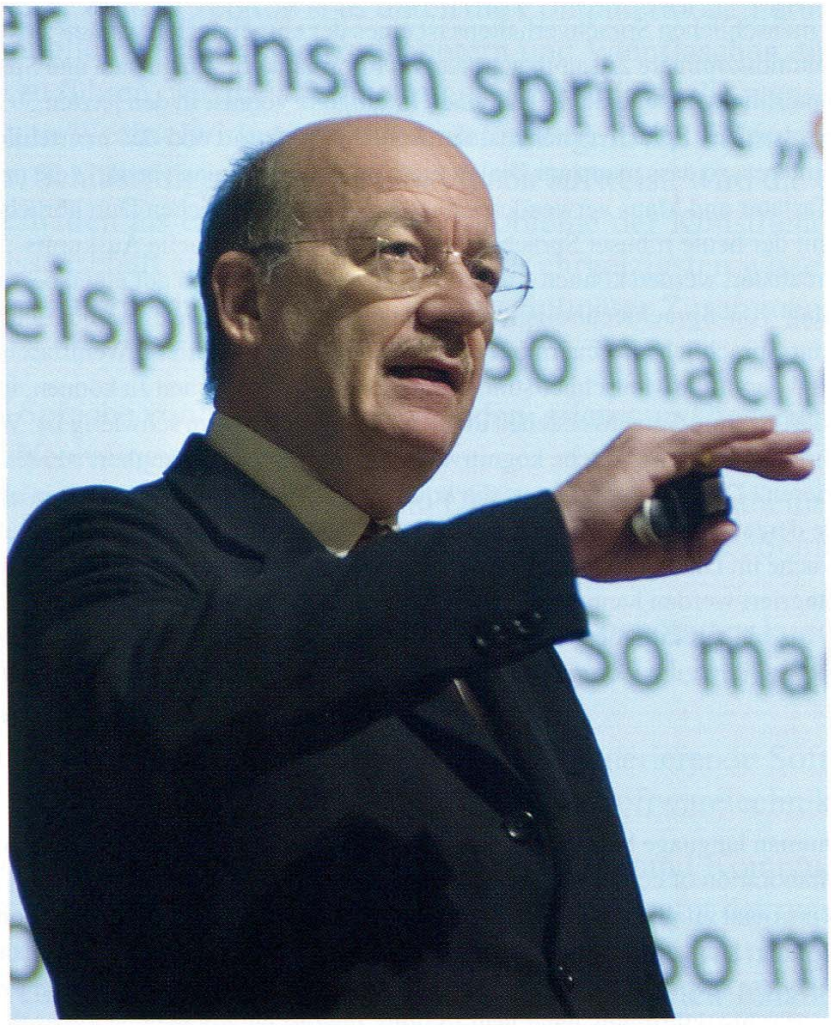
in: Lengauer, Th. (ed.) Computermodelle in der Wissenschaft- zwischen Analyse, Vorhersage und Suggestion. Nova Acta Leopoldina, Band 110, Nummer 377, Wissenschaftliche Verlagsgesellschaft: Stuttgart, pp. 119-141

Nova Acta Leopoldina NF *110*, Nr. 377, 119–141 (2011)

Mit den Dingen sprechen: Autos, Roboter und Weinflaschen als Dialogpartner?[1]

Wolfgang WAHLSTER ML (Saarbrücken)

Mit 19 Abbildungen



1 Die multimediale Präsentation mit Videodemonstrationen zu dem Vortrag kann auf der beigelegten DVD bzw. in der Online-Publikation des Bandes auf der Homepage der Akademie betrachtet werden.

Zusammenfassung

Die Computersimulation des menschlichen Sprachverhaltens ist eines der ehrgeizigsten wissenschaftlichen Ziele unseres Zeitalters. Sie setzt die multidisziplinäre Zusammenarbeit von Informatikern, Linguisten und Sprachpsychologen sowie Spezialisten der Computerlinguistik und der Künstlichen Intelligenz voraus. In den letzten 30 Jahren ist es uns gelungen, die komplexen Prozesse des Sprachverstehens, der Sprachproduktion und des menschlichen Dialogverhaltens immer exakter zu simulieren, so dass man zum Dialog mit Computersystemen verstärkt die natürliche Spracheingabe anstatt umständlich Tastatur und Maus verwenden kann. Die wissenschaftlichen Durchbrüche haben zu einer Sprachtechnologie geführt, mit der heute robuste Sprachdialogsysteme für maschinelle Auskunfts-, Buchungs-, Bestell- und Beratungssysteme realisiert werden können – allerdings stets eingeschränkt auf einen bestimmten Diskursbereich. Durch die Einbettung von Sprachtechnologie in Alltagsgegenstände wird es möglich, beispielsweise Fahrerassistenzsysteme im Auto sprachlich zu steuern oder entsprechend instrumentierte Weinflaschen im Verkaufsregal anzusprechen. Damit wird der uralte Menschheitstraum, mit den Dingen sprechen zu können, in greifbare Nähe gerückt. Zunächst wird motiviert, warum Sprachverstehen durch Computer so extrem schwierig ist: Wie kann Sprachverhalten algorithmisch beschrieben werden? Welche kognitiven Leistungen müssen simuliert werden, um natürliches Sprachverhalten zu reproduzieren? Der Beitrag erläutert die komplexen Verarbeitungsprozesse in aktuellen Sprachdialogsystemen und macht die derzeitigen Grenzen bei der Simulation von Sprachverhalten deutlich. Es wird gezeigt, wie Mehrdeutigkeiten der Sprache im Dialogkontext aufgelöst werden können und wie Sprache mit Gestik und Mimik in multimodalen Dialogen integriert werden kann. Es wird die Entwicklung der Forschung zur Sprachtechnologie in Deutschland in vier Dekaden bis hin zur kommerziellen Nutzung aufgezeigt. Schließlich wird auch gezeigt, wo der weitere massive Forschungsbedarf liegt, um Computer zu echten Dialogpartnern machen zu können.

Abstract

The computer simulation of human language behavior is one of the most ambitious scientific goals of this century. It requires multidisciplinary collaboration of computer scientists, linguists and psycholinguists as well as specialists in computational linguistics and artificial intelligence. During the last 30 years, our simulations of the complex processes underlying language understanding, language production and human dialog behavior became more and more precise, so that today we can increasingly use speech input instead of the cumbersome keyboard and mouse for the interaction with computer systems. Scientific breakthroughs have lead to quite mature language technology that enables us to realize spoken dialog systems for automated help, booking, ordering and consulting systems – however always restricted to a limited discourse domain. Embedding language technology into everyday items allows us, for example, to control driving assistance systems in cars by speech or to talk to wine bottles in instrumented shopping shelves. Thus the ancient dream of mankind, to be able to converse with things, is coming into reach. First, we motivate why language understanding is so extremely difficult for computer systems: How can we describe language behavior computationally? Which cognitive abilities must be simulated to reproduce natural language behavior? This paper discusses the complex processing schemes of current spoken dialog systems and presents today's limits of these simulations. We show how ambiguity can be resolved in the dialog context and how we can combine speech with gestures and facial expressions in multimodal dialogs. We present the development of human language technology in Germany during four decades up to the commercial deployment. Finally, we discuss the massive need for further research, if we want to use computers as real dialog partners.

1. Die Computermodellierung von Sprachverhalten als Schlüsselproblem

Sprachverhalten ist eine definierende Eigenschaft unserer Spezies. Seit etwa 60 Jahren beschäftigen sich Informatiker mit den Fragen:

(*1.*) Können Computermodelle natürliches Sprachverhalten analysieren, vorhersagen und simulieren?
(*2.*) Können Computersysteme zu echten Dialogpartnern des Menschen werden?

Die Fortschritte im Bereich der Computerlinguistik (vgl. ALLEN 1994, COLE et al. 1997, JURAFSY und MARTIN 2000) und der Künstlichen Intelligenz (vgl. RUSSEL und NORVIG 2009) haben uns in den letzten 30 Jahren einer positiven Beantwortung dieser Fragen näher gebracht. Damit erscheint der uralte Menschheitstraum, mit den Dingen sprechen zu können, realisierbar zu werden. Dieser Beitrag zeigt, wie es durch die Computermodellierung von Sprachverhalten schon heute möglich wird, mit Autos, Robotern und Weinflaschen einfache Dialoge in Umgangssprache zu führen, wie dies im Titel suggeriert wird. Da die menschliche Sprachfähigkeit eine Vielfalt von Wechselwirkungen mit anderen kognitiven Fähigkeiten wie Bildverstehen, Problemlösen, Schlussfolgern, Lernen und Planen aufweist, wird die Computermodellierung von Sprachverhalten als eines der Schlüsselprobleme der Künstlichen Intelligenz (KI) und Kognitionswissenschaft angesehen. Um eines der ehrgeizigsten wissenschaftlichen Ziele unseres Zeitalters zu erreichen, werden in interdisziplinärer Zusammenarbeit folgende Fragen erforscht:

(*1.*) Wie wird Sprache vom Menschen verstanden, erlernt und generiert? (zusammen mit Psycho- und Neurolinguistik)
(*2.*) Wie können Sprach-, Bedeutungs- und Dialogstrukturen mathematisch-logisch beschrieben werden? (zusammen mit Linguistik, Sprachphilosophie, Semiotik und Kommunikationswissenschaft)
(*3.*) Wie können die für intelligentes Sprachverhalten benötigten kognitiven Leistungen maschinell verfügbar gemacht werden? (zusammen mit anderen Teilgebieten der Künstlichen Intelligenz und Kognitionswissenschaft)
(*4.*) Wie können effiziente sprachverstehende und -generierende Softwaresysteme realisiert werden? (zusammen mit Computerlinguistik und Softwaretechnik)

Sprache ist nicht nur das Hauptmedium für Information und Kommunikation, sondern auch Werkzeug bei der Arbeit und sozialem Handeln. Daher kommt der maschinellen Analyse von Sprache und Sprachdialogsystemen in der Informations- und Internetgesellschaft eine Schlüsselrolle zu, weil nur so die Mehrzahl der sprachlich codierten Inhalte einer automatisierten Weiterverarbeitung zugeführt werden können. Sprache kann nicht nur akustisch und schriftlich, sondern auch gestisch (z. B Gebärdensprache für Gehörlose) und haptisch (z. B. Blindenschrift) codiert werden, wobei bis auf den Fall der Tastatureingabe zunächst immer eine akustische oder visuelle Signalverarbeitung stattfinden muss, bevor die Struktur und Bedeutung einer Äußerung durch Symbolverarbeitung auf dem Computer verarbeitet werden kann. Mit der heute verfügbaren Sprachtechnologie können prinzipiell all diese Modalitäten der Sprache mit Computersystemen analysiert und generiert werden, wobei die maschinellen Systeme auf der Analyseseite hauptsächlich aufgrund der Unsicherheiten bei der Signal-Symbol-Transformation und des Umfangs an erforderlichem Kontext- und Weltwissen allerdings bei weitem noch nicht die Robustheit und Qualität menschlichen Verstehens erreichen.

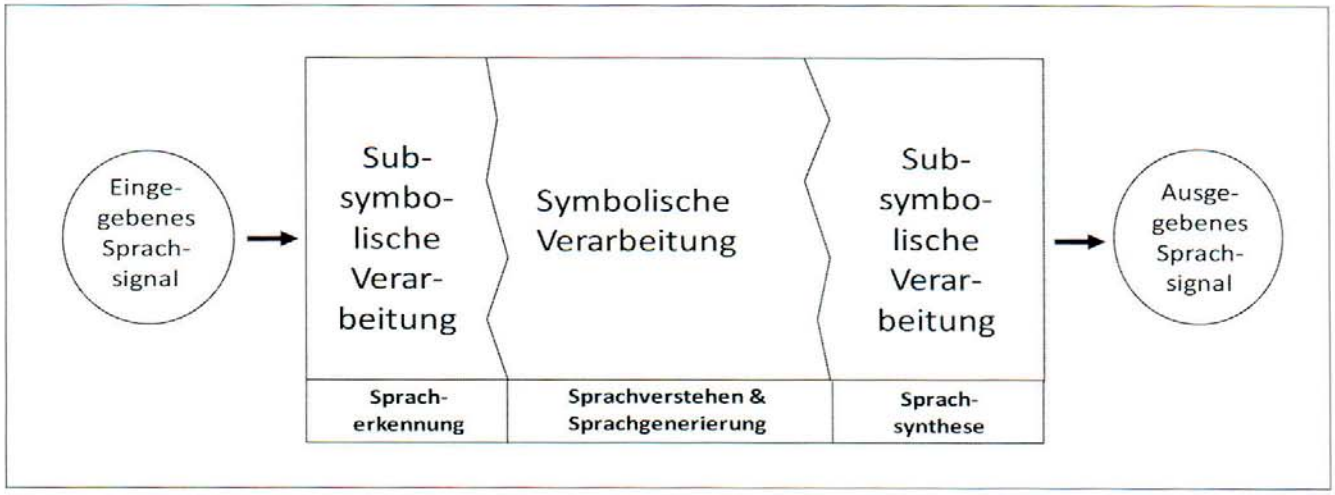


Abb. 1 Die Signal-Symbol-Signal-Transformation in Sprachdialogsystemen

Abbildung 1 zeigt den stark vereinfachten Verarbeitungsablauf in einem Sprachdialogsystem, das auf gesprochene Spracheingaben mit gesprochenen Sprachausgaben reagiert, etwa wenn eine Anfrage eines Benutzers über sein SmartPhone mit GPS-Ortung wie „Wo finde ich den nächsten Briefkasten mit Sonntagsleerung?" mit „In der Saargemünder Str. 25. 350 m entfernt von hier." von einer synthetischen Stimme beantwortet wird. Nach der Spracherkennung, welche das akustische Eingabesignal in symbolische Worthypothesengraphen transformiert, erfolgt das symbolische Sprachverstehen mit dem Ziel der formalen Bedeutungsrepräsentation für die Eingabe. Diese wird dann zur Berechung einer Dialogreaktion verwendet, die danach durch Sprachgenerierung verbalisiert wird. Schließlich erfolgt über die Sprachsynthese die Ausgabe des Sprachdialogsystems. Oft wird eine verbale Antwort des Systems auch mit einer graphischen Ausgabe kombiniert, wie dies in Abbildung 2 mit unserem Dialogsystem SmartWeb illustriert wird (vgl. auch WAHLSTER 2008).

Abb. 2 Mobile Fragebeantwortung als Webdienst

Die einzelnen Subprozesse in solchen Sprachdialogmodellen sind äußerst komplex und führen auf vielen Ebenen an die technologischen Grenzen der Computermodellierung. Woher diese Komplexität kommt, soll im folgenden Abschnitt gezeigt werden.

2. Warum ist Sprachverstehen für den Computer so schwer?

Bei der Computermodellierung von Sprachverstehen stellt die schrittweise Reduktion von Unsicherheiten bei der Interpretation von sprachlichen Äußerungen die größte Herausforderung dar. Dies zeigt sich bereits bei der Abbildung von Schallwellen auf Wörter, da es dafür keine eineindeutige Funktion gibt. Die Äußerung „Urlauber wollen wieder me:r ans me:r" enthält in der phonetischen Umschrift zweimal „me:r" mit langem „e", muss aber im Kontext richtig als die Aussage „Urlauber wollen wieder mehr ans Meer" interpretiert werden. Wie Abbildung 3 im Spektrogramm für den zeitlichen Verlauf des Schallsignals in einer Darstellung der Frequenzverteilung deutlich zeigt, sind die Muster für „mehr" und „Meer" in dieser Äußerung nahezu identisch.

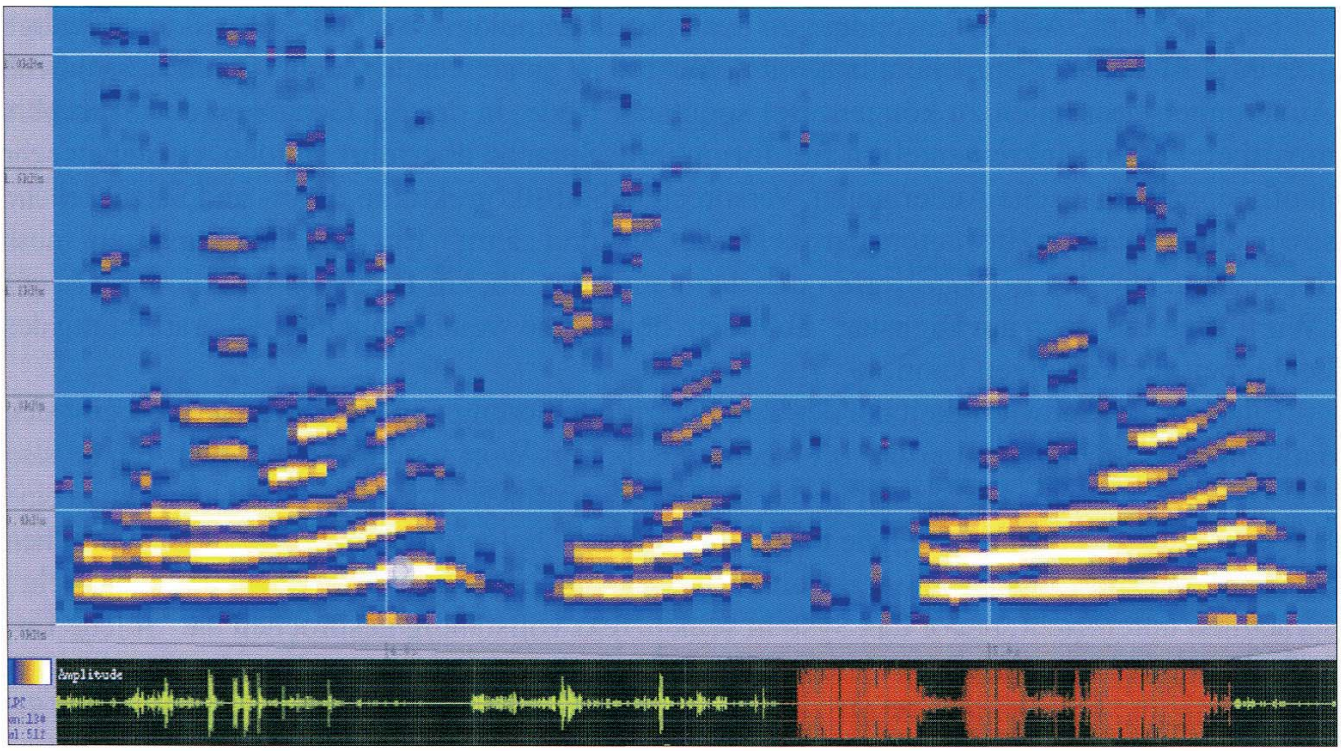


Abb. 3 Spektrogramm mit der Frequenzverteilung für „me:r ans me:r"

Der naive Ansatz, einfach eine Datenbank aller Spektren für alle Wörter einer Sprache anzulegen und danach einen Musterabgleich mit den Eingabespektren durchzuführen, ist zum Scheitern verurteilt und kann für eine Computermodellierung des Sprachverstehens nicht verwendet werden. Ein und dasselbe Wort kann, auch wenn es vom selben Menschen gesprochen wird – z. B. je nach dem emotionalem Zustand des Sprechers und der Situation – unterschiedlich laut, in verschiedenen Tonlagen und Geschwindigkeiten ausgesprochen werden. Hinzu kommen die großen Ausspracheunterschiede von Person zu Person, u. a. auch in Abhängigkeit vom Geschlecht und Alter sowie der Muttersprache des Gesprächspartners. Diese hohe phonetische Variabilität lautsprachlicher Äußerungen muss daher bei der Computermodellierung

in statistischen Modellen berücksichtigt werden. Erschwerend wirkt, dass viele Menschen einen Dialekt sprechen und damit eine Vielzahl weiterer Aussprachevarianten entstehen, die zu Interpretationsunsicherheiten führen. So kann die Äußerung „Ich find das nätt." je nach Heimat des Sprechers als „Ich finde das nett" oder „Ich finde das nicht" interpretiert werden.

Eine weitere Schwierigkeit bei der Computermodellierung von Sprachverstehen besteht darin, dass Wortgrenzen oft im Sprachfluss untergehen und damit eine automatische Segmentierung einer Äußerung keine triviale Aufgabe ist. So muss die durch Koartikulation entstandene Lautfolge „amontag" zunächst auf Wortebene zu „am Montag" segmentiert werden, bevor eine weitere Verarbeitung der Eingabe erfolgen kann. Der Mensch spricht auch „ohne Punkt und Komma", so dass neben der Segmentierung auf Wortebene auch die Segmentierung auf Satzebene und die Erkennung des Satzmodus beim automatischen Sprachverstehen oftmals weitere Interpretationsunsicherheiten erzeugt.

So kann die Wortfolge „So machen wir das vielleicht klappt es" nach einer Satzsegmentierung beispielsweise in die beiden Aussagesätze „So machen wir das. Vielleicht klappt es." oder einer Aussage mit nachfolgender Frage „So machen wir das vielleicht. Klappt es?" transformiert werden. Bei der automatischen Segmentierung auf Satzebene spielen die Satzmelodie, die Akzentuierung und der Sprechrhythmus als prosodische Merkmale eine wesentliche Rolle.

Für praktische Alltagsanwendungen der Sprachtechnologie wie mobile Auskunftsdialoge über Mobilfunk oder Sprachdialoge mit dem Auto ist es notwendig, auch spontansprachliche Äußerungen bearbeiten zu können. Bei spontaner Rede, in der die Gedanken des Sprechers unvorbereitet und inkrementell in frei formulierte Sprache überführt werden, entstehen sehr häufig ungrammatische Äußerungen, Versprecher, Selbstkorrekturen, Wort- und Satzabbrüche sowie eine Fülle von Häsitationspartikeln wie „äh" und „ähm" Ein Sprachdialogsystem muss Spontansprache wie in „Wir treffen uns dann am Mon, äh, am Dienstag" letztlich intern auf „Wir treffen uns dann am Dienstag" abbilden, bevor eine weitere inhaltliche Verarbeitung des Dialogbeitrages erfolgen kann.

Eine weitere Schwierigkeit bei der Computermodellierung von Dialogverstehen ergibt sich dadurch, dass sich Dialogpartner oft wechselseitig „ins Wort fallen" sich Sprachsignale also überlappen (vgl. Abb. 4). Typisch ist bei Telefonauskunftssystemen, dass der Nutzer etwa eine lange Ansage des Systems, welche beispielsweise die Dialogoptionen mit einer synthetischen Stimme vorstellt, mit der Auswahl einer ihm passenden Option wie „Buchungsübersicht" sprachlich unterbricht, bevor das System seinen Text vollständig gesprochen hat. Solche Unterbrechungen eines Sprechers durch einen Dialogpartner sind in der Alltagskommunikation sehr häufig und führen meist zu effizienterer Dialogführung, auch wenn sie in bestimmten Situationen als unhöflich gelten.

Abb. 4 Überlappende Sprachäußerungen in spontansprachlichen Dialogen

Um selbst einen menschlichen Dialogpartner jederzeit unterbrechen zu können, muss ein Computersystem dessen Eingaben schritthaltend analysieren und zu jedem Zeitpunkt einer Äuße-

rung eine zumindest partielle Interpretation bereithalten. Diese sogenannte Inkrementelle Analyse stellt eine besondere Herausforderung bei der Modellierung dar, weil der Redefluss zunächst häufig syntaktisch und semantisch in die Irre leiten kann. Wenn wir das Segment „Die Staatssekretärin lobt" gehört haben, gehen wir aufgrund der Wahrscheinlichkeiten von Satzkonstruktionen im Deutschen zunächst davon aus, dass die Staatssekretärin als Subjekt des Satzes fungiert und erwarten als Nächstes die Angabe, wer von ihr gelobt wird. Wenn der Satz dann aber als „Die Staatssekretärin lobt der Ministerpräsident" fortgesetzt wird, müssen wir „umdenken" und nun den Ministerpräsidenten als Subjekt und die Staatssekretärin als Objekt des Satzes interpretieren. Die Notwendigkeit der Korrektur fehlerhafter Erstinterpretationen bei einer inkrementellen Sprachverarbeitung ergibt sich oft bei ungewöhnlichen Satzstellungen, die – wie im obigen Beispiel – zur besonderen Fokussierung eines Aussageteils durch Topikalisierung verwendet werden.

Die größte Interpretationsunsicherheit bei der Computermodellierung von Sprachverhalten ergibt sich durch die vielfältigen Arten der Mehrdeutigkeit natürlicher Sprache (Abb. 5). So bleibt ohne weiteren Kontext bei einer Aussage wie „Wir telefonierten mit Freunden in Japan" mehrdeutig, ob damit gemeint ist, dass wir in Japan mit Freunden telefonierten oder ob wir Freunde in Japan aus einem anderen Land angerufen haben. Eine solche Mehrdeutigkeit der Anbindung einer Präpositionalphrase wie „in Japan" ist eine typische Form der syntaktischen Ambiguität, die nur unter Hinzuziehung weiterer Wissensquellen beim maschinellen Verstehen aufgelöst werden kann.

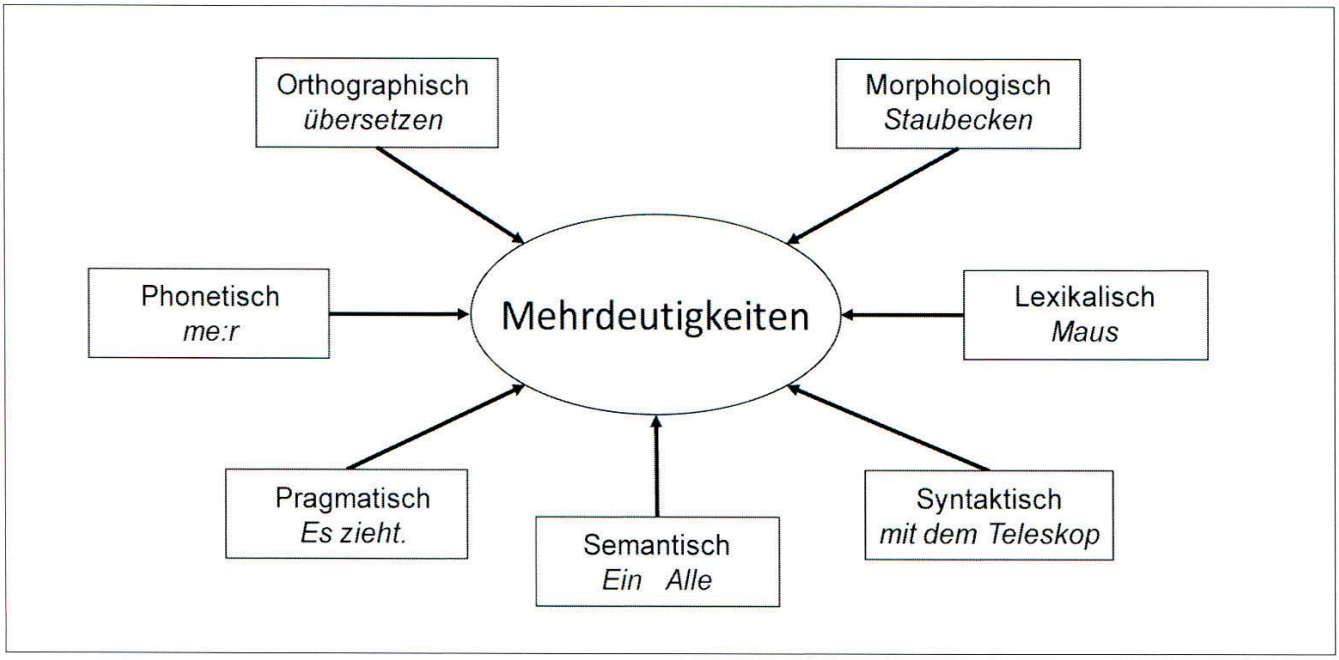


Abb. 5 Übersicht zu wichtigen Formen der Mehrdeutigkeit natürlicher Sprache

Abbildung 5 zeigt, dass es u. a. auch pragmatische Mehrdeutigkeiten wie bei der Äußerung „Es zieht" gibt, wo ohne Kontext unklar ist, ob es sich um eine Aufforderung, die Zugluft zu verhindern, oder um eine reine Tatsachenbeschreibung handelt. Besonders schwierig und ohne zusätzliches Weltwissen meist unmöglich ist die Auflösung von semantischen Mehrdeutigkeiten im Bereich der natürlichsprachlichen Quantoren und ihrer Gültigkeitsbereiche, wie das in der Äußerung „Allen Studenten gehört ein Computer" deutlich wird, wo unklar ist, ob ein einziger Computer allen Studenten gehört oder ob jeder Student einen eigenen Computer hat.

Da sich die alternativen Interpretationen durch Mehrdeutigkeiten über alle Verarbeitungsebenen kumulieren, entsteht bei der Computermodellierung die Gefahr einer kombinatorischen Explosion von Lesarten. Diese muss durch die Ausfilterung unsinniger Kombinationen im jeweiligen Kontext während des Sprachverstehensprozesses frühzeitig eingedämmt werden. Nimmt man die oben skizzierten Unsicherheiten bei der Worterkennung und der Satzsegmentierung während der Spracherkennung hinzu, so ergibt sich insgesamt eine Hypothesenflut, die letztlich auf eine einzige Bedeutungsrepräsentation reduziert werden muss, um Sprachverhalten sinnvoll zu modellieren. Im nächsten Abschnitt werden Verarbeitungsprozesse aktueller Sprachverarbeitungssysteme erläutert, die eine schrittweise Unsicherheitsreduktion ermöglichen.

3. Die Reduktion der Interpretationsunsicherheit beim Sprachverstehen

Grob werden in aktuellen Computermodellen des Sprachverstehens drei Verarbeitungsstufen unterschieden, die eine schrittweise Reduktion von Interpretationsunsicherheiten vornehmen (vgl. Abb. 6).

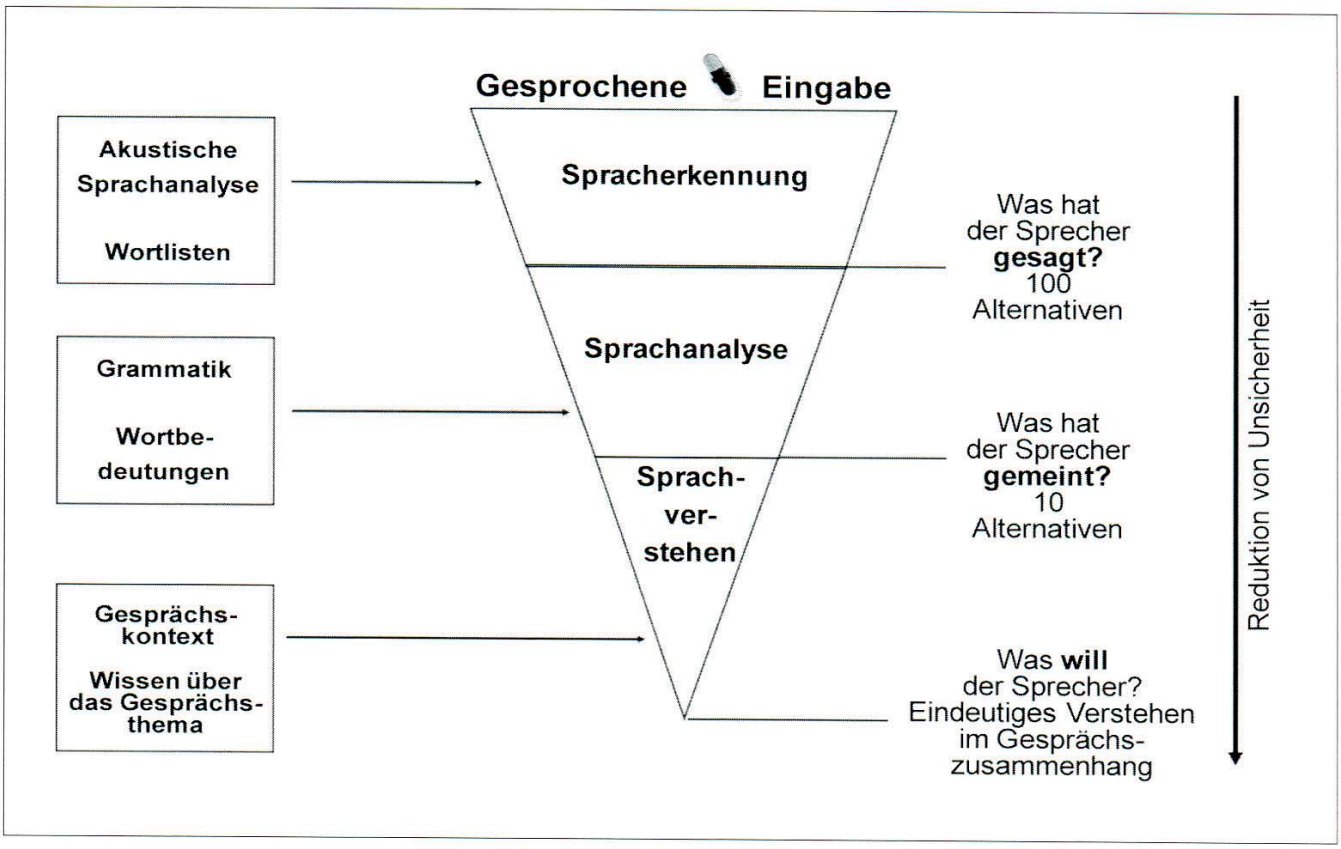


Abb. 6 Die drei Stufen der Sprachverarbeitung

Bei der Spracherkennung geht es zunächst darum, was der Sprecher mit seiner Äußerung gesagt haben könnte, in der Sprachanalyse darum, was er gemeint haben könnte, und schließlich beim Sprachverstehen darum, was er mit seiner Äußerung eigentlich wollte. Hierbei werden alle semiotischen Ebenen der Zeicheninterpretation von der syntaktischen über die semantischen bis hin zu den pragmatischen Aspekten einer Äußerung durchlaufen. Auf jeder Ebene wird die Flut der Interpretationshypothesen durch die Auswertung von zusätzlichen Wissens-

quellen wie Grammatiken und Wissensbasen zum Gesprächsthema um eine Größenordnung verringert, um schließlich zu einem eindeutigen Verstehen im Gesprächszusammenhang zu kommen (vgl. Abb. 6).

Die Berücksichtigung des Kontextes beim Sprachverstehen ist eine der wesentlichen Voraussetzungen für anspruchsvolle Anwendungen der Sprachtechnologie. Besonders deutlich wird dies beim automatischen Dolmetschen von Telefongesprächen, wie dies in einem Anwendungsszenario des Verbmobil-Systems für die Domäne der Reiseplanung realisiert wurde (vgl. WAHLSTER 2000). Das Wort „nächste" muss bei der Übersetzung ins Englische für die Eingaben „Wann fährt der nächste Zug nach Halle ab?" und „Wo befindet sich das nächste Hotel?" im ersten Falle mit „next" und im zweiten Kontext mit „nearest" wiedergegeben werden, je nachdem, ob ein zeitlicher oder räumlicher Bezug besteht. Bei einer Wort-für-Wort-Übersetzung wäre dagegen kaum eine Verständigung möglich. Es muss zunächst die Eingabe inhaltlich im Diskurskontext verstanden werden, bevor in der jeweiligen Zielsprache versucht wird, die intendierte Bedeutung und das Äußerungsziel sprachlich umzusetzen. In Verbmobil werden mehrere Übersetzungsstränge parallel ausgewertet, wobei die Übersetzung aufgrund einer expliziten Bedeutungsrepräsentation im Sinne einer Dialogsemantik der quellsprachlichen Eingabe die anspruchvollste, aber auch aufwendigste Methode ist.

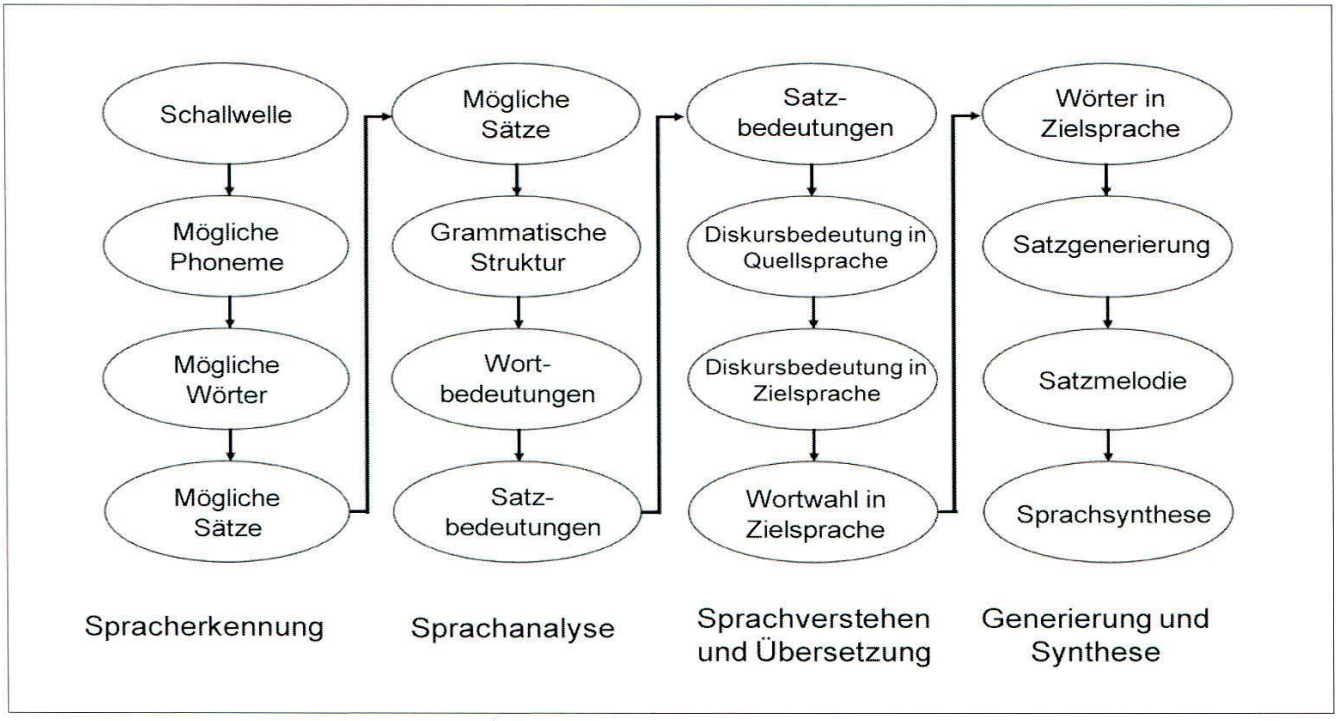


Abb. 7 Von der Spracheingabe zur Sprachausgabe in einem Dolmetschsystem

Die Verarbeitungskaskade in Abbildung 7 ist eine starke Abstraktion von den real verwendeten Computermodellen, weil der Kontrollfluss nicht einfach linear verläuft, sondern in der Künstlichen Intelligenz und Computerlinguistik komplexe Softwarearchitekturen entwickelt wurden, um effiziente Sprachdialogsysteme durch die wechselseitige Disambiguierung verschiedener Verarbeitungshypothesen zu realisieren. Bewährt hat sich z. B eine sogenannte Multi-Nachrichtentafel-Architektur (vgl. Abb. 8) in der verschiedene Verarbeitungskomponenten lesend und schreibend auf gemeinsam verwendete Datenstrukturen unterschiedlicher Nachrichtentafeln zugreifen können.

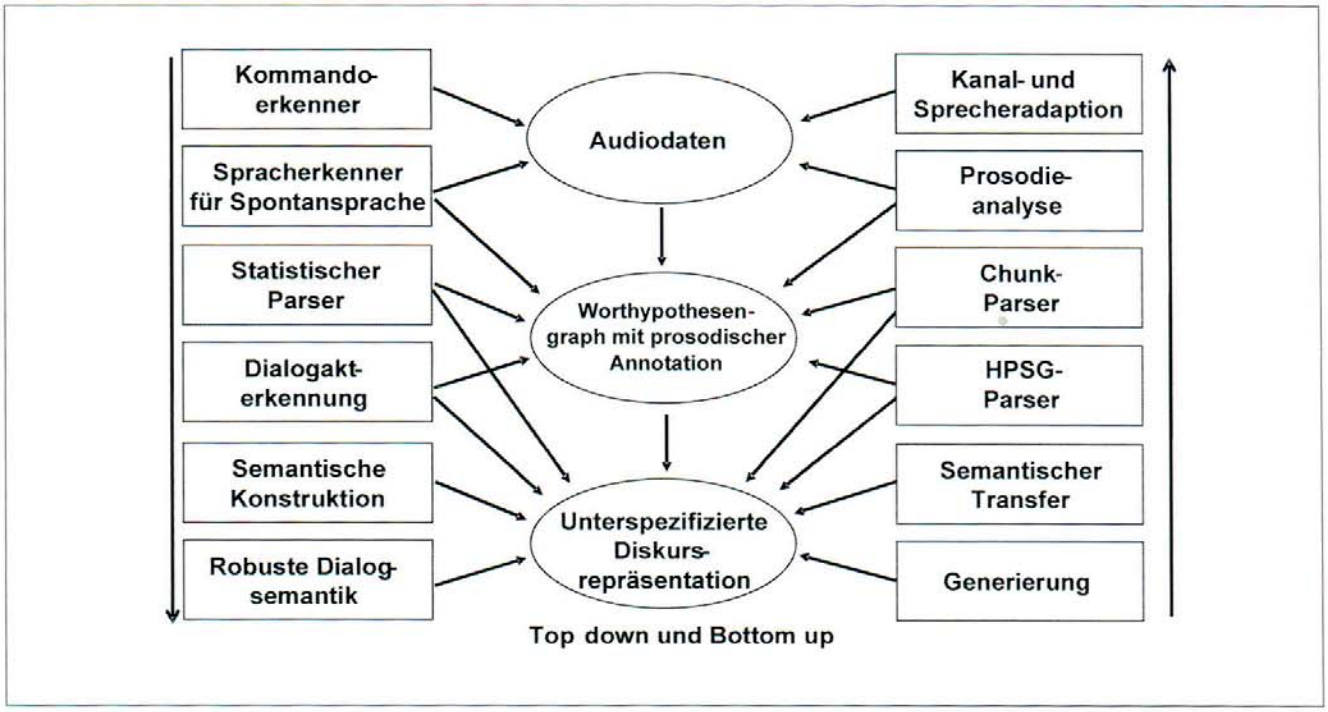


Abb. 8 Multi-Nachrichtentafel-Architektur für ein Sprachdialogsystem

In Abbildung 8 werden drei Nachrichtentafeln (als Ellipsen dargestellt) verwendet, wobei Information zwischen den Komponenten von oben nach unten (*top down*) und von unten nach oben (*bottom up*) fließen kann. Dieses Beispiel aus dem Verbmobil-System zeigt, dass drei verschiedene Parser an der syntaktischen Analyse über dem Worthypothesengraphen beteiligt sind, um eine Diskursrepräsentation zu erzeugen. Es ergibt sich eine enge Verzahnung von probabilistischen und wissensbasierten Verarbeitungsmethoden in einem hybriden und über die Nachrichtentafeln eng vernetzten Gesamtsystem.

Aktuelle Sprachdialogsysteme können auch die jeweils gesprochene Sprache anhand von statistischen Sprachmodellen (vgl. MANNING und SCHÜTZE 2000) erkennen. So haben wir in unserem Dolmetschsystem Verbmobil (vgl. WAHLSTER 2000) die Möglichkeit, nach 3 s Spracheingabe zu entscheiden, ob der Sprecher eine der Sprachen Deutsch, Japanisch oder Englisch spricht. Dabei werden zwei Spracherkenner für Deutsch und Englisch parallel benutzt und über Schwellwerte der Erkennungskonfidenzen für die Wörter in dem betrachteten kurzen Zeitintervall wird dann eine Entscheidung getroffen, wobei in 87 % der Fälle im Testkorpus schon mit diesem einfachen Ansatz die richtige Sprache automatisch korrekt identifiziert werden kann.

Häufig wird ein Worthypothesengraph als Ausgabe der Spracherkennung und als Eingabe für die Sprachanalyse verwendet. Dabei werden unterschiedliche Worthypothesen den jeweiligen Zeitintervallen zugeordnet und mit einem Wahrscheinlichkeitswert versehen.

Jeder Pfad durch den Graphen, der das gesamte Eingabesignal zeitlich überspannt, stellt dann eine mögliche Interpretation der akustischen Eingabe dar. Die Unsicherheit bei der Interpretation wird durch die Anzahl der verschiedenen Pfade in dem Graphen visualisiert, wie Abbildung 9 für die intendierte Lesart „Dann möchte ich doch lieber ins Kino gehen" zeigt. Hier wurden z. B. statt „dann" auch „da" und statt „möchte" auch „werde", „will" und „würde" als Worthypothesen generiert.

Erfolgreiche sprecherunabhängige Spracherkennung mit geringer Wortfehlerrate basiert heute fast ausschließlich auf Hidden-Markov-Modellen (HMM), die mit Hilfe maschineller

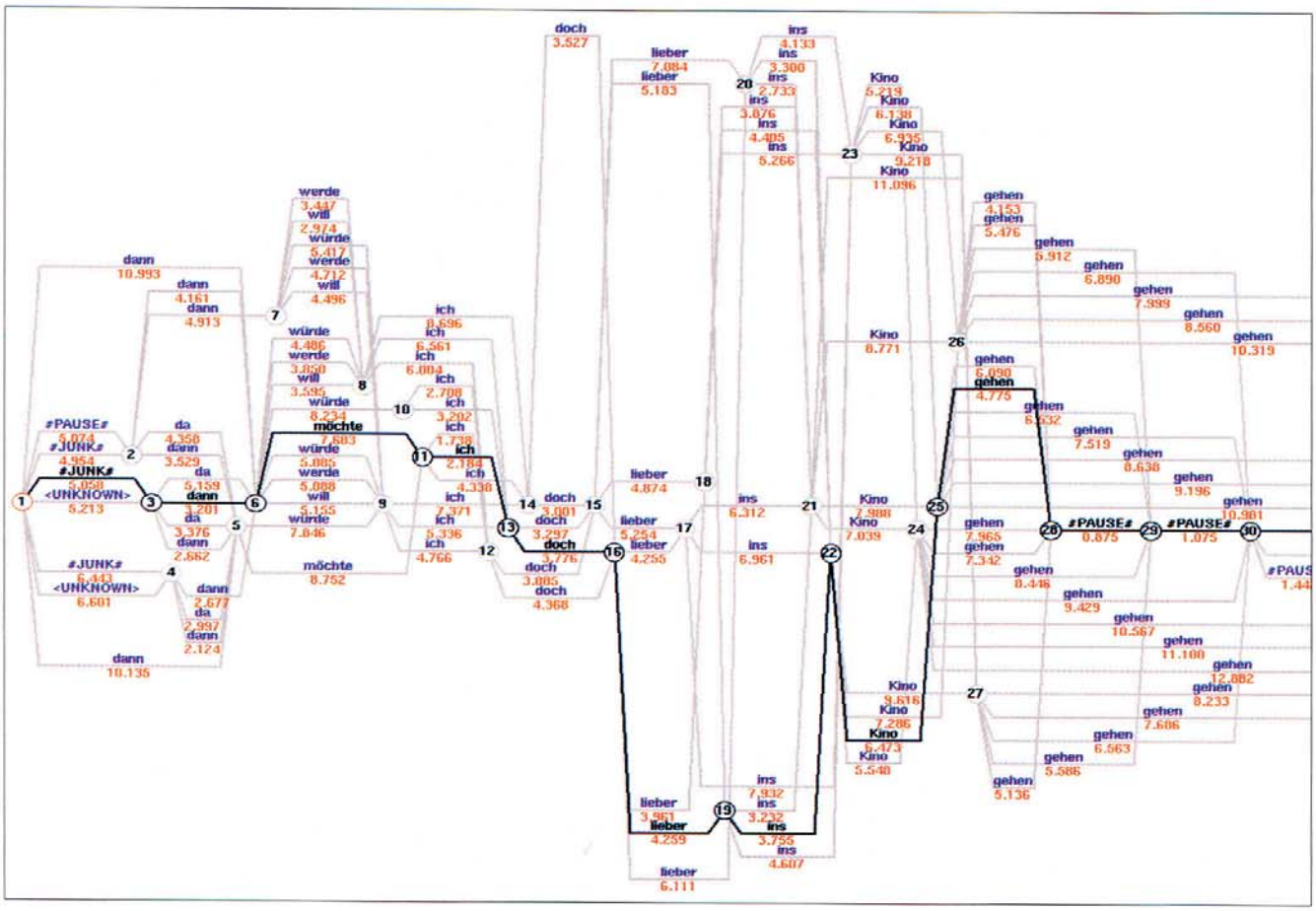


Abb. 9 Worthypothesengraph für „Dann möchte ich doch lieber ins Kino gehen“

Lernverfahren trainiert werden. Dabei entscheiden heute hauptsächlich der Umfang und die Güte der Annotationen der zum Training verwendeten Datensammlungen über die Qualität des Spracherkenners. Diese Datensammlungen enthalten vielfältig annotierte Sprachsignale, um u. a. die prosodisch markierten Satzgrenzen und Dialogakte auf der Basis gelernter probabilistischer Verfahren erkennen zu können. In Verbmobil haben wir eine der weltweit größten und sehr feingranular annotierten Datensammlungen von 21,5 GB mit 1685 Sprechern und 3200 Dialogen erstellt. Bisher stellen solche Datensammlungen als Grundlage der maschinellen Lernverfahren hohe Arbeitsaufwände dar, da die meisten Annotationen im besten Falle semi-automatisch erstellt werden können.

Für die Sprachanalyse kann man heute auf eine große Auswahl von formalen Grammatiken weitverbreiteter Sprachen wie Englisch, Deutsch und Französisch zurückgreifen, die mithilfe von für den jeweiligen Grammatikformalismus geeigneten Parsern für die syntaktische Analyse der wahrscheinlichsten Pfade in Worthypothesen benutzt werden.

Dabei entstehen im Gegensatz zur sehr einfachen Syntaxanalyse von künstlichen Sprachen wie den Programmiersprachen für Computer allerdings durch die große Variabilität in der Syntax natürlicher Sprachen, deren Kontextsensitivität sowie durch die Vielzahl elliptischer Konstruktionen, Selbsteinbettungen, lokaler Umstellungen und Fernabhängigkeiten herausfordernde Komplexitätsprobleme (vgl. Abb. 10). Während ein Parser für kontextfreie Sprachen nur $O(n^3)$ Verarbeitungsschritte für eine Syntaxanalyse benötigt, muss man für die schwach kontextfreien Baumadjunktionssprachen, die als einer der ausdrucksstärkeren Grammatikformalismen für natürliche Sprachen geeignet sind, bereits $O(n^5)$ Rechenschritte bei der Syntaxanalyse aufwenden, wobei n die Länge des eingegebenen Satzes ist.

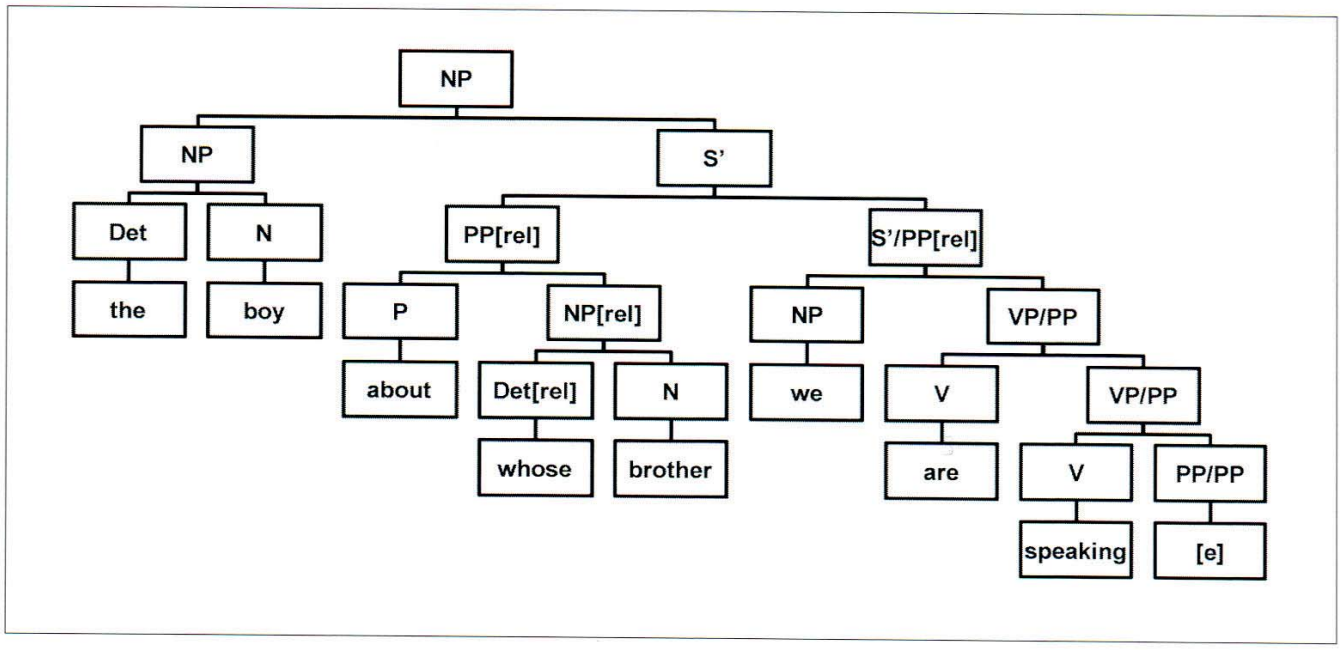


Abb. 10 Von einem Parser erstellte Syntaxanalyse mit rekursiven Einbettungen und Elision

Da sich die Bedeutung komplexer sprachlicher Äußerungen – von idiomatischen Wendungen abgesehen – kompositionell und geleitet durch die syntaktische Struktur bestimmen lässt, ist das Ergebnis der Syntaxanalyse entscheidend für die Konstruktion einer semantischen Repräsentation einer Äußerung. Allerdings gibt es heute weltweit noch kein System, das alle (zumindest wohlgeformten) Äußerungen einer Person mithilfe von formalen Grammatiken bis ins Detail linguistisch korrekt analysieren kann. Man behilft sich in praktischen Anwendungen damit, in den Segmenten des Worthypothesengraphen, wo die detailgenaue grammatische Analyse z. B mit einer HPSG-Grammatik versagt, eine grobe Phrasenstruktur zumindest approximativ etwa mit einem Chunk-Parser oder statistischen Parser zu ermitteln (vgl. Abb. 8). Man spricht dann von hybrider Verarbeitung, in welcher eine tiefe syntaktische Analyse von Fragmenten einer Äußerung mit einer flachen Analyse der restlichen Segmente mithilfe approximativer Verfahren kombiniert wird.

Eine der klassischen Methoden zur Disambiguierung ist die Verwendung von Selektionsrestriktionen und Weltwissen, wie folgendes einfache Beispiel veranschaulicht. Wenn ein Vater im künftigen Cyber-Restaurant dem Service-Roboter sagt: „Die Apfelschorle trinkt meine Tochter, die Weinschorle meine Frau", dann führt die syntaktische Präferenz zunächst zu einer nicht sinnvollen Lesart wie *trinkt* (*Agens. Apfelschorle*, *Objekt. Tochter*) ∧ *trinkt* (*Agens. Weinschorle, Objekt: Frau*). Wenn das System die Selektionsrestriktion *trinkt* (*Agens. Mensch, Objekt. Getränk*) auf das ontologische Weltwissen Apfelschorle, Weinschorle ⊂ Getränk und Tochter, Frau ⊂ Mensch anwendet, wird die richtige Lesart des Eingabesatzes als *trinkt* (*Agens. Tochter Objekt: Apfelschorle*) ∧ *trinkt* (*Agens. Frau, Objekt: Weinschorle*) selektiert. Verbmobil war das erste Dialogübersetzungssystem, das neben Weltwissen auch den Kontext des vorausgehenden Satzes zur lexikalischen Disambiguierung und adäquaten Übersetzung nutzt. So wird nach „Nehmen wir dieses Hotel, ja?" das Wort Platz in „Ich reserviere einen Platz" mit *room* übersetzt, während nach „Gehen wir zum Abendessen" das englische Wort *table* für Platz verwendet wird. Im Extremfall wird sogar die interne Systemuhr zur Äußerungszeit herangezogen, wenn vor 13 Uhr der Eingabesatz „Lassen Sie uns zusammen Essen gehen" mit „Let's have lunch together" und nach 17 Uhr als „Let's have dinner together"

übersetzt werden muss. Ein wesentliches Ergebnis der Forschung im Bereich Computerlinguistik ist, dass kein universeller Disambiguierungsalgorithmus existiert, sondern nur mithilfe der Kombination von Evidenzen aus verschiedenen Wissensquellen und durch die Fusion von Ergebnissen konkurrierender Verarbeitungspfade eine eindeutige Interpretation natürlichsprachlicher Dialogbeiträge erreicht werden kann.

Große Fortschritte hat es in der letzten Dekade auch auf dem Gebiet der Dialogmodellierung gegeben. Nur auf der Basis der von Sprachdialogsystemen schrittweise aufgebauten Diskursmodelle ist es möglich, komplexe Diskursphänomene wie die Verwendung von elliptischen Äußerungen und pronominalen Referenzen, Klärungsfragen, die Steuerung des Rednerwechsels oder metakommunikatives Verhalten maschinell zu bearbeiten. Um sich an einen einzelnen Benutzer anpassen zu können, verwenden viele Sprachdialogsysteme auch sogenannte Benutzermodelle (vgl. WAHLSTER und KOBSA 1989) die dynamisch aufgebaut werden, um anschließend bei der Sprachanalyse und der Sprachgenerierung verwendet zu werden. Damit kann sich das Sprachdialogsystem während eines Gesprächs an einen speziellen Benutzer anpassen, wobei nicht nur der Sprechstil, das geschätzte Alter und das Geschlecht, sondern auch dessen vermutete Ziele, Vorlieben und Vorkenntnisse das Dialogverhalten des Systems bestimmen.

Als sehr fruchtbar hat sich auch die Idee erwiesen, die Dialogführung als planbasiertes Verhalten (vgl. WAHLSTER et al. 1993) zu modellieren, wobei eine Sequenz von Dialogakten abhängig vom Verhalten des Dialogpartners so geplant wird, dass letztlich das eigene Gesprächsziel erreicht wird. Dabei werden in den am weitesten entwickelten Computermodellen Algorithmen aus dem Bereich der Künstlichen Intelligenz für die Planerkennung, die Plangenerierung und die Überwachung der Planausführung eingesetzt.

Komplexes Sprachverhalten mit emotionalen, ironischen, humoristischen oder persuasiven Dialogbeiträgen kann heute nur sehr begrenzt in Laborsystemen modelliert werden. Hier zeigt sich, dass die Interpretationsunsicherheit über längere Segmente eines Dialoges bestehen bleibt und im Extremfall erst nach Ende des Gesprächs durch zusätzliche Wissensquellen aufgehoben werden kann.

Nach der Planung einer angemessenen Systemreaktion erfolgt zunächst die Sprachgenerierung ausgehend von einer semantischen Repräsentation des geplanten Dialogbeitrages.

Hierzu werden in Dialogsystemen häufig schablonenbasierte Ansätze oder lexikalisierte Baumadjunktionsgrammatiken verwendet. Eine besondere Herausforderung liegt dabei in der inkrementellen Sprachgenerierung. Oft erfordert es der situative Zeitdruck, bereits ein Fragment zu äußern, obwohl der gesamte Inhalt einer Äußerung noch nicht feststeht. Typisch hierfür sind z. B alle Arten von Warn-, Alarmierungs- und Steuerungssystemen, in denen der Hörer unter Realzeitbedingungen auf sprachliche Systemausgaben reagieren muss oder eine zu einem laufenden Prozess schritthaltende Verbalisierung erfolgen muss (etwa in der Funktion eines Sportreporters). Ein Äußerungsanfang, der wegen Zeitdrucks schon artikuliert wurde, obwohl er auf der Basis unvollständiger und damit nachträglich änderbarer Eingabeinformation erzeugt worden war, kann nicht immer ohne Revision in Form von Selbstkorrekturen des Systems fortgesetzt werden. Dazu wurden in Laborsystemen Rechtfertigungsnetze implementiert, welche die bei der inkrementellen Sprachgenerierung notwendigen nicht-monotonen Verarbeitungsverfahren effizient unterstützen.

Schließlich muss eine geplante Äußerung des Systems noch akustisch realisiert werden. Hierzu werden heute bei der Computermodellierung konkatenative Sprachsyntheseverfahren verwendet.

Bei der konkatenativen Sprachsynthese werden aus großen annotierten Sprachdatenbanken zur Laufzeit die im Aussprachekontext geeignetsten Bausteine ausgewählt und zusammengesetzt.

Dabei werden für jedes Wort eine Vielzahl von Sprachaufnahmen des gleichen Sprechers in ganz unterschiedlichen Betonungen und Satzpositionen benötigt. Dabei soll das originale Sprachsignal so wenig wie möglich manipuliert werden, indem das am besten passende Synthesesegment verwendet wird (vgl. Abb. 11). Bei einer solchen korpusbasierten Sprachsynthese, wie sie heute in den meisten kommerziellen TTS-Systemen (*Text-to-Speech*-Systemen) angeboten wird, kann natürlich nur dann auf Wortebene konkateniert werden, wenn das zu sprechende Wort in der passenden Kontextvariante im Korpus gefunden wird. Sonst muss auf Subwort-Einheiten (z. B Silben oder Phoneme) ausgewichen werden. Bei sehr großem Korpus und entsprechend viel verfügbarem schnellem Hauptspeicher zur Korpussuche ergibt sich eine hohe Natürlichkeit und eine gute Sprachqualität, die fast kaum noch von menschlichen Sprechern zu unterscheiden ist.

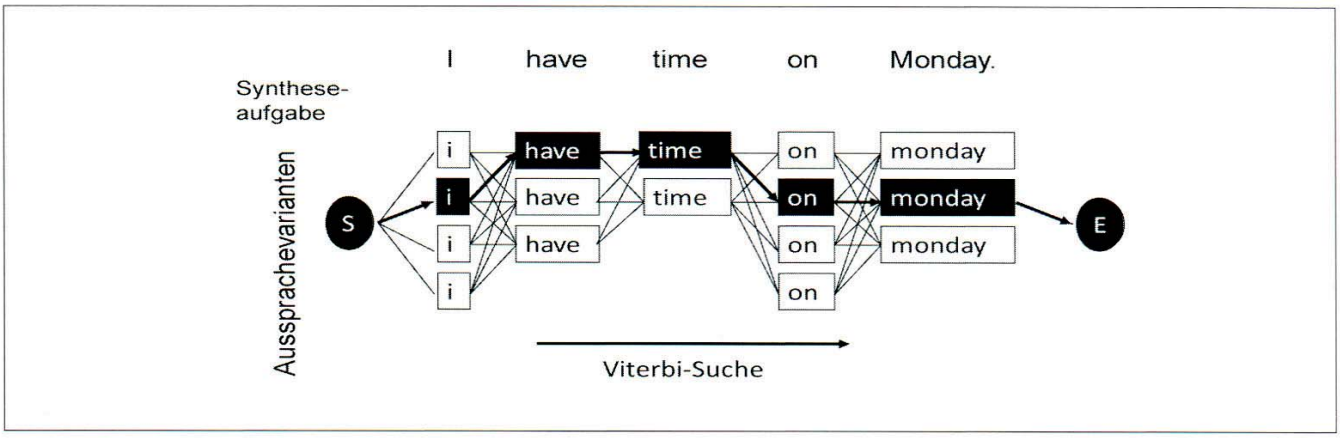


Abb. 11 Konkatenative Sprachsynthese als Selektion von Aussprachevarianten

Da natürliche Sprachdialoge immer von anderen Kommunikationsmodalitäten wie Gestik, Mimik, Körperhaltung und oft auch von physischen Aktionen begleitet werden, wurde die Computermodellierung in den letzten 20 Jahren auch verstärkt auf multimodalen Dialog verallgemeinert. Dabei hat sich gezeigt, dass oft eine wechselseitige Disambiguierung multipler Eingabe- und Ausgabemodalitäten bei der Reduktion von Interpretationsunsicherheiten große Vorteile bietet. Im nächsten Abschnitt werden daher multimodale Dialogsysteme präsentiert.

4. Multimodale Mensch-Computer-Interaktion: Sprache, Gestik, Mimik und Agieren

Bei der Computermodellierung multimodalen Dialogverhaltens müssen die Ergebnisse der akustischen und visuellen Analyse zur Sprach-, Gestik- und Mimikerkennung zu einer gemeinsamen Bedeutungsrepräsentation zusammengeführt werden. Diesen Prozess nennt man multimodale Fusion (vgl. BUNT et al. 2005). Wichtig für die Fusionsverfahren ist, dass alle unimodalen Erkenner Konfidenzwerte für ihre Erkennungsresultate liefern.

Die multimodale Fission bietet die inverse Funktionalität auf der Ausgabeseite, wobei zunächst für eine semantischen Repräsentation des vom System geplanten Dialogbeitrags eine Allokation der Informationseinheiten auf die verschiedenen Modalitäten und danach eine synchronisierte Ausgabe koordinierter Sprache, Gestik und Mimik realisiert werden muss. Man

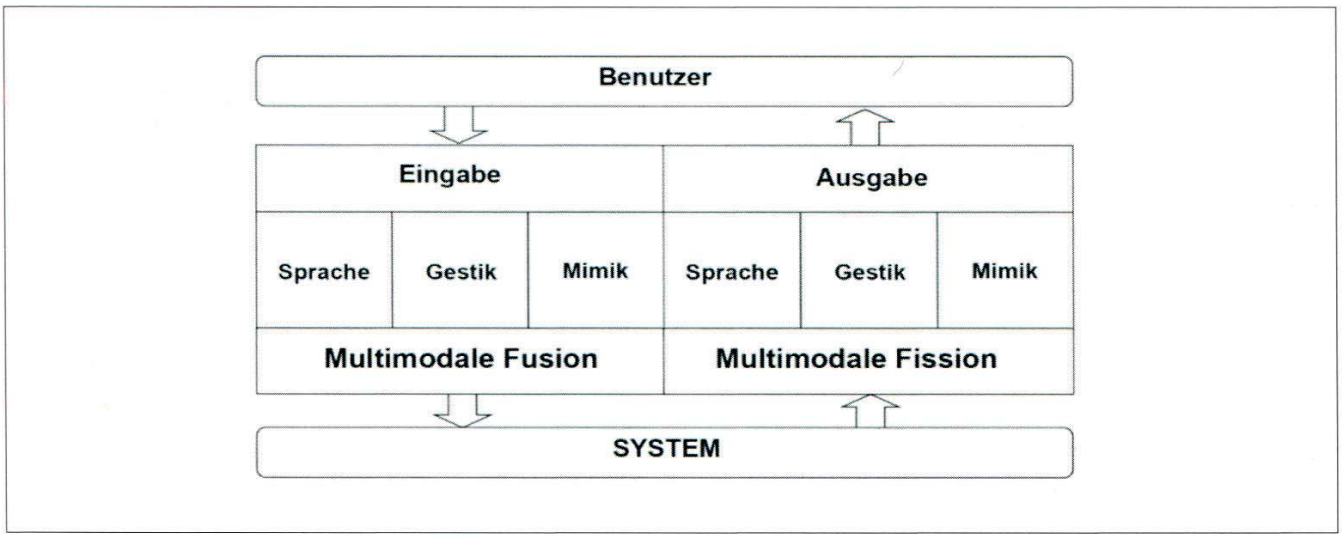


Abb. 12 Multimodale Fusion und Fission

spricht von symmetrischer Multimodalität, wenn alle Eingabemodalitäten auch für die Ausgabe verfügbar sind und umgekehrt (vgl. Abb. 12).

Um auf der Ausgabeseite Sprache mit Gestik und Mimik zu koordinieren, werden oft virtuelle Charaktere eingesetzt (vgl. NDIAYE et al. 2005). Dabei handelt es sich um autoanimierte graphische Figuren, die mit modernen Verfahren der Computergraphik oft eine fotorealistische Anmutung haben. Beispielsweise nutzt das System Ligabot (vgl. Abb. 13) einen solchen virtuellen Charakter zur multimodalen Ausgabe von Dialogbeiträgen. Der Ligabot ist ein virtueller Fußballreporter und ermöglicht es, mit gesprochener Sprache aktuelle Information über Spielergebnisse, Spielorte, nächste Begegnungen und die Tabellensituation der laufenden Fußballbundesliga-Saison zu erfragen.

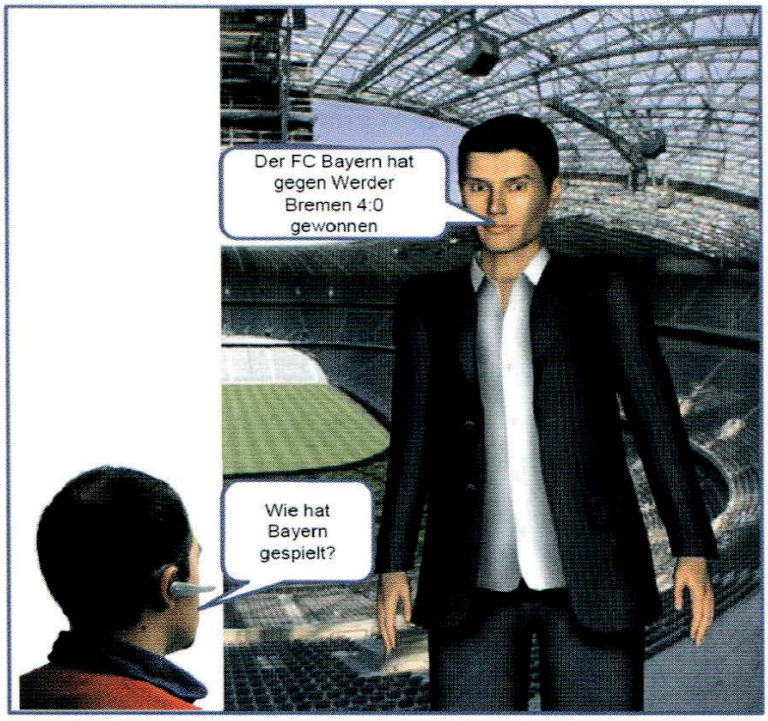


Abb. 13 Ligabot: Dialoge mit einem virtuellen Fußballreporter

Das System wird seit 2006 als Dauerexponat des Deutschen Museums in München im Zusammenhang mit dem Deutschen Zukunftspreis von den Museumsbesuchern benutzt. Ligabot greift über Internet auf die neuesten Ergebnisse und den Spielplan der Bundesliga zu und setzt diese in umgangssprachlich formulierte Antworten um. Das Dialogsystem kann in der beschränkten Diskursdomäne auch auf viele idiomatische Formulierungen wie „Wer hat die rote Laterne?" oder „Wie haben die Königsblauen gespielt?" korrekt antworten. Hierbei werden die Dialogbeiträge des virtuellen Gesprächspartners durch passende Mimik und Gestik unterstützt, und eine intelligente Kameraführung in einem 3D-Modell des Münchner Fußballstadions erhöht die Natürlichkeit der Gesprächssituation. Ligabot zeigt, dass natürliche Dialoge mit virtuellen Gesprächspartnern mit den aktuellen Computermodellen über ein eingeschränktes Themenfeld in hoher Qualität und Robustheit möglich sind. Ligabot versteht aber überhaupt nichts außerhalb der Fußballdomäne. Die Beschränkung der Domäne wird im Gegenteil beim Sprachverstehen genutzt, um die Interpretationsunsicherheiten drastisch zu reduzieren.

Es hat sich herausgestellt, dass eine kombinierte Sprach- und Gestikverarbeitung die Robustheit von Dialogsystemen erhöht. Wenn man den akustischen Kanal mit einem Videokanal verbindet und die Analyseergebnisse einer Fusion unterwirft, wird die Disambiguierung oftmals erheblich vereinfacht. So erhöht sich die Robustheit von Spracherkennern bei gestörtem Sprachsignal und niedriger Worterkennungsrate durch gleichzeitiges Lippenlesen. Durch eine Kombination von Spracherkennung und Prosodieanalyse konnte im Verbmobil-System sowohl eine verbesserte lexikalische als auch syntaktische Disambiguierung von Dialogbeiträgen erreicht werden. Im SmartKom-System (vgl. WAHLSTER 2006) wurden die referenzsemantische Disambiguierung und die Aufmerksamkeitssteuerung durch die kombinierte Sprach- und Gestikerkennung wesentlich verbessert.

Durch die Kombination von Spracherkennung und Mimikerkennung wurde im SmartKom-System erstmals sogar die Erkennung von Ironie möglich. Dort kann die gleiche Spracheingabe des Benutzers je nach dem klassifizierten Gesichtsausdruck (ärgerlich *versus* neutral), der über ein Mustererkennungsverfahren aus den Daten einer Gesichtskamera analysiert wird, zu einer ironischen oder nicht-ironischen Interpretation einer Äußerung führen.

Wichtig ist die Erkenntnis, dass nicht nur die Sprache durch die Gestik und Mimik disambiguiert werden kann, sondern auch umgekehrt eine mehrdeutige Geste oder ein unklarer Gesichtsausdruck durch die damit kombinierte sprachliche Äußerung auf eine eindeutige Bedeutungsrepräsentation abgebildet werden kann. Deiktische Referenzen durch sprachbegleitende Gesten, die über eine Kamera erfasst werden, können sich beispielsweise im SmartKom-System auf dynamisch generierte Visualisierungen beziehen (vgl. Abb. 14).

Dabei kann die Geste selbst keine eindeutige Objektreferenz liefern, sondern nur die sprachliche Formulierung „Den Film" versetzt das System in die Lage „Matrix: Reloaded" als intendiertes Diskursobjekt zu identifizieren (es könnte ja u. a. auch der Buchstabe „d" gewesen sein). Beim Beispiel der Sitzplatzauswahl in Abbildung 14 wird die vage Referenz einer umkreisenden Geste erst durch die Formulierung „drei Sitze" eindeutig und durch eine visuelle Rückkopplung an den Benutzer durch grüne Kreuze vom System bestätigt. In etwas vereinfachter Form können solche mit Spracheingaben kombinierten Gesten auch ohne Bildauswertung einfach durch Fingerberührung kapazitiver Tastbildschirme von modernen Mobiltelefonen erfolgen.

Wenn Hände des Benutzers durch direktes Greifen und Ablegen von RFID-instrumentierten Objekten in ihrer Umgebung (z. B ein intelligentes Regal mit Leseantennen, die feststellen, wann ein Objekt mit einem *Radio Frequency Identification* (RFID)-Chip entfernt wird) eine

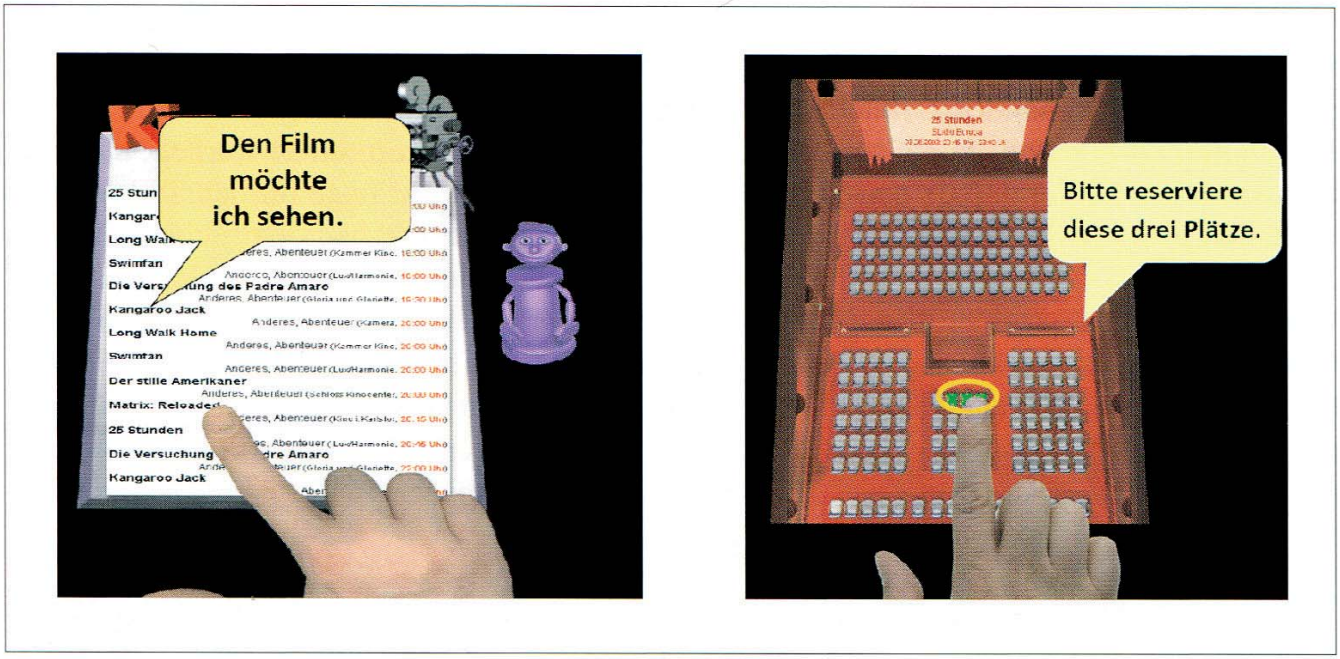


Abb. 14 Die multimodale Kombination von Sprach- und Gestikeingabe

referentielle Geste auslösen können, kann das einen sprachlichen Dialog mit dem Produkt initiieren, wie dies in Abbildung 15 illustriert wird. Die direkte Interaktion mit anthropomorphisierten Produkten (vgl. WAHLSTER et al. 2010), die aufgrund physischer Aktionen des Benutzers scheinbar selbst die Dialoginitiative ergreifen und sich sprachlich vorstellen, führt in diesem Fall zum Eindruck eines sprechenden Produktes in Form einer Weinflasche. Mit den aktuellen Verfahren zur Sprachsynthese lässt sich erreichen, dass ein französischer Rotwein die Fragen des Kunden scheinbar mit französischem Akzent beantwortet, während sich eine Canon-Digitalkamera mit dem typischen japanischen Tonfall präsentiert. Die digital veredelte Weinflasche enthält in unserer Installation des digitalen Sommeliers auch einen Funksensor, der laufend die Temperatur und Helligkeit erfasst, so dass auch Fragen über das „digitale Tagebuch" einer einzelnen Weinflasche beantwortet werden können.

Abb. 15 Sprechende Produkte als anthropomorphisierte Alltagsobjekte

Durch die Kombination von physischen Handlungen mit Spracheingaben lässt sich eine semantisch-pragmatische Disambiguierung im situativen Kontext durchführen. So kann z. B im Auto das Kommando „Schneller" als Aufforderung zur Einstellung einer höheren Geschwindigkeit des Scheibenwischers oder des Tempomaten interpretiert werden, abhängig davon, ob der Fahrer gerade den Scheibenwischerhebel oder die Taste für den Tempomaten bedient hat (vgl. Abb. 16).

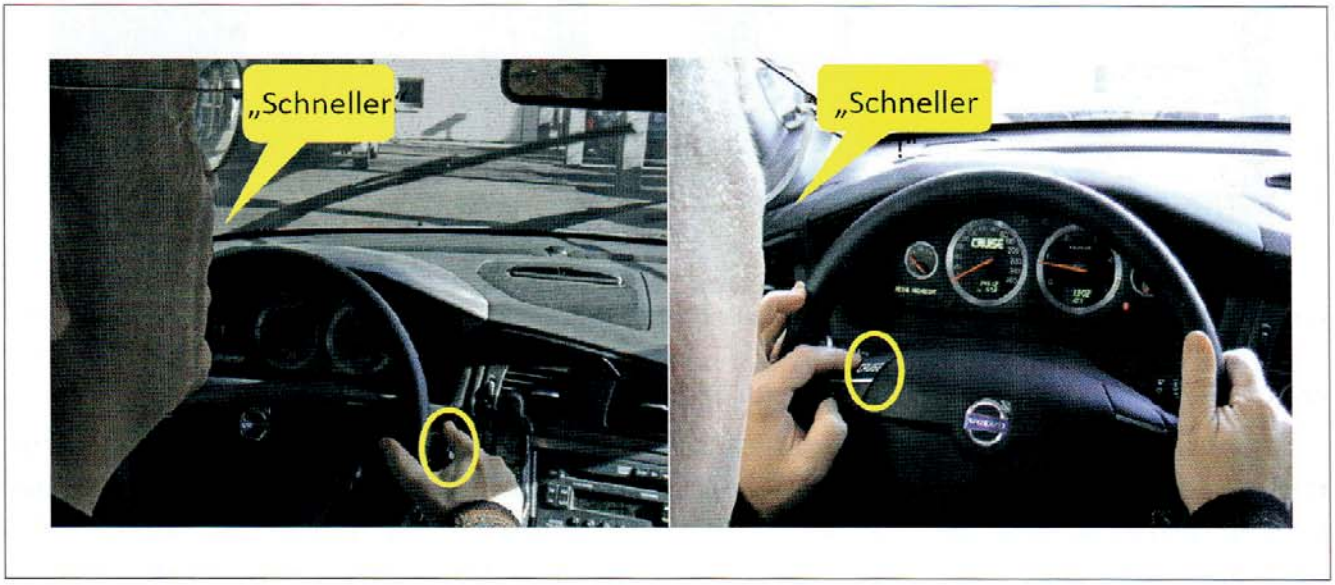


Abb. 16 Disambiguierung der Spracheingabe durch Bedienaktionen

Während solche komplexen Computermodelle für die sprachliche Interaktion in instrumentierten Umgebungen mit ambienter Intelligenz heute erst in Forschungslaboren erprobt werden, haben einfachere Sprachdialogsysteme den Weg in den Alltag gefunden. Dadurch haben sich Anwendungen der Sprachtechnologie als ein neuer Wirtschaftszweig etabliert.

5. Sprachtechnologie im Alltag

Sprachtechnologie begegnet uns heute bereits überall im Alltag. Telefonbasierte Auskunfts- und Kundendienstzentren setzen Sprachdialogsysteme verstärkt ein, zumindest um das Gespräch mit dem persönlichen Kundenberater schon durch eine automatisierte Vorqualifikation des Anrufenden vorzubereiten und zu verkürzen. Die neueste Generation von SmartPhones, wie das iPhone oder das Nexus One von Google, bieten das sprachgesteuerte Wählen, die Suche in Kontaktlisten, die Stichworteingabe per Sprache sowie das Diktieren von SMS- und E-Mail-Nachrichten in mehreren Sprachen an.

Viele Banken bieten heute ein Sprachdialogsystem zur automatisierten Abwicklung einfacher Finanztransaktionen über Telefon an. Die Systeme der ersten Generation mit geringer Kundenakzeptanz durch Einzelworterkennung, schlecht verständlicher Sprachsynthese und rudimentärem Sprachverständnis wurden inzwischen fast vollständig durch Systeme der zweiten Generation mit akzeptablem Dialogverhalten sowie hochwertiger Erkennung und Synthese kontinuierlicher Spontansprache abgelöst. Durch die rasche Verbreitung fortgeschrittener Sprachapplikationen außerhalb der telefonbasierten Anwendungen – beispielsweise im Bereich der Adresseingabe für Navigationssysteme – haben immer mehr Kunden bereits Erfahrungen

mit der Bedienerfreundlichkeit von aufgabenorientierten Sprachdialogsystemen und lehnen daher die umständlichen Auswahlmenüs und unkooperative Dialogführung von Systemen der ersten Generation vermehrt ab. Das Erkennungsvokabular ist in den besten Systemen von wenigen hundert auf mehrere zehntausend Wörter angewachsen. Die Eingabe mehrerer Zielparameter in einer einzigen Formulierung und die Möglichkeit zum spontanen Wechsel der Dialoginitiative zwischen Nutzer und System werden daher zur Grundvoraussetzung für kommerziell erfolgreiche Sprachdialogsysteme.

Vermehrt wird auch die Datenerfassung vor Ort über Mobilfunk in Sprachdialogen möglich. So ermöglicht beispielsweise E.ON seinen Strom- und Gaskunden rund um die Uhr die jährliche Zählerstandserfassung im Selbstservice per Sprachdialogsystem mit der vollständigen Erfassung von Kundennummer, Zählernummer, Ablesedatum und Zählerstand sowie mit einer Plausibilisierung des Zählerstandes durch Datenbankabgleich. Das System versteht auch gruppierte Ziffern (z. B „zehn, zwölf" als Alternative zu „eins, null, eins, zwei") und kann relative Datumsangaben (z. B „am Sonntag") in konkrete Datenbankeinträge umsetzen. Viele nützliche Serviceangebote werden heute über Sprachdialogsysteme angeboten. So bietet die Bundesvereinigung der Apotheker eine lokale Suche nach Apotheken mit Notdienst an, die über 15000 deutsche Ortsnamen versteht, und bei IKEA kann man über ein Sprachdialogsystem feststellen, ob ein gewünschtes Produkt in einem bestimmten Einrichtungshaus im Lager vorhanden ist. Durch biometrische Stimmidentifikation können Sprachdialogsysteme auch in sicherheitskritischen Anwendungen eingesetzt werden, beispielsweise zur automatisierten Rücksetzung eines Zugangspasswortes.

Im Fahrzeug können immer mehr sicherheitsunkritische Funktionen über Sprachdialoge gesteuert werden. In einem R-Klasse-Mercedes-Testfahrzeug des Deutschen Forschungszentrums für Künstliche Intelligenz (DFKI) können alle Komfortfunktionen, die über den sogenannten CAN-Bus (*Controller Area Network*) in Realzeit gesteuert werden können, vom Fahrer per Sprache über die Freisprechanlage des Fahrzeuges ausgelöst werden (vgl. Abb. 17). Durch die Verwendung von natürlichsprachlichen Quantoren („öffne alle Fenster") und sprachlichen Raumreferenzen („die vorderen Fenster", „das Fenster hinten links") wird eine

Abb. 17 Sprachsteuerung von Komfortfunktionen eines Automobils

intuitive Bedienung ohne Zugriff auf Handbücher und ohne komplexe Bedienoperationen möglich, so dass sich der Fahrer voll auf das Fahrgeschehen konzentrieren kann.

Natürlich bietet es sich in vielen Situationen an, auch mit Robotern sprachlich zu interagieren. Obwohl es sehr viele Forschungsprojekte zu dieser Aufgabenstellung gibt, existieren sprachgesteuerte Roboter auf dem Markt fast ausschließlich im Bereich der Unterhaltungsroboter.

Da in diesem Fall die gesamte Sprachverarbeitungssoftware auf dem eingebetteten Prozessor des Spielroboters zusätzlich zur Bilderkennung und Bewegungsplanung laufen muss, sind die Sprachdialoge bei diesen Systemen allerdings sehr eingeschränkt. Den kommerziellen AIBO-Roboterhund (vgl. Abb. 18) kann man nach seinem Namen und Alter fragen, ihn loben, ermahnen und einfache Anweisungen wie „Gib Pfötchen" „Steh auf" „Suche den Ball" und „Lass uns tanzen" geben. In der einfachen Grundversion versteht AIBO nur 180 Eingabevarianten und reagiert darauf kontextabhängig mit leuchtenden Augen, Schwanzwedeln und den befohlenen Bewegungen.

Der humanoide QRIO-Roboter, der wie AIBO ebenfalls von Sony entwickelt wurde, aber nicht kommerziell angeboten wird, kann einfache Gespräche mit simplen Dialogmustern mit seinen menschlichen Mitspielern führen und wendet sich dabei bei Antworten in Richtung des Fragenden, weil er über eine einfache akustische Erkennung der Position von Sprechern im Raum verfügt.

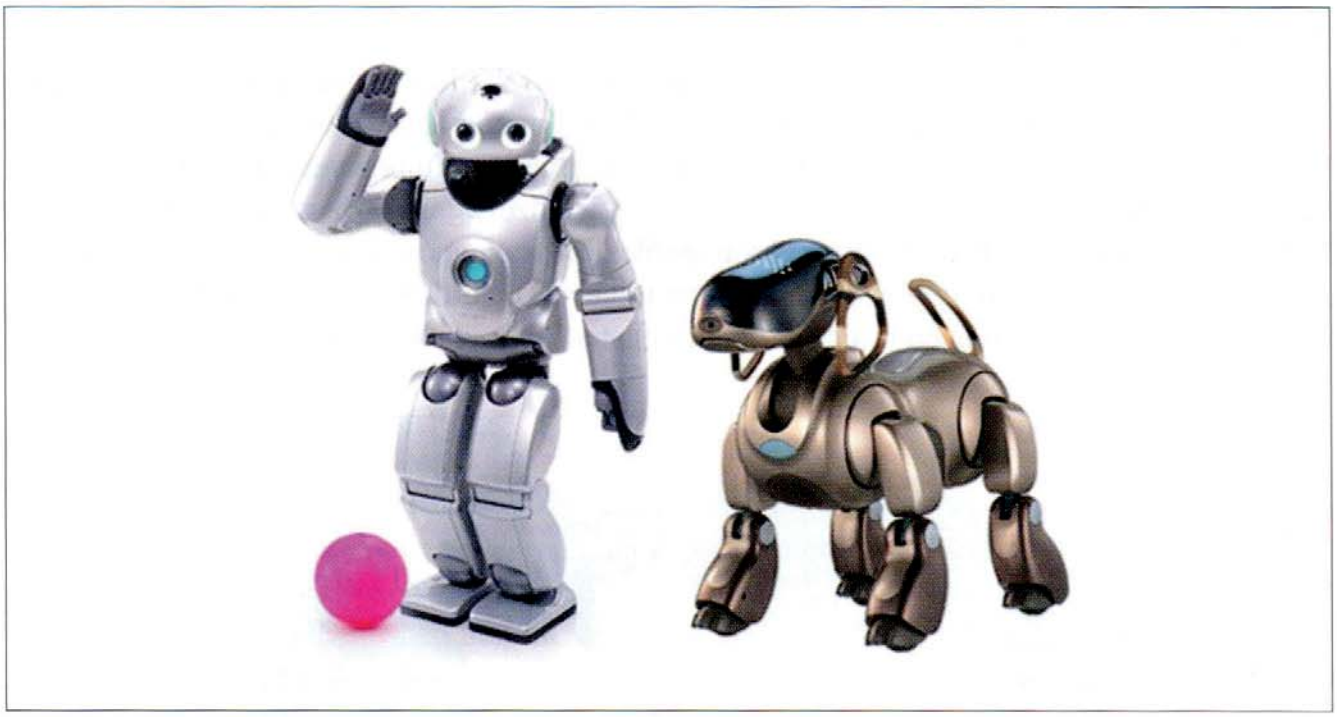


Abb. 18 Sprachgesteuerte Unterhaltungsroboter QRIO und AIBO

6. Rückblick und Ausblick: Vier Dekaden der deutschen Sprachtechnologieforschung

Die Forschung zur Computermodellierung von Sprachverhalten wurde in Deutschland seit Mitte der 1970er Jahre intensiv gefördert (vgl. REUSE und VOLLMAR 2008). Man hatte erkannt, dass die maschinelle Verarbeitung der deutschen Sprache überwiegend in Deutschland vorangetrieben werden muss, wenn man bei der weiteren Entwicklung der Informations- und

Internetgesellschaft nicht vom englischsprachigen Sprachraum dominiert werden möchte. Außerdem gab es durch die lange Tradition exzellenter deutscher Forscher in der Sprachwissenschaft, der Sprachpsychologie, der Sprachphilosophie und durch interdisziplinäre Zusammenarbeit mit der aufblühenden neuen Disziplin der Informatik beste Voraussetzungen für eine erfolgreiche „elektronische Sprachforschung", wie der erste DFG-Sonderforschungsbereich (1972–1986) zu diesem Thema am Standort Saarbrücken hieß.

Man kann die Forschungen zur Sprachtechnologie in Deutschland grob in vier Dekaden einteilen, die jeweils durch große Verbundvorhaben geprägt wurden (vgl. Abb. 19). Die ersten vollständig implementierten Dialogpartnermodelle wurden mit HAM-RPM (vgl. VON HAHN et al. 1980) und HAM-ANS an der Universität Hamburg entwickelt, konnten aber nur einfache Frage-Antwort-Dialoge in Miniwelten (z. B Hotelbeschreibung, Verkehr an einer Kreuzung) führen. Anschließend wurden in weiteren DFG-Sonderforschungsbereichen wie „Künstliche Intelligenz – Wissensbasierte Systeme" (1985–1995) und „Ressourcenadaptive kognitive Prozesse" (1996–2007) speziell in Saarbrücken die Grundlagen der natürlichsprachlichen Systeme weiterentwickelt. Während bis Mitte der 1980er Jahre noch reine Grundlagenforschung betrieben wurde, begannen danach auch Forschungsgruppen in großen Unternehmen wie Siemens, Philips, AEG, Daimler und SEL anwendungsorientiert an der Computermodellierung von Sprachverhalten zu arbeiten. Gleichzeitig wurden in der Grundlagenforschung von Deutschland aus die neuen Themen des wissensbasierten Textverstehens (vgl. HERZOG und ROLLINGER 1991), der Benutzermodellierung und der sprachlichen Bildbeschreibung (vgl. HERZOG et al. 1989) vorangetrieben, mit denen deutsche Forscher als Pioniere weltweit bekannt wurden. Durch die großen und langfristig angelegten BMBF-Verbundprojekte Verbmobil und SmartKom wurden dann die Dialogübersetzung und die multimodale Interaktion stark vorangetrieben.

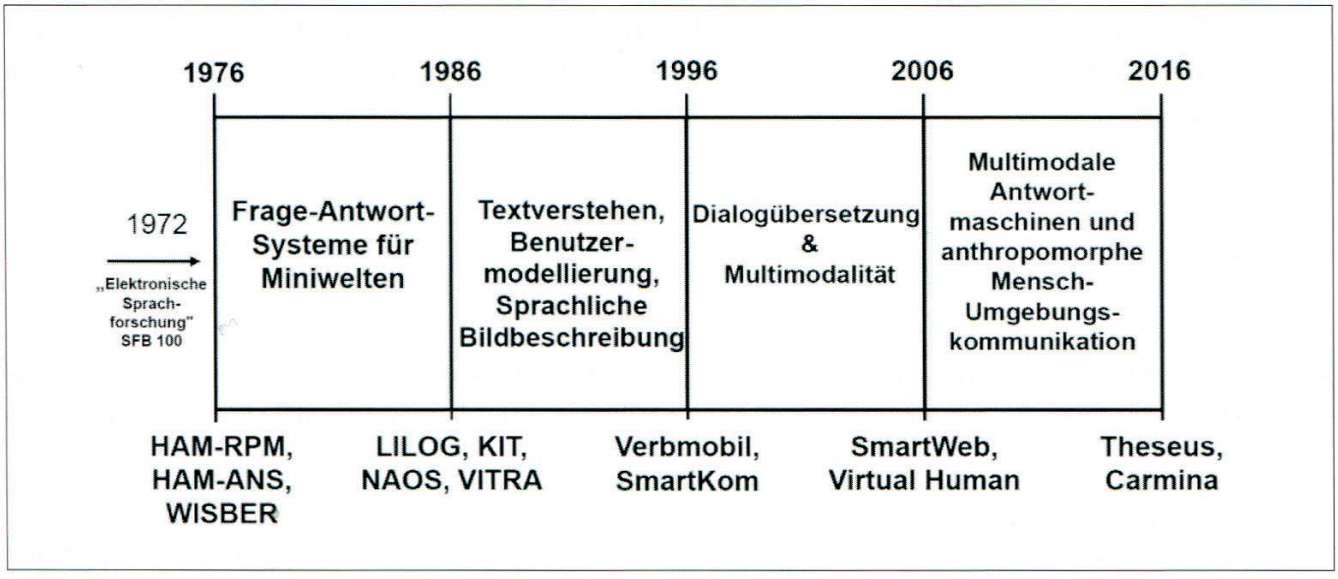


Abb. 19 Vier Dekaden der Forschung zur Sprachtechnologie in Deutschland

Zunächst mit dem BMBF-Verbundprojekt SmartWeb und derzeit in dem BMWi-Konsortium Theseus wurde der Übergang der Forschung von abgeschlossenen Diskursbereichen zu offenen Anfragen an das Web in Form von semantischen Antwortmaschinen (vgl. FENSEL et al. 2003) geschaffen. In den letzten 20 Jahren wurde es analog zur Experimentalphysik notwendig, immer größere Forschungsteams in Verbundvorhaben arbeiten zu lassen, wobei zur Realisie-

rung von komplexen Systemen wie Verbmobil bis zu 400 Forscher zu koordinieren waren. Das BMBF-Projekt Virtual Human (vgl. REITHINGER et al. 2006) erreichte einen Durchbruch bei der Mehrpersonenkommunikationen mit virtuellen Charakteren und legte die Basis für heute laufende Forschungen im Bereich der anthropomorphen Mensch-Umgebungskommunikation. Inzwischen sind in Deutschland aus den großen Verbundprojekten des BMBF zahlreiche Firmenneugründungen wie Sympalog, SemanticEdge, SemVox und Yocoy im Bereich der deutschen Sprachdialogsysteme entstanden, die kommerzielle Lösungen für den rasch wachsenden Markt der Sprachtechnologie anbieten. Die Grundlagenforschung wird heute besonders im Bereich der multimodalen Dialogsysteme in den beiden DFG-Exzellenzclustern „Multimodal Computing and Interaction" in Saarbrücken und „Cognitive Interaction Technology" in Bielefeld vorangetrieben.

Ein aktueller Forschungstrend in der Sprachtechnologie entsteht durch den Übergang von Nahbesprechungsmikrophonen zu Raummikrophonen, die häufig als Mikrophonarrays realisiert werden, um Anwendungen mobiler Sprecher ohne eigenes Mikrophon (z. B Besprechungssituationen, Fernsehbedienung vom Sofa aus) realisieren zu können. Im Dialogbereich wird verstärkt an der Verallgemeinerung der Ansätze für dyadische Dialoge auf Mehrpersonengespräche geforscht. Statt dem sprachlichen Zugang zum Web der Information steht derzeit der Sprachdialog im zukünftigen Internet der Dienste und der Dinge im Vordergrund zahlreicher Forschungsvorhaben. Erfolgreiche monolinguale Ansätze für die wichtigsten internationalen Verkehrssprachen werden immer mehr für multilinguale Systeme auch auf weniger verbreitete Sprachen adaptiert.

Die Frage im Titel dieses Beitrages „Autos, Roboter und Weinflaschen als Dialogpartner?" kann heute mit „Ja, aber " beantwortet werden. Die Dialogfähigkeiten sind im Vergleich zu menschlichen Gesprächspartnern immer noch stark eingeschränkt. Ein angemessenes und robustes Verstehen natürlicher Spontansprache durch virtuelle Dialogpartner ist derzeit nur für eng begrenzte Themenbereiche möglich. Trotzdem werden bereits heute alleine in Deutschland mit Sprachcomputern pro Tag über 400000 Gespräche mit einer Fallabschlussquote von 80 % geführt.

Literatur

ALLEN, J Natural Language Understanding. Addison-Wesley 1994

BUNT H., KIPP, M., MAYBURY M., and WAHLSTER, W.. Fusion and coordination for multimodal interactive information presentation. In: STOCK, O., and ZANCANARO, M. (Eds.): Multimodal Intelligent Information Presentation; pp. 325–340. Dordrecht: Kluwer 2005

COLE, R. A., MARIANI, J., USZKOREIT H., ZAENEN, A., and ZUE, V (Eds.): Survey of the State of the Art in Human Language Technology. Cambridge: Cambridge University Press 1997

FENSEL, D., HENDLER, J., LIEBERMAN, H., and WAHLSTER, W. (Eds.): Spinning the Semantic Web. Bringing the World Wide Web to Its Full Potential. Cambridge: MIT Press 2003

HAHN, W VON, HOEPPNER, W., JAMESON, A., and WAHLSTER, W The anatomy of the natural language dialogue system HAM-RPM. In: BOLC, L. (Ed.): Natural Language Based Computer Systems; pp. 119–253. München: Hanser/ Macmillan 1980

HERZOG, O., and ROLLINGER, C.-R. (Eds.): Text Understanding in LILOG: Integrating Computational Linguistics and Artificial Intelligence. Berlin, Heidelberg, New York: Springer 1991

HERZOG, G., SUNG, C.-K., ANDRÉ, E., ENKELMANN, W., NAGEL, H.-H., RIST T., WAHLSTER, W., and ZIMMERMANN, G.. Incremental natural language description of dynamic imagery. In: FREKSA, C., und BRAUER W. (Eds.): Wissensbasierte Systeme. S. 153–162. Berlin, Heidelberg, New York: Springer 1989

JURAFSKY D., and MARTIN, H. H.. Speech and Language Processing. Upper Saddle River, NJ: Prentice Hall 2000

MANNING, C. D., and SCHÜTZE, H.. Foundations of Statistical Natural Language Processing. Cambridge: MIT Press 1999

MÜLLER, C. (Ed.): Speaker Classification – Fundamentals, Features, and Methods. Berlin, Heidelberg, New York: Springer 2007

NDIAYE, A., GEBHARD, P., KIPP, M., KLESEN, M., SCHNEIDER, M., and WAHLSTER, W Ambient intelligence in edutainment: Tangible interaction with life-like exhibit guides. In: MAYBURY M., STOCK, O., and WAHLSTER, W (Eds.): Intelligent Technologies for Interactive Entertainment. Lecture Notes in Artificial Intelligence Vol. 3814, pp. 104–113. Berlin, Heidelberg, New York: Springer 2005

REITHINGER, N., GEBHARD, P., LÖCKELT M., NDIAYE, A., PFLEGER, N., RUMPLER, M., KIPP, M., BLOCHER, A., and WAHLSTER, W.. Dialogic and affective interaction with VirtualHuman. In: Proc. of the 29th Annual German Conference on Artificial Intelligence; pp. 45–48. Berlin, Heidelberg, New York: Springer 2006

REUSE, B., und VOLLMAR, R. (Eds.): Informatikforschung in Deutschland. Berlin, Heidelberg, New York: Springer 2008

RUSSELL, S., and NORVIG, P Artificial Intelligence: A Modern Approach. Prentice Hall 2009

WAHLSTER, W (Ed.): Verbmobil: Foundations of Speech-to-Speech Translation. Berlin, Heidelberg, New York: Springer 2000

WAHLSTER, W (Ed.): SmartKom: Foundations of Multimodal Dialogue Systems. Berlin, Heidelberg, New York: Springer 2006

WAHLSTER, W.. SmartWeb: Ein multimodales Dialogsystem für das semantische Web. In: REUSE, B., und VOLLMAR, R. (Eds.): Informatikforschung in Deutschland. S. 300–311. Heidelberg, Berlin, New York: Springer 2008

WAHLSTER, W., ANDRÉ, E., FINKLER, W., PROFITLICH, H.-J., and RIST T.. Plan-based integration of natural language and graphics generation. Artificial Intelligence *63*, 387–427 (1993)

WAHLSTER, W., FELD, M., GEBHARD, P., HECKMANN, D., JUNG, R., KRUPPA, M., SCHMITZ, M., SPASSOVA, L., and WASINGER, R.. The shopping experience of tomorrow: Human-centered and resource-adaptive. In: CROCKER, M. W., and SIEKMANN, J (Eds.): Resource-Adaptive Cognitive Processes; pp. 205–237 Berlin, Heidelberg, New York: Springer 2010

WAHLSTER, W., and KOBSA, A.. User models in dialog systems. In: KOBSA, A., and WAHLSTER, W. (Eds.): User Models in Dialog Systems; pp. 4–34. Berlin, Heidelberg, New York: Springer 1989

Prof. Dr. Dr. h.c. mult. Wolfgang WAHLSTER
Deutsches Forschungszentrum für Künstliche Intelligenz
und Fachrichtung Informatik der Universität des Saarlandes
Stuhlsatzenhausweg 3, Campus D3 2
66123 Saarbrücken
Bundesrepublik Deutschland
Tel.. +49 681 3025252
Fax: +49 681 3025383
E-Mail: wahlster@dfki.de