# Optional Deep Case Filling and Focus Control with Mental Images: ANTLIMA-KOREF

**Anselm Blocher**
SFB 314, Project VITRA
Department of Computer Science
University of Saarbrücken
D-66123 Saarbrücken
Germany
anselm@cs.uni-sb.de

**Jörg R.J. Schirra**
International Computer
Science Institute

Berkeley, California 94704
USA
joerg@icsi.berkeley.edu

## Abstract

The connection between vision and natural language systems in AI research relies on what is often called reference semantics. In the situation of a radio reporter for soccer games, an utterance must be perceptually anchored and coherent in order to be understandable to a listener not able to see the scene. Accordingly, the speaker must be able to anticipate the listeners' understanding by means of mental images. In this paper, we demonstrate the comparison of mental images and visual perception on the level of spatial relations and show how to employ the results for cooperatively filling optional deep cases, and for controlling the use of underspecific definite descriptions.

## 1 Introduction

The project VITRA started in 1985 as part of the German special collaboration programme SFB 314, *AI & Knowledge-Based Systems*. It deals with the relations between speaking and seeing, and it aims at a completely operational form of reference semantics for what is visually perceived. The system SOCCER (cf. [André *et al.*, 1989]) constructed in VITRA demonstrates the computational link between visual perception and natural language. Here, we concentrate on the interaction between the 'speaker system' SOCCER and its listener model ANTLIMA (cf. [Schirra and Stopp, 1993]), mediated by the component ANTLIMA-KOREF.

## 2 Three Communicative Questions

An objective description of soccer events essentially consists of a sequence of assertions. For the listener of the description, it is crucial to understand each assertion in its corresponding context. Following the logical distinction between the contextually given anchor points of an utterance[1] and the other components of the assertion which are the actually informative parts,[2] we can distinguish the following two questions:

[1]These are essentially given by the definite noun phrases and their proforms.

[2]In the case we are interested in, these are essentially the verbs and prepositions.

**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)?

For a speaker who wants to be understood, these questions have to be considered in advance: Because speakers usually try to give 'economical' descriptions, they need to check whether short but underspecific noun phrases (such as, in the extreme case, the proforms) uniquely identify the intended object. The speakers additionally have to consider a third communicative question, since they have a certain intention with their utterance, e.g., to inform the listeners objectively about something perceived:

**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?

In the framework of these questions, we are essentially interested in two particular instances. (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, as in example (1b)? (2) Can the speaker also use the anticipated mental images to extend his ability to use underspecific definite descriptions in a legitimated way, as in example (2b)?

**(1)** *(a)* M. attacks S. $\longrightarrow$ *(b)* M. attacks S. near the left penalty spot.
**(2)** *(a)* B. crosses the upper side line. $\longrightarrow$ *(b)* B. crosses the line.

## 3 A Glance at SOCCER and ANTLIMA

SOCCER generates descriptions of short soccer scenes (in German), similar to a live radio coverage, i.e., simultaneously and in an objective manner, to an audience not able to see the game. The input, called the 'geometrical scene description', consists of the shape and configuration of the soccer field and its parts, plus, at every time quantum, the set of spatial locations and velocity vectors of each mobile object perceived, as provided by the motion analysis system ACTIONS (cf. [Nagel, 1988], [Herzog *et al.*, 1989]). Currently, SOCCER uses a bird's-eye

projection of the perceived 3D scene onto a 2D plane with mobile objects idealized to mere points.

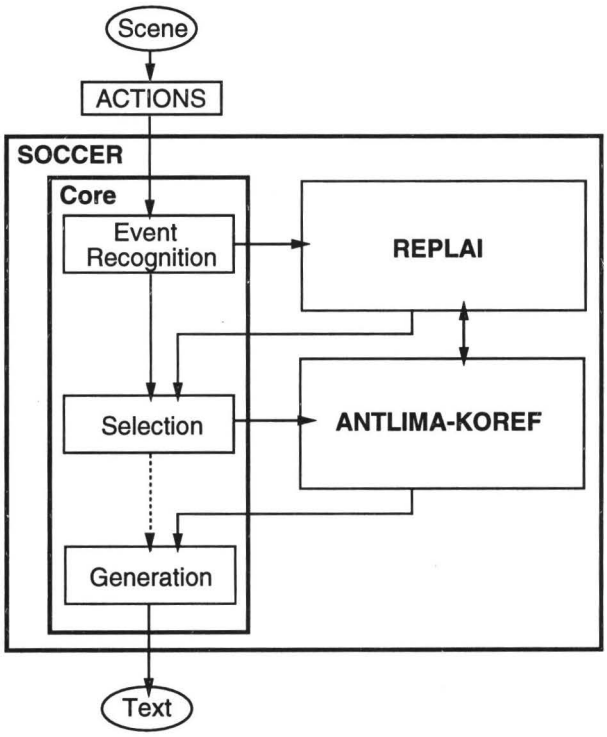



Figure 1: Extended Architecture of SOCCER

The core system of SOCCER consists of three components (cf. Fig. 1).[3] First, the Event Recognition component analyses the incoming part of the geometrical scene description, based on graded classification functions for elementary static spatial relations projecting spatial configurations to *degrees of applicability* $\in$ [0.0, 1.0] (cf. Fig. 2 and [Schirra, 1992]). The resulting propositions – describing instances of spatial relations and spatio-temporal events (e.g., *(left player-1 player-2)* and *(double-pass player-1 ball player-2)*) – are passed to the Selection component. Here, the continuation of the report is planned by choosing some of the event propositions in question as relevant to be communicated. The criteria of relevance depend only on the scene perceived, e.g., time elapsed since perception, or salience of event type; they do not consider the listeners. Finally, the Generation component transforms the event proposition into an appropriate utterance in German. In particular, the descriptions of the objects in question have to be determined.

An additional component with the particular task of improving the text of the core system concerning communicative aspects is given by the listener model ANTLIMA – ANTicipation of the Listeners' IMAgery (cf. [Schirra, 1992; Schirra and Stopp, 1993; Schirra, 1994]). 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. [Grice, 1975]), 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. [Schirra, 1994, Chapter 6]). Therefore in ANTLIMA 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.

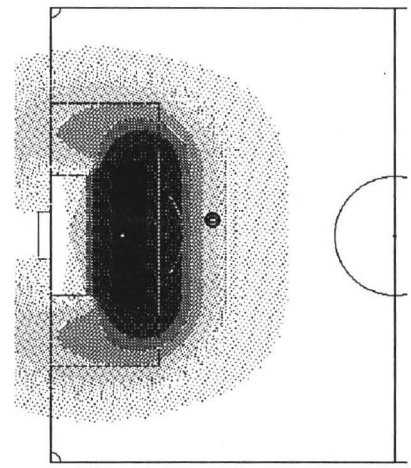


Figure 2: Graphical representation of the classification functions for 'being in front of' a goal area (darkness corresponds to applicability)

Three steps of processing 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. The first two steps have already been described in sufficient detail elsewhere (cf. [Schirra and Stopp, 1993], [Schirra, 1994]); we include here merely a short overview.

ANTLIMA starts with the event proposition chosen by the Selection component to be verbalized next. Following the definition of the event type in question, an equivalent temporally ordered sequence of sets of elementary spatio-temporal relations is constructed, describing the scene by consecutive snapshots. In a 'cinematographic procedure', each of these sets is transformed to a corresponding part of a geometrical scene description depending on the proper contextual snapshot: A hill-climbing algorithm is employed to locate the mobile objects concerned at the positions most typical for the given restrictions. The degrees of applicability of the classification functions are re-interpreted as *typicality values* guiding the search for the optimal position (cf. Fig. 3). When an object is first mentioned, its contextual position is taken to be a standard position associated with the player's function: The goal keeper typically is in his goal area. Due to the influence of the particularly given context, the resulting anticipated mental image may differ from the speaker's perception. It therefore is analyzed again in the second step by means of the classifi-

[3] An extension for recognizing plans and intentions of the observed agents is realized by the component REPLAI; cf. [Retz-Schmidt, 1991; 1992].

cation procedures used in the core system. These recognition procedures distinguish between relevant and irrelevant spatio-temporal differences by either classifying two scenes under distinct categories or putting them together under the same one.

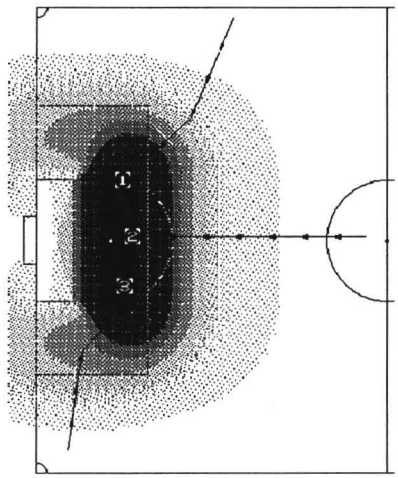


Figure 3: Hill-climbing

In this article, we focus on the third step, which defines the task of the ANTLIMA-KOREF component (A-KOREF, for short; cf. [Blocher, 1994]). Under the precondition that the listener could construct a plausible mental image of the utterance planned, A-KOREF decides whether additional locative deep case fillers could enhance the report, and whether a given object may be referred to by means of an underspecific definite description.

## 4 The Component ANTLIMA-KOREF

The data packages used in A-KOREF are called 'contexts'. We distinguish *intended*, *realized*, and *difference* contexts, or **I-**, **R-**, and **D-**contexts (cf. Fig. 4). Looking at time point $i$ when the Selection component proposes an event proposition to be verbalized, we cope with one corresponding context for each of these types: I-Context ($i$), R-Context ($i$), and D-Context ($i$). The I-Context contains the relevant data that is delivered by SOCCER's core system: This set of current event propositions is sorted according to criteria like state of recognition and salience. Furthermore, locative propositions describing the positions of those mobile objects that are *visually focused* belong to the I-Context: Their calculation is described below. The R-Context is defined analogously to the I-Context, but it evolves from ANTLIMA's Re-Analysis. Finally, the D-Context describes the *difference* between the I- and R-Contexts, and thus, presents the relevant differences between the speaker's intention (a true description of his perception) at moment $i$, and the listeners' assumed understanding with respect to the proposition chosen to be communicated next. Therefore, the D-Context is the basis of A-KOREF's response: Depending on this data, a modified proposition is constructed and sent to the component Generation to be uttered.

Due to restrictions of space, we concentrate in the following sections on the differences with respect to the elementary spatial relations and ignore the processing of event propositions.[4]

[4]Descriptions of the latter are to be found in [Blocher, 1994] and [Schirra, 1994, Chapter 12]; essentially, initiating a clarification interruption of the description in order to recover from a severe misunderstanding and changes of the priorities of the event propositions already selected (the speech plan of the Selection component) are dealt with; currently, only parts are realized in the implementation of SOCCER.

Under the precondition that the event proposition under consideration was rated plausible,[5] we also can assume that this event can be identified again in the image, and thus, is part of both the corresponding I- and R-Context: However, the spatial details of understanding may still be different. We therefore now describe how corresponding locative propositions in the I- and R-Contexts are associated to form the D-Context, i.e., the set of pairs of propositions expressing the relevant differences.

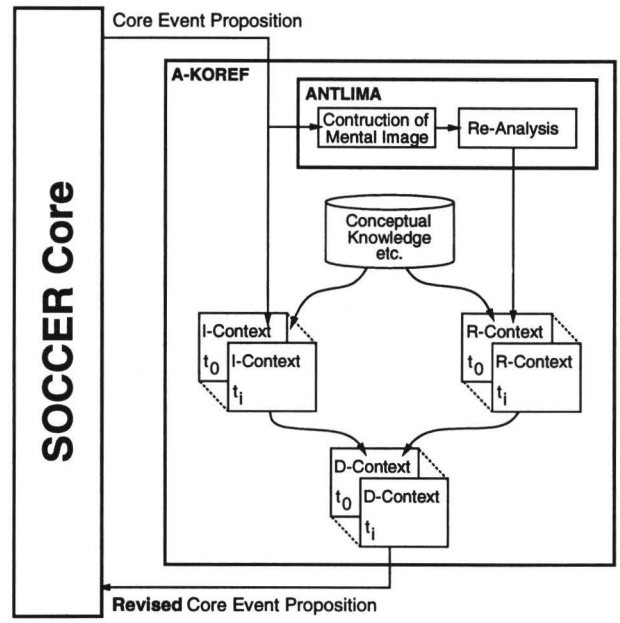



Figure 4: Architecture of ANTLIMA-KOREF

But first, let us take a look at how the static locative propositions are obtained. The recognition of elementary static spatial relations in SOCCER is initiated on request only, e.g., during event recognition in order to establish further evidence for a particular event instance. The construction of the locative parts of I- and R-Contexts therefore has to request for corresponding propositions, as well. This obviously poses a problem of efficiency: As we consider more than 20 movable objects – not to mention the static parts of the field – and about 10 static locative binary relations (excluding for the moment the non-intrinsic readings of the projective prepositions and relations like 'between' which all take three arguments), we have to face for every utterance 10 * 20 * 20, i.e., 4000 as a lower bound on the number of possible propositions, the degrees of applicability of which would have to be calculated twice, for the speaker's percept and for the anticipated mental image. Although the calculation is relatively fast for one such proposition, the high number calls for some intelligent reduction (cf. [Blocher, 1994]).

First, we consider only 'simple' binary relations: (a) the topological relations 'in', 'at', 'close to', and 'near', and (b)

[5]I.e., an image with high typicality could be constructed by ANTLIMA.

the projective relations 'in front of', 'to the right of', 'behind', and 'to the left of' in their intrinsic readings. Furthermore, as subject of a relation (LO – object to be localized), only those objects are considered that participate in the event proposition considered: That is, the objects used as the fillers of the corresponding obligatory deep cases. The number of these objects is usually highly restricted (three on the average). Finally, merely those reference objects (ROs) should be looked for that are in the same 'visual focus of attention' as the subjects of the relations, i.e., located within a certain diameter.

In the cases we examined, the number of reference objects focused was five or less; thus, these strategies sum up to reduce the effort to $(8 * 3 * 5) = 120$ propositions to be calculated twice, and then set into relation. Fig. 5 illustrates part of the resulting descriptions for the object in the circle (S2-Right-Defender): Only the propositions printed boldly would be used in the comparison between the intended and the anticipated understanding.

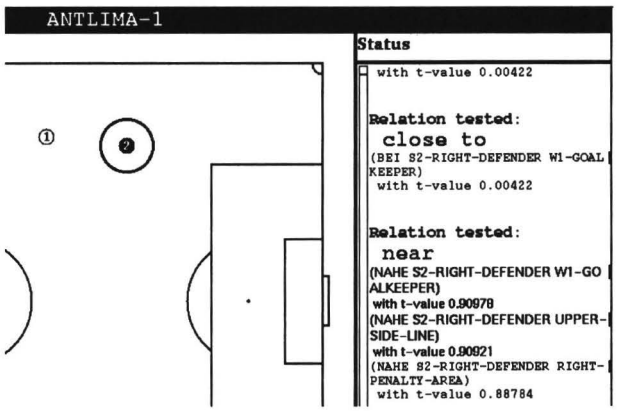



Figure 5: Example of 'focused' calculation of relevant spatial relations (only partial result)

## 5 Comparing the Intended and the Anticipated Understanding

For relating the propositions forming the intended understanding with the equivalent propositions composing the anticipated effects, precisely one of three types of criteria is taken into consideration: Two propositions are associated as an element of the D-Context if they differ either in the relation, the reference object or the degree of applicability DA (cf. Table 1).

| type of difference | | source | relation | LO | RO | DA |
|---|---|---|---|---|---|---|
| 1 | relation | SOCCER<br>ANTLIMA | near<br>at<br>≠ | Miller<br>Miller<br>= | Smith<br>Smith<br>= | 0.95<br>0.92<br>= |
| 2 | RO | SOCCER<br>ANTLIMA | near<br>near<br>= | Miller<br>Miller<br>= | Smith<br>Russel<br>≠ | 0.95<br>0.92<br>= |
| 3 | DA | SOCCER<br>ANTLIMA | near<br>near<br>= | Miller<br>Miller<br>= | Smith<br>Smith<br>= | 0.95<br>0.48<br>≠ |

Table 1: Examples of Criteria of Equivalence for Spatial Relations

More precisely in the first case, only those propositions are associated whose relations are members of the same *conceptual neighbourhood*, i.e., either 'topological' (i.e., depending essentially on distance) or 'projective' (i.e., depending essentially on relative direction). Thus, if the I-Context has high applicability for an object to be *at* some line in the I-Context, and the R-Context has high applicability for the same object to be *near* that same line, A-KOREF has detected a relevant difference (cf. Fig. 7). In this case, usually an equivalence of the third type will be found as well: If that object's being *near* that line is highly applicable, the corresponding proposition with 'at' must have a lower degree of applicability in the R-Context.

The second case is particularly interesting, since it allows the system to notice 'barrier objects', as they are called in linguistics (cf. [Pribbenow, 1993]): An object $C$ is a barrier object for a spatial relation concerning objects $A$ and $B$ if the existence of $C$ at its location decreases the usability of that relation in a description although objects $A$ and $B$ have not changed their positions.

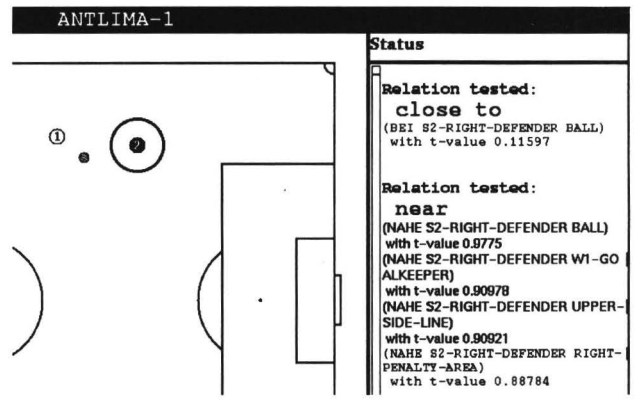



Figure 6: Ball as barrier object for '(close S2 W1)'

For example, a wall between two objects usually breaks down the applicability of 'being near' between those two objects, even if they would be rated as good examples of nearness without that wall (compare Fig. 5 with Fig. 6). This is an effect of the context on the interpretation of a locative phrase, hence part of pragmatics.

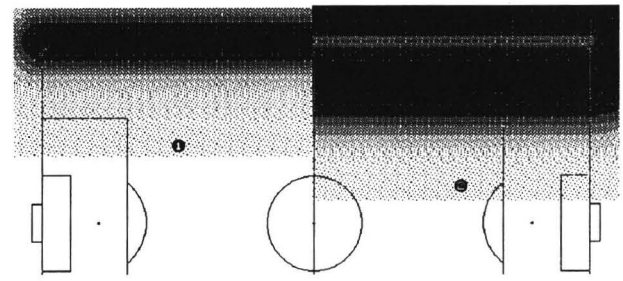


Figure 7: 'at' (left) and 'near' (right) belong to the same conceptual neighbourhood (graphical representation of the corresponding classification functions)

All pairs of associated propositions form the D-Context of the utterance under consideration. They 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 of difference lead to changes in the sentence finally uttered.

## 6 Determining Optional Deep Case Fillers

Following the general pragmatic maxim of SOCCER, i.e., that all deviations from the contextually most typical instances have to be mentioned explicitly, proposition pairs with a sufficient degree of difference should initiate modifications of the proposition chosen. The prominent opportunity for such changes are optional deep cases, particularly the locative forms, like Source, Goal, or Location (cf. [Fillmore, 1978], [Rauh, 1988], and especially [Marburger and Wahlster, 1983]). In fact, those pairs of differing propositions already tell us what kind of fillers can repair the utterance's communicative problem.

If the SOCCER part of the proposition pair has a higher degree of applicability than the ANTLIMA part, 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 unmodified event proposition. In the other case – the proposition from the R-Context has the higher degree of applicability of the pair – the proposition from the I-Context 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; e.g., "Miller stands *to the left of* Scott, *not behind* him."

Thus, our example for a communicative difficulty with respect to correctness is solved; with A-KOREF, we can explain why and when a radio reporter may use optional locative deep cases to insure the correct understanding of his audience. However, there remains one question: Is it not necessary to run the whole anticipation and test cycle again on the modified utterance? Note that the anticipated understanding of the listeners, and particularly the mental image, obviously have to be modified according to the changes in the utterance: After all, it has to provide the actual context for the next utterance; that is, at least the process of image construction has to be repeated. However, a further run of the other components of ANTLIMA, including A-KOREF, seems unnecessary, if the modifications are compatible with all the other (correct) propositions. A complete justification of this problem has not yet been developed. An ad hoc solution to the latter problem is proposed in the course of the example in section 8.

## 7 Anticipated Visual Focus for Attribute Elision

Construction and evaluation of the listeners' mental images were based on the assumption that the problem of reference mentioned in the beginning has already been solved: The listeners are able to uniquely identify the objects in question.[6] The general means of referring to objects that are already mutually known to all participants of a verbal exchange are definite descriptions.[7] For any such object, there is a minimal unambiguous definite description (or even several of them), which usually consists of a set of attributes that separates this particular object from all the others of that context. Sometimes, this set may be rather large; shorter descriptions, however, become ambiguous. In general, the speaker can (and should) use a definite description that is ambiguous in the discourse universe in question if the object actually meant is *focused* by the listener and the alternatives are not. The paradigmatic case for such a focusing occurs if the object meant has already been referred to in earlier utterances of the exchange (cf., e.g., [Jameson and Wahlster, 1982]).

A-KOREF demonstrates a different type of focusing that does not depend on previous mentioning. It is given by the *imaginative visual focus* of an utterance introduced above while calculating the set of relevant spatial relations: By means of this focus, the R-Context contains exactly those spatial implicatures the listener assumedly was able to derive. These implicatures, in turn, form the context of the next utterance: All objects included here are focused by the listeners as well, whether mentioned explicitly before or not. Therefore, the uniqueness of the reference of a definite description need not be bound to the whole discourse universe with all the players, the ball, and the parts of the soccer field; the much smaller set of objects involved in the actual imaginative visual focus 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. As was mentioned above, this set of focused objects usually included around five objects in the cases we examined.

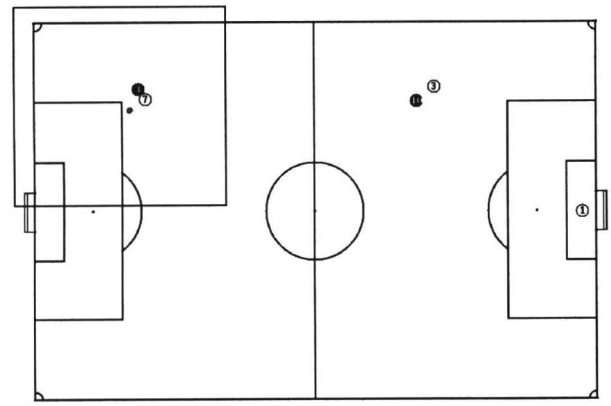


Figure 8: Player W7 and the ball are the anchor points of the focus

Thus, the speaker can use, for example, the expression 'the penalty area' instead of 'the left penalty area' without lack of understanding, even if the penalty area meant was never explicitly referred to before. As long as this particular penalty area is the single penalty area in the imaginative visual focus that was ascribed to the listener while determining whether the *previous* utterance needs some additional deep cases (cf. Fig. 8), the attribute 'left' can be elided.

## 8 An Example Instantiation of A-KOREF

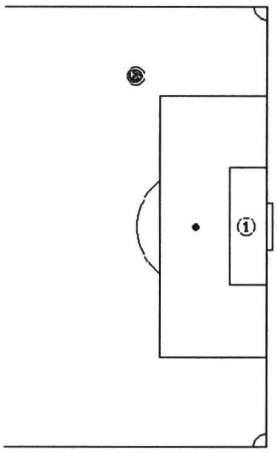
The following three sentences are part of a longer example produced by SOCCER. The modifications initiated by ANTLIMA and A-KOREF are emphasized:

[6]Without that precondition, the listeners would not be able to construct propositions to be interpreted referentially; of course, the speaker must have these propositions before he has the complete utterance; therefore, he may deal with the problem of reference after the problems of plausibility and correctness.

[7]We here ignore reference by proper names and deictic or indexical expressions.

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.*

As mentioned above, ANTLIMA uses a set of standard positions when objects are mentioned first; this happens when a description is commenced. Thus, the image for the first utterance uses as its context standard positions of the goal keeper (in his goal area) which happens to be correct, and of the ball (at the middle point) as the context: The ball's position then is correctly changed by the hill- climbing procedure. In this case, A-KOREF merely states that everything works fine with the utterance proposed. For the second utterance, Moll's standard position has to be used, since this player was not mentioned before: However, as is indicated by Fig. 10, this position is quite different from the one perceived (cf. Fig. 9).



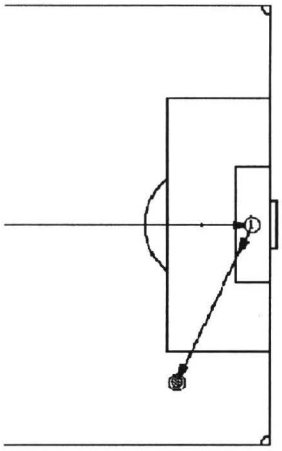


Figure 9: What is seen and... Figure 10: ...what is imagined

Correspondingly, A-KOREF is able to construct a non-empty D-Context, since the same relation ('near') is applicable to different reference objects in the two contexts compared:

I-Context: *((near W3 upper-side-line) 0.79)*
R-Context: *((near W3 lower-side-line) 0.84)*

This leads, as explained above, to the additional prepositional phrase (PP) *"near the upper side line"* in the second utterance. Since the mentioned side line is not part of the visual focus of the previous utterance, attribute elision cannot apply.

In order to be usable as the correct context for the next utterance, these modifications have now to be integrated in the anticipated understanding of the listeners. By hill-climbing, the position of the object in question is changed so that not only do the additional relation hold, but also all those relations that have already held before correctly. Thus, no new difference propositions are constructed.[8]

In the example, additional to relation *(near W3 upper-side-line)*, the relations *(in W3 right-half-field)* and *(near W3 right-penalty-area)* are taken into account. These three relations generate the compound typicality distribution of Fig. 11 and result in the revised position of player W3, shown in Fig. 12.

[8]This mechanism has not yet been integrated in A-KOREF.

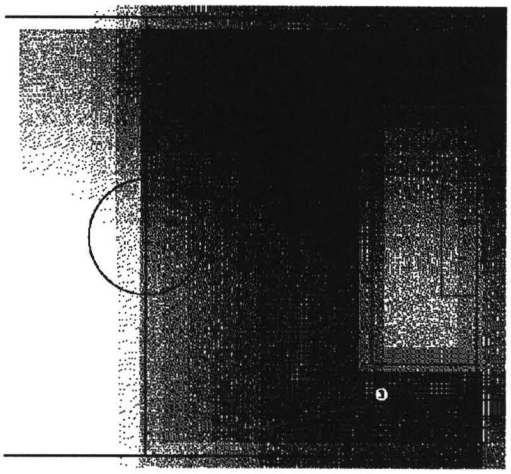


Figure 11: Combined Typicality Distribution for Revising the Position of Player W3

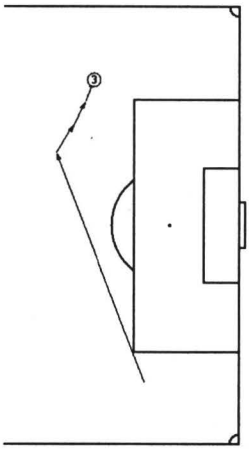


Figure 12: Resulting Revised Position of W3

The position found obviously corresponds extremely well to the speaker's perception. If the two additional propositions are not taken into account, player W3 – being unmentioned and contextually unbound – would be located first in the center of the upper side line since in his standard position hill-climbing based on *(near W3 upper-side-line)* cannot be started (the gradient is 0.0 – we use the center of all given reference objects as a heuristic for such cases). From this interim position, the player would be moved by hill-climbing to the closest 'near'-location, either inside or outside the soccer field, violating at least one of the two other restrictions.

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-KOREF finds the highly relevant difference (first case of classification in Table 1):

I-Context: *((close W11 left-penalty-area) 0.97)*
R-Context: *((at W11 left-penalty-area) 0.99)*

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 – *at* the left

penalty area (cf. Fig. 13). 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.

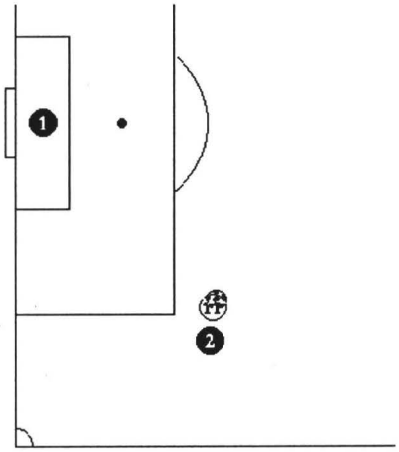



Figure 13: Focused Region for utterance 7

Correspondingly, the definite description 'the penalty area' can be assumed to be uniquely identifying in this situation the object intended, although it is ambiguous with respect to the whole discourse universe. Thus, A-KOREF suggests for utterance 8 the additional PP 'at the penalty area', although none of the penalty areas ever were mentioned before during the description.

## 9 Conclusion

The interaction between the core of SOCCER and its listener model ANTLIMA as mediated by ANTLIMA-KOREF allows us to demonstrate how reference semantics and the resulting conception of mental images can be used to answer some communicative questions. Motivations for adding necessary locative information or reducing superfluous referential attributes can be based elegantly on a purely imagined visual focus of attention applied to the anticipated understanding of the listeners. Further effects on the selection of event propositions and the initiation of self-repair were investigated but not yet realized; they have been described elsewhere (cf. [Schirra, 1994, Section 12.1.1f.], [Blocher, 1994, Section 5.1.2f.]). The systems are implemented in Common Lisp with CLOS and CLIM. Concerning temporal effort, the complete cycle of recognition, selection, anticipated imagination, re-analysis, comparison, and generation of the modified text (without the low-level vision parts of ACTIONS) ran for the extended version of the example sketched in section 8 with factor 3 with respect to the real time in the scene described.

## References


[André *et al.*, 1989] E. André, G. Herzog, and T. Rist. Natural language access to visual data: Dealing with space and movement. Report 63, Project VITRA, Department of Computer Science, University of Saarbrücken, Saarbrücken, 1989. Presented at the 1$^{st}$ Workshop on Logical Semantics of Time, Space and Movement in Natural Language, Toulouse, France.

[Blocher, 1994] A. Blocher. KOREF: Zum Vergleich intendierter und imaginierter Äußerungsgehalte. Memo 61, Project VITRA, Department of Computer Science, University of Saarbrücken, Saarbrücken, 1994.

[Fillmore, 1978] 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.

[Grice, 1975] H. P. Grice. Logic and conversation. In P. Cole and J. L. Morgan, editors, *Syntax and Semantics*, volume 3, pages 41–58. Academic Press, New York, 1975.

[Herzog *et al.*, 1989] G. Herzog, C.-K. Sung, E. André, W. Enkelmann, H.-H. Nagel, T. Rist, W. Wahlster, and G. Zimmermann. Incremental natural language description of dynamic imagery. In W. Brauer and C. Freksa, editors, *Wissensbasierte Systeme (3. Internationaler GI-Kongreß, München)*, pages 153–162. Springer, Berlin, 1989.

[Jameson and Wahlster, 1982] A. Jameson and W. Wahlster. User modelling in anaphora generation: Ellipsis and definite description. In *Proc. of the 5 $^{th}$ ECAI*, pages 222–227, Orsay, 1982.

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

[Nagel, 1988] H.-H. Nagel. From image sequences towards conceptual descriptions. *Image and Vision Computing*, 6:59–74, 1988.

[Pribbenow, 1993] S. Pribbenow. *Räumliche Konzepte in Wissens- und Sprachverarbeitung: Hybride Verarbeitung von Lokalisierung*. Deutscher Universitäts-Verlag, Leverkusen, 1993.

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

[Retz-Schmidt, 1991] G. Retz-Schmidt. Recognizing intentions, interactions, and causes of plan failures. *User Modeling and User-Adapted Interaction*, 1:173–202, 1991.

[Retz-Schmidt, 1992] G. Retz-Schmidt. *Die Interpretation des Verhaltens mehrerer Akteure in Szenenfolgen*. Springer, Berlin, 1992.

[Schirra and Stopp, 1993] J. R. J. Schirra and E. Stopp. ANTLIMA – a listener model with mental images. In *Proc. of the 13 $^{th}$ IJCAI*, Chambéry, 1993.

[Schirra, 1992] J. R. J. Schirra. A contribution to reference semantics of spatial preposition: The visualization problem and its solution in VITRA. In C. Zelinsky-Wibbelt, editor, *The Semantics of Prepositions - From Mental Processing to Natural Language Processing*. Mouton de Gruyter, Berlin, 1992.

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