In: HNF (ed.): Alltag der Zukunft, Paderborner Podium 3, Paderborn: Schöningh 1999, pp. 18-37.

Sprachtechnologie im Alltag

Der Computer als Dialogpartner

Wolfgang Wahlster

Deutsches Forschungszentrum für Künstliche Intelligenz GmbH
Stuhlsatzenhausweg 3
66119 Saarbrücken
email: wahlster@dfki.de

Zusammenfassung: Die Rolle der Sprachtechnologie bei der Verwirklichung der Wissensgesellschaft wird anhand zahlreicher praktischer Einsatzbeispiele beschrieben. Der heutige Entwicklungsstand und die technischen Herausforderungen der Sprachtechnologie werden zusammengefaßt. Neben Systemen zur Sprachsteuerung von Geräten, automatischen Auskunftssystemen, Dolmetschsystemen und mobilen Kommunikationsassistenten werden auch zukünftige Einsatzfelder der Sprachtechnologie im Alltag diskutiert. Es wird gezeigt, welche Probleme sich bei der robusten Verarbeitung von Spontansprache ergeben und wie diese durch die Kombination von Sprach-, Kontext- und Weltwissen gelöst werden können.

Schlüsselwörter: Sprachtechnologie, Künstliche Intelligenz, Computerlinguistik, Dialogsysteme, Übersetzungssysteme, Multimodale Benutzerschnittstellen

1. Die Rolle der Sprachtechnologie in der Wissensgesellschaft

Durch die Fortschritte auf dem Gebiet der Sprachtechnologie ist die Vision des Computers als Dialogpartner, der menschliche Alltagssprache versteht und selbst auch spricht, in greifbare Nähe gerückt. Die Sprachtechnologie zählt zu den Schlüsseltechnologien bei der Verwirklichung der Wissensgesellschaft, da sich bislang nur die menschliche Sprache zur Formulierung, Speicherung und Weitergabe komplexer Sachverhalte, Gedanken und Wissensinhalte eignet. Der weltweite Zugriff auf das gesamte digital gespeicherte Wissen für Jedermann, zu jeder Zeit und an jedem Ort würde daher im Zeitalter des Internet ohne Einsatz von Sprachtechnologie eine Fiktion bleiben. Nur wenn es prinzipiell für jeden Menschen möglich wird, in seiner Muttersprache spontan eine Anfrage oder ein Kommando in Computersysteme zu sprechen, und wenn die entsprechende Antwort oder Reaktion wiederum für ihn verständlich in Alltagssprache ertönt, wird die Mensch-Computer-Interaktion den Stand erreicht haben, der den Computer zum integralen Bestandteil einer universalen Kulturtechnik für die Wissensgesellschaft macht.

Die starke Zunahme an informationstechnischen Anwendungen in allen Lebensbereichen verursacht einen hohen Bedarf an effektiveren, effizienteren und natürlicheren Schnittstellen, um den Zugriff auf Information und Anwendungen zu erleichtern. Dieser Bedarf wird weiter gesteigert durch die rasch zunehmende Komplexität der Anwendersysteme, durch die immer geringere Zeit, welche die Benutzer zum Ausführen von Aufgaben und das Erlernen von Bedienkonzepten haben und die Notwendigkeit, die Kosten für die sehr aufwendige Entwicklung produktspezifischer Benutzerschnittstellen zu reduzieren. Da elektronische Interaktion ein integraler Bestandteil des täglichen Lebens, der Arbeit und der Erziehung sein wird, könnten rasch Nachteile für diejenigen Menschen entstehen, die nicht in der Lage sind, solche Interaktionen auszuführen. Um diesen Personenkreis von der Wissens-

gesellschaft nicht auszuschließen, müssen mithilfe der Sprachtechnologie Dialogschnittstellen geschaffen werden, die möglichst jedermann völlig intuitiv bedienen kann.

2. Die drei Stufen der Sprachverarbeitung

Die maschinelle Sprachverarbeitung (vgl. [1],[3],[4]) ist wissenschaftlich eines der ehrgeizigsten Ziele unseres Zeitalters. Sie setzt die enge multidisziplinäre Zusammenarbeit von Informatikern, Linguisten, Sprachpsychologen, Nachrichtentechnikern, Kommunikationswissenschaftlern sowie Spezialisten der Computerlinguistik und der Künstlichen Intelligenz voraus. Für die Konstruktion von natürlichsprachlichen Dialogsystemen ist es notwendig, vom akustischen Signal durch Spracherkennung zunächst zu einer symbolischen Repräsentation der eingegebenen Äußerung zu kommen (vgl. Fig. 1). Darauf setzt dann der Prozeß der Sprachanalyse sowie des Sprachverstehens auf und nach der Interpretation des Dialogbeitrags erfolgt die Sprachgenerierung für die Rückäußerung des Systems. Schließlich wird die symbolische Form der geplanten Systemausgabe durch die Sprachsynthese wieder in ein akustisches Sprachsignal verwandelt. Der „hörende und sprechende Computer“ setzt also eine komplexe Signal-Symbol-Signal Transformation voraus (vgl. Fig. 1).

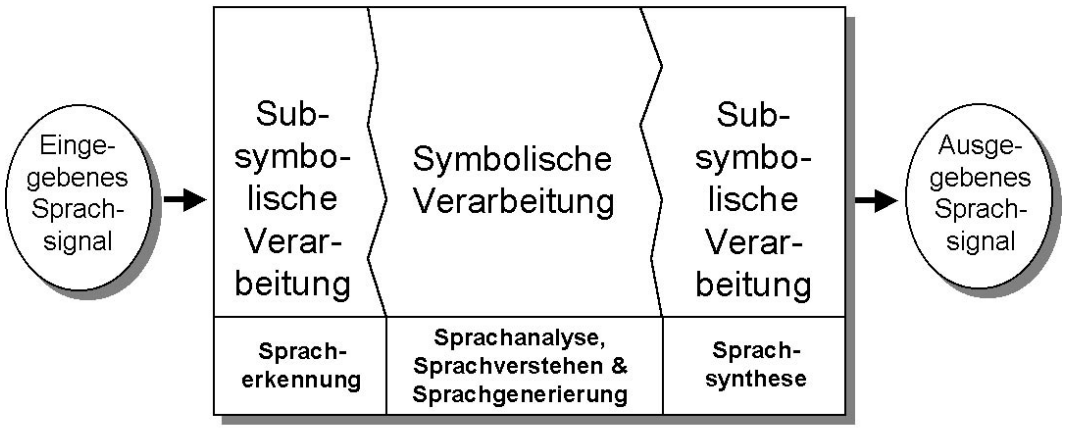

Fig. 1: Signal-Symbol-Signal Transformationen
in natürlichsprachlichen Dialogsystemen

Bei der Sprachverarbeitung für eine gesprochene Eingabe können drei Stufen unterschieden werden (vgl. Fig. 2): die Spracherkennung, die Sprachanalyse und das Sprachverstehen. Für einfache Anwendungen wie die Sprachsteuerung von Geräten im Alltag (z.B. Abstellen eines Weckers, Telephonbedienung, Auswahl eines Radiokanals im Auto) reicht es aus, wenn das System richtig erkennt, was der Sprecher gesagt hat. Eine weitergehende Sprachanalyse über die Bedeutung der Wörter im grammatischen Zusammenhang erübrigt sich, da der Interpretationsspielraum durch den vorgegebenen Anwendungszusammenhang so klein ist, daß schon auf der Wortebene klar ist, was gemeint ist. Einzelne Wörter, Ziffern oder Wortfolgen können heute sprecherunabhängig effizient und zuverlässig erkannt werden. Da solche Spracherkenner für ein kleines Vokabular (20 – 100 Wörter) auf einem Chip integrierbar sind, werden sie als eingebettete Systeme immer mehr in Gegenstände des täglichen Gebrauchs integriert werden. Alternativ zum Tastendruck wird man also der

Kaffeemaschine vom Sofa aus zurufen können „Einen Capuccino bitte“. Im Auto kann man heute schon mithilfe der Linguatronic von DaimlerChrysler seine Hände am Steuer lassen, wenn man seinen Chef sprechen will, weil das Spracherkennungssystem trotz der Nebengeräusche seinen Namen im Wahlkommando erkennt.

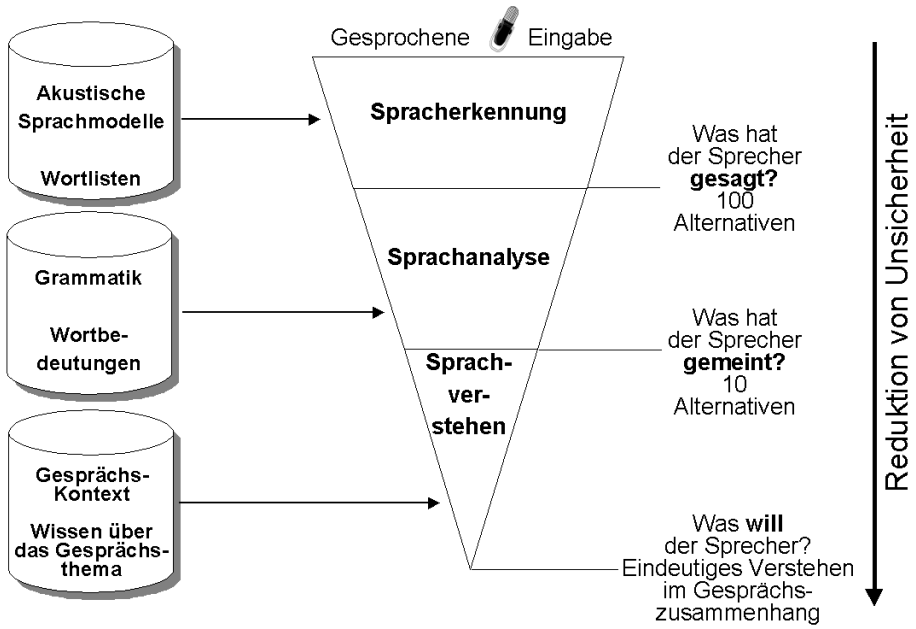

Fig. 2: Drei Stufen der Sprachverarbeitung

Die meisten Aufforderungen, Fragen und Wünsche lassen sich aber nicht mit wenigen Schlüsselwörtern formulieren. Wenn die Anwendung anspruchsvollere Eingabemöglichkeiten erfordert, muß durch Sprachanalyse versucht werden, den Sinnzusammenhang zu erfassen und in einer semantischen Repräsentation festzulegen, was der Sprecher mit seiner Äußerung gemeint hat. Durch die extrem hohe Variabilität des Sprachsignals bei unterschiedlichen Sprechern, aber auch bei demselben Sprecher in verschiedenen Sprechsituationen und Äußerungskontexten ist es den heutigen, ausschließlich auf statistischen Verfahren beruhenden Spracherkennern nicht möglich, fließend gesprochene Sprache in eine eindeutige Wortfolge zu überführen. Neben der Vielzahl der Aussprachevarianten für ein Wort müssen dadurch, daß Wortgrenzen im Sprachsignal nicht immer klar erkennbar sind und Laute verschliffen werden, selbst in den besten heutigen Spracherkennern Tausende von Worthypothesen alternativ überprüft werden.

Wie Fig. 2 zeigt, werden auf den drei Ebenen der Sprachverarbeitung schrittweise immer mehr Wissensquellen in die Verarbeitung eingeführt, so daß die Unsicherheit darüber, was der Sprecher mit seiner Äußerung letztlich will, auf jeder Ebene weiter reduziert wird. Wegen der für natürliche Sprachen charakteristischen starken Mehrdeutigkeit im lexikalischen und syntaktischen Bereich kann oft nur über ein explizites Modell des Gesprächskontextes oder Wissen über das Gesprächsthema ein eindeutiges Verstehen erreicht werden. Oft erweist sich die zunächst verfolgte Satzhypothese, die aufgrund von Hidden-Markov-Modellen als wahrscheinlichste Wortfolge eingestuft wurde, auf späteren Ebenen der Sprachverarbeitung durch das Hinzuziehen von syntaktischen und semantischen Modellen als falsch, so daß ein alternativer Pfad durch den vom Spracherkenner erzeugten Worthypothesengraphen als Interpretation gewählt werden muß.

3. Technische Herausforderungen für die Sprachtechnologie

Es wäre ein Trugschluß zu glauben, mit der Verfügbarkeit einer Vielzahl von Diktiersystemen als preiswerte PC-Massenprodukte (u.a. FreeSpeech von Philips, ViaVoice von IBM, SpeechBase von Siemens, Naturally Speaking von Dragon, VoiceXpress von Lernout & Hauspie) sei das Spracherkennungsproblem gelöst. Diese Systeme liegen nach einem sprecherabhängigen Training bei Wortschätzen von 50 000 – 120000 Wörtern und bis zu 140 gesprochenen Wörtern pro Minute noch bei Worterkennungsraten um die 90%. Da trotz individuellen Trainings solcher Diktiersysteme selbst bei Wortfehlerraten von nur 5%-10% kaum eine längere Äußerung fehlerfrei in Text transformiert wird, ergibt sich ein sehr begrenzter Einsatzbereich für diese Systeme, die sich auf die erste Sprachverarbeitungsebene der Spracherkennung beschränken und damit die inhaltliche Ebene unberücksichtigt lassen. Vom amerikanischen Markt für Diktiersoftware wird bereits berichtet, daß schon ein Jahr nach der Anschaffung nur rund 10% der Käufer von Diktiersystemen diese überhaupt noch einsetzen, da sich die Kauferwartungen nicht erfüllt haben. Diktiersysteme „raten“ zu oft falsch und kommen häufig zu Worthypothesen, die für den Sprecher im Sinnzusammenhang völlig absurd wirken. Es zeigt sich, daß diese Systeme nur bei professioniellen Diktierern in begrenztem fachsprachlichen Kontext ein hohes Ratiopotential haben, aber beim gelegentlichen Diktieren von stark variierenden Textsorten gerade im privaten Bereich zu schnell an ihre Grenzen stoßen. Es bedarf der Kombination von Sprachtechnologie und Wissensverarbeitung, um in Zukunft universell einsetzbare Diktiersysteme zu entwickeln.

Als Komponenten für Dialogsysteme sind die heutigen Diktiersysteme auch deshalb nicht geeignet, weil sie spezielle Eingabebedingungen wie das Verwenden eines Nahbesprechungsmikrophons und einer expliziten Aktivierungstaste für die Spracheingabe voraussetzen. Das Haupteinsatzgebiet von Dialogsystemen liegt aber im Bereich der Telephonie, wo die Signalqualität erheblich schlechter ist und vom Benutzer keine technische Segmentierung seiner Äußerung erwartet werden kann. Während man für Diktiersysteme eine Trainingsphase akzeptieren mag, ist dies bei telephonischen Auskunfssystemen ausgeschlossen. Eine noch höhere Stufe der Komplexität für die Spracherkennung ergibt sich bei der Verwendung von Mobiltelephonen auf GSM-Basis, wobei besonders beim Wechsel zwischen Funkzellen die Signalqualität selbst für den Menschen kaum noch eine sichere Erkennung des Gesagten ermöglicht. Wie Fig. 3 zeigt, wird heute in der Forschung von einem „offenen Mikrophon“ ausgegangen. Dies bedeutet, daß das System selbst den Beginn und das Ende von Äußerungen erkennen muß und auch die Segmentierung längerer Äußerungen in Sätze und Phrasen durch sprachtechnologische Komponenten vornehmen muß.

Telephonische Auskunftssysteme müssen Spontansprache verarbeiten können, die im Gegensatz zur fließend vorgelesenen Sprache eine Vielzahl von Häsitations- und Selbstkorrekturphänomenen enthält. Dies muß sprecherunabhängig geschehen, da man natürlich nicht erwarten kann, daß jeder Benutzer zunächst eine initiale Trainingsphase zur Ermittlung seiner Stimm- und Aussprachecharakteristika durchlaufen kann. Fig. 3 zeigt, daß man derzeit an sog. sprecheradaptiven Systemen arbeitet, die zunächst in einem sprecherunabhängigen Modus starten und im Verlaufe des Dialogs – für den Benutzer unmerklich – Adaptionen mit dem Ziel vornehmen, die Verstehensleistung und Verarbeitungsgeschwindigkeit zu erhöhen.

Fig. 3: Herausforderungen für die Sprachtechnologie

Während bei Diktier- und Kommandodialogen das System selbst keine sprachlichen Äußerungen erzeugt, muß für Auskunftsdialogsysteme auch automatische Sprachgenererierung und Synthese betrieben werden. Eine noch höhere Komplexität für die Dialogsteuerung ergibt sich in Verhandlungsrunden, wie sie z.B. im Umfeld des elektronischen Handels zu erwarten sind, wobei System und Benutzer z.B. beim Aushandeln eines Preises unterschiedliche Dialogziele verfolgen und ein Dialog mit gemischter Initiative zwischen System und Benutzer entsteht.

In dem BMBF-Verbundprojekt Verbmobil (Laufzeit Phase 1: 1993-1996; Phase 2: 1997 - 2000) zur maschinellen Übersetzung spontansprachlicher Dialogsprache wird versucht, für alle vier in Fig. 3 verwendeten Dimensionen der Sprachtechnologie die höchste Komplexitätsstufe zu realisieren(vgl. [2],[5]). Das Verbmobil-System erkennt gesprochene Spontansprache, analysiert die Eingabe, übersetzt sie in eine Fremdsprache, erzeugt einen Satz und spricht ihn aus. Für ausgewählte Themenbereiche (z.B. Terminverhandlung, Reiseplanung, PC-Fernwartung) soll Verbmobil Übersetzungshilfe in Gesprächssituationen mit ausländischen Partnern leisten.

4. Auskunftssysteme und die Automatisierung von Call Centern

Telephonbasierte Auskunftssysteme sind heute bereits im praktischen Einsatz. Eines der ersten verfügbaren Systeme für das Deutsche war TABA von Philips, das unter der Telephonnummer 0241-604020 Auskünfte zu Zugverbindungen gibt. Folgender Dialog mit dem System verdeutlicht die Leistungsfähigkeit dieses sprecherunabhängigen Auskunftssystems:

(1) System: „Von wo nach wo möchen Sie fahren?“
(2) Benutzer: „Ich möchte von Saarbrücken nach Paderborn fahren.“
(3) System: „Wann möchten Sie von Saarbrücken Hauptbahnhof nach Paderborn Hauptbahnhof fahren?“
(4) Benutzer: „Übermorgen.“

(5) System: „Um welche Uhrzeit möchten Sie übermorgen fahren?“
(6) Benutzer: „Ich möchte mittags fahren.“
(7) System: „Sie möchten also zwischen 11.00 und 13.00 fahren?“
(8) Benutzer: „Ja, das ist richtig.“
(9) System: „Es bestehen folgende Verbindungen: Mit Interregio Abfahrt in Saarbrücken um 11.47....“

Der Dialog zeigt, daß das TABA in seinen Antworten jeweils klar macht, wie es die vorangegangenen Benutzereingaben für die Zugauskunft interpretiert. So wird z.B. die Eingabe in (6) „mittags“ vom System in (7) als „zwischen 11.00 und 13.00“ interpretiert. Auch Standardannahmen, z.B. daß jeweils der Hauptbahnhof gemeint ist, wenn für eine Stadt keine spezielle Bahnhofsangabe gemacht wird (vgl. (2)), werden vom System expliziert (vgl. (3)), so daß der Benutzer noch widersprechen kann, wenn er z.B. von einem Stadtteilbahnhof abfahren möchte und nur vergaß, dies in seiner Anfrage direkt mitzuteilen. Solche Klärungsdialoge sind wesentlich für das Gelingen komplexer Mensch-Maschine-Dialoge. Die nächste Generation spontansprachlicher Beratungssysteme wird komplexe dialogische Interaktionen ermöglichen, wobei sowohl der Benutzer als auch das System Interaktionen initiieren, Rück- und Klärungsfragen stellen, Verstehensprobleme signalisieren oder den Dialogpartner unterbrechen können.

Das von DaimlerChrysler entwickelte OSCAR-System gibt wie TABA ebenfalls Bahnauskünfte und ist telephonisch unter 0180 5 99 66 22 erreichbar. Die genauen Abflugzeiten und Ankunftszeiten der Lufthansa kann man von dem Sprachdialogsystem ALF unter 0180 3 00 00 74 erfahren. Während bislang Wartezeiten von bis zu 7 Minuten beim Anruf sog. Call Center zu verzeichnen sind, können durch den Einsatz der Sprachtechnologie gerade für einfache Auskünfte Wartezeiten vermieden und eine konstante Service-Qualität erreicht werden.

Allerdings ist der Einsatz vollautomatischer Dialogsysteme heute noch durch mehrere Faktoren eingeschränkt: die Information, die der Kunde sucht, muß in einer Datenbank verfügbar sein; das Dialogthema muß mit einem Vokabular von 1000 - 10000 Wörtern abzudecken sein und der Dialogverlauf muß einer gewissen Systematik folgen. Zudem ist die Entwicklung spezieller automatischer Dialogsysteme für Call Center noch recht zeit- und kostenintensiv, so daß sich der Einsatz nur für Themenfelder mit sehr hohem Transaktionsvolumen lohnt.

Daher kommt in vielen Einsatzfällen nur eine Kombination von Sprachtechnologie und Dialog mit einem menschlichen Call Center-Agenten in Frage. Dabei kann durch ein Dialogsystem zu Beginn des Gesprächs Routineinformation abgefragt und eine Klassifikation des Kundenwunsches durchgeführt werden, die dann zur Weitervermittlung an einen menschlichen Agenten für das vertiefende Gespräch führt. Damit wird der Durchsatz und die Effizienz der menschlichen Agenten erhöht und durch eine gezielte Auswahl eines problemkundigen Beraters die Service-Qualität verbessert.

5. Dialogübersetzung durch Verbmobil

Die Berücksichtigung des Kontextes beim Sprachverstehen ist eine der wesentlichen Voraussetzungen für anpruchsvolle Anwendungen der Sprachtechnologie. Beson-

ders deutlich wird dies beim Dolmetschen von Telephongesprächen, wie dies in einem Anwendungsszenario von Verbmobil für die Domäne der Reiseplanung vorgesehen ist. Das Wort „nächste“ muß bei der Übersetzung ins Englische einmal mit „next“ und in anderem Satzkontext mit „nearest“ wiedergegeben werden, je nachdem, ob ein zeitlicher oder räumlicher Bezug besteht (vgl. Fig. 4). 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.

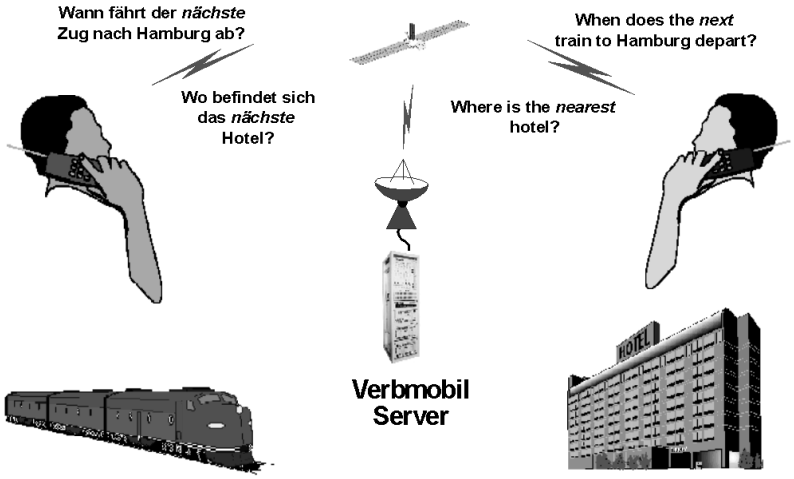

Fig. 4: Dialogübersetzung durch das System Verbmobil

5.1. Robuste Verarbeitung von Spontansprache

Verbmobil war weltweit das erste System, das auch Spontansprache als frei formulierte Alltagssprache verarbeiten kann. Dabei werden Gedankengänge fortlaufend in Sprache umgesetzt, wobei sehr häufig auch ungrammatische Sätze entstehen. Verbmobil muß deshalb mit abgebrochenen Sätzen, Einschüben und Selbstkorrekturen umgehen können. Nicht bedeutungstragende Äußerungselemente wie Räuspern, Schmatzen, "äh" und "ehm" werden von der Spracherkennung zunächst wie spezielle Wörter behandelt und für die weitere Analyse aus der Eingabe entfernt. Wenn der Sprecher sagt "Ja, ich weil also würde mal sagen äh vorschlagen, wir könnten uns am äh 7. treffen so im Mai", so würde dieser Satz von einem an der Schriftsprache orientierten System abgelehnt und der Sprecher müßte den Satz wiederholen. Durch Kombination von statistischen und linguistischen Verfahren wird Verbmobil jedoch so fehlertolerant und robust, daß der Dialogakt "suggest_date" mit der Datumsangabe "7. Mai" aus der oben zitierten Äußerung extrahiert und die Übersetzung "How about the seventh of May?" ausgegeben wird.

Während die heutigen Diktiersysteme die Eingabe einfach transkribieren und damit auch fehlerhafte Äußerungen unverändert übernehmen, erkennt Verbmobil Repa-

raturen und extrahiert die intendierte Bedeutung, nachdem der Worthypothesen-Graph entsprechend transformiert wurde (vgl. Fig. 5).

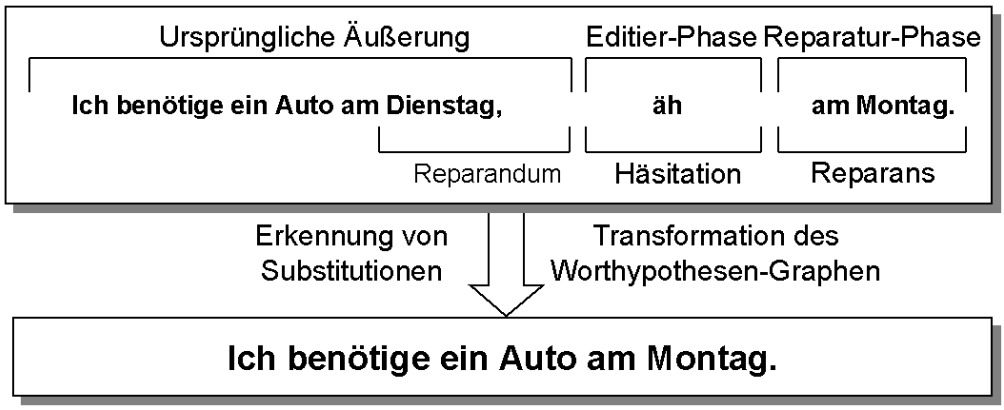

Fig. 5: Das Verstehen von Selbstkorrektur im Verbmobil-System

Oft werden in schneller Rede kurze Funktionswörter wie „in“ oder „an“ vom Sprecher „verschluckt“ oder so undeutlich gesprochen, daß der Spracherkenner sie nicht als Worthypothese ausgibt. Um trotz der dadurch entstehenden fragmentarischen Eingabe die Äußerung inhaltlich verstehen zu können, wurde für Verbmobil eine robuste Dialogsemantik entwickelt, die partielle Analysen mithilfe heuristischer Regeln ergänzt. Fig. 6 zeigt zwei Beispiele, in denen jeweils das Funktionswort „in“ nicht erkannt wurde. Verbmobil erkennt, daß zwischen den ersten beiden Fragmenten eine temporale Relation besteht und zwischen den folgenden zwei Fragmenten eine räumliche Beziehung vorliegen muß.

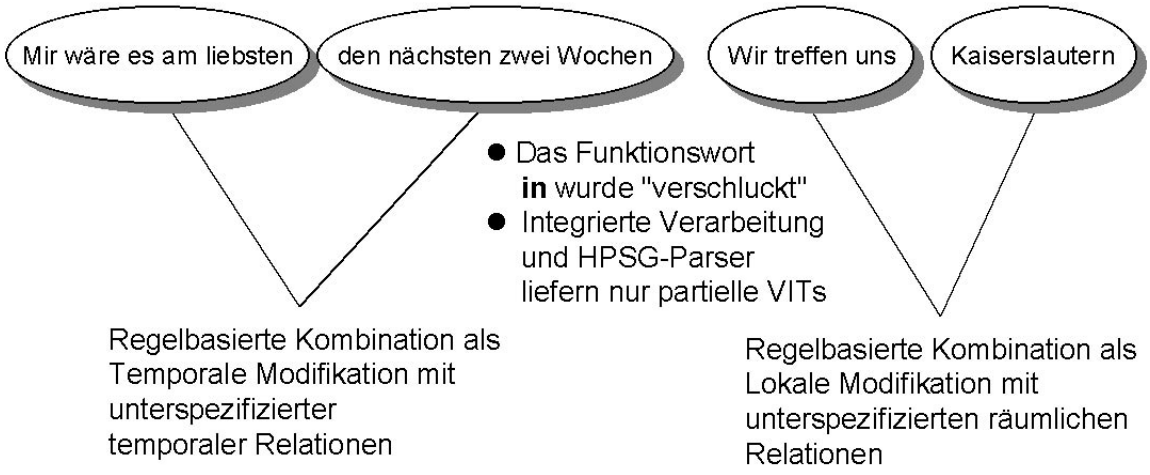

Fig.6: Ein Beispiel für die robuste Dialogsemantik

Eine zusätzliches Problem stellen dialektale Färbungen dar. So ist bei vielen Sprechern aus dem Saarland und der Pfalz in der Äußerung "Ich finde das nätt" rein akustisch "nett" kaum von "nicht" zu unterscheiden. Auch ein menschlicher Dialogpartner kann in diesem Fall nur durch Einbeziehung des Kontextes und der Betonung ermitteln, ob der Satz als Zustimmung oder als Ausdruck einer ergebnislosen Suche gemeint ist.

5.2. Die kombinierte Verarbeitung von Sprach-, Kontext- und Weltwissen

Gesprochene Sprache kennt keine Interpunktion; Betonung und Phrasierung ersetzen Punkt und Komma. Die Wortfolge "Ja-zur-Not-geht-es-auch-am-Samstag" kann je nach Betonung als Bestätigung des Termins "Samstag" interpretiert werden ("Ja, zur Not geht es auch am Samstag.") oder als eingeschränkte Annahme eines Termins mit Gegenvorschlag: "Ja, zur Not! Geht es auch am Samstag?" Nur durch die Berücksichtigung der Prosodie können Mehrdeutigkeiten auch von einzelnen Wörtern wie "noch" für die Übersetzung aufgelöst werden. Lautet die Eingabe "Wir brauchen noch einen Termin" ohne prosodischen Akzent auf "noch", so übersetzt Verbmobil mit "We still need a date". Wird "noch" jedoch betont, so wählt Verbmobil aufgrund der anderen Satzbedeutung die Übersetzung "We need another appointment". Ohne Weltwissen über den Gesprächsgegenstand ist eine Übersetzung oft nicht möglich. Die Transferregeln von Verbmobil müssen daher in Sortentests auf Wissen zurückgreifen, das in einer terminologischen Logik codiert ist, um z.B. das Wort "vor" in dem Satz "Wir treffen uns vor dem Hotel" durch "in front of" zu übersetzen, aber bei der Eingabe "Wir treffen uns vor der Tagung" die Übersetzung "before" zu wählen. Die Übersetzung muß auch kontextabhängig erfolgen und den Dialogverlauf berücksichtigen. So muß der Satz "Geht es bei Ihnen?" von Verbmobil als "Do we meet at your place?" übersetzt werden, wenn vorher gefragt wurde "Wo treffen wir uns?". Dagegen lautet die korrekte Übersetzung der identischen Eingabe "Is it possible for you?", wenn vorher "Sollen wir uns im April treffen?" geäußert wurde.

Nur durch Weltwissen in der Gesprächsdomäne können subtile Bedeutungsunterscheidungen, die in der Quellsprache nicht vorhanden in der Zielsprache richtig generiert werden. Selbst bei Sprachpaaren mit großer Verwandtschaft wie Deutsch-Englisch treten solche Probleme oft auf. So muß die im Deutschen zeitneutrale Formulierung „sich zum Essen treffen“ im Englischen abhängig von der aus dem Kontext hervorgehenden Zeit mit „meet for dinner“ oder „meet for lunch“ übersetzt werden. Verbmobil übersetzt daher:

Benutzer: „Wir könnten uns um zwanzig Uhr zum Essen treffen.“
Verbmobil: „We could meet for dinner at eight o’clock.“

In diesem Fall würde die falsche Übersetzung „We could meet for lunch at eight o’clock“ beim Zuhörer Unverständnis auslösen, weil ein „lunch“ normalerweise um die Mittagszeit eingenommen wird. Das Beispiel zeigt deutlich, daß nur durch die Kombination von Sprachwissen, Weltwissen und Kontextwissen eine hohe Qualität der automatischen Sprachverarbeitung erreicht werden kann.

5.3. Die Übersetzung und Protokollierung von Telephongesprächen

Technisch ist es mit Verbmobil erstmals gelungen, durch Sprachkommandos eine ISDN-Dreierkonferenz für eine anschliessende Dialogübersetzung aufzubauen. Dabei wählt z.B. ein deutscher Teilnehmer zunächst den Verbmobil-Sprachserver an und gibt das Kommando „Neuen Teilnehmer hinzunehmen“. Danach wird er von Verbmobil nach einer Telephonnummer gefragt, die das System als dritten Gesprächsteilnehmer der Dreierkonferenz anwählt (vgl. Fig. 7). Dieser Teilnehmer wird dann von Verbmobil auf Englisch begrüßt und das Gespräch kann beginnen. Dabei

fungiert Verbmobil als Dolmetschhilfe, nachdem es vorher die Rolle eines Telephon-Operators für Auslandsgespräche übernommen hatte.

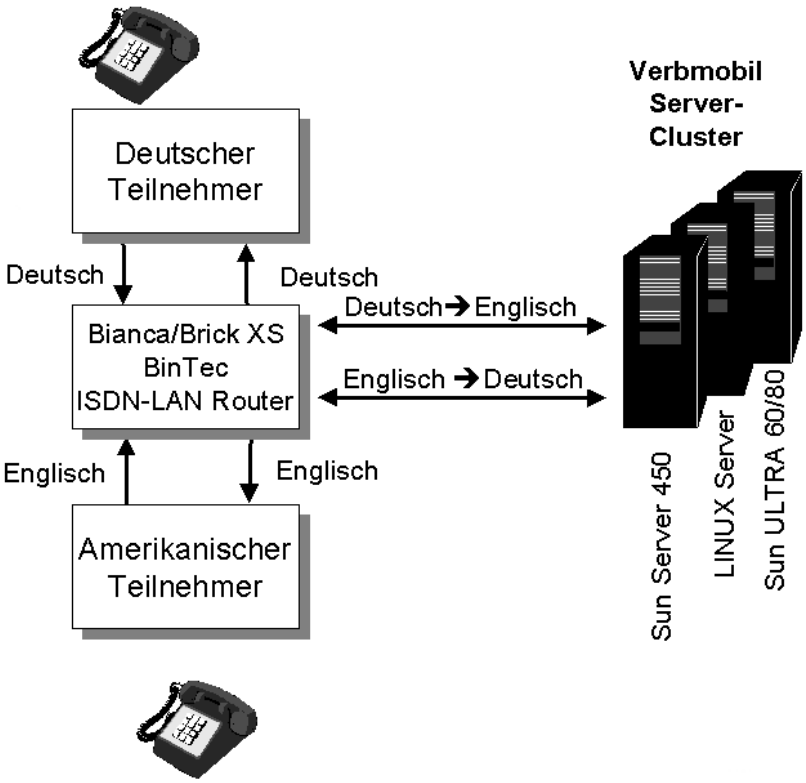

Fig. 7: Eine per Sprachdialog aufgebaute Dreierkonferenz mit dem Verbmobil-Server

Eine im Alltag sehr nützliche Innovation stellt die in Verbmobil erstmals realisierte Dialogprotokollierung dar. Das System ist in der Lage, nach Beendigung eines Telephondialogs entweder ein knappes Ergebnisprotokoll oder ein ausführlicheres Verlaufsprotokoll zu erzeugen. Mit der Sprachtechnologie wird es also möglich werden, eine Zusammenfassung von Gesprächsergebnissen automatisch schriftlich erstellen zu lassen (vgl. Fig. 8). Dies bietet für Geschäftsverhandlungen, juristische Diskussio-

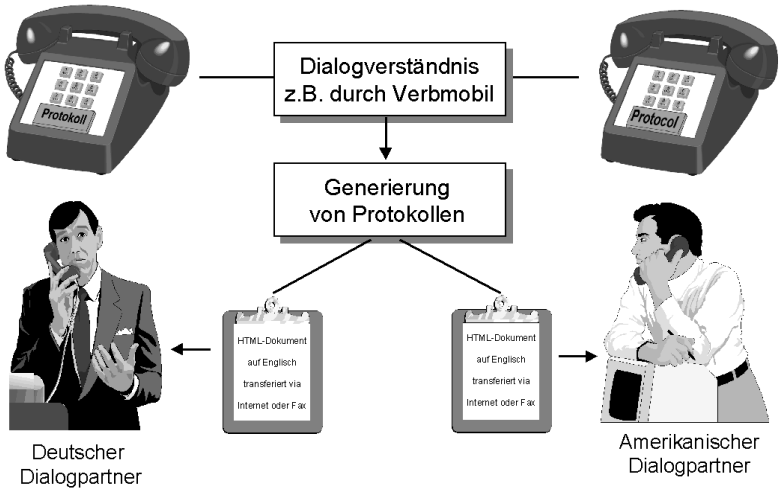

Fig. 8: Automatische Gesprächsprotokollierung als Mehrwertdienst

nen und alle Arten von Abstimmungsgesprächen große Vorteile, weil nach Gesprächsende über Email, Fax oder WWW ein schriftliches Protokoll angefordert werden kann. Obwohl diese Funktionalität von Verbmobil derzeit nur für Terminabsprachen und Reiseplanungen als telephonischer Mehrwertdienst bereitgestellt werden kann (vgl. Fig. 9), ist davon auszugehen, daß in der nächsten Dekade die schriftliche Zusammenfassung von Gesprächsergebnissen auch für andere Themengebiete möglich werden wird. Bei der Protokollgenerierung wird die Verstehensleistung besonders deutlich: Nur wenn das System den Dialogverlauf nachvollziehen kann und alle Äußerungen im Wesentlichen verstanden hat, kann eine vernünftige Zusammenfassung eines Dialoges generiert werden. Verbmobil kann mithilfe seiner Textgeneratoren für Deutsch und Englisch solche Zusammenfassungen dann in zwei Versionen erzeugen, so daß jeder Gesprächspartner zum Schluß einer Verhandlung sich ein Gesprächsprotokoll in seiner Muttersprache ausdrucken lassen kann .

VERBMOBIL ERGEBNISPROTOKOLL Nr. 1

Teilnehmer: Herr Schneider, Herr Thompson
Datum: 7.12.1999
Uhrzeit: 8:02 Uhr bis 8:03 Uhr
Thema: Reise mit Unterkunft und Freizeitgestaltung

GESPRÄCHSERGEBNISSE

Terminabsprache:
Das Arbeitstreffen findet in Hannover statt. Es findet am 20. Januar 2000 um 11 Uhr am Vormittag statt. Herr Schneider und Herr Thompson treffen sich am Bahnhof in München. Dieses Treffen findet am 19. Januar 2000 um halb 10 statt.

Reiseplanung:
Die Hinreise mit der Bahn nach Hannover von München beginnt am 19. Januar um 4 Uhr am Nachmittag.

Unterkunft:
Ein gutes Hotel wurde vereinbart. Die Einzelzimmer kosten 80 Euro pro Nacht. Herr Thompson kümmert sich um die Reservierung.

Freizeit:
Herr Schneider und Herr Thompson vereinbarten am 19. am Abend essenzugehen.

Protokollgenerierung automatisch am 7.12.1999 10:26:27 h

Fig. 9: Beispiel für ein durch Verbmobil erzeugtes Dialogprotokoll

5.4. Korpus-basierte Lernverfahren als Grundlage statistischer Sprachmodelle

Eine fundamentale Einsicht der Sprachtechnologie war es in den letzten Jahren, daß nur mit maschinellen Lernverfahren über möglichst großen Datensammlungen und Sprachkorpora die statistischen Spracheigenschaften in Systemen ausreichend genau modelliert werden können, um die für anspruchsvolle Anwendungen notwendige Verarbeitungsqualität zu erreichen. Oft ist die effiziente Steuerung von symbolischen Sprachverarbeitungsverfahren nur durch probabilistische Verfahren möglich. Außerdem können große Regelmengen meist nur mit statistischen Verfahren aus Korpora gelernt werden. Probabilistische Verfahren spielen daher auf allen Ebenen der Sprachverarbeitung heute ein wichtige Rolle. Fig. 10 zeigt das Spektrum aufbe-

reiteter Sprachdaten und deren Nutzung zum Lernen verschiedenartiger statistischer Sprachmodelle, wie sie im Verbmobil-System genutzt werden. Die Kombination statistischer und symbolischer Verfahren in hybriden Systemen wird in Zukunft zu einem großen Anwendungsspektrum der Sprachtechnologie im Alltag führen.

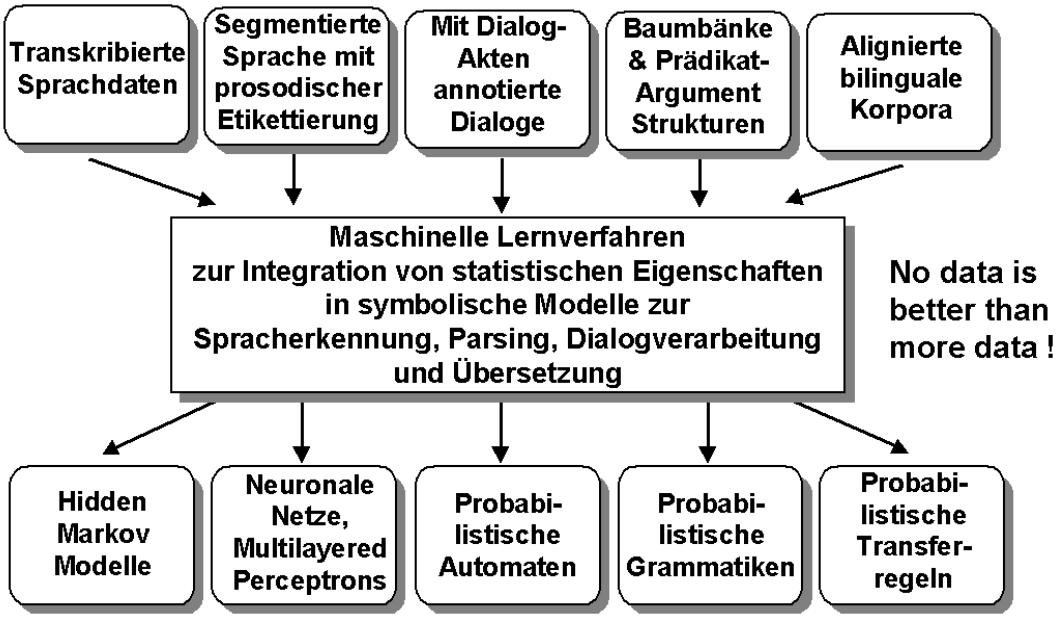

Fig. 10: Korpus-basierte Lernverfahren zur Erzeugung statistischer Sprachmodelle

6. Die zukünftige Weiterentwicklung der Sprachtechnologie

Da in Zukunft immer mehr eingebettete Computersysteme die Steuerung und das Zusammenwirken von Geräten des alltäglichen Gebrauchs übernehmen werden, wird der Endbenutzer sich häufig mit der Aufgabe konfrontiert sehen, daß er diese Systeme konfigurieren und programmieren muß. Dies stellt für viele potentielle Benutzer schon heute eine unüberwindbare Barriere dar (z.B. Programmierung von privaten ISDN-Anlagen, Videorekordern, Modemanschlüssen). Hier bietet es sich an, die Sprachtechnologie zur „natürlichsprachlichen Programmierung“ zu nutzen. Hierzu muß die eingegebene Alltagssprache zunächst wie beim automatischen Dolmetschen verstanden werden, um dann aber nicht in eine Fremdsprache sondern in eine Programmiersprache übersetzt zu werden. Damit wird es möglich, eine intelligente Haussteuerung alltagssprachlich zu konfigurieren und bei Bedarf individuell umzuprogrammieren (vgl. Fig. 11).

Derzeit wird in der Forschung an der automatischen Indexierung und Klassifikation von Fernsehsendungen gearbeitet, wobei dort die gesprochene Sprache mit eventuell vorhandenen Untertiteln und einer groben Segmentierung des Bildmaterials abgeglichen und zu Zugriffsindices verdichtet wird. Damit wird mittelfristig die Vision einer inhaltsbasierten Suche über alle Fernsehkanäle hinweg realisierbar (vgl. Fig. 12).

Fig. 11: Natürlichsprachliche Programmierung eingebetteter Computersysteme

Durch die Online-Suche in aufgezeichneten Kommentaren, Interviews und Diskussionen könnte auch der gezielte sprachbasierte Zugriff auf global verfügbaren, digitalen Multimedia-Archiven ermöglicht werden und durch die Kombination von Fernsehgerät und Internet-Zugang in WebTVs im Alltag für jedermann nutzbar gemacht werden.

Fig. 12: Sprachbasierter Informationsabruf aus laufenden Fernsehsendungen

Eine Leitvorstellung in der aktuellen Forschung zu intuitiven Benutzerschnittstellen ist es, die Vorteile sprachlich dialogischer Kommunikation mit den Vorteilen graphischer Bedienoberflächen und taktiler Interaktionsformen zu einem höherwertigen, oft multimodal genannten Bedienparadigma zu verschmelzen (vgl. [4],[6]).

So streben wir in dem vom BMBF geförderte Leitvorhaben SmartKom (1999 – 2003) die nahtlose Integration und koordinierte semantische Verarbeitung sich wechselseitig ergänzender Eingabemodalitäten wie Sprache, Gestik, Stifteingabe, Graphik und Biometrie an.

Durch Benutzermodellierung, Planerkennung und Lernverfahren soll sich SmartKom an den indivuellen Nutzer adaptieren und zum persönlichen Assistenten werden. Durch Auswertung des Bedien- und Aufgabenkontextes kann SmartKom auch fehlerhafte, unvollständige und gestörte Eingaben noch sinnvoll interpretieren. Der Kom-

munikationsagent SmartKom wird personalisiert, indem er als animierter Life-like Character visualisiert wird. Der mobile Kommunikationsassistent SmartKom-Mobil ist als eine Ausprägung der SmartKom-Architektur ein persönlicher, ständiger Begleiter eines Benutzers im Büro, der Wohnung, im Auto und zu Fuß.

In Fig. 13 ist eine Designstudie zu SmartKom-Mobil dargestellt. Die eigentliche Verarbeitung kann auf einen tragbaren Rechner ausgelagert werden, der wie in Fig. 12 (am Gürtelklipp) angedeutet, beispielsweise über eine Funkverbindung kommunizieren kann.

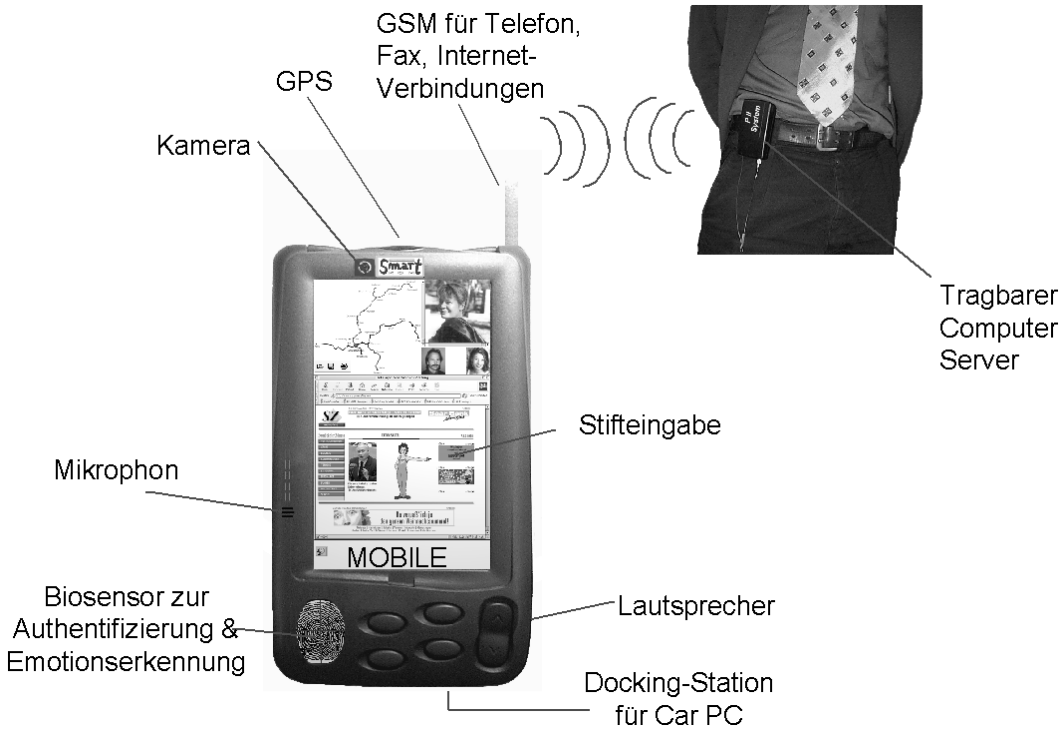

Fig. 13: Designstudie zu SmartKom-Mobil

SmartKom-Mobil dient als mobile Plattform für zahlreiche Informationsdienste, wie etwa als Zugang zum Internet, der über eine GSM-Verbindung hergestellt wird. Durch ein integriertes GPS kann die Eigenbewegung auf digitalen Karten verfolgt werden. Für die Ausgabe sind ein Display und Lautsprecher vorgesehen. Um einen Dialog in gesprochener Sprache führen zu können, verfügt SmartKom-Mobil über ein Mikrophon, das in der Designstudie am linken Rand des Gerätes zu sehen ist. Weiterhin ist eine Kamera integriert (siehe oberen Rand des Gerätes der Designstudie), mit der etwa die Mimik erfaßt werden kann. Mit einem Stift können Handschrift oder Zeigegesten koordiniert mit Sprache eingegeben werden.

7. Zusammenfassung und Ausblick

Neben der im letzten Abschnitt diskutierten Multimodalität für intuitive Benutzerschnittstellen wird besonders in Europa und Asien die Multilingualität eines der wichtigsten Ziele der weiteren Forschung zur Sprachtechnologie sein. Dabei wird das gesamte Spektrum multilingualer Sprachtechnologie von Dialogübersetzungssystemen wie Verbmobil, über multilinguales Indexieren von Multimedia-Archiven sowie den sprachgesteuerten Zugriff auf multilinguale Web-Seiten bis hin zu mobilen

Kommunikationsassistenten wie SmartKom so weiterentwickelt werden, daß diese Funktionen mittelfristig im Alltag als Massenprodukte verfügbar werden.

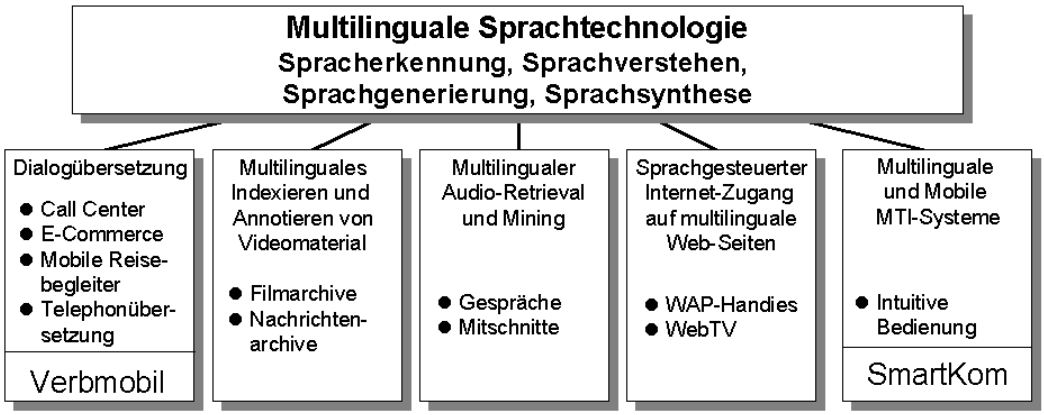

Fig. 14: Das Einsatzspektrum für multilinguale Sprachtechnologie

Die Sprachtechnologie wird eine wesentliche Säule der Informationstechnologie bilden, da

- die verbale Interaktion die einfachste, ausdrucksstärkste und beliebteste Kommunikationsform des Menschen bleiben wird.
- Computer standardmäßig mit Audioschnittstellen ausgerüstet sein werden.
- die Prozessorleistung und der Speicher selbst kleinster Computer die Realzeitverarbeitung von Sprache ermöglichen wird.
- sich die Sprachkommunikation über IP-Netze explosionsartig ausdehnen wird.
- besonders Sprachdialogsysteme die bestehenden Hemmschwellen von Computerlaien bei der Nutzung der Informationstechnologie abbauen können und so einen Beitrag leisten zur Nutzerfreundlichkeit und Nutzerzentrierung der Technik in der Wissensgesellschaft.

Literatur

[1] James F. Allen (1994): Natural Language Understanding. Benjamin Cummings, Second Edition.

[2] Thomas Bub, Wolfgang Wahlster, Alex Waibel, A. (1997): Verbmobil: The Combination of Deep and Shallow Processing for Spontaneous Speech Translation. In: Proceedings of the International Conference on Acoustics, Speech and Signal Processing, München, pp. 71-74.

[3] Ronald Cole , Joseph Mariani, Hans Uszkoreit, Annie Zaenen, Victor Zue (eds.) (1997): Survey of the State of the Art in Human Language Technology. Cambridge University Press (Studies in Natural Language Processing)

[4] Mark Maybury, Wolfgang Wahlster, W. (eds.) (1998): Readings in Intelligent User Interfaces. San Francisco: Morgan Kaufmann.

[5] Wolfgang Wahlster (1993): Verbmobil: Translation of Face-to-Face Dialogs. In: Proceedings of the Fourth Machine Translation Summit, Kobe, Japan, pp. 127-135.

[6] Wolfgang Wahlster, Elisabeth André, Wolfgang Finkler, Hans-Jürgen Profitlich, Thomas Rist (1993): Plan-Based Integration of Natural Language and Graphics Generation. Artificial Intelligence 63(1-2): pp. 387-427.

erscheint in:

Heinz-Nixdorf-Museums-Forum (ed.) (1999):
Alltag der Zukunft -Informationstechnik verändert unser Leben.
Paderborn: Schöningh