A Contribution to Reference Semantics of Spatial Prepositions: The Visualization Problem and its Solution in VITRA

Jörg R.J. Schirra
SFB 314, VITRA
Fachbereich 14 - Informatik
Universität des Saarlandes
D - 6600 Saarbrücken 11
Federal Republic of Germany
e-mail: joerg@cs.uni-sb.de

Second Revised Version, May 1, 1992

Abstract

The cognitive function of mental images with respect to the referential aspect of language is examined and used in the listener model ANTLIMA of the natural language system SOCCER. An operational realization of the reference relation used to recognize instances of spatial concepts in the results of a vision system and also to visualize locative expressions is presented and compared to A. Herskovits' analysis of the semantics of spatial prepositions.

1 On Reference Semantics in AI

In AI research concerning natural language systems, the reference aspect of verbal expressions very often plays only a minor role. Nevertheless one has to consider that every verbal expression refers to something, and that the structure of this *something* has an influence on the use of that verbal expression, i.e., on its meaning. In other words: the meaning of any verbal expression is somehow *anchored* in the corresponding referents. This extra-linguistic influence is especially recognizable if we study the interactions of a natural language system with the world by means of sensor and motor systems, e.g., a vision system. But also in machine translation, considering the referents will help to overcome the gap between the conceptual systems of two languages which has so often trapped even approaches with a relatively deep semantic analysis of the source texts.[1] The (more or less) unique referent might serve as a *fixed point* during the transformation to the goal language: changes in the meaning structure necessary due to the

[1] cf. [ZW90a], [ZW90b], [Bat90], and [Gra90];

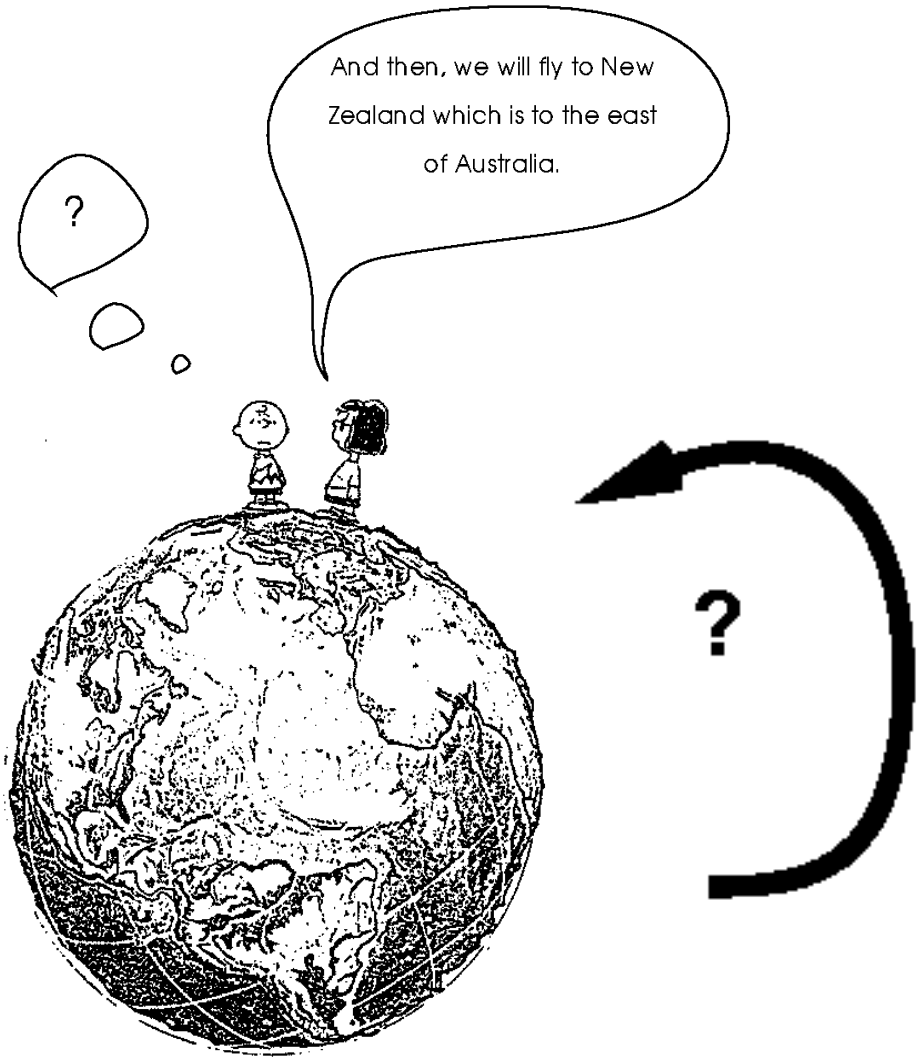

Figure 1: How Can the Reference Relationship Overcome the Distance?

different conceptual systems underlying the two languages are restricted to valid interpretations of that referent.

The major question arising in this context is: *What is the nature of the reference relationship?* which can be split into (a): *What are the referents?* and (b): *How does their influence on verbal behavior work?* I will concentrate herein on a primitive form of spatial prepositions and their referents – essentially geometric relations – although some of these considerations are more general.[2]

Analyses of spatial prepositions in the framework of reference semantics usually view geometric relations as in some sense objectively given, i.e., existing independently in the 'real world', external to any mind.[3] However, a more careful examination leads to the conclusion that the needed referents cannot be provided by the world *per se*. Expressions of fictitious things show us one argument against this *objectivist* view of the reference relationship: there

[2]cf. [Sch90b];
[3]e.g., cf. [DWP81];

are no unicorns in the real world, hence no objectively given referents for the expression *The last unicorn went back into the silent forest.* Is it possible that we can understand this expression even without referents?

There are other simpler disadvantages of the objectivist view. Imagine that we speak about a journey around the world we want to take in the future. While standing in the middle of Europe, we might use a sentence like: *Afterwards, we will fly to New Zealand which is to the east of Australia.* Here, we do not have the problem of nonexistent referents: 'to the east' clearly refers to that de facto spatial relation we could perceive if only we would look for example from a spaceship at that part of the earth. But how can this referent have any influence on the dialog, i.e., on the use or meaning of verbal expressions, miles away on the other side of the earth? Thus, not only is the nature of the referents in some cases unclear – to say the least (cf. Fig. 1). Also our second question about the kind of influence referents have on verbal behavior remains essentially unanswered.[4]

The alternative *experiential* view of the reference relationship assumes that referents are essentially *percepts* and therefore always mental constructs.[5] Thus, spatial prepositions refer mainly to *visually perceptible* relations between objects. Now, there are no obvious problems concerning the mechanisms of influence, as long as we only speak about perceived things. But the examples above still remain problematic. How can the spatial relation between Australia and New Zealand influence a dialog *without being perceived*? We seemingly need some kind of pseudo-percepts, if we speak about something not present or fictitious. It is this very nature of percepts as being mentally constructed which allows for constructing other mental entities and using them as substitutes for percepts. Obviously, we have to speak about these hypothetical mental entities in just the same manner as about percepts – which is exactly the way we speak about *mental images*.[6] Thus, if we consider the reference relationship experientially rather than from an objectivist point of view, it is possible to better understand visual mental images and their cognitive function: mental images have the cognitive function of making available visual referents in the case where, as for radio reporting of a sports event, they are not directly perceptible (cf. Fig. 2).

Dealing with reference semantics in AI, one has to comprehend that the reference relationship consists of three parts: (a) the referents, (b) the propositions, which are the mental representatives of the meaning of verbal expressions, and (c) the connection between the two, the reference relation proper, so to speak.[7] These three entities all have to be modeled in an AI system in the framework of reference semantics (cf. Fig. 3).

The propositional level is already well examined in AI: most formalisms of Knowledge Representation can be used for this purpose. On the other hand, the referent level – because of its

[4]General arguments against every realistic approach of reference can be found in [Emp85] and [Wit63]; both authors base their arguments on a reflexive turn: asking how we could refer to the reference relationship itself they are led to sceptical consequences which 'destroy' the presuppositions of objectivist reference theories, namely the possibility of access to an independently given world;

[5]cf. [Lak87] and [Joh87];

[6]I use the expression not only to refer to visual images although only those are considered in this paper; since they are connected so closely to perception, mental images seemingly can arise in all modalities of perception; cf. [Lak87, p. 444]: "The term 'image' is not intended here to be limited to visual images. We also have auditory images, olfactory images, and images of how forces act upon us."

[7]By the way: 'propositions' in this context might best be related to 'symbolic cognitions' in the sense of Leibniz (cf. [Lei37, §24]); in connection with corresponding referents, they become 'adequate cognitions'; the expression 'reference relation proper' corresponds closely to the term 'notion' in the traditional terminology; cf. [Sch93, Part I];

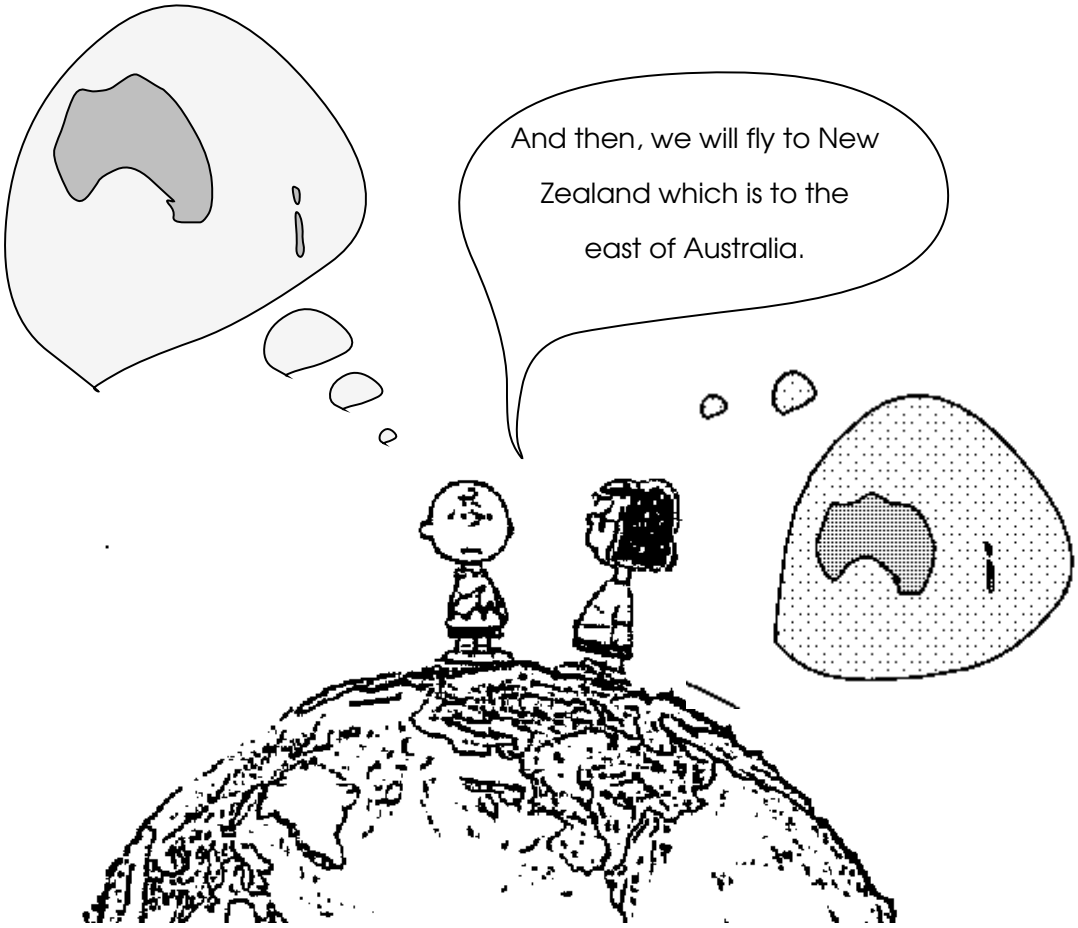

Figure 2: Mental Images as Substitutes for Percepts

connection to perception – has to be based on data structures used in corresponding perceptual systems, e.g., vision systems.[8] At this point, one confusion of notions very often appears, blurring further discussions: since the results of perceptual systems in AI even on a very low level of processing are usually represented in a form similar to the above-mentioned formalisms used for the propositional level – sometimes even the very same knowledge representation languages are used – they are called propositional, as well. This use of the word 'propositional' is different from the one introduced above. In this very broad sense, everything expressed in the knowledge representation formalism is called a proposition.[9] In my terminology, only those expressions which are connected to referents by a reference relation are propositions. The difference between referents and propositions is not a question of form but of use. Referents have meaning, namely the propositions associated with them, but they do not refer to something else, e.g., an extra-mental set-theoretical model named 'the world'. Similarly, propositions do not have meaning; they only refer to something. There is neither a meta-level above the propositions which might include the meanings of propositions, nor is there a sub-level which could keep

[8] e.g., cf. [Mar82], [Sun88], [HSE+89], and [MF88];

[9] cf. [Pyl81]; this usage is closely connected to the procedural/declarative distinction; cf. [Win75]; cf. also the distinction of different uses of the expression 'proposition' in [Joh87, p. 3];

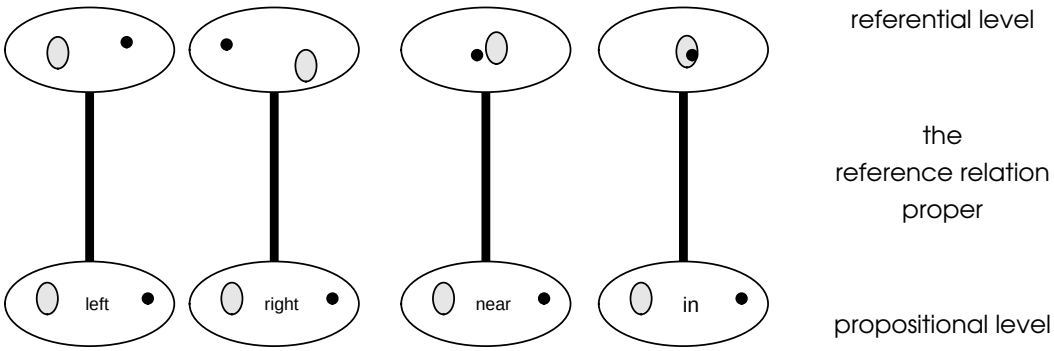

Figure 3: The Reference Relationship — A First View

ready those objects the referents refer to.[10]

The third and most important entity of the trinity of reference, the connection between referents and propositions, quite obviously has a very different character. In order to avoid an infinite regress, the realization of the reference relation proper cannot use propositions or referents (percepts) again. Actually, the system does not necessarily know anything about the structure of the reference relation, e.g., for a spatial proposition. It can only use its realization of the reference relation in order to establish the connection between referent and proposition, i.e., recognize spatial relations in a given percept or visualize spatial relations (construct the referent). Therefore, a *procedural* realization of the reference relation proper seems adequate.

In the following sections, I will present an overview of the system SOCCER with special emphasis on how language can be grounded in visual perception using a simple kind of artificial percepts and how mental images of the same simple kind can be reconstructed in a model of anticipated understanding. The reference relation is used for, and its realization restricted by, both purposes. Section 6, finally, presents the realization of the reference relation we have chosen in VITRA.

2 The Project VITRA and the System SOCCER

The project VITRA (VIsual TRAnslator) which started in 1985 as part of the German special collaboration programme SFB 314, *AI & Knowledge-Based Systems*, examines the relations between speaking and seeing: a completely operational form of reference semantics for the visually perceived is to be developed. CITYTOUR and SOCCER are two systems constructed in VITRA which – broadly speaking – transform visual perceptions into language. Here, we will concentrate on SOCCER.[11]

[10]This does not exclude (mental) entities to be used at one time as referents and at another time as propositions. This is typically demonstrated by (percepts of) sketches, e.g., of the sun: the sketch itself is a concrete thing which can be perceived and used as referent – as for example in this sentence. But normally, we do not look at the sketch itself, but at what it represents. We use the sketch as a set of propositions about something else – the referent of the sketch which is the sun in our example (cf. [Eco72, esp. p. 208]). Similarly, when speaking about percepts the corresponding mental entities are used as propositions: the percepts *of something*. In this article, no such metamorphosis is considered.

[11]for Citytour cf. [ABHR85], [ABHR86a], [ABHR86b], [SBSZ87], and [RS88];

SOCCER simultaneously analyses and describes in German short scenes from soccer games similar to a live radio report, i.e., simultaneously and in an objective manner to an audience which is not able to see the game themselves. For this purpose, a large number of quite complicated 'cognitive' activities has to be performed: e.g., perceiving the locations and movements of ball and players in the scenery, interpreting these movements with respect to the conventions of soccer games, especially assuming the intentions and plans of the agents in the field, and – last but not least – selecting which events to utter in which sequence and with which words.

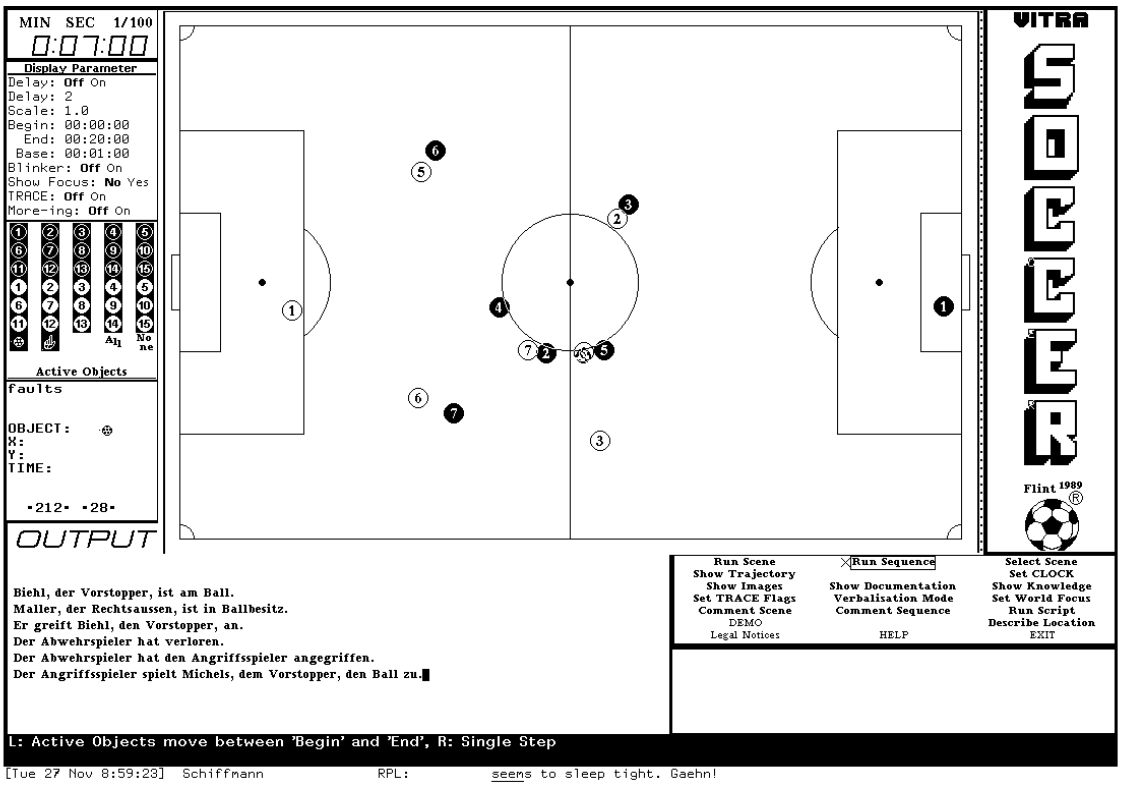

Figure 4: What SOCCER Shows the User

The input data of SOCCER, which in a way are its percepts, are called mobile object data – *MOD* – and generated by the motion analysis system ACTIONS.[12] These percepts consist of the set of the two-dimensional spatial locations and the velocity vectors of every mobile object perceived in the soccer field from a bird's eye view. At every time quantum, ACTIONS delivers the corresponding data which, then, is entered successively into SOCCER. At present, all mobile objects are perceived as ideal points (zero-dimensional). The MOD is analysed successively as soon as it is entered. It implicitly refers to the geometry of the soccer field which is known by the system as *StaB* – static background (cf. Fig. 4).

SOCCER does not know the whole scene at once. Like a radio reporter, it has to analyse the scene during its occurrence. Therefore, all processing steps have to be done incrementally, i.e., a selection of already recognized events is verbalized simultaneously with further event recognition. Indeed, SOCCER already recognizes events before they have occurred completely.

[12] cf. [Nag88], [Sun88], and [HSE+89];

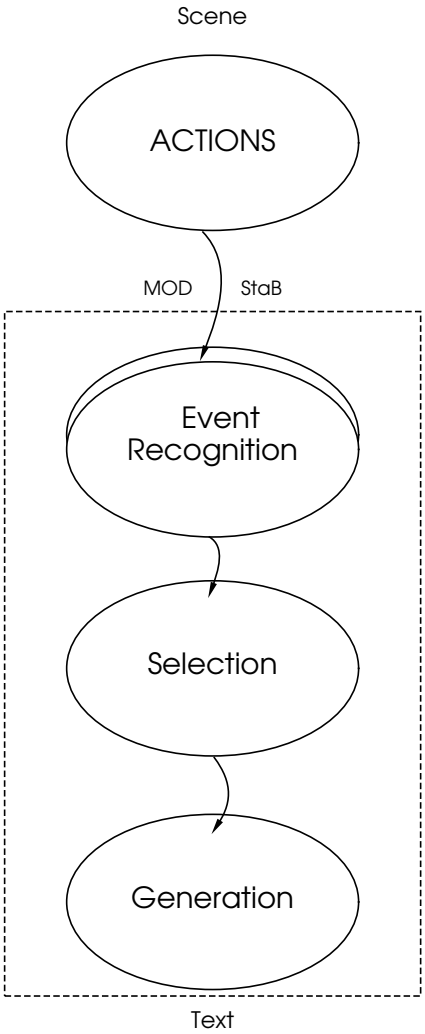

Figure 5: SOCCER: The Core System

The power, but also the limitation of a rigorous grounding of language in perception is already demonstrated by the core system (cf. Fig. 5): in a kind of pipelining, three components transform the perceptual data into a text similar to the so-called protocol sentences of Vienna Circle philosophy:[13] only directly sensed impressions are reported – indeed still too primitive a type of sports report. For example, compare text a with b, both of which describe the same referent.

a: *Schmidt, the goal keeper of the Blacks, is standing in the left penalty area. Meier, the captain of the Reds, is running along the middle line. The ball is close to him. Now, the distance between the ball and Meier increases. The ball crosses the left half of field and comes near to Schmidt. Schmidt starts moving towards the ball. Now, the ball stops moving when it reaches Schmidt.*

b: *Meier, the captain of the Reds, has the ball and is running along the middle line. Now, he tries to score with a long shoot, but Schmidt, the goal keeper of the Blacks, catches the ball in his penalty area.*

The first component of the core system, SOCCER's EVENT RECOGNITION, has two parts: first, elementary spatio-temporal relations are recognized in the current percept: propositions like (left player-5 player-7) or (greater-velocity ball 45) are constructed by algorithms which are procedural to SOCCER. The core of these algorithms is always a graded classification function that associates visual percepts used as input to SOCCER with abstract spatial relations by so-called applicability degrees (cf. Section 4).

As a second step, these elementary relations are chunked into propositions describing more abstract relations: (running-with-the-ball agent: player-5 place: (in-front-of RightGoal)) or (running-parallel agent: player-7 co-agent: player-2 direction: (along MiddleLine)). In contrast to the elementary relations, SOCCER here uses knowledge in a declarative form: for each composed event, a so-called event model defines how elementary relations have to be combined to yield events of that type. The event recognition component works in a quasi-parallel way, thus recognizing simultaneously all spatio-temporal relations or events similar to a human observer.[14]

[13] cf. [Car33] and [Neu33];
[14] SOCCER event recognition: cf. [HR88];

While the event recognition is working, the SELECTION component already chooses subsets of the recognized propositions to be uttered and orders them into a queue. The head of this queue is passed to the GENERATION component as soon as it is idle. The time of generating the sentences is important since the order in the queue might be changed if new events which seem to be more important have been recognized in the meantime. Previously selected items might even be removed from the queue.

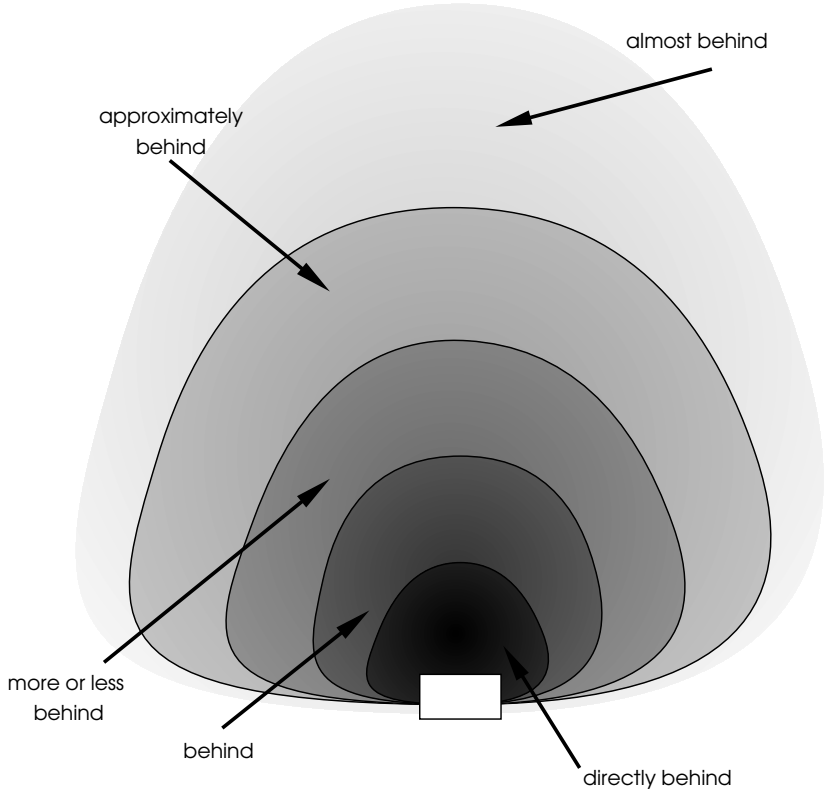

Figure 6: Applicability Degrees and Linguistic Hedges

The generation component transforms the event proposition chosen by the selection component into a continuation of the report. Here, the applicability degrees of elementary spatio-temporal relations are verbalized, using linguistic hedges (cf. Fig. 6).[15] Localizing phrases and even previously mentioned events can be used to disambiguate objects. Furthermore, anaphora are used to increase the coherence of the text.[16]

Obviously, the SOCCER core system is rather primitive compared to human perception and cognition: it sees only two spatial dimensions, namely soccer fields from a bird's eye view, and perceives players as zero-dimensional without inherent orientations. Correspondingly, its language use is restricted: spatial prepositions for example cannot be used in the differentiated way we use them. Especially, there are no metaphoric extensions of spatial prepositions. On the other hand, these limitations simplify the problems of realizing the reference relationship of spatial prepositions to a treatable complexity, and thus, even may serve as a base for further studying spatial metaphors.

[15] cf. [Lak72];

[16] A detailed description of both selection and generation components is to be found in [André 88];

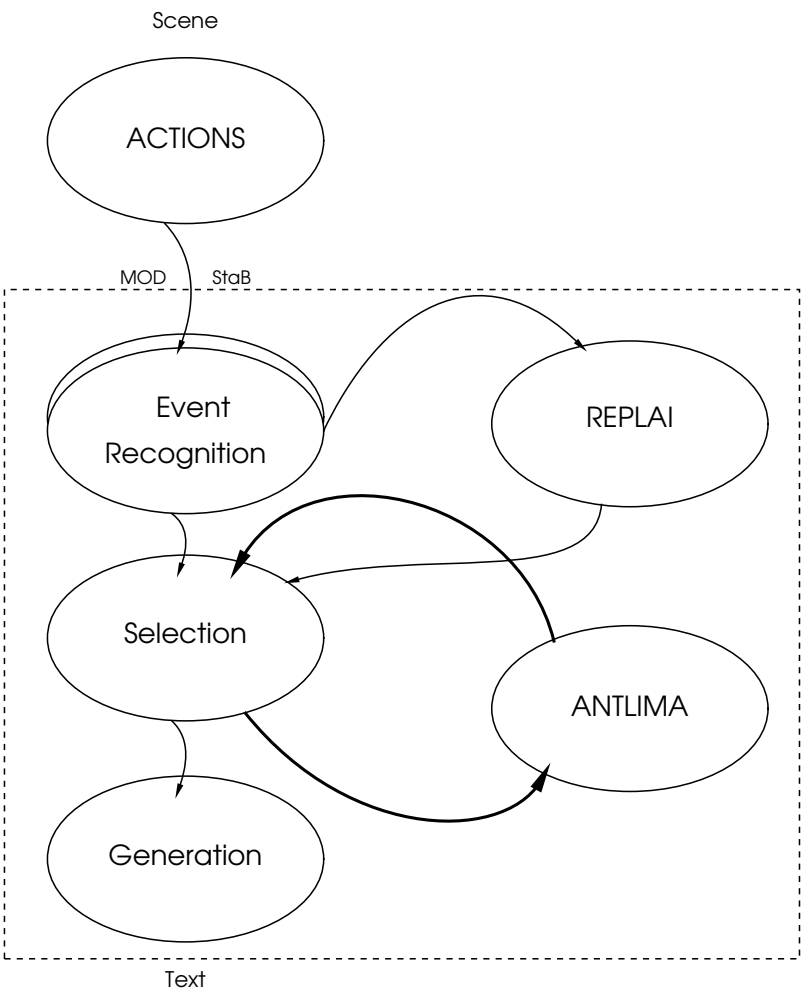

Figure 7: Extended Architecture of SOCCER

To overcome the handicaps of pure protocol sentences, the intention and plan recognition component REPLAI-II[17] is added to the SOCCER core system (cf. Fig. 7): we can actually speak about soccer games only if we assume that the players behave according to internal and not directly observable *mental* states like intentions and plans. If the ball moves into the goal after changing its direction while being very close to a player, SOCCER can only describe this event as *That player scored* if it presumes and tries to verify in its percepts that this player has certain soccer-specific intentions and plans. Although intentions are based in the perceived, as well, they are not totally grounded in percepts and require additional knowledge to be (hypothetically) recognized. Thus, REPLAI-II extends the strict reference semantics of the SOCCER core system.[18]

Another extension of SOCCER is provided by its listener model.

[17]cf. [RS91b];
[18]More details cf. [RS91a];

3 The Listener Model ANTLIMA

In order to follow Grice's Cooperative Principle,[19] a speaker has to know how his utterance is understood by the listeners in the present context. He needs a model of the listener, e.g., to make sure that despite of the maxim of economy[20] the listeners are still able to recognize all the relevant structures even from elliptic descriptions. Thus, a listener model serves as a device which balances between the divergent demands of economy and completeness.[21] With this knowledge about the listeners, the speaker can rate how much information actually is *required* in the given case.

Correspondingly, SOCCER also needs a component that can construct a model of the listeners' knowledge of the events that have already been described. This listener model enables the system to continue its description in a cooperative way by anticipating the listeners' understanding of the utterance just planned. With these anticipations, the *plausibility* of that utterance in the context already known can be rated and used in an anticipation feedback loop to improve the coherence (cf. Fig. 7). Therefore, the main task is to find out whether and to what degree the listeners might be able to understand the planned utterance at all, and, as a second step, whether they understand it as intended.[22]

There is little evidence that the listeners and the speaker use different kinds of semantics. Thus, we assume that the listeners understand the soccer report by reference semantics, as well. What does that mean? As a German linguist wrote in 1969, "the radio reporter has solved his task only if he describes the reality of a sports event so vividly and obviously to the listener that the listener believes he sees that reality."[23] The reporter shall induce – so to speak – a *cinema in the heads* of his audience. This clearly refers to mental images just in the way we mentioned them in Section 1: if the listeners want to have a 'deep' understanding of the report, they need access to the referents and should be able to reconstruct them, i.e., to construct (visual) mental images corresponding to the speaker's percepts.[24]

Since the listener model of SOCCER anticipates how the listeners understand the planned continuation of the report, it also has to ground the meaning of these utterances and especially of spatial prepositions referentially: ANTLIMA – ANTicipation of the Listeners' IMAgery – must be able to reconstruct corresponding visual percepts – albeit in the limited sense of SOCCER as MOD and StaB (cf. Fig. 8). In other words, it must *visualize* the abstractly described situation. It is our thesis that those *pseudo-percepts* reconstructed by SOCCER's listener model correspond to the listeners' visual mental images.[25]

As the backbone of ANTLIMA, the event proposition chosen by the selection component of

[19] "Make your conversational contribution such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged." cf. [Gri74];

[20] "Make your contribution as informative as is required (for the current purposes of the exchange)!" cf. [Gri74];

[21] "Do not make your contribution more informative than is required!" cf. [Gri74];

[22] Additionally, but not described in this report, the listener model is used to control the generation of noun phrases, anaphora, and ellipses; cf. [JW82], [André 88], and [Sch91];

[23] cf. [Dan69, p. 94 (transl. J.S.)];

[24] This does not mean that listeners of broadcasted sports reports *always* generate visual mental images. The reconstruction of the referents constitutes an already very high level of understanding (cf. [CL72], and also [Lei37, §24]).

[25] Though, these listeners are assumed to function mentally in a way corresponding to the simplicity of SOCCER; i.e., they perceive similarly to SOCCER and know more or less the same about spatio-temporal relations and composed events. If radio reporters could not rely on such an assumption of *cognitive equivalence* with their listeners, sports reports would have to be of a very different form than they actually are.

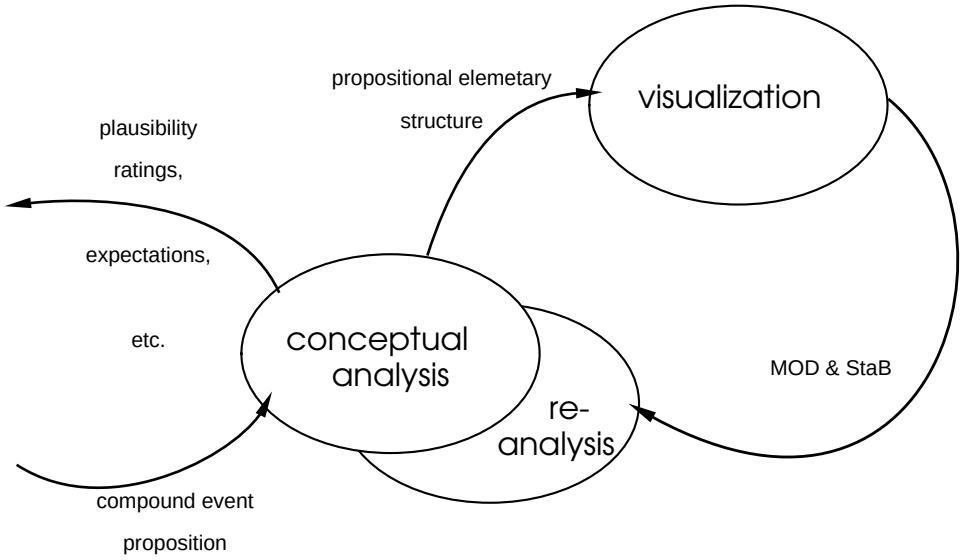

Figure 8: Architecture of ANTLIMA

SOCCER has to be analysed conceptually, i.e., dealing only with propositions. The conceptual analysis provides the basis for the visualization and for the feedback to SOCCER: the definition of the considered event, i.e., its subevents and the spatio-temporal relations between them, must be *expanded* and adapted to the situational constraints in order to achieve spatio-temporal coherence with the context.[26] Additionally, it might be necessary to integrate modifications given by optional deep case fillers.[27] The construction of a corresponding mental image serves as a kind of focusing device for this reasoning.[28] The intermediate result of the conceptual analysis, which is the basis for the visualization, is called the *propositional elementary structure*, the temporally ordered sequence of sets of elementary spatio-temporal relations. Visualization transforms this data to a sequence of MOD in the StaB, thus fixing specific locations and velocities for all considered objects at each time point.

The generated mental image should not be compared directly with the original percept of SOCCER – coordinate by coordinate, so to speak. Although such a comparison seems to be necessary if we want to know whether the listeners will have got the correct referent, the system only compares propositions, not images. Why? In general, we lose information transforming an image to propositions. Therefore, we cannot expect that the listeners will reconstruct the very same picture from that selected set of propositions actually communicated. Their images – and equally the one generated by ANTLIMA – will usually be only more or less similar to the original percept. The question now is: which deviations are essential and which are not? If a player is standing in the middle of the right field – nobody near him – 50 pixels' difference normally will not matter. But if he stands at the edge of the field, or near some other player, or very close to the ball, even 10 pixels' deviation of his location might change the whole interpretation of the scene. Since the propositions just ignore irrelevant details by

[26] cf. [Sch90b];
[27] cf. [Son78] and [MW83];
[28] cf. [Pri88] and [Sch91];

definition (cf. Section 4), ANTLIMA has to re-analyse its mental images first, i.e., describe them propositionally again (cf. Fig. 8, component Re-analysis). Then, it can compare this new set of propositions with the set SOCCER 'found' in the original percept or the set which was actually communicated. Now, substantial deviations result in a different set of propositions, e.g., an additional (outside player-5 SoccerField) or a missing (at player-7 ball). Since these differences might have several causes, among them: (a) error of ANTLIMA during conceptual analysis or visualization, (b) error of SOCCER during selection, (c) expectation of an event which will be communicated next, and (d) implicit event which SOCCER expects ANTLIMA to know about without communicating it, the further processing is quite complicated. As it does not effect the discussion in this report, we do not deal with it herein.[29] Finally, plausibility ratings, correctly expected continuations of the report, and – if necessary – errors are given back to SOCCER's selection component, thus closing the anticipation feedback loop.

Before the elementary visualization is described in greater detail in Section 5, we devote our attention to its counterpart, the recognition of elementary spatio-temporal relations, since this first use of the reference relation gives a clue for the solution of the visualization problem.

4 Recognition of Static Spatial Relations

Dealing with the connection between seeing and speaking, the first of the problems we have to consider is establishing the reference relation at all: we have the task of finding a connection between perceptual and propositional level. In general, the transformation to the propositional level has the function of reducing the amount of information included in the percepts to those features important for further acting – reporting in our case. Therefore, the spatial relations we consider correspond closely to (German) prepositions. The location of an object is relevant only relative to other objects' positions.[30] Other information in the percepts, e.g., the precise coordinates of objects, is regarded as irrelevant, and ignored.

The elementary level of recognition in SOCCER is formed by static spatial relations verbally described by prepositions such as (being) *left*, *at*, and *between*[31] and represented as SPATIAL CONCEPTS. Individual occurrences of a relation are represented by instances of the Spatial Concept and called (spatial) *propositions*. More precisely, spatial propositions are combinations of one spatial relation – the type of the proposition - and a set of objects forming the arguments of the relation. Fig. 9 shows a typical (static) percept for SOCCER. Corresponding to the human uses of spatial prepositions, SOCCER should be able to interpret this percept as a near situation, i.e., by creating a proposition (near player-5 player-7). But it can instantiate the Spatial Concept left-of – with respect to the direction of movement –, as well. Obviously, recognition is not a one-to-one association, i.e., simply combining one percept with one proposition (cf. Fig. 3). Different propositions can stand for the same geometric configuration. The reference relationship in VITRA associates every percept with a set of propositions, all describing that percept (cf. Fig. 10). But which of them should be used to describe the percept verbally?

This decision is supported by the radial structure of spatial concepts: we have to consider that some referents are good, others bad examples of a proposition. In most cases, we can

[29] cf. [Sch91];

[30] The first object is called LO – located object –, the others ROs – reference objects; cf. [HP88]; other terms used in linguistics are listed in [RS88];

[31] actually, of course, in German;

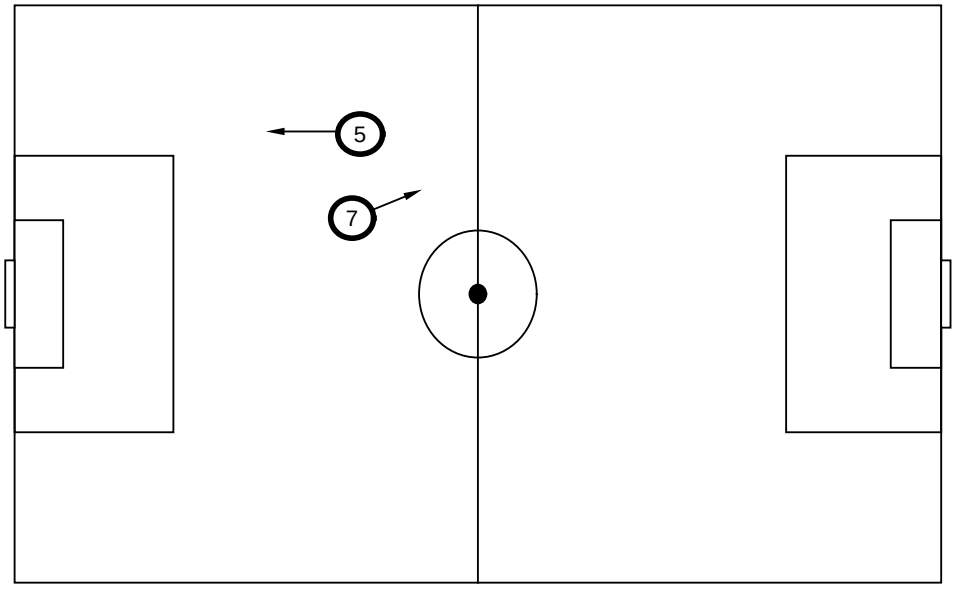

Figure 9: SOCCER Percept with *near* and *left of* Occurrences

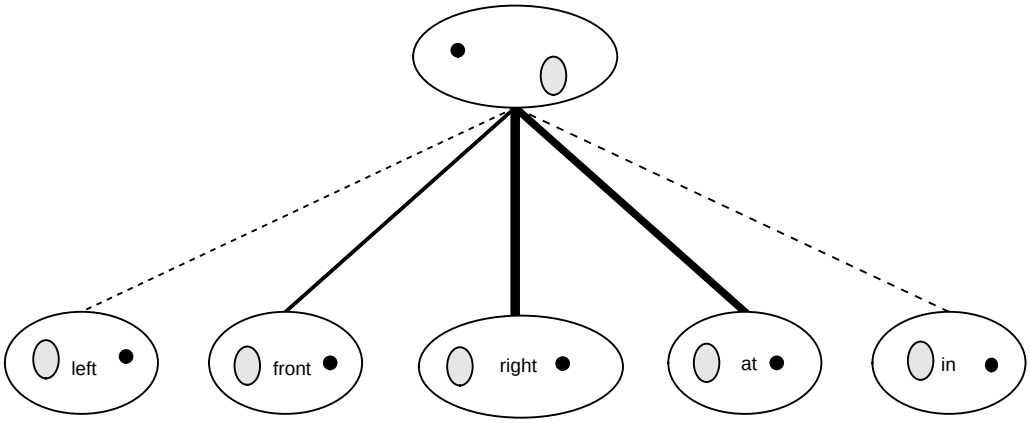

Figure 10: The Reference Relationship (Part 2: Recognition)

move the LO a little bit without changing the propositional description. But the referents gradually become poorer examples of that proposition. For most spatial relations, we find that the probability of their being used to describe a situation changes *gradually* as the LO shifts. Fig.s 11, 12, and 13 illustrate this phenomenon for *in*, *near*, and *in front of* by means of a kind of *probability clouds*, each drawn for two different types of ROs. The dense centers of these clouds mark those positions rated as good examples for the relations.

In VITRA, the probability of use of a spatial proposition which is associated with every possible position of the LO with respect to some given ROs is interpreted as a measure of applicability for the proposition. In Fig. 10, the degree of applicability is indicated by the thickness of the connection. The higher the applicability degree of a proposition is for a given percept, the better this proposition can be used to describe the percept. Therefore, an essential

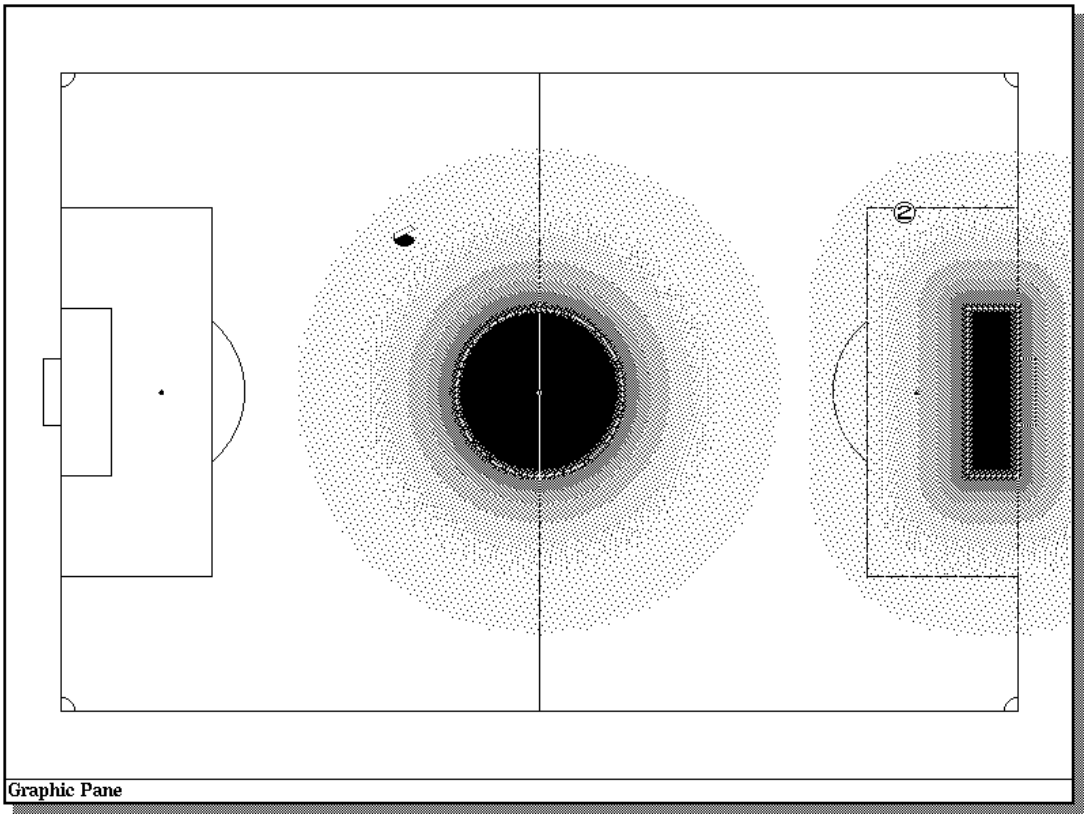

Figure 11: Probability Cloud Representation of 'in' for Two Different Kinds of Objects

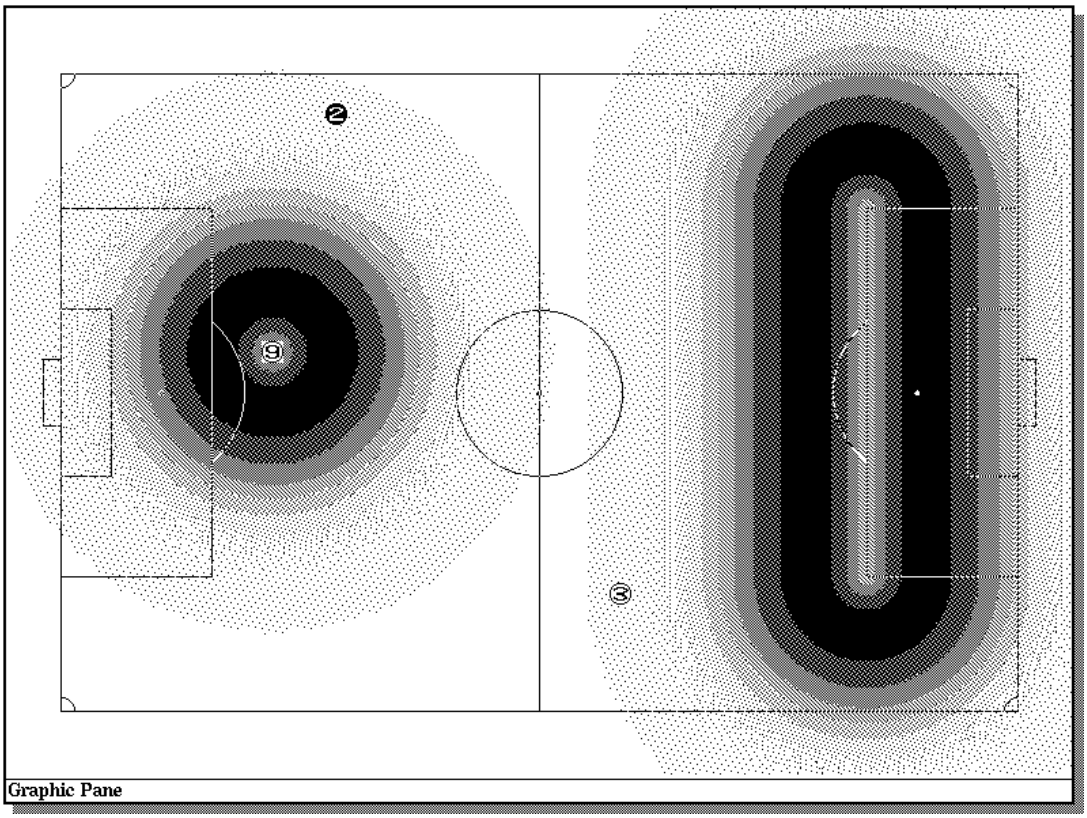

Figure 12: Probability Cloud Representation of 'near' for Two Different Kinds of Objects

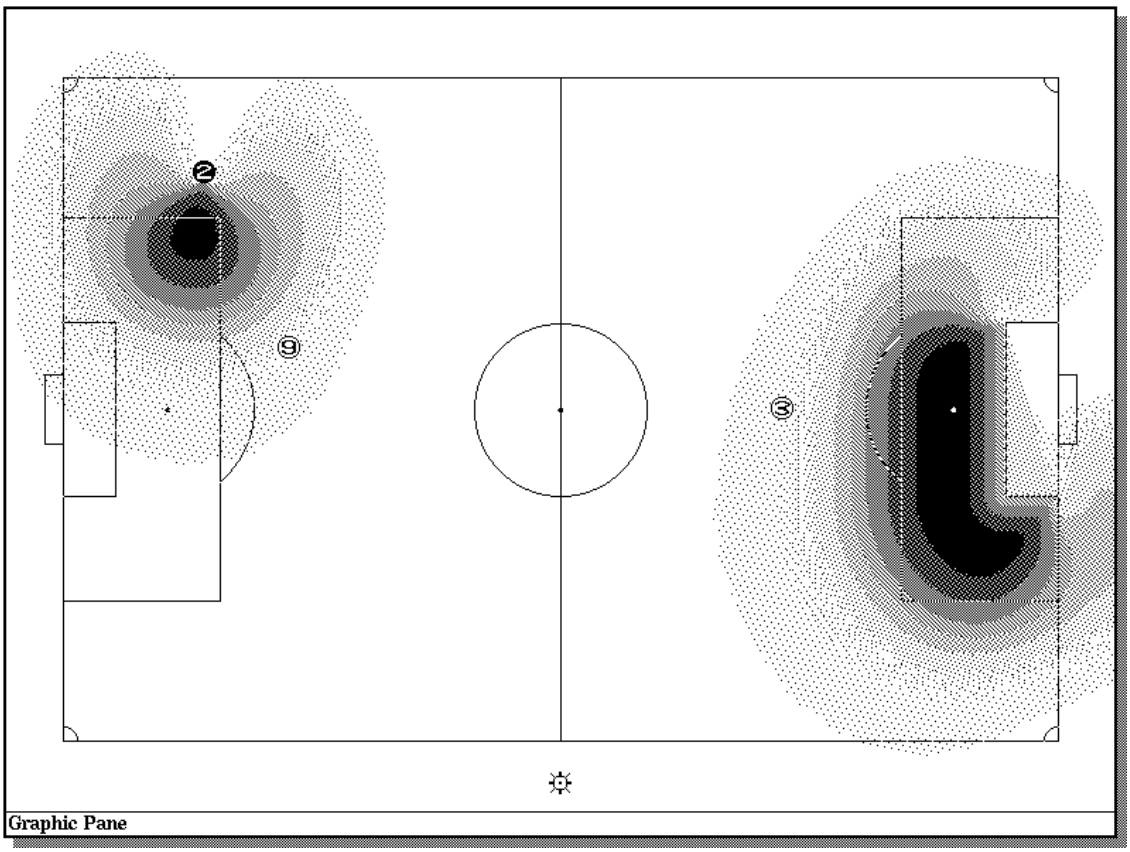

Figure 13: Probability Cloud Representation of 'in front of' (extrinsic use) for Two Different Kinds of Objects

part of the reference relation is encoded by a graded classification function associated with every proposition. Given a percept, these functions yield the applicability degree[32] of the corresponding proposition. For example, the applicability degree of 'at' is calculated by means of Formula 1: the RO is assumed to be at (0,0), the LO at (x, y); d is a scaling factor.

$$\mathrm{A}_{at}\,(\mathrm{x},\,\mathrm{y}) = \mathrm{e}^{-\frac{\sqrt{x^2+y^2}}{d}} \tag{1}$$

The connections between spatial concepts, propositions, percepts, and applicability degrees are shown in Fig. 14 as signatures of abstract data types (ADT). Note that the reference relation and the above-mentioned classification functions are hidden in the operation *recognize*. For reasons described in the next section, the applicability degrees are called *T-values*.

Ambiguity and gradation are both well described in linguistic literature: many investigations about the connection between object locations and applicable spatial relations have been carried out and can be used for our purpose.[33] But, to be sure, the simplicity of SOCCER's percepts does not allow for the full range of use German prepositions show: for example, metaphoric uses even in the spatial domain[34] are totally excluded. Nevertheless, some of the features described by linguists can be and have been considered in SOCCER (cf. Sections 6 and 7).[35]

[32] in [0.0 .. 1.0];

[33] E.g., cf. [Saf66], [Fil71], [MJL76], [Moi79], [Tal83], [Her85], [Van86], [Lak87], [Bie87], and [HHR89];

[34] cf. [SHfc];

[35] This influence is documented for example in: [ABHR85], [ABHR86a], [ABHR86b], [ARH87], [Wah88],

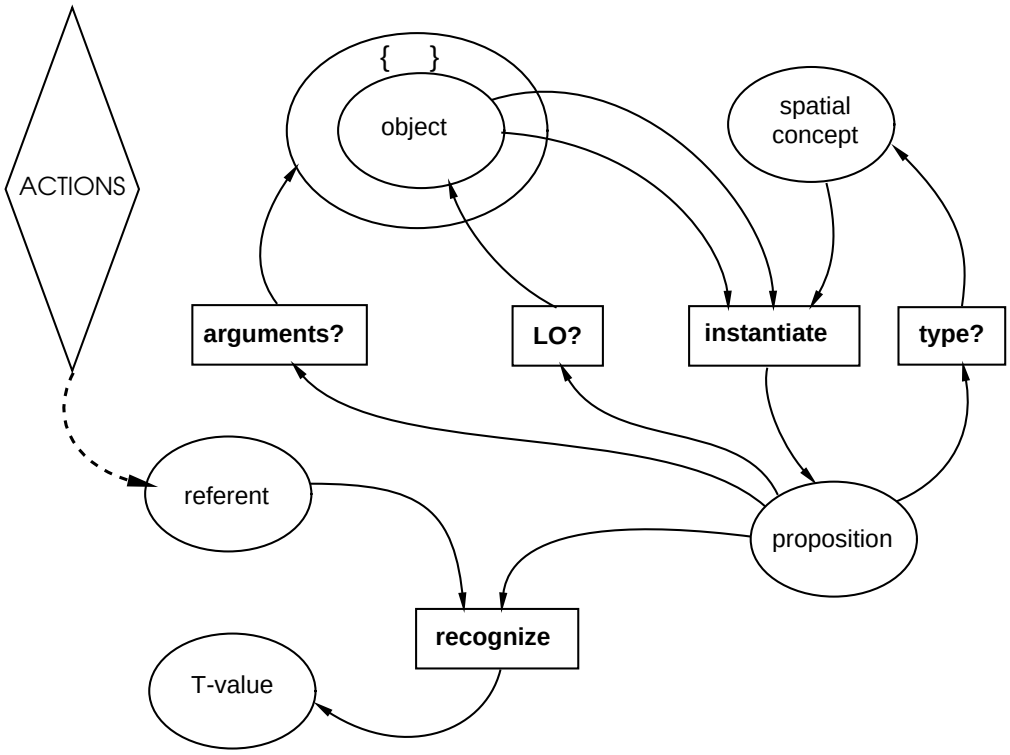

Figure 14: Signatures of the Spatial Concept System (I): Recognition

5 Visualization of Static Spatial Relations

Corresponding to the elementary recognition, the most elementary step of the visualization task is constructing the static image of a set of elementary static spatial propositions which should hold simultaneously,[36] e.g.,

(left player-5 RightPenaltyArea), (in-front player-5 player-7),
(in player-3 RightHalfField), (near player-3 MiddleLine),
(near player-7 OutfieldSide1), (at player-7 player-3),
(near player-7 ball), (between player-7 player-3 player-5),
...

or organized with respect to the LO's:

player-5: [left RightPenaltyArea], [in-front player-7]

player-7: [between player-3 player-5], [at player-3], [near ball], [near OutfieldSide1]

player3: [in RightHalfField], [near MiddleLine]

... ...

[SBSZ87], [RS88], [Sch90a], [Sch89], [Hay90], [Hay89], [AHR89], and [HRA90];

[36] cf. Section 3; the input for the complete visualization task is the temporally ordered sequence of those sets (plus elementary velocity restrictions), called the propositional elementary structure; the complete referent of a composed event proposition is constructed by chunking static images appropriately together to form an image sequence;

This task, called the visualization problem in ANTLIMA, is just the reverse of the aforementioned classification task that is at the heart of perception: instead of abstracting away from a concrete situation, an abstract (propositional) description must be augmented – filled with life – and transformed into a plausible concrete form. Unfortunately, the advantage of classification, namely the possibility of ignoring irrelevant details of the referent, produces for the visualization task a corresponding disadvantage: how, for example, can we fix a concrete (i.e., precise) position of a ball of which we only know that it is to the left of the penalty area?

One way out of this problem is given by the already mentioned gradation of the Spatial Concepts: if we consider the reference relationship not from the viewpoint of a given percept with respect to which every spatial proposition is *applicable* to a certain degree, but conversely from the viewpoint of one such proposition, then several percepts are to varying degrees *typical* examples of that proposition. Whereas the problem of perceiving is essentially the question: *Which (spatial) relation is most applicable to a (visually) given situation?*, the visualization problem can be summarized as: *Which situation (visual pseudo percept) is most typically intended by a given abstract description?* (cf. Fig. 15).

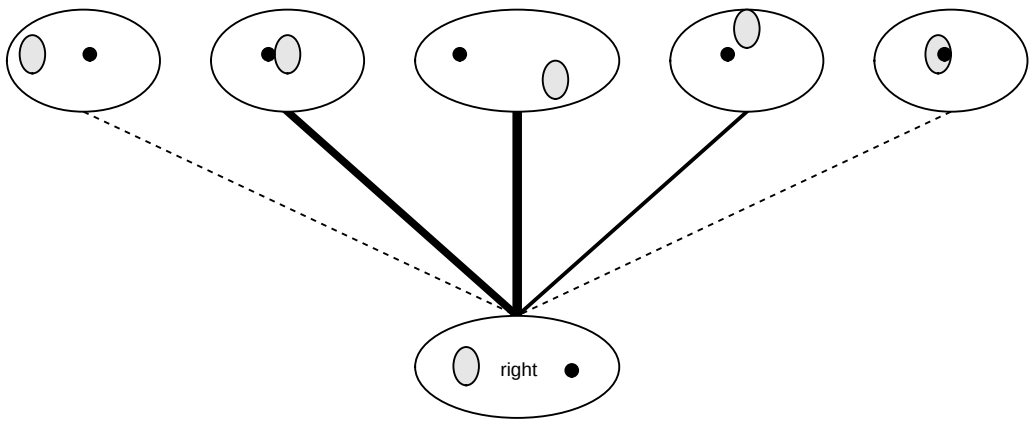

Figure 15: The Reference Relationship (Part 3: Visualization)

Therefore, I assume that listeners – and similarly ANTLIMA – always try to give an utterance its most typical interpretation. Furthermore, they expect that the speaker will explicitly mention any important deviation from the typical case. This again reflects the aforementioned criterion for deciding which of all applicable propositions should be used to describe the percept to the listeners: SOCCER chooses the propositions with the highest degrees of applicability for verbalization.

Starting from a propositional description, that mental image must be constructed that realizes all of the given spatio-temporal relations with maximal typicality. For a given set of ROs, ANTLIMA has to locate the LO at a position where the classification functions, or typicality functions as they better might be called in this context, are maximal. The essential task, then, is finding the maxima of the typicality functions. This task is not too difficult for one proposition. However, we have to consider sets of propositions which should hold simultaneously. Some of the given relations may happen to conflict: they cannot be maximally typical for the same situation. Look, for example, at Fig. 9: if the ball is to be located in the left goal *and simultaneously* near player-7, we are not able to find any location with maximal typicality for both restrictions. In those cases, compromises must be calculated: a renormalized addition of all the

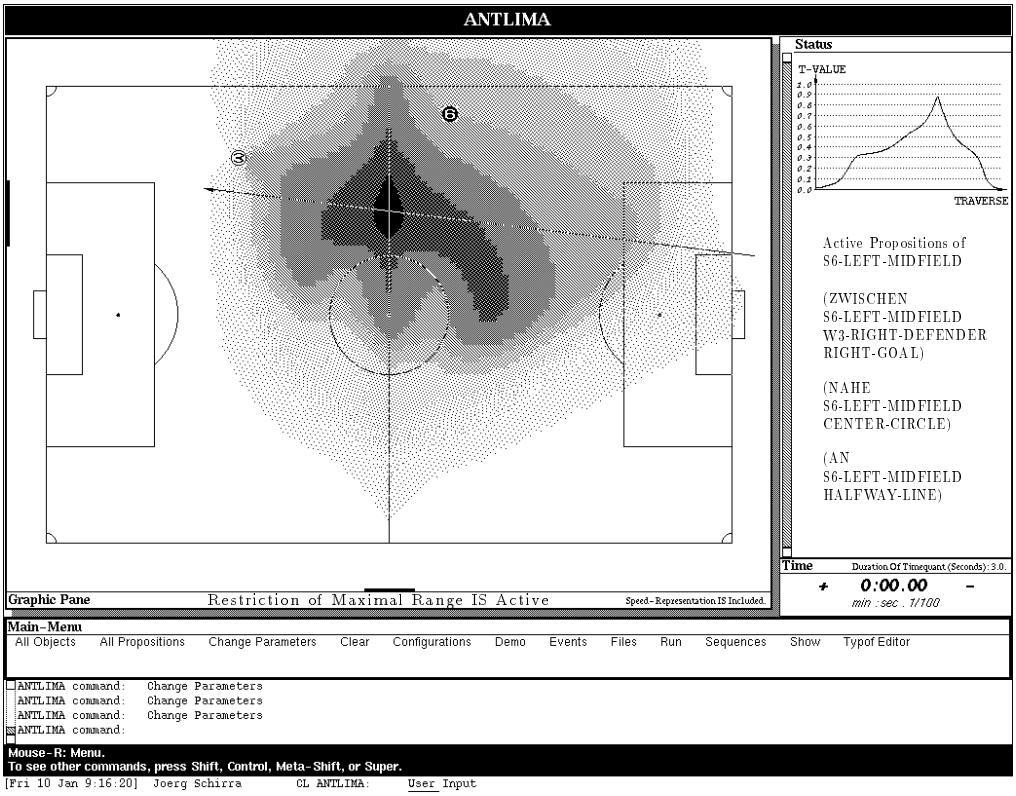

Figure 16: Combination of Three Compatible Restrictions with Sectional View on the Right Side

typicality functions (e.g., algebraic average) which are associated with each LO – meaning in this case: object *to be* located – describes the typicality distribution for the conjunction of the restrictions. We can illustrate these calculations as a combination of several probability clouds (cf. Fig.s 11 to 13): only where several dense centers overlap does the combined typicality reach really high values. Fig. 16 shows the combination of three localizations: [near CenterCircle], [between player W3 RightGoal], and [at MiddleLine].[37]

With such combinations, compatible restrictions result in a combined typicality function with a quite high typicality maximum, as shown in Fig. 17 for the combination of [in front of LeftGoal] and [near player-7], whereas incompatible propositions, e.g., [in front of RightGoal] combined with [in front of LeftGoal], yield maxima of typicality with an extremely low value, since the maxima of the components do not overlap (cf. Fig. 18 with a maximum at 0.5). Thus, the maximal degree of typicality – or T-value – reachable for the set of propositions can be used as a rating for the plausibility the considered utterance has for the listeners. If only a low T-value can be reached while generating the mental image, i.e., reconstructing the referent of that utterance, incompatibilities must have been included. If a high typicality value can be reached, all involved restrictions could be satisfied in the mental image and the listeners found (one of) the typical referent(s) of the utterance.

[37]In the right window, a sectional view cutting through the typicality cloud at its maximum is shown, called 'Traverse' and following the arrow in the graphic pane from the right to the left; the maximum for this localizing combination is reached at a T-value of approximately 0.9;

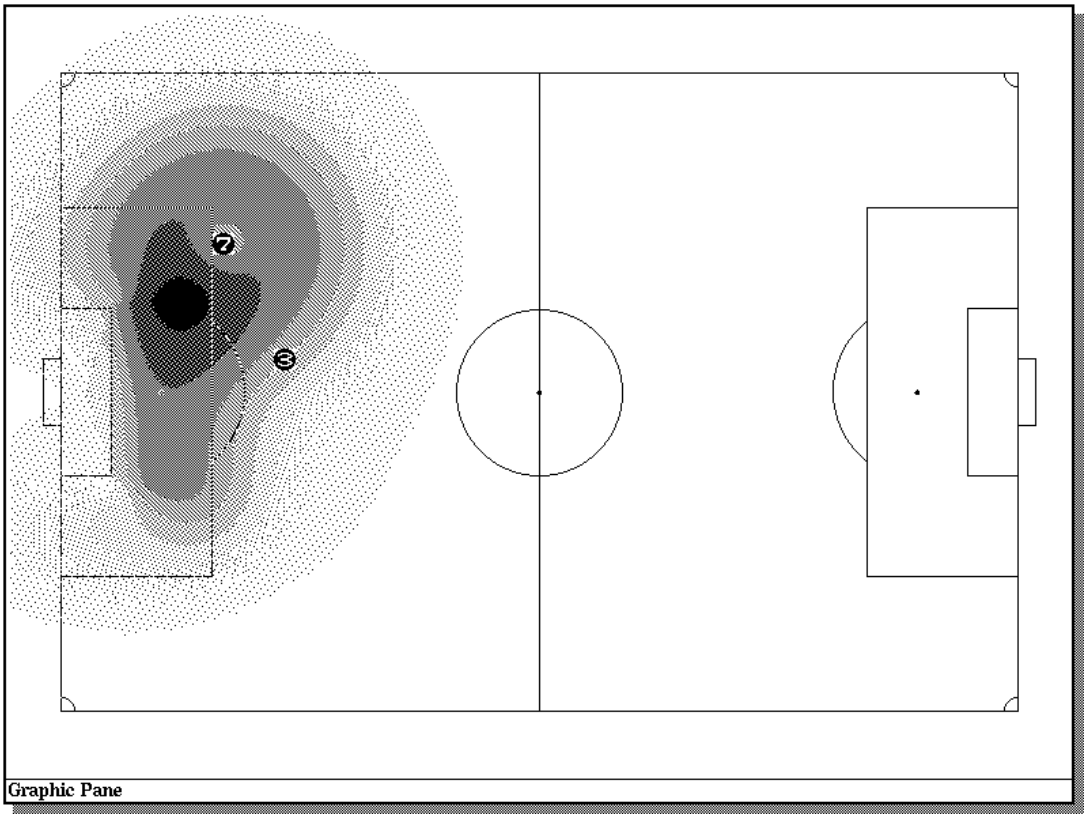

Figure 17: Combination of Applicability Clouds of Compatible Spatial Propositions

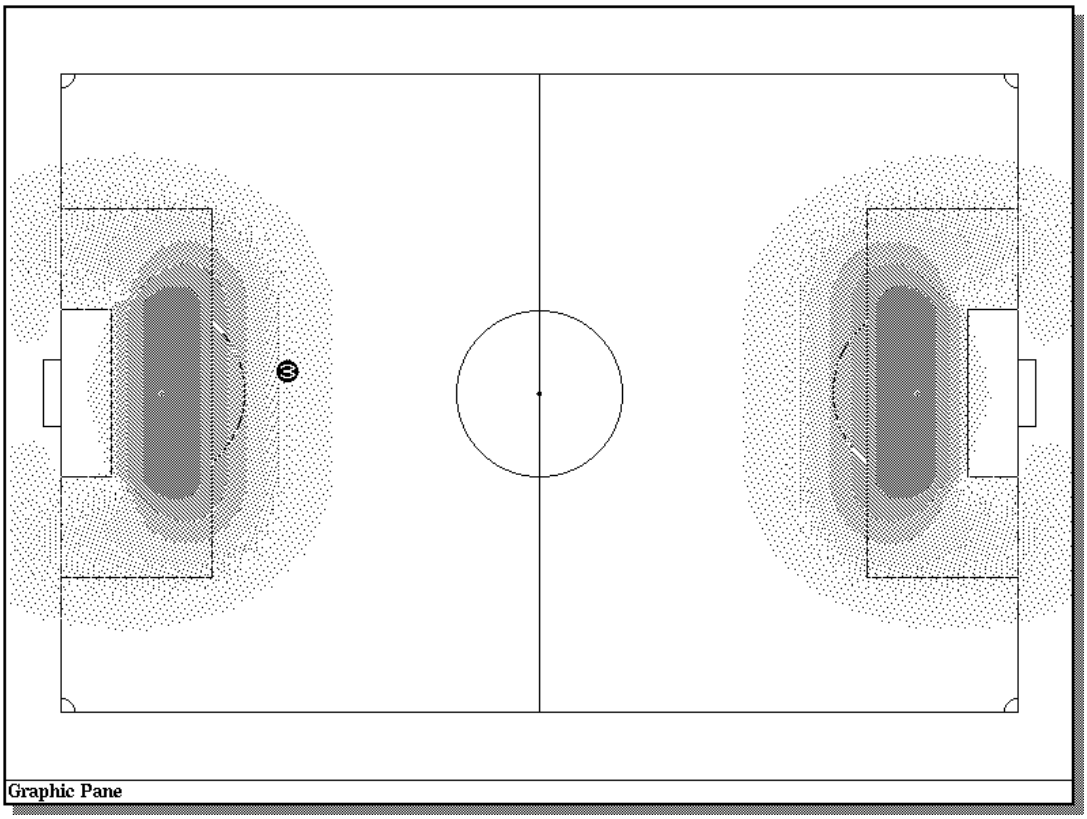

Figure 18: Combination of Applicability Clouds of Incompatible Spatial Propositions

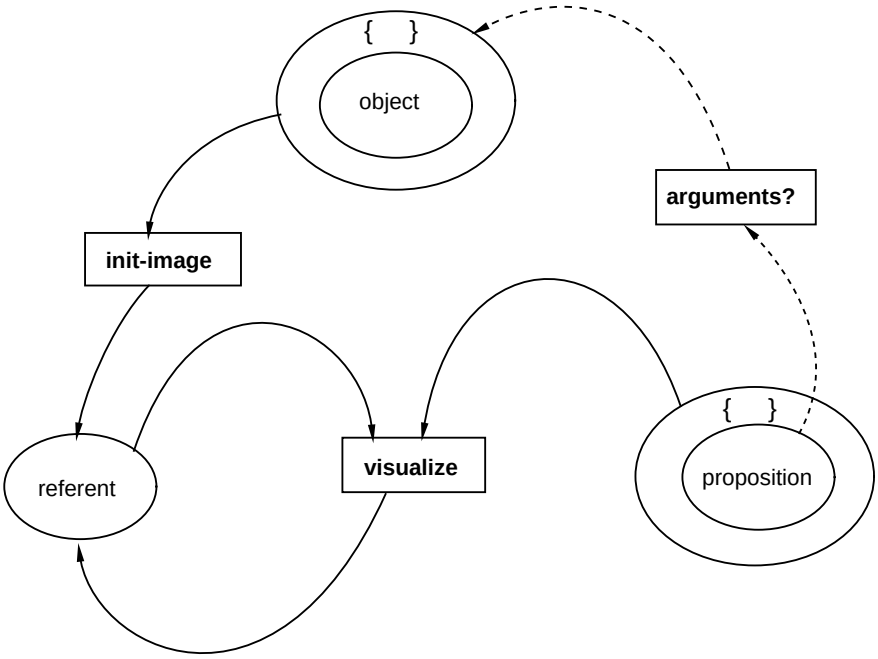

Figure 19: Signatures of the Spatial Concept System (II): Visualization

Using again the signatures of abstract data types, we can represent the visualization as an operation of the same set of ADTs already used in Section 4. Fig. 19[38] shows that the operation *visualize* takes an initial image and a set of propositions and yields another image. The initial image is generated by an operation *init-image* which takes the set of all considered objects and locates them at arbitrary positions.[39] The special attribute of the resulting image cannot be represented in the signature. Instead, I use the following logical formula (cf. Formula 2). Or less formally: if we sum up the T-values (typicality/applicability values) of all propositions p in the given set of propositions P with respect to the result of the *visualize* operation on P, this sum has to be at least as great as the sum with respect to any other possible image. In other words: the result is the (or one of the) most typical image(s) for the set of propositions P.

$$\forall P \in 2^{propositions}, \forall I \in referent: \tag{2}$$
$$\sum_{p \in P} (\text{recognize } (\text{visualize } (\text{init-image } \left(\bigcup_{p' \in P} (\text{arguments?}\,(p')) \right), P), p)) \geq \sum_{p \in P} (\text{recognize } (\text{I}, p))$$

[38]Dotted arrows mark those operations described earlier;

[39]This is not completely true, but it will suffice here; cf. Section 7;
In SOCCER, all mobile objects are zero-dimensional. Thus it is easy to define the position of such an object. Objects of higher dimensionality would be idealized to their 'centers of gravity' which represent their location. Their extension is considered by different mechanisms (cf. I-rules in Section 6).

6 The Operational Form of the Reference Relation in VITRA

In the preceding two sections, the reference relation was used in two directions, for recognition and for visualization. To that purpose, we introduced graded classification functions which encode substantial parts of the reference relation for SOCCER on an abstract level of description, hidden in the operations *visualize* and *recognize* (Fig.s 14 and 19). Can we bring them to a more concrete level which also expresses the relationship between these two operations?

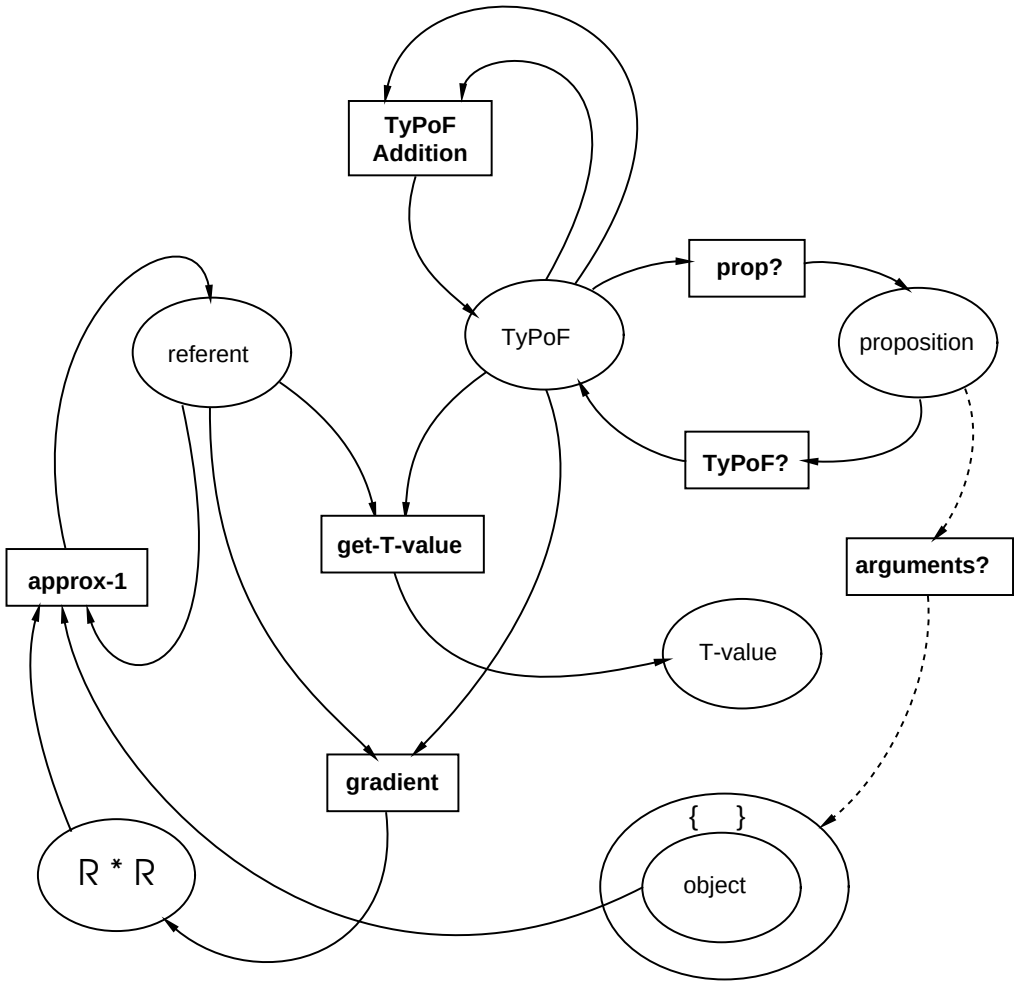

Figure 20: Signature of the Spatial Concept System (III): TyPoFs

To that purpose, a new data type is introduced which explicitly represents the graded classification functions. It is called TyPoF – TYpicality POtential Field (cf. Fig. 20). This data type stands for the third part of the reference relationship, the 'arch' between the two 'columns of the bridge', namely images and propositions. As mentioned in Section 1, for SOCCER, this part has to be essentially procedural; SOCCER does not know (in the strict sense) anything about this data type and only can apply the corresponding operations. But both the *recognize* and the *visualize* operations can now be expressed simply by using TyPoFs.

Each proposition is associated with exactly one TyPoF (cf. operations *prop?* and *TyPoF?* in Fig. 20). The *recognize* operation simply can be expressed now by asking for the value of the classification function (Formula 3).

$$\text{recognize (proposition, referent)} \equiv \text{get-T-value (referent, TyPoF? (proposition))} \tag{3}$$

As mentioned above, the *visualize* operation is to find the maxima of the classification function: therefore, an operation *gradient* is defined. It yields the component-wise differentiation of the classification function ($\vec{\nabla} T = (\frac{\partial T}{\partial x}, \frac{\partial T}{\partial y})$) for the position of the LO in the percept. This two-dimensional vector always points in the direction of the closest local maximum. Its length is proportional to the local slope of the classification function. At the maximum, we get the zero vector. If the initial position of the LO is favorable, we find the maximum of the classification function by moving the LO in the direction of the local gradient. One step of this hill climbing algorithm for one proposition is indicated in Fig. 20 by the operation *approx-1*: the considered object is shifted in the referent a little bit in the direction of the given vector.[40] If we iterate the combination of these two operations as shown in Formulas 4 to 6 until the gradient is zero, we find the maximum, or at least a local maximum depending on the initial position of the LO.

$$\forall i \geq 0 : \text{recognize (prop, image}_{i+1}) \geq \text{recognize (prop, image}_i) \tag{4}$$

$$\text{image}_0 \equiv \text{init-image (arguments? (prop))} \tag{5}$$

$$\text{image}_{[i+1]} \text{ (prop)} \equiv \text{approx-1 (gradient (TyPoF? (prop), image}_i \text{ (prop)) image}_i \text{ (prop), LO? (prop))} \tag{6}$$

To get faster and better results, we can use typical positions associated with the object type of LO. For example, the goal keeper typically will be located near or in the goal, which usually is already close to the most typical position in a specific case. Thus, the approximation of his position should start with a position in the goal. In ANTLIMA, we usually will use the positions of the objects at the time quantum before as the initial position. Then, velocity restrictions can give further hints as to where the objects will be. [41]

Since the proposed algorithm reacts sensitively on the starting position, the influence of the context conditions on the localization is rather naturally included. Fig. 21 demonstrates the approximation for [in front of LeftGoalArea] for several starting positions. Depending on the starting positions which play the role of conditions of context, different solutions of the visualization problem are constructed. Furthermore, an object to be located with respect to a second object to be located will *follow* that second LO until both have reached their optimal

[40] cf. [YND88]; the term typicality potential field is motivated by interpreting the classification function physically as a potential field whose associated force field (the vector field resulting the gradient operation) pulls the LO into its optimal position (with maximal potential energy); similar to a negative field of gravity, the TyPoFs pull the LOs always 'uphill to the summit of the typicality mountains'.
Note that the moving of the imaginative LO during approximation also can be used to direct (visually) searching for the perceived LO: if we want to find an object which position is described verbally, our visual focus of attention moves – as if controlled by the typicality distribution – toward the most typical positions associated with the verbal description;

[41] These hints for finding good starting positions have to be used in the aforementioned *init-image* operation; cf. [Sch94];

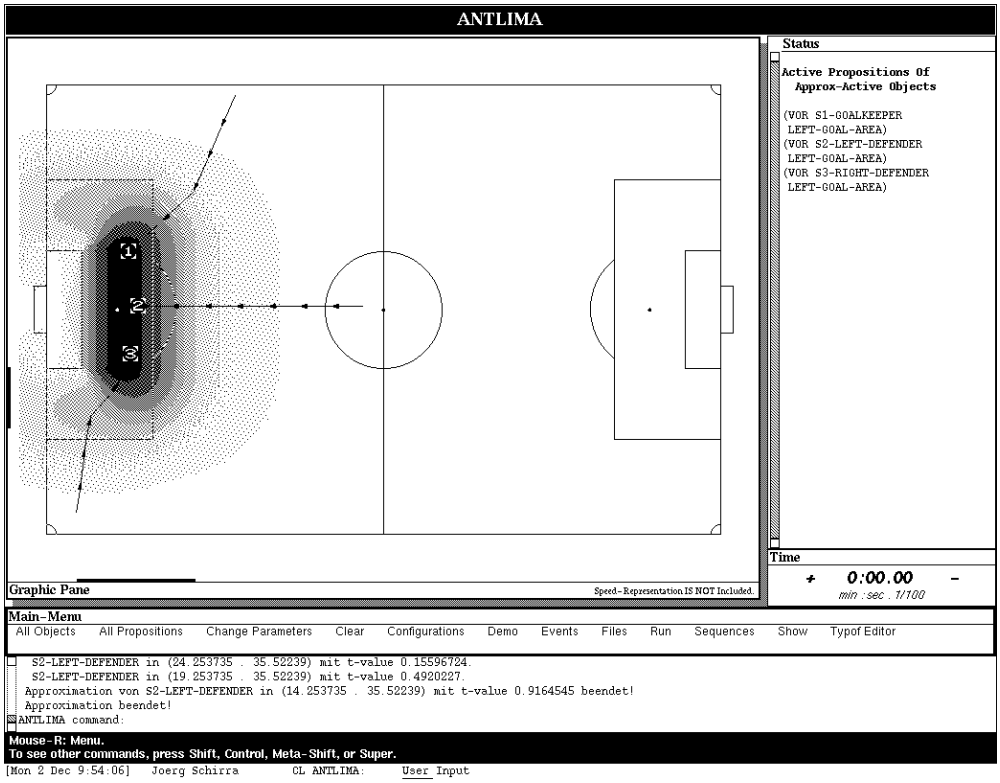

Figure 21: Demonstration of the Approximation for Different Starting Points

positions. Fig. 22 shows the approximation paths for two players localized by (left of player-5 player-8 (extrinsic from the lower left corner)) and (in front of player-8 LeftGoal). After about 5 steps, player-5 already reached a position left to player-8, and has to follow him until he also has reached his final position.

The operation *TyPoF-Addition* combines several TyPoFs for one LO by arithmetic average in order to simplify the approximation by reducing the number of TyPoFs to be considered simultaneously. The basis for this operation was described in Section 5.

Remember that TyPoFs are associated with individual propositions, like (left player-3 player-5). Since there are quite a lot of these individual propositions even in such a simple system as SOCCER, the task of determining every single TyPoF would be too complicated. In fact, the reference relation depends on all parts of the proposition: that is, both the type of the relation, i.e., the Spatial Concept, and the arguments influence the shape of the typicality function. Can we *extract* from their instances unique descriptors of the influence of the Spatial Concepts, and thus, separate both kinds of influences in order to get a formalism which is easier to deal with? Although the typicality distribution for different propositions of one type are more or less similar, the kinds of objects and especially their dimensionality, size, and shape modify the typicality distribution of a single proposition, stretching it, adapting it to the shape of the RO (cf. Fig.s 11 to 13). Therefore, the assumed influence of the arguments, i.e., the objects involved, has to transform the proposed influence of the Spatial Concept underlying all of its instances.

To cover the similarity between all instances of a Spatial Concept, we assume one function for each Spatial Concept which describes the typicality distribution not with respect to the

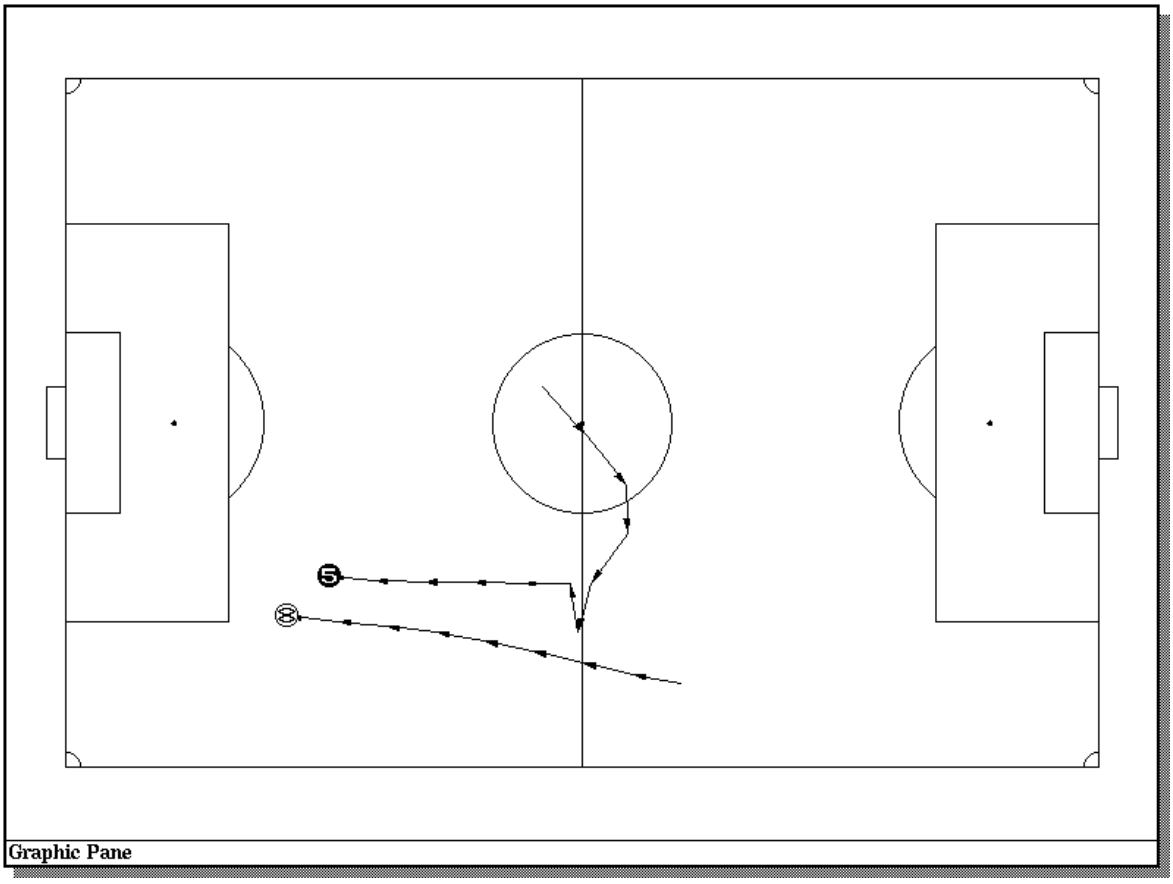

Figure 22: Demonstration of Approximation for Depending Objects

coordinates of the objects, but with respect to so-called *essential parameters*. The typicality distribution of every instance of a spatial relation is derived from this function. It differs from the distributions of other instances only with respect to how the essential parameters are calculated from the objects' coordinates. The ADT for these functions is called *Typicality Schema* (cf. Fig. 23). The essential parameters, e.g., distance, angles, and scaling factors, are abstractions from the concrete coordinates. Fig. 24 visualizes three Typicality Schemata — i.e., functions from essential parameters to T-values. For example, the Typicality Schemata for proximity and contiguity (or the *near* and *at* concepts) both depend on two essential parameters: the distance between LO and RO and a scaling factor.[42] How both parameters are connected to the actual coordinates of LO and RO depends on the dimensionality and size of the objects.

Typicality schemata can be combined by multiplication. In ANTLIMA, the Typicality Schemata of the projective relations *left, right, behind,* and *in front* are defined as combinations of the Direction and the Proximity Schemata, and thus, have five essential parameters: distance r, distance scaling G, the reference system δ, the angle ϕ with respect to δ, and the scaling of the angle G'. In summary, VITRAs Spatial Concepts are essentially defined as combinations of simple graded functions (multiplication of basic Typicality schemata) of essential parameters.[43]

[42]Thus, the Typicality Schemata are actually two-dimensional; the scaling dimension was skipped in the sketch as an autonomous axis for the sake of simplicity;

[43]Obviously, the combination of TyPoFs (which is actually a renormalized addition of functions) and the combination of Typicality Schemata (which is a plain multiplication of functions) serve different purposes: the first combination integrates the simultaneous influences of several restrictions for one LO, whereas the second combination allows for defining compound Spatial Concepts;

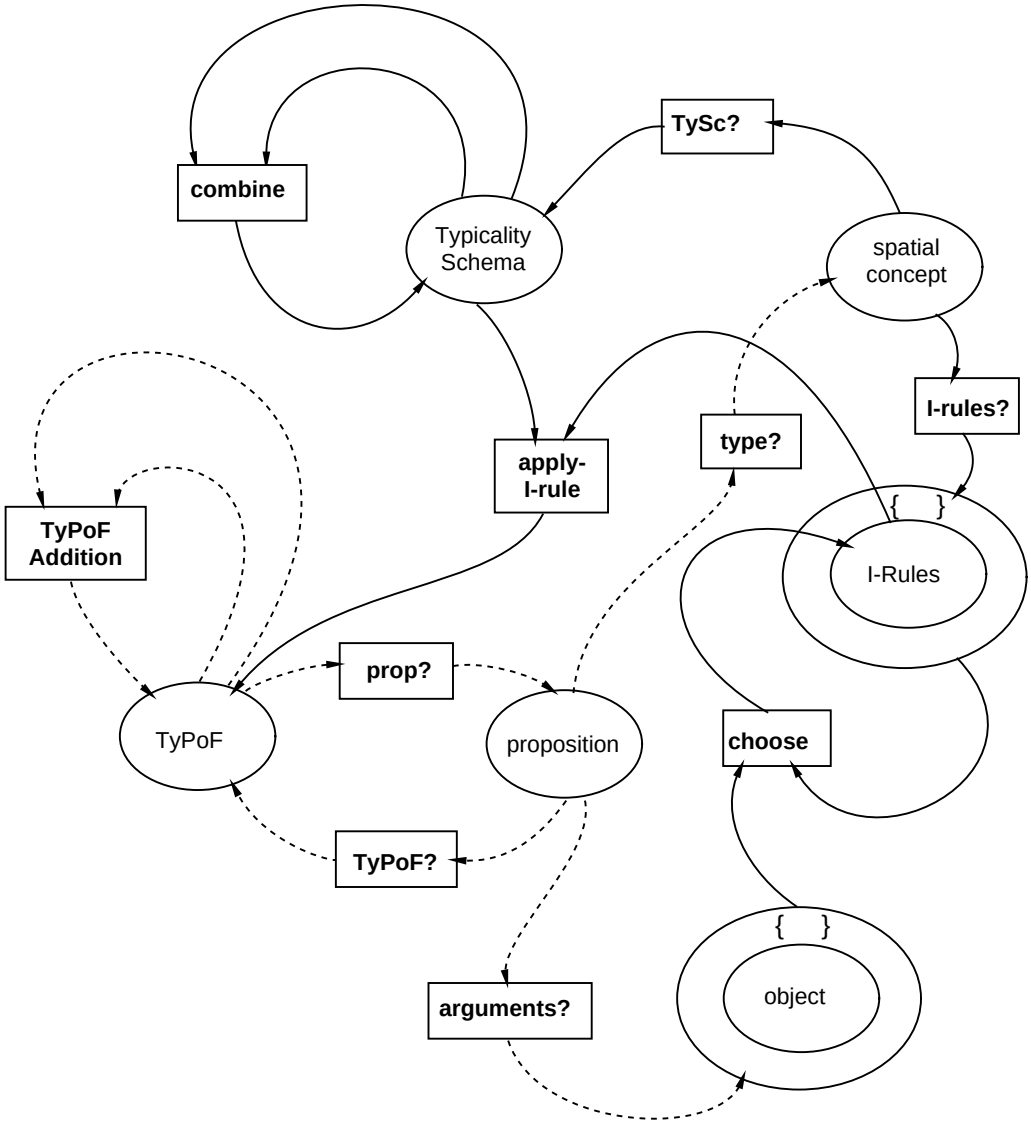

Figure 23: Signature of the Spatial Concept System (IV): Typicality Schemata

The influence of the objects is encoded by *TyPoF Instantiation Rules* (for short: *I-rules*). In addition to its Typicality Schema, each spatial concept is associated with a set of I-rules. We can say that, in a way, I-rules *spread* the typicality distributions encoded in the Typicality Schema around the ROs in the percept, thus developing the appropriate Typicality Potential Field, which, then, directs the LO towards its optimal position. Each I-rule specifies, according to the object attributes, a set of functions for calculating the essential parameters. These so-called *parameter functions* transform the coordinates of the objects, i.e., the information in the percept, to the essential parameters needed by the Typicality Schemata. Since, for example, *distance*, the essential parameter of the Proximity Schema, is defined only between zero-dimensional objects, we have to reduce – or idealize – higher-dimensional objects to points

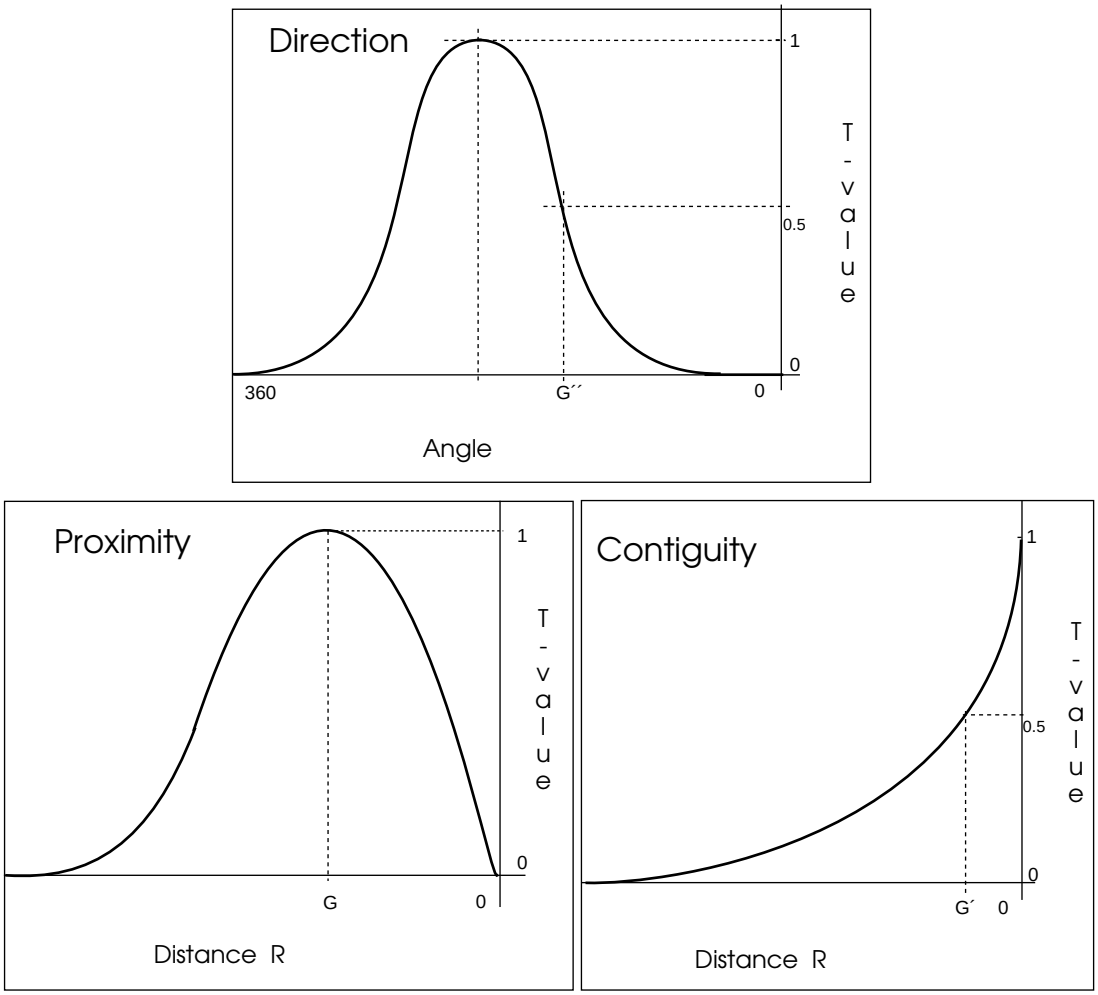

Figure 24: Three Examples for Typicality Schemata

in order to apply the Proximity Schema. In other words, to find the distance between two two-dimensional objects, we first have to look for those two points of the borders of both objects which are closest to each other (cf. Fig. 25, and also Fig.s 11 to 13). These points are *viewed as* those objects, their distance *is* the distance between the objects. Actually, this object coercion is the important part of the parameter function, since the calculation of distance *per se* remains unchanged in every I-rule: the Euclidean distance between two points.

Similarly, *angle* in the Direction Schema is only defined for two lines – the zero line of the reference system and the line between RO and LO. The geometric situations in the percept have to be adapted to this restriction: e.g., the zero direction of the reference system can be derived either from inherent object properties of the RO (intrinsic use) or from another object in the context (extrinsic use) which defines exactly one line to the RO.[44] Again, the coercion of the original percept to an idealization plays the major part of the parameter functions used for the Direction Schema. In every specific case, the angle is calculated in the same way from the idealization.

[44]A special case of extrinsic use is the deictic use: the object that determines the orientation is the speaker/listener and is normally not explicitly mentioned; cf. [RS88];

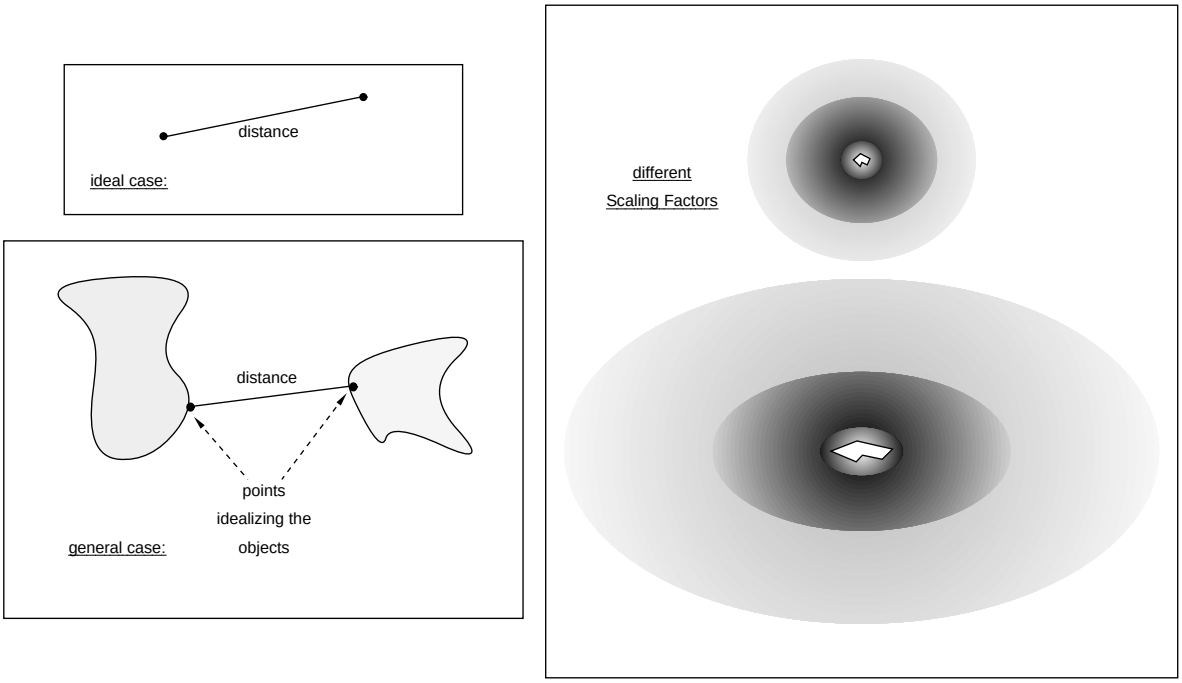

Figure 25: Illustrations for Distance Functions and Scaling Factors

For both cases, I-rules have to be defined. The size of the objects, which was ignored by the essential parameters *distance* and *angle*, has an influence by means of the scaling factors. For [near player-5] and [near PenaltyArea], the Typicality Schema of Proximity not only has been spread differently around the ROs with respect to their shape, but also 'stretched' to different diameters corresponding to the size of the ROs (cf. Fig. 25, and also Fig.s 11 to 13).

In summary, hidden in the I-rules of a spatial concept in VITRA are functions which coerce parts of the percept to a more abstract, sketch-like form which is the basis for calculating the essential parameters for the Typicality Schemata. This sketch is not explicitly modeled in VITRA yet, though, and further studies have to be carried out about their relation to reference semantics of spatial concepts (cf. Section 1).

The whole algorithm for visualizing elementary static spatial relations can be sketched as in Fig. 26. Originally starting from the temporally ordered sets of elementary spatio-temporal propositions, the so-called propositional elementary structure, we consider here only one time quantum and the corresponding set of static spatial propositions. For each proposition in this set, we find the type and the associated Typicality Schema (TySc? (type? (p))); additionally, we get the set of I-rules of the Spatial Concept (I-rules (type? (p))). The arguments of the proposition choose one of the I-rules, which, then, transforms the Typicality Schema to the appropriate TyPoF by specifying the appropriate parameter functions:

$$\begin{aligned} \text{TyPoF? (p)} \;\equiv\; & \text{apply-I-rule (choose (I-rules? (type? (p)),} \\ & \qquad\qquad \text{Arguments? (p)),} \\ & \qquad \text{TySc? (type? (p)))} \end{aligned} \tag{7}$$

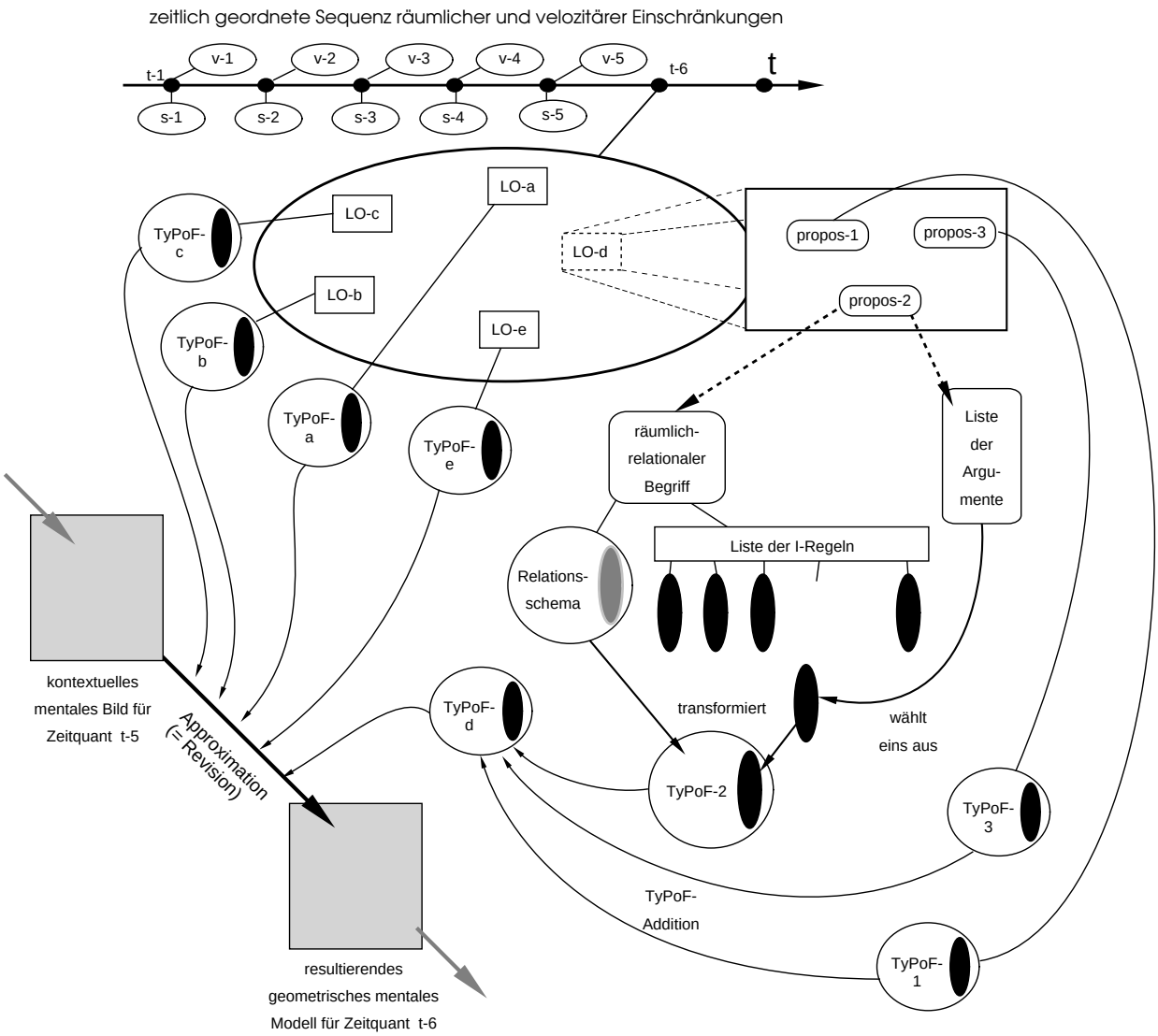

Figure 26: Algorithm of ANTLIMA's Visualization of Static Spatial Propositions

By TyPoF-addition, all TyPoFs belonging to one LO in the set are combined. Thus, we finally consider a set of LO's, each associated with one TyPoF. With the *init-image* operation, we, then, construct a first mental image with all the objects used as arguments in the original set of propositions.[45] The aforementioned hill climbing approximation (cf. Formula 6) transforms this initial image to the most typical image, the referent we looked for. Then, the sum of all T-values finally reached is used as a first approximation of the plausibility of the description for the listeners.

[45] If a whole image sequence is to be constructed, this step is only needed at the beginning or when a new object enters the scene; otherwise, the image for one time quantum is used as the initial image for the approximation of the next time quantum;

7 A Comparison with Herskovits' Analysis of the Semantics of Spatial Prepositions

On first view, the separation of the influences on the typicality distributions described in the previous section seems to reflect the distinction between several aspects A. Herskovits introduced in her analysis of the semantics of spatial prepositions in English. Is this similarity only superficial, or can we also interpret the operational form of the reference relationship in SOCCER as a simplified realization of Herskovits' analysis?[46]

In order to avoid polysemy for spatial prepositions, Herskovits suggests a so-called *ideal meaning* of a spatial preposition underlying all uses of that preposition. In her own words:

> The ideal meaning of a preposition is a geometrical idea, from which all uses of that preposition derive by means of various adaptations and shifts. An ideal meaning is generally a relation between two or three ideal geometric objects (e.g., points, lines, surfaces, volumes, vectors) – in fact, ideal meanings are usually those simple relations that most linguists and workers in artificial intelligence have proposed as meanings of the prepositions. These relations play indeed an important role, but as something akin to prototypes, not as truth-conditional meanings.
>
> [Her86, p. 39]

Coincidence of points, inclusion of a point in a line or in4 an area, contiguity of two surfaces – these are some examples of relations used as ideal meanings. A substantial point is made in the last sentence of the quote: although the ideal meaning underlies every occurrence of a spatial preposition, it may be 'deformed', 'weakened', partially 'overwritten' or 'extended' and serves merely as a 'gravitational center' for the meanings of different occurrences, not as an accurate and complete description. Actually, for Herskovits, the meaning of single occurrences of a spatial preposition is derived from the ideal meaning by three steps (cf. Formula 8):

> In a particular use of a preposition, the ideal meaning may have been *transferred* to another relation, one that is in some way closely related; [step 1]
> this new relation may in turn be only approximately true. [step 2]
> Moreover, the objects related are mapped onto geometric objects (matching the categories specified for the arguments of the ideal meanings) by processes of geometric imagination, idealization and selection. [step 3]
> These mappings onto geometric descriptions, corresponding to various geometric conceptualizations and metonymies, are accomplished by a variety of functions.
>
> [Her86, p. 40, with comments by JS]

[46] The following comparison is still on a very coarse level and will be elaborated further in the future.

And later, defining technical terms for the transformations:

> The geometric meaning of a locative expression is thus a proposition involving a relation applying to geometric descriptions [3] of the objects, and that relation may be the result of two transformations applying in succession to the ideal meaning, which I call sense shifts [1] and tolerance shifts [2].
>
> [Her86, p. 40, numbers added by JS]

$$\left[\mathrm{T}\ \left[\mathrm{S}\ [\mathrm{IM}]\right]\right]\ \left(\mathrm{G}_1\ (\mathrm{O}_1),\ \mathrm{G}_2\ (\mathrm{O}_2)\right) \tag{8}$$

with

T	—	Tolerance Shift	—	(step 2)
S	—	Sense Shift	—	(step 1)
G_i	—	Geometric Idealization of Object O_i	—	(step 3)
IM	—	Ideal Meaning		

Although the spatial prepositions in SOCCER's language are quite restricted compared to our everyday language use, the reference-semantical mechanism controlling this reduced language is already rather complicated, as we saw. It is my thesis that this mechanism can be matched with parts of Herskovits' analysis.

Already on first view, one notices a relationship between Spatial Concepts in SOCCER and ideal meanings. Spatial concepts represent the meaning of spatial prepositions *per se*, so to speak; they function as reference points with respect to which propositions about the perceived spatial environment are formulated. In general, they are defined by combinations of Typicality Schemata. Similarly, an ideal meaning stands for the meaning of a spatial prepositions itself, and is defined as a conjunction of elementary spatial relations which are "perceptually salient relations", i.e., "easily, quickly and accurately perceivable" ([Her86, p. 54]).[47] Of course, it is not just by chance that my examples for Typicality Schemata – Proximity, Contiguity – all can be found in the list of examples for elementary relations used as ideal meanings. Indeed, also the procedural aspect of ideal meanings is stressed:

> We can conceive of this core schema [i.e., the ideal meaning of 'in' (JS)] as a logical predicate, but its psychological realization is most certainly a routine [Ull85], a procedure which checks whether spatial inclusion holds.
>
> [Her89, p. 13]

[47] On p. 55, one finds a list of relations Herskovits takes into consideration: "enclosure, contiguity with line or surface, order of three points on a line, order of two points on an oriented line, coincidence of two points, line in/on a plane, alignment of points, parallelism of lines, alignment with direction, orthogonality of lines, support, on line of sight, on orthogonal to line of sight";

Though I could not find an explicit record, it seems probable to me that Herskovits' elementary relations correspond to mathematical concepts which encode clear, binary classifications: a set of arguments either stands in a given relation, or it does not.[48] The second step of adaptation mentioned above, the tolerance shift, supports this assumption. We have already considered the fuzziness of the applicability of spatial prepositions in Section 4. To cover this phenomenon of ordinary language with the mathematically accurate definitions of ideal meanings, the sharp borders of applicability of the mathematical relations have to be softened. Then, coincidence of points can be true in a weaker sense – or in other words: the relation is less applicable – for two points which do not actually coincide but only are close together.

In contrast to that, SOCCER already considers the fuzziness on the level of Spatial Concepts: Typicality Schemata describe/define weak, fuzzy forms of some of the mathematical relations mentioned above. This may reflect a more empiricist view, assuming that the spatial concepts somehow evolve from perception, which is assumed to be always vague or fuzzy. There is no obvious reason why spatial concepts should loose this quality. Herskovits, on the other hand, seems to adopt a more rationalist point of view, starting by a clear, simple, and especially binary classification which is somehow innate and which afterwards has to be blurred in order to fit the phenomena.[49] As a consequence, there is no need for a distinction of the second type of transformation Herskovits mentioned, the tolerance shifts, from other kinds of transformations in SOCCER. Every transformation of Typicality Schemata necessarily changes the 'tolerance' for the applicability of that preposition somehow, e.g., spreading and stretching the Typicality Schema around the RO.

In principle, something equivalent to sense shifts, the first step in adapting ideal meanings to a particular instance, can occur on combinations of Typicality Schemata, as well. Skipping one of the conjuncts or adding one more are the most prominent examples for sense shifts (cf. [Her86, p. 94]) and both can be performed similarly on Typicality Schemata. Although, the domain of SOCCER is still too simple (or better: SOCCER perceives it in a too impoverished manner) for an example of such a shift to be found yet.

Most interesting for this comparison is the remaining step in Herskovits' derivation scheme (no. 3), called geometric description. As mentioned above, the objects used as arguments of the prepositions have to be transformed to the simple geometric objects used as arguments of ideal meanings. Since the relations defining the ideal meanings or their derivates by sense and tolerance shifts only apply to simple geometric objects, coercions have to be found which also can be viewed as different conceptualizations of the objects: when an object is described as 'at' some RO, both objects are conceptualized as points. To this purpose, a broad set of *geometric description functions* was collected by Herskovits.[50] Usually, several coercions apply in sequence until the appropriate conceptualization is found. Pragmatic factors like salience, relevance, and tolerance influence the decision as to which sequence of coercions is applied.[51]

[48] Remember that these relations apply to geometric concepts derived from the argument objects; Herskovits' discussion of 'at' ([Her86, p. 51]) and 'to the right' (p. 184) as 'graded concepts' points toward an understanding similar to VITRA's; although it remains unclear whether the gradation is part of the ideal meaning or not. At least, the ideal meaning of 'at' (p. 128) is simply defined as: "for a point to coincide with another". Similarly, the ideal meanings finally proposed for the projective prepositions (p. 190) do not include gradation: the LO has to be located (exactly) on the corresponding axis of the reference system;

[49] This presentation of Herskovits' handling of tolerance phenomena is rather simplified; cf. especially [Her86, Chapter 6.3, 'Tolerance and Idealization']; the 'blurring' is assumed to be either idiosyncratic or controlled by pragmatic factors;

[50] cf. [Her86, Chapter 5];

[51] cf. also [Hay87];

Obviously, something similar is hidden in the I-rules in VITRA; as discussed in the previous section, I-rules specify parameter functions which calculate the essential parameters from the percept – the coordinates of the objects. To that purpose, the objects first have to be transformed to idealized forms appropriate to calculate the essential parameters – points, lines etc.: a sketch-like representation of parts of the referent is generated implicitly, so to speak. However, there is one main difference between Herskovits and VITRA: the coercion in VITRA always depends on all objects *simultaneously*, whereas in Herskovits' scheme the geometric description functions apply to each object separately. I assume that this difference is a consequence of the selection principles. Whereas Herskovits considers in her much broader framework pragmatic principles which coordinate the idealizations of all involved objects, in VITRA until now only idiosyncratic rules are defined: the condition part of I-rules refers to attributes of the involved objects like dimensionality, size, type, etc., and specifies explicitly for each relevant class of combinations one set of parameter functions with the corresponding coercions. General rules have not been considered yet (cf. Formulas 9 and 10).

$$\mathrm{TS}\left[\mathrm{PF}_1\ (\mathrm{LO},\ \mathrm{RO}),\ \mathrm{PF}_2\ (\mathrm{LO},\ \mathrm{RO}),\ \cdots \mathrm{PF}_n\ (\mathrm{LO},\ \mathrm{RO})\right] \tag{9}$$

$$\mathrm{PF}_i\ (\mathrm{LO},\ \mathrm{RO}) = \mathrm{F}\ \left(\mathrm{Idealize}_{i_1}\ (\mathrm{LO}),\ \mathrm{Idealize}_{i_2}\ (\mathrm{RO})\right) \tag{10}$$

with

TS	—	Typicality Schema
PF_i	—	Parameter Function of the ith Essential Parameter
F	—	the *pure* parameter function (e.g., distance of points)

To conclude, the solution of the visualization problem in VITRA has led us to a conception of the reference semantics of spatial prepositions which is comparable to a simplified form of A. Herskovits' theoretical framework (cf. Formulas 8, 9 and 10). As in her approach, we assume a core meaning of each spatial preposition – the Spatial Concept with its defining Typicality Schema. Individual occurrences are derived by spreading the schema appropriately over the mental image by means of I-rules. This spreading can also be viewed as idealizing the objects in the mental image to simple geometric objects – which corresponds closely to Herskovits' derivation step 3. Because VITRA alludes to a totally operational form of reference semantics for spatial prepositions, the comparison offers a computational realization of a substantial part of Herskovits' analysis. Furthermore, using this realization to derive additional hints for future elaborations of the framework seems a plausible path to follow.

8 Summary

The problem of integrating vision and natural language systems has led us in the project VITRA to investigate the nature of the reference relationship (cf. Fig. 27). This resulted in a rejection of the objectivist view due to its rather obvious weak points: essentially, objectivist reference theories cannot explain adequately how the objects in the world, which are viewed as the referents, conduct their influence on verbal behavior. Instead, we view mental entities like percepts and mental images as referents, thus adopting in VITRA an experientialist theory of the reference relationship.[52]

[52] Further investigations considering some problems inherent to the mentalist point of view of the reference relation are to be found in [Sch93]; cf. also [Tug76, Section 20];

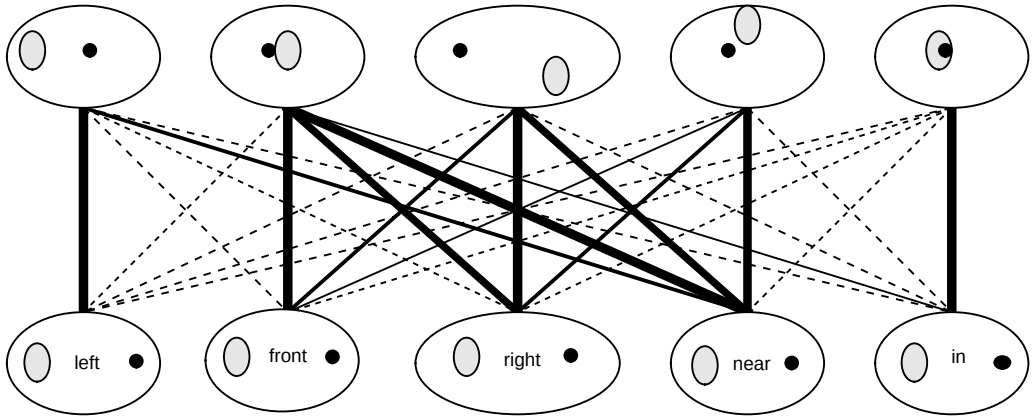

Figure 27: The Reference Relationship — Final View

In the system SOCCER, this decision resulted in an operational realization of the reference relation proper: algorithms are being developed that bind exactly in the spirit of experiential reference semantics the usages of spatial prepositions to SOCCER's percepts, the results of an image understanding system. The core of these algorithms is always a classification function which associates the percepts to a graded Spatial Concept. A degree of applicability allows us to distinguish between good and less good occurrences of such a graded concept and is reflected on the verbal surface by linguistic hedges.

Combining the demands of experiential reference semantics with the Gricean Cooperative Principle furthermore compelled us to reconstruct in the form of mental images the referents of the text SOCCER plans to generate in order to anticipate the understanding of the listeners. SOCCER, thus, is able to improve its texts. This leads directly to the visualization problem: since we usually lose information when we transform percepts to propositions, the inverse transformation from propositions to mental images has to produce additional information somehow. In fact, both transformations have to use the reference relation: reinterpreting it, i.e., the classification functions, as typicality functions and correspondingly the degrees of applicability as degrees of typicality allows us to solve this task by generating the most typical mental images as reconstructed referents.

The operational realization of the reference relation in VITRA rests on the classification/typicality functions: Typicality Potential Fields encode these functions, and are used both for recognizing spatial relations from percepts and for reconstructing mental images from propositional descriptions. Furthermore, we distinguish between two components which define the TyPoF for a specific case: Typicality Schemata bring in the general influence of the spatial concept and are transformed by I-Rules that depend on the objects involved. Although technical reasons originally forced this distinction, the separation of the two components seems to reflect a deeper, cognitively important structure of the reference relation; a comparison with the analysis of A. Herskovits revealed a significant degree of similarity between ideal meanings and Typicality Schemata on the one hand, and geometric descriptions of objects and parts of the transformations in the I-rules on the other. Sketch-like idealizations which are hidden in the I-rules in VITRA and rather overtly included in Herskovits' framework point to structures integral to the reference relationship which are connected to visual abstraction and metaphoric extension; these will have to be examined further.

In summary, with its system SOCCER, VITRA demonstrates – albeit still on a very coarse and primitive level – the benefits of investigating reference phenomena in an operational form.

I would like to thank heartily Ellen Hays who not only had a critical look at my English but 'obstetrically' helped me to clarify my ideas during hours of discussion. Also my colleagues in Saarbrücken are to be thanked for comments and critiques on earlier versions. Furthermore, all participants of the workshop on the Semantics of Prepositions in Saarbrücken 1990 contributed with their discussion to the present form of the paper.

References

[André 88] Elisabeth André. Generierung natürlichsprachlicher Äußerungen zur simultanen Beschreibung von zeitveränderlichen Szenen: Das System SOCCER. Memo 28, Fachbereich Informatik, Universität des Saarlandes, Saarbrücken, August 1988. Diplomarbeit.

[ABHR85] Elisabeth André, G. Bosch, G. Herzog, and T. Rist. CITYTOUR - Ein natürlichsprachliches Anfragesystem zur Evaluierung räumlicher Präpositionen. Abschlußbericht der Fortgeschrittenenpraktikums, Fachbereich Informatik, Universität des Saarlandes, Saarbrücken, 1985.

[ABHR86a] Elisabeth André, G. Bosch, G. Herzog, and T. Rist. Characterizing trajectories of moving objects using natural language path descriptions. In *Proceedings of the 7th European Conference on Artificial Intelligence, Brighton, England*, pages 1–8, 1986.

[ABHR86b] Elisabeth André, G. Bosch, G. Herzog, and T. Rist. Coping with the intrinsic and deictic use of spatial prepositions. In P. Jorrand and V. Sgurey, editors, *AIMSA-86*, pages 375–382, Amsterdam, 1986. North-Holland.

[ARH87] Elisabeth André, T. Rist, and G. Herzog. Generierung natürlichsprachlicher Äußerungen zur simultanen Beschreibung von zeitveränderlichen Szenen. In *GWAI-87*, pages 330–338, Berlin, 1987. Springer.

[AHR89] Elisabeth André, G. Herzog, and T. Rist. Natural language access to visual data: Dealing with space and movement. Bericht 63, SFB 314, VITRA, Universität des Saarlandes, Saarbrücken, 1989.

[Bat90] J. Bateman. Finding translation equivalents: An application of grammatical metaphor. In *Proceedings of the 13th International Conference on Computational Linguistics, Helsinki*, volume 2, pages 13–18, 1990.

[Bie87] Manfred Bierwisch. On the grammar of local prepositions. In M. Bierwisch, W. Mosch, and I. Zimmermann, editors, *Syntax, Semantik und Lexikon*, pages 1–65. Akademie-Verlag, Berlin, 1987.

[Car33] Rudolf Carnap. Über Protokollsätze. *Erkenntnis*, 3:215–228, 1933.

[CL72] F.J.M. Craik and R.S. Lockheart. Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, 11?:671–684, 1972.

[Dan69] H. Dankert. *Sportsprache und Kommunikation – Untersuchungen zur Struktur der Fußballsprache und zum Stil der Sportberichterstattung.* Tübinger Vereinigung für Volkskunde e.V., Tübingen, 1969.

[DWP81] D. Dowty, R. Wall, and S. Peters. *Introduction to Montague Semantics.* D. Reidel, Dordrecht, 1981.

[Eco72] Umberto Eco. *Einführung in die Semiotik.* Wilhelm Fink Verlag, München, 1972.

[Emp85] Sextus Empiricus. *Grundriß der phyrronischen Skepsis.* Suhrkamp, Frankfurt/M., 1985.

[Fil71] Charles Fillmore. Toward a theory of deixis. Technical report, Univ. of Hawaii, 1971. Paper presented at the Pacific Conference on Contrastive Linguistics and Language Universals.

[Gra90] Michael Grabski. Transfer statements as conditional constraints. EUROTRA-D Working Paper 18, iai, Saarbrücken, 1990.

[Gri74] Herbert Paul Grice. Logic and Conversation. In P. Cole and J.L. Morgan, editors, *Syntax and Semantics*, volume 3, pages 41–58. Academic Press, New York, NY, 1974.

[Hay87] Ellen M. Hays. A computational treatment of locative relations in natural language. Technical Report MS-CIS-87-31, Department of Computer and Information Science, Univ. of Pennsylvania, Philadelphia, PA, 1987.

[Hay89] Ellen M. Hays. Two views of motion: On representing move events in a language-vision system. In D. Metzing, editor, *GWAI-89*, pages 312–317, Berlin, 1989. Springer.

[Hay90] Ellen M. Hays. On defining motion verbs and spatial prepositions. In C. Freksa and C. Habel, editors, *Repräsentation und Verarbeitung räumlichen Wissens*, pages 192–206. Springer, Berlin, 1990.

[Her85] Annette Herskovits. Semantics and pragmatics of locative expressions. *Cognitive Science*, 9:341–378, 1985.

[Her86] Annette Herskovits. *Language and cognition – An interdisciplinary study of the prepositions in english.* Cambridge University Press, Cambridge, UK, 1986.

[Her89] Annette Herskovits. The Linguistic Expression of Spatial Knowledge. L.A.U.D. Report A 248, Linguistic Agency University of Duisburg, Duisburg, 1989.

[HHR89] Christopher Habel, M. Herweg, and K. Rehkämper, editors. *Raumkonzepte in Verstehensprozessen – Interdisziplinäre Beiträge zu Sprache und Raum.* Niemeyer, Tübingen, 1989.

[HP88] Christopher Habel and S. Pribbenow. Gebietskonstituierende Prozesse. LILOG-Report 18, LILOG, IBM Deutschland, Stuttgart, 1988.

[HR88] Gerd Herzog and T. Rist. Simultane Interpretation und natürlichsprachliche Beschreibung zeitveränderlicher Szenen: Das System SOCCER. Memo 25, Fachbereich Informatik, Universität des Saarlandes, Saarbrücken, 1988. Diplomarbeit.

[HRA90] Gerd Herzog, T. Rist, and E. André. Sprache und Raum: natürlichsprachlicher Zugang zu visuellen Daten. In C. Freksa and C. Habel, editors, *Repräsentation und Verarbeitung räumlichen Wissens*, pages 207–220. Springer, Berlin, 1990.

[HSE+89] Gerd 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*, pages 153–162, Berlin, 1989. Springer.

[Joh87] Mark Johnson. *The body in the mind – The bodily basis of meaning, imagination, and reason.* Chicago University Press, Chicago, IL, 1987.

[JW82] A. Jameson and W. Wahlster. User modelling in anaphora generation: Ellipses and definite descriptions. In *Proceedings of the 5th European Conference on Artificial Intelligence, Orsay, France*, pages 222–227, 1982.

[Lak72] George Lakoff. Hedges: A study in meaning criteria and the logic of fuzzy concepts. In P.M. Perantean, J.N. Levi, and G.C. Phares, editors, *Papers from the 8th Regional Meeting*, pages 183–228. Chicago Linguistics Society, Univ. of Chicago, Chicago, IL, 1972.

[Lak87] George Lakoff. *Women, fire, and dangerous things – What categories reveal about the mind.* Chicago University Press, Chicago, IL, 1987.

[LB] LIFE and L. Banett. *Die Welt in der Wir Leben.* Droemersche Verlagsanstalt, München, Zürich.

[Lei37] Gottfried W. Leibniz. Discourse on metaphysics. In *Discourse on metaphysics, correspondence with Arnauld, and Monadology.* Open Court Publishing Company, Lasalle, Illinois, 1937. Transl.: George R. Montgomery.

[Mar82] David Marr. *Vision: A computational investigation into human representation and processing of visual information.* Freeman and Co., New York, San Francisco, 1982.

[MF88] Michael Mohnhaupt and D. Fleet. Raum-zeitliche Filter für eine top-down Steuerung der Bewegungsanalyse. In W. Hoeppner, editor, *GWAI-88*, pages 296–305, Berlin, 1988. Springer.

[MJL76] George A. Miller and P.N. Johnson-Laird. *Language and perception.* Cambridge University Press, Cambridge, UK, 1976.

[Moi79] Markku Moilanen. *Statische lokative Präpositionen im heutigen Deutsch – Wahrheits- und Gebrauchsbedingungen.* Max Niemeyer Verlag, Tübingen, 1979.

[MW83] Heinz Marburger and W. Wahlster. Case role filling as a side effect of visual search. In *Proceedings of the EACL-83*, pages 188–195, 1983.

[Nag88] H.-H. Nagel. From image sequences towards conceptual descriptions. *Image and Vision Computing*, 6(2):59–74, 1988.

[Neu33] Otto Neurath. Protokollsätze. *Erkenntnis*, 3:204–214, 1933.

[Pri88] Simone Pribbenow. Verträglichkeitsprüfungen für die Verarbeitung räumlichen Wissens. In W. Hoeppner, editor, *GWAI-88 – German Workshop on AI*, pages 226–235, Berlin, 1988. Springer.

[Pyl81] Zenon W. Pylyshyn. The imagery debate: Analogue media vs. tacit knowledge. *Psychological Review*, 88(1):16–45, 1981.

[RS88] Gudula Retz-Schmidt. Various views on spatial prepositions. *AI Magazine*, 9.2:95–105, 1988.

[RS91a] Gudula Retz-Schmidt. *Die Interpretation des Verhaltens mehrerer Akteure in Szenenfolgen.* Doctoral thesis, Technische Fakultät, Universität des Saarlandes, Saarbrücken, 1991. (reprint by Springer 1992).

[RS91b] Gudula Retz-Schmidt. Recognizing intentions, interactions, and causes of plan failures. *User Modeling and User-Adapted Interaction*, 1(1):173–202, 1991.

[Saf66] Hans Saf. *Der Ausdruck des Ortes in der deutschen Sprache der Gegenwart.* Doctoral thesis, Philosophische Fakultät der Friedrich-Schiller-Universität, Jena, 1966.

[SBSZ87] Jörg R.J. Schirra, G. Bosch, C.K. Sung, and G. Zimmermann. From image sequences to natural language: A first step towards automatic perception and description of motions. *Applied AI*, 1(3):287–305, 1987.

[Sch89] Jörg R.J. Schirra. Ein erster Blick auf ANTLIMA – Visualisierung statischer räumlicher Relationen. In D. Metzing, editor, *GWAI-89 – German Workshop on AI*, pages 301–311, Berlin, 1989. Springer.

[Sch90a] Jörg R.J. Schirra. Einige Überlegungen zu Bildvorstellungen in kognitiven Systemen. In C. Freksa and C. Habel, editors, *Repräsentation und Verarbeitung räumlichen Wissens*, pages 68–82. Springer, Berlin, 1990.

[Sch90b] Jörg R.J. Schirra. Expansion von Ereignis-Propositionen zur Visualisierung – Die Grundlagen der begrifflichen Analyse von ANTLIMA. In H. Marburger, editor, *GWAI-90 – German Workshop on AI*, pages 246–256, Berlin, 1990. Springer.

[Sch91] Jörg R.J. Schirra. Zum Nutzen antizipierter Bildvorstellungen bei der sprachlichen Szenenbeschreibung. VITRA-Memo 49, SFB 314, VITRA, Universität des Saarlandes, Saarbrücken, 1991.

[Sch93] Jörg R.J. Schirra. Connecting visual and verbal space – Preliminary considerations concerning the concept 'mental image'. In M. Aurnague, A. Borillo, M. Borillo, and M. Bras, editors, *Semantics of Time, Space, and Movement*, pages 105–121, Toulouse, 1993. Gruppe "Langue, Raisonnement, Calcul", CNRS und die Universitäten Paul Sabatier und Le Merail. Working Papers of the 4th International Workshop, Château de Bonas; (also as VITRA Report No. 90, 1992).

[Sch94] Jörg R.J. Schirra. *Sprachliche Bildbeschreibung als Verbindung von visuellem und sprachlichem Raum – Eine interdisziplinäre Untersuchung von Bildvorstellungen in einem Hörermodell.* Doctoral thesis, Technische Fakultät, Universität des Saarlandes, Saarbrücken, 1994, (to be reprinted by INFIX-Verlag, St. Augustin, 1995).

[SHfc] Jörg R.J. Schirra and Ellen M. Hays. Solvitur ambulando — an imaginative walk toward an understanding of spatial metaphor. techn. report, ICSI, Berkeley. forthcoming.

[Son78] Norm K. Sondheimer. A semantic analysis of reference to spatial entities. *Linguistics and Philosophy*, (2):235–280, 1978.

[Sun88] C.-K. Sung. Extraktion von typischen und komplexen Vorgängen aus einer langen Bildfolge einer Verkehrsszene. In H. Bunke, O. Kübler, and P. Stucki, editors, *Mustererkennung 88*, Berlin, 1988. Springer.

[Tal83] Leonard Talmy. How language structures space. In H. Pick and L. Acredolo, editors, *Spatial orientation: Theory, Research, and Application*, pages 225–282. Plenum, New York, 1983.

[Tug76] Ernst Tugendhat. *Vorlesungen zur Einführung in die sprachanalytische Philosophie.* Suhrkamp, Frankfurt/M., 1976.

[Ull85] Shimon Ullman. Visual routines. In S. Pinker, editor, *Visual Cognition*, pages 97–159. MIT Press, Cambridge, MA, 1985.

[Van86] Claude Vandeloise. *L'espace en français.* Editions du Seuil, Paris, 1986.

[Wah88] Wolfgang Wahlster. One word says more than a thousand pictures – on the automatic verbalization of the results of image sequence analysis systems. *T.A. Informations*, Special Issue: Linguistique et Informatique en République Fédérale Allemande, 1988.

[Win75] Terry Winograd. Frame representations and the declarative-procedural controversy. In D.G. Bobrow and A. Collins, editors, *Representation and Understanding*, pages 35–82. Academic Press, New York, NY, 1975.

[Wit63] Ludwig Wittgenstein. *Tractatus logico-philosophicus.* Suhrkamp, Frankfurt/M., 1963. (es12).

[YND88] Atsushi Yamada, T. Nishida, and S. Doshita. Figuring out most plausible interpretations from spatial descriptions. In *Proceedings of the 12th International Conference on Computational Linguistics, Budapest*, pages 764–769, 1988.

[ZW90a] Cornelia Zelinsky-Wibbelt. The semantic representation of spatial configurations: a conceptual motivation for generation in machine translation. In *Proceedings of the 13th International Conference on Computational Linguistics, Helsinki*, volume 3, pages 299–303, 1990.

[ZW90b] Cornelia Zelinsky-Wibbelt. Using cognitive principles for the interpretation of spatial configurations in machine translation. In W. Hoeppner, editor, *Workshop Räumliche Alltagsumgebungen des Menschen – RAUM-90*, pages 209–220, Koblenz, 1990. Univ. Koblenz-Landau.

(Figures 1 and 2 are produced on the base of two drawings of Charles M. Schultz and a picture of the earth taken from [LB])

published in: Cornelia Zelinsky-Wibbelt (ed.), *The Semantics of Prepositions – From Mental Processing to Natural Language Processing*, p. 471–515. Mouton de Gruyter, Berlin, 1993.

(this is a slightly changed version for American paper format, arranged in Oct. 94; JRJS)