Kombinierte multimodale Mensch-Rechner-Interaktionen

Hans-Jörg Bullinger, Klaus-Peter Fähnrich, Karl-Heinz Hanne
Fraunhofer-Institut für Arbeitswirtschaft und Organisation, Stuttgart

Zusammenfassung

Notepad-Computer und Eingabestifte, die „direkt" auf der Benutzerschnittstelle agieren, ermöglichen neue Mensch-Rechner-Interaktionsformen. Neben den traditionellen Interaktionsformen, wie direkte Manipulation, natürlichsprachliche Interaktion und formalsprachliche Interaktionsformen lassen sich kombinierte (multimodale) Interaktionsformen entwerfen. Diese Interaktionsformen haben Vor- und Nachteile. Insbesondere Kombinationen von direkter Manipulation und natürlicher Sprache führen zu interessanten Ergebnissen. „Bild und Wort" werden damit integriert. Realisierungen lassen sich auf unterschiedliche Wissensrepräsentationsarchitekturen entwerfen.

1 Einleitung

Notepad-Computer sind mittlerweile auf dem Markt verfügbar und geben dem traditionellen Begriff der direkten Manipulation in der Benutzungsoberfläche eine weitergehende Bedeutung. Multimedia-Systeme und Systeme der virtuellen Realität, verbunden mit natürlichsprachlichen Systemen und direkter Manipulation, bieten die technische Basis zur komfortablen, aufgabengerechten Mensch-Rechner-Interaktion. Neue Zeigegeräte, z.B. drahtlose Stifte, Datenhandschuhe sowie Displaytechnologien, z.B. berührempfindliche Flachbildschirme, erlauben die Einbindung von gestischen Interaktionen in die Mensch-Rechner-Schnittstelle und, darauf aufbauend, z.B. die Erkennung von Handschrift und Kommandogesten.

2 Generische Mensch-Rechner-Interaktionsmodi

In der Mensch-Rechner-Interaktion lassen sich generische Interaktionsmodi, wie formale Interaktionssprachen, direkte Manipulation und natürlichsprachliche Interaktion isolieren. Detaillierte Einteilungen nach DIN 66234 Teil 8, VDI 5005 und dem ISO-Vorstandard 9241 [10] führen als Interaktionsformen zusätzlich menügesteuerte Dialoge, formularbasierte Techniken und Frage-/Antwortsysteme ein.

3 Ziel kombinierter multimodaler Mensch-Rechner-Interaktion

Das Ziel der Entwicklung kombinierter multimodaler Benutzerschnittstellen ist es, Vorteile der verschiedenen Interaktionsmodi zusammenzufügen und Nachteile zu vermeiden. Insbesondere die Kombination von natürlichsprachlicher und direkter graphischer Mensch-Rechner-Interaktion bietet wesentliche Vorteile (vgl. z.B. [3], [5], [6], [8]).

Die Kombination von verschiedenen Interaktionsmodi kann im allgemeinen nicht allein auf dem Darstellungsniveau gelöst werden ([6], [7]). Es ist ein internes konzeptuelles Modell, eine gemeinsame Repräsentation des zugänglichen Interaktionswissens nötig. Jedes System muß eine Verbindung zwischen dargestelltem und ggf. selektiertem Objekt oder Objektteil und seiner Benennung oder Bedeutung herstellen und aufrechterhalten. Ob die Priorität der natürlichsprachlichen Komponente und der dort implementierten Wissensrepräsentation [11], der graphischen Objektrepräsentation oder einem unabhängigen konzeptuellen Modell [4] gegeben wird, ist global nicht zu entscheiden.

4 Gestische und deiktische Interaktionsformen

Kommunikation im allgemeinen kann sowohl auf verbaler als auch auf nichtverbaler Ebene durchgeführt werden. Das Spektrum der multimodalen menschlichen Kommunikation und den Zusammenhang zwischen Deixis und Gesten zeigt Abb. 1.

4.1 Gestische Interaktionsformen

Gesten sind Bewegungen der Arme und Hände, die kommunikativen Charakter besitzen und nicht aus einer Arbeitsbewegung der Arme oder Hände entstehen. Im Sinne der Mensch-Rechner-Interaktion können Gesten als bewußt zur Interaktion vorgesehene Bewegungen eines Zeigegerätes, geführt durch die Arme oder Hände eines Interaktionspartners im dreidimensionalen Raum oder auf einer Fläche, definiert werden. Ein Beispiel für gestische Interaktion ist die Referenz auf ein Objekt, dessen Namen dem Benutzer nicht bekannt ist. Solche Referenzen können durch eine Zeigegeste, begleitet von einer natürlichsprachlichen (deiktischen) Äußerung, hergestellt werden. Verbale Beschreibungen dienen in erster Linie zur Kategorisierung. Zeigeaktionen (Gesten) erlauben das einfache Lokalisieren von Orten und Objekten. Gesten lassen sich durch eine zeitlich aufeinander folgende, feste Reihenfolge (Sequenz) von Grundgesten bestimmen. Diese können symbolisch abgebildet und durch Gestiksprachen beschrieben werden.

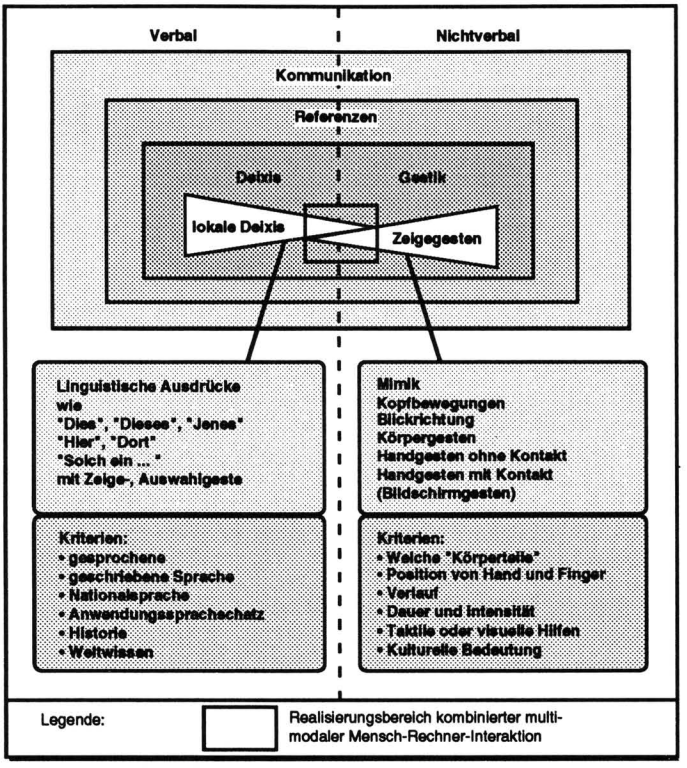

Abb. 1: Multimodale Interaktion. Deixis versus Gesten

4.2 Deiktische Interaktionsformen

Bei Gestik im Zusammenwirken mit verbalen Äußerungen in natürlicher Sprache entstehen deiktische Effekte. Nach der Definition von Lyons ist Deixis die:

> „(...) Eigenschaft bzw. Funktion sprachlicher Ausdrücke, die sich auf die Person-, Raum- und Zeitstruktur von Äußerungen in Abhängigkeit von der jeweiligen Ausgangssituation bezieht. Solche deiktischen Ausdrücke sind Personalpronomen (ich, du) Adverbialausdrücke (hier, dort) und Demonstativpronomen (dieser, jener). Im Unterschied zu Eigennamen und Kennzeichnungen (...) verweisen deiktische Ausdrücke entweder auf andere sprachliche Zeichen innerhalb eines gegebenen Textes (textuelle Deixis, Rededeixis) oder auf außersprachliche Elemente in Relation zur jeweiligen Sprechsituation (personale Deixis...temporale Deixis). Deixis gilt als Bindeglied zwischen Semantik und Pragmatik, insofern als die referentielle Bedeutung deiktischer Ausdrücke nur aus der jeweils pragmatisch situierten Sprechsituation heraus interpretierbar sind. (...)" ([9] S. 83-84).

4.3 Spezifikation zugelassener Zeigehandlungen in der Mensch-Rechner-Interaktion

Zeigehandlungen werden in der Mensch-Rechner-Interaktion auf einem geeigneten graphischen Bildschirm mit einem Zeigegerät gleichzeitig oder in engem zeitlichen Zusammenhang zur sprachlichen Äußerung durchgeführt.

An erster Stelle muß bei der Interpretation versucht werden, die Handlung selbst als Geste zu klassifizieren und falls möglich, in ihrer gestischen Bedeutung zu erkennen. Die menschliche Interaktion wird von einer Vielzahl von Zeigehandlungen begleitet ([1], [6], [12]). Ein Interaktionssystem muß in der Lage sein, aus der Kombination von erkannter Geste und sprachlicher Äußerung das Intendierte zu erschließen.

In der vorliegenden Realisierung werden nur punktuelle Zeigegesten im Zusammenspiel mit natürlichsprachlicher Interaktion zugelassen. Nichtpunktuelle Gesten im Zusammenwirken mit natürlichsprachlichen Formularabfragen werden z.B. in der Komponente TACTILUS zugelassen ([1], [12]).

5 Wissensrepräsentation kombinierter Mensch-Rechner-Interaktion

Die Kombination verschiedener Mensch-Rechner-Interaktionsmodi erfordert neben einer Wissensrepräsentation einen steuernden und koordinierenden Mechanismus zum gleichzeitigen, gleichberechtigten Zugriff auf interaktionsbezogenes Wissen auf der Basis standardisierter Interaktionswerkzeuge.

Die natürlichsprachlichen Komponenten stellen einerseits erweiterte Anforderungen an die Interaktionssteuerung, an die Repräsentationsmechanismen und an die Inferenzmechanismen zur Ableitung sprachlicher Ausdrücke. Andererseits besitzen gerade diese Systeme effektive Auflösungs- und Steuerungsmechanismen auf der Basis der zugrunde liegenden linguistischen Theorien.

In der Realisierung der Wissensrepräsentation lassen sich verschiedene Architekturen definieren:

- Wissens(-re-)präsentation mittels eines Präprozessors,
- durch eine Blackboardarchitektur und
- durch Auftrennung in „sichten"-spezifisches Wissen.

5.1 Wissensrepräsentation mit Präprozessor

Ein Präprozessor wird in die Interaktion zwischen ein bestehendes, natürlichsprachliches Anwendungssystem und die graphische Interaktionskomponente eingefügt. Es werden keine Eingriffe in das Anwendungssystem und nur wenige in das natürlichsprachliche System vorgenommen. Aus der Sicht des natürlichsprachlichen Systems handelt es sich um ein erweitertes intelligentes Terminal. Ein Präprozessor übernimmt die Ausgaben des Systems und die Eingaben des Benutzers und verändert sie entsprechend.

Die graphische Präsentation wird statisch auf einem geeigneten graphischen System erzeugt und nur im Präprozessor verwaltet. Änderungen an der Benutzungsoberfläche haben keinen direkten Einfluß auf das natürlichsprachliche System oder die Anwendungssysteme. Diesen werden nur die ihnen bekannten Anfrageformate zugeleitet, so daß keine grundlegende Ergänzung nötig wird. Ausgewählte Objekte werden durch einfache graphische Symbole oder Aktionen, wie z.B. durch Hervorhebung oder durch Einblenden eines Fadenkreuzes, kenntlich gemacht.

Der Einsatz dieses Modells ist beschränkt. Es können nur die direkt im Anwendungssystem zugänglichen Objekte auf andere Art identifiziert werden.

5.2 Wissensrepräsentation mit Blackboardarchitektur

Die Blackboardkopplung (vgl. z.B. [2]) folgt dem Gedanken der unabhängigen interagierenden Komponenten. Als gedachtes Medium wird eine gemeinsame Wissensbasis, die alle Informationen enthält, eingeführt. Ein ‹‹Blackboard›› ist eine globale Datenstruktur, auf die durch ein definiertes Protokoll von allen an der Wissensverarbeitung beteiligten Komponenten zugegriffen wird. In der Realisierung bedeutet dies, daß eine globale gemeinsame Wissensrepräsentationsstruktur, die alle Aspekte der Anwendung, der natürlichsprachlichen Komponenten und der graphischen Interaktionskomponenten beinhaltet, eingerichtet und aufrecht erhalten werden muß.

Neben den Repräsentationsaspekten der Anwendung und der Komponenten des natürlichsprachlichen Systems müssen ebenfalls alle Präsentationsdetails der interaktiven und passiven graphischen Objekte in einem gemeinsamen konzeptuellen Modell und in einer Wissensbasis gehalten und manipuliert werden. Im Prinzip ist dies ein Entwurf mit den geringen Problemen der Synchronisation und Kommunikation. Es lassen sich logikbasierte Systeme entwerfen, die dieser Architektur folgen. Tatsächlich widerspricht dieses Modell jedoch in verschiedenen Aspekten einem geordneten Ablauf.

Die Effizienz der Wissensrepräsentation ist nicht gegeben. Alle anwendungsbezogenen, interaktionsbezogenen, graphischen und natürlichsprachlichen Fakten müssen in diesem Entwurf in einer Wissensbasis gehalten werden. Insbesondere für graphische, oft gerätebezogene Details ist dies ineffizient. Änderungen sind nur mit großem Aufwand möglich.

Blackboardmodelle sind typisch für wissensbasierte Systeme. Ein Charakteristikum dieser Denkwelt ist der Nichtdeterminismus der Abläufe. Im Gegensatz dazu sind jedoch die Mensch-Rechner-Interaktionen keineswegs nichtdeterministisch. Es muß eine Kontrolle über den Dialogstand und dessen Historie gehalten werden.

Üblicherweise wird versucht, eine gemeinsame Sprache im Sinne einer semantischen Repräsentation sowohl für Wissen des Anwendungssystems, der natürlichsprachlichen Komponenten als auch der direktmanipulativen Interaktionskomponente einzuführen. Es bestehen grundsätzliche Schwierigkeiten, eine geeignete semantische Repräsentation, die alle logischen und zeitbezogenen Effekte umfaßt, zu finden. Die mögliche Nichtmonotonie des Schlußfolgerungsprozesses stellt bis heute nicht lösbare Probleme. Es existieren vereinzelte Ansätze, die jedoch nicht ausreichend in entwickelten Systemen berücksichtigt werden.

5.3 Auftrennung der Wissensrepräsentation in Sichten

In diesem Ansatz wird im gemeinsamen konzeptuellen Modell eine Teilung vorgenommen. Das Anwendungssystem (natürlichsprachliches System, traditionelle An-

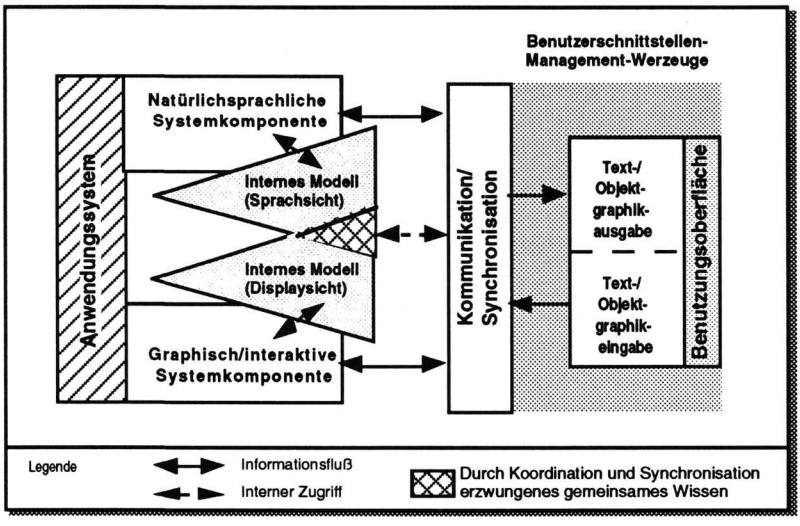

Abb. 2: Auftrennung von Display- und Sprachwissen

wendung) hält sein Wissen in geeigneter aber unabhängiger Form von der Wissensbasis der Interaktionsschicht (vgl. Abb. 2).

Die Grundidee der Verknüpfung der Wissensrepräsentation in einer internen Form mit einer natürlichsprachlichen und einer graphischen Oberflächenstruktur ist die Überlegung, daß beide Aspekte nur unterschiedliche Sichten auf die zugrunde liegende, gemeinsame Wissensbasis sind.

Relevante Änderungen in der einen Wissenswelt müssen für abgeschlossene Anwendungssysteme „unbemerkt" als neuer Fakt in die Wissensbasis eingetragen werden und damit die entsprechenden Abläufe anstoßen.

Die Effizienz der Wissensrepräsentation ist in diesem Modell gegeben. Da das interaktionsbezogene Wissen, z.B. Wissen über die interaktive Objektgraphik, in einer geeigneten Repräsentation gehalten werden kann, sind prozedurale, interaktionsbezogene Aspekte effizient repräsentierbar und vom deklarativen Wissen der Anwendung getrennt. Beide Wissensarten können unterschiedlich modelliert und ggf. mit unterschiedlichen Paradigmen realisiert werden.

Änderungen und Erweiterungen sind mit geringerem Aufwand möglich. Deklaratives Wissen der Anwendung kann in der zugehörigen, geeigneten Wissensrepräsentation geändert werden. Objekte und Attribute der graphischen Benutzungsoberfäche können in der anderen Wissensrepräsentation geändert werden. Es können Werkzeuge des Benutzerschnittstellen-Managements eingesetzt werden.

Die Interaktionskontrolle ist in diesem Modell von der Wissensdarstellung und dem Ablaufparadigma der Anwendung getrennt.

Die Konsistenz der semantischen Repräsentation der beiden unterschiedlichen Sichten ist durch die geeignete Realisierung gewährleistet.

Das Ziel dieses, auf einem aufgetrennten, konzeptuellen Modell der Wissensrepräsentation beruhenden Modell ist es, daß ein Teil des Anwendungssystemwissens, z.B. eine Datenbankanwendung, eine Sicht eines Teiles der „realen (Anwendungs-) Welt" abbildet und umgekehrt.

6 Realisierte Anwendungssysteme

6.1 Systeme zur deiktischen Formularinteraktion

In zwei Beispielanwendungen wurden Formulare und technische Skizzen als prototypische Testfälle für die Untersuchung des Zusammenwirkens zwischen natürlicher

Sprache und Zeigehandlungen in einer virtuellen Umgebung gewählt. Es sind drei Klassen von Referenzen möglich:

- Referenzen auf direkt sichtbare Objekte, z.B. Eingabefelder, Texte.
- Referenzen auf Daten, die eingegeben werden oder eingegeben wurden.
- Referenzen auf Konzepte der Formulare.

Die Zeigehandlung wird in den Satz eingebunden, d. h. sie erfolgt zwischen zwei Wörtern der Eingabe und wird mit Hilfe eines Handsymbols angezeigt. Die Zeigehandlung wird nach einer Heuristik aufgelöst ([5], [6]).
Ausgaben des Systems können auf verschiedene Weise durchgeführt werden:

- durch ein Ausgabefenster, welches wiederum die Dialoghistorie aufzeigt;
- durch spezielle Erklärungen, die in dafür geöffneten Ausgabebereichen angeboten und präsentiert gehalten werden;
- Durch Verknüpfung beider Ausgabeströme mit einem Spracherzeugungssystem, das akustisch Texte ausgibt.

Die Wissensrepräsentation wurde in einer geteilten, gekoppelten Lösung, entsprechend Abb. 2 realisiert. Die natürlichsprachliche Komponente enthält die bewertete Hierarchie möglicher Referenten und löst diese entsprechend der linguistischen Strategie auf.

Eine weitere Applikation erlaubt, über Skizzen technischer Zeichnungen, z.B. für Trainingsanwendungen, Fragen in deutsch zu stellen.

6.2 Systeme zur Erkennung von Handschriften und Korrekturgesten

Erkennung von Handschriften und Korrekturgesten sind insbesondere durch die neuen Notepad-Computer wirtschaftlich interessante Anwendungen der nichtverbalen, gestischen Mensch-Rechner-Interaktion. Besondere Eignung findet gestikbasierte Zeichenerkennung in Sprachen ohne Alphabet und mit hoher Zeichenzahl, die nicht auf einer Tastatur untergebracht werden können. Ebenso findet diese Technik Eingang in die Mensch-Rechner-Interaktion mit Notepad-Computern, die keine Tastatur integrieren. Der wesentliche Unterschied zu traditionellen Systemen der optischen Zeichenerkennung liegt in der Nutzung der zeitlichen Abfolge der Zeichenerzeugung, der Sequenz der Zeichenerzeugung.

Ein Beispiel dieser realisierten Systeme zeigt der Bildschirmabzug der Benutzungsoberfläche in Abb. 3. Die Implementierung umfaßt [7]:

- die Erkennung von 498 chinesischen Schriftzeichen;
- die Erkennung aller 46 japanischen Hiragana-Zeichen;

- die Erkennung aller 46 japanischen Katakana-Zeichen;
- die Erkennung der Buchstaben und Ziffern der westlichen Schriften in Blockschrift;
- die Erkennung freier Korrekturzeichen aus einem benutzerspezifischen Repertoire und
- die Erkennung von standardisierten Korrekturzeichen nach DIN 16511.

7 Erfahrungen und weitere mögliche Anwendungen

Die deiktischen und gestischen Mensch-Rechner-Interaktionstechniken sind in den herausgearbeiteten Anwendungsbereichen nützlich und effizient einsetzbar. Die realisierten, prototypischen Mensch-Rechner-Interaktions-Systeme zeigen die Anwendbarkeit kombinierter multimodaler Mensch-Rechner-Interaktionen in technischen Anwendungsbereichen. Deiktische Interaktions-Systeme sind in hohem Maße von den natürlichsprachlichen Systemen abhängig.

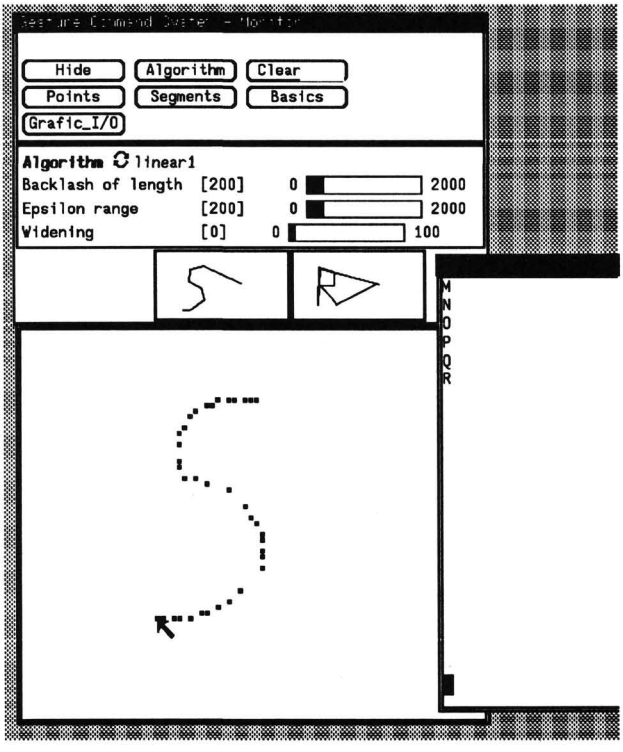

Abb. 3: Bildschirmabzug einer Beispielanwendung

Die direktmanipulativen Interaktionskomponenten müssen sich in der Regel diesen Systemen anpassen und ihnen zuarbeiten. Die Kontrolle liegt am besten bei der natürlichsprachlichen Komponente, die auch die Syntax der Wissensrepräsentation bestimmt. Gestische Interaktionssysteme dagegen lassen sich frei von linguistischen Überlegungen in verschiedenste Anwendungssysteme integrieren.

Werden die Konzepte der Gestenerkennung von der menschlichen Kommunikation auf andere Bereiche übertragen, ergeben sich interessante weiterführende Anwendungsfelder:

- Andere konventionelle oder genormte Zeichensysteme, beispielsweise aus der Elektrotechnik oder der Mathematik, können zum Drafting und Skizzieren auf berührempfindlichen Rechnerbildschirmen genutzt werden.
- Bestimmte technische, statistische oder medizinische Signale und Signalreihen können als „Gesten" interpretiert, mit symbolischen Sprachen beschrieben, repräsentiert und klassifiziert werden.
- Trajektorien von bewegten Objekten auf zweidimensionalen Bildern oder in dreidimensionalen Strukturen können ggf. als „Gesten" in erweitertem Sinne interpretiert und mit den beschriebenen Mechanismen erkannt werden.

8 Literatur

[1] Allgayer, J.; Harbusch, K.; Kobsa, A.; Reddig, C.; Reithinger, N.; Schmauks, D.: XTRA: A Natural Language Access System to Expertsystems. In: International Journal of Man-Machine Studies (1989) Vol. 31, S. 161-195.

[2] Barr, A.; Feigenbaum, E.: The Handbook of Artificial Intelligence. Los Altos: Kaufmann, 1981.

[3] Cohen, P. R.; Dalrymple, M.; Moran, D.; Pereira, F. C. N.; Sullivan, J.; Gargan, R.; Schlossberg, J.; Tyler, S.: Synergistic Use of Direct Manipulation and Natural Language. In: Tagungsband CHI '89, Human Factors in Computing Systems. New York: ACM, 1989, S. 227-233.

[4] Görner, C.; Vossen, P.; Ziegler, J.: Direct Manipulation Interface. In: Methods and Tools in User-Centred Design for Information Technology/ Hrsg. von M. Galer, S. Harker und J. Ziegler. Amsterdam u.a.: Elsevier, 1992, Kapitel 8, S. 237-279.

[5] Hanne, K.-H.: Multimodal Communication, Natural Language and Direct Manipulation (Gestures) in Human-Computer Interaction. In: Multimedia Interface Design in Education/ Hrsg. von A. D. N. Edwards und S. Holland. Berlin u. a.: Springer, NATO ASI Serie F: Computer and Systems Sciences, Vol. 76, 1992, Kapitel 11, S. 157- 175.

[6] Hanne, K.-H.; Bullinger, H.-J.: Multimodal Communication: Integrating Text and Gestures. In: Multimedia and Multimodal Interface Design/ Hrsg. von M. Blattner und R. Dannenberg. Reading: ACM Press, Addison Wesley, 1992, Kapitel 8, S. 127-138.

[7] Hoepelman, J.; Hanne, K.-H.: Neue Entwicklungen der Mensch-Rechner-Interaktion. In: Handbuch des Informationsmanagements im Unternehmen/ Hrsg. von H.-J. Bullinger. München: Beck, 1991, Band 1, Kap. 29, S. 867-893.

[8] Kobsa, A.; Allgayer, J.; Reddig, C.; Reithinger, N.; Schmauks, D.; Harbusch, K.; Wahlster, W.: Combining Deictic Gestures and Natural Language for Referent Identification. In: Tagungsband International Conference on Computational Linguistics. Bonn, 1986, S. 356-361.

[9] Lyons, J.: Semantik. München: Beck, 1983.

[10] Norm ISO 9241 (Draft International Standard): Ergonomic Requirements for Office Work with Visual Display Terminals. Teil 10: Dialogue Principles. ISO/ TC 159/ SC 4/ WG 5, 1992.

[11] Ostler, N. D.: LOQUI: How Flexible Can a Formal Prototype Be? In: The Structure of Multimodal Dialogue/ Hrsg. von M. M. Taylor, F. Néel und D. G. Bouwhuis. Amsterdam u. a.: Elsevier, 1989, S. 407-416.

[12] Schmauks, D.: Deixis in der Mensch-Maschine-Interaktion, Multimediale Referentenidentifikation durch natürliche und simulierte Zeigegesten. Tübingen: Niemeyer, 1991. Zugl. Saarbrücken, Universität des Saarlandes, Dissertation, 1990.

Hans-Jörg Bullinger, Klaus-Peter Fähnrich, Karl-Heinz Hanne
Fraunhofer-Institut für Arbeitswirtschaft und Organisation (IAO),
Nobelstr. 12
D-7000 Stuttgart 80

Tel. +49 711 970 2413
Fax +49 711 970 2401