). One Word says More Than a Thousand Pictures. On the Automatic Verbalization of the Results of Image Sequence Analysis Systems. *Computers and Artificial Intelligence 8*, 479–492.