MITTEILUNG 123

SZENENBESCHREIBUNG UND IMAGINATION IN NAOS

Hans-Joachim Novak und Bernd Neumann

IFI-HH-M-123/84

April 1984

ZUSAMMENFASSUNG

NAOS[1] ist ein System zur NAtürlichsprachlichen Beschreibung von Objektbewegungen in einer Straßenverkehrsszene. Die Architektur dieses Systems wird dargestellt. Dabei wird im Besonderen die Repräsentation der Szene, welche die Grundlage aller weiterverarbeitenden Prozesse ist, die Repräsentation der zu erkennenden Ereignisse, sowie der Prozeß der Ereigniserkennung ausführlich beschrieben. Im letzten Kapitel wird dargestellt, wie eine Antizipation des Hörerverständnisses die Planung einzelner Aussagen unterstützen kann, um so zu einer zusammenhängenden Beschreibung der Bildfolge zu gelangen.

[1]) Die Arbeiten an NAOS werden teilweise von der DFG unterstützt.

SZENENBESCHREIBUNG UND IMAGINATION IN NAOS

H.-J. Novak und B. Neumann

Fachbereich Informatik, Universität Hamburg
Schlüterstraße 70, D-2000 Hamburg 13

1. Übersicht

Dieser Beitrag handelt von einem System zur natürlichsprachlichen Beschreibung von Realwelt-Bildfolgen. Der Name NAOS - NAtürlichsprachliche Beschreibung von Objektbewegungen in einer Straßenverkehrsszene - verweist auf die Domäne und das besondere Interesse an zeitabhängigen Vorgängen. Es geht also darum, eine durch eine Kamera aufgenommene Bildfolge von einer Verkehrsszene mit Methoden des Bildverstehens auszuwerten und in eine sprachliche Beschreibung zu überführen.

Einer der Schwerpunkte der Untersuchungen ist die "Höhere Bilddeutung", also die Phase der Bildfolgenanalyse, wo zeitübergreifende Vorgänge, spezielle Situationen, besondere Beziehungen zwischen Objekten, kurz: Zusammenhänge erkannt werden, die über Objekterkennung hinausgehen. Welche höheren Zusammenhänge in einem bildverstehenden System im einzelnen erkannt werden sollten, ist bisher wenig geklärt, siehe hierzu die Diskussion in NEUMANN 82. NAOS beschränkt sich auf eine besondere Klasse von höheren Konzepten, auf Ereignisse. Ein Ereignis ist definiert als eine Teilmenge einer vierdimensionalen (weil zeitabhängigen) Szene, die mit einem Bewegungsverb beschrieben werden kann, z.B. dem Verb 'abbiegen'. Durch Ereigniserkennung wird also gleichzeitig eine Verbalisierung vorbereitet. Es zeigt sich, daß Ereigniskomponenten mit den Tiefenkasus des zugehörigen Verbs in einfacher Beziehung stehen, so daß ein vollständiger Kasusrahmen erzeugt werden kann. Aus einem Kasusrahmen wird dann mithilfe des Generierungsmoduls von BUSEMANN 83 eine sprachliche Äußerung erzeugt. Der Prozeß der höheren Bilddeutung bis hin zur Verbalisierung wird in Abschnitt 2

ausführlich beschrieben.

Abschnitt 3 behandelt Aspekte der Sprechplanung, also eine Komponente von NAOS, die aus einzelnen Äußerungen eine kohärente Beschreibung zusammenstellen soll. Hier spielen die Vorstellungen eine entscheidende Rolle, die sich ein Hörer zu dem Gesagten machen kann. Ist die Vorstellung unvollständig oder weicht sie zu stark von der tatsächlichen Szene ab, so ist dies Anlaß für ergänzende oder präzisierende Äußerungen. In NAOS wird deshalb eine Simulation der hörerseitigen Imagination angestrebt.

2. Architektur von NAOS

Im Folgenden werden die Wissensquellen und Repräsentationsebenen sowie die darauf aufbauenden Prozesse beschrieben, die in NAOS zur Erzeugung einzelner Aussagen verwandt werden. Zunächst werden die verschiedenen Wissensquellen dargestellt, anschließend (in 2.3) wird der Prozeß der Ereigniserkennung anhand eines Beispiels erläutert. Im letzten Abschnitt wird gezeigt, wie ein erkanntes Ereignis verbalisiert wird.

2.1 Geometrische Szenenbeschreibung (GSB)

Um eine Szene, einen Ausschnitt der realen Welt, der Rechnerverarbeitung zugänglich zu machen, wird mit einer Schwarz/Weiß-Fernsehkamera eine Folge von Einzelbildern aufgenommen. In NAOS umfassen die betrachteten Bildfolgen 50-500 Einzelbilder, was bei einer Aufnahmerate der Kamera von 25 Bildern pro Sekunde einer Länge von bis zu 20 Sekunden entspricht. Die analogen Einzelbilder werden anschließend für die Weiterverarbeitung digitalisiert.

Läßt man die Zeitdimension außer acht, so ist eine Bildfolge eine 2D-Projektion der ursprünglichen Szene. Aufgabe der Szenenanalyse ist es, Objekte in der Bildfolge zu erkennen, zu klassifizieren und die bei der Projektion verlorengegangene Tiefeninformation zu rekonstruieren. Im Prinzip besteht eine Bildfolge aus einem unbewegten Anteil, z.B. Straßen, Häuser, etc. und bewegten Objekten. Die unbewegten Objekte (stationärer Hintergrund) werden mithilfe eines detaillierten Straßenmodells erkannt und klassifiziert. Die Erkennung der Form und Trajektorie der bewegten Objekte erfordert spezielle Prozesse, wie sie z.B. in DRESCHLER und NAGEL 82 dargestellt sind. Eine Klassifikation der bewegten Objekte wie z.B. Autos, Fußgänger oder Radfahrer kann derzeit von unserem Szenenanalysesystem nicht geleistet werden und erfolgt daher interaktiv. Eine genauere Darstellung der Prozesse, die notwendig sind, um aus der Bildfolge eine 3D-Szenenbeschreibung (im Folgenden Geometrische Szenenbeschreibung (GSB) genannt) zu erstellen, ist in

NEUMANN 82 enthalten.

Die GSB hat die Aufgabe, alle Informationen aus der ursprünglichen Szene für die verbale Beschreibung zur Verfügung zu stellen. Sie besteht aus den zwei Teilen:

a) Stationärer Hintergrund (instantiiertes Straßenmodell)
b) bewegte Objekte.

Die GSB ist eine objektorientierte Repräsentation, in der die für ein Objekt relevante Information mit diesem Objekt assoziiert ist. Konkret enthält die GSB:

je Einzelbild der Bildfolge
- Zeitpunkt
- Liste aller Objekte
- Betrachterposition und -orientierung
- Beleuchtung

je Objekt
- Identität (über die Bildfolge)
- 3D-Form und -Aussehen
- 3D-Position und -Orientierung
- Klassenzugehörigkeit
- Name
- Farbe
- funktionale Merkmale

Die Bedeutung der einzelnen Attribute ist wie folgt:

Der Zeitpunkt beschreibt, wann das Bild aufgenommen wurde und wird durch die Bildnummer angegeben.

Die Liste aller Objekte enthält die in dem entsprechenden Bild sichtbaren Objekte.

Die Betrachterposition und -orientierung entspricht dem Kamerastandpunkt und wird mit x, y und z Koordinaten für den

Standpunkt und drei Winkeln für die Auslenkung um jede der drei Achsen angegeben.

Die Beleuchtung kann von unserem System noch nicht bestimmt werden, ist aber aus Gründen der Vollständigkeit mit aufgeführt. Vollständigkeit wird in dem Sinne erreicht, daß die Bildfolge mithilfe der Beleuchtungsinformation aus der GSB rekonstruiert werden kann.

Die Identität eines Objektes innerhalb der Bildfolge wird durch einen eindeutigen Bezeichner, z.B. AUTO1 dargestellt.

3D-Form und -Aussehen eines Objektes werden durch Angaben über die Oberflächenform und Oberflächenfarbe repräsentiert. Oberflächen können aus ebenen Polygonen oder Kegelmantelflächen zusammengesetzt sein, wobei jedes Oberflächenelement weiter in Farbflächen gegliedert sein kann. Die Farbflächen geben die Rot-, Grün- und Blaureflektivität an.

Die 3D-Position und -Orientierung bezeichnet die Lage eines Objekts in dem entsprechenden Bild. Sie wird durch die x, y und z Koordinaten des Schwerpunktes des Objektes sowie durch einen Orientierungsvektor angegeben. Der Orientierungsvektor bezeichnet die Richtung, in welche die Vorderseite des Objektes weist. Da die Lage eines Objektes eine zeitabhängige Größe ist, wird zusätzlich noch ein Zeitintervall angegeben, das die Dauer der Gültigkeit dieser Lage angibt.

Die Klassenzugehörigkeit gibt an, zu welcher Klasse von Objekten ein von der Szenenanalyse erkanntes Objekt gehört.

Ein eventueller Name eines Objektes, z.B. "Alte-Post" wird ebenso aufgeführt wie die Farbe des Objektes.

Die Vorderseite eines Objektes ist z.B. ein funktionales Merkmal, daß auch für andere Prozesse von Bedeutung ist, z.B. für die Auswertung räumlicher Präpositionen.

Im Folgenden ist ein Ausschnitt aus der GSB dargestellt. Der LAGE-Eintrag hat die allgemeine Form:

(LAGE <interner Objektbezeichner> <Positionstripel>
<Orientierungsvektor> <Zeit1> <Zeit2>).

Das angegebene Zeitintervall ist dabei als rechtsseitig offen zu verstehen, also [Zeit1, Zeit2). Die anderen Einträge sind selbsterklärend.

```
(KLASSE VW1 VW)
(FARBE VW1 GELB)
(KLASSE LKW1 LKW)
(KLASSE HAUS1 HAUS)
(NAME HAUS1 "Fachbereich Informatik")
(LAGE HAUS1 (100 -60 70) (0 1 0) 1 40)
(LAGE VW1 (-100 70 8) (4 1 0) 1 2)
(LAGE VW1 (-80 75 8) (4 1 0) 2 3)
                .
                .
                .
(LAGE VW1 (875 50 8) (1 0 0) 31 32)
(LAGE VW1 (880 50 8) (1 0 0) 32 40)
(LAGE LKW1 (50 50 15) (1 0 0) 1 2)
(LAGE LKW1 (60 50 15) (1 0 0) 2 3)
                .
                .
                .
```

Die GSB ist die grundlegende Repräsentation in NAOS, auf der alle weiteren Prozesse aufbauen, die "höhere Konzepte" berechnen. In NAOS sind diese "höheren Konzepte" Änderungen in der Szene, die mit einem Fortbewegungsverb ausgedrückt werden können. Insofern sind die "höheren Konzepte" auf Verbalisierung angelegt. Um solche verbalisierbaren Teile der Szene zu erkennen, greift der Erkennungsprozeß auf Ereignismodelle zurück. In den Ereignismodellen ist festgelegt, welche Daten in der GSB vorhanden sein müssen, um ein bestimmtes Fortbewegungsverb zu benutzen. Im

folgenden Abschnitt wird die Repräsentation der Ereignismodelle dargestellt.

2.2 Ereignismodelle

Im Kontext der von uns betrachteten Verkehrsszene sind Ereignisse konkrete Vorkommnisse, z.B., daß ein Auto anhält, ein Fußgänger die Straße überquert oder ein PKW um die Ecke rast. Das zugrundeliegende Konzept eines Ereignisses wird in Form eines Ereignismodells repräsentiert. Ereignismodelle sind mit den Verben assoziiert, die zu ihrer Beschreibung benutzt werden können; sie sind also "verbzentriert" repräsentiert.

Als Beispiel ist hier das Ereignismodell 'überholen' aufgeführt:

```
(ÜBERHOLEN OBJ1 OBJ2 T1 T2)
     (BEWEGEN OBJ1 T1 T2)
     (BEWEGEN OBJ2 T1 T2)
     (HINTER OBJ1 OBJ2 T1 T3)
     (HINTER OBJ2 OBJ1 T4 T2)
     (NEBEN OBJ1 OBJ2 T3 T4)
     (NÄHERN OBJ1 OBJ2 T1 T3)
     (ENTFERNEN OBJ1 OBJ2 T4 T2)
```

Informell ist das obige Ereignismodell wie folgt zu lesen. Wenn OBJ1 OBJ2 in einem Zeitintervall von T1 bis T2 überholt, muß gelten: Beide Objekte bewegen sich in dem Zeitintervall. In einem Teilintervall von T3 bis T4, das in dem Intervall (T1 T2) liegt, sind beide Objekte nebeneinander. Davor ist OBJ1 hinter OBJ2 und nähert sich diesem, danach ist OBJ2 hinter OBJ1, und OBJ1 entfernt sich von OBJ2.

Aus obigem Beispiel wird deutlich, daß Ereignismodelle relational notiert sind. Die einzelnen Relationen (im Folgenden auch Propositionen genannt) bestehen aus dem Relationenbezeichner z.B. ÜBERHOLEN, im Allgemeinen mehreren variablen Platzhaltern (OBJ1 OBJ2) sowie Zeitvariablen, die angeben, in welchem Zeitintervall die

Proposition gültig ist z.B. von T1 bis T2.

Wir unterscheiden drei Typen von Propositionen: primitive, zusammengesetzte und spezielle. Primitive Propositionen werden direkt anhand der GSB berechnet, sind also prozedural definiert. Ein Beispiel ist die Proposition BEWEGEN. Zusammengesetzte Propositionen bestehen wie in obigem Beispiel aus einer Menge einzelner Propositionen, die selbst wiederum zusammengesetzt sein können. Spezielle Propositionen schließlich dienen zum Evaluieren von Beziehungen wie "zeitlich innerhalb", die nicht unmittelbar auf die GSB zugreifen.

Propositionen werden ausgewertet, indem Variablenbelegungen für die Platzhalter generiert werden, so daß die Proposition wahr ist. Damit eine zusammengesetzte Proposition wie 'ÜBERHOLEN' wahr ist, müssen alle Teilpropositionen, aus denen sie besteht wahr sein.

Ist ein Ereignismodell wie (BEWEGEN OBJ1 T1 T2) anhand der GSB verifiziert worden, so liefert der Erkennungsprozeß ein instantiiertes Ereignismodell als Ergebnis. Eine solche Instanz kann z.B. (BEWEGEN AUTO1 12 37) sein. Wie der Erkennungsprozeß Ereignismodelle instantiiert, wird im nächsten Abschnitt beschrieben.

2.3 Ereigniserkennung

Prinzipiell besteht Ereigniserkennung aus dem Vergleich zwischen einem Ereignismodell und den Daten der GSB. Da beide Repräsentationen relational notiert sind, kann man unter Ereigniserkennung einen Vergleich zwischen Relationalstrukturen verstehen (BARROW and POPPLESTONE 71). Man kann den Ereigniserkennungsprozeß allerdings auch als Beweisfindung auffassen, die z.B. in PROLOG durch Eingabe einer entsprechenden Anfrage realisiert werden könnte.

Ist für ein Ereignismodell keine Instanz in der GSB vorhanden, so muß sie berechnet werden und wird dann als weiteres Datum in die GSB

eingetragen. Im allgemeinen Fall muß zur Ereigniserkennung (Instantiierung eines Modells) eine Liste von Propositionen ausgewertet werden. Diese Liste wird von der Auswertungsprozedur rekursiv abgearbeitet. Die rekursive Auswertung ergibt eine klare Struktur, die bei der Auswertung zusammengesetzter Prädikate besonders deutlich wird, da in diesem Fall die Auswertungsprozedur auf die Liste der Teilpropositionen angewandt wird.

Die Reihenfolge, in der eine Liste von Propositionen ausgewertet wird, bestimmt der Verzweigungsfaktor der einzelnen Propositionen. In NAOS unterscheiden wir zwischen einem intrinsischen und einem effektiven Verzweigungsfaktor. Der intrinsische Verzweigungsfaktor einer Proposition entspricht der Wahrscheinlichkeit mit der diese Proposition bei einer beliebigen Belegung der Variablen wahr ist. Dieser Wert ist domänenabhängig und wird durch Erfahrung bestimmt. Er ist mit dem Relationenbezeichner assoziiert. Der effektive Verzweigungsfaktor berechnet sich aus dem intrinsischen, durch Multiplikation des intrinsischen Verzweigungsfaktors mit der Anzahl der Möglichkeiten, die Variablen der Proposition zu belegen. Dies geschieht zur Laufzeit des Systems. Eine geschickte Wahl der Reihenfolge ist die, zuerst die Proposition auszuwerten, die mit der geringsten Wahrscheinlichkeit wahr ist. Ist diese nicht erfüllbar, so spart man sich die Auswertung der restlichen Propositionen. Die Berechnung der effektiven Verzweigungsfaktoren wird nach jeder Auswertung für die verbleibenden Propositionen wiederholt, um die erfolgten Instantiierungen richtig zu berücksichtigen.

Die Auswertung folgt einer 'Generiere-und-Suche' Strategie, wobei für die Suche eine Backtracking-Kontrollstruktur benutzt wird. Diese Strategie bedeutet im Prinzip, daß in einem ersten Zyklus alle Instanzen einer Proposition generiert und in die GSB eingetragen werden. Dann wird im Such-Zyklus für die erste Instanz durch rekursiven Aufruf geprüft, ob die restlichen Propositionen ebenfalls instantiiert werden können. Zusammengesetzte Propositionen werden zuerst in ihre Teilpropositionen expandiert und dann ausgewertet. Alle zusammengesetzten Propositionen werden also letztlich auf primitive Propositionen zurückgeführt.

Da für die Zeitvariablen einer Proposition mehrere Instanzen aufgrund verschiedener Zeitintervalle vorhanden sein können z.B. (BEWEGEN LKW1 1 23) und (BEWEGEN LKW1 29 40), wird durch Backtracking dafür gesorgt, daß jedes der Intervalle auf Zeitkompatibilität mit den restlichen Propositionen geprüft wird.

Die Auswertung ist erfolgreich, wenn für eine konkrete Variablenbelegung alle Teilpropositionen wahr sind. Kann eine Proposition nicht erfüllt werden oder ist sie nicht zeitkompatibel mit den anderen, so liefert die Auswertung einen Mißerfolg.

Die Zeit spielt bei der Auswertung von Ereignismodellen eine besondere Rolle. Stellt man an die GSB die Anfrage

(BEWEGEN ?OBJ 20 30),

und enthält die GSB einen Eintrag

(BEWEGEN AUTO1 20 30),

so ist der Vergleich erfolgreich und die Variable ?OBJ instantiiert zu AUTO1. Stellt man jedoch die Anfrage

(BEWEGEN ?OBJ 21 29),

so ist bei üblicher Behandlung eines Relationalvergleichs kein erfolgreicher Vergleich möglich. Die Ursache liegt darin, daß die implizite Bedeutung der Zeitkomponenten als untere und obere Schranke eines durativen Vorgangs unberücksichtigt bleibt. Der obige Vergleich des Modelltupels mit dem Datentupel sollte auch dann erfolgreich sein, wenn das Zeitintervall des Modelltupels in dem des Datentupels enthalten ist. Das impliziert, daß Zeitvariable bei einem Vergleich nicht instantiiert, sondern bezüglich ihres Wertebereiches eingeschränkt werden müssen. So sollte der Vergleich von (BEWEGEN ?OBJ ?T1 ?T2) mit (BEWEGEN AUTO1 20 30) die folgenden Ungleichungen liefern:

$$20 \leqslant T1 \leqslant T2 \leqslant 30$$

Durative Ereignisse sind solche, die auch für alle Teilintervalle ihre Gültigkeit behalten. Für sie können die jeweiligen Beschränkungen der Zeitvariablen einfach repräsentiert werden, indem die Grenzwerte des Maximalintervalls festgehalten werden (wie in obiger Ungleichung).

Zusammengesetzte Ereignisse wie 'überholen' verlieren in Teilintervallen ihre Gültigkeit. Jedoch genügen die Intervallgrenzen eines jeden Ereignisses, unabhängig von der Anzahl der Teilereignisse, einem Ungleichungssystem mit folgendem Aufbau:

$$TB\text{-}MIN \leqslant TB \leqslant TB\text{-}MAX$$
$$TE\text{-}MIN \leqslant TE \leqslant TE\text{-}MIN$$
$$TB < TE$$

Die Grenzwerte von TB (Beginn des Ereignisses) und TE (Ende des Ereignisses) sind jeweils Konstante. Wir verwenden hier in der Regel die Notation:

(<Ereignis> ... (<TB-MIN> <TB-MAX>) (<TE-MIN> <TE-MAX>))

Ein duratives Ereignis ist ein Spezialfall, für den TB-MIN = TE-MIN und TB-MAX = TE-MAX gilt. Hierfür verwenden wir die vereinfachte Notation:

(<duratives Ereignis> ... <TB-MIN> <TE-MAX>)

In NAOS ist der Relationalvergleich derart erweitert worden, daß anstelle einer Instantiierung bei einem Zeitintervall eine Modifikation des zugehörigen Beschränkungssystems stattfindet. Sobald die Ungleichungen keine Lösung mehr zulassen, sind die beteiligten Propositionen zeitinkompatibel und es wird Backtracking eingeleitet.

Das Ergebnis der Ereigniserkennung ist die Instanz eines Ereignismodells, z.B. (ÜBERHOLEN AUTO1 LKW1 (10 12) (34 37)). Um von dieser Instanz zu einer natürlichsprachlichen Äußerung zu gelangen, bedarf es einer weiteren Wissensquelle, der Kasussemantik.

Diese wird im Folgenden dargestellt.

2.4 Kasussemantik

Auf Grund der verbzentrierten Darstellung der Ereignismodelle ist die Abbildung einer Instanz in eine natürlichsprachliche Äußerung in einfacher Weise mithilfe des Kasusrahmens des Verbs möglich. Das Verb wird in der Instanz bereits explizit genannt. Die anderen Komponenten einer Instanz sind systeminterne Bezeichner, z.B. AUTO1 und LKW1, die bei der Verbalisierung sprachlich referenziert werden müssen. Zu diesem Zweck wird für jedes Verb eine Liste seiner obligaten und fakultativen Tiefenkasus (FILLMORE 68) geführt sowie Vorschriften, wie diese Tiefenkasus zu verbalisieren sind. 'Überholen' hat z.B. als obligate Tiefenkasus Agent und Objektiv sowie einen fakultativen Lokativ. Das instantiierte Ereignismodell (ÜBERHOLEN PKW1 LKW1 (10 12) (34 37)) enthält als Agent und Objektiv die internen Objektbezeichner PKW1 bzw. LKW1, die bei der Verbalisierung des Ereignisses sprachlich referenziert werden müssen. Um einen Lokativ auszudrücken, muß der entsprechende Ort bekannt sein. In der Kasussemantik wird festgelegt, daß der Lokativ eines Überholen-Ereignisses der Ort von OBJ1 im Intervall (T1 T2) ist. Zusätzlich enthält der Lokativ eine Vorschrift, die zu diesem - zunächst in Koordinatenform vorliegenden - Ort eine Referenzierung berechnet (REF-LOC, s.u.).

Schließlich enthält die Kasussemantik noch Angaben zur Modalität. Das Tempus der Äußerung wird aus der Lage des Ereignisintervalls innerhalb der Gesamtlänge der Szenenfolge bestimmt (MARBURGER et al. 81). Als genus verbi ist im Moment nur aktiv vorgesehen.

Das Folgende ist die Struktur der Kasussemantik für 'überholen':

```
(ÜBERHOLEN OBJ1 OBJ2 T1 T2)
      Verb:    überholen
     Agent:    (REF OBJ1)
  Objektiv:    (REF OBJ2)
   Lokativ:    (REF-LOC OBJ1 T1 T2)
```

Tempus: (INNERHALB T1 T2 TBEG TEND)
Genus: Aktiv

Die Angabe (REF OBJ1) ist eine Anweisung, die zum internen Bezeichner OBJ1 eine Referenzierung berechnet. Dazu kann seine Klassenzugehörigkeit, Farbe etc. ausgenutzt werden.

(REF-LOC ...) leistet entsprechendes für die Koordinatenwerte der Objektpositionen.

(INNERHALB ...) bestimmt aus der Lage des Ereignisintervalls innerhalb der Gesamtlänge der Szene die Verbzeit der Äußerung.

Für obige ÜBERHOLEN-Instanz ergibt sich der folgende Kasusrahmen:

Verb: überholen
Agent: "e- gelb- VW"
Objektiv: "e- LKW"
Lokativ: "vor d- Fachbereich Informatik"
Tempus: Präteritum
Genus: Aktiv

Ein Generierungssystem wie z.B. das System SUTRA (BUSEMANN 83) kann aus diesem Kasusrahmen unter Zugriff auf ein entsprechendes Lexikon einen korrekt flektierten Oberflächensatz erzeugen:

"Ein gelber VW überholte einen LKW vor dem Fachbereich Informatik."

3. Sprechplanung

Das NAOS-System hat zum Ziel, eine Bildfolge zu beschreiben. Grundlage dafür ist die im vorigen Kapitel beschriebene Ereigniserkennung, die es ermöglicht, einzelne Aussagen zu generieren.

Es ist nun die Aufgabe der Sprechplanung, aus mehreren Aussagen eine zusammenhängende Beschreibung zu erstellen. Dazu muß z.B. entschieden werden, in welcher Reihenfolge Aussagen gemacht werden, welches der Verben, die für eine Beschreibung einer Bewegung benutzt werden können, tatsächlich ausgewählt wird, und welche fakultativen Tiefenkasus eines Verbs sprachlich realisiert werden sollen.

Ein und derselbe Vorgang kann sprachlich sehr verschieden ausgedrückt werden, wie aus den Beschreibungen eines Unfalls gegenüber der Polizei, dem gegenerischen Anwalt oder gegenüber einem Freund ersichtlich wird. Die Art der Beschreibung hängt sowohl von den Intentionen des Sprechers als auch von seinem Modell des Hörers ab. Die Erzeugung einer Beschreibung ist also in einen kommunikativen Kontext eingebettet, d.h. sie wird für einen potentiellen Hörer erstellt. In NAOS machen wir die folgenden grundlegenden Annahmen über den Hörer:

a) Er/Sie kennt den statischen Hintergrund der Szene.
b) Er/Sie kann die Szene nicht sehen.
c) Er/Sie hat keine speziellen Interessen geäußert, außer:
   "Beschreibe die Szene!"

Eine Beschreibung kann das Ergebnis so verschiedener Sprechakte wie z.B. 'informieren', 'überzeugen', 'bitten' oder 'versprechen' sein. Von NAOS wird nur der Sprechakt 'informieren' erzeugt.

Im Folgenden wird der Sprechakt 'informieren' definiert. Danach werden Konsequenzen dieser Definition bezüglich der Konstruktion eines Hörermodells dargestellt.

3.1 Hörermodell

Jemanden informieren heißt: jemandem etwas mitteilen, was wahr und neu ist. Geht man von dieser Definition des Sprechaktes "informieren" aus, so erfordert Sprechaktplanung offenbar eine Bewertung von potentiellen Äußerungen bezüglich dieser beiden Kriterien.

Betrachtet man die Definition von "wahren Äußerungen" näher, so ergeben sich gewisse Voraussetzungen für die Kommunikationspartner. Unsere Definition baut auf der Situationssemantik von BARWISE und PERRY 83 auf. Dort versteht man unter der Bedeutung einer Äußerung eine Beziehung, die zwischen Äußerung und beschriebener Situation besteht - in unserem Fall als algorithmischer Zusammenhang zwischen Szenen und ihren verbalen Beschreibungen realisiert. Die Interpretation einer Äußerung (durch einen Hörer) ist in der Regel eine Menge möglicher Situationen, die mit der Äußerung durch eine Bedeutungsbeziehung verbunden sind. Wir definieren nun eine Äußerung als wahr, wenn diese Menge von Situationen die tatsächlich vorliegende Situation einschließt. Äußerungen, die zu "Mißverständnissen" führen, sind in diesem Sinne also nicht wahr.

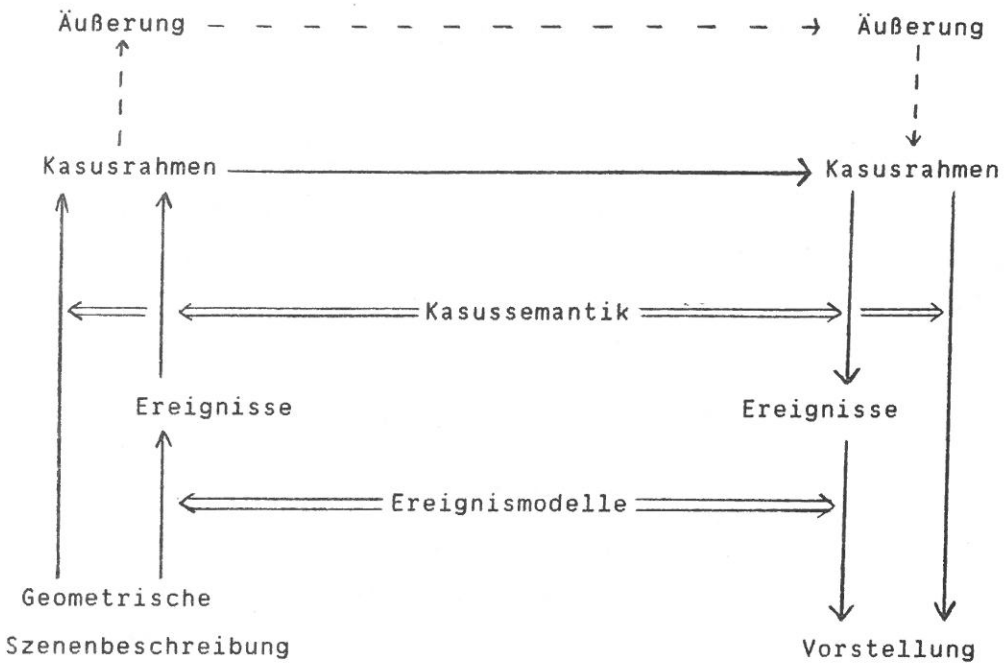
Für unser System bedeutet die Forderung nach Wahrheit zweierlei. Zum einen muß das Verfahren zur Szenenbeschreibung auf den gedachten Hörer abgestimmt sein - es muß seine Bedeutungsbeziehungen realisieren. Dies entspricht der naheliegenden Kommunikationsregel, daß man Äußerungen auf das Sprachverständnis des Hörers abstellen sollte.

Der zweite Gesichtspunkt ist eine Konsequenz des ersten. Wenn man davon ausgeht, daß man dasselbe Bedeutungssystem wie der Hörer hat, so kann man simulieren, wie der Hörer eine Äußerung interpretieren wird. Man kann also die Situationen generieren, die allein durch die Äußerung, ohne Kenntnis der tatsächlichen Situation, impliziert werden. Bei einer Szenenbeschreibung sind diese Situationen gleichbedeutend mit den "Vorstellungen", die sich der Hörer von der (ihm unbekannten) Szene machen kann.

Eine Äußerung muß für den Hörer neu sein, wenn sie informieren soll. Hinterfragt man den Begriff "neu", so erkennt man, daß das Vermeiden von Wort- oder Satzwiederholungen offenbar nicht gemeint ist - die Interpretation einer Äußerung muß Neuigkeitswert haben, nicht die Wortwahl. Oben wurde dargelegt, daß unter der Interpretation einer Äußerung die Menge der dadurch beschriebenen Situationen verstanden werden kann. Wir definieren entsprechend eine Äußerung als "neu", wenn sie die durch vorherige Äußerungen umrissene Menge von Situationen weiter einschränkt. Dies ist mengentheoretisch eine Durchschnittsbildung. Man kann auch auf eine mengenbezogene Formulierung verzichten, indem man von der Menge aller möglichen Situationen als der unspezifizierten Situation und von eingeschränkten Mengen entsprechend als einer partiell spezifizierten Situation spricht. Eine Äußerung mit Neuigkeitswert muß also zusätzliche Spezifikationen für die beschriebene Situation bringen.

Hieraus ergibt sich eine konkrete Aufgabe für einen Sprecher, der informieren möchte. Er muß durch Hörersimulation überprüfen, ob die von ihm geplante Äußerung zusätzliche Spezifikationen im obigen Sinne mit sich bringt. Besser noch: er muß Äußerungen so auswählen, daß sie im obigen Sinne informieren.

Im Folgenden wird die für unser System beabsichtige Hörersimulation in etwas mehr Detail erläutert. Der grundsätzliche Ablauf ist im Diagramm illustriert.



Links sind Repräsentationsebenen und Prozesse des Sprechers gezeigt. Rechts wird ein Hörer simuliert, der die Prozesse - soweit es geht - invertiert, um aus einer Äußerung eine Szenenbeschreibung zu erschließen. Das Ergebnis ist eine "Vorstellung" von der Szene, die in der Regel weniger spezifiziert ist als die "Geometrische Szenenbeschreibung" links. Wir gehen davon aus, daß die sprachliche Tiefenstruktur - der "Kasusrahmen" - aus der Äußerung zurückgewonnen werden kann. Daher beginnt die Simulation mit dem Kasusrahmen, der der Generierung zugrunde liegt.

Interessanterweise ist der Prozeß der Generierung einer Vorstellung ebenso eine 1:n Abbildung wie der Prozeß der Beschreibung eines Ereignisses. Zu einer sprachlichen Beschreibung kann man mehrere Vorstellungen entwickeln ebenso kann man zu einer Szene mehrere Beschreibungen generieren. Nach (OLSON 72) beinhaltet jede sprachliche Beschreibung eines konkreten Ereignisses einen Genauigkeitsverlust, in unserem Fall z.B. bezüglich der genauen Position der beteiligten Objekte. Analog dazu liefert der inverse Prozeß eine Menge von Alternativen. Auf unsere Domäne bezogen bedeutet dies, daß z.B. die Präpositionalphrase 'vor dem Fachbereich

Informatik' eine Menge möglicher Orte beschreibt, an denen sich ein Objekt konkret befinden kann.

Das Diagramm, das den grundsätzlichen Ablauf der Hörersimulation illustriert, impliziert die Benutzung derselben semantischen Konzepte sowohl für die Generierung von Aussagen als auch für das Erzeugen von Vorstellungen. Im Folgenden betrachten wir beispielhaft, wie unsere Repräsentation der Präposition 'vor' für diese beiden Zwecke benutzt werden kann.

Die Präposition 'vor' ist in NAOS ebenso relational dargestellt wie die Ereignismodelle, konkret (VOR OBJ1 OBJ2 T1 T2). Diese Darstellung bedeutet, daß sich OBJ1 im Zeitintervall von T1 bis T2 vor OBJ2 befindet. Soll z.B. ein Lokativ mithilfe der Präposition 'vor' ausgedrückt werden, so wird dazu folgender Algorithmus angewandt:

> Suche ein Referenzobjekt, OBJ2, derart, daß die Position von OBJ1 Element der Menge der Positionen ist, die innerhalb eines 60 Grad-Kegels liegen, der von OBJ2 aus in Richtung seiner Vorderseite aufgespannt wird.

Offenbar können alle Positionen von OBJ1, die im Kegel von OBJ2 liegen, zu derselben Verbalisierung führen.

Betrachten wir nun die Umkehrung dieses Vorgangs, d.h. die Relation, z.B. (VOR AUTO1 HAUS1 3 20) ist gegeben. Bei der Generierung einer Vorstellung wird nun die oben angegebene Vorschrift in umgekehrter Richtung ausgewertet. Ausgehend von HAUS1 wird in Richtung seiner Vorderseite ein 60 Grad-Kegel aufgespannt, der die Menge der möglichen Positionen für AUTO1 enthält. Es wird also eine Menge von Alternativen generiert.

Die Invertierung anderer Relationen wie z.B. der Ereignismodelle 'überholen' oder 'fahren', führt zu Alternativen anderer Art. Eine Vorstellung zu der Äußerung 'Das gelbe Auto fährt auf der Schlüterstraße' beinhaltet z.B. Unsicherheiten bezüglich der Lage der Trajektorie, der Bewegungsrichtung und der Geschwindigkeit.

Ein Hörer schränkt die algorithmisch generierte Menge von Alternativen häufig durch die Bevorzugung bestimmter Alternativen oder durch die Annahme von Standardwerten ein. Beispiele derartiger Standards sind z.B. Geschwindigkeit, Ort, Zeitdauer. Standardwerte sind abhängig von der Domäne ('fahren auf der Autobahn' hat einen höheren Geschwindigkeitswert als 'fahren in der Stadt'), vom Objekt (Auto vs. Radfahrer) und ebenso von der Situation ('ein Auto auf dem Parkplatz' vs. 'ein Auto auf der Straße'). Verschiedene Verben haben dabei unterschiedliche Standardwerte, z.B. 'fahren' vs. 'rasen'. Zudem kann der Standardwert für die Geschwindigkeit von 'fahren' explizit außer Kraft gesetzt werden, z.B. durch die Angaben 'schnell' bzw. 'langsam'. 'Rasen' dagegen hat einen Standardwert, der nicht durch die genannten Adjektive modifiziert werden kann.

Die Annahme von Standards verringert die Anzahl möglicher Vorstellungen, läßt jedoch in der Regel einen Unsicherheitsbereich bestehen. Wir gehen nicht davon aus, daß mithilfe von Standards stets eine eindeutige Vorstellung erzeugt werden kann, wie etwa in ADORNI und DI MANZO 83.

Die Generierung von Vorstellungen dient der Antizipation des Hörerverständnisses und soll damit gleichzeitig die Sprechplanung unterstützen. Für die Sprechplanung ist ein Vergleich der tatsächlichen Szene mit der Vorstellung des Hörers notwendig. Herrscht z.B. beim Hörer eine Unsicherheit bezüglich der Bewegungsrichtung, so ist dies ein Hinweis darauf, eine Äußerung zu wählen, in der die Bewegungsrichtung explizit genannt wird.

Zur Unterstützung des Vergleichs von Szene und Vorstellung scheint uns eine explizite Repräsentation der Unsicherheiten der Vorstellung erforderlich. Die von uns vorgestellte Repräsentation muß in dieser Hinsicht noch ausgebaut werden. In diesem Zusammenhang sollte untersucht werden, inwieweit analogische Repräsentationen (FORBUS 83) für die genannten Aufgaben hilfreich sein können. Denkbar wäre z.B. eine analogische Repräsentation für die Unsicherheitsbereiche von Trajektorien.

LITERATURVERZEICHNIS

Adorni and Di Manzo 83
Top-Down Approach to Scene Interpretation
G. Adorni, M. Di Manzo
Proc. CIL 83, Barcellona, Spain, June 1983

Barrow und Popplestone 71
Relational Descriptions in Picture Processing
H.G. Barrow und R.J. Popplestone
Machine Intelligence 6 (B. Meltzer, D. Michie, eds.)
University Press Edinburgh, 1971, 377-396

Barwise und Perry 83
Situations and Attitudes
J. Barwise, J. Perry
Bradford Books, 1983

Busemann 83
Oberflaechentransformationen bei der Generierung
geschriebener deutscher Sprache
St. Busemann
in B. Neumann (Hrsg.), GWAI-83, Informatik Fachberichte 76,
Springer, Berlin/Heidelberg/New York 1983, 90-99

Dreschler und Nagel 82
Volumetric Model and 3D-Trajectory of a Moving Car Derived
from Monocular TV-Frame Sequences of a Street Scene
L. Dreschler and H.-H. Nagel
Computer Vision, Graphics and Image Processing 20 (1982),
199-228

Fillmore 68
The Case for Case
Charles J. Fillmore
in E. Bach and R.T. Harms (eds.), Universals in Linguistic
Theory. Holt, Rinehart and Winston, New York 1968, 1-88

Forbus 83
Qualitative Reasoning About Space and Motion
K.D. Forbus
in: D. Gentner and A.L. Stevens (eds.), Mental Models,
Lawrence Erlbaum Associates, 1983

Jameson und Wahlster 82
User Modelling in Anaphora Generation:
Ellipsis and Definite Description
A. Jameson, W. Wahlster
Proc. First European Conference on AI ECAI-82, 222-227 (1982)

Marburger et al. 81
Natural Language Dialogue about Moving Objects in an
Automatically Analyzed Traffic Scene
H. Marburger, B. Neumann, H.-J. Novak
IJCAI-81, 49-51

Neumann 82
Towards Natural Language Description of Real-World Image Sequences
B. Neumann
GI - 12. Jahrestagung, Informatik Fachberichte 57, Springer Berlin/Heidelberg/New York 1982, 349-358

Olson 72
Language Use for Communicating, Instructing and Thinking
David R. Olson
in R.O. Freedle, J.B. Carroll (eds.), Language Comprehension and the Acquisition of Knowledge, Washington 1972

Waltz 81
Toward a Detailed Model of Processing for Language Describing the Physical World
D.L. Waltz
IJCAI-81, 1-6