EVAR: Ein sprachverstehendes Dialogsystem

W. Eckert[+], G. Fink[*], A. Kießling[+], R. Kompe[+], T. Kuhn[+], F. Kummert[*], M. Mast[+], H. Niemann[+], E. Nöth[+], R. Prechtel[+], S. Rieck[+], G. Sagerer[*], A. Scheuer[+], G. Schukat-Talamazzini[+], B. Seestaedt[*]

[*]: AG Angewandte Informatik, Universität Bielefeld
e-mail: sagerer@techfak.uni-bielefeld.de

[+]: Lehrstuhl für Informatik 5 (Mustererkennung), Friedrich-Alexander-Universität Erlangen-Nürnberg, e-mail: niemann@informatik.uni-erlangen.de

Kurzfassung: Dieser Artikel befaßt sich mit dem sprachverstehenden Dialogsystem EVAR, insbesondere mit der linguistischen Verarbeitung des Systems. Aufgabe von EVAR ist die Führung eines informationsabfragenden Dialogs über das deutsche InterCity-Zugsystem. Das linguistische Wissen ist einheitlich in einem semantischen Netz repräsentiert. Die Wissensbasis ist gemäß einem geschichteten linguistischen Modell wohlstrukturiert. Schnittstelle zur Spracherkennung ist die Worthypothesen-Ebene. Der Kontrollalgorithmus ist anwendungsunabhängig formuliert und erlaubt das dynamische Umschalten zwischen den beiden grundlegenden Analysestrategien *top-down* und *bottom-up*. Das im System repräsentierte Wissen wird sowohl zur Steuerung der Erkennungsphase als auch in der Verstehensphase benutzt. Das System ist in der Lage, Anfragen trotz fehlerhafter Erkennungsergebnisse zu bearbeiten. Ergebnisse für eine sprecherabhängige- und eine Mehrsprecher-Version der Erkennung werden vorgestellt.

Abstract: This article presents the speech understanding and dialog system EVAR and concentrates on the linguistic processing of the system. The task of EVAR is to lead an information retrieval dialogue about the German InterCity train system. The linguistic knowledge is uniformly represented in a semantic network structure. The knowledge base is well structured following the layered linguistic model. Interface to the speech recognition is the level of word hypotheses. The control algorithm is independent of the application and allows to alternate dynamically between the two fundamental control strategies *top-down* and *bottom-up*. The linguistic knowledge represented in the system is used to control the recognition phase and for the understanding phase. The system can handle defect input due to recognition errors. Results are presented for speaker-dependent and multi-speaker versions of the recognition module.

1 Motivation

Eine der wichtigsten Anwendungen für das Forschungsgebiet "Automatische Spracherkennung" ist die Abfrage von Information. Häufig ist hierbei ein Klärungsdialog notwendig, z.B. falls der Benutzer unterspezifizierte Fragen stellt. Leider können die Ergebnisse der Forschungen zum Gebiet "natürlichsprachlicher Datenbankzugang" [8] nicht direkt übernommen werden. Zwei Hauptgründe lassen sich hierfür anführen:

- Im Gegensatz zu getipptem Input (NL-System) kann bei gesprochenem Input nicht von syntaktisch wohlgeformten Sätzen ausgegangen werden. Typische Eigenschaften spontaner Sprache, wie Häsitationen und (für Schriftsprache) ungewöhnliche Anordnung der Satzglieder müssen für ein reales System modelliert werden.

- Während bei NL-Systemen von einer 100-prozentigen "Erkennung" ausgegangen werden kann, muß bei Systemen mit gesprochener Eingabe die Unsicherheit, bedingt durch Aussprachevariationen und fehlerhafte Erkennung, modelliert werden. Da der Wortschatz eines Erkennungssystems immer begrenzt ist, und die Erkennungsverfahren immer ein Ähnlichkeitsmaß zwischen einer lexikalischen Repräsentation eines beliebigen Wortes und einem Sprachsignal berechnen, kann das System

nicht feststellen, ob der Benutzer ein dem System unbekanntes Wort äußert oder nicht.

Das in Erlangen und Bielefeld entwickelte System EVAR versucht, diese Gegebenheiten zu berücksichtigen: Das System versucht zunächst, die gesamte Äußerung zu interpretieren. Ist es dazu nicht in der Lage, z.B. aufgrund von Fehlern in der Worterkennung oder wegen der Verwendung von dem System unbekannten Wörtern, so wird versucht, die Anfrage mit der bis dahin erstellten partiellen Interpretation zu verarbeiten. Somit können die oben genannten Probleme zumindest zum Teil modelliert werden. Die syntaktische Korrektheit wird auf der Konstituenten-, jedoch nicht auf der Satzebene gefordert. Somit können einige für Spontansprache typische Satzkonstrukte verarbeitet werden.

Der Beitrag ist folgendermaßen aufgebaut: Abschnitt 2 stellt die Systemkomponenten *akustische Verarbeitung, linguistische Wissenbasis* sowie *Dialog* und *Anwendungsdatenbank* vor. In Abschnitt 3 wird die integrierte Verarbeitung der Spracherkennungsebene und der linguistischen Ebenen behandelt. Abschnitt 4 stellt erste Erkennungsergebnisse zur dialogfähigen Version von EVAR vor. Zum Schluß wird ein Ausblick auf weitere Arbeiten gegeben.

Schwerpunkt dieses Beitrages ist die Darstellung der Interaktion zwischen der akustischen und den linguistischen Analyse-Ebenen. Die Analyse verläuft nicht sequentiell, da Vorerwartungen und Analyse-Zwischenergebnisse einer Wissensebene jederzeit an die anderen Ebenen weitergegeben werden. Die einzelnen System-Komponenten werden aus Platzgründen jeweils nur soweit, wie es für die Darstellung der Interaktion notwendig ist, vorgestellt (und somit unterschiedlich detailliert). Für weitere Einzelheiten muß auf die angegebene Literatur verwiesen werden.

2 Architektur und Systemkomponenten

2.1 Architektur

Das System EVAR gliedert sich in die Module *Worterkennung, Syntax, Semantik, Pragmatik und Dialog.* Die Wissensbasis ist gemäß einem geschichteten linguistischen Modell strukturiert ([5]). Einen Überblick über das System gibt Bild 1, siehe auch [6].

Die vier Module der linguistischen Analyse sind mit einem einheitlichen Formalismus realisiert, dem **ER**langer Semantischen **NE**tzwerk Sy**ST**em **ERNEST** (Ein Überblick über das System und seinen Einsatz bei verschiedenen Anwendungen der Mustererkennung findet sich in [7]). Dieses System erlaubt die Modellierung von Begriffen, wie Präpositionalgruppe und Dialogschritt, und die Darstellung von Beziehungen zwischen diesen Begriffen. Jedes der Module bildet innerhalb des Netzwerks eine Abstraktionsebene. Es stehen drei Kantentypen zur Verfügung:

- *Spezialisierungen (spez)* erlauben Verfeinerungen von allgemeinen Konzepten, wobei die Vererbung von Eigenschaften (Attributen, Bestandteilen, Konkretisierungen) möglich ist.
- *Bestandteile (bst)*: Beziehungen zwischen einem Konzept und anderen Konzepten der Wissensbasis, aus denen sich dieses Konzept zusammensetzt. So ist z.B. *Ankunftsort* ein Bestandteil des Konzeptes, das eine Fahrplanauskunft modelliert.
- *Konkretisierungen (kon)* ermöglichen die Verbindung von Wissen aus höheren Abstraktionsebenen mit niedrigeren Ebenen. Mit ihnen wird z.B. die Modellierung des pragmatischen Konzeptes *P_ANKUNFTSORT* durch das semantische Konzept

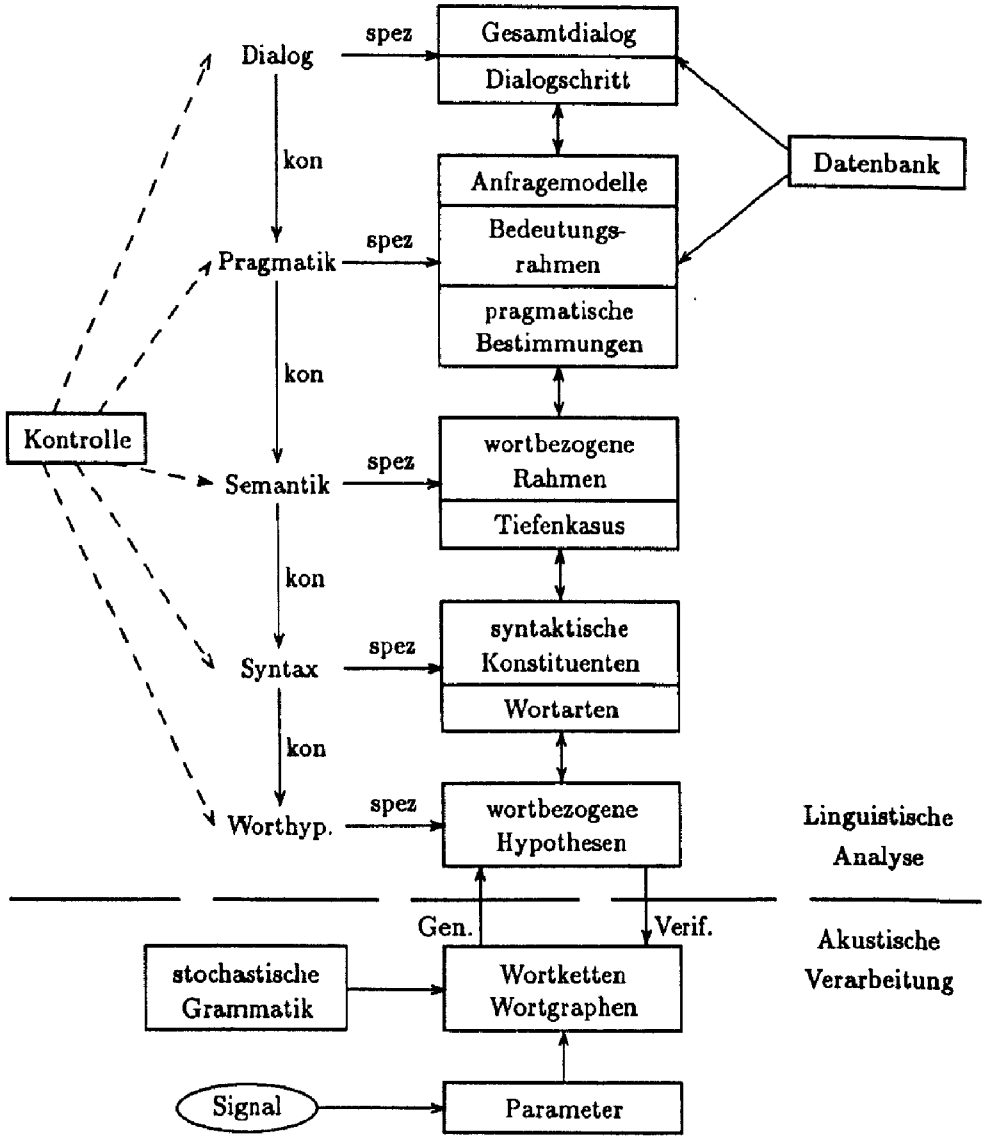

Bild 1: Struktur des Systems EVAR

S_GOAL gekennzeichnet. Diese Kanten dienen während der Analyse zur top-down Prädiktion und zur bottom-up Interpretation von signalnahen Konzepten durch abstraktere Konzepte.

Das Kontrollmodul ist problemunabhängig und wird von ERNEST bereitgestellt. Es basiert auf dem A^*-Algorithmus.

Die Schnittstelle zwischen der linguistischen Analyse und der Worterkennung ist über eigens dafür geschaffene Konzepte realisiert. Diese dienen zur Anforderung von Worthypothesen. Die Anforderung kann gemäß dem aktuellen Status der Analyse auf bestimmte Signalbereiche und eine Teilmenge des Lexikons restringiert sein.

Das EVAR-System wurde in C auf einer DECstation 5000/200 implementiert. Die Spracheingabe erfolgt über ein Desklab der Firma Gradient, das über eine SCSI-Schnittstelle mit der DECstation verbunden ist. Es wird keine Spezialhardware verwendet.

2.2 Akustische Verarbeitung

Die akustische Verarbeitung in EVAR analysiert das Sprachsignal der Benutzeräußerung hinsichtlich der gesprochenen Wörter. Eine sprecherunabhängige Worterkennung mit großem Wortschatz und einer den Anwendungsbereich komfortabel abdeckenden Grammatik ist nach dem gegenwärtigen Stand der Technik nicht völlig fehlerfrei möglich. Daher wurde die akustische Analyse in ein von der linguistischen Verarbeitung gesteuertes heuristisches Suchverfahren (A^*) eingebettet, in deren Verlauf sie zwei unterschiedliche Aufgaben zu erfüllen hat:

- Die initiale Erzeugung von Worthypothesen, welche der nachfolgenden Suche als Saatpunkte dienen.
- Die akustische Verifikation partieller Interpretationen der Benutzereingabe.

Das tiefpaßgefilterte (6.4 kHz) Signal wird mit 16 kHz abgetastet und mit 14 Bit quantisiert. Anschließend wird die Eingabe in Sprach- und Stille-Intervalle aufgeteilt. Während der Kurzzeitanalyse (10 msec Fenster) werden 12 Cepstralparameter des *mel*-Spektrums sowie, zur Erfassung der dynamischen Eigenschaften der Sprachproduktion, deren zeitliche Ableitungen berechnet.

Akustische Wortmodelle werden aus Modellen kleinerer Spracheinheiten zusammengesetzt. Die Rolle der Wortuntereinheiten übernehmen hier kontextunabhängige sowie, zur genaueren Modellierung phonetischer Koartikulation, kontextabhängige Phone. Zur akustischen Beschreibung durch ein Markovmodell (HMM, [9]) gelangen all jene Spracheinheiten, deren hinreichende Verfügbarkeit in der Lernstichprobe des Erkenners gesichert ist.

Die Erzeugung von Hypothesen über vermutlich gesprochene Wörter W_i und ihre zeitlichen Begrenzungen im Signal t_i^{anf}, t_i^{end} geschieht in einem zweistufigen, HMM-basierten Verfahren [3]. Zuerst wird die Spracheingabe nach einem Inventar kontextabhängiger Phone (Tri-, Bi- und Monophone) klassifiziert und segmentiert. Die resultierende Segmentfolge dient als Eingabe der Links-Rechts-Suche auf Wortebene, deren Ergebnis eine Menge gut passender Worthypothesen $H_i = (W_i, t_i^{anf}, t_i^{end})$ ist. Die Elemente dieser Hypothesenmenge dienen der integrierten Analyse als initiale Ankerpunkte. Der Suchraum wird von einem Vokabular von 1081 Wortformen aufgespannt. Er wird durch eine stochastische Grammatik eingeschränkt, die zur Zeit 95 syntaktisch-semantische Wortkategorien unterscheidet [1] und eine Perplexität von 111 besitzt.

Die Verifikation partieller Interpretationen des Eingabesignals besteht in einer akustischen Bewertung hypothetischer Wortketten, welche unter Umständen mit zeitlichen Unterbrechungen vorliegen können. Bewertungsgrundlage sind Markovmodelle mit kontinuierlichen Ausgabedichten; in dieser Phase werden außer Phonen diverse Lautverbindungen, Halbsilben, Silben, Morpheme und selbst komplette Wörter als Entscheidungseinheiten herangezogen, um eine möglichst detaillierte Erfassung kontextueller Aussprachevariation zu garantieren [11]. Als Kettenbewertung der Gesamtkontrolle fungiert dann der negative Logarithmus der bedingten Produktionswahrscheinlichkeit der akustischen Evidenz X_K aus dem Wortkettenmodell λ_K:

$$q(H_K) = -\log P(X_K \mid \lambda_K)$$

2.3 Linguistische Wissensbasis

Die Beschreibung des linguistischen Wissens ist im semantischen Netzwerk in die folgenden vier Abstraktionsebenen unterteilt:

- Die *Hypothesenebene* stellt die Schnittstelle zur Worterkennung dar.
- Die *Syntaxebene* enthält Konzepte, die zum einen syntaktische Konstituenten wie Verbalgruppe oder Präpositionalgruppe (*SY_PNG*) und zum anderen spezielle Zeitangaben wie Datum und Uhrzeit modellieren. Auf die Erstellung einer kompletten Satzgrammatik wurde verzichtet, da Stellungsregularitäten in gesprochener Sprache praktisch nur *innerhalb* von Konstituenten auftreten und die Anordnung der Konstituenten innerhalb einer Äußerung relativ frei ist. Die Konstituenten können diskontinuierlich sein, wie z.B. die Verbalgruppe in "Wann *fährt* der Zug ... *ab*?"
- Die Modellierung von Bedeutungen auf der *Semantikebene* beruht auf der Tiefenkasustheorie [2]. Sie geht davon aus, daß ein Verb für eine gewisse Bedeutung Leerstellen eröffnet, denen eine funktionale Rolle oder Tiefenkasus wie z.B. `Goal` (*S_GOAL*) zugeordnet wird. Dieses Vorgehen läßt sich auch auf Nomina übertragen.
- Die *Pragmatikebene* dient der Beschreibung anwendungsabhängiger Begriffe wie `Fahrplanauskunft`, `Ankunftsort` (*P_ANKUNFTSORT*) oder "`mit dem Zug fahren`". Die Modellierung orientiert sich stark an der Semantikebene. Die dort vorhandenen Konzepte werden hier auf ihre spezifische Bedeutung im Anwendungsbereich eingeschränkt.

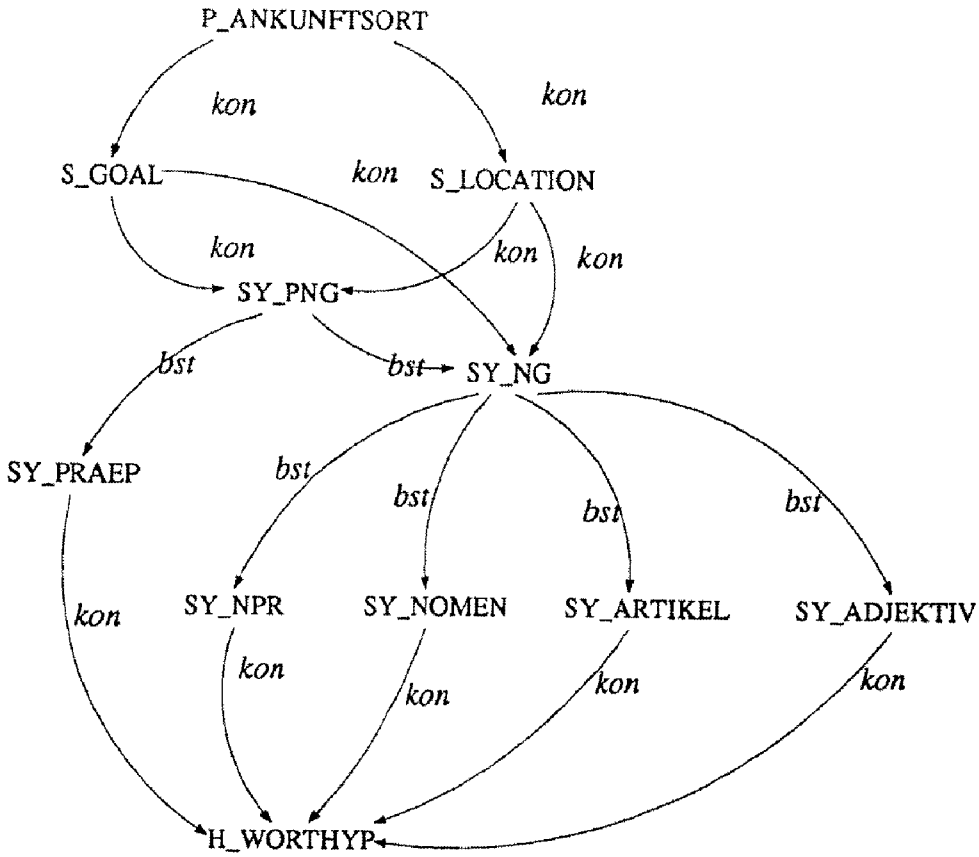

Bild 2: Modellierung des Begriffs `Ankunftsort` im semantischen Netz

Auf allen genannten Beschreibungsebenen können Abfolgerelationen zwischen Bestandteilen komplexerer Konstituenten spezifiziert werden. Bild 2 zeigt beispielhaft die Modellierung des Begriffs `Ankunftsort`. Syntaktische, semantische und anwendungsabhängige Merkmale werden durch Attribute der Netzwerkkonzepte beschrieben. Eine ausführliche Beschreibung der Wissensbasis findet sich in [4].

2.4 Dialog

Das *Dialogmodul* soll die Steuerung des Auskunftdialogs über einen begrenzten Aufgabenbereich übernehmen. Ein Dialogmodell wurde manuell aus einem Korpus von echten Telefondialogen mit Diskursbereich Reiseauskunft extrahiert. Grundelemente des Dialogmodells sind Dialogschritt-Typen, die in etwa Sprechakten entsprechen.

Dialoge lassen sich im allgemeinen grob in die Phasen *Einleitung, Hauptteil* und *Abschluß* gliedern. Einleitung und Abschluß bestehen hauptsächlich aus metakommunikativen Äußerungen wie Gruß- oder Dankformel. Der Hauptteil dient hier der Informationsgewinnung und läßt sich in weitere Abschnitte untergliedern. Zuerst stellt der Informationssuchende seine Anfrage, die dann bearbeitet und wenn möglich beantwortet wird. Diese Abschnitte lassen sich bis auf die Dialogschrittebene untergliedern. So kann z.B. die Bearbeitungs- oder Klärungsphase einer Anfrage aus einer Nachfrage nach einem fehlenden Parameter durch das System und der Ergänzung dieses Paramters durch den Benutzer bestehen.

Um eine Interpretation der Benutzeräußerungen relativ zum Dialogmodell zu ermöglichen, wurden die möglichen sprachlichen Realisierungen festgelegt, wobei die Besonderheiten gesprochener Sprache berücksichtigt wurden. Charakteristisch für Dialoge ist ein möglichst ökonomischer Sprachgebrauch, wie z.B. die Verwendung von Ellipsen und Proformen. Ellipsen sind syntaktisch oder semantisch unvollständige Sätze, die jedoch aus dem Zusammenhang interpretiert werden können. Dafür wird die Information vergangener Äußerungen in einem Dialoggedächtnis abgelegt.

Zur Beantwortung von Anfragen bzgl. des Anwendungsbereichs "InterCity-Auskunft", wurde der Zugriff auf eine Datenbank, die den Fahrplan enthält, ermöglicht. Um EVAR sinnvoll mit "naiven" Benutzern testen zu können, ist eine reale Datenbank notwendig. Aus diesem Grund wurde in EVAR das von der Deutschen Bundesbahn (DB) verwendete und von der Firma HaCon entwickelte Fahrplanauskunftssystem HaFas integriert.

Der HaFas-Fahrplan umfaßt sämtliche Verbindungen (über 40.000) der DB. Eine Fahrplanauskunft erfolgt in den Teilschritten *Wegesuche* (Ermittlung einer Menge möglicher Wege) und *Verbindungssuche* (Bestimmung der im Sinne von Reisezeit, Umsteigehäufigkeit, Weglänge, Fahrzeughierarchie, Fußwegen und Aufenthaltszeiten günstigsten Verbindung). Für eine Fahrplanauskunft müssen vom Benutzer Abfahrts- und Ankunftsort, Abfahrts- oder Ankunftszeitpunkt oder -intervall und optional Restriktionen auf bestimmte Zugklassen (z.B. IC), Routen oder Serviceleistungen (z.B. Schlafwagen) angegeben werden. Diese Parameter werden von EVAR aus der Benutzeräußerung ermittelt.

3 Integrierte Verarbeitung

3.1 Kontrollstrategie

Um die Vorerwartungen der linguistischen Wissensbasis möglichst umfassend zu nutzen, wird eine Hypothese nicht aufgrund einer sequentiellen Abarbeitung des Sprachsignals (Links-Rechts-Analyse, Inselstrategien) erweitert, sondern aufgrund von strukturellen Beziehungen. Dies bedeutet, daß für die Erweiterung einer Hypothese nicht die aktuelle Überdeckung des Sprachsignals mit Worthypothesen entscheidend ist, sondern die Analyse durch Vorerwartungen, die im semantischen Netz modelliert sind, gesteuert wird. Dadurch wird eine Worthypothese in jedem noch nicht überdeckten Abschnitt des Sprachsignals akzeptiert, sobald sie den Anforderungen der Wissensbasis genügt.

Ziel der linguistischen Analyse ist die Instantiierung eines Konzepts, das eine zulässige Be-

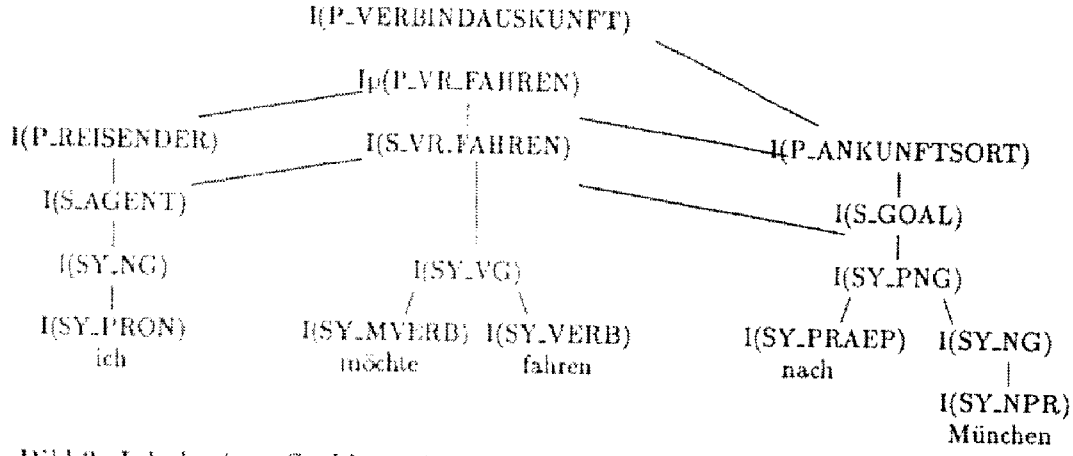

Bild 3: Inhalt eines Suchbaumknotens nach vollständiger linguistischer Analyse

nutzeranfrage repräsentiert, z.B. Verbindungsauskunft, Fahrplanauskunft. Wegen der unsicheren Worterkennung und wegen der vielfältigen Ausdrucksmöglichkeiten von Sprache erscheint weder ein rein datengetriebener Ansatz noch eine rein erwartungsgesteuerte Analyse erfolgversprechend. Deshalb verfolgen wir eine Strategie, die sowohl die akustischen Daten als auch die Vorerwartungen des linguistischen Modells berücksichtigt [4].

Der einheitliche Formalismus erlaubt trotz einer modularen Strukturierung des Wissens eine hohe Effizienz der Analyse durch frühzeitigen Einsatz von Wissen aller Ebenen (*constraint propagation*). Dabei werden während der Analyse Konzepte modifiziert und dadurch dem aktuellen Stand der Analyse angepaßt (weiß man z.B., daß die Worthypothese *München* Bestandteil einer Präpositionalgruppe (PNG) sein muß und daß es sich um einen Ankunftsort handeln muß, so sind lediglich die Präpositionen *in* und *nach* für die Realisierung dieser PNG zulässig und sie können sich im Signal nur unmittelbar vor dem Wort *München* befinden.

Bild 3 zeigt den Inhalt eines Suchbaumknotens nach der vollständigen linguistischen Analyse. Über die darin enthaltenen Instanzen werden den gesprochenen Wörtern syntaktische, semantische und aufgabenspezifische Interpretationen im Rahmen der linguistischen Wissensbasis zugeordnet. Der Satz "ich möchte nach München fahren" wurde als Verbindungswunsch mit dem Ankunftsort München analysiert. Da kein Abfahrtsort detektiert wurde, wird als Default der aktuelle Standort eingesetzt (in Bild 3 nicht dargestellt).

3.2 Bewertungen

Die *Zulässigkeit* einer Hypothese ist durch die strukturellen Beziehungen der zugeordneten Interpretation gegeben. Sie ist ein binäres Maß und überprüft, ob die linguistischen Restriktionen, wie zum Beispiel Kongruenz von Kasus, Numerus und Genus innerhalb einer Nominalgruppe, erfüllt sind. Da hier nur grundlegende Beziehungen getestet werden, die i.a. selbst bei spontan gesprochener Sprache zutreffen, wird eine Hypothese bei Verletzung dieser Restriktionen verworfen.

Die *Qualität* einer Hypothese ist durch die akustische Ähnlichkeit zwischen den zugrunde liegenden Wortketten und dem Sprachsignal definiert [11]. Um die Vergleichbarkeit der Bewertung unterschiedlich langer Interpretationen zu gewährleisten, wird eine Restabschätzung durchgeführt, die auf statistischen Annahmen über die Verteilung der Qualität korrekter Hypothesen basiert. Wie in [10] empirisch verifiziert wurde, ist die Qualität kor-

rekter Hypothesen q_k, die L Längeneinheiten umfassen, folgendermaßen normalverteilt:

$$\mu_k(L) = \mathcal{E}(q_k|L, \text{korrekt}) = \mu_k L \qquad \sigma_k^2(L) = \mathcal{E}((\mu_k(L) - q_k)^2|L, \text{korrekt}) = \sigma_k^2 L$$

Damit ergibt sich die Restschätzung für die Qualität eines nicht überdeckten Signalbereichs der Länge L zu $\tilde{q}(L) = \mu_k L - C\sigma_k\sqrt{L}$. Über die Konstante C wird die Wahrscheinlichkeit eingestellt, mit der die Restschätzung einen optimistischen Wert liefert. Da das für die akustische Qualität verwendete Bewertungsmaß additiv ist, ergibt sich für eine Hypothese H, die aus N Wortketten $K_i, 1 \leq i \leq N$ besteht und L Längeneinheiten nicht überdeckt, die folgende Qualitätsbewertung:

$$q(H) = \sum_{i=1}^{N} q(K_i) + \tilde{q}(L)$$

Die *Sicherheit* einer Hypothese orientiert sich an der Tatsache, daß längere Worthypothesen mit größerer Sicherheit korrekte Hypothesen darstellen [10]. Daneben werden benachbarte Hypothesen zu einer Kette zusammengefaßt und als Einheit verifiziert. Das heißt, es wird für K_i auf der Grundlage des Sprachsignals die akustische Qualität $q(K_i)$ bestimmt. Demzufolge läßt sich $s(H)$ als ein Maß für die Sicherheit einer Hypothese wie folgt definieren:

$$s(H) = \max_{1 \leq i \leq N}\{L(K_i)\}, \quad L(K_i) := \text{Länge der Kette } K_i$$

Als Maß für die *Relevanz* einer Hypothese bietet sich der Aufwand an, der benötigt wird, um eine vollständige Interpretation zu erreichen. Somit werden Hypothesen, die bereits einen Großteil des Sprachsignals überdecken, für die weitere Analyse bevorzugt. Damit gilt als Maß für die Relevanz einer Hypothese $r(H)$ = Anzahl der von H überdeckten Einheiten.

Zu einem Vektor zusammengefaßt läßt sich die Bewertung $b(H)$ einer Hypothese wie folgt darstellen $b(H) = (z(H), \hat{q}(H), s(H), r(H))$. Da der Bewertungsvektor monoton in jeder Komponente ist, ist das Bewertungsschema für den A*-Algorithmus zulässig, d.h. es wird die im obigen Sinne bestbewertete Äußerung gefunden. Die Vergleichbarkeit zwischen zwei Bewertungsvektoren wird durch einen komponentenweisen Vergleich erreicht, wobei bei den ersten beiden Komponenten Gleichheit über Intervalle definiert ist:

$$(x_1, ..., x_k) < (y_1, ..., y_k) \Leftrightarrow \exists x_i[x_i < y_i], 1 \leq i \leq k \wedge \forall x_l[x_l = y_l], l < i$$

4 Experimente und Ergebnisse

4.1 Test1: Sprecherabhängige Version

Bei den Experimenten zur sprecherabhängigen akustischen Verarbeitung wurden pro Dialogschritt bis zu 100 Worthypothesen (abhängig von der Länge der Benutzeräußerung) berechnet. Das Training wurde mit 100 anwendungsabhängigen und 200 phonetisch balancierten Sätzen durchgeführt. Zur Erzeugung der Worthypothesen wurde ein stochastisches Bigrammodell mit der Perplexität 111 verwendet [3].

Jeder Dialog kann aus bis zu fünf Dialogschritten bestehen: der Benutzer beginnt mit einer Anfrage, wobei möglicherweise eine Begrüßungsfloskel vorausgeschickt wird. Daraufhin erfragt das System weitere Parameter, die für einen Datenbankzugriff notwendig sind, oder erwartet die Bestätigung der Daten. Nachdem der Benutzer die Parameter bestätigt oder korrigiert hat, wird eine Datenbankanfrage erzeugt und das Resultat in

einer Antwortschablone dargestellt. Schließlich wird die akustische Ausgabe durch einen Sprachsynthesegerät der Firma Daimler-Benz generiert.

In 68 Fällen der 85 Testdialoge mit insgesamt 170 Benutzeräußerungen wurde der Test erfolgreich beendet. Davon war in drei Fällen eine Korrektur bei einer Bestätigungsfrage notwendig. 17 Dialoge konnten nicht erfolgreich durchgeführt werden, das heißt die Datenbankanfrage lieferte nicht die erwarteten Ergebnisse. Dies ist auf Analysefehler oder Speicherbegrenzung zurückzuführen. Im Mittel dauerte ein Dialog 3:57 Minuten, die mittlere Rechenzeit für die linguistische Analyse betrug 1:32 Minuten. Die akustische Verarbeitung bis zur Generierung der Worthypothesen geschah in 3,5-facher Echtzeit. Der Suchbaum bestand durchschnittlich aus 1390 Knoten.

4.2 Test2: Mehrsprecher-Version

Zur Worterkennung und -verifikation wurde das ISADORA–System [11] ohne Sprachmodell verwendet. Daher ist die Perplexität etwa genauso groß wie die Kardinalität des Lexikons. Von jedem Sprecher wurden 500 anwendungsabhängige Sätze zum Training des Worterkennungssystems verwendet. Die linguistische Analyse basiert auf den Worthypothesen, die aus den jeweils 10 besten Wortketten gebildet wurden.

Aufgrund der hohen Perplexität wurde lediglich eine Wortakkuratheit von 74.6% im 4–Sprecher–Modus erreicht, wodurch in einigen Sätzen für die linguistische Analyse nicht alle gesprochenen Wörter hypothetisiert werden konnten. Daher wurde die für eine Interpretation notwendige Überdeckung des Sprachsignals durch Worthypothesen reduziert und ein erweitertes Dialogmodell in das System integriert. Um eine erfolgreiche linguistische Analyse zu gewährleisten, wurde die Überdeckung auf 2/3 eingestellt, was die korrekte Interpretation von Äußerungen trotz fehlender Worthypothesen erlaubt. Desweiteren können fehlerhafte Interpretationen durch zusätzliche Dialogschritte korrigiert werden, bis alle notwendigen Werte für eine Datenbankabfrage bestätigt sind.

Das System wurde von einem in der Trainingsstichprobe vertretenen Sprecher mit 50 Dialogen getestet. 60% der Dialoge wurden ohne Korrekturdialog erfolgreich beendet, wobei im Mittel ein Dialog aus 2.7 Äußerungen bestand. 14% der Dialoge konnten unter Verwendung eines Korrekturschrittes beendet werden, wobei jeweils durchschnittlich 3.4 Benutzeräußerungen notwendig waren.

Die linguistische Analyse einer Äußerung benötigte durchschnittlich 25.5 Sekunden Rechenzeit, wobei 478 Suchbaumknoten erzeugt wurden.

Der Dialog konnte in den übrigen 26% nicht erfolgreich beendet werden. Tabelle 1 faßt die Ergebnisse der beiden Experimente zusammen.

	sprecherabhängig	Mehrsprecher
Anzahl der Dialoge	85	50
erfolgreich beendete Dialoge	68 (80%)	37 (74%)
davon mit Korrekturen	3 (4%)	7 (14%)
erfolglos	17 (20%)	13 (26%)

Tabelle 1: Zusammenfassung der Ergebnisse zur sprecherabhängigen und zur Mehrsprecher-Version der Erkennung

5 Zusammenfassung und Ausblick

Es wurde ein System-Ansatz zum Führen eines informationsabfragenden Mensch-Maschine-Dialogs vorgestellt. Wichtige Charakteristika des Systems sind die uniforme Repräsentation des linguistischen Wissens in einem semantischen Netz sowie die Verarbeitung defekter Eingabe aufgrund fehlerhafter Erkennung oder Verwendung unbekannter Wörter. Bei ersten Experimenten mit einer sprecherabhängigen (Mehrsprecher-) Version der Erkennung konnten 80 (74) % der gesprochenen Dialoge erfolgreich abgeschlossen werden.

Ein Schwerpunkt der zukünftigen Arbeiten ist die Erweiterung der linguistischen Kompetenz des Systems. Besonderes Augenmerk liegt dabei auf der Verwendung prosodischer Information zur Unterstützung der Erkennung und der linguistischen Analyse, der Verbesserung der Anaphernauflösung, der Interpretation von Mehrsatzäußerungen, der verbesserten Modellierung von Phänomenen spontaner Sprache sowie der Verwendung linguistisch motivierter Sprachmodelle.

Das diesem Bericht zugrundeliegende Vorhaben wurde mit Mitteln des Bundesministers für Forschung und Technologie, des Esprit-Projekts P 2218 (SUNDIAL) und der Deutschen Forschungsgemeinschaft gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt bei den Autoren.

Literatur

[1] F. Andry, P. Baggia, G. Bakenecker, F. Charpentier, A. Cozannet, G. Niedermair, S. Thornton C.Rullent, and H. Tropf. *Linguistic Knowledge Bases and Software Realisation Specification for the Linguistic Processing Component.* Technical Report, Esprit P 2218 SUNDIAL WP5 Report, 1991.

[2] Ch. Fillmore. A case for case. In E. Bach and R. T. Harms, editors, *Universals in Linguistic Theory*, pages 1–88, Holt, Rinehart and Winston, New York, 1968.

[3] T. Kuhn, E.G. Schukat-Talamazzini, and H. Niemann. Context-dependent modeling in a two-stage hmm word recognizer for continuous speech. *European Signal Processing Conference (to appear)*, 1992.

[4] F. Kummert. *Flexible Steuerung eines sprachverstehenden Systems mit homogener Wissensbasis.* PhD thesis, Lehrstuhl für Informatik 5 (Mustererkennung), Universität Erlangen-Nürnberg, 1991.

[5] H. Niemann, A. Brietzmann, R. Muehlfeld, P. Regel, and G. Schukat. The speech understanding and dialog system evar. In De Mori, Suen R., and C. Y., editors, *New Systems and Architectures for Automatic Speech Recognition and Synthesis*, pages 271–302, Springer, NATO ASI Series, Berlin, 1985.

[6] H. Niemann, G. Sagerer, U. Ehrlich, E.G. Schukat-Talamazzini, and F. Kummert. The interaction of word recognition and linguistic processing in speech understanding. In P. Laface and R. De Mori, editors, *Speech Recognition and Understanding. Recent Advances, Trends, and Applications*, pages 425–453, Springer, 1992.

[7] H. Niemann, G. Sagerer, S. Schröder, and F. Kummert. Ernest: a semantic network system for pattern analysis. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 9:883–905, 1990.

[8] William Ogden and Ann Sorknes. What do users say to their natural language interface? In H.J. Bullinger and B. Schedul, editors, *Human-Computer interaction - INTERACT-87*, pages 561–566, North Holland, Amsterdam, 1987.

[9] L.R. Rabiner. Mathematical foundations of hidden markov models. In H. Niemann, M. Lang, and G. Sagerer, editors, *Recent Advances in Speech Understanding and Dialog Systems*, pages 183–205, Springer, 1988.

[10] E. G. Schukat-Talamazzini. *Generierung von Worthypothesen in kontinuierlicher Sprache.* Volume 141 of *Informatik-Fachberichte*, Springer-Verlag, Berlin, 1987.

[11] E.G. Schukat-Talamazzini, H. Niemann, W. Eckert, T. Kuhn, and S. Rieck. Acoustic modelling of subword units in the isadora speech recognizer. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing*, pages 577–580, San Francisco, 1992.