Ein Wort sagt mehr als 1000 Bilder

Zur automatischen Verbalisierung der Ergebnisse von Bildfolgenanalysesystemen

Wolfgang Wahlster
Fachbereich 10 - Informatik
Universität des Saarlandes
6600 Saarbrücken 11

Abstract

Die Arbeit gibt eine kompakte Einführung in die Zielsetzung des Forschungsprojektes VITRA, in dem Möglichkeiten der sprachlichen Bildbeschreibung untersucht werden, und stellt anhand von Beispielen einige der Resultate des Vorhabens dar. Bei der Kopplung bildverstehender und sprachverstehender Systeme werden zwei wichtige Forschungsrichtungen der Künstlichen Intelligenz zusammengeführt. Nach einer Darstellung der verschiedenen Diskursbereiche von VITRA wird erläutert, wie räumliche Relationen systemintern repräsentiert werden. Danach wird an Beispielen gezeigt, wie das System sein Wissen über die Bedeutung sprachlicher Ausdrücke, die sich auf räumliche Konzepte beziehen, bei der Generierung deutschsprachiger Bildbeschreibungen verwendet. Da die kurze Darstellung nur eine erste Orientierung ermöglichen kann, werden für den an Details interessierten Leser zahlreiche Hinweise auf ausführlichere Forschungsberichte gegeben.

1. Einleitung

Wenn wir eine Folge von Fernsehbildern von einem Autobahnabschnitt sehen, auf dem sich mehrere hundert Fahrzeuge hintereinanderreihen und nur im Kriechtempo vorankommen, so sind wir in der Lage, das Gesehene mit der sprachlichen Beschreibung 'Stau' zusammenzufassen. Dies ist ein typisches Beispiel für eine große Klasse von Situationen, die in Umkehrung eines klassischen Sprichwortes durch die Feststellung 'Ein Wort sagt mehr als 1000 Bilder' beschrieben werden können.
Ein Teilziel der Forschungen im Bereich der Künstlichen Intelligenz (KI) ist es, Programmsysteme zu entwickeln, die visuelle Information sprachlich umsetzen können. In dieser Arbeit wird über den derzeitigen Entwicklungsstand des KI-Systems VITRA (VIsual TRAnslator) berichtet, das im Rahmen des Sonderforschungsbereichs (SFB) 314 'Künstliche Intelligenz -Wissensbasierte Systeme' am Fachbereich Informatik der Universität des Saarlandes unter meiner Leitung aufgebaut wird. Das seit 1985 von der Deutschen Forschungsgemeinschaft geförderte Projekt soll einen Beitrag zur Grundlagenforschung auf dem Gebiet der Kopplung bildverstehender und sprachverstehender Systeme leisten.
Langfristig verfolgt man mit dieser Forschungsrichtung zwei Hauptziele:

(A1) Die komplexen Informationsverarbeitungsprozesse des Menschen, die der Interaktion von Sprachproduktion und visueller Wahrnehmung zugrundeliegen, sollen mit informatischen Mitteln exakt beschrieben und erklärt werden.

(A2) Durch die sprachliche Bildbeschreibung sollen dem Benutzer die Ergebnisse eines bildverstehenden Systems besser zugänglich und verständlich gemacht werden.

Charakteristisch für die KI-Forschung ist, daß neben dem kognitionswissenschaftlichen Erkenntnisinteresse (A1) auch eine ingenieurwissenschaftliche Zielsetzung (A2) verfolgt wird. Ein großer praktischer Vorteil der sprachlichen Bildbeschreibung besteht in der Möglichkeit zur anwendungsspezifischen Wahl unterschiedlicher Verdichtungsgrade für visuelle Information. So kann die beispielsweise in der Medizintechnik, der Fernerkundung und der Verkehrssteuerung anfallende Flut von Bilddaten nur noch maschinell bewältigt werden. Im Gegensatz zu einer Repräsentation der Verarbeitungsergebnisse von digitisierten Bildfolgen in Form von Computergraphiken kann eine sprachliche Bildbeschreibung dem Anwender mehr Information in

weniger Zeit liefern. Wenn ein KI-System in der Lage ist, das Interpretationsergebnis für eine Bildfolge in einer medizinischen Anwendung mit 'Verengung der linken Nierenarterie' zu beschreiben, so kann der Arzt diese Aussage zunächst direkt in den diagnostischen Zusammenhang einordnen und kann dann später bei Bedarf gezielt auf Ausschnitte relevanter Einzelbilder zurückgreifen.

2. Die Diskursbereiche von VITRA

Derzeit ist man von einem universell einsetzbaren KI-System, das beliebige Bildfolgen sprachlich beschreibt, noch sehr weit entfernt und muß sich bei der Systementwicklung jeweils auf eingeschränkte Diskursbereiche konzentrieren (Möglichkeiten und Grenzen wissensbasierter Systeme zum Bild- und Sprachverstehen werden in [Nagel 1985] bzw. in [Wahlster 1982] dargestellt). Im Projekt VITRA werden vier verschiedene Diskursbereiche und zwei unterschiedliche Kommunikationssituationen betrachtet, um möglichst frühzeitig die Übertragbarkeit der entwickelten Konzepte und Methoden auf andere Domänen prüfen zu können:

Kommunikationssituation KI:	Beantwortung natürlichsprachlicher Anfragen über räumliche Relationen und Bewegungsverläufe nach Ablauf einer Bildsequenz
Diskursbereich Dl:	CITYTOUR Stadtplanausschnitt von Saarbrücken mit Trajektorien bewegter Objekte
Diskursbereich D2:	UNITOUR Lageplan des Campus der Universität des Saarlandes
Diskursbereich D3:	DURLACHER TOR Straßenverkehrsszene aus Karlsruhe mit Trajektorien bewegter Objekte
Kommunikationssituation K2:	Simultane Berichterstattung über beobachtete Ereignisse während des Ablaufs einer Bildsequenz
Diskursbereich D4:	SOCCER Ausschnitte aus Fernsehübertragungen von Fußballspielen

Während in (KI) die Rolle des KI-Systems der eines Ortskundigen ähnelt, der Auskünfte erteilt, ist sie in (K2) mit einem Radioreporter vergleichbar. Wie in den früheren Systemen HAM-ANS (vgl. [Wahlster et al. 1983]) und NAOS (vgl. [Neumann/Novak 1986]) erfolgt in der Kommunikationssituation (KI) die Beschreibung a posteriori, während in (K2) durch die simultane Beschreibung völlig neuartige Problemstellungen auftreten (vgl. [André et al. 1987], [Rist et al. 1987]). Für beide Typen von Situationen gibt es zahlreiche realistische Anwendungsszenarios. So könnte beispielsweise ein Biologe im ersten Fall aufgrund einer Folge ausgewerteter Luftbilder fragen 'Wo wurden Schädigungen von Birkenbeständen festgestellt?'. Im zweiten Fall erwartet z.B. der Bediener eines Leitstandes für ein komplexes technisches System eine Beschreibung einer sich anbahnenden Fehlfunktion oder eine Warnung vor einer potententiellen Betriebsstörung.
Während es sich bei (Dl) und (D2) um synthetisches Bildmaterial handelt, für das eine geometrische Szenenbeschreibung vorgegeben ist, handelt es sich bei (D3) und (D4) um Bildfolgen aus der natürlichen Umwelt. Für (D3) werden die Trajektorien der bewegten Objekte durch ein von einer ebenfalls im SFB 314 tätigen Forschungsgruppe des Fraunhofer-Instituts für Informations- und Datenverarbeitung (FhG-IITB) in Karlsruhe entwickeltes Bildfolgenanalysesystem (vgl. [Zimmermann et al. 1987]) ermittelt, so daß ein Teil der geometrischen Szenenbeschreibung bereits algorithmisch erzeugt wird (die geometrische Beschreibung des statischen Hintergrunds ist noch fest vorgegeben). Dies wird für (D4) langfristig auch angestrebt, ist aber u.a. dadurch wesentlich schwieriger, daß hier im Gegensatz zu (D3) keine starre Kameraposition vorausgesetzt werden kann und die bewegten Objekte meist nicht-starre Körper sind. Zur Zeit werden Beispiele für die SOCCER-Domäne noch mit einem graphischen Trajektorien-Editor menügesteuert aufgebaut (vgl. [Herzog 1986]).
Mit der geglückten Kopplung des am IITB entwickelten bildverstehenden Systems mit VITRA ist es im Sonderforschungsbereich 314 weltweit erstmals gelungen, ein KI-System zu entwickeln, das eine sprachliche Bildbeschreibung von Bewegungen in einer von einer Video-Kamera aufgenommenen Bildfolge ohne jeglichen manuellen Eingriff durchführt. Dabei werden die Bild- (zwischen 250KB und 1MB pro Bild) und Ergebnisdaten (7 Dateien mit je 12 KB) über drei miteinander verbundene Computernetze (DFN, CANTUS, ETHERNET) von einem VAX-Rechner des IITB in Karlsruhe auf einen SYMBOLICS-Rechner des KI-Labors in Saarbrücken übertragen.
Für den Diskursbereich (D3) wurde vom Dach eines 35m hohen Gebäudes mit einer starren Video-

Kamera das Geschehen auf der Straßenkreuzung am Durlacher Tor aufgezeichnet (vgl. [Zimmermann et al. 1987]). Eine Sequenz von 130 Bildern (5.2 sec) wurde digitisiert (512 * 512 pixel, 8 bit Grauwerte) und vom Bildfolgenanalysesystem der FhG verarbeitet. Das System fand 10 Kandidaten für bewegte Objekte und deren Trajektorien trotz Verdeckungen durch Bäume und Straßenlampen. Auf den zentralen Fahrbahnen bewegten sich eine Straßenbahn von links nach rechts, zwei kleine LKWs, drei PKWs und ein Fahrrad von rechts nach links. Auf der Fahrbahn im oberen Bildabschnitt wurden drei PKWs erkannt. Ein viertes Auto stand bereits am Anfang der ausgewerteten Bildfolge und wurde daher durch das Bewegungsdetektionsverfahren nicht berücksichtigt. VITRA erkennt auf der Grundlage der Trajektorienverläufe als höhere Bewegungskonzepte, daß u.a. drei Autos vor einer Ampel anhalten und die Straßenbahn entlang der Kaiserstraße fährt. Abb. 1 zeigt einen Bildschirmabzug des hochauflösenden Graphik-Monitors der LISP-Maschine, auf der VITRA implementiert ist. Das große rechte Graphikfenster zeigt ein digitisiertes Bild aus der verarbeiteten Sequenz. Die Trajektorien der erkannten Fahrzeuge sind graphisch mit dem Originalbild überlagert. Jede systemintern als Liste von Ort-Zeit-Paaren repräsentierte Objekttrajektorie wird graphisch durch einen benannten Kantenzug dargestellt, auf dem Zeitmarken durch ihren Abstand und ihre Numerierung die Geschwindigkeit bzw. die Richtung der beobachteten Bewegung codieren. Das kleine innerhalb des Graphikfensters eingeblendete Menü erlaubt die mausgesteuerte Selektion eines anderen Diskursbereichs.
Das linke Drittel des Bildschirms enthält drei Textfenster: Eingabe, Dialog und Trace. Die in deutscher Sprache formulierten Anfragen des Benutzers werden in das Eingabefenster eingetippt. Im Dialogfenster erscheint die Antwort von VITRA zusammen mit einigen vorangegangenen Frage/Antwort-Paaren, die einen Teil des Dialogkontextes darstellen. Im Trace-Fenster hat der Benutzer u.a. die Möglichkeit, die internen Verarbeitungsabläufe des Systems zu verfolgen oder sich die von VITRA verwendeten Wissensquellen anzeigen zu lassen. Das Trace-Fenster und das Graphikfenster sind für die Weiterentwicklung von VITRA besonders wichtig, weil mit diesen Hilfsmitteln Wissenslücken des Systems aufgespürt und die Ursachen für fehlerhaftes oder unangemessenes Antwortverhalten analysiert werden können.

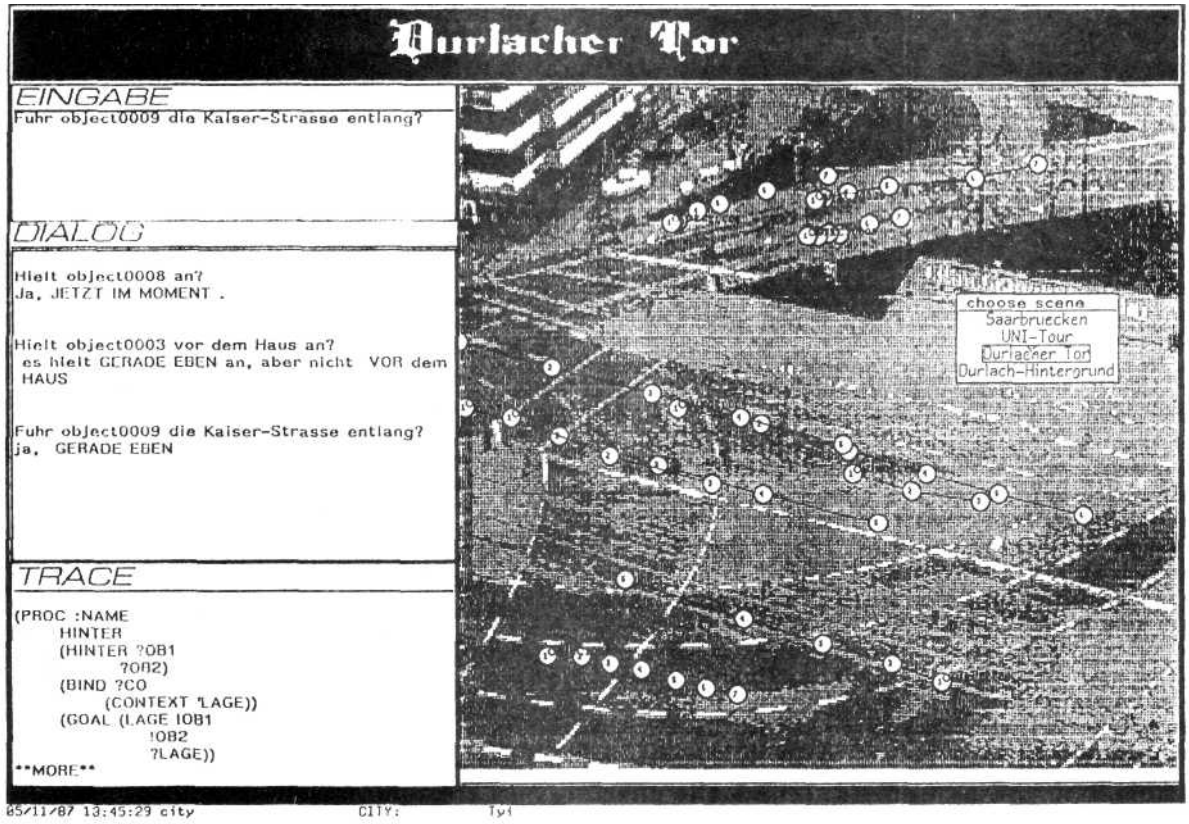


Abb. 1: Beispieldialog mit VITRA über Bewegungsabläufe am Durlacher Tor

In realistischen Anwendungsumgebungen für die Kommunikationssituation (K1) sind natürlich nur das Eingabe- und das Dialogfenster von Interesse.
In Abb. 1 werden bei der Eingabe systeminterne Objektbezeichner verwendet (z.B. 'object0009' statt 'die Straßenbahn'). Wie die Beispieldialoge in den Abb. 2 und 3 zeigen, ist dies keine Einschränkung von VITRA, sondern beruht allein auf der Tatsache, daß von der kooperierenden Forschungsgruppe am IITB bisher für diese Szene keine Objektidentifikation durchgeführt worden ist. Prinzipiell gibt es mehrere modellgesteuerte KI-Verfahren, welche die erkannten Objekte den Klassen PKW, LKW und Straßenbahn zuordnen könnten. Im bisherigen Projektverlauf haben wir uns bei der Kopplung aber auf die Bewegungsanalyse konzentriert, da hier in der Grundlagenforschung größere Methodendefizite vorlagen. Die Antworten des Systems in den Dialogfenstern der Abb. 1-3 zeigen, daß neben direkten Antworten auf Fragen des Benutzers auch Zusatzinformation generiert wird (z.B. Zeitangaben, erkannte Präsuppositionsverletzungen, Quelle und Ziel einer Bewegung, Modifikation der erfragten Prädikation). Solch eine Überbeantwortung von Entscheidungsfragen (vgl. [Wahlster et al. 1983]) ist eine wesentliche Voraussetzung für die Entwicklung kooperativer Zugangssysteme, die ihr eigenes Verhalten dynamisch an die Handlungsziele und das Vorwissen des Benutzers anpassen können (vgl. [Wahlster 1984]).
Das Trace-Fenster in Abb. 1 zeigt einen Ausschnitt einer Wissensquelle von VITRA, in der die formale Semantik der räumlichen Präpositionen definiert ist, welche VITRA verstehen und generieren kann. Es ist ein Auszug aus der Definition der Relation 'hinter' zu sehen, die mithilfe der KI-Programmiersprache FUZZY codiert ist. FUZZY ist in die Symbolverarbeitungssprache LISP einbettet und enthält zusätzlich u.a. Mechanismen zur automatischen Schlußfolgerung, zum Zugriff auf assoziative Datenbasen und zum pattern-gesteuerten Prozeduraufruf.
Bevor die Repräsentation der Semantik solcher Präpositionen in Abschnitt 4 an Beispielen illustriert wird, soll im folgenden Abschnitt anhand der Diskursbereiche CITYTOUR und UNITOUR zunächst die in VITRA verwendete geometrische Szenenbeschreibung als Grundlage für die Verarbeitung räumlicher Relationen skizziert werden.

3. Die Repräsentation räumlicher Relationen

In CITYTOUR wird zwischen statischen (d.h. unbeweglichen) und dynamischen (d.h. beweglichen) Objekten unterschieden. Als statische Objekte können u.a. Straßen, Häuser und Plätze vorkommen (vgl. Abb. 2). Straßen sind durch ihre rechten und linken Ränder sowie Mittellinien repräsentiert. Alle anderen statischen Objekte werden in der geometrischen Szenenbeschreibung als Polygonzüge dargestellt. Aus den Polygonzügen können umschreibende Rechtecke und deren Schwerpunkte berechnet werden, die als gröbere Repräsentationsformen für schnelle, approximative Berechnungen räumlicher Relationen verwendet werden. Als weitere Eigenschaft eines Teils der statischen Objekte ist eine ausgezeichnete Vorderseite (z.B. Haupteingang der Post in Abb. 2) definiert, die im Graphikfenster durch eine fett gezeichnete Objektkante hervorgehoben wird (vgl. Abb. 2). Durch ein maus-sensitives Menü, das im Graphikfenster eingeblendet werden kann, wird die Menge der jeweils dargestellten Objektklassen definiert. In Abb. 2 sind bis auf die Strassen alle intern repräsentierten Objekte sichtbar.
Als dynamische Objekte können Fußgänger, Radfahrer und Fahrzeuge (Autos, Busse, Straßenbahnen) auftreten. Diese werden in ihrer Repräsentation nicht unterschieden, sondern sind alle gleichermaßen als Punkte (Objektschwerpunkte) repräsentiert. Ihre Trajektorien sind als Listen von Ort-Zeit-Paaren repräsentiert, wobei die Orte wiederum als Paare von x- und y-Koordinaten dargestellt sind. Auf dem Bildschirm sind die Trajektorien über den gesamten Szenenzeitraum hinweg in ein Bild des statischen Hintergrundes projiziert (vgl. Abb. 2). Räumliche Relationen werden als atomare Formeln im Sinne der Prädikatenlogik repräsentiert, die aus einem Prädikat und mehreren Termen bestehen. Das Prädikat entspricht einer Präposition (oder einem gleichwertigen Ausdruck) in der natürlichen Sprache (z.B. "vor", "bei", "links von"). Das erste Argument wird *Subjekt* genannt. Es ist dasjenige Objekt, das in bezug auf ein (in den meisten Fällen) oder mehrere (z.B. im Falle von "zwischen", vgl. Abb. 4) andere Bezugsobjekte lokalisiert werden soll.
Es wird unterschieden zwischen *statischen Relationen*, bei denen sowohl Subjekt als auch Bezugsobjekt(e) unbewegt sind, und *dynamischen Relationen*, bei denen das Subjekt bewegt ist. Bisher können nur unbewegte Bezugsobjekte behandelt werden.
Während statische Relationen den Ort von unbewegten Objekten angeben, können durch dynamische

Relationen die Richtung (z.B. "hinter das Rathaus gehen") und der Pfad (z.B. "am Saarcenter vorbei gehen", "abbiegen") von Trajektorien bewegter Objekte beschrieben werden. Außer den dynamischen Relationen ist die Semantik einiger Bewegungsverben wie "anhalten" und "anfahren" implementiert. Die durch Bewegungsverben charakterisierten Ereignisse können ebenso wie unbewegte Objekte durch statische Relationen lokalisiert werden (z.B. "vor der Post anhalten").

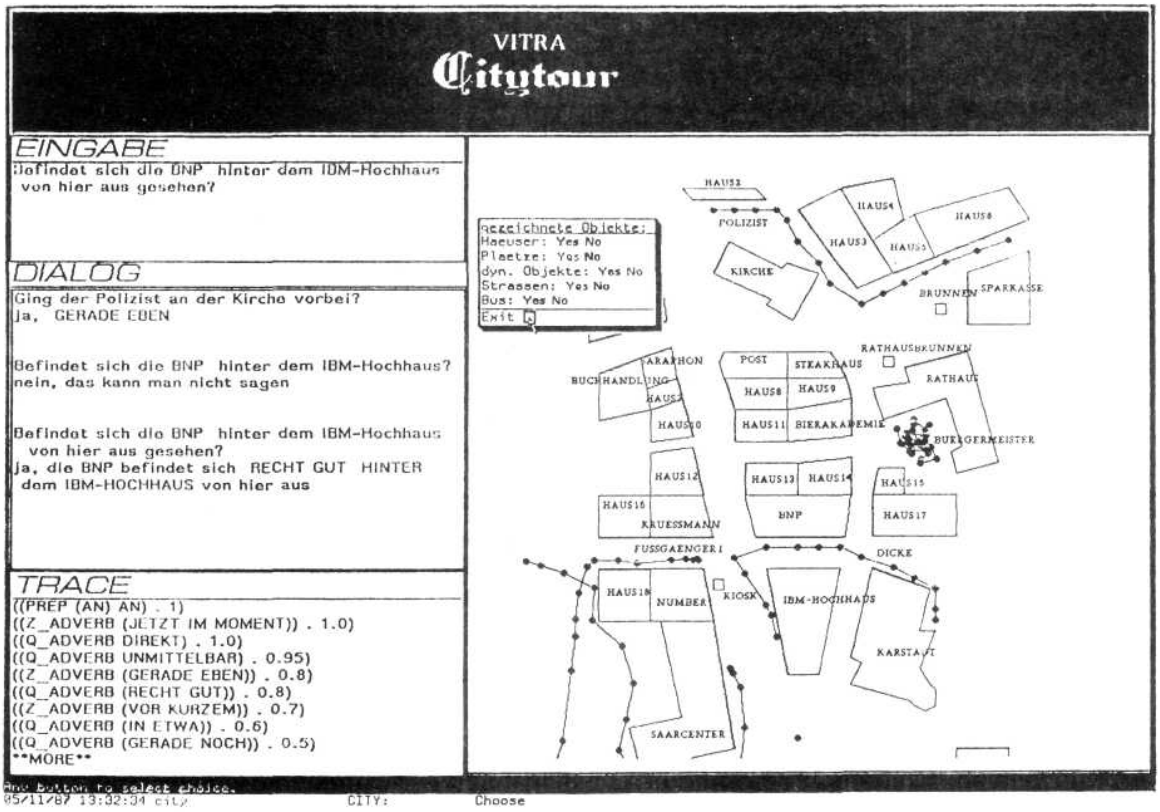


Abb. 2: Beispieldialog für den Diskursbereich CITYTOUR

Einige Präpositionen erlauben es, das Subjekt entweder in bezug auf intrinsische Eigenschaften des Bezugsobjektes oder aber bezüglich eines anderen Beobachterstandortes zu lokalisieren. Im ersten Fall sprechen wir von *intrinsischem,* im zweiten von *extrinsischem Gebrauch* der Präposition. Ist der Beobachterstandort der Standort des Sprechers (oder Hörers), so sprechen wir von *deiktischem Gebrauch* (vgl. [Retz-Schmidt 1986b] für eine ausführliche Diskussion dieser Unterscheidungen).
Ein Beispiel, das diesen Unterschied verdeutlicht, ist in Abb. 2 zu sehen. Sprecher wie Hörer werden als an dem durch einen einzelnen dicken schwarzen Punkt gekennzeichneten Ort befindlich angenommen (vgl. den Punkt unterhalb des mit 'IBM-Hochhaus' bezeichneten Polygons in der Mitte des unteren Randes im Graphikfenster von Abb. 2). Man kann sie sich z.B. als Tourist und Führer bei einer Stadtrundfahrt vorstellen, wobei dem Führer vom Touristen Fragen über die räumliche Lage der Gebäude in der Stadt gestellt werden. In der Kommunikationssituation (KI) erlaubt VITRA die beliebige Positionierung des fiktiven Sprecher/Hörer-Standortes durch graphische Operationen (Verschiebung des entsprechenden Bildpunktes), um die Situationsabhängigkeit räumlicher Beschreibungen testen und vorführen zu können. VITRA ermöglicht intrinsischen und deiktischen Gebrauch bei folgenden Präpositionen: "vor", "hinter", "rechts von", "links von" und "neben" in ihren lokativen und direktionalen Gebräuchen sowie in Kombination mit "vorbei".

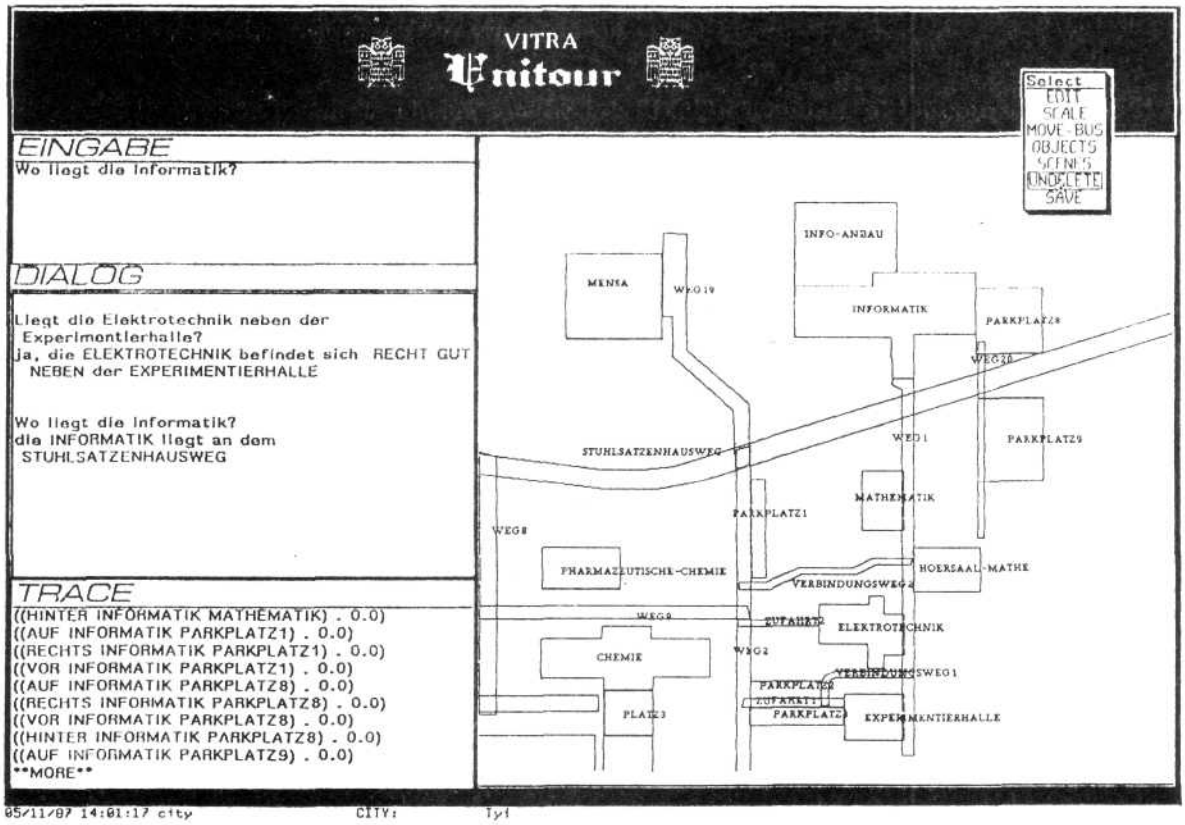
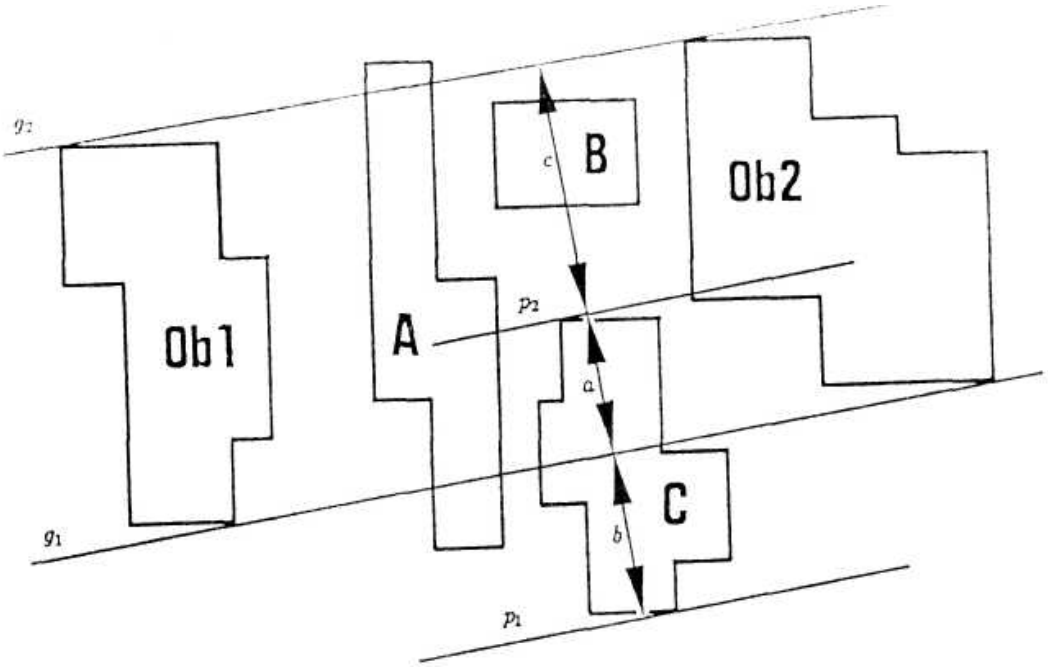


Abb. 3: Beispieldialog für den Diskursbereich UNITOUR

Im Falle des lokativen Gebrauchs prüft das System wie folgt, ob die jeweilige Relation erfüllt ist: Die Fläche um das Bezugsobjekt wird in vier Halbebenen unterteilt: eine vordere, eine hintere, eine rechte und eine linke. Diese vier Halbebenen werden an den Seiten eines umschreibenden Rechtecks des Bezugsobjektes orientiert. Im Falle des intrinsischen Gebrauchs wird das umschreibende Rechteck an der ausgezeichneten Vorderseite des Bezugsobjektes orientiert. Im Falle des deiktischen Gebrauchs wird das umschreibende Rechteck an der Blickrichtung vom Beobachterstandort zum Bezugsobjekt orientiert (für Details vgl. [André et al. 86b]). In beiden Fällen kann dann bestimmt werden, in welcher/n der Halbebenen das Subjekt liegt.
Bei der Entscheidung über intrinsischen oder deiktischen Gebrauch verwendet VITRA folgende - an [Miller/Johnson-Laird 76] angelehnte - Strategie: Es wird vom intrinsischen Gebrauch ausgegangen, solange ein deiktischer Gebrauch nicht explizit (z.B. durch "von hier aus") sprachlich markiert ist und das Bezugsobjekt eine ausgezeichnete Vorderseite besitzt. Andernfalls wird vom deiktischen Gebrauch ausgegangen.
Wie die Beispiele in Abb. 2 zeigen, stellt das System fest, daß bei einer intrinsischen Interpretation die BNP nicht hinter dem IBM-Hochhaus liegt, während der deiktische Gebrauch zu einer positiven Beantwortung der eingegebenen Frage führt.

4. Zur Semantik räumlicher Präpositionen

Obwohl die vier Halbebenen um die Bezugsobjekte unendlich groß sind, erscheint es nicht plausibel, daß z.B. die Relation "vor" noch gelten soll, wenn das Subjekt sehr weit vom Bezugsobjekt entfernt liegt. Andererseits erscheint es unangemessen, eine scharfe Grenzlinie zwischen dem Bereich, in dem eine räumliche Relation gilt, und dem Bereich, in dem sie nicht mehr gilt, zu ziehen.



1. Schritt:	Berechne die beiden Tangenten g_1 und g_2 der Bezugsobjekte Ob_1 und Ob_2
2. Schritt:	Berechnung der Anwendbarkeit:
Fall A:	Beide Tangenten g_1 und g_2 schneiden Objekt A => Anwendbarkeit = 1
Fall B:	Objekt B liegt vollständig innerhalb des von g_1 und g_2 aufgespannten Zwischenraums => Anwendbarkeit = 1
Fall C:	Objekt C schneidet eine Tangente (g_1) => Anwendbarkeit = $max\left[\frac{a}{a+b}, \frac{a}{a+c}\right]$

Abb. 4: Anwendbarkeitsgrade für die räumliche Relation 'zwischen'

Es werden daher sogenannte *Anwendbarkeitsgrade* für die Relationen berechnet, die von der Entfernung zwischen Subjekt und Bezugsobjekt sowie von der Größe des Bezugsobjektes abhängen. Die Anwendbarkeitsgrade werden systemintern durch Werte aus dem reellen Intervall [0,1] dargestellt. Die Anwendbarkeitsgrade werden bei der Beantwortung von Entscheidungsfragen als linguistische Hecken verbalisiert, wie in dem Beispiel:

"Liegt der Rathausbrunnen vor der Bierakademie?"
"Ja, der Rathausbrunnen liegt IN ETWA vor der Bierakademie."

Zum anderen können sie bei der Beantwortung von Wo-Fragen (vgl. Fig. 3) zur Auswahl des geeignetsten Bezugsobjektes dienen. Im Falle von Wo-Fragen wird die Anwendbarkeit der vier Grundrelationen für verschiedene benachbarte Bezugsobjekte überprüft, wobei noch ein Auffälligkeitswert in die Berechnung einfließt, bevor das Bezugsobjekt mit der höchsten Anwendbarkeitsgrad für die Lokationsbeschreibung gewählt wird.

"Wo liegt die Informatik?"
"Die Informatik liegt an dem Stuhlsatzenhausweg."

Als Beispiel für die Berechnung der Anwendbarkeitsgrade veranschaulicht Abb. 4 das Verfahren für die statische Relation "zwischen". Für die drei möglichen Subjekte A, B und C wird geprüft, inwiefern bezüglich der Objekte Ob1 und Ob2 die Relation "zwischen" besteht. Wenn keiner der Fälle A, B oder C vorliegt, wird der Anwendbarkeitsgrad 0 berechnet, was der Negation der entsprechenden atomaren Formel entspricht.
Für die Bestimmung der Gültigkeit dynamischer Relationen wird in Kommunikationssituation (KI) davon ausgegangen, daß die Trajektorien vollständig bekannt sind. Die Trajektorien sind in die statische Szene hineinprojiziert (vgl. Abb. 2) und ihr letzter Punkt entspricht dem Ort des bewegten Objekts zum aktuellen Beobachtungszeitpunkt.
In früheren Arbeiten zur sprachlichen Bildbeschreibung wurde eine im Vergleich zu VITRA wesentlich einfachere geometrische Szenenbeschreibung zugrundegelegt, in der auch die statischen Objekte durch Schwerpunktkoordinaten repräsentiert sind (vgl. z.B. [Wahlster et al. 1978]). Dies hatte zur Folge, daß die Semantik von Pfadpräposition wie "vorbei" und "entlang", welche die Trajektorie eines bewegten Subjekts in Relation zur Begrenzungslinie oder -fläche eines anderen Objektes setzen, nicht adäquat beschreiben werden konnte. In VITRA konnte dagegen erstmals die Feinsemantik der Pfadpräpositionen "entlang" und "vorbei" genauer untersucht werden, wobei sich folgende Unterschiede ergaben:
In beiden Fällen darf der Abstand zwischen der Trajektorie des bewegten Subjekts und dem Bezugsobjekt einen bestimmten, von der Größe des Bezugsobjektes abhängigen Grenzwert nicht überschreiten. Dieser ist im Falle von "entlang" jedoch niedriger als im Falle von "vorbei". Weiterhin muß für "entlang" die Trajektorie der Kontur des Bezugsobjektes genauer folgen als für "vorbei". Während für "vorbei" daher die gröbere Repräsentationsform *umschreibendes Rechteck* ausreicht, wird für die Berechnung von "entlang" der *Polygonzug* des Bezugsobjektes verwendet. Ein weiterer Unterschied ist, daß im Falle von "entlang" die Trajektorie nicht zwischendurch ihre Richtung ändern darf. Während im Falle von "hinter/vor/rechts/links ... vorbei" die gesamte Fläche zwischen den jeweils gegenüberliegenden Halbebenen durchquert werden muß, braucht für "entlang" nur ein genügend langer Weg längs des Bezugsobjektes zurückgelegt zu werden. Beispiele für Trajektorien, die den Unterschied zwischen "entlang" und "vorbei" verdeutlichen, sind in Abb. 5 zusammengestellt. Die hier skizzierten Unterschiede flossen in die Realisierung der Semantik von "vorbei" und "entlang" in VITRA ein (siehe [André et al. 86a]).

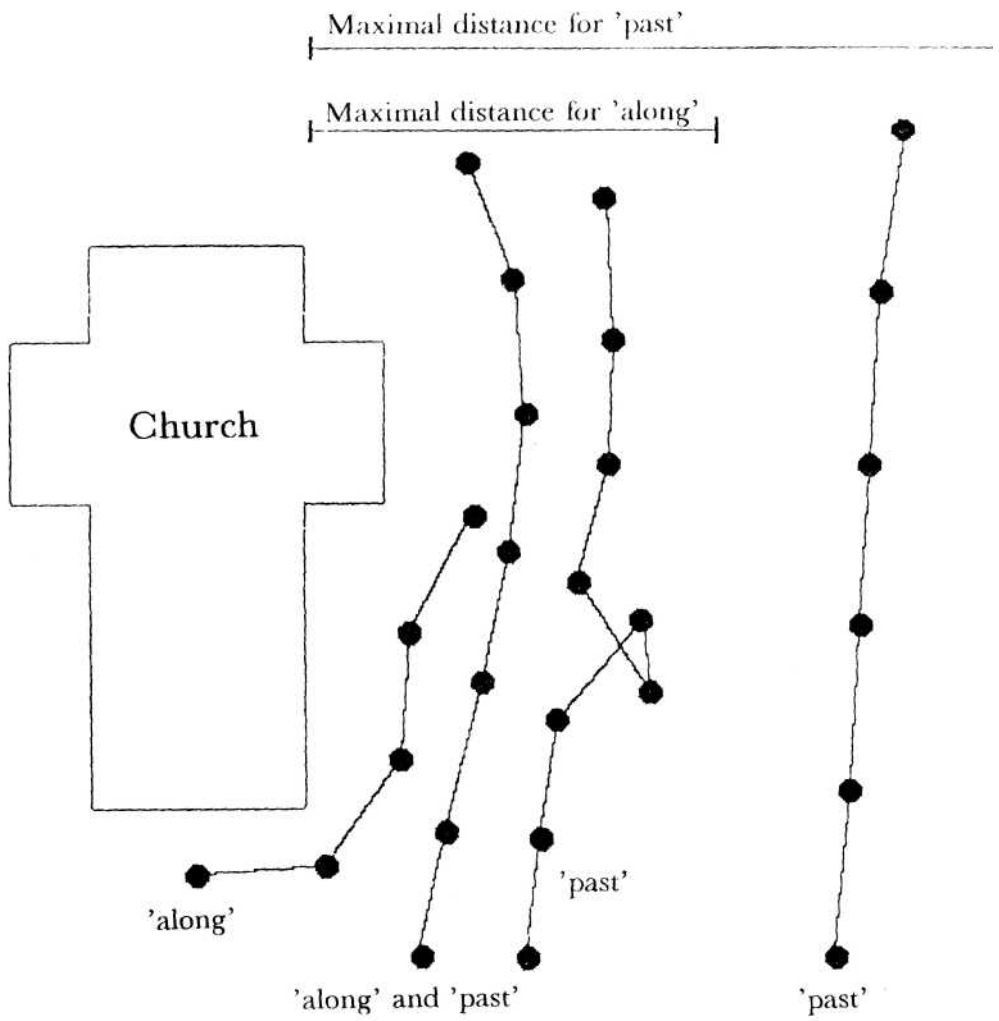


Abb. 5: Beispiel-Trajektorien für zwei Pfadpräpositionen

5. Aktuelle Forschungsaufgaben

Die vorangegangenen Abschnitte haben gezeigt, daß die im Projekt VITRA entwickelten KI-Systeme bereits in der Lage sind, einfache sprachliche Beschreibungen von Bildsequenzen zu erzeugen. Von den sprachlichen, kommunikativen und kognitiven Leistungen eines menschlichen Beobachters ist die Performanz der bisher entwickelten Systemkomponenten jedoch noch weit entfernt. Daher müssen während der weiterhin im Rahmen des SFB 314 geförderten zweiten Projektphase von 1988-1990 noch zahlreiche offene Probleme einer formalen Rekonstruktion des Zusammenspiels zwischen "Sehen und Sprechen" gelöst werden. Im folgenden sollen drei dieser aktuellen Fragestellungen erläutert werden:

(a) die Erkennung und Beschreibung von Gruppen bewegter Objekte
(b) Der Einfluß von vermuteten Handlungsintentionen bei der Beschreibung von Bewegungen
(c) die Entwicklung einer Imaginationskomponente für ein Benutzermodell, das eine Antizipationsrückkopplungsschleife ermöglicht.

Auf einem Fußballfeld können sich 26 Objekte gleichzeitig bewegen: 22 Spieler, 1 Ball, 1 Schiedsrichter und 2 Linienrichter. Ein während einer Reportage geäußerter Satz kann in den Extremfällen die Bewegung eines einzelnen Objektes (1) oder aller 26 Objekte (2) umfassen.

(1) Der Ball rollt ins Aus.
(2) Das Spiel wird langsamer.

In den meisten Fällen beziehen sich die Beschreibungen auf Elemente der Potenzmenge aller bewegten Objekte (3).

(3) Die Verteidung des FC baut eine Mauer auf.

Dies verdeutlicht die enorme Selektionsaufgabe, die der Mensch in extrem kurzer Zeit bewältigt, da rein rechnerisch zwischen $2^{26} - 1$, d.h. 67.108.863 potentiellen Gruppierungen auszuwählen ist.

Ein weiteres, im allgemeinen Fall ungelöstes Problem, das schon bei der Beschreibung von Vorgängen in Straßenverkehrsszenen auftrat, besteht darin, daß nicht allein der Verlauf einer Trajektorie in Zeit und Raum für die Auswahl einer adäquaten Beschreibung entscheidend ist. So können zwar alle temporalen und lokalen Bedingungen eines 'Park-Ereignisses' für eine beobachtete Trajektorie eines Fahrzeuges erfüllt sein, aber dennoch wird eine Beschreibung wie (4) als inadäquat empfunden.

(4) Das Auto parkt vor der Ampel.

Nur wenn man die Handlungsabsicht berücksichtigt (vgl. [Retz-Schmidt 1986a]), kommt man bei der gleichen Trajektorie zu einer angemessenen Beschreibung wie (5).

(5) Das Auto wartet vor der Ampel.

Ein Kriterium zur Auswahl des Diskursbereichs 'Fußballreportage' war die Tatsache, daß hier der Einfluß der vermuteten Intentionen der Handelnden bei der Beschreibung besonders klar hervortritt. So beschreiben (7) und (8) den gleichen zeitlich-räumlichen Vorgang, unterstellen aber jeweils eine andere Mannschaftszugehörigkeit des Spielers A.

(7) A klärt durch einen Schuß hinter die Torauslinie.

(8) A verpaßt das Tor knapp.

In (7) hat der Spieler nicht die Absicht, den Ball ins Tor zu befördern, sondern schießt ihn absichtlich ins Aus. Dagegen galt der Schuß des angreifenden Spielers in (8) sicherlich dem Tor, was in der Formulierung 'verpaßt' zum Ausdruck gebracht wird.

Ein Vorteil des gewählten Diskursbereichs besteht darin, daß mit der Spielerposition, der Mannschaftszugehörigkeit und der Rollenverteilung in Standardsituationen (z.B. Strafstoß, Eckstoß) jeweils eine stereotype Intention vorgegeben ist, so daß beim derzeitigen Forschungsstand zur Planerkennung bessere Erfolgsaussichten als in den meisten anderen, weniger schematisierten Handlungssituationen für unser Arbeitsziel (b) bestehen.

Ein drittes Problem, das wir z.Zt. untersuchen, entsteht dadurch, daß sich das System für die Generierung kommunikativ adäquater Beschreibungen ein Modell von den visuellen Vorstellungen machen muß, die es beim Hörer seines Berichtes hervorruft. Ein solches Benutzermodell (vgl. [Wahlster/Kobsa 1986]) kann z.B. für die Entscheidung relevant werden, ob statt einer definiten Kennzeichnung auch ein Pronomen in der nächsten zu generierenden Äußerung für den Hörer verständlich ist. Nehmen wir an, das System hat soeben folgenden Text als Beschreibung für einen beobachteten Spielzug generiert:

(9) In der linken Hälfte läuft Jones mit dem Ball auf das Tor zu. Meyer verfolgt ihn und versucht ihn anzugreifen. Aber Meyer ist zu langsam.

Da für den Hörer das Spielgeschehen visuell nicht präsent ist, kann er anhand des Berichtes nur eine ungefähre räumliche Vorstellung entwickeln. Entscheidend ist nun, daß das System selbst in der Lage sein muß, sich in die Rolle des Hörers zu versetzen und dessen mögliche Imagination bei der weiteren Sprachproduktion zu berücksichtigen.

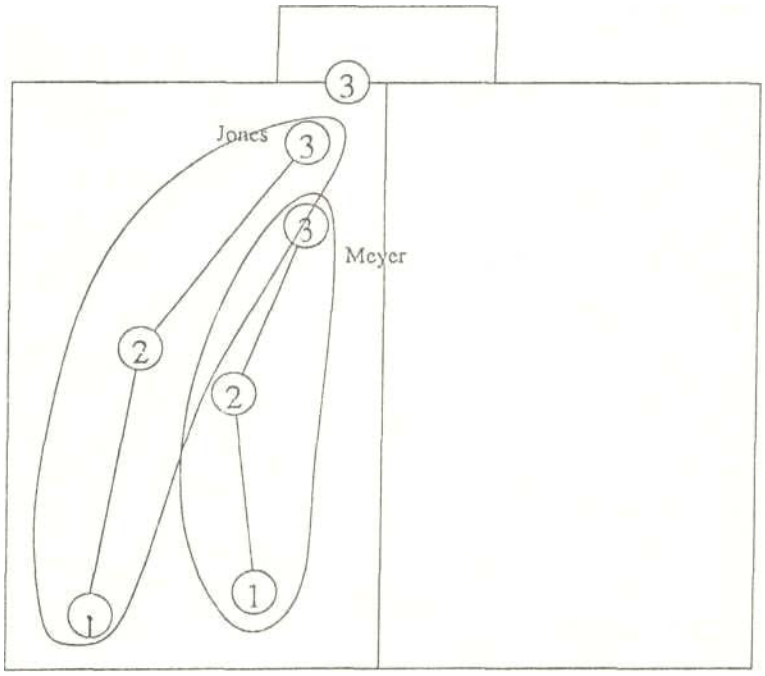


Abb. 6: Eine durch antizipiertes Verstehen abgeleitete Imagination

Abb. 6 zeigt eine mögliche graphische Repräsentation der durch (9) erzeugten Imagination. Dabei ist der gezeigte Trajektorienverlauf als prototypisch zu deuten und die elliptischen Flächen weisen auf den räumlichen Interpretationsspielraum hin. Falls das System nun plant, als nächsten Satz (10) zu äußern, so muß es im Sinne seiner

(10) Jetzt ist nur noch der Torhüter zwischen ihm und dem Tor.

kooperativen Kommunikationsabsicht prüfen, ob das Pronomen 'ihm' für den Hörer eindeutig auflösbar ist.

Als Bezugsobjekte für das Pronomen kommen aus dem vorangegangenen Text nur 'Jones' und 'Meyer' in Frage. Da 'Meyer' zuletzt erwähnt wurde, wird diese Referenzauflösung rein textuell für den Hörer näher liegen. Im Sinne einer Antizipationsrückkopplungsschleife (vgl. [Jameson/Wahlster 1982]) könnte das System aber durch den Zugriff auf die Imaginationskomponente des Benutzermodells feststellen, daß diese Auflösung der Anapher inkonsistent zu der angenommenen räumlichen Vorstellung des Hörers sein wird. Somit bleibt als eindeutige und mit dem Benutzermodell verträgliche Dereferenzierungsmöglichkeit nur 'Jones'. Erst nach dieser Antizipation eines erfolgreichen Verstehensprozesses, sollte der geplante Satz wirklich geäußert werden. Anderenfalls müßte das System auf die Verkürzung seiner Äußerung durch Pronominalisierung verzichten und z.B. einen Eigennamen benutzen.

Danksagung

Ohne die erfolgreiche Kooperation mit meinem Kollegen H.-H. Nagel vom IITB der FhG in Karlsruhe, die schon vor rd. 10 Jahren an der Universität Hamburg begann, wäre die hier beschriebene Kopplung von zwei KI-Systemen sicherlich nicht möglich gewesen. Seinen Mitarbeitern G. Zimmermann und C.K. Sung gilt unser Dank für die gute Zusammenarbeit in den letzten drei Jahren. Als Mitarbeiter im VITRA-Projekt lieferten G. Retz-Schmidt und J. Schirra grundlegende Beiträge zur Konzeption und Durchführung des Projektes. Wichtige Ideen und die gesamte Implementation gehen auf E. André, G. Bosch, G. Herzog, T. Rist und I. Wellner zurück, die als studentische Hilfskräfte und Diplomanden entscheidend zum Projektfortschritt beitrugen.

André, E.; Bosch, G.; Herzog, G.; Rist, T. (1986a): Characterizing Trajectories of Moving Objects Using Natural Language Path Descriptions. Memo Nr. 5, SFB 314, Fachbereich Informatik, Universität des Saarlandes. Auch in: Proc. of the 7th ECAI, July 1986, Brighton, England, Vol. 2, 1-8.

André, E.; Bosch, G.; Herzog, G.; Rist, T. (1986b): Coping with the Intrinsic and Deictic Uses of Spatial Prepositions. Bericht Nr. 9, SFB 314, Fachbereich Informatik, Universität des Saarlandes. Auch in: Proc. of AIMSA 1986, Varna, Bulgarien.

André, E.; Rist, T.; Herzog, G. (1987): Generierung natürlichsprachlicher Äußerungen zur simultanen Beschreibung von zeitveränderlichen Systemen. Bericht Nr. 18, SFB 314, auch in: Morik, K. (ed.): GWAI-87, 11th German Workshop on Artificial Intelligence. Berlin/Heidelberg/New York/Tokyo: Springer.

Herzog, G. (1986): Ein Werkzeug zur Visualisierung und Generierung von geometrischen Bildfolgenbeschreibungen. Memo Nr. 12, SFB 314, Fachbereich Informatik, Universität des Saarlandes.

Jameson, A.; Wahlster, W. (1982): User Modelling in Anaphora Generation: Ellipsis and Definite Descriptions. In: Proc. of 1st ECAI, Orsay 1982, 222-227.

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

Nagel, H.-H. (1985): Wissensgestützte Ansätze beim maschinellen Sehen: Helfen Sie in der Praxis? in: Brauer, W., Radig, B. (eds.): Wissensbasierte Systeme. GI-Kongreß 1985. Informatik-Fachberichte, Berlin/Heidelberg/New York/Tokyo: Springer.

Neumann, B.; Novak, H.-J. (1986): NAOS: Ein System zur natürlichsprachlichen Beschreibung zeitveränderlicher Szenen. In: Informatik - Forschung und Entwicklung (1986) 1, 83-92.

Retz-Schmidt, G. (1986a): Script-Based Generation and Evaluation of Expectations in Traffic Scenes. In: H. Stoyan (ed.) (1986): GWAI-85, 9. Fachtagung über Künstliche Intelligenz, Informatik-Fachberichte, Berlin/Heidelberg/New York/Tokyo: Springer.

Retz-Schmidt, G. (1986b): Deictic and Intrinsic Uses of Spatial Prepositions: A Multidisciplinary Comparison. Memo Nr. 13, SFB 314, Fachbereich Informatik, Universität des Saarlandes. Auch in: Proc of the Workshop on Spatial Reasoning and Multi-Sensor Fusion, St. Charles, Illinois, Oct. 1987, Morgan Kaufmann.

Rist, T.; Herzog, G.; André, F. (1987): Ereignismodellierung zur inkrementellen High-level Bildfolgenanalyse. Bericht Nr. 19, SFB 314, Fachbereich Informatik, Universität des Saarlandes, auch in: Buchberger, E.; Retti, J. (eds.): ÖGAI-87. 3. Österreichische AI-Tagung. Informatik-Fachberichte, Berlin/Heidelberg/New York/Tokyo: Springer.

Wahlster, W. (1982): Natürlichsprachliche Systeme. Eine Einführung in die sprachorientierte KI-Forschung. In: Bibel, W.; Siekmann, J.H. (eds.): Künstliche Intelligenz. Frühjahrsschule der GI. Informatik-Fachberichte. Berlin/Heidelberg/New York/Tokyo: Springer.

Wahlster, W. (1984): Cooperative Access Systems. In: Future Generations Computer Systems, Vol. 1, No. 2, 103-111.

Wahlster, W.; Jameson, A.; Hoeppner, W. (1978): Glancing, Referring and Explaining in the Dialogue System HAM-RPM. In: American Journal of Computational Linguistics, 53-67.

Wahlster, W.; Marburger, H.; Jameson, A.; Busemann, S. (1983): Over-answering Yes-No Questions: Extended Responses in a NL Interface to a Vision System. In: Proc. of the 8th IJCAI, Karlsruhe.

Wahlster, W.; Kobsa, A. (1986): Dialog-based User Models. In: Proceedings of the IEEE 74(7), 948-960 (Special Issue on Natural Language Processing).

Zimmermann, G.; Sung, C.K.; Bosch, G.; Schirra, J.R.J. (1987): From Image Sequences to Natural Language: Descriptions of Moving Objects. Gemeinsamer Zwischenbericht der FhG-IITB und FB Informatik der Universität des Saarlandes, Bericht Nr. 17, SFB 314.