Kognitionswissenschaft (1999) 8: 137–142

Kognitions-
wissenschaft


# Ein Situierter Künstlicher Kommunikator für Konstruktionsaufgaben*

**Bernd Hildebrandt, Alois Knoll, Christian Scheering, Jianwei Zhang**

SFB 360, Universität Bielefeld, Postfach 100 131, 33501 Bielefeld
(e-mail: {berndh, knoll, pcscheer, zhang}@techfak.uni-bielefeld.de)

## A Situated Artificial Communicator for Assembly Tasks

**Summary.** In this article we describe a *Situated Artificial Communicator* for assembly tasks. The main components of the system we are developing are a speech understanding module and a two-arm-robot module. The robot system can be instructed using spontaneous speech. The speech understanding module is based on *Combinatory Categorial Grammar*, which makes incremental and interactive speech understanding possible. The robot module is provided with multiple sensors and it masters complex assembly operations like peg-in-hole or screwing a nut into a bolt. The architecture and the underlying cognitive principles enable interactive processing that depends on the actual situation and allows the system to take advantage of redundant items of information. Due to these principles our Situated Artificial Communicator is highly robust.

**Zusammenfassung.** In diesem Beitrag stellen wir einen *Situierten Künstlichen Kommunikator* vor, im vorliegenden Fall ein Robotersystem für Konstruktionsaufgaben. Das Robotersystem kann durch spontansprachliche Anweisungen gesteuert werden. Die Hauptkomponenten des Systems sind eine Sprachverstehenskomponente und eine Zwei-Arm-Roboterkomponente. Die Sprachverstehenskomponente basiert auf der *Combinatory Categorial Grammar* und ermöglicht eine inkrementelle und interaktive Sprachverarbeitung. Die Roboterkomponente verfügt über eine Vielzahl von Sensoren und beherrscht Montageoperationen wie Stecken und Schrauben. Durch die gewählte Architektur und die zugrundegelegten kognitiven Verarbeitungsprinzipien können Teilkomponenten des Systems der aktuellen Situation entsprechend interagieren und Informationsredundanz nutzen. Das System erhält dadurch eine hohe Robustheit.

* Diese Arbeit wurde von der Deutschen Forschungsgemeinschaft (DFG) im Rahmen des Sonderforschungsbereichs 360 gefördert.

## 1 Einleitung

Der Sonderforschungsbereich *Situierte Künstliche Kommunikatoren* (SFB 360) hat sich neben der Schaffung theoretischer Grundlagen auch die prototypische Realisierung eines Situierten Künstlichen Kommunikators (SKK) zum Ziel gesetzt (vgl. z.B. Rickheit & Wachsmuth, 1996). Um die Vielfalt und Komplexität natürlicher Kommunikationssituationen überschaubar und modellierbar zu halten, wurden als Forschungsgegenstand aufgabenorientierte Dialoge gewählt und als gemeinsame empirische Basis ein Konstruktionsszenario. Das Szenario deckt alle relevanten Aspekte wie sprachliches Handeln, visuelle Perzeption und Objektmanipulation ab. Der menschliche Instrukteur erteilt anhand eines Bauplans einem künstlichen Konstrukteur (einem Robotersystem) Anweisungen, um aus Holzbauteilen ein Modellflugzeug zu konstruieren.

Innerhalb des SFB gibt es sich ergänzende Ansätze, um sich der Realisierung eines SKK zu nähern. Mit *CODY*, dem virtuellen Konstrukteur, lassen sich aus Bauteilen, die 3D-computergraphisch visualisiert werden, beliebige Aggregate auf einer virtuellen Montagefläche konstruieren (vgl. z.B. Wachsmuth & Jung, 1996). Hierzu wird ein hybrider Repräsentationsformalismus für imaginale und attributive Information entwickelt. *CoRA* ist ein virtueller Roboter, der in einer hybriden Architektur Sprache, Perzeption und Handlung integriert, und der auf Anweisungen unterschiedlicher Komplexität situiert und flexibel reagiert (vgl. z.B. Peters, Strippgen & Milde, 1998). Zusätzlich werden bis zu 15 heterogene Software-Module, die von den jeweiligen Teilprojekten erstellt wurden, zu einem Gesamtsystem integriert (vgl. z.B. Fink et al., 1998).

Von zentraler Bedeutung ist die konkrete Realisierung eines SKK. Im SFB wird deshalb ein Robotersystem für Konstruktionshandlungen entwickelt, das durch spontan gesprochene Anweisungen gesteuert werden kann. Die zugrundegelegten Verarbeitungsprinzipien und die Architektur des Systems sind kognitiv motiviert (vgl. z.B. Hildebrandt et al., 1995; Knoll, Hildebrandt & Zhang, 1997), wobei die technischen Rahmenbedingungen teilweise enge Grenzen setzen. So verfügt das System beispielsweise über zwei Roboterarme und ist diesbezüglich einem menschlichen Konstrukteur nachempfunden (vgl. Abbildung 1). Dies gilt aber bei-

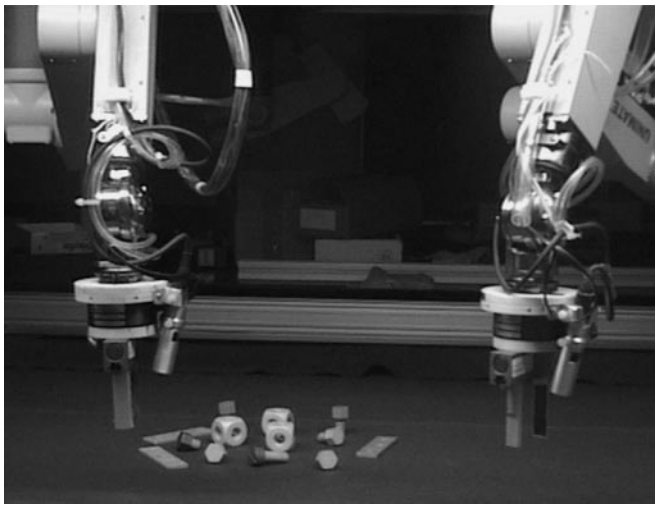

**Abb. 1.** Kooperierende Roboterarme

spielsweise nicht für die Greifer, die nur über zwei Finger verfügen. Im folgenden wird dieses Robotersystem für Konstruktionsaufgaben und seine kognitiven Grundlagen vorgestellt und gezeigt, wie Robustheit erreicht wird. Die Entwicklung des Systems erfolgt vor allem in enger Kooperation der Teilprojekte *Handlungsanweisungen* (C3) und *Multisensorgestützte Exploration und Montage* (D4).

## 2 Prinzipien kognitiver Informationsverarbeitung

Um die Prinzipien menschlicher Text- und Sprachverarbeitung zu verstehen, wurde eine Vielzahl psycholinguistischer Experimente durchgeführt. Häufig lagen den Experimenten sprachlich ambige Strukturen zugrunde, die Rückschlüsse auf unterschiedliche Verarbeitungsprozesse zulassen. Verfeinerte Meßmethoden ermöglichen inzwischen multimodale Experimente, bei denen beispielsweise die Wechselwirkungen zwischen visueller und sprachlicher Information untersucht werden (vgl. beispielsweise die Experimente in Kessler et al., im gleichen Heft; Weiß et al., im gleichen Heft). Die Ergebnisse legen eine sprachunabhängige Verallgemeinerung der Verarbeitungsprinzipien nahe. Demnach sollte ein kognitiv motiviertes Robotersystem zumindest folgende Prinzipien berücksichtigen:

- die Verarbeitung ist inkrementell,
- die Verarbeitung ist interaktiv und
- die Verarbeitung ist modular.

Bei einer *inkrementellen* Verarbeitung wird Information möglichst umgehend verarbeitet, ohne diese zuvor zu sammeln oder auf bestimmte Schlüsselinformationen zu warten. Bezüglich der Sprachverarbeitung heißt dies, daß Wörter in der Regel unmittelbar und interaktiv verarbeitet werden, ohne zuvor auf das Satzende zu warten (vgl. z.B. Hildebrandt & Rickheit, 1997). Verarbeitungsrelevante Informationseinheiten, die Inkremente, können neben Wörtern auch Konstituenten im syntaktischen Bereich oder Konzepte bzw. Referenten im mentalen Modell im semantischen Bereich sein. Inkremente im visuellen Bereich können objektbezogene Regionen, die durch Farbe oder Form eine Einheit bilden, oder Aufmerksamkeitsbereiche sein. Obwohl die Granularität gegenwärtig unterspezifiziert ist, wird von einer vom Verarbeitungsstand und -ziel abhängigen Einteilung ausgegangen.

Eng verknüpft mit der inkrementellen Verarbeitung ist das Prinzip der *interaktiven* Verarbeitung, bei der unterschiedliche Informationsbereiche frühzeitig berücksichtigt werden. Dies können neben Interaktionen zwischen syntaktischer und semantischer Information auch Interaktionen zwischen sprachlicher und visueller Information sein. Beispielsweise konnte Sichelschmidt (1995) durch Augenbewegungsexperimente zeigen, daß die Reihenfolge von Lokalangaben durch Adjektive (*das rechte obere Objekt* bzw. *das obere rechte Objekt*) die Blickbewegungen beeinflussen können, wenn Versuchspersonen in einer symmetrischen Konstellation das benannte Objekt ansehen sollten (vgl. auch Spivey-Knowlton et al., 1994). In welchem Maße und unter welchen Bedingungen sprachliche Verstehensprozesse mit den verschiedenen Informationsbereichen interagieren, ist größtenteils noch unerforscht. Dennoch muß ein kognitiv motiviertes Robotersystem Möglichkeiten zur Interaktion und Integration vorsehen; technisch bedeutet letzteres die Möglichkeit zur Datenfusion.

Wenn man menschliche Informationsverarbeitung als interaktive Verarbeitung auffaßt, dann gibt es Verarbeitungsebenen, Komponenten oder Module, die miteinander interagieren. *Modularität* wird seit Fodor (1983) meist als ein kognitives Architekturmodell verstanden, bei dem einzelne kognitive Module isoliert voneinander sind und sequentiell durchlaufen werden. Wir unterscheiden hingegen funktional unterschiedliche Verarbeitungsbereiche, wie beispielsweise Syntax und Semantik, Diskurs- und Weltwissen oder visuelle und taktile Wahrnehmung.

## 3 Systemkomponenten und Architektur

Die Hauptkomponenten der Realisierung sind ein Spracherkenner und eine Sprachverstehenskomponente sowie das Zwei-Arm-Robotersystem, das über eine Vielzahl von Sensoren verfügt. Neben einem externen Multi-Kamera-System zur Objekterkennung und groben Lagepositionierung ist auf jedem Arm eine „Eye-on-Hand"-Kamera montiert, durch die im Nahbereich mittels eines visuellen Regelungsprozesses Bauteile exakt gegriffen werden können (vgl. Zhang, Schmidt & Knoll, 1999). Zusätzlich sind an den Handgelenken der Arme Kraftmomentsensoren montiert. Sie erlauben dem System komplexe Montageoperationen wie Stecken und Schrauben. Durch die schritthaltende Auswertung sowohl der Kräfte als auch der Bilder der Handkameras werden dabei Fehler in Position und Orientierung der Greifer weitgehend ausgeglichen.

Der Spracherkenner basiert auf dem Diktiersystem IBM *Via Voice Gold*. Um die Erkennungsleistung zu erhöhen, wurde das Vollformen-Lexikon dem Vokabular des Szenarios angepaßt und das System um grammatische Bedingungen und einfache Wortkompositionsregeln erweitert (vgl. Reinsch, Hildebrandt & Zhang, 1998). Der Spracherkenner ist als „Frontend" zur Sprachverstehenskomponente angelegt. Falls eine Äußerung nur partiell korrekt erkannt wurde, ist in der Regel dennoch eine zumindest teilweise Interpretation möglich.

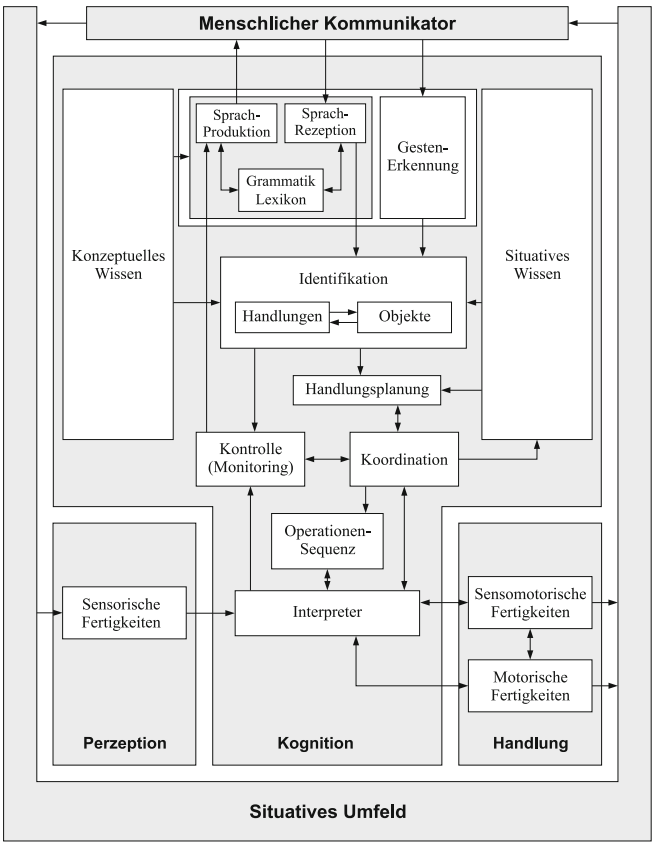


**Abb. 2.** Architektur für den Situierten Künstlichen Kommunikator

Die Sprachverstehenskomponente basiert auf einer von Mark Steedman entwickelten Variante der Kategorialgrammatik, der *Combinatory Categorial Grammar* (z.B. Steedman, 1987, 1993, 1996). Um den kognitiven Verarbeitungsprinzipien der Sprachrezeption, wie z.B. der inkrementellen Verarbeitung, und der flexiblen Konstituentenstellung im Deutschen zu entsprechen, wurde die Grammatik modifiziert und an das Deutsche angepaßt. Dadurch können selbst komplexe und (lokal) ambige Anweisungen (z.B. *steck die lange Schraube in das zweite Loch von links der siebenlöchrigen Leiste*) effizient interpretiert werden. Aus den sprachlichen Anweisungen leitet die Sprachverstehenskomponente semantische Strukturen ab, aus denen wiederum elementare Operationen für den Roboter generiert werden.

Entscheidend ist die Architektur, die die Berücksichtigung des situativen Umfelds ermöglicht (vgl. Abbildung 2). Die passiven Interaktionen (Perzeption) und aktiven Interaktionen (Handlung) sind angedeutet. Der menschliche Instrukteur kann sprachliche Anweisungen erteilen und Objekte durch Zeigegesten spezifizieren. Die Interpretation einer Anweisung basiert zusätzlich zum linguistischen Wissen auch auf konzeptuellem Wissen über Objekte und Handlungen. Gegenwärtig werden daher vollständig interpretierte Anweisungen an ein Modul *Identifikation* weitergereicht. Das heißt, alle Objekte, die zur Ausführung der Handlung benötigt werden, sind explizit angegeben, auch wenn eines der Objekte sprachlich nicht benannt wurde. Im Modul Identifikation werden unter Berücksichtigung situativen Wissens Objekte für die maßgebliche Handlung identifiziert und diese dadurch spezifiziert. Hierbei können sprachlich ambige Anweisungen aufgelöst werden. Geplant ist eine inkrementelle Schnittstelle, durch die handlungsrelevante Inkremente übergeben werden. Im Anschluß an die Identifikation wird situativ abhängig eine Sequenz elementarer Handlungen generiert (*Handlungsplanung*), die auf die einzelnen Handlungsmöglichkeiten des Roboters, wie Greifen oder Schrauben, abgestimmt ist und die den Zustand des Roboters einbezieht. So kann beispielsweise die Handlungssequenz für eine Schraubanweisung um Greifhandlungen ergänzt werden, wenn sich die dazu benötigten Objekte noch nicht in den Roboterhänden befinden. Die Bereiche *Kontrolle* (*Monitoring*), *Koordination*, *Operationensequenz* und *Interpreter* dienen der Informationsverwaltung und Ablaufsteuerung und können als eigener Bereich des *Scheduling* von Handlungen aufgefaßt werden (vgl. Zhang, von Collani & Knoll, 1998). Darin werden entsprechend des aktuellen Roboterzustands aus elementaren Handlungen Roboteroperationen generiert. So wird aus der Handlung schrauben(Nut,Bolt) die folgende Operationensequenz:

1. find_contact(nut, bolt) (Stelle Kontakt her),
2. find_drill() (Finde die Bohrung),
3. find_notch_point() (Finde den Gewindegang),
4. screw() (Schraub ein).

Zur Abarbeitung einer solchen Operationensequenz verwendet der Interpreter sogenannte Fertigkeiten (*skills*). Aus dem perzeptiven Bereich stellen sensorische Fertigkeiten Informationen über die Situation zur Verfügung. Einige davon müssen echtzeitfähig sein, um die technischen Randbedingungen für geschlossene Regelkreise der Robotersteuerung zu erfüllen. Als positiver Nebeneffekt wird eine unmittelbare Reaktion des Systems auf menschliches Eingreifen möglich und somit eine wirklichkeitsnahe Interaktion mit dem Menschen. Die Handlungen werden über motorische Fertigkeiten umgesetzt, die einzelne Roboterbewegungen wie das Öffnen, Schließen oder Drehen des Greifers steuern. Als direktes Bindeglied zwischen Perzeption und Aktion dienen sensomotorische Rückkopplungen, die beispielsweise für kraftüberwachte Bewegungsregelungen benötigt werden. Unerwartete Ereignisse, wie etwa eine Kraftüberschreitung oder der Ausfall eines Moduls, werden an den Interpreter gemeldet, der im Zusammenspiel mit Kontrolle und Kooperation geeignete Maßnahmen wie z.B. Rückfragen an den Instrukteur veranlassen kann. (Allerdings ist eine Komponente zur Sprachproduktion gegenwärtig noch nicht integriert.) Über den Interpreter wird darüber hinaus eine ständige Aktualisierung des Wissens über die Situation vorgenommen. So zählt zum Wissen über die Szene unter anderem, ob sich ein Objekt auf dem Tisch oder in der Hand des Konstruktionsroboters befindet. Die Veränderung der Objektposition erfolgt in solchen Fällen meist durch intendierte Handlungen. Eine Veränderung der Szene, die eine Aktualisierung der Szenenrepräsentation erfordert, kann allerdings auch unbeabsichtigt auftreten, beispielsweise wenn eine bisher stehende Schraube umfällt, während ein anderes Objekt in ihrer Nähe gegriffen wird.

Um eine sprachliche Anweisung durch gestische Information unterstützen zu können, ist in der Architektur eine Komponente zur Gesteninterpretation vorgesehen. Dadurch soll die korrekte Umsetzung deiktischer Anweisungen wie

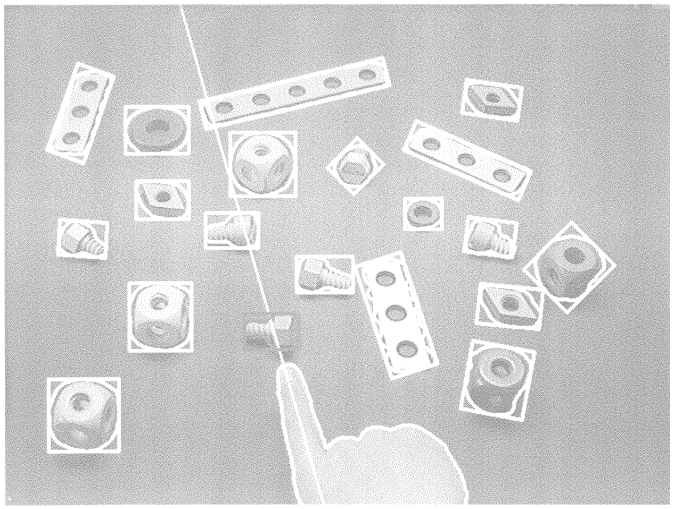

**Abb. 3.** Interpretation einer Zeigegeste

*nimm die rote da* möglich werden. Während zur Zeit die Ausrichtung einer Zeigegeste durchaus erkannt werden kann und zur Auswahl zwischen zwei Objekten genutzt wird, ist die Bestimmung eines beliebigen Objektes, auf das die Geste zielt, noch relativ ungenau (vgl. Abbildung 3). Um dies zu verbessern, wird versucht, die ungefähre 3D-Raumrichtung der Gestik (vgl. Cipolla & Hollinghurst, 1996, 1997) zu bestimmen und in den Auswahlprozeß zu integrieren. Erschwerend kommt hinzu, daß zumindest für längere Anweisungen eine temporale Synchronisation zwischen Geste und sprachlichem Ausdruck erfolgen muß.

## 4 Robustheit durch Interaktion und Redundanz

Während der Handlungsinterpretation und -ausführung kommt es zu einem komplexen Interaktionsgeflecht zwischen den Modulen. Die vorgestellte Architektur und die zugrundegelegten Verarbeitungsprinzipien bilden die Basis für ein robustes System, das unerwartete, unvollständige und auch widersprüchliche Informationen verarbeiten kann, ohne handlungsunfähig zu werden. Voraussetzung hierfür ist, daß das System

- instabile Zustände erkennt,
- die Situation diagnostizieren kann und
- eigenständig Fehlerbehandlungsroutinen einleitet oder über den Zustand berichtet und den Instrukteur interaktiv um Unterstützung bittet.

Ein Grund für die Notwendigkeit einer robusten Informationsverarbeitung ist die Erkennungsquote bei den sprachlichen Anweisungen. Daraus folgt, daß das Sprachverstehenssystem lexikalische Einfügungen, Ersetzungen und Auslassungen bewältigen muß. Außerdem zeigen die im SFB erhobenen Sprachdaten, daß eine Äußerung nicht immer einen Satz bildet, zumindest nicht gemäß einer üblichen Grammatik. Ein Sprachverstehenssystem muß also selbst bei einem perfekten Spracherkenner robust bezüglich der sprachlichen Eingabe sein. Ähnliches gilt für die Perzeptions- und die Handlungskomponente des SKK. Insgesamt muß das System eine Reihe qualitativ unterschiedlicher Schwierigkeiten überwinden, die auf im folgenden eingegangen wird:

*Wortersetzungen.* Der Spracherkenner ist oft unsicher bezüglich der Flexionsendungen *-em* und *-en*, so daß beide leicht vertauscht werden. Da durch die Flexion u.a. Kasus und Genus markiert sind, kann bei einer Vertauschung die vorliegende Nominalphrase nicht als syntaktisch kongruent analysiert werden. Als Fehlerbehandlungsroutine liegt es daher nahe, auf Kasus- und Genuskongruenz innerhalb einer Nominalkonstituente ganz zu verzichten, wenn eine alternative kongruente Konstituente fehlt. Um Übergeneralisierungen zu vermeiden, kann eine solche Strategie allerdings nur vereinzelt eingesetzt werden. Experimentell wird zur Zeit versucht, geeignete Kriterien für eine constraint-tolerante Grammatik zu finden.

*Semantische Inkonsistenz.* Eine Anweisung kann semantisch inkonsistent sein oder zumindest dem System so erscheinen. Im Zweifelsfall muß das System beim Instrukteur um eine Korrektur oder eine detaillierte Spezifikation nachfragen. Manchmal kann es eine vollständige Handlungsanweisung ableiten, indem das System weitere Informationen aggregiert und fehlende Information inferiert. Wörtlich genommen kann beispielsweise die Anweisung *schraub die Leiste auf den Würfel* nicht ausgeführt werden, da weder die Leiste noch der Würfel die Funktion einer Schraube ausübt. Wird allerdings das benötigte Instrument vom System inferiert (*schraub die Leiste mit der roten Schraube auf den Würfel*), ist die Handlung möglich. Bei einer unterspezifizierten Anweisung wird also aufgrund semantischen Wissens das benötigte Instrument inferiert.

*Sprachliche Ambiguität.* Das Sprachverstehenssystem muß darüber hinaus verschiedene Formen der Ambiguität bewältigen. Auf der lexikalischen Verarbeitungsebene kann ein Wort unterschiedliche syntaktische Kategorien haben, z.B. kann *schrauben* sowohl Nomen als auch Verb sein (die entsprechende Orthographie liefert der Spracherkenner nicht, da hierzu eine syntaktische Voranalyse nötig wäre). In den meisten Fällen wird eine solche lexikalische Ambiguität durch den syntaktischen Kontext aufgelöst; d.h. es wird jene Kategorie gewählt, mit der eine syntaktische Analyse möglich ist. Außerdem kann auf der syntaktischen Verarbeitungsebene die Anbindung der Konstituenten ambig sein. Ein klassisches Beispiel hierfür ist die Anbindung der Präpositionalphrase (vgl. Hildebrandt & Rickheit, 1997). Hierbei nutzt das Sprachverstehenssystem eine Vielzahl von Informationsquellen. Unter anderem wird geprüft, ob für das polyseme Verb eine präferierte Lesart vorliegt, ob das Objekt semantisch mit der Funktionalität des Verbarguments übereinstimmt und für welche Interpretation es ein Denotat im situativen Kontext gibt. Zur Zeit erfolgt die Disambiguierung erst, nachdem der gesamte Satz verarbeitet wurde, d.h. am Äußerungsende liegen mindestens zwei Satzinterpretationen vor. Angestrebt ist allerdings eine inkrementelle Disambiguierung. Kann das Sprachverstehenssystem die Ambiguität nicht selbst auflösen, sollte es beim Instrukteur nachfragen. Die Roboterkomponente erhält jedenfalls eine eindeutige Sequenz elementarer Handlungen.

*Situative Ambiguität.* Eine Anweisung, die von der Sprachverstehenskomponente semantisch vollständig interpretiert wurde, führt nicht unbedingt zu einer gleichfalls erfolgreichen Handlung. Beispielsweise können bei einer Anweisung

wie *nimm eine Schraube* mehrere Schrauben in der Szene als Referent zur Verfügung stehen. Zur Auflösung kann gegebenenfalls Zeigegesteninformation herangezogen werden. Im Regelfall folgt die Handlungskomponente allerdings einem Ökonomieprinzip: Es wird jenes Objekt gegriffen, das am leichtesten zu erreichen ist und dessen Positionierung die wenigsten Greifprobleme bereitet. Je nach Verteilung der Objekte in der Szene sollte das System auch nachfragen können, ob beispielsweise eine lange oder eine kurze Schraube benötigt wird.

*Sprachliche Intervention.* Aus einer Vielzahl von Ursachen kann es dazu kommen, daß die Handlungskomponente eine Handlung ausführen will, die bezüglich der Intention des Instrukteurs falsch ist. Dies ist beispielsweise immer dann der Fall, wenn das System eigenständig entscheidet und ein Objekt greifen will, das der Instrukteur nicht gemeint hat. Während der Greifarm sich auf das Objekt zubewegt, hat der Instrukteur allerdings die Möglichkeit, verbal zu intervenieren. Gegenwärtig kann die Handlung nur durch den Befehl *stop* unterbrochen werden. Ergänzend sollen künftig auch korrigierende Information wie *eine andere Schraube* bzw. *besser die vordere Schraube* einfließen können. Die Handlung wird hierdurch lediglich modifiziert, womit vermieden werden soll, daß die Handlung völlig neu einsetzen muß.

*Objektposition.* Die Detektion von Objekten und die Bestimmung ihrer jeweiligen Position sowie Ausrichtung in einer komplexen Szene mit Objektüberdeckungen und Hintergrundobjekten ist überaus schwierig. Die visuelle Komponente benutzt daher mehrere Kameras, die die Szene aus unterschiedlichen Perspektiven aufnehmen. Hierdurch wird einerseits eine höhere Toleranz gegenüber schlechter Signalqualität, wie sie z.B. bei ungünstigen Lichtverhältnissen gegeben ist, oder gar gegenüber Bildausfall erreicht. Andererseits besteht die Hoffnung, daß durch ein verteilt operierendes Multi-Sensoren-Agenten-Netzwerk (vgl. Scheering & Knoll, 1998) verdeckte Objekte sicherer erkannt und ihre Lage präziser bestimmt werden kann.

*Objektmontage.* Um Ungenauigkeiten in der Objektpositionsschätzung auszugleichen, wird zusätzlich die (redundante) lokale Sicht der „Eye-on-Hand"-Kameras verwendet (vgl. Zhang, Schmidt & Knoll, 1999). Dabei wird zuerst die geschätzte Objektlage mit dem Roboter angefahren (grobe Zielannäherung). Anschließend wird durch eine visuell geregelte Positionskontrolle versucht, die optimale Greifpostur einzunehmen. Nach Kontaktherstellung wird zusätzlich mit der Handkamera geprüft, ob das Objekt gegriffen wurde. Um sowohl Ungenauigkeiten der Positionierung als auch potentielle Griffvarianzen auszugleichen, z.B. durch „Verspringen" des Objekts im Moment des Zugreifens, sind Schrauben und Stecken nicht als festprogrammierte Bewegungsfolgen, sondern als kraftpositionsgeregelte und zum Teil visuell geleitete Prozesse realisiert.

Die Anforderung an das Gesamtsystem, echtzeitfähig zu sein, hat für die Sprachkomponente zur Folge, daß sie nicht beliebig lange Zeit für die Interpretation benutzen darf. Sie muß vielmehr jederzeit in der Lage sein, zumindest Teilinterpretationen an andere Komponenten liefern zu können (*any-time capability*). Hierzu benötigt sie Kriterien, um eine laufende Analyse und Interpretation abbrechen zu können.

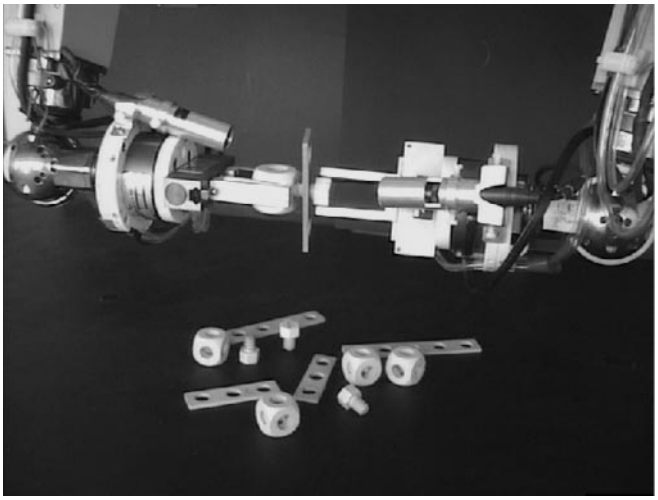

**Abb. 4.** Beginn der Einschraubphase nach der Kontaktherstellung von Bohrung und Gewinde

Ein quantitatives Abbruchkriterium ist die bisher verstrichene Zeit und die Anzahl konkurrierender Lösungswege. Ein qualitatives Abbruchkriterium ist erreicht, wenn die zur Handlung benötigte Information vorliegt, also beispielsweise der Typ der Handlung und alle obligatorischen Argumente. Zugrundegelegt wird hierbei eine inkrementelle und interaktive Verarbeitung sowie das Prinzip, nicht allen möglichen Interpretationen zu folgen, sondern nur der stabilsten und informativsten (*intelligent pruning*). Die Operationalisierung dieses Prinzips ist allerdings nicht unproblematisch und hängt oft von spezifischen Konstellationen ab.

Die Robustheit des gesamten Systems beruht auf Redundanz, da sich verschiedene Informationsquellen auf denselben Bereich beziehen, wie beispielsweise bei der visuellen Perzeption. Darüber hinaus bedient sich das System reaktiver und interaktiver Robustheitsstrategien:

- systemextern, indem mit dem Instrukteur kommuniziert wird, wobei die Initiative vom Situierten Künstlichen Kommunikator oder vom Instrukteur ausgehen kann,
- systemintern bzw. intermodular, indem verschiedene Systemkomponenten interagieren, und
- modulintern, indem verschiedene Arten linguistischer Information oder visuelle Information aus unterschiedlichen Perspektiven integriert werden.

## 5 Zusammenfassung und Ausblick

Der im Rahmen des SFB 360 entwickelte Konstruktionsroboter wird durch spontansprachliche Anweisungen gesteuert. Er kann beliebige Objekte des Szenarios greifen und durch Fügehandlungen wie Stecken und Schrauben miteinander verbinden (s. Abbildung 4). Aufgrund der gewählten Architektur, der zugrundegelegten kognitiven Verarbeitungsprinzipien und der Berücksichtigung von Situativität erhält das System ein hohes Maß an Robustheit.

Weiterführende Arbeiten beziehen sich sowohl auf die Sprachverstehenskomponente als auch auf den Konstruktionsroboter. So kann die Sprachverstehenskomponente unterspezifizierte Handlungsanweisungen zwar in Roboterdirektiven übertragen, die der Konstruktion komplexer Aggregate

dienen. Doch kann die Struktur der Aggregate oder allein ihr Gewicht unter Umständen dazu führen, daß eine Realisierung mit zwei Greifbacken und ohne weitere Hilfsmittel nicht möglich ist. Dasselbe gilt für das rechtwinklige Ausrichten zweier Objekte: Der Winkel zwischen den Objekten kann so ungünstig sein, daß mit den Greifbacken eine Ausrichtung nicht möglich ist. Für solche Fälle ist vorgesehen, daß der SKK den Instrukteur um aktive Hilfe bittet, d.h. der SKK erkennt seine Grenzen und fordert den Instrukteur auf, die gewünschte Aktion zumindest partiell selbst auszuführen. Die dafür benötigte Dialogkomponente muß noch entwickelt werden, wobei die Ergebnisse und die Erkenntnisse anderer Teilprojekte des SFB einfließen werden. Die Grammatik und das Lexikon werden fortlaufend ausgebaut und den erweiterten Möglichkeiten des Konstruktionsroboters angepaßt. Hierbei werden die empirischen Befunde der psycholinguistischen Experimente zum Sprachverstehen berücksichtigt, um ein Gesamtsystem zu erhalten, das im Sinne der experimentell-simulativen Methode neue Fragestellungen aufwirft und gegebenenfalls seinerseits als Experimentalumgebung dienen kann.

## Literatur


Cipolla, R. & Hollinghurst, N. (1996). Human-robot interface by pointing with uncalibrated stereo vision. *Image and Vision Computing*, *14*, 171-178.

Cipolla, R. & Hollinghurst, N. (1997). Visually guided grasping in unstructured environments. *Robotics and Autonomous Systems*, *19*, 337-346.

Fink, G. A., Jungclaus, N., Ritter, H. & Sagerer, G. (1998). Integration verteilter Systeme mit DACS. *Mustererkennung 1998, 20. DAGM-Symposium*. Berlin: Springer.

Fodor, J. (1983). *The modularity of mind*. Cambridge, MA: Bradford.

Hildebrandt, B., Moratz, R., Rickheit, G. & Sagerer, G. (1995). Integration von Bild- und Sprachverstehen in einer kognitiven Architektur. *Kognitionswissenschaft*, *4*, 118-128.

Hildebrandt, B. & Rickheit, G. (1997). *Verarbeitung von Präpositionalphrasen in der Combinatory Categorial Grammar* (Report 97/6 - Situierte Künstliche Kommunikatoren, SFB 360). Bielefeld: Universität Bielefeld.

Kessler, K., Hoffhenke, M., Rickheit, G. & Wachsmuth, I. (1999). Dynamische Konzeptverarbeitung mit imaginalen und assoziativen Strukturen. *Kognitionswissenschaft*, *8*, 115–122.

Knoll, A., Hildebrandt, B. & Zhang, J. (1997). Instructing cooperating assembly robots through situated dialogues in natural language. *Proceedings, IEEE Conference on Robotics and Automation, Albuquerque, New Mexico, April 1997*. New Mexico: ICRA.

Peters, K., Strippgen, S. & Milde, J.-T. (1998). CoRA - An instructable robot. In T. Lueth, R. Dillmann, P. Dario & H. Wörn (eds.), *Distributed Autonomous Robotic Systems 3* (pp. 247-256). Berlin: Springer.

Reinsch, M., Hildebrandt, B. & Zhang, J. (1998). *RCRC- Ein flexibles System zur Spracherkennung* (Report 98/11 - Situierte Künstliche Kommunikatoren, SFB 360). Bielefeld: Universität Bielefeld.

Rickheit, G. & Wachsmuth, I. (1996). Collaborative Research Center „Situated Artificial Communicators“at the University of Bielefeld, Germany. *Artificial Intelligence Review*, *10*, 165-170.

Scheering, C. & Knoll, A. (1998). Framework for implementing self-organized task-oriented multisensor networks. *Proceedings of SPIE International Symposium on Intelligent Systems and Advanced Manufacturing, Vol. 3523, Boston, November 1998*.

Sichelschmidt, L. (1995). *Der Augenblick des Verstehens: Motorische Indikatoren unmittelbarer semantischer Verarbeitung* (ZIF Mitteilungen). Bielefeld: Universität Bielefeld, 4/95, 3-15.

Spivey-Knowlton, M., Sedivy, J., Eberhard, K. & Tanenhaus, M. (1994). Psycholinguistic study of the interaction between language and vision. *AAAI-94 Workshop on Integration of Natural Language and Vision Processing. Twelfth National Conference on Artificial Intelligence*. Seattle, WA: American Association for Artificial Intelligence, 189-192.

Steedman, M. (1987). Combinatory grammars and human language processing. In J. L. Garfield (Ed.), *Modularity in knowledge representation and natural-language understanding* (pp. 187-205). Cambridge, MA: MIT Press.

Steedman, M. (1993). Categorial grammar. *Lingua*, *90*, 221-258.

Steedman, M. (1996). *Surface structure and interpretation*. Cambridge, MA: MIT Press.

Weiß, P., Kessler, K., Hildebrandt, B. & Eikmeyer, H.-J. (1999). Konzeptualisierung in inkrementell-integrativer Sprachverarbeitung. *Kognitionswissenschaft*, *8*, 108–114.

Wachsmuth, I. & Jung, B. (1996). Dynamic conceptualization in a mechanical-object assembly environment. *Artificial Intelligence Review*, *10*, 345-368.

Zhang, J., Schmidt, R. & Knoll, A. (1999, im Druck). Appearance-based visual learning in a neuro-fuzzy model for fine-positioning of manipulators. *Proceedings of the IEEE International Conference on Robotics and Automation, Detroit, MA, USA*.

Zhang, J., von Collani, Y. & Knoll, A. (1998). Development of a robot agent for interactive assembly. In T. Lueth, R. Dillmann, P. Dario & H. Wörn (eds.), *Distributed Autonomous Robotic Systems 3* (pp. 277-286). Berlin: Springer.