Verbmobil: Multilinguale Verarbeitung von Spontansprache

Reinhard Karger, Wolfgang Wahlster

Verbmobil ist ein langfristig angelegtes, interdisziplinäres Leitprojekt im Bereich der Sprachtechnologie. 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, Fernwartung) soll Verbmobil Übersetzungshilfe in Gesprächssituationen mit ausländischen Partnern leisten. Das Verbundvorhaben, in dem Unternehmen der Informationstechnologie, Universitäten und Forschungszentren kooperieren, wird vom Bundesministerium für Bildung, Wissenschaft, Forschung und Technologie (BMBF) in zwei Phasen (Laufzeit Phase 1: 1993-1996; Phase 2: 1997 - 2000) gefördert. Nachdem in der ersten Phase Terminverhandlungsdialoge zwischen einem deutschen und japanischen Geschäftspartner mit Englisch als Zwischensprache verarbeitet wurden, steht in der zweiten Phase von Verbmobil die robuste und bidirektionale Übersetzung spontansprachlicher Dialoge aus den Domänen Reiseplanung und Hotelreservierung für die Sprachpaare Deutsch-Englisch (ca. 10.000 Wörter) und Deutsch-Japanisch (ca. 2.500 Wörter) im Vordergrund.

Übersetzung gesprochener Spontansprache

Spontansprache ist frei formulierte Alltagssprache, bei der ein Sprecher nicht etwa vorbereitete Texte vorliest. Gedankengänge werden 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. 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. 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.

Im Bereich Sprachsynthese ist das Hauptziel der Forschungen in Verbmobil eine möglichst natürlich klingende Aussprache der übersetzten Dialogbeiträge. Um nicht „roboterhaft“ zu klingen, muß Verbmobil Verschleifungen richtig aussprechen, so daß „am Montag“ als „amontag“ synthetisiert wird. Vor allem aber muß Verbmobil die richtige, zum Inhalt des Redebeitrages passende Satzmelodie berechnen. Außerdem wird in Verbmobil u.a. mithilfe von neuronalen Netzen versucht, den Stimmcharakter des jeweiligen Sprechers auch bei der automatisch erzeugten Übersetzung nachzubilden, so daß nicht etwa die deutsche Eingabe einer Frauenstimme in der englischen Übersetzung als eine tiefe Männerstimme ertönt.

Die Entwicklungsstufen von Verbmobil

Das erste integrierte System, der sog. Verbmobil-Demonstrator, konnte 1995 während der CeBIT von Bundesforschungsminister Dr. Jürgen Rüttgers der Öffentlichkeit vorgestellt werden. Der Verbmobil-Demonstrator (Umfang 1292 Wörter) erkennt gesprochene deutsche Eingaben aus der Domäne Terminverhandlung, analysiert sie, übersetzt sie und äußert die englische Übersetzung. Der Verbmobil-Forschungsprototyp 1.0 (Umfang 2461 Wörter), der auf der CeBIT 1997 vorgestellt wurde, erkennt auch japanische Eingaben, um sie ins Englische zu übersetzen, und kann auch auf Deutsch Klärungsdialoge mit dem Benutzer führen. Für die erste Projektphase wurden bis Ende 1996 64,9 Millionen DM Fördermittel des BMBF eingeplant. Zusätzlich brachten die Industriepartner 31 Millionen DM auf.

Nach der erfolgreichen Abnahme des Verbmobil-Forschungsprototyps im Oktober 1996 bewilligte das BMBF für die zweite Phase (1997 - 2000) 50,2 Mio. DM; die Industriepartner stellen 20,4 Mio. DM an Eigenmitteln zur Verfügung. In der zweiten Phase wird Verbmobil auf einem zentralen Sprachserver implementiert (Umfang ca. 10000 Wörter für Deutsch-Englisch und 2500 Wörter für Deutsch-Japanisch), der über ISDN-Telephone, ATM-basierte Telekooperationsdienste oder GSM-Mobilfunk in Anspruch genommen werden kann. Dieser Sprachserver identifiziert die Eingabesprache und übernimmt die Spracherkennungs-, Übersetzungs- und Sprachgenerierungsleistung. Da mehrere Nutzer die Übersetzungsdienstleistung gleichzeitig in Anspruch nehmen können, werden bei dem Sprachserverkonzept parallele Kanäle vorgesehen. Verbmobil wird dadurch auch in mehrsprachigen Telekonferenzen mit mehr als zwei Partnern eingesetzt werden können (Multiparty-Situation).

Der Forschungsprototyp von Verbmobil

Alle technischen Ziele der ersten Phase von Verbmobil wurden voll erreicht und in einem Forschungsprotoypen realisiert:

- Erkennung fließend gesprochener Spontansprache für Deutsch, Japanisch und Englisch über Nahbesprechungsmikrophon
- Wortschatz von ca. 2500 Wörtern für die Übersetzungrichtung Deutsch nach Englisch
- Sprecheradaptives System mit sprecherunabhängigem Kern
- Linguistisch fundierte deutsche Basisgrammatik für Spontansprache mit tiefer und flacher semantischer Analyse
- Gesprochene Klärungsdialog zwischen dem Benutzer und dem Verbmobil-System bei Spracherkennungs- und Verstehensproblemen
- Semantischer Transfer für Deutsch - Englisch und Japanisch - Englisch
- Sprachgenerierung für Englisch und für deutsche Paraphrasen
- mehr als 70% approximativ korrekte Übersetzungen bei der End-to-End Evaluation in der Domäne Terminverhandlung
- Reine Softwarelösung für alle Module auf Standardhardware
- Netto-Verarbeitungszeit < sechsfache Echtzeit, bezogen auf die Länge des Eingabe-Sprachsignals.

Wie die Architekturübersicht in Abb. 1 zeigt, wurde Verbmobil als hochgradig nebenläufiges System nach dem Multiagenten-Prinzip mit zahlreichen Kommunikationsschnittstellen zwischen den Verarbeitungsmodulen vollständig objektorientiert realisiert. Die Benutzeroberfläche, durch die auch der Verarbeitungsablauf visualisiert wird, zeigt nur die Hauptmodule der insgesamt 43 Systemkomponenten.

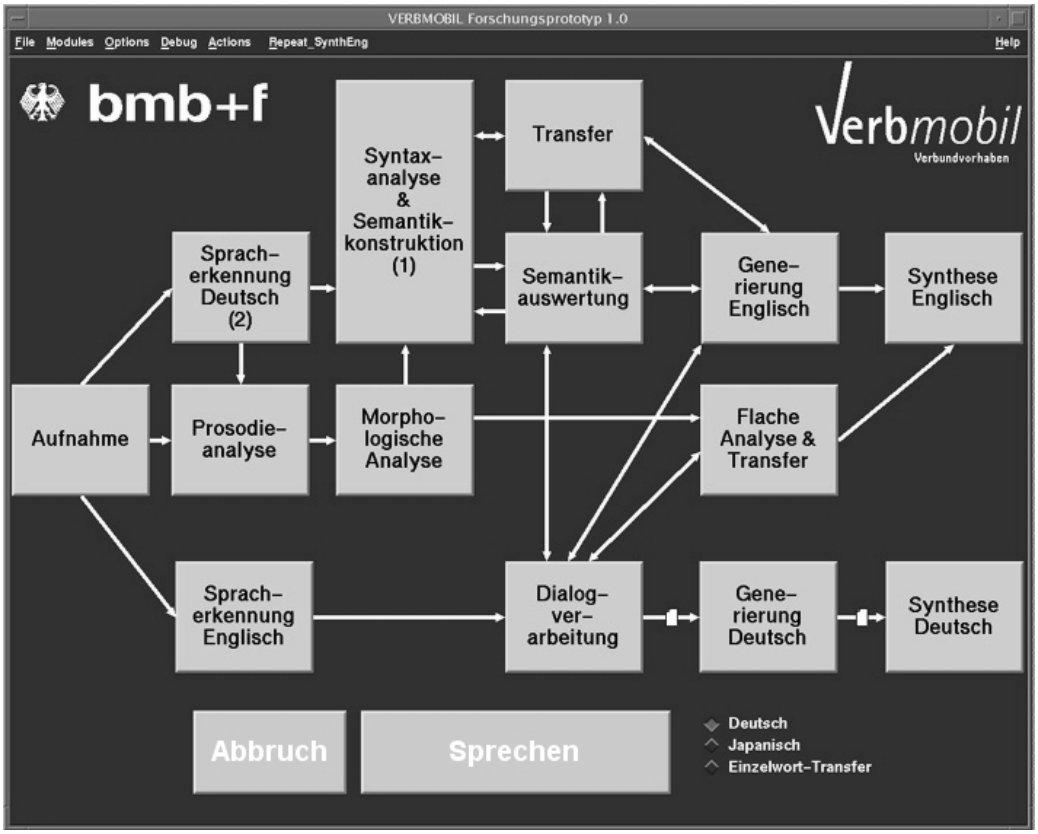

Abb.1: Die Benutzeroberfläche des Forschungsprototypen von Verbmobil

In der ersten Phase von Verbmobil (siehe [5]) wurde erstmals ein echtzeitfähiges sprecherunabhängiges System für deutsche Spontansprache mit hoher Erkennungsleistung realisiert. Die als möglich erkannten Wörter werden mit Wahrscheinlichkeiten bewertet und in einem Worthypothesengraphen dargestellt. Dabei konnte die Wortfehlerrate in der Testdomäne von über 50% zu Beginn des Projektes auf 14% bei der letzten Evaluation am Ende der ersten Phase reduziert werden, was derzeit weltweit die beste Erkennungsleistung für Spontansprache darstellt (vgl. Intern. Conf. on Acoustics, Speech and Signal Processing, ICASSP 97, Session Speech-to-Speech Translation, Vol. 1, p. 71-115, München). Der Forschungsprototyp kann extrem lange Sätze verarbeiten, wie sie spontan in Verhandlungsdialogen häufig gebildet werden (im Gegensatz zur Kommando-Eingabe in sprachgesteuerten Systemen oder in Datenbank-Anfragesystemen).

Verbmobil ist das einzige System, das prosodische Information zur Interpretation von Äußerungen auf mehreren Verarbeitungsebenen verwendet (siehe [3]), u.a. zur Segmentierung langer Gesprächsbeiträge, zur grammatischen Verarbeitung, Bedeutungsdeterminierung, Übersetzung und Dialogverarbeitung. Da keine Interpunktion zur Segmentierung verwendet werden kann, ist die durch die Prosodie-Komponente von Verbmobil erzeugte Satzgrenzeninformation sehr wichtig, da sie die syntaktische Analyse um 92% beschleunigt und die Anzahl der Lesarten um 96% reduziert.

Im Verbmobil-Forschungsprototypen wurde erstmals ein durchgängiger Analyseweg von der spontansprachlichen Eingabe bis zur Diskurssemantik realisiert, die mithilfe von DRS (Discourse Representation Structures) eine Integration von Sprecher-, Hörer- und Kontextbezug ermöglicht. Die verschränkte syntaktisch- semantische Analyse (siehe [1]) auf der Basis von linguistisch-fundierten Unifikationsgrammatiken und kompositioneller Bedeutungsrepräsentation sichert das frühzeitige Einbeziehen von Bedeutungsinformation. Die linearisierte, minimal-rekursive Bedeutungsdarstellung in Form sog. VIT (Verbmobil Interface Terms) wurde in Hinblick auf einen effizienten Transfer optimiert. Die Unterspezifikation von Mehrdeutigkeiten ermöglicht eine kompakte und ambiguitätserhaltende Darstellung von Bedeutungsstrukturen, ohne unnötig Disambiguierungsaufwand bei parallelen Formen der Mehrdeutigkeit in Quell- und Zielsprache zu erzeugen.

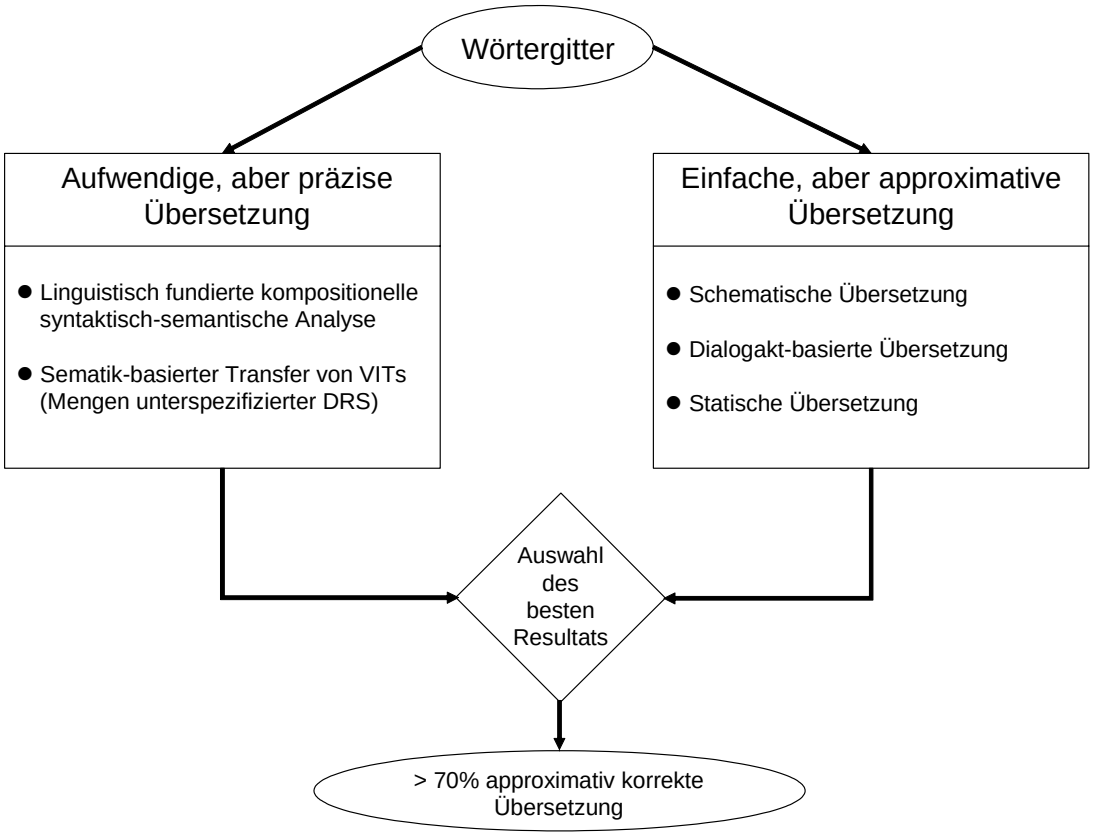

Abb. 2: Die Kombination flacher und tiefer Übersetzungsverfahren

Neuartig im Forschungsprototypen ist auch, daß ein hocheffizienter Sprachgenerator auf der Basis einer reversiblen HPSG-Grammatik für das Englische realisiert werden konnte, der eine durchschnittliche Generierungszeit pro Satz von nur 0.7 Sekunden aufweist. In einer Vorverarbeitungsphase wird dabei die am Center for the Study of Language and Information (CSLI) an der Stanford University im Auftrag für Verbmobil entwickelte HPSG-Grammatik in lexikalisierte TAG-Bäume im Baumadjunktionsformalismus transformiert, so daß zur Generierungszeit nur extrem schnelle Baumadjunktionen und vereinfachte Merkmalsunifikationen stattfinden. Der lexikalisch gesteuerte Sprachgenerator VM-GECO kann auch aus unterspezifizierten semantischen Strukturen mithilfe seiner 2730 Mikroplanungsregeln und seines hierarchischen Constraint-Progagierungssystems Dialogäußerungen erzeugen.

Die Gesamtverarbeitungszeit von der Eingabe bis zur Ausgabe teilt sich im Mittel folgendermaßen auf: Spracherkennung 38%, Prosodie 17%, Syntax und Semantik 25%, Semantische Auswertung und Dialog 14%, Transfer 3% und Generierung 3%. Es wird also derzeit noch über 50% der Verarbeitungszeit pro Dialogbeitrag in die signalnahe Erkennung der akustischen Eingabe investiert. Bemerkenswert ist, daß die eigentliche Übersetzung durch die semantik-basierte Transferkomponente sehr wenig Verarbeitungsaufwand erfordert, da hierbei lediglich noch eine quellsprachliche Bedeutungsrepräsentation in eine Darstellung in der Zielsprache transformiert werden muß.

Bei einer Gesamtevaluation der Übersetzungsleistung des Systems (siehe [2]), bei der mehr als 25000 Übersetzungsbeispiele durch Dolmetscher bewertet wurden, zeigte sich, daß derzeit ca. 70% der angebotenen Übersetzungen approximativ korrekt sind, d.h. den vom Sprecher intendierten Inhalt des Dialogebeitrags in der Zielsprache verständlich wiedergeben. Dieses Ergebnis konnte nur mithilfe des hybriden Übersetzungsansatzes von Verbmobil realisiert werden, der beispielorientierte Verfahren, Dialogakt-basierte Übersetzung und statistische Übersetzung mit einer tiefen linguistisch fundierten Analyse fallweise kombiniert (vgl. Abb. 2).

Die Ziele der zweiten Phase von Verbmobil

In der zweiten Phase von Verbmobil steht die robuste und direkte Übersetzung spontansprachlicher Dialoge für die Sprachpaare Deutsch-Englisch (10000 Wörter) und Deutsch-Japanisch (2500 Wörter) (Multilingualität) im Vordergrund. Die angestrebte Multilingualität des Gesamtsystems setzt weitgehend sprachenunabhängige und möglichst reversible Verarbeitungsverfahren und Wissensquellen sowohl bei der Sprachanalyse als auch beim Transfer und der Generierung

voraus. Bei der Übersetzung werden unterspezifizierte Repräsentationen systematisch verwendet, so daß ein Transfer gepackter Strukturen möglich wird. Es werden sprachtechnologische Werkzeuge entwickelt, die auf die multilingualen Anforderungen abgestimmt sind und die auch für große Wortschätze z.B. durch semiautomatische Adaptions- und Lernverfahren eine zeit- und kostengünstige Systemrealisierung ermöglichen.

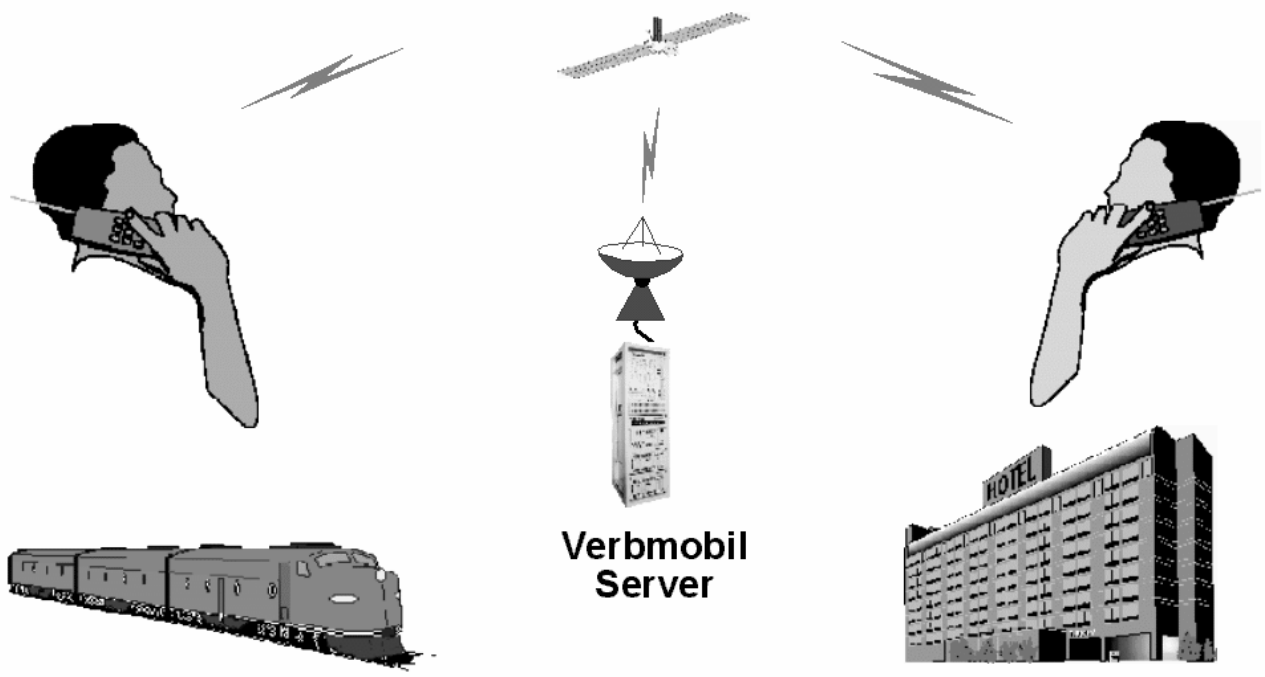

Fig.3: Anwendungsszenario Reiseplanung für die zweite Phase von Verbmobil

Verbmobil bleibt auch in der zweiten Phase domänenabhängig, soll aber rasch auf verschiedene Gesprächsthemen umschaltbar sein (Multifunktionalität). Es wird untersucht, ob durch die automatische Erkennung des Hauptgesprächsthemas oder von Themenwechseln (Topic Detection) eine explizite Umschaltung auf ein anderes Domänenmodell durch den Benutzer entfallen kann. Ein Schwerpunkt liegt auf der effizienten, weitgehend automatisierten Adaptierbarkeit der sprachlichen Wissensquellen wie Sprachmodelle und Lexika an neue Domänen.

Am Ende der zweiten Phase soll ein System verfügbar sein, das nicht mehr von der Spracheingabe über ein Nahbesprechungsmikrophon abhängt, so daß selbst Freisprechen ermöglicht wird. Für die Spracherkennung ergibt sich hiermit die Herausforderung der Verarbeitung von Spontansprache in Telephon- oder sogar Funkqualität unter Berücksichtigung von Mikrophonwechsel. In einem ersten Anwendungsszenario (vgl. Abb. 3) kann ein ausländischer Besucher z.B. über ein GSM-Funktelephon mittels des Verbmobil-Servers seinen Dialog zur Planung einer Reise nach Hannover (Verkehrsverbindungen, Hotelreservierung, Eintrittskarten) mit dem deutschsprachigen Partnervollständig in Englisch führen. Dabei soll durch eine automatische Anfangs- und Endedetektion von Redebeiträgen der explizite Aufnahme- und Analysestart durch den Benutzer im Gegensatz zum Forschungsprototyp entfallen. Insgesamt wird angestrebt, Verbmobil selbst als vollständig sprachgesteuertes System zu realisieren.

In einem zweiten Anwendungsszenario wird Verbmobil in mehrsprachigen multimedialen Telekonferenzen zur Reiseplanung mit mehr als zwei Partnern getestet (Multiparty-Situation). Die Übersetzung erfolgt bidirektional für die einzelnen Sprachpaare. Geht man von einer multilingualen Telekonferenz aus, z.B. mit einem Deutschen, einem Japaner und einem Amerikaner, so wird Verbmobil einen deutschen Dialogbeitrag parallel ins Englische und ins Japanische übersetzen, um bei sämtlichen Dialogpartnern den gleichen Informationsstand zu garantieren.

Dialogakte modellieren die intendierte Interpretation von Äußerungen in Dialogen und stellen Informationstypen dar, die von spontansprachlichen Performanzphänomenen abstrahieren und nur die relevante Information einer Äußerung repräsentieren. Statistische Methoden aus der Sprachmodellierung werden benutzt, um Dialogakte zu erkennen (z.Z. ca. 70% korrekte Erkennung), aber auch um Dialogakte vorherzusagen. Es resultiert eine Makrostruktur des Dialogs, in der jeder Turn nur noch durch seinen zentralen Gehalt repräsentiert ist. Diese Information wird auch zur Top-Down-Steuerung der mikrostrukturellen Analyseebenen verwendet. Die kondensierte Form des Dialogs, die eine Art Dialogprotokoll darstellt, bietet das Rohmaterial für eine schriftliche Fixierung des Verhandlungsverlaufs. Zur Erstellung der Protokolle, die am Ende

einer Telekooperationssitzung jeder Teilnehmer in seiner Muttersprache anfordern kann, wird neben der Generierung von gesprochener Sprache auch die Generierung von Schriftsprache erforderlich.

In Phase 2 wird eine entscheidende Verbesserung der Qualität der Sprachausgabe ein vordringliches Ziel sein, wobei im Gegensatz zur ersten Phase weniger die artikulatorische Qualität, sondern die Prosodie und satzübergreifende Phänomene den Schwerpunkt bilden. Dabei wird konsequent der Weg einer engen Kopplung von Sprachgenerierung und Sprachsynthese beschriten, um zu leistungsfähigen „Content-to-Speech"-Komponenten für die vorgesehenen Zielsprachen zu kommen.

Die Qualitätsbarrieren der klassischen „Text-to-Speech-Synthese" sollen auf diese Weise überwunden werden. Dabei stellt die Generatorausgabe bereits zielgerichtet für die prosodie-orientierte Synthese annotierte Information bereit, welche die Reanalyse innerhalb der Synthesekomponente überflüssig macht. Vom Sprecher intendierte Hervorhebungen werden in der integrierten Sprachgenerierungs- und Synthesekomponente alternativ durch syntaktische Mittel (etwa topikalisierte Konstruktionen), durch intonatorische/prosodische Mittel (Satzbetonung, Lautdauern, Pausen) oder deren Kombinationen realisiert.

In der zweiten Phase von Verbmobil werden die beiden Verarbeitungsstränge - die tiefe wissensbasierte und die reduktionistische flache Analyse - weiter integriert. So wird auch die Verarbeitung von syntaktisch und semantisch deformierten Äußerungen möglich. Eine neue Komponente zum partiellen Parsing fügt in den Interpretationsgraphen partielle syntaktisch-semantische Elemente ein, die Folgen von Worthypothesen überspannen. Die generelle Vorgehensweise besteht darin, daß die Äußerung inkrementell durch stochastische endliche Automaten in partiell syntaktisch und semantisch interpretierbare Einheiten aufgebrochen wird.

Zusammenfassend kann man festhalten, daß der Forschungsprototyp von Verbmobil (VM-I) derzeit im internationalen Vergleich einen Spitzenplatz einnimmt und die sehr anspruchsvolle Zielsetzung der zweiten Phase (VM-II) auch im Jahr 2000 den Konsortialpartnern einen internationalen Vorsprung im Bereich der Sprachtechnologie sichert. Die zweite Phase von Verbmobil kann zusammenfassend wie folgt gekennzeichnet werden:

- *Multifunktionalität*: Verbmobil soll rasch auf neue Gesprächsdomänen einstellbar sein.
- *Multilingualität*: Verbmobil soll spontane Dialoge in mehrere Sprachen übersetzen können.
- *Multimedialität*: Verbmobil soll in internationalen Multimedia-Anwendungen Übersetzungshilfe anbieten.
- *Mobilitä:*Durch einen Sprachserver soll Verbmobil auch über Handy nutzbar sein.
- *Multiparty-Funktionalität*: Verbmobil soll nicht nur in Dialogsituationen, sondern auch in Telekooperations-anwendungen mit vielen Gesprächspartnern Übersetzungshilfe leisten.

Auch in der zweiten Phase von Verbmobil liegt die Gesamtprojektleitung beim Deutschen Forschungszentrum für Künstliche Intelligenz (DFKI GmbH). Industriepartner sind: Daimler-Benz, DASA, Philips, Siemens.

Forschungspartner sind: DFKI, LMU München, RWTH Aachen, TU Berlin, TU Dresden, TU München, Uni Bielefeld, Uni Bochum, Uni Bonn, Uni Braunschweig, Uni des Saarlandes, Uni Erlangen, Uni Hamburg, Uni Karlsruhe, Uni Stuttgart, Uni Tübingen.

Die bisherigen Projektergebnisse von Verbmobil wurden in 375 Publikationen dokumentiert. Eine umfassende Übersicht zum aktuellen Projektstand kann über das World Wide Web unter der URL:

http://www.dfki.de/verbmobil

abgerufen werden. Daher wird in der folgenden Literaturliste nur auf einige wenige Überblicksarbeiten verwiesen.

Literatur

[1] Hans Ulrich Block (1997): The Language Components in Verbmobil. In: Intern. Conf. on Acoustics, Speech and Signal Processing, ICASSP 97, Vol. 1, p. 79-83, München

[2] Thomas Bub, Wolfgang Wahlster, Alex Waibel (1997): VERBMOBIL: The Combination Of Deep And Shallow Processing For Spontaneous Speech Translation. In: Intern. Conf. on Acoustics, Speech and Signal Processing, ICASSP 97, Vol. 1, p. 71-74, München

[3] Heinrich Niemann, Elmar Nöth, Andreas Kießling, Rolf Kompe, Anton Batliner (1997): Prosodic Processing and its Use in Verbmobil. In: Intern. Conf. on Acoustics, Speech and Signal Processing, ICASSP 97, Vol. 1, p. 75-78, München.

[4] Elmar Nöth, Anton Batliner, Andreas Kiessling, Ralf Kompe, Heinrich Niemann, Volker Warnke: Spracherkennung und Prosodie (in diesem Band)

[5] Wolfgang Wahlster (1993): Verbmobil: Translation of Face-to-face Dialogs. In: 3rd European Conference on Speech Communication and Technology, Eurospeech'93, Berlin, p. 29-38.