Experimentelle Evaluation des situierten Kommunikators — eine Pilotstudie

Katharina J. Rohlfing und Christian Bauckhage

0. Motivation / Abstract

Die Tatsache, dass im SFB ein weltweit einzigartiges System geschaffen wurde, das sich mit dem Menschen über reale Objekte verständigt, birgt zugleich auch eine große Herausforderung, wenn es darum geht, dieses System zu evaluieren. Die meisten Evaluationsstudien beziehen sich auf eine dyadische Form der Mensch-Maschine Kommunikation. Im SFB-Szenario handelt es sich aber um eine triadische Form. Das heißt, dass sich Mensch und Maschine auf Objekte in der realen Welt beziehen und über diese kommunizieren. Im Folgenden werden erste Überlegungsschritte zur Evaluation des situierten Kommunikators und der Hintergrund der Überlegungen dargestellt. Die Pilotstudie hat das Ziel, die bestehenden Evaluationskriterien für künstliche Systeme auf den situierten Kommunikator anzuwenden. Zugleich zeigt sie jedoch auch auf, dass es notwenig ist, bei triadischen Formen von Kommunikation über die bestehenden Kriterien hinauszugehen. Dementsprechend geht die Diskussion der Studie auf neuere Entwicklungen der Evaluationsstudien ein, und ermuntert dazu, Kriterien zu schaffen, die die vollen Leistungen der heutigen multimodalen Systeme berücksichtigen.

1. Stand der Forschung

1.1 Was ist Evaluation?

Folgt man den Erklärungen der Organisationspsychologie, so bedeutet Evaluation die Bewertung der Effizienz, Güte und Wirksamkeit von Programmen, Techniken, Entscheidungen und Handlungen (Weinert, 1987). Bislang stand bei der Übertragung des Prozesses der Evaluation auf die Bewertung künstlicher Systeme die Tatsache im Vordergrund, dass computerbasierte Systeme wie ein Fahrkartenautomat letztendlich vom Menschen bedient werden. Bereits in den 60er Jahren wurde daher vorhergesagt, dass in der Zukunft nicht die computer-orientierten Menschen als vielmehr die menschen-orientierten Computer (Nickerson, 1969) benötigt werden. Der Gedanke, Produkte menschengerechter zu entwickeln, begründete einen wissenschaftlichen Zweig zu Mensch-Maschine Studien, der besonders in den 90er Jahren im Zusammenhang mit der Überprüfung der Benutzbarkeit (*Usability*) populär wurde. In Usability-Studien bedeutet Evaluation konkret: „The gathering of information within a specified context on the usability or potential usability of an

interface, and the use of that information either to improve features within the interface and the supporting material or to assess the completed interface (Preece, 1993). Als Produkte wurden zu diesem Zeitpunkt hauptsächlich Benutzeroberflächen evaluiert, um Informationen zu bekommen, die der Verbesserung sowohl der Einfachheit in der Bedienbarkeit der Oberfläche wie auch der Geschwindigkeit beim Erlernen der Bedienbarkeit dienen sollten.

Nach Dix und Kollegen (1998) verfolgt die Evaluation künstlicher Systeme drei grundsätzliche Ziele:

- Die Bewertung des Ausmaßes an Funktionalität des Systems
- Die Bewertung der Schnittstelle zum Benutzer
- Die Identifizierung jeglicher spezifischer Probleme mit dem System.

Allerdings hängen die Ziele zusätzlich vom Zeitpunkt ab, zu dem eine Evaluation durchgeführt wird. Diesbezüglich werden in der Literatur der 90er Jahre in erster Linie zwei Typen, die formative und die summative Evaluation, diskutiert (s. z.B. Preece, 1993) — s. Tabelle 1.

Tabelle 1: Die Unterscheidung zwischen formativer und summativer Evaluation (nach Wottawa & Thierau, 1998)

Formative Evaluation	**Summative Evaluation**
stellt Informationen bereit für Systeme / Maßnahmen, die verbessert werden sollen	soll die Qualität und den Einfluss bereits fertiger Systeme/ stattgefundener Maßnahmen feststellen und abschließend bewerten
Zielgruppe	
Entwickler	interessierte Öffentlichkeit, Geldgeber
Datensammlung	
Klärung der Ziele, Probleme	Dokumentation
Präsentation	
Diskussion / Treffen, informelle Interaktion	formale Berichte

Während die formative Evaluation den Entwicklern dient und innerhalb des gesamten Designprozesses möglich und hilfreich ist, hat die summative Evalutation das Ziel, die Qualität des Systems zu kontrollieren und abschließend zu bewerten. Dieser Unterschied im Zeitpunkt des Evaluationsprozesses wird von Dix und Kollegen Ende der 90er Jahre noch deutlicher herausgestellt, indem sie begrifflich zwischen *Designevaluation* und *Evaluation der Implementierung* (Dix u.a., 1998)

unterscheiden.

1.2. Methoden der Evaluierung

Unabhängig vom Evaluationstyp erstreckten sich die Methoden und Techniken bereits vom Anfang der Diskussionen über Evaluation in Mensch-Maschine Studien über ein ganzes Spektrum, das sowohl qualitative wie auch quantitative Herangehensweisen enthielt (Preece, 1993). Grundsätzlich findet man folgende Unterscheidungen in Methoden der Evaluierung:

- *Analytische Evaluation:* Dabei handelt es sich um die Analyse von Aufgaben, die die Benutzer erfüllen müssen und die Analyse von Operationen, die nötig sind, um jene Aufgaben zu erfüllen. Diese Evaluationsmethode hat das Ziel, das Verhalten und die Leistung der Benutzer vorauszusehen und zu beschreiben. Bei dieser Methode sind die Benutzer nicht direkt in die Evaluation involviert.
- *Evaluation durch Beobachtung:* Die Benutzer werden beobachtet, während sie das System nutzen; dadurch werden Informationen gewonnen, was sie bei der Interaktion tun. Solche Informationen können anschließend für eine aufgaben- oder leistungsbasierte Analyse genutzt werden. Das Ziel ist, Bedürfnisse zu identifizieren, die sich bei der Interaktion ergeben.
- *Experimentelle Evaluation:* Eine Anzahl von experimentellen Bedingungen wird aufgestellt und im Rahmen bestimmter kontrollierter Variablen variiert. Jede Veränderung im Verhalten der Probanden wird unterschiedlichen Bedingungen zugeschrieben. Durch das Messen von Verhaltensmerkmalen der Probanden ermöglicht diese Methode, Informationen zu gezielten Fragen und Hypothesen zu sammeln (Dix u.a., 1998).
- *Evaluation durch Befragungen:* Bei dieser Methode werden schriftliche oder mündliche Befragungen mit Benutzern durchgeführt. Ziel dieser Befragung ist es, die subjektive Meinung der Benutzer zu der Schnittstelle zu erfahren.
- *Evaluation durch Expertenbefragung:* Bei dieser Methode verlässt man sich auf das Wissen von Experten, die zur Begutachtung eines Produkts, einer Maßnahme hinzugezogen werden. Diese versuchen, sich in die Rolle der Benutzer hinein zu versetzen, wobei sie häufig vielfältige Erfahrungen über Probleme von Mensch-Maschine Schnittstellen haben. Die Daten, die aus solch einer Evaluation gewonnen werden, sind meist qualitativer Natur.

Das experimentelle Vorgehen wird aus mindestens zwei Gründen als eine der erfolgversprechendsten Methoden dargestellt (Dix u.a., 1998: 416). Zum einen vermittelt es empirische Belege um eine spezielle Hypothese zu unterstützen. Zum anderen kann es angewendet werden, um verschiedene Studien auf verschiedenen Detailebenen durchzuführen. Diese Evaluationsmethode findet daher im Zuge der

immer zahlreicheren Studien zur Mensch-Maschine-Kommunikation große Beachtung.

1.3 Usability Kriterien

Nicht nur die Wahl der Methode der Evaluation, sondern vor allem die Gesichtspunkte, unter denen sie durchgeführt wird, sind für das Design einer Evaluationsstudie entscheidend. Die Gesichtspunkte hängen wiederum von den Leistungen der zu evaluierenden Systeme oder Produkte ab. Die meisten aktuellen Evaluationsstudien konzentrieren sich auf Systeme, die eine dyadische Kommunikation ermöglichen: In Preece (2002) werden Studien zur Entwicklung von Internetseiten aber auch zur Entwicklung von Produkten wie Funktelefone, Kopiergeräte usw. angesprochen, bei denen die *Benutzung* dieser Produkte im Vordergrund steht. Der Aspekt der Benutzbarkeit (Usability) wurde Anfang der 90er Jahre intensiv diskutiert. Der Begriff sollte die Seite der Benutzer charakterisieren und ihre Bedürfnisse im Umgang mit Systemen erfassen. Preece (1993: 131) definiert usability als „a measure of the ease with which a system can be learned or used, its safety, effectiveness and efficiency, and the attitude of its users towards it“. An einer anderen Stelle (Lindgaard, 1994) wird ein System dann als ‚usable' bezeichnet, wenn seine Bedienung leicht und schnell zu erlernen ist. Die Spezifizierung dieser Aussage jedoch, d.h. die Frage wann ein System leicht und schnell zu lernen ist, brachte eine Menge an Operationalisierungsideen, die auch als Ziele von Usability-Studien gelten (Preece, 2002: 18). Im Folgenden möchten wir drei Ansätze vorstellen, die die Evaluationskriterien spezifizieren. Die Darstellung dieser Ansätze soll deutlich machen, dass einerseits die Ideen nicht weit voneinander entfert sind. Andererseits ist jedoch auch kein Konsens zu Evaluationskriterien zu erwarten. Letztlich hängt es vom jeweiligen System ab und von den genaueren Zielen der Evaluation, welche Kriterien von Relevanz sind.

Usability: Learnability, Flexibility und Robustness

Dix und Kollegen (1998: 162 ff.) spezifizieren den Oberbegriff der Usability durch drei Unterbegriffe, die wiederum jeweils in weitere, diese Phänomene beeinflussende Prinzipien gefasst werden. Ohne hier auf die einzelnen Prinzipien eingehen zu wollen, werden die Usability spezifizierende Begriffe aufgezählt:

Learnability steht für die Leichtigkeit, mit der neue Benutzer eine effektive Interaktion beginnen können und maximale Leistung erreichen. Die die Learnability unterstützenden Prinzipien gehen auf die Fähigkeiten des Systems zurück, die Benutzer bei ihren Handlungen zu unterstützen, sei es durch Ähnlichkeit der Aufgaben / der Interaktion, Berücksichtigung ihres Vorwissens / ihrer Erfahrung oder Transparenz der Interaktion (vgl. Dix u.a., 1998: 163).

Flexibility steht für die Vielfältigkeit der Wege, auf der die Benutzer und das System Informationen austauschen können. Ähnlich wie Learnability wird sie durch

beeinflussende Prinzipien konkretisiert. Diese gehen auf die Interaktion beschränkende Vorgaben des Systems ein und suchen zu erfassen, in wie weit sich das System den Benutzern anpassen kann.

Als drittes Kriterium der Usability wird bei Dix und Kollegen (1998: 172) *Robustness* erwähnt. Das ist der Grad der Unterstützung, den die Benutzer bekommen und dadurch ihre Aufgaben und Ziele erfolgreich zu Ende führen können. Unterstützt wird *Robustness* durch Fähigkeiten des Systems zur Transparenz der einzelnen Schritte einer Aufgabe, die auch ein späteres Eingreifen ermöglicht und Stabilität des Systems bei allen Aufgaben, die die Benutzer mit ihm lösen wollen.

Auffällig bei den die Usability definierenden Begriffen ist, dass sie auf Fähigkeiten eines künstliches Systems zielen, die die Benutzbarkeit charakterisieren sollen. Die Hervorhebung der Fähigkeit wird in der Begriffbildung deutlich – nicht nur der Begriff *Usability*, sondern auch die weiteren Spezifizierungen wie *Learnability, Predictability, Generalizability, Flexibility, Customizability* usw. enden nach Möglichkeit auf –*ability*.

Usability: ein multidimensionales Konzept

Die Tatsache, dass Usability ein multi-dimensionales Konzept ist, hebt ebenfalls Paternò (1999: 141) hervor und ist der Meinung, dass es mehr beinhaltet als nur die Leichtigkeit des Benutzens und des Lernens. Paternò schlägt daher eine umfangreichere Definition vor, zu der er die folgenden Unterbegriffe zählt: die *Relevanz* des Systems (d.h. wie gut dient das System den Bedürfnissen der Benutzer), seine *Effizienz* (wie effizient die Benutzer die Aufgaben mit dem System erledigen können), die *Einstellung der Benutzer* zum System (ihre subjektiven Empfindungen), die *Lernbarkeit* des Systems (wie leicht ist das System zu erlernen für anfängliche Bedienung und wie gut können sich die Benutzer daran erinnern, wie man das System benutzt[1]) und seine *Sicherheit* (wenn sich das System den Benutzern gegenüber nicht destruktiv verhält und ihnen die Möglichkeit dazu gibt, vorherige Schritte rückgängig zu machen).

Zusätzlich zu diesen Kriterien, gibt Paternò (1999: 143) konkrete Untersuchungs- und Messziele an, die der Evaluation der Benutzerschnittstelle dienen sollen:

- *Aufgaben, die die Benutzer fähig waren zu lösen* als Anzeichen dafür, dass die Benutzerscnittstelle ‚usable' ist
- *Aufgaben, die die Benutzer nicht fähig waren zu lösen* als wichtigen Hinweis darauf, welche Probleme die Benutzer mit der Schnittstelle haben, und welche Hilfe sie benötigen, um sie zu lösen.
- *Wie viele Male eine Aufgabe von Benutzern erledigt wurde* als Hinweis auf mögliche Makros.

[1] Ob die Bedienung eines System leicht zu merken ist, fasst Preece (2002: 14) unter einem speziellen Kriterium zusammen und nennt es: *memorability*.

- *In welcher Reihenfolge wurden die Aufgaben erledigt* um den Benutzern eventuelle Verkürzungen anzubieten.
- *Unterschiedliche Fehler, die die Benutzer machen* um die unterschiedlichen Quellen der Fehler charakterisieren zu können.
- *Subjektive Meinung der Benutzer über die Schnittstelle* als Möglichkeit für die Entwickler, Anregungen zu bekommen, was die Benutzer nicht akzeptieren und warum.
- *Wie viel Hilfe benötigen die Benutzer* als Anzeichen für den Schwierigkeitsgrad der Schnittstelle.
- *Wie viele Male müssen die Benutzer von Anfang an beginnen* als Hinweis darauf, wie schwierig es für die Benutzer ist, durch das Programm zu navigieren.
- *Wie lange brauchen die Benutzer für das Erreichen ihres Ziels*, da eine relativ kurze Zeit bei manchen Produkten ein wichtiger Benutzerfaktor ist.

Usability: Bewertung von Produkten

Ähnlich den konkreten Messzielen, gibt Jameson (2001) Kriterien an, denen sich jedes zu evaluierende Produkt stellen soll. Im Vergleich zu den Kriterien von Paternò, erscheinen Jamesons Kriterien konkreter, da sie an bestimmte Wertvorstellungen geknüpft sind. So drückt das Kriterium *Schnelligkeit der Ausführung* aus, dass eine schnelle Ausführung von Aufgaben wichtig ist. Solche Wertevorstellungen stehen bei Paternò nicht im Vordergrund, sodass die Evaluation nach seinen Kriterien eher einen beschreibenden als bewertenden Charakter bekommt. Dagegen lassen sich die Kriterien von Jameson auch gut auf einen Vergleich zweier Systeme anwenden.

1. Schnelligkeit der Ausführung (Wie schnell kann die Benutzerin / der Benutzer $\mathcal{B}$ mit dem System $\mathcal{S}$ die Aufgaben erledigen?)
2. Schnelligkeit des Lernens (Wie schnell kann $\mathcal{B}$ lernen, eine Aufgabe mit $\mathcal{S}$ zu erledigen?)
3. Anzahl der ernsthaften Fehler (Wie oft entstehen Fehler in der Benutzung von $\mathcal{S}$ und wie ernsthaft sind sie?)
4. ‚Mental load' (Wie sorgfältig muss $\mathcal{B}$ nachdenken und wie viele Informationen im Kopf behalten während der Benutzung von $\mathcal{S}$?)
5. Funktionalität (Wie viele verschiedene Aufgaben kann $\mathcal{B}$ mit $\mathcal{S}$ erledigen?)
6. Qualität der Ergebnisse (Wie gut sind die Ergebnisse von $\mathcal{B}$ wenn $\mathcal{S}$ benutzt wird?)
7. Robustheit (Wie gut kann $\mathcal{S}$ mit ungewöhnlichen Situationen umgehen?)

8. Zufriedenheit der Benutzer (Wie gern benutzt $\mathcal{B}$ das $\mathcal{S}$?)
9. Produktivität (Verbessert $\mathcal{S}$ die Arbeit von $\mathcal{B}$, d.h. seinen Output relativ zu seinem Input?)

Zu den Usability-Kriterien lässt sich abschließend sagen, dass sich ein Bewertungskriterium dadurch charakterisiert, dass seine anfängliche Abstraktheit mit Indikatoren verbunden wird, die messbar gemacht werden sollen. Das Kriterium wird also mit beobachtbaren Phänomenen verbunden, die gemessen werden können um Daten für die Bewertung eines künstlichen Systems zu bekommen. Dies ist ein Vorgehen, dass aus Grundlagen des empirischen Forschens über den Begriff des *Operationalisierens* bekannt ist. Insofern handelt es sich bei der Wahl der Kriterien, die Usability erfassen sollten, um nichts anderes als um Operationalisierungsideen. Durch eine analytische Definition eines Bewertungskriteriums – wie z.B. Funktionalität durch die Frage *Wie viele verschiedene Aufgaben kann $\mathcal{B}$ mit $\mathcal{S}$ erledigen?* (Jameson, 2001) – kann nachvollzogen werden, ob die operationalen Indikatoren sinnvoll sind oder nicht. Davon unbeschadet können andere Wissenschaftler zu anderen Konzepten gelangen, welche dann in der Forschungs- und Anwendungspraxis miteinander konkurrieren können (Trimmel, 1997). Um ein Beispiel für konkurrierende Konzepte zu geben, fasst Preece (1993: 47) einzelne Kriterien von Jameson (2001) wie *Schnelligkeit der Ausführung* und *Anzahl der ernsthaften Fehler* zusammen mit der Anzahl der erfolgreich gelösten Aufgaben unter *Throughput* zusammen, das die Leichtigkeit des Lernens charakterisieren soll.

1.4 Evaluation multimodaler und mobiler Systeme

Jeder Fortschritt in der Interaktion mit künstlichen Systemen bringt neue Herausforderungen für Usability-Studien. Zu den Fortschritten der letzten Zeit zählt sowohl die Multimodalität wie auch Mobilität. Multimodalität bringt zugleich die Möglichkeiten einer besonders vielfältigen Art, mit dem System Informationen auszutauschen, wie auch Schwierigkeiten mit sich. Denn für die erkennenden Systeme werfen sich Fragen auf, wie die unterschiedlichen modalen Komponenten gegeneinander abgewogen werden können (Beringer u.a., 2002a). Was die Evaluationsstudien solcher Systeme anbetrifft, so sprechen Beringer und ihre Kollegen (2002b) an dieser Stelle ebenfalls die Schwierigkeit der Auswertung mulimodalen Inputs oder Outputs und die Schwierigkeit der Synkronizität an. Das System muss dementsprechend fähig sein, auch mit einem nicht-synchronen mulimodalen Input zurrecht zu kommen. Sie präsentieren die Methode PROMISE, die eine Evaluationsprozedur für mulimodale interaktive Systeme beschreibt (Behringer, 2002a und 2002b). Diese Methode stellt eine Liste von Qualitäts- und Quantitätsmaßen dar und bezieht sich speziell auf die Evaluation von Systemen wie SmartKom, die sowohl mit gesprochener Sprache wie auch anderen physischen

Input gesteuert werden. Der Evaluationsrahmen, den Behringer und Kollegen (2002a) vorstellen, vermittelt einen Eindruck von großer Komplexität, da viele einzelne Kriterien eine Rolle spielen.

Ähnlich komplex erscheint die Aufgabe einer Evaluierung von mobilen Systemen. Denn werden künstliche Systeme mobil, spielt die sie umgebende Umwelt eine große Rolle. Es muss daher beachtet werden, in wie weit die Systeme auf die Beschaffenheit und Veränderungen der Umwelt eingehen können. Im Hinblick auf Evaluationsstudien spricht Jameson (2002) Usability-Anforderungen an, die durch die Verknüpfung beider Fähigkeiten, der Multimodalität und Mobilität, entstehen. Nach Jameson (2002) schafft die Fähigkeit zur Multimodalität eine besondere Form der Flexibilität (*Flexibility afforded by multimodality*). Allerdings erschwert die Multimodalität auch die Datensammlung für die Evaluationsstudien, da unterschiedliche, auch schwer aufnehmbare Informationen (wie die Gestik) zusammenfließen. Die Mobilität des Systems verursacht wiederum einen ständigen Wettbewerb um die Aufmerksamkeit der Benutzer zwischen dem System und der Umwelt. Die *System-environment competition* drückt also aus, dass die Aufmerksamkeit einer Benutzerin / eines Benutzers von anderen Reizen der Umgebung abgelenkt werden kann. Zum Beispiel kann die Aufmerksamkeit auf einem Flughafen durch einen plötzlichen Zeitmangel abgelenkt werden, da die Benutzerin oder der Benutzer eine Durchsage gehört hat und nun plötzlich aufbrechen muss. Diese ‚gestresste' Aufmerksamkeit kann sich sowohl auf die Perzeption wie auch die kognitive Verarbeitung und die physikalische Handlung der mit einem System Benutzer auswirken. Die Auswirkungen des System-Umwelt Wettbewerbs sind um so größer, wenn das System einen multimodalen, flexiblen Input erlaubt (Jameson, 2002). Das heißt, es wird noch schwieriger sein, die Reaktionen der Benutzer in solch einer Situation vorherzusagen, vor allem weil das Verhalten von Benutzern selbst in einfachen Situationen nicht immer rational und optimal ist (Jameson, 2002).

> [...] where a task analysis shows that users have some choices among alternative methods, their behavior is relatively hard to predict. If there is only one way of successfully performing a task, it is reasonable to assume that users will learn that method sooner or later and henceforth execute it in a more or less predictable fashion. Where free choices are available, various factors can determine which alternative a given user will select in a given situation [...] (Jameson, 2002).

Letztlich ist das Aussuchen einer Input-Methode selbst ein kognitiver Akt, der einer Anstrengung bedarf. Jameson (2002) fasst zusammen, dass wenn beide Variablen *System-environment competition* und *Flexibility afforded by multimodality* in einer Situation, in dem das System benutzt wird eine Rolle spielen, sich die Aufgabe sowohl für Entwickler und Evaluatoren wie auch für die Benutzer verkompliziert. Eine Lösung besteht darin, das Systems mit einer Fähigkeit auszustatten, die

perzeptuellen und kognitiven Resourcen einer Benutzerin / eines Benutzers in Abhängigkeit von einer Situation zu erkennen und sich ihnen anzupassen (Jameson, 2002). An dieser Stelle sagt Jameson (2002) erhebliche Schwierigkeiten für die Realisierung voraus: Was die Begrenzung der Aufmerksamkeitskapazität von Benutzern betrifft, so wird die Einschätzung es Systems immer fehlerbehaftet und konkurrierend mit situationsabhängigen Zielen sein. Ein Vorschlag ist daher, den Benutzern die Kontrolle über die Art und zeitliche Regulierung der Systemadapation zu überlassen. Jameson und Schwarzkopf (2002) nennen diese Fähigkeit (die zwar von dem Benutzer gewollt und ausgeführt, aber von der Systemseite ermöglich wird) die *Controllability*[2]. Die Kontrollierbarkeit eines Systems macht es möglich, es an die jeweilige Benutzerin / den jeweiligen Benutzer anzupassen. Jameson (2001b) spricht in Folge dessen auch von *Personalization*. Dieses Kriterium referiert auf eine weite Klasse von Methoden, mit welchen interaktive Systeme eine oder mehrere Bedingungen erfüllen: (a) sich an verschiedene Eigenschaften (wie Interessen, Wissen oder gegebenen Kontext) ihrer Benutzer anpassen; (b) erlauben den Benutzern das Aussehen oder die Art des Arbeiten zu verändern; (c) verhalten sich wie Menschen, in dem sie z.B. natürliche Sprache verstehen oder in menschenähnlicher Form erscheinen.

1.5 Evaluation von Systemen in triadischen Kontexten: Interactionability

Im Folgenden beziehen sich unsere Überlegungen auf das System COUSIN (COmmUnicator for Situated Interactions) im SFB 360. Versucht man die Besonderheiten dieses künstlichen situierten Kommunikators herauszustellen, so steht die Interaktion des Systems mit dem Menschen im triadischen Kontext im Vordergrund (Bauckhage u.a., 2002). Der triadische Kontext ergibt sich aus der Tatsache, dass der Mensch mit dem System über Objekte und Ereignisse kommuniziert, die in der Welt stattfinden. Die Referenz auf die Objekte (im SFB-Szenario sind das Baufix®-Einzelteile und Aggregate) und Ereignisse findet durch Sprache und Bild statt. Das heißt, durch die Kombination von Spracherkennung, Dialogkompetenz und Bilderkennung ist es möglich, dass das System Konstruktionsanweisungen des Menschen versteht, die gemeinten Objekte erkennt, und den Instruktionen folgt. Das System arbeitet also mit multimodalem Input, was bei einer Evaluation die oben erwähnten Probleme mit sich bringt. Zusätzlich kommt jedoch die Tatsache hinzu, dass die Aufgaben in einem triadischen Kontext stattfinden. Für eine Evaluation in triadischen Kontexten gibt es bislang keine detaillierten Vorgehensweisen. Betrachtet man die Usability-Kriterien, so wird deutlich, dass sie nicht allen Anforderungen, die an ein System im triadischen Kontext gestellt werden können, genügen. Nichtsdestotrotz können sie eine

[2] „A key usability issue with systems that adapt to their users is controllability: the ability of the user to determine the nature and timing of the adaptation“ (Jameson & Schwarzkopf, 2002).

Pilotstudie leiten, die eine Grundlage für weitere, spezifischere Kriterien sein kann. Daher war das Vorgehen der Pilotstudie, die bestehenden Überlegungen zu Evaluationskriterien aufzunehmen und sie auf das komplexe System zu übertragen.

Eine große Herausforderung wird in der Evaluation des künstlichen situierten Kommunikators darin bestehen, den zentralen Begriff der *Interaktion* zu charakterisieren und zu berücksichtigen. In der Literatur wird bereits ein der Idee nach verwandter Terminus verwendet: *Interaction Design* (Preece, 2002). Auch wenn sich dieser Begriff darauf bezieht, wie interaktive Produkte entworfen werden können, die Allltags- und Berufstätigkeiten des Menschen unterstützen, beschränkt er sich bislang auf Produkte, die eine Kommunikation über eine Bildschirmoberfläche möglich machen und wird somit in dyadischen Kontexten benutzt. Der Begriff der Interaktion soll jedoch auch die sozialen und kommunikativen Aspekte hervorheben: z.B. Benutzung der natürlichen Sprache, Rückmeldung des Systems, Monitoring des Aufmerksamkeitsfokusses. Eine Diskussion darüber, wie *Interactionability* greifbar gemacht werden kann, wird für eine ganze Klasse von Systemen relevant sein, die Dautenhahn und Kollegen (2002) als *social robots* bezeichnen. Damit sind Systeme gemeint, die mit Menschen interagieren. Die Systeme können im Ausmaß ihrer Interaktionskapazitäten variieren. Ein Roboter, der größere senomotorische Fähigkeiten vorweist, kann mehr mit seiner Umwelt interagieren. Dautenhahn und Kollegen (2002) übertragen die Fähigkeit der Interaktion sowohl auf die soziale wie auch die physikalische Welt: Je nachdem wie ausgeprägt die Fähigkeit ist, sprechen sie von einem kleineren oder größerem Grad an *embodiment*. Die Variable *Embodiment* soll die Interaktionskapazität eines Systems mit seiner Umwelt charakterisieren. Sie wird dadurch gemessen, wie stark sich das System und die Umwelt beeinflussen können:

> A system $\mathcal{S}$ is embodied in an environment $\mathcal{E}$ if perturbatory channels exist between the two. That is, $\mathcal{S}$ is embodied in $\mathcal{E}$ if for every time t at which both $\mathcal{S}$ and $\mathcal{E}$ exist, some subset of $\mathcal{E}$'s possible states with respect to $\mathcal{S}$ have the capacity to perturb $\mathcal{S}$'s state, and some subset of $\mathcal{S}$'s possible states with respect to $\mathcal{E}$ have the capacity to perturb $\mathcal{E}$'s state.

Auf eine detaillierte Diskussion dieses Kriteriums kann hier nicht eingegangen werden. Es sei lediglich darauf verwiesen, dass zwei Messkriterien genannt *perturbatory bandwidth* und *structural variability* die Stärke der Beeinflussung ausmachen (Dautenhahn u.a., 2002).

Dieser kurze Verweis auf Embodiment als Interaktionskapazität eines Systems macht jedoch deutlich, dass wenn man die Fähigkeit der *Interaktion* eines Systems betrachtet, zunächst ein vielschichtiges Problem erscheint, das weiterer Forschung bedarf, um zum konkreten Kriterium zu werden. Usability-Kriterien können für eine Bewertung von social robots nur einen Ausgangspunkt bilden, d.h. die Benutzbarkeit solcher Systeme wird mit Sicherheit eine zentrale Rolle spielen. Doch ob und wie

solche Systeme in ihrer Umwelt aufgenommen werden, hängt von Kriterien ab, die über die Usability hinausgehen. Daher ist das Nachdenken über weitreichende Kriterien wie Embodiment notwendig. Auch können Kriterien, die im Moment in anderen Bereichen entstehen, auf künstliche multimodale Systeme übertragen werden. So erwähnt Preece (2002) ganz kurz *Socialability* als ein Kriterium, das auf den Bereich des Internets angewandt messen soll, wie gut die soziale Interaktion unterstützt wird. Als Beispiel gibt Preece (2002: 417 f.) an, dass mit diesem Kriterium vor allem online Gruppen evaluiert werden können, die eine Möglichkeit zum Austausch und Kommunikation unter den Benutzern ermöglichen. Auf die social robots übertragen, könnte das Kriterium generell die Fähigkeit zu erfassen suchen, in wie weit die sozialen Bedürfnisse einer Benutzerin / eines Benutzers in einer Kommunikation unterstützt werden.

2. Beschreibung des Systems COUSIN

Der prototypische künstliche Kommunikator COUSIN (COmmUnicator for Situative INteraction), auf den wir uns hier im Speziellen beziehen, ist in einem Konstruktionsszenario situiert und unterstützt die flexible Aggregierung einfacher Elemente eines Spielzeugbaukastens zu komplexeren Einheiten: Den Anweisungen eines menschlichen Instrukteurs folgend manipuliert er in seiner Umgebung verteilte Objekte und montiert sie zu Modellen von z.B. Flugzeugen oder Lastwagen.

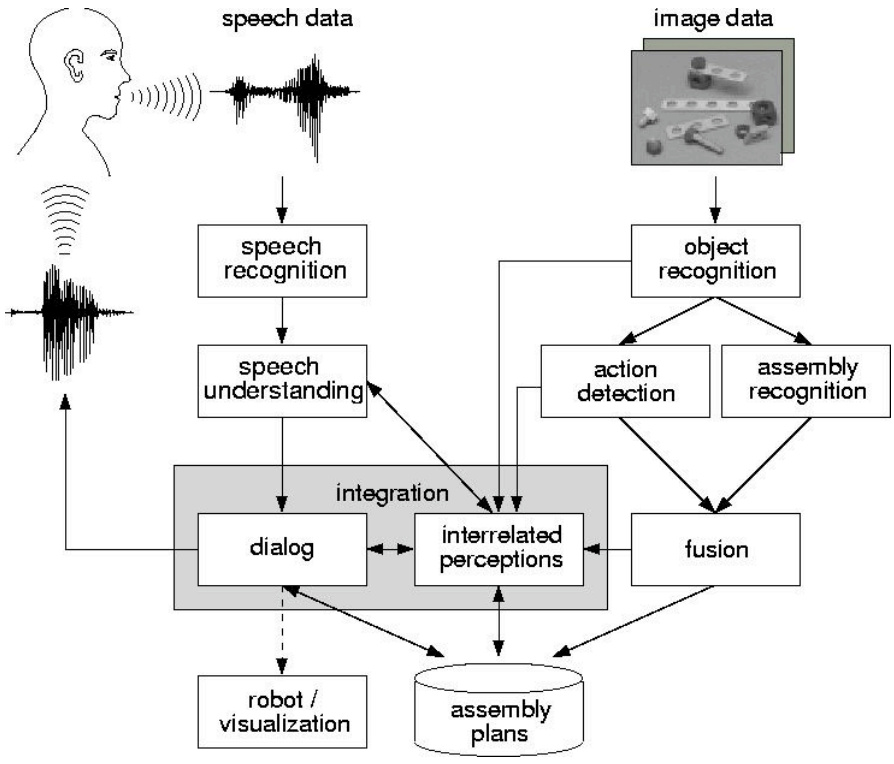

Abbildung 1: COUSIN – Systemübersicht

Abbildung 1 zeigt den Aufbau des perzeptiven Front-Ends unseres künstlichen Kommunikators. Es besteht im wesentlichen aus zwei Strängen signalverarbeitender Module, die in einer Integrationskomponente zusammengeführt werden. Im links abgebildeten Sprachverarbeitungsstrang erfolgt zunächst eine sprecherunabhängige, robuste Erkennung deutscher Spontansprache. Die Besonderheit des hierzu realisierten Moduls besteht darin, dass es statistische Spracherkennung durch Hidden-Markov-Modelle mit einer wissensbasierten Komponente integriert (Wachsmuth u.a., 1998).

Ein nachgeschalteter LR(1)-Parser, der auf einer Unifikationsgrammatik basiert, ermöglicht die Auflösung von Reparaturen oder Nachträgen und dient als Sprachverstehenskomponente des Systems (Kronenberg & Kummert, 1999). Zur Erkennung von elementaren Objekte und Objektaggregaten in Bildfolgen werden semantische Netzwerke eingesetzt, die domänenspezifisches Objektwissen modellieren (Kummert u.a., 1998, Bauckhage u.a., 2000). Gleichzeitig erkennt ein regelbasierter Algorithmus elementare Montageoperationen; durch Fusion der dabei anfallenden Ergebnisse mit denen aus der Aggregaterkennung lassen sich Baupläne der in der Szene sichtbaren Aggregate ableiten, so dass umfangreiches Wissen über zuvor unbekannte komplexe Objekte aus Bilddaten erlernt werden kann (Bauckhage u.a., 1999). Die Resultate aus Bild- und Sprachverarbeitung laufen in einem Modul zusammen, das Bayes-Netze einsetzt, um integrierte Interpretationen zu berechnen. Gleichzeitig realisiert eine zustandsbasierte Dialogkomponente robuste Strategien zur Dialogführung und ist verantwortlich für die Generierung von Rückfragen im Falle unverständlicher oder mehrdeutiger Anweisungen an das System (Brandt-Pook u.a., 1999). Das Integrationsmodul, das Bild- und Sprachverarbeitungsresultate zusammenführt und aufeinander bezieht, nimmt in der skizzierten Architektur eine zentral Rolle ein. Aufgrund von Schnittstellen zur Dialogkomponente sowie zu einer Datenbank, die aus Bildern extrahierte Aggregatstrukturen speichert, ist es möglich, sowohl auf explizite wie auch auf implizite Art und Weise Benennungen für Aggregate einzuführen, die im Verlauf eines Konstruktionsdialoges entstehen. Um Tests des perzeptiven Front-Ends zu erleichtern, ist es möglich, die Robotikkomponente durch ein Visualisierungsmodul zu ersetzen, das die aktuelle Interpretation, die das Systems über Objekte und Ereignisse in seiner Umgebung hergeleitet hat, mit Mitteln der virtuellen Realität veranschaulicht. Technische Einzelheiten zu den einzelnen Komponenten und ihrer Verschaltung zu einem multimodalen kognitiven System können in (Bauckhage u.a., 2001) nachgelesen werden.

Die Leistung der hier skizzierten Architektur zur Integration verschiedener perzeptiver Modalitäten liegt darin, dass sie eine intuitive und natürliche Mensch-Maschine Kommunikation ermöglicht, die in Ansätzen an die Kommunikation zwischen Menschen in dieser speziellen Aufgabe erinnert. So mag bei der in Abbildung 2

gezeigten Szene der Instrukteur, die lange Leiste in der Mitte des Bildes als die *7-Loch-Leiste*, *die lange Leiste, die Leiste vor dem Heckteil* oder *das lange Ding* bezeichnen. Laufen alle vorverarbeitenden Berechnungen korrekt ab, bildet das System jede dieser zunehmend abstrakteren Bezeichnungen auf das intendierte Objekt ab.

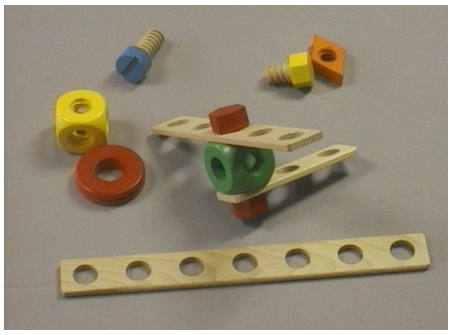

Abbildung 2: Beispiel einer Szene in COUSIN

Wie die Auflösung von Raumrelationen, unterspezifizierten oder vagen Äußerungen durch die Integration von visueller Objekterkennung und Sprachverarbeitung konkret erfolgt, ist im Wachsmuth und Sagerer (2002) genauer dargestellt. Zu betonen sei, dass die Integration von Sprach- und Bildverstehen stellt eine Grundlage natürlicher Kommunikation darstellt, die neuerdings auch in der Forschung zur Mensch-Maschine Kommunikation vermehrt Beachtung findet. Dabei gehen bekannte technische Integrationsansätze wie etwa regelbasierte Übersetzer (Takahashi u.a., 1998) oder integrierte wissensbasierte Systeme (Bronstedt u.a., 1998) bislang davon aus, dass alle Erkennungsergebnisse korrekt sind und dass sich verbale eindeutig in visuelle und visuelle sich eindeutig in verbale Beschreibungen transformieren lassen. Solcherlei Annahmen gelten allerdings nicht generell. In unserem Szenario beispielsweise wird die Bedeutung der Bezeichnung *lange Stab* von der aktuellen Szene abhängen. Je nach visuellem Kontext kann damit eine 3-, 5- oder 7-Loch-Leiste oder auch etwas ganz anderes gemeint sein. Andererseits kann es zu einer *5-Loch-Leiste* je nach Szenenkontext hunderte verschiedener verbaler Referenzen geben und die vom Sprecher gewählte wird von seinen oder ihren Vorlieben abhängen.

Wie aus der kurzen Darstellung ersichtlich wird, stellt der künstliche Kommunikator COUSIN ein hochkomplexes System dar, dessen kommunikative Leistungen zum jetzigen Zeitpunkt noch nicht in ihrer Vollständigkeit beschrieben werden können. Fähigkeiten wie die Integration von Sprach- und Bildverarbeitung, sowie das Führen einfacher Klärungsdialoge bzw. Stellen von Rückfragen, wenn keine konsistente Interpretation der aufgenommenen Bild- und Sprachsignale möglich ist, machen das System flexibel in der Kommunikation. Selbst wenn zu einem Zeitpunkt in einer längeren Interaktion keine zufrieden stellende Systemreaktion erzielt werden kann,

so ist es durchaus denkbar, dass sich durch Nachfragen und daraus ergebenden Klärungen doch noch die gewünschte Reaktion einstellt. D.h. dass zur Bewertung der Leistungsfähigkeit des Systems längere Interaktionsphasen beobachtet und beurteilt werden sollten. Für diese Zwecke gibt es jedoch keine allgemein akzeptierte Methodik. Hinzu kommt die Tatsache, dass multimodale Systeme, an denen man derartige Probleme untersuchen könnte, bisher kaum verfügbar waren. Deshalb stellt zur Zeit die Frage der Evaluation komplexer kognitiver Systeme ein eigenes Forschungsfeld dar. Der nächste Abschnitt beschreibt, wie Interaktionsexperimente in Bezug auf einige Usability-Kriterien ausgewertet werden können, um erste Schritte in Richtung der Evaluation komplexer Systeme zu machen.

3. Pilotstudie

3. 1 Ausgewählte Kriterien

Oben wurde erwähnt, dass Usability-Kriterien einen Ausgangspunkt für Evaluationsstudien mit multimodalen künstlichen Systemen in triadischen Kontexten darstellen können. Im Folgenden wird erläutert, welche Usability-Kriterien für eine Pilotstudie ausgewählt wurden und wie die Studie durchgeführt wurde. Im Vorfeld ist es jedoch wichtig zu betonen, dass vor der experimentellen Evaluation bereits andere Evaluationsmethoden auf den künstlichen situierten Kommunikator angewandt wurden (s. z.B. Bauckhage u. a., 2002). Sie betrafen hauptsächlich die Leistung der einzelnen Module entlang einzelner Aufgabenschritte. Was jedoch bislang fehlte, war ein Ansatz, der die Mensch-Maschine Kommunikation in der Gesamtheit der Aufgabe bewerten sollte. Da sich die Pilotstudie an zwei Kriterien der Usability-Studien orientiert, stand die Leistung der Benutzerin / dem Benutzer mit dem System im Fokus der Untersuchung. In Abhängigkeit von einem unterschiedlichen Grad an der Komplexität der Aufgaben, wurde die *Schnelligkeit der Ausführung* und die *Funktionalität des Systems* getestet. Die Aufgabe bestand darin, mit dem System ein Aggregat zu bauen. Ihre Komplexität wurde durch zwei Faktoren festgelegt: die Einzelteile des Zielobjekts, d.h. je mehr Schrauben und Einzelobjekte das Zielobjekt enthielt, desto komplexer war die Aufgabe. Der zweite Faktor der Komplexität war die Vorgabesituation. Für sie galt: je mehr Objekte sich in der Szene befinden, desto komplexer wird die Aufgabe.

Die *Schnelligkeit der Ausführung* wurde durch die Zeit gemessen, die für eine Aufgabe benötigt wurde. Weiterhin gab die Qualität der Ergebnisse in Abhängigkeit von einem unterschiedlichen Grad an Komplexität Aufschluss über die *Funktionalität des Systems*, da somit getestet wurde, wie viele verschiedene Aufgaben vom System erfüllt werden können.

Neben der Zeit, die für die Aufgaben benötigt wurde und der Qualität der Ergebnisse als abhängige Variable wurden ebenfalls zwei weitere Variablen beobachtet: die

Anzahl der Instruktionen, die die Testpersonen benötigten, um zum Ziel zu gelangen wie auch die Anzahl der Nachfragen seitens des Systems. Das System hat nämlich die Fähigkeit nachzufragen, wenn die Referenz unklar erscheint. Diese Beobachtungen sollten Hinweise darauf geben, wie das System in die Qualität der Ergebnisse selbst involviert ist.

3.2 Vorgehen und Design

Die Pilotstudie wurde mit acht computererfahrenen Testpersonen durchgeführt, davon waren 2 weiblich und 6 männlich. Das Experiment bestand für jede Testperson aus zwei Aufgaben, ein unterschiedlich komplexes Zielaggregat mit dem System zu bauen (Abb. 3). Bei der weniger komplexen Aufgabe handelte es sich um ein Aggregat mit einer Schraube und einem Einzelteil (s. Abb. 3a); bei der komplexeren Aufgabe um ein Aggregat mit zwei Schrauben und zwei Einzelteilen (s. Abb. 3b).

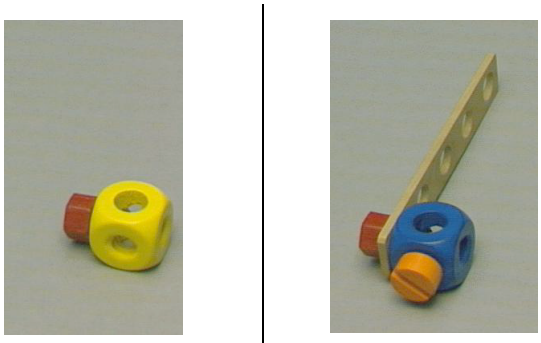

Abbildung 3: Zwei unterschiedlich komplexe Zielaggregate

Das Zielobjekt sollte zusammen mit dem System in zwei unterschiedlich komplex gestalteten Vorgabesituationen gebaut werden. Für die Variation in der Vorgabesituation waren einmal sechs Objekte, das andere Mal zehn Objekte in der Szene. Insgesamt musste jede Testperson vier Mal ein Aggregat bauen lassen (Abb. 4).

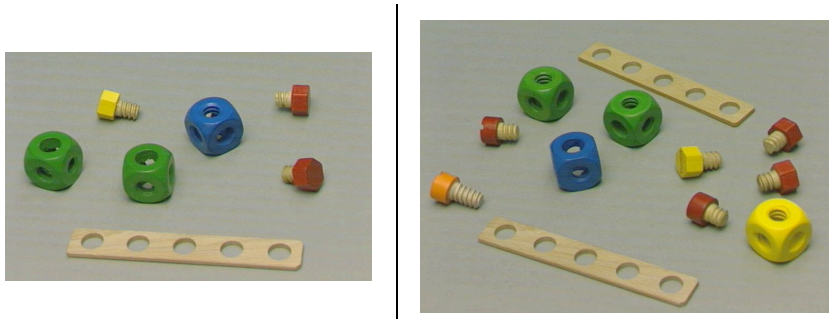

Abbildung 4: Unterschiedlich komplexe Vorgabesituationen

Die Testpersonen wurden zunächst über die sprachlichen Möglichkeiten des Systems instruiert. Dieses Vorgehen erwies sich schon im Vorfeld als notwendig, da auch computererfahrene Testpersonen nicht wussten, wie viel Leistung das System bietet. Anschließend bekamen die Testpersonen das Ziel der Aufgabe auf einem Bild gezeigt. Jede Testperson baute mit dem System zwei Aggregate in jeweils 2 unterschiedlich komplexen Situationen, wobei die Reihenfolge der Aufgaben und Situationen durchvariiert wurde (s. Tabelle 2 linke Seite).

Die Dauer der einzelnen Aufgaben beinhaltete auch die Zeit, die das System benötigt, um die Aufgabe zu erfüllen. Den Beginn markierte daher die erste Instruktion der jeweiligen Testperson, das Ende wurde durch die Erkennung der Objekte auf dem Tisch durch das System bestimmt.

Um letztlich die Qualität der Ergebnisse zu testen, wird eine Aufgabe als Erfolg kodiert, wenn das System das gewünschte Aggregat baut und zum Schluss als solches erkennt.

3.3 Ergebnisse

Die Tabelle 2 gibt einen Überblick über das Design wie auch die Ergebnisse der Pilotstudie. Die Zahlen, die als Ergebnisse aufgelistet sind (s. Tab. 2 rechte Seite) sind Durchschnittswerte.

Tabelle 2: Experimentelles Design und Ergebnisse (Durchschnittswerte) der Pilotstudie

Experimentelles Setting			Ergebnisse			
Komplexität der Vorgabesituation (Objekte in der Szene)	Komplexität des Aggregats					
	Einzeteile im Aggregat	Schrauben im Aggregat	Zeit in sek.	Anzahl der Instruktionen	Anzahl der Nachfragen	Erfolgsrate
6	2	1	49	2.6	0,4	87,5%
6	4	2	184	8.0	2,2	75%
10	2	1	57	3.1	1,1	100%
10	4	2	220	9.8	3,0	62,5%

Wegen der wenigen Testpersonen kann die statistische Analyse (t-Test) der Ergebnisse nur richtungsweisend gedeutet werden. Wie jedoch erwartet zeigte sich im Bezug auf die Schnelligkeit der Ausführung, dass wenn ein Objekt mit mehreren Schrauben und Einzelteilen gebaut werden soll – unabhängig von der Vorgabesituation – signifikant mehr Zeit benötigt ($p < 0{,}001$) wird.

Die weiteren Ergebnisse, die im Folgenden dargestellt werden, umfassen jeweils nur (wenn nicht explizit anders angemerkt) die gelösten Aufgaben. Bei ungelösten Aufgaben kam es vor, dass die Testpersonen auf Grund der misslungenen Kommunikation viele verschiedene Instruktionen versuchten (je nach Geduld und Ehrgeiz), worauf das System mit vielen Nachfragen reagierte. Daher ist eine objektive Angabe darüber, wie viele Instruktionen in einer misslungenen Aufgabe gebraucht werden unmöglich, ebenso wie die Angabe darüber, ab wann eine Testperson erkennt, dass die Kommunikation nicht mehr funktioniert.

Parallel zu der wachsenden Zeit steigt mit der Anzahl an Schrauben und Einzelteilen im Zielobjekt die Anzahl der Instruktionen der Testpersonen an ($p < 0{,}01$). Es fällt dabei auf, dass wie auch die Schnelligkeit der Ausführung, dieses Ergebnis von der Vorgabesituation unabhängig ist. Somit scheint die Anzahl der Objekte in der Vorgabesituation die Komplexität der Aufgabe nicht zu beeinflussen — wenn man die Komplexität als eine Aufgabe, die größeren Zeitaufwand und mehr einzelne Aufgabenschritte erfordert, charakterisiert. Vielmehr scheint die Komplexität der Aufgabe durch die Anzahl der Einzelteile und Schrauben im Zielobjekt bedingt zu werden. Dass die Anzahl der Objekte in der Szene keine Auswirkung auf die Komplexität der Aufgabe hat, bestätigt auch die Analyse der Qualität der Ergebnisse: Betrachtet man die gelösten und ungelösten Aufgaben in den jeweiligen Bedingungen, so kann kein Unterschied vermerkt werden. Die einzige auffindbare Tendenz betrifft die Vorgabesituation mit zehn Objekten, in der das Bauen eines einfaches Objektes eine größere Erfolgsrate zu haben scheint als das Bauen eines Objektes aus 2 Schrauben und 4 Einzelteilen ($p = 0{,}08$). Es heißt also, dass auch wenn die Prozentzahlen (s. Tab. 2, Spalte „Erfolgsrate“), die sich aus der geringen Teilnehmerzahl ergeben, deutliche Unterschiede zeigen, sie dennoch statistisch nicht relevant sind.

Eine weitere Analyse zeigt, dass die Vorgabesituation auch auf eine weitere Variable einen geringen Einfluss zeigt. Hinsichtlich der Nachfragen des Systems konnte lediglich in der Vorgabesituation mit sechs Objekten beobachtet werden ($p < 0{,}05$), dass das Nachfragen ansteigt, wenn mehrere Schrauben und Einzelteile im Zielobjekt verbaut werden. In weiteren Situationen zeigte die Analyse keinen Unterschied in der Anzahl der Systemnachfragen ($p < 0{,}094$).

Dialogsystem

Welche Interaktionsschritte für das Dialogsystem schwierig waren, konnte daran beobachtet werden, dass sich einige Probleme bei unterschiedlichen Teilnehmern wiederholten. Diese wurden offensichtlich durch die Möglichkeit hervorgerufen, in dieser Aufgabe spontan zu kommunizieren.

Als wichtigstes Problem sei hier die Deixis genannt. Diese fiel im Verlauf eines Dialogs auf: Nachdem die Benutzer ein Objekt aus der Szene ausgewählt hatten und

sprachlich darauf referierten z.B. „Nimm die rote Schraube“, bestätigte das System diese Aktion. Auf diese Bestätigung hin instruierten wiederum die Benutzer, indem sie sich auf das bekannte bezogen und z.B. Pronomina für die Objektnamen (hier Schraube) benutzten: „Stecke sie in die Leiste“. Das System jedoch hatte Schwierigkeiten, das Pronomen zu ‚verstehen’, was häufig zu Missverständnissen und letztlich zum Dialogabbruch führte.

Eine weitere Ausprägung von Deixis ist z.B. dieses Interaktionsbeispiel ($\mathcal{B}$ = Benutzerin / Benutzer, $\mathcal{S}$ = System):

(1) $\mathcal{B}$: Nimm die kleine rote Schraube!

(2) $\mathcal{S}$: Ok. Ich nehme die rote Schraube. Was soll ich als nächstes tun?

(3) $\mathcal{B}$: Schraube darauf den grünen Würfel!

Hier steht „darauf“ für „auf die Schraube“, die bereits im ersten Schritt des Dialogs eingeführt wurde. Auch in dieser Situation fragte das System nach, weil „darauf“ unbekannt und / oder ungenau war. Somit war die Instruktion für das System unvollständig. Der Benutzer korrigierte daraufhin seine Instruktion und äußerte „Schraube den grünen Würfel auf die rote Schraube!“, was letztlich zu einer korrekten Lösung der Aufgabe führte. Wie aus der Formulierung deutlich wird, begibt sich der Benutzer mit dieser Reparatur auf eine Ebene, auf der er nicht von einem sprachlich referenziellen, sondern einem aktionalen Wissen ausgehen kann. Eine weitere Vermutung ist in diesem Beispiel, dass die Rückmeldung des Systems in (2), d.h. „Was soll ich als nächstes tun?“ die Inferenz, die die Äußerung des Benutzers in (3) beinhaltet, provoziert.

Bilderkennung

Die Leistung dieser Integrationskomponente lässt sich durch die Anzahl korrekt etablierter Abbildung zwischen Bild- und Sprachsignal charakterisieren. Diese hängt zum einen wesentlich von den Erkennungsergebnissen der vorverarbeitenden Module ab. Andererseits wirkt sich aber auch die Sprachwahl des Sprechers auf das Ergebnis aus.

Wie viele Computer-Vision Systeme realisiert auch unser System die visuelle Erkennung von Objekten als einen hierarchischen Prozess: Zunächst werden in den von der Kamera aufgenommenen Bildern homogene Farbregionen segmentiert. Diesen Regionen werden dann Objektbezeichner zugeordnet, und schließlich werden zweidimensionale Cluster dieser Objekterkennungsergebnisse daraufhin untersucht, ob sie ein Aggregat abbilden. Gleichzeitig protokolliert das System, ob (und wenn ja, welche) Objekte aus der Arbeitsumgebung entfernt werden bzw. neu in dieser auftauchen. Nimmt man an, dass solche Ereignisse nur dann auftreten, wenn eine Konstruktionshandlung stattfindet, so lässt sich auch aus der Beobachtung aufgenommener bzw. abgelegter Bauteile schließen, welche Aggregate bzw. welche

Aggregatstrukturen im Laufe einer längeren Interaktion mit dem System konstruiert werden (Bauckhage u.a., 1999b).

Leider ist es durchaus möglich, dass jeder dieser Verarbeitungsschritte fehlerhafte Ergebnisse liefert. Zwar wurden sowohl für die Erkennung elementarer Bauteile als auch für die Erkennung von Aggregaten hohe durchschnittliche Erkennungsraten gemessen (92 % bzw. 82 %), die Erfahrung zeigt aber, dass selbst derart zuverlässige Ergebnisse keine reibungslose Interaktion mit unserem System garantieren. Schon kleinste Veränderungen der Beleuchtungsverhältnisse in der Arbeitsumgebung oder einfaches Kamerarauschen führen dazu, dass in aufeinander folgenden Bildern selbst bei statischen Szenen nur selten absolut identische Regionen segmentiert werden. Dies wiederum kann zu einer Verkettung von Fehlinterpretationen seitens des Systems führen: Die Merkmale, anhand derer Objekte klassifiziert werden, degenerieren, so dass die Ergebnisse der Objekt- und Aggregaterkennung von einem zum nächsten Bild nicht reproduzierbar sind; Weitere Konsequenzen betreffen die Module zur Handlungsbeobachtung. Da diese von der Heuristik ausgehen, dass Objekte, die gegriffen wurden, nicht mehr sichtbar sind, werden Objekte, die über einen längeren Zeitraum sichtbar waren, aber plötzlich nicht mehr im Bild gefunden werden, als gegriffen verstanden. Dies bewirkt, dass die Interpretation, die das System von seiner Umgebung aufbaut, nicht mehr mit der tatsächlichen Situation übereinstimmt. Folglich kann es durchaus passieren, dass selbst Anweisungen in Klärungsdialogen nicht mehr aufgelöst werden können und das System in einen Zustand vollständiger Konfusion gerät. Oftmals hilft hier zwar die Anweisung *Stop!*, durch die das System veranlasst wird, die Dialoghistorie zu löschen. Wenn aber (z.B. wegen sehr ungünstiger Beleuchtungsverhältnisse) die Ergebnisse der Bildverarbeitungskomponenten derart unbeständig sind, dass permanent fehlerhafte Interpretationen aufgebaut werden, kann auch das wiederholte Zurücksetzen des Dialogmoduls keine robuste Interpretation der Konstruktionssituation garantieren.

3.4 Diskussion

Diskussion der Ergebnisse

Die Ergebnisse der Pilotstudie geben Aufschluss darüber, ob der künstliche Kommunikator über ausreichend transparente Interaktionsfähigkeiten verfügt, die es einem Benutzer möglich machen, mit ihm zu arbeiten. Die Ergebnisse deuten darauf hin, dass der künstliche situierte Kommunikator eine hohe Zuverlässigkeit in seiner Leistung bei einfacheren Aufgaben zeigt, wenn er von computererfahrenen Personen instruiert wird — wie es in der Pilotstudie der Fall war. Bestehen die Zielobjekte, die mit dem System gebaut werden sollen, aus mehr als 2 Einzelteilen, steigt die Dauer der Ausführung, und die Anzahl der Instruktionen, die zur fertigen Konstruktion führt,

wächst. Zugleich bietet die steigende Anzahl der Instruktionen mehr Raum für Missverständnisse und fehlerhafte Leistungen des Systems. Inwieweit sich die Ergebnisse zur stabilen Leistung des Systems auch für 'roboterunerfahrene' Personen bestätigen lassen, wird eine weitere, umfangreichere Evaluationsstudie zeigen. Im Moment bleibt festzuhalten, dass der künstliche Kommunikator natürliche Sprache versteht und flexibel gegenüber Benutzern ist, die individuell bestimmen wie viele Instruktionen sie formulieren um eine Aufgabe zu erfüllen.

Aus der Tatsache heraus, dass bei den meisten Analysen nur die gelösten Aufgaben betrachtet wurden, ergibt sich der Nachteil, dass nicht eindeutig festgestellt werden kann, welchen Effekt die Nachfragen des Systems auf die Erfolgsquote haben. Dieser Effekt lässt sich nur indirekt vermuten: Dem Kriterium der Flexibilität folgend, scheint das System gegenüber den Vorgabesituationen in seiner Leistung stabil zu sein. Steigt die Komplexität der Aufgabe – wie in der Studie durch die Anzahl der Schrauben und Einzelteile im Zielobjekt bestimmt wurde – hat das System die Tendenz, mehr Nachfragen zu stellen. Dies deutet darauf hin, dass das System in der Lage ist, die Komplexität zu bewältigen und entstehende Ambiguitäten durch Nachfragen zu lösen. Das Nachfragen spielt also eine Rolle bei der Sicherung der Qualität der Ergebnisse.

Diskussion der Prozedur

Gerade für die Evaluierung der Fähigkeit des Systems, natürliche Sprache zu verstehen, ist es essentiell zu erkennen, dass die Testpersonen auf die gleiche Art und Weise in die Leistung des Systems eingeführt werden, so dass sie identische Informationen erhalten, die womöglich ihr Sprachverhalten verändern. Eine solche Einführung am Anfang des Experiments hat sich als schwierig erwiesen, da die Anleitung eine kommunikative Situation war, in der sich die Experimentatoren je nach Testperson anders verhielten. Auch wer immer für die gleiche Menge an Informationen sorgen möchte, kann nicht sicher gehen, ob sie auch auf die gleiche Art geäußert werden. Daher ist es notwendig, eine objektivere Maßnahme für die Einführung der Testpersonen zu finden. Eine Lösung wäre, den Testpersonen am Anfang des Experiments einen Film zu zeigen, der darüber informiert, was das System kann und woraus die Aufgabe besteht. Ein solcher Film wird zur Zeit für eine weitere umfangreichere Studie vorbereitet. In diesem Film erfahren die Testpersonen neben der Beschaffenheit und den Leistungen des Systems, wie man mit dem System kommunizieren kann. Die Prozedur hat sich bereits in den ersten Tests mit dem künstlichen situativen Kommunikator als nützlich erwiesen, da roboterunterfahrene Personen nicht wissen, wie sie ihr Sprach- und Denktempo modulieren sollen. Beispielsweise kam es vor, dass Testpersonen zu unsichere Instruktionen formulierten, die sich darin äußerten, dass das Sprechtempo langsam und die Artikulation gedehnt war. Bezogen auf das Denktempo war vielen Testpersonen nicht klar, dass sie über die Instruktionen nachdenken müssen, bevor

sie sie äußern. Jede Denkverzögerung, die sich z.B. im Gebrauch von Diskurspartikeln äußert, wird vom System als Störung wahrgenommen und bringt Probleme in der Dialogerkennung mit sich.

Diskussion der Evaluationskriterien

Wie bereits oben erwähnt, können die untersuchten Kriterien lediglich ein erstes Bild der Leistungen des Systems zeichnen. Das Kriterium der Funktionalität muss in weiteren Studien zum künstlichem situierten Kommunikator auf die sprachliche Interaktion erweitert werden. Dann können Evaluationsstudien darüber Aufschluss geben, auf Grund wie vieler natürlichsprachiger Instruktionen das System die Aufgaben erfüllt. Dazu sind mutigere Schritte in Richtung des Kriteriums der *Interactionability* notwendig. Denkbar ist im Moment den Fokus auf die Art und Weise zu richten, wie Testpersonen ihre Instruktionen an das System formulieren. D. h. es kann analysiert werden, wie komplex die Testperson ihre Instruktionen formuliert hat, und wie der Erfolg des Systems auf diese Instruktionen hin war. Für diese Analyse könnte eine Skala entwickelt werden, auf der eine einfache Instruktion nur aus einem Hauptsatz besteht (z.B. *Nimm die rote Schraube*), in dem ein Verb und ein Nomen genannt werden. Je komplexer die Sätze, desto mehr Konjunktionen und Pronomina enthalten sie (z. B. *Nimm sie und steck sie in die Leiste*). Das System hat die Möglichkeit, im Dialog durch das Nachfragen Informationen zu spezifizieren. Daher ist ein weiteres Ziel der Evaluation, die Anzahl der Nachfragen im Zusammenhang mit der Komplexität der Instruktionen auszuwerten um so Informationen über die Stärken und Schwächen des Dialogs zu bekommen.

4. Ausblick

Die Besonderheit des Systems im Sonderforschungsbereich besteht in der Möglichkeit einer Mensch-Maschine Kommunikation im triadischen Kontext: Der Mensch und die Maschine beziehen sich auf Objekte in der realen Welt, um aus diesen Aggregate zu konstruieren. Der Begriff der Interaktion steht daher im Zentrum der Kommunikation. Oben wurde angemerkt, dass ein Evaluationskriterium, das diesen Begriff zu erfassen sucht, auch die sozialen und kommunikativen Aspekte hervorheben muss wie z.B. Benutzung natürlicher Sprache, Rückmeldung des Systems, Monitoring des Aufmerksamkeitsfokusses. Das Nachdenken über ein mögliches und zugleich breites Evaluationskriterium für multimodale künstliche Systeme in triadischen Kontexten wie das der *Interactionability* bringt daher nicht nur neue Erkenntnisse für die Mensch-Computer-Studien. Auch für die Linguistik eröffnet sich ein Feld, das durch die Infragestellung der Natürlichkeit der Kommunikation sichtbar wird – was man an den Problemen des Dialogsystems erkennt. Im Moment ist es der Mensch, der die Strategien der Interaktion überblicken und die Klarheit der Referenz sichern muss. Aber auch zur Perspektive auf die Strategien, von denen der Mensch Gebrauch macht, gibt es erste Studien, die z.B. in Tenbrink, Fischer und

Moratz (2002) zu finden sind. Diese Studien geben Einblick darin, wie Menschen verfahren, wenn sie mit künstlichen Systemen kommunizieren. Es erscheint plausibel, dass erst die Erforschung der Art und Weise wie die Mensch-Maschine-Kommunikation – abhängig von der Aufgabenorientierung – funktioniert, uns Informationen darüber vermitteln wird, was die Menschen von einer Interaktion mit (mobilen) künstlichen Systemen erwarten, und was ihr so genanntes natürliches Verhalten in solch einer Situation ist. Aus diesen Informationen können Adaptationsanforderungen für das System gewonnen werden, da nicht nur die Strategien der Benutzer im Vordergrund stehen sollen. Wünschenswert wäre, dass auch das System in der Lage ist, sich an die Interaktionsart des Menschen anzupassen, wobei die Anpassung bei der Erkennung der individuellen Sprache erst anfangen kann. Ebenfalls dazu gehören Strategien der Dialogführung. Im Sonderforschungsbereich sind bereits einige Arbeiten zu Strategien der Verständnissicherung entstanden. Wichtig erscheint weiterhin, der Frage nachzugehen, wie solche Strategien auf ein künstliches System übertragen werden können, so dass roboterunerfahrene Benutzer spontan mit ihm kommunizieren können, ohne sie vorher über die kommunikativen Leistungen eines Systems aufklären zu müssen. Diesen Aspekt zu realisieren wäre erstrebenswert, vor allem weil die meisten künstlichen Systeme heutzutage lediglich mit ihren Autoren oder speziell trainierten Benutzern interagieren können. Im Moment berücksichtigt die Evaluation des künstlichen Kommunikators die Tatsache, dass Testpersonen von den kommunikativen Möglichkeiten des Systems erfahren, bevor sie gemeinsam mit ihm eine Aufgabe lösen. Im Rahmen der näheren Analyse der Interaktionsfähigkeit des Systems könnte erfasst werden, wie sich Benutzer in ihrer Kommunikationsweise an das System anpassen. Dabei ist von besonderem Interesse, wie sich eine Adaptationsphase zeitlich erstrecken kann und wie man diese bestimmen kann. Dass die Adaptationsprozesse auch in der Kommunikation zwischen Menschen innerhalb eines Konstruktionsszenarios gültig sind, zeigen Studien über menschliche Informationsverarbeitung (z.B. Vorweg & Voß, 2002). Daher wäre es von der linguistischen Seite wichtig, die Strategien der Verständnissicherung nicht nur in Situationen zu beobachten, in denen die Beteiligten feststellen, dass sie ein Verständigungsproblem haben, sondern auch als Mittel der kommunikativen Adaptation.

Zusammenfassend lässt sich sagen, dass eine Mensch-Maschine-Kommunikation im triadischen System neue Fragestellungen aufwirft. Die Komplexität dieser kommunikativen Situation zeigt, dass *interaktionable* Systeme einerseits eine Interaktion ermöglichen, die einer Mensch-Mensch-Kommunikation mehr gleicht und daher vielleicht intuitiver ablaufen kann. Andererseits eröffnen sie aber auch ein Spannungsfeld, auf dem der Aufwand und der Nutzen gegeneinander abgewogen werden müssen. Zu vermuten ist, dass jede Benutzung von Sprache einem (künstliche oder menschlichen) Partner gegenüber einer Adaptation benötigt. Je nachdem wie viel Anstrengung diese Phase erfordert, kann es sinnvoll sein, auf

bestimmte Modi des Inputs zu verzichten um Gesamtinteraktion dadurch zu erleichtern.

Referenzen:

Bauckhage, C. / Kummert, F. / Sagerer, G. (1999): Learning Assembly Sequence Plans Using Functional Models. In: Proceedings of the IEEE International Symposium on Assembly and Task Planning (ISATP'99): 1-7.

Bauckhage, C. / Fritsch, J. / Sagerer, G. (1999): Erkennung von Aggregaten aus Struktur und Handlung. In: Künstliche Intelligenz 3: 4—11.

Bauckhage, C. / Kronenberg, S. / Kummert, F. / Sagerer, G. (2000): Grammars and discourse theory to describe and recognize mechanical assemblies. In: Proceedings of the 8th International Workshop on Structural and Syntactic Pattern Recognition (S+SSPR2000). Springer Verlag: 173-182 (=Lecture Notes in Computer Science 1876).

Bauckhage, C. / Fink, G. A. / Fritsch, J. / Kummert, F. / Lömker, F. / Sagerer, G. / Wachsmuth, S. (2001): An Integrated System for Cooperative Man-Machine Interaction. In: Proceedings of the IEEE International Symposium on Computationa Intelligence in Robotics and Automation. Banff, Canada: 328-333.

Bauckhage, C. / Fritsch, J. / Rohlfing, K. J. / Wachsmuth, S. / Sagerer, G. (2002): Evaluating integrated speech and image understanding. In: Proceedings of the IEEE International Conference on Multimodal Interfaces (ICMI'02), 2002

Beringer, N. / Kartal, U. / Louka, K. / Schiel, F. / Türk, U. (2002a): PROMISE - A Procedure for Multimodal Interactive System Evaluation. In: Proceedings of the Workshop 'Multimodal Resources and Multimodal Systems Evaluation'. Las Palmas, Gran Canaria, Spain.

Beringer, N. / Louka, K. / Penide-Lopez, V. / Türk, U. (2002b): End-to-End Evaluation of Multimodal Dialogue Systems -can we Transfer Established Methods? In: Proceedings of the Third International Conference on Language Resources and Evaluation. Las Palmas, Gran Canaria, Spain.

Brandt-Pook, H. / Fink, G. A. / Wachsmuth, S. / Sagerer G. (1999): Integrated recognition and interpretation of speech for a construction task domain. In: Bullinger, H.-J. & Ziegler, J. (Hrgs.): Proceedings of the 8th International Conference on Human-Computer Interaction. München: 550-554.

Bronstedt, T. / Larsen, L.B. / Manthey, M. / McKevitt, P. / Moeslund, T. / Olesen, K.G. (1998): The intelligent workbench -- a generic environment for multimodal systems: In: Proceedings of the International Conference on Spoken Language Processing (ICSLP): 273—276.

Dautenhahn, K. / Ogden, B. / Quick, T. (2002): From embodied to socially embedded agents – implications for interaction-aware robots. *Cognitive Systems Research*.

Dix, A. / Finlay, J. / Abowd, G. / Beale, R. (1998): Human-computer interaction. Prentice Hall Europe.

Jameson, A. (2001) Human-computer interaction / interactive systems (slides from class 1, 11. September 2001). Retrieved Mai 9, 2002, from German Research Center for Artificial Intelligence GmbH Web site: http://www.dfki.de/~jameson/hci/

Jameson, A. (2002a): Usability issues and methods for mobile multimodal systems. In Proceedings of the ISCA Tutorial and Research Workshop on Multi-Modal dialogue in Mobile Environments, Kloster Irsee, Germany.

Jameson, A. / Schwarzkopf, E. (2002b): Pros and cons of controllability: an empirical study. In: De Bra, P. (Hg.): Adaptive hypermedia and adaptive web-based systems: Proceedings of AH 2002, Berlin: Springer.

Kronenberg, S. / Kummert, F. (1999): Incremental processing and syntactic disambiguation of extrapositions. In: Proceedings of the IASTED International Conference on Artificial Intelligenceand Soft Computing, 9.- 12. August 1999. Honolulu, USA: 343-347.

Kummert, F. / Fink, G. A. / Sagerer, G. / Braun, E. (1998): Hybrid object recognition in image sequences. In: Proceedings of the 14th International Conference on Pattern Recognition, volumne 2. Brisbane: 1165—1170.

Nickerson, R.S. (1969): Man-computer interaction: A challange for human factors research. In: Ergonomics 12: 501-517.

Preece, J. / Benyon, D. / Davies, G. / Keller, L. (1993): A guide to usability. Addison Wesley Publishing Company.

Preece, J. / Rogers, Y. / Sharp, H. (2002): Interaction design: beyond human-computer interaction. John Wiley & Sons.

Rosson, M. B. / Carroll, J. M. (2001): Usability engineering: scenario-based development of human computer interaction. Morgan Kaufmann.

Takahashi, T. / Nakanishi, S. / Kuno, Y / Shiari Y. (1998): Helping computer vision by verbal and nonverbal communication. In: Proceedings of the International Conference on Pattern Recognition (ICPR): 1216-1218.

Tenbrink, T. / Fischer, K. / Moratz, R. (2002): Spatial strategies in linguistic human-robot communication. In: Freksa, C. (Hg.): Künstliche Intelligenz – Themenheft *Spatial Cognition*. arenDTaP Verlag.

Trimmel, M. (1997): Wissenschaftliches Arbeiten. Ein Leitfaden für Diplomarbeiten und Dissertationen in den Sozial- und Humanwissenschaften mit besonderer Berücksichtigung der Psychologie. Wien: WUV-Universitätsverlag.

Vorwerg, C / Voß, I (Hrgs) (2002): Gedächtnisprozesse in Interaktion. SFB 360 Report 03/02, Universität Bielefeld.

Wachsmuth, S. / Fink, G. A. / Sagerer, G. (1998): Integration of parsing and incremental speech recognition. In: Proceedings of the the European Signal Processing Conference, volumne 1. Rhodes: 371-375.

Wachsmuth, S. / Sagerer, G. (2002): Bayesian Networks for Speech and Image Integration. In: Proceedings of 18th National Conference on Artificial Intelligence (AAAI-2002), Edmonton, Alberta, Canada: 300-306.

Weinert, A. W. (1987): Lehrbuch der Organisationspsychologie. München-Weinheim: Psychologie Verlangs Union.

Wottawa, H. / Thierau, H. (1998): Lehrbuch Evaluation. Bern [u.a.]: Huber.