Sonderforschungsbereich 378
Ressourcenadaptive kognitive Prozesse

KI-Labor am Lehrstuhl für Informatik IV

Leitung: Prof. Dr. W. Wahlster

REAL

Universität des Saarlandes
FB 14 Informatik IV
Postfach 151150
D-66041 Saarbrücken
Germany

Tel. 0681 / 302-2363

Bericht Nr. 137

Construction of Mental Images and their Use in a Listener Model

Eva Stopp, Anselm Blocher

Januar 1997

Sonderforschungsbereich 378
Ressourcenadaptive
Kognitive Prozesse

ISSN 0944-7814

137

Construction of Mental Images and their Use in a Listener Model

Eva Stopp and Anselm Blocher
SFB 378 – Project REAL
Department of Computer Science
Saarland University
D–66123 Saarbrücken
Germany
{eva,anselm}@cs.uni-sb.de

January 23, 1997

Abstract

The research in Artificial Intelligence concerning the connection between vision and natural language systems relies on what is often called reference semantics. In the situation of a listener not able to see a reported scene, the utterances of the speaker must be perceptually anchored and coherent in order to be understandable. Accordingly, the reporter must be able to anticipate the listeners' understanding by means of mental images. In this paper, we demonstrate the construction of mental images and their comparison to visual perception on the level of spatial relations. The results are used to improve the understandability of planned utterances.

1 Introduction

In Artificial Intelligence (AI), the research areas computer vision and natural language processing have generally been studied intensively but independently from each other. Only few attempts have been made towards the integration of image understanding and natural language descriptions for real word image sequences.

The project VITRA – **VI**sual **TRA**nslator, 1985-1995, cf. [7, 6]) – as part of the German special collaboration programme SFB 314, *AI & Knowledge-Based Systems*, dealt with the relations between speaking and seeing, and aimed at a completely operational form of reference semantics for what is visually perceived.

In: Proceedings of the Symposium on Conceptual and Semantic Knowledge in Language Generation, Heidelberg, Germany, 15-17 November, 1995, 270–280.

2 The VITRA Project

Within VITRA, several scenarios were considered in order to construct a system for incremental natural language description of image sequences:

- Answering questions about traffic scenes
- Simultaneously describing video scenes
- Incremental multi-media route descriptions
- Natural language interaction with an autonomous robot system

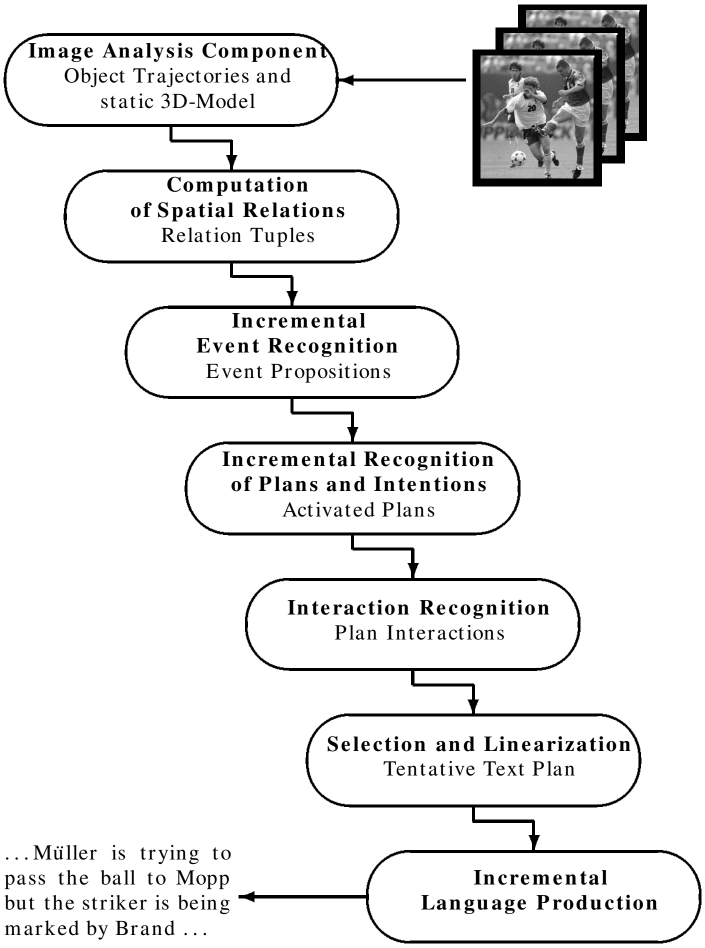


Figure 1: Cascaded Processing in VITRA

The processing in the VITRA system forms a kind of cascade which is shown in Fig. 1. An image sequence in form of digitized video frames builds the input for the processes on the sensory level. Stating the locations of the visible objects at consecutive points in time, the image analysis component incrementally constructs a geometrical scene description. Its contents is further interpreted by the processes on the cognitive level: spatial relations, interesting motion events as well as presumed intentions, plans, and interactions of the observed agents are extracted. The produced conceptual structures form the connection between visual data and natural language concepts such as spatial prepositions, motion verbs, temporal adverbs, and purposive or causal clauses. After a suitable description of the scene is chosen by the selection component, the linearization module forms a tentative text plan which builds the basis for the transformation into natural language utterances. In terms of reference semantics, explicit links between sensory data and natural language expressions are established.

3 Spatial Relations in VITRA

Processing of spatial relations is described in more detail in the section written by Klaus-Peter Gapp. There, he describes the distinction made in VITRA between the "basic meaning" of a spatial preposition (cf. [5]) represented by spline functions and a concrete instantiation which depends on the object to be located (LO) and one or more reference objects (RO). According to *instantiation rules* depending on the reference object(s), a spatial proposition is created for every instance of a spatial concept. It builds the basis for the computation of the degree of applicability of the considered concept with respect to the actual position of the LO (cf. [3, 13]). The applicability distribution of a whole region is called *applicability region* or *potential field*. We distinguish several classes of spatial relations, as, e.g., topological or projective. This kind of computation can be used both for the generation of suitable spatial description and for the analysis of spatial expressions.

4 A Listener Model with Mental Images

4.1 Anticipation of the Listeners' Imagery

For the listeners of an objective description, it is crucial to understand each assertion in its corresponding context. In the following, the incremental generation of descriptions of short soccer scenes is taken as an example for the use of a listener model with mental images. The resulting report is similar to a live radio coverage, i.e., simultaneously and in an objective manner, to an audience not able to see the game. Currently, we use a bird's-eye projection of the perceived 3D scene onto a 2D plane with mobile objects idealized to mere points.

The logical distinction between the contextually given anchor points of an utterance – essentially given by the definite noun phrases and their pro-forms – and the other components of the assertion which are the actually informative parts – here, we are especially interested in verbs and prepositions – leads to the following questions:

Question of Reference: Can the listener of an assertion uniquely identify the contextual objects in question by means of the given noun phrases?

Question of Plausibility: Is the listener able to integrate the newly communicated information into the present context (which thereby becomes the context for the next assertion)?

And finally:

Question of Correctness: Does the listener's interpretation of the assertion correspond to the speaker's intention? Or is it necessary to modify the assertion in order to reach the communicative goal intended?

Here, we emphasize on two concrete aspects of these questions:

(1) Can the speaker use anticipations of the listeners' *mental images* evoked by the former description in order to choose optional locative deep case fillers as modifications of assertions that otherwise would lead to a correctness problem?

(a) Miller attacks Smith.
$\longrightarrow$ *(b)* Miller attacks Smith near the left penalty spot.

(2) Can the speaker also use the anticipated mental images to extend his ability to use under-specific definite descriptions in a legitimated way?

(a) Brown crosses the upper side line.
$\longrightarrow$ *(b)* Brown crosses the line.

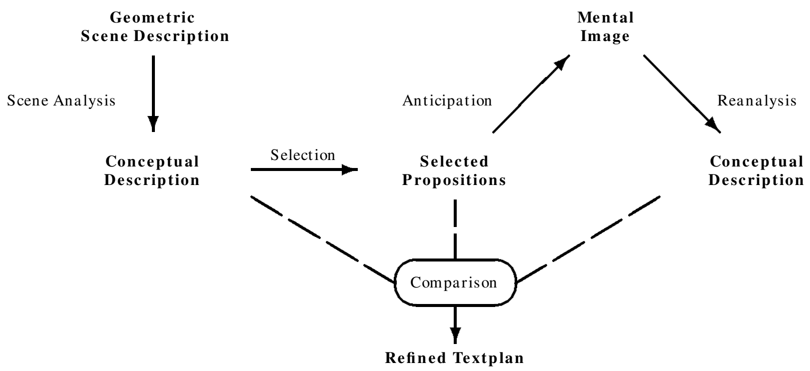


Figure 2: Anticipation Feedback Loop

An additional component with the particular task of improving the text of the speaker system concerning communicative aspects is given by a listener model (cf. [13, 12, 1]): It enables the system to deal with the above-mentioned questions of plausibility, correctness, and reference by anticipating the listeners' understanding of a planned statement. According to the approach of reference semantics, the listeners have reached a deep understanding of the description only if they are able to create mental images structurally corresponding to percepts. In generalizing Grice's Maxim of Quality to graded concepts (cf. [4]), it is postulated that a listener interprets an utterance of the speaker as a description of a maximally typical occurrence of the mentioned event in the given context; deviations have to be communicated explicitly (cf. [12, Chapter 6]). Therefore, the successful construction of a highly typical image is taken as a proof of the planned assertion's plausibility: The listeners are assumed to be able to integrate the newly communicated event within the given context.

Three steps of processing within the *anticipation feedback loop* (cf. Fig. 2) of the listener model are distinguished: (a) the construction of a visual *pseudo-percept*, (b) the analysis of this *mental image*, which is necessary to explicitly represent the implicatures the listeners derive from the utterance, and (c) the comparison of this anticipated understanding with the intended effects of the utterance and the resulting changes in the generation processes.

4.2 Construction of Mental Images

The listener model starts with the event proposition – in Fig. 3 this is *(*pass* W1 Ball W3)* – chosen by the incremental event selection component of VITRA: it represents the core of the planned utterance.

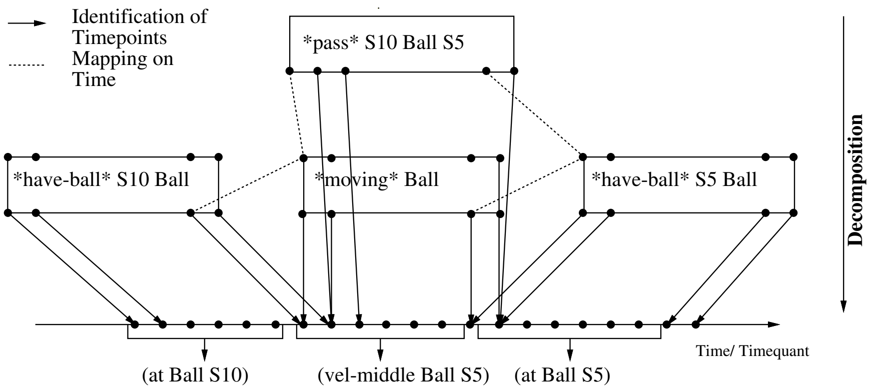


Figure 3: Decomposition of a **pass** Event

The construction of mental images is done in two steps: the temporal concretization which results in the *propositional elementary structure* and, based on that, the spatial concretization which is competent for the generation of the mental images. An event proposition is recursively defined by its subevents (with their temporal relations) down to primitive propositions. The definition of the considered event must be expanded and adapted to the situational constraints in order to achieve spatio-temporal coherence between the required spatial restrictions of the involved objects during each phase of the event and the context (cf. [11]). Further constraints given by optional deep case fillers also have to be associated with the appropriate part of the event (cf. [8] and [10]). I.e., in order to find a basis for the spatial concretization, the temporal concretization explicits the spatial and temporal relations between objects, relations which are implicitly contained in the utterances. Fixing the duration of the elementary sub-events then results in the propositional elementary structure, the temporally ordered sequence of sets of elementary spatio-temporal relations. Each set restricts the scene for exactly one instant (time quantum).

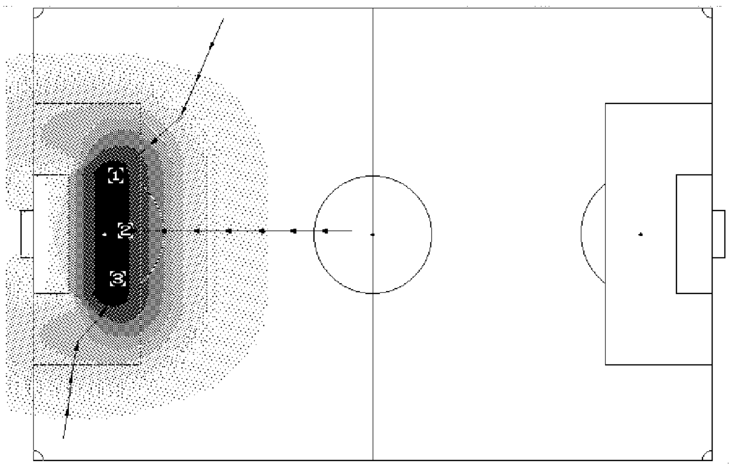


Figure 4: Context-sensitive Approximation of each Player Standing in Front of the Goal Area

This data is transformed in the spatial concretization to a sequence of snapshots by approximating particular locations and velocities for all objects considered successively for each time quantum. To do so, we apply the so-called *hill-climbing algorithm* to the potential fields that correspond to the relations concerning the LOs respectively. If there are several propositions at one time they are combined using algebraic average. The features of the hill-climbing algorithm cause the approximation to be performed in a context-sensitive manner (cf. Fig. 4), i.e., the resulting optimal position of an object depends on its starting position at the beginning of the approximation.

4.3 Comparing the Intended and the Anticipated Understanding

The above descibed process generates an imaginative cinematographic scene which can be reanalysed by the same means as the visual percept. The resulting sets of propositions, then, are compared to those based on the geometrical scene description. In order to relate the propositions forming the intended understanding with the equivalent propositions representing the anticipated effects, precisely one of three types of criteria is taken into consideration: Relation, reference object or degree of applicability (DA) (cf. Table 1).

All pairs of associated propositions are additionally rated and ordered with respect to the relevance of the digression they express: The numerical difference between the two degrees of applicability and a measure of the "conceptual distance" between the involved relations and ROs are combined to a general ranking of the three types. Those pairs exceeding a given threshold

Type of Difference		Source	Relation	LO	RO	DA
1	Relation	Visual Percept	near	Miller	Smith	0.95
		Mental Image	at	Miller	Smith	0.92
			$\neq$	=	=	=
2	RO	Visual Percept	near	Miller	Smith	0.95
		Mental Image	near	Miller	Russel	0.92
			=	=	$\neq$	=
3	DA	Visual Percept	near	Miller	Smith	0.95
		Mental Image	near	Miller	Smith	0.48
			=	=	=	$\not\approx$

Table 1: Examples of Criteria of Equivalence for Spatial Relations

of difference lead to changes in the sentence finally uttered. The prominent opportunity for such changes are optional deep cases, particularly the locative forms, like Source, Goal, or Location (cf. [2], [9], and especially [8]). In fact, those pairs of differing propositions already tell us what kind of fillers can repair the utterance's communicative problem.

If the part of the proposition pair derived from the geometrical scene description has a higher degree of applicability than the one originated by analysing the mental image then this proposition obtained by the speaker's perception must be uttered because it assumedly is not present in the mental model of the listeners after understanding the event proposition. In the other case – the proposition from the mental image has the higher degree of applicability of the pair – the proposition from the visual percept should be mentioned, too. However, a correction of the location falsely deduced may be added, indicating that the speaker is aware of this contextual difficulty: "Miller stands *to the left of* Scott, *not behind* him."

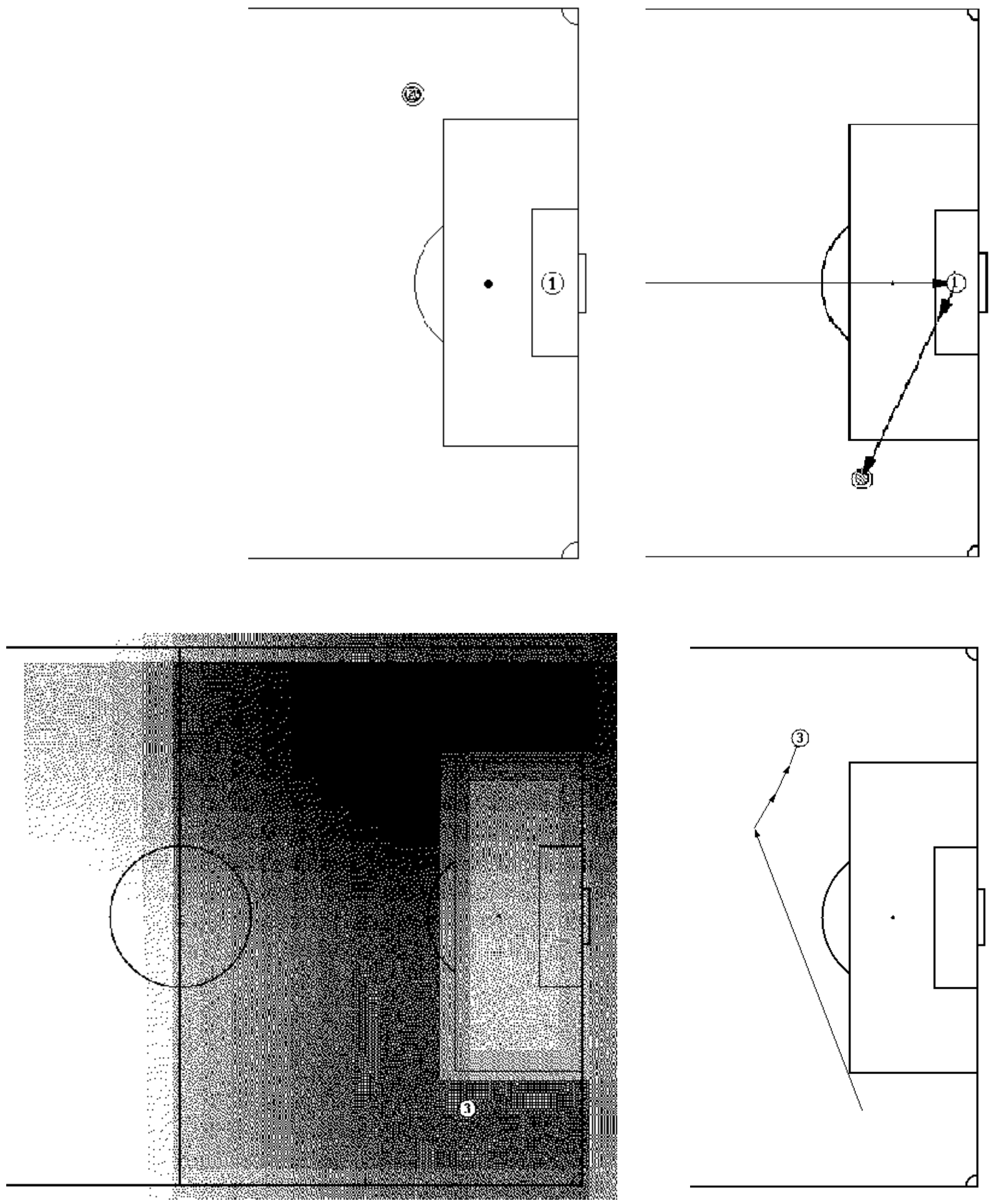
The following sentences are part of a longer example produced by the speaker system. The modifications initiated by the listener model are emphasized:

1. Miller, the goal keeper, has the ball.
2. He plays the ball to Moll, the defender, *near the upper side line.*
3. – 7. ... *(no penalty area is mentioned)*
8. Michels, the left-wing, has got the ball *at the penalty area*.

Since at the beginning of a report there is no previous information given concerning the involved objects, the image for the first utterance uses standard positions of the goal keeper and the ball as its context: While it happens to be correct that the goalie is in his goal area, the ball's position (at the middle point) has to be changed by the hill-climbing procedure. In this case, everything works fine with the utterance proposed. For the second utterance, Moll's standard position has to be used, too, since this player had not been mentioned before: But in contrast to the previous utterance this location (cf. Fig. 5b) differs a lot from the one perceived (cf. Fig. 5a).

(a) What is Seen and...

(b) ...what is Imagined



Combined Typicality Distribution for Revising the Position of Player W3 (c)

Resulting Revised Position of W3 (d)

Figure 5: Visual Percept, Mental Image and its Revision

We can extract the following pair of propositions which applies the same relation ("near") to different reference objects:

Visual Percept: *((near W3 upper-side-line) 0.79)*

Mental Image: *((near W3 lower-side-line) 0.84)*

This leads to the additional prepositional phrase (PP) *"near the upper side line"* in the second utterance.

These modifications in the planned statement now have to be integrated in the anticipated understanding of the listeners. The position of the object in question is changed using the hill-climbing algorithm: Hereby, not only the additional relation must hold, but also all those relations that have already held before correctly. This prevents the system from constructing *new* difference propositions. In our example, additional to *(near W3 upper-side-line)*, the relations *(in W3 right-half-field)* and *(near W3 right-penalty-area)* have to be taken into account. These three relations generate the compound typicality distribution of Fig. 5c and result in the revised position of player W3, shown in Fig. 5d. The position found obviously corresponds extremely well to the speaker's perception.

In the eighth utterance, the decision to use an additional locative prepositional phrase for communicating an optional deep case is motivated as above. In this case, the position of Michels (W11) resulting from the previous understanding – not his standard position – comes out to be not exactly as was observed. A highly relevant difference is found (first case of classification in Table 1):

Visual Percept: *((close W11 left-penalty-area) 0.97)*
Mental Image: *((at W11 left-penalty-area) 0.99)*

A listener model with mental images demonstrates a type of focusing that does not – directly – depend on previous mentioning but on the *imaginative visual focus* of an utterance. The corresponding mental image – forming the context of the next utterance – exactly contains those spatial implicatures the listener assumedly was able to derive: Therefore, the uniqueness of the reference of a definite description need not be bound to the whole discourse universe but can be restricted to the much smaller set of objects involved in the actual imaginative visual focus. This suffices to determine whether a description is ambiguous or not, and therefore shrinks the set of attributes necessary to uniquely distinguish an object from the others.

In the image anticipated for utterance 7, Michels happens not to stand close enough to that part of the soccer field to be considered really – as in the speaker's percept (cf. Fig. 6) – *at* the left penalty area. Therefore, a corresponding modification should be added. In contrast to the former example, the analysis of the mental image constructed for the previous sentence here did focus on the reference object to be used in the additional PP, i.e., the left penalty area; simultaneously, the alternative right penalty area was not focused. Correspondingly, the *under-specified* definite description "the penalty area" can be assumed to be uniquely identifying in this situation the RO intended, although it is ambiguous with respect to the whole discourse universe. Thus, the listener model suggests for utterance 8 the additional PP "at the penalty area", although none of the penalty areas ever were mentioned before during the description.

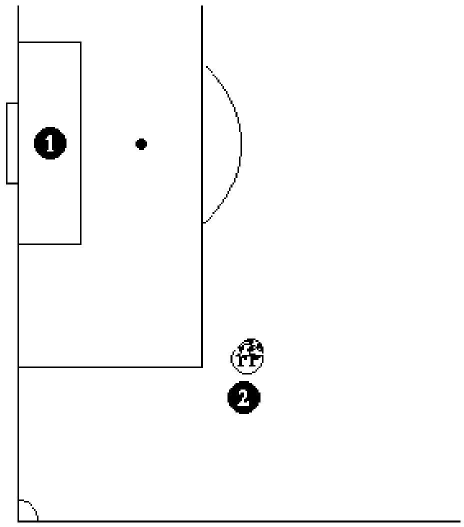


Figure 6: Focused Region of Utterance 7

5 Conclusion

In this paper, we demonstrated how an anticipation feedback loop based on reference semantics can be used to improve speakers' reports to an audience unable to see the scene. Based on the planned utterance, the corresponding event proposition is expanded to a temporally ordered sequence of sets of elementary spatio-temporal relations. For each of these sets a mental image is constructed representing the listeners' presumed contextual information. This process generates an imaginative cinematographic scene which is re-analysed and compared to the visually perceived one. The text plan is adapted to the results of this comparison of the intended and anticipated understanding of the utterance. Optional deep case fillers may be added, under-specified definite descriptions may be applied. By this anticipation of the listener's imagery, redundancies, misunderstandings and false inferences can be reduced and coherence and understandability of the report can be improved.

References

[1] A. Blocher and J.R.J. Schirra. Optional deep case filling and focus control with mental images: ANTLIMA-KOREF. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pages 417–423, Montréal, Canada, August 19-25, 1995.

[2] C. J. Fillmore. The case for case reopened. In P. Cole and J. M. Sadock, editors, *Syntax and Semantics*, volume 8. Academic Press, New York, 1978.

[3] K.-P. Gapp. Basic meanings of spatial relations: Computation and evaluation in 3d space. In *Proc. of AAAI-94*, Seattle, WA, 1994.

[4] H. P. Grice. Logic and conversation. In P. Cole and J. L. Morgan, editors, *Speech Acts*, pages 41–58. Academic Press, London, 1975.

[5] A. Herskovits. *Language and Spatial Cognition. An Interdisciplinary Study of the Prepositions in English*. Cambridge University Press, Cambridge, London, 1986.

[6] G. Herzog, A. Blocher, K.-P. Gapp, E. Stopp, and W. Wahlster. VITRA: Verbalisierung visueller Information. *Informatik – Forschung und Entwicklung*, 11(1):12–19, 1996.

[7] G. Herzog and P. Wazinski. VIsual TRAnslator: Linking perceptions and natural language descriptions. *Artificial Intelligence Review Journal*, 8(2):175–187, 1994. Special Volume on the Integration of Natural Language and Vision Processing, edited by P. Mc Kevitt.

[8] H. Marburger and W. Wahlster. Case role filling as a side effect of visual search. In *Proc. of the 1st EACL*, pages 188–195, Pisa, 1983.

[9] G. Rauh. *Tiefenkasus, thematische Relationen und Thetarollen. Die Entwicklung einer Theorie von semantischen Relationen*. Gunter Narr, Tübingen, 1988.

[10] P. Sablayrolles. Analyse conceptuelle et représentation des verbes de déplacement du français. Rapport de Recherche IRIT/91-23-R, Université Paul Sabatier, Toulouse, 1991.

[11] J. R. J. Schirra. Einige Überlegungen zu Bildvorstellungen in kognitiven Systemen. In Ch. Freksa and Ch. Habel, editors, *Repräsentation und Verarbeitung räumlichen Wissens*, pages 68–82. Springer, Heidelberg, 1990.

[12] J. R. J. Schirra. *Bildbeschreibung als Verbindung von visuellem und sprachlichem Raum: Eine interdisziplinäre Untersuchung von Bildvorstellungen in einem Hörermodell*. infix, St. Augustin, 1994.

[13] J. R. J. Schirra and E. Stopp. Antlima – a listener model with mental images. In *Proc. of the 13th IJCAI*, pages 175–180, Chambery, France, 1993.