Verstehen natürlicher Sprache für den Mensch-Maschine-Dialog

Johannes Müller, Manfred Lang
Lehrstuhl für Mensch-Maschine-Kommunikation
Technische Universität München

Zusammenfassung

In diesem Beitrag wird ein System zum automatischen Verstehen sowie zum automatischen Übersetzen natürlicher, gesprochener Sprache beschrieben. Das Kernstück ist hierbei ein semantischer Decoder, welcher aus einem vorverarbeiteten Sprachsignal mit einer stochastischen maximum-a-posteriori Klassifikation unmittelbar eine semantische Darstellung bestimmt. Dabei wird akustisches, phonetisches, syntaktisches und semantisches Wissen gleichermaßen herangezogen. Ein separater Spracherkenner (Umwandlung des Sprachsignals in eine Wortkette) ist aufgrund der Integration des bewährten Viterbi-Algorithmus` und eines probabilistischen Chart-Parsers nicht notwendig! Des weiteren wird die semantische Gliederung als Repräsentation der Bedeutung einer Äußerung dargestellt. Sie kann sowohl innerhalb eines sprachverstehenden Systems als Zwischenebene für einen nachfolgenden Intentionsdecoder als auch innerhalb eines sprachübersetzenden Systems als Interlingua-Ebene für eine nachfolgende Sprachproduktion eingesetzt werden. Mit den entwickelten Algorithmen wurden bisher sprachverstehende und sprachübersetzende Schnittstellen für die Domänen „Grafikeditor“, „Serviceroboter“, „medizinische Bildverarbeitung“ und „Terminvereinbarung“ implementiert.

1. Einleitung

Spracherkennung, *Sprachverstehen* und *Sprachübersetzung* sind drei Disziplinen der automatischen Sprachverarbeitung, welche eine benutzeradäquate, intuitive und multimodale Mensch-Maschine-Kommunikation sehr unterstützen. Natürliche Sprache ist das meistverwendete Kommunikationsmittel zwischen den Menschen. Sie wird in der Regel von jedem beherrscht, funktioniert ohne den Einsatz von Händen oder

Füßen und ist ohne visuelle Rückkopplung und damit auch im Dunklen möglich. Dabei muß der sprachverstehende Rechner nicht in die Lage versetzt werden, natürliche Sprache wie ein Mensch kognitiv zu erfassen und damit Emotionen oder Assoziationen zu wecken, sondern soll lediglich Sprache als Informationsquelle zur Erledigung vorgegebener Aufgaben verwenden. In diesem Sinne seien drei Anwendungsgebiete exemplarisch dargestellt, die mit den oben genannten Disziplinen ausgeführt werden können.

Spracherkennung ist die alleinige Umwandlung von gesprochener Sprache in geschriebenen Text, das kann ein einzelnes Wort oder eine aus mehreren Wörtern bestehende Wortkette sein. Ein mögliches Anwendungsbeispiel ist die „hörende Schreibmaschine“, bei welcher der Computer ein gesprochenes Diktat als graphemische Wortkette *W* niederschreiben, jedoch nicht interpretieren muß. Hierbei ist der Bedeutungsinhalt des gesprochenen oder geschriebenen Textes für die Reaktion des Programms völlig irrelevant. Ein Einzelworterkenner kann nur einzelne oder getrennt gesprochene Wörter erkennen. Beim Diktieren eines zusammenhängenden Textes müssen dabei zwischen den einzelnen Wörtern kurze Pausen gemacht werden müssen. Anspruchsvoller sind Systeme zur Erkennung fließend gesprochener Sprache, wobei zwischen den Wörtern keine Pausen notwendig sind, jedoch u.a. das Problem der Koartikulation (phonetische Verschmelzung benachbarter Worte bei der Aussprache) zu bewältigen ist. *Sprecherabhängige Systeme* sind nur auf die Besonderheiten eines bestimmten Sprechers angepaßt, während *sprecherunabhängige Systeme* Eingaben von beliebigen Sprechern verarbeiten.

Sprachverstehen bedeutet die Interpretation von gesprochener bzw. geschriebener Sprache, d.h. die Extraktion der zugrundeliegenden Benutzerintention. Im Falle der sprachlichen Interaktion mit einer Applikation wäre die zu lösende Aufgabe die Umwandlung der sprachlichen Eingabe in computerverständliche Anweisungen, die das vom Benutzer sprachlich angeordnete Kommando ausführen. Diese

Umwandlung kann dabei unmittelbar oder indirekt über eine oder mehrere Zwischenebenen (z.B. Wortebene, Semantikebene) ablaufen.

Sprachübersetzung ist als automatische Übersetzung einer natürlichen Sprache in eine andere die Umsetzung von gesprochener oder geschriebener Quellsprache in eine geschriebene Zielsprache mit nachfolgender Sprachsynthese. Analog zum Sprachverstehen kann dies unmittelbar oder indirekt über eine oder mehrere Zwischenebenen (z.B. Wortebene, Interlingua-Ebene) erfolgen.

2. Semantische Gliederung

Zum Einsatz innerhalb einer sprachverstehenden Applikation wird eine formale Repräsentation des Bedeutungsinhaltes einer natürlichsprachlichen gesprochenen oder geschriebenen Äußerung benötigt, die

- **wortnah** genug ist, um einen direkten und probabilistischen Bezug zur Wortkette zu ermöglichen,
- **hierarchisch aufgebaut** ist, um eine Struktur nach festen (rekursiven) Regeln zu ermöglichen und um bestehende Abhängigkeiten probabilistisch zu erfassen,
- **formal-logisch korrekt** ist, im Sinne, daß der Bedeutungsinhalt konsistent und logisch nachvollziehbar repräsentiert wird,
- **generalisierend** ist, daß semantisch äquivalente, aber unterschiedliche Wortketten identisch repräsentiert werden und
- **maschinennah** genug ist, um diese formale Repräsentation mit möglichst einfachen Mechanismen in maschineninterpretierbare Kommandos umzuwandeln.

2.1 Definition der semantischen Gliederung

Aufgrund der aufgezählten Anforderungen wird die *semantische Gliederung S* als semantische Repräsentation einer gesprochenen Äußerung eingeführt. Sie kann als eine baumartige und damit hierarchische Struktur aufgefaßt werden und besteht aus *N* kleineren bedeutungs-

tragenden Einheiten, welche im folgenden *semantische Untereinheiten* (oder kurz *Semune*) s_n genannt werden:

$$S = \{s_1 , s_2, ..., s_n, ..., s_N\} \quad (1)$$

Jedes Semun s_n wird mit *(X+2)* Komponenten durch
- seinen Typ $t[s_n]$,
- seinen Wert $v[s_n]$ und
- Verweise auf $X \geq 1$ Nachfolger $q_1[s_n] , ..., q_X[s_n] \in \{s_{n+1} , ..., s_N , \text{leer}\}$

beschrieben:

$$\boxed{s_n = (t [s_n], v [s_n], q_1 [s_n] , ..., q_X [s_n])} \quad (2)$$

- Der **Typ** $t[s_n]$ gibt die Anzahl X der Nachfolger fest vor, schränkt die Menge möglicher Typen der Nachfolger ein und trifft eine sinnvolle Auswahl möglicher ihm zuzuordnender Werte $v[s_n]$.
- Der **Wert** $v[s_n]$ gibt in der Regel die eigentliche Bedeutung des Semuns s_n an.
- Jeder **Nachfolger** $q_x[s_n]$ mit $1 \leq x \leq X$ spezifiziert einen bestimmten Sachverhalt des Semuns s_n.
  - Ist eine solche Spezifizierung in der Äußerung vorhanden, ist dieser Nachfolger mit einem weiteren Semun innerhalb der semantischen Gliederung S identisch und wird als *Nachfolger-Semun* $q_X[s_n] \in \{s_{n+1} ,..., s_N\}$ bezeichnet.
  - Fehlt in der Äußerung die diesbezügliche Spezifizierung, wird auf einen *leeren Nachfolger* $q_X[s_n]$ = leer verwiesen.

Abb. 1 veranschaulicht eine semantische Gliederung S_1 der ‚Grafikeditor'-Domäne. Der Verweis auf ein Nachfolger-Semun wird durch die Kante „ ⟶ " markiert. Im Gegensatz dazu kennzeichnet die Kante „ ⟙ " den Verweis auf einen leeren Nachfolger.

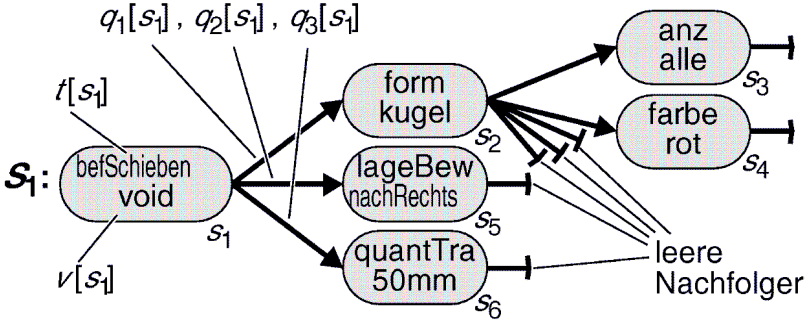

Abb. 1: Semantische Gliederung S_1 zur Wortkette „schiebe alle rote kugeln fünf zentimeter nach rechts“

Alle Semune s_2, ..., s_N besitzen genau ein Vorgänger-Semun. Damit bildet die gesamte semantische Gliederung S einen ‚Baum‘ mit dem Semun s_1 als ‚Wurzel‘ und den leeren Nachfolgern als ‚Blätter‘. Ein ‚Ast‘ $S(s_n)$ bezeichnet ein Semun s_n mit (rekursiv betrachtet) allen seinen Nachfolgern bis hin zu den Blättern:

$$S(s_n) \quad = \{s_n , s_{n+1}, s_{n+2},...\} \subseteq S \tag{3}$$

2.2 Bezug zur korrespondierenden Wortkette

Die semantische Gliederung ist eine wortnahe Darstellung des Bedeutungsinhaltes einer Äußerung. Trotzdem besitzt sie die Fähigkeit zur Generalisierung: Unterschiedliche, jedoch bedeutungsgleiche Wortketten korrespondieren zu identischen semantischen Gliederungen. Dazu werden folgende Festlegungen getroffen:

- Jedes Wort der Wortkette W korrespondiert zu genau einem Semun s_n aus der semantischen Gliederung S.
- Jedem Semun s_n der semantischen Gliederung S wird genau ein bedeutungstragendes Wort $w^+[s_n]$ und maximal ein bedeutungsloses Wort $w^-[s_n]$ aus der Wortkette W zugeordnet.

- Jeder Ast korrespondiert zu einer zusammenhängenden Teilwortkette .

Als Beispiel sei in Abb. 2 der Bezug zwischen der Wortkette „bitte den schönen grünen quader neben der ähm roten kugel löschen“ und deren semantischen Gliederung dargestellt. Auf das bedeutungstragende Wort eines Semuns sei mit dem dunklen Pfeil „ ⟶ “ hingewiesen, auf das optionale bedeutungslose Wort („bitte“, „schönen“, „ähm“) mit dem hellen Pfeil „ ⟶ “. Die zu dem jeweiligen Ast $S(s_n)$ gehörende, lückenlose Teilwortkette $W(s_n)$ wird darunter aufgezeigt.

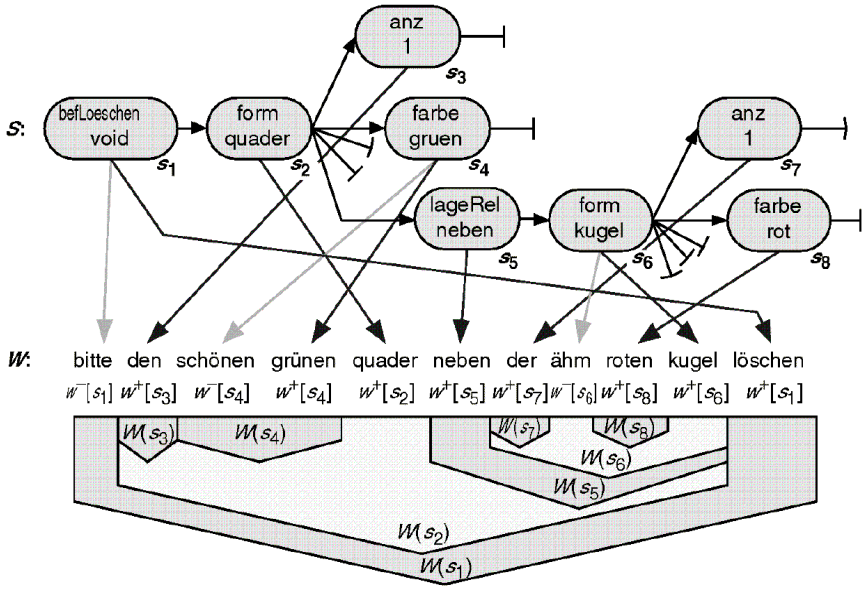

Abb. 2: Semantische Gliederung S_2 und Teilwortketten $W(s_n)$

3. Signal-Vorverarbeitung

Wie ein Spracherkenner verarbeitet unser semantischer Decoder akustische Merkmale, die aus dem Sprachsignal gewonnen werden. Sie sollen möglichst viel Information über den Inhalt der Äußerung enthalten, jedoch möglichst wenig von der Person des Sprechers, vom Eingabekanal und von Hintergrundgeräuschen beeinflußt sein. In der

Forschung wurden verschiedene Methoden entwickelt, um diese Merkmale aus dem Sprachsignal durch Analyse im Zeit- und Frequenzbereich zu gewinnen. In Abb. 3 ist ein Verfahren zur Signalvorverarbeitung dargestellt, das sich in vielen Spracherkennern bewährt hat und auch im hier beschriebenen sprachverstehenden System eingesetzt wird.

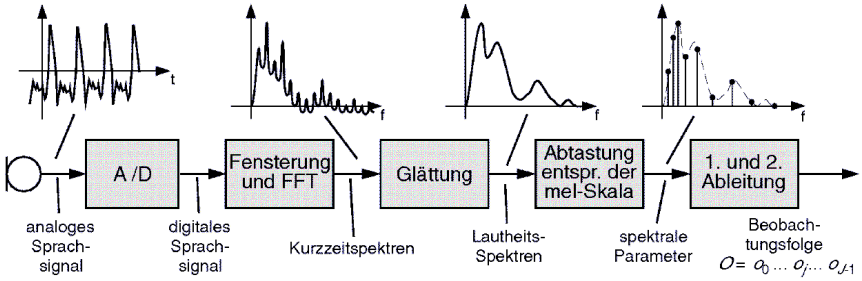

Abb. 3: Blockdiagramm der Signal-Vorverarbeitung [7]

Das analoge Sprachsignal wird digitalisiert, einer Zeitfensterung der Dauer 10 bis 30 ms und anschließend einer Fast-Fourier-Transformation (FFT) unterzogen. Diese Kurzzeitspektralanalyse wird typischerweise alle 10 ms ausgeführt, um Instationaritäten im Sprachsignal hinreichend genau zu erfassen. Die Feinstruktur des Spektrums enthält Information über den Sprecher, trägt jedoch zur weiteren Decodierung kaum bei. Deshalb werden die logarithmierten Kurzzeitspektren geglättet, um z.B. den Einfluß der Grundfrequenz des Sprechers zu eliminieren. Das geglättete Spektrum wird abgetastet, um die Anzahl der Merkmale auf etwa 20 bis 50 zu reduzieren. Man berücksichtigt die Frequenzauflösung des menschlichen Gehörs, indem die Stützstellen dieser Abtastung nicht äquidistant, sondern entsprechend der Mel-Skala [10] gewählt werden. Eine abschließende auf der (nicht)-linearen Diskriminanzanalyse (LDA bzw. NLDA) beruhende Koordinatentransformation bewirkt, daß die somit erhaltenen Merkmale eine maximale Unterscheidbarkeit bezüglich ihrer Klassifizierung besitzen.

Die Signal-Vorverarbeitung erzeugt somit alle 10 ms einen mehrdimensionalen Merkmalsvektor o_j, der den zugehörigen Signalabschnitt beschreibt, und komprimiert gleichzeitig die Daten um einen Faktor 10 bis 20. Die zeitliche Aneinanderreihung der Vektoren o_j ergibt die Beobachtungsfolge $O = o_0 \ldots o_j \ldots o_{J-1}$ [5].

4. Semantische Decodierung

Als semantische Decodierung wird im vorliegenden Beitrag die Umwandlung einer Beobachtungsfolge in eine formale syntaktisch-semantische Darstellung verstanden.

Der derzeit meistverfolgte Ansatz zur semantischen Decodierung ist eine Aufteilung dieses Prozesses in zwei voneinander getrennte Module. Ein *Spracherkenner* liefert eine bestimmte Anzahl der besten Wortketten-Hypothesen oder eine Wort-Lattice, d.h. ein Netzwerk aus Worthypothesen mit dazwischenliegenden Übergängen. Eine nachfolgende *semantische Textanalyse* analysiert dies und generiert daraus eine korrespondierende syntaktisch-semantische Darstellung. Diese klassische Aufteilung hat ihren Ursprung durch die klare Trennung der ingenieurnahen Disziplin Signalverarbeitung und der geisteswissenschaftsnahen Disziplin Linguistik, welche mit der eindeutig definierten Schnittstelle der Wortebene miteinander verbunden sind.

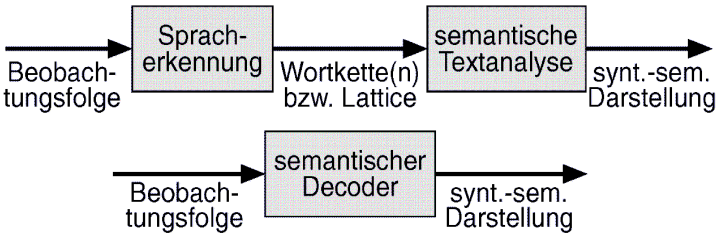

Abb. 4: Zwei- bzw. einstufiger Ansatz zur semantischen Decodierung von Sprache

Ansätze, welche die semantische Decodierung innerhalb einer Stufe vollbringen, werden in der aktuellen Forschung erstaunlicherweise sehr selten verfolgt. Dies hängt nicht zuletzt damit zusammen, daß innerhalb eines Algorithmus‘ Probleme der Signalverarbeitung, Spektralanalyse, Phonetik, Linguistik, Wissensrepräsentation und der Wissensverarbeitung bewältigt werden müssen. Ein in sich geschlossener, einstufiger Ansatz vermeidet allerdings Inkonsistenzen zwischen mehreren Modulen. Es können mehrere Wissenbasen, welche die Zusammenhänge zwischen den jeweiligen Repräsentationsebenen enthalten, im Sinne einer gesamten, integrierten Wissensbasis mit allen zur Verfügung stehenden Parametern zusammenwirken. Eventuell auftretende Fehler des ersten Moduls können sich nicht mehr auf ein nachfolgendes, davon losgelöstes Modul auswirken.

Im Rahmen zweier Dissertationen [6][7] konnte am Lehrstuhl für Mensch-Maschine-Kommunikation eine einstufige semantische Decodierung entwickelt und in ein sprachverstehendes bzw. übersetzendes System implementiert werden (Abb. 5):

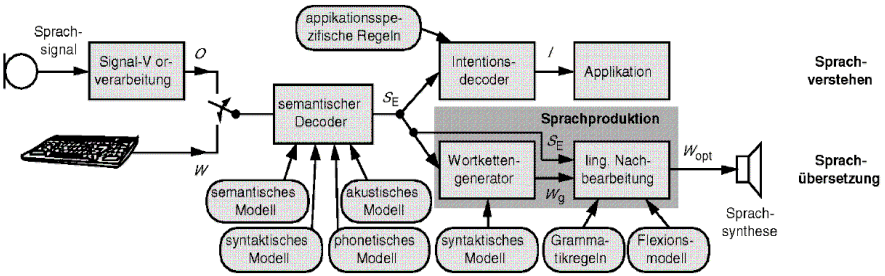

Abb. 5: Systemarchitektur: Der einstufige semantische Decoder arbeitet mit rein stochastischem und automatisch trainiertem Wissen. Die semantische Gliederung S_E dient als Zwischenebene zum Sprachverstehen und als Interlingua-Ebene zur Sprachübersetzung.

4.1 Maximum-a-posteriori Klassifikation

Der semantische Decoder ermittelt aus einer vorliegenden Beobachtungsfolge *O* diejenige erkannte semantische Gliederung S_E, welche unter Vorliegen genau dieser Beobachtungsfolge *O* und den zur Verfügung stehenden Wissensbasen die höchste Wahrscheinlichkeit *P(S|O)* besitzt. Diese Wahrscheinlichkeit wird mit der Bayes'schen Regel umgeformt.

$$S_E = \arg\max_S P(S \mid O) = \arg\max_S \frac{P(O \mid S) \cdot P(S)}{P(O)} \tag{4}$$

Die a-priori-Wahrscheinlichkeit braucht für die Maximierung nicht berücksichtigt zu werden, da sie bei gegebener Beobachtungsfolge *O* konstant ist.

$$S_E = \arg\max_S [P(O \mid S) \cdot P(S)] \tag{5}$$

Die direkte Bestimmung von ist aufgrund der Vielfalt möglicher Kombinationen aus *O* und *S* nicht ohne weiteres möglich. Deshalb werden die Wortebene *W* und die Phonemebene *Ph* als weitere Repräsentationsebenen eingeführt.

$$\begin{aligned} S_E &= \arg\max_S \max_W \max_{Ph} [P(O \mid Ph) P(Ph \mid W) P(W \mid S) P(S)] \\ &= \arg\max_S \max_W \max_{Ph} P(O, Ph, W, S) \end{aligned} \tag{6}$$

Gl. (6) beschreibt die maximum a-posteriori Klassifikation zur Extraktion derjenigen semantischen Gliederung S_E, welche in der wahrscheinlichsten Kombination aus einer semantischen Gliederung, einer Wortkette, einer Phonemkette und der vorliegenden Beobachtungsfolge enthalten ist.

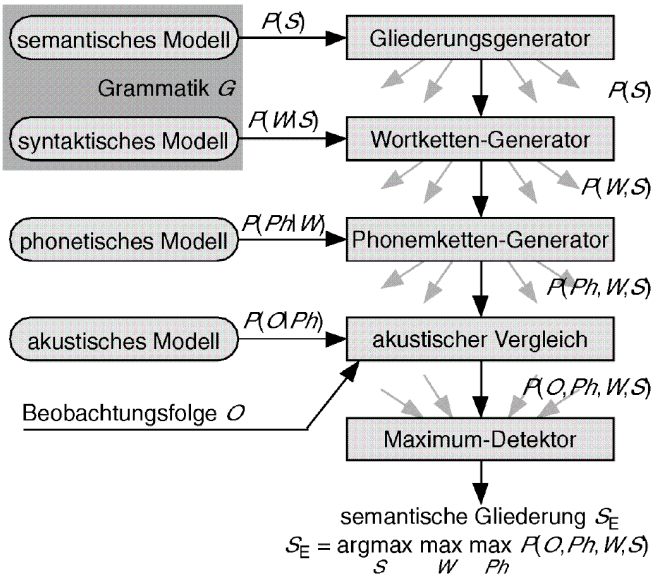

Abb. 6: Prinzip der 'top-down'-semantischen Decodierung

Dieser Ansatz benötigt vier stochastische Wissensbasen (Modelle), die gleichzeitig und gleichermaßen zur semantischen Decodierung herangezogen werden:

- Das **semantische Modell** liefert die a-priori-Wahrscheinlichkeit $P(S)$ für das Auftreten einer semantischen Gliederung S.
- Das **syntaktische Modell** liefert die bedingte Wahrscheinlichkeit $P(W|S)$ für das Auftreten einer Wortkette W gegeben eine semantische Gliederung S.
- Das **phonetische Modell** liefert die bedingte Wahrscheinlichkeit $P(Ph|W)$ für das Auftreten einer Phonemkette Ph gegeben eine Wortkette W. (Diese Wissenbasis ist prinzipiell mit derjenigen eines Spracherkenners identisch.)
- Das **akustische Modell** liefert die bedingte Wahrscheinlichkeit $P(O|Ph)$ für das Auftreten einer Beobachtungsfolge O gegeben eine Phonemkette Ph. (Diese Wissenbasis kann unverändert von einem Spracherkenner übernommen werden.)

Im Falle von textueller Eingabe sind phonetische und akustische Modelle nicht notwendig. Somit vereinfacht sich Gl. (6) zu

$$S_E = \arg\max_S [P(W \mid S) \cdot P(S)] \qquad (7)$$

Sämtliche Wahrscheinlichkeiten werden im sogenannten *Training* abgeschätzt (d.h. *trainiert*), indem auftretende Häufigkeiten in einer möglichst großen Trainingsstichprobe ermittelt werden. Das Trainingsmaterial besteht aus domänenspezifischen Äußerungen, von denen jede durch semantische Gliederung, Wortkette, Phonemkette und Beobachtungsfolge repräsentiert ist. Eine detaillierte Beschreibung des Trainings und der Integration aller stochastischer Wissensbasen sowie des semantischen Decoders, der eine Kombination eines modifizierten Earley-Parsers [1] (auf der Syntax- und Semantik-Ebene) mit einem Viterbi-Suchalgorithmus [8] (auf der Akustik-Ebene) darstellt, kann der Dissertation von H. Stahl [7] entnommen werden.

4.2 Die stochastische Grammatik

4.2.1 Wahrscheinlichkeiten im semantischen Modell

Unter der Annahme statistischer Unabhängigkeit läßt sich *P(S)* als Produkt von bedingten Wahrscheinlichkeiten erster Ordnung berechnen.

$$P(S) = f_0 \cdot \prod_{n=1}^{n} (e_n \cdot f_n) \qquad \text{mit:} \qquad (8)$$

- Die **Wurzelwahrscheinlichkeit** $f_0 = P(t[s_1])$ bezeichnet die a-priori-Wahrscheinlichkeit, daß das Wurzel-Semun s_1 den Typ $t[s_1]$ besitzt.
- Die **Wertwahrscheinlichkeit** $e_n = P(v[s_n] \mid t[s_n])$ bezeichnet die bedingte Wahrscheinlichkeit, daß in einem Semun s_n des Typs $t[s_n]$ der Wert $v[s_n]$ auftritt.

- Die **Folgewahrscheinlichkeit** $f_n = P(t[q_1[s_n]],..., t[q_X[s_n]] \mid t[s_n])$ steht für die bedingte Wahrscheinlichkeit, daß zu einem Semun s_n vom Typ $t[s_n]$ die Nachfolgersemune mit den Typen $t[q_1[s_n]]$, ..., $t[q_X[s_n]]$ auftreten:

4.2.2 Wahrscheinlichkeiten im syntaktischen Modell

Die semantische Gliederung setzt sich baumförmig und damit hierarchisch aus N kleineren Einheiten, den semantischen Untereinheiten (abgekürzt Semunen) s_n mit $1 \le n \le N$ zusammen. Analog dazu setzt sich das zugehörige syntaktische Netzwerk ebenfalls hierarchisch aus N kleineren Einheiten, den sogenannten *syntaktischen Modulen* (abgekürzt SMen) $A(s_n)$ mit $1 \le n \le N$ zusammen. Jedes Semun s_n steuert dem syntaktischen Modell ein ihm zugeordnetes SM $A(s_n)$ bei.

Jedes SM $A(s_n)$ steuert die Wortstellung und die Wortwahl der dem Semun s_n zugeordneten Worte. Ein SM ist ein probabilistischer Zustandsautomat, der aus $(X+4)$ Zuständen bzw. Knoten $beg(s_n)$, $A(q_1[s_n])$, ..., $A(q_X[s_n])$, $B(s_n)$, $C(s_n)$ und $end(s_n)$ besteht. Dabei markiert X die vom Semun-Typ $t[s_n]$ fest vorgegebene Nachfolger-Anzahl des entsprechenden Semuns s_n.

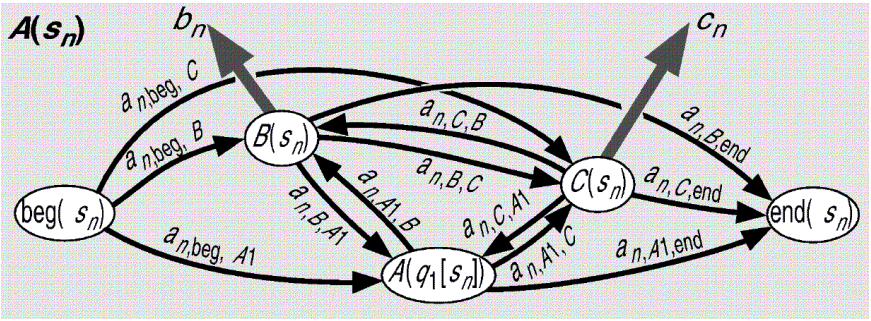

Abb. 7: Syntaktisches Modul $A(s_n)$ für ein Semun s_n mit $X=1$

- Der **Knoten *beg*(s_n)** wird vom Vorgänger-SM angesprungen. Alle weiteren Zustände des SMs können nur nach *beg*(s_n) betreten werden. Bei $n = 1$ bildet *beg*(s_1) den Startpunkt des gesamten syntaktischen Netzwerks.
- Jeder **Zustand *A*(q_x[s_n])** mit $1 \leq x \leq X$ muß unterscheiden:
  - Falls $q_x[s_n]$ ein Nachfolger-Semun ist, steht dieser Zustand stellvertretend für das SM *A*(q_x[s_n]) des entsprechenden Nachfolger-Semuns $q_x[s_n]$.
  - Falls $q_x[s_n]$ ein leerer Nachfolger ist, wird *A*(q_x[s_n]) unmittelbar nach Betreten wieder verlassen.
- Der **Zustand *B*(s_n)** emittiert ein bedeutungstragendes Wort $w^+[s_n]$ mit der Emissionswahrscheinlichkeit b_n.
- Der **Zustand *C*(s_n)** emittiert ein bedeutungsloses Wort $w^-[s_n]$ mit der Emissionswahrscheinlichkeit c_n.
- Der **Knoten *end*(s_n)** springt zu demjenigen Zustand des Vorgänger-SMs, aus dem das aktuelle SM angesprungen wurde. Bei $n = 1$ bildet *end*(s_1) den Endpunkt des gesamten syntaktischen Netzwerks.

Der Übergang zwischen zwei aufeinanderfolgenden Zuständen wird primär durch die jeweiligen Übergangswahrscheinlichkeiten bestimmt, z.B. markiert die Übergangswahrscheinlichkeit $a_{n,B,C}$ die Wahrscheinlichkeit für einen Übergang von *B*(s_n) nach *C*(s_n). Um die in Kap. 2.2 getroffenen Festlegungen einzuhalten, müssen zusätzlich auf dem Pfad von *beg*(s_n) nach *end*(s_n) die folgenden Bedingungen zwingend eingehalten werden:

- Keiner der Zustände *A*(q_1[s_n]), ..., *A*(q_X[s_n]), *B*(s_n) oder *C*(s_n) kann jeweils zweimal betreten werden.
- Der Knoten *end*(s_n) kann nicht vor den Zuständen *A*(q_1[s_n]), ..., *A*(q_X[s_n]) und *B*(s_n) betreten werden.

Dadurch wird sichergestellt, daß jeder der Zustände *A*(q_1[s_n]), ..., *A*(q_X[s_n]) und *B*(s_n) genau einmal und der Zustand *C*(s_n) maximal einmal betreten wird, was bedeutet, daß jeder Nachfolger genau einmal

angesprungen wird sowie genau ein bedeutungstragendes Wort und maximal ein bedeutungsloses Wort emittiert werden.

Zur Berechnung der Wahrscheinlichkeit $P(W|S)$ werden in guter Näherung die folgenden bedingten Wahrscheinlichkeiten erster Ordnung herangezogen:

- Die **Pfadwahrscheinlichkeit a_n** ist die bedingte Wahrscheinlichkeit, daß in einem zu einem Semun des Typs $t[s_n]$ gehörenden SM genau der vorliegende Pfad durchlaufen wird. Sie berechnet sich aus dem Produkt aller Übergangswahrscheinlichkeiten $a_{n,\mu,\nu}$ entlang des Pfades von $beg(s_n)$ nach $end(s_n)$ durch das SM.

$$a_n = P\,(\text{bestimmter Pfad } \underline{\text{mit max. } P(W|S)}|t[s_n]) \quad = \prod_{\substack{\text{alle Übergänge}\\ \text{von } \mu \text{ nach } \nu \text{ ent-}\\ \text{lang Pfad durch } A(s_n)}} a_{n,\mu,\nu} \qquad (9)$$

- Die **Emissionswahrscheinlichkeit b_n** ist die bedingte Wahrscheinlichkeit, daß der Zustand $B(s_n)$ des SMs, welches zu einem Semun des Typs $t[s_n]$ und des Werts $v[s_n]$ gehört, das bedeutungstragende Wort $w^+[s_n]$ emittiert:

$$b_n = P\,(\text{Emission von } w^+[s_n] \,|\, t[s_n],\ v[s_n]) \qquad (10)$$

- Die **Emissionswahrscheinlichkeit c_n** ist die bedingte Wahrscheinlichkeit, daß der Zustand $C(s_n)$ des zu einem Semun des Typs $t[s_n]$ gehörenden SMs das bedeutungslose Wort $w^-[s_n]$ emittiert.

$$c_n = \begin{cases} P(Emiss.von\, w^-[s_n] \,|\, t[s_n]) & , \text{falls } C(s_n) \\ 1 & , \text{sonst} \end{cases} \qquad (11)$$

Unter der Annahme der statistischen Unabhängigkeit ergibt sich die bedingte Wahrscheinlichkeit $P(W|S)$ als Produkt der obigen Wahrscheinlichkeiten von allen N SMen:

$$P(W|S) = \prod_{n=1}^{N} (a_n \cdot b_n \cdot c_n) \qquad (12)$$

5. Intentionsdecodierung

Da die semantische Gliederung die Umgebungskonstellation sowie den Dialogstatus nicht miteinbezieht, reicht die bereitgestellte Information unter Umständen nicht zur unmittelbaren Steuerung einer laufenden Applikation aus. Vielmehr muß eine nachfolgende Instanz jenes Wissen liefern, welches in der semantischen Gliederung nicht zwingend vorhanden ist. Als triviales Beispiel seien die Wortketten *„mache ihn größer“* bzw. *„größer machen“* betrachtet. Das Wissen, auf welches Objekt sich die Anweisung bezieht, ist in der semantischen Gliederung nicht vorhanden. Weiterhin kann es bei der Äußerung *„lösche den roten quader“* zu Problemen kommen, falls überhaupt kein Quader existiert oder falls mehrere Quader in der Grafik vorhanden sind. Im letzteren Fall müßte der Quader mit weiteren Eigenschaften (z.B. Farbe, Größe usw.) näher spezifiziert werden, ansonsten liegt eine Fehleingabe vor. Der Intentionsdecoder verbindet eine semantische Gliederung mit der aktuellen Dialog- und Umgebungskonstellation und erzeugt daraus eine lineare Abfolge von maschineninterpretierbaren Befehlen, die als Intention *I* bezeichnet werden.

Für die Umsetzung einer semantischen Gliederung in die entsprechende Intention wird eine Kombination aus Präprozessor und Compiler vorgeschlagen. Die mittels eines *Präprozessors* aufbereitete semantische Gliederung wird als Quellsprache für den *Compiler* aufgefaßt. Ziel der Compilierung ist ein maschineninterpretierbarer Code, wie z.B. eine Sequenz von UNIX-, SQL- oder speziellen applikationsspezifischen Kommandos.

Um die baumförmige semantische Gliederung zu verarbeiten, muß sie also in eine lineare Sequenz von Befehlen umgeformt werden, die in der nachfolgenden Applikation abgearbeitet wird. Hierbei muß von den Standardansätzen des Compilerbaus (Abarbeitung in ‚top-down‘- bzw. in ‚bottom-up‘-Richtung) abgewichen werden. Denn um den aktuellen Zustand eines jeden Semuns zu bestimmen, muß der Baum ‚top-down‘

abgearbeitet werden. Die Elementarbefehle können aber nur ‚bottom-up' generiert werden, da sie in höheren Ebenen des Baumes nur ausgeführt werden können, wenn die Elementarbefehle der Nachfolgersemune ausgeführt worden sind, da diese nähere Informationen liefern. Deshalb werden beide Verfahren zu einem ‚top-down-bottom-up'-Ansatz, kurz ‚top-up'-Ansatz, kombiniert.

In Abb. 8 ist der Flußplan für die Abarbeitung eines Unterbaums abgebildet mit den wesentlichen Schritten „HINAB" für die ‚top-down'-Abarbeitung des Baumes, „HERAUF" für die ‚bottom-up'-Generierung

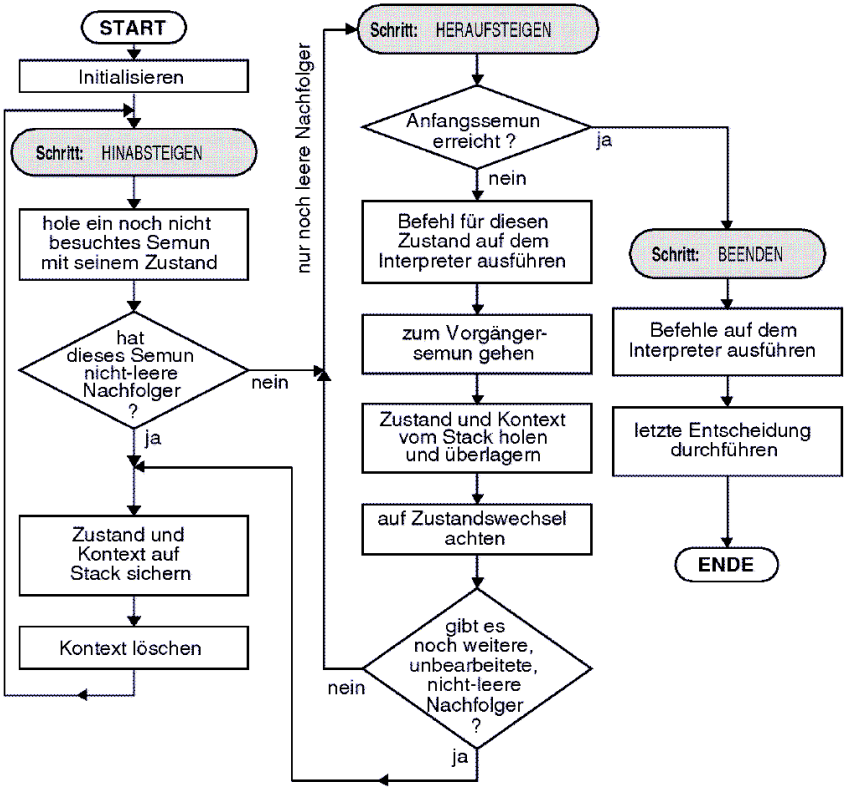

Abb. 8: Flußplan des Linearisierungsalgorithmus'

der Elementarbefehle und „BEENDEN“ für die ultimative Weitergabe an die ausführende Applikation. Eine ausführliche Beschreibung des Vorgehens wird in [2] und [6] aufgezeigt.

6. Sprachproduktion

6.1 Wortketten-Generator

Im Gegensatz zum Wortketten-Generator der semantischen Decodierung, welcher zu einer vorliegenden semantischen Gliederungshypothese viele Wortkettenhypothesen entstehen läßt, erzeugt der an dieser Stelle beschriebene Wortketten-Generator aus einer vorliegenden semantischen Gliederung S_E nur eine einzige korrespondierende Wortkette W_g:

$$W_g = \arg\max_W P(W|S_E) \tag{13}$$

Dabei spielt es prinzipiell keine Rolle, ob die generierte Wortkette natürlichsprachlich ist oder nicht. Der stochastische Prozeß der Wortkettengenerierung mit der in Kap. 4.2.2 beschriebenen Grammatik kann als komplexes Zustands-Übergangs-Netzwerk, ähnlich einem hierarchischen Hidden-Markov-Modell, aufgefaßt werden. Jedes Semun korrespondiert zu einem syntaktischen Modul (SM), welches entsprechend der vorliegenden semantischen Gliederung mit anderen SMen zu einem hierarchischen syntaktischen Netzwerk verbunden ist. Innerhalb dieses Netzwerks beeinflussen die Übergänge zwischen einzelnen Zuständen die Wortstellung, die Emissionen von Worten aus bestimmten Zuständen die Wortwahl.

6.2 Linguistische Nachbearbeitung

Da die semantische Gliederung keinerlei grammatikalische Information enthält, können Genus, Kasus bzw. Numerus mancher Worte aus der generierten Wortkette W_g falsch sein. Dennoch wird der semantische Inhalt dieser syntaktisch mitunter unkorrekten Wortkette von einem

menschlichen Informationsempfänger verstanden. So kann beispielsweise die Wortkette

W_g: *„dessine deux sphère rouge“*

vom einem mit einem französischsprachigen syntaktischen Modell ausgestatteten Wortketten-Generator anstatt der grammatikalisch korrekten Wortkette *„dessine deux sphères rouges“* ausgegeben werden. In diesem Fall werden anstatt der korrekten Pluralworte *„sphères“* und *„rouges“* fälschlicherweise die jeweiligen Singularformen emittiert. Somit muß eine nachfolgende Instanz die auftretenden grammatikalischen Inkonsistenzen beseitigen. Eine generierte Wortkette W_g wird also linguistisch nachbearbeitet, um somit eine grammatikalisch korrekte (d.h. „optimierte“) Wortkette W_{opt} zu erhalten.

Ein wesentlicher Vorteil des angewandten Verfahrens ist, daß zur Nachbearbeitung nicht nur die Wortkette W_g, sondern auch die semantische Gliederung S_E zur Verfügung steht. Ausgehend von deren Wurzel bis hin zu deren Blättern wird nun die semantische Gliederung S_E rekursiv abgearbeitet. Wenn mindestens eines der zu einem Semun s_n gehörenden Worte $w^+[s_n]$ bzw. $w^-[s_n]$ ein Substantiv, Adjektiv oder Artikel ist, muß die entsprechende Flexion dieses Wortes, welche aufgrund der jeweiligen grammatikalischen Eigenschaften (d.h. Genus, Kasus, Numerus) gewählt wird, anhand des vorliegenden Flexionsmodells gefunden und dem zu verbessernden Wort angehängt werden.

Die grammatikalischen Eigenschaften werden dabei mittels sprachspezifischer Grammatikregeln innerhalb der semantischen Gliederung S_E bestimmt und zunächst dem jeweiligen Semun s_n zugeordnet. Nur die Genusbestimmung wird von einem emittierten Substantiv beeinflußt, alle anderen Eigenschaften gründen ausschließlich auf der semantischen Gliederung. Dies ist natürlich im Vergleich zur klassischen Linguistik, welche syntaktische und semantische Bindungen und Beschreibungen nur aufgrund der Worte und Sätze vornimmt, ein vollständig verschiedener und neuartiger Ansatz!

7 Realisierte Domänen

Die strikte Trennung in domänenspezifisches Wissen und domänenunabhängige Algorithmen ermöglicht eine leichte Portierung auf andere Domänen. Bisher wurden semantische Gliederungen der in Kap. 2 beschriebenen Art für vier Anwendungsbereiche entwickelt:

- **NASGRA** (natürlichsprachlicher Grafikeditor) zum Erzeugen, Verändern und Löschen dreidimensionaler Objekte auf dem Bildschirm (Verstehen von Deutsch und Slowenisch sowie Übersetzung von Deutsch und Slowenisch in Deutsch, Englisch, Französisch und Slowenisch) [6]. Eine interaktive Demonstration mit textueller Ein- und Ausgabe kann über das WWW unter folgender Adresse aufgerufen werden:

  http://www.mmk.e-technik.tu-muenchen.de/~mue/nasgra/

- **ROMAN**[1] (roving manipulator), ein sprachverstehender Serviceroboter zur Ausführung mobiler Handhabungsaufgaben in Innenraumumgebungen (Verstehen von Deutsch) [3].

- **invitoVR**[2] (interactive visualization tool for virtual reality), sprachliche Interaktion innerhalb einer virtuellen Umgebung zur dreidimensionalen Visualisierung von anatomischen CT-Daten (Verstehen von Deutsch) [4]. Hierbei wurde insbesondere ein multimodaler Mensch-Maschine-Dialog implementiert, wobei Sprache, Handgesten und konventionelle Eingabegeräte gleichermaßen ausgewertet werden.

- **TERMINator** zur Übersetzung von gesprochenen Terminabsprache-Dialogen von Deutsch in Englisch bzw. Griechisch - die Domäne ähnelt prinzipiell derjenigen des Verbundprojektes VERBMOBIL [9].

---

[1] In Zusammenarbeit mit dem Lehrstuhl für Steuerungs- und Regelungstechnik, Technische Universität München.

[2] In Zusammenarbeit mit dem Institut für Medizinische Informatik und Systemforschung, Forschungszentrum für Umwelt und Gesundheit (GSF), Neuherberg.

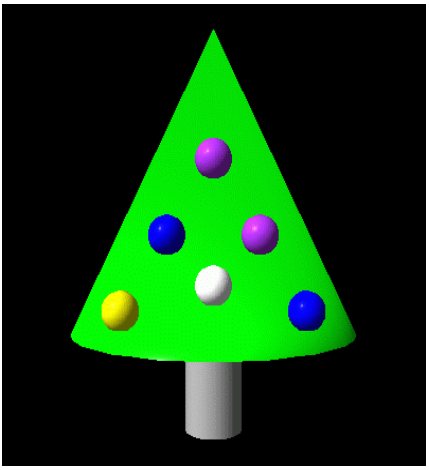

Abb. 9: Typische Szene aus der NASGRA-Domäne

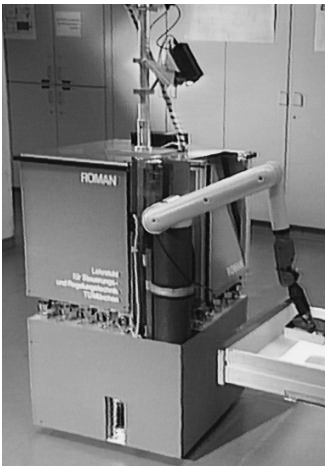

Abb. 10: Sprachverstehender Serviceroboter ROMAN

8. Ergebnisse

8.1 Semantische Decodierung

Abb. 11 zeigt den Gegensatz zwischen Suchaufwand und Trefferrate für unseren einstufigen semantischen Decoder. Ein beeinflussender Parameter ist dabei die Pruningschwelle, welche ein Maß für das Verwerfen von Hypothesen darstellt. Die beste Trefferrate von etwa 92 % wird bei

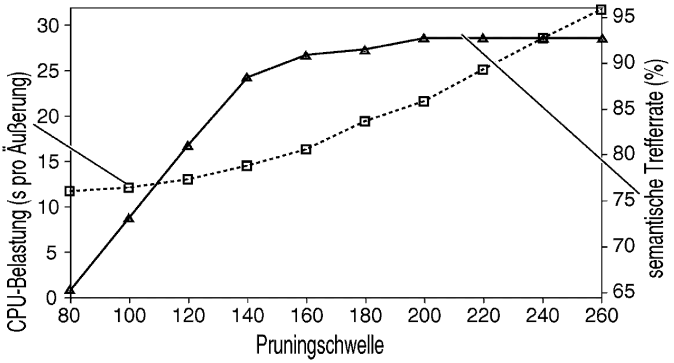

Abb. 11: Semantische Trefferrate und CPU-Belastung in Abhängigkeit von der Pruningschwelle [7]

einer Pruningschwelle von 200 erzielt, was pro Äußerung eine CPU-Belastung von durchschnittlich 22 s (auf einer SUN™-*UltraSPARC*™, 168 MHz) bedeutet.

8.2 Intentionsdecodierung

Der Intentionsdecoder wurde getestet mit 1843 semantischen Gliederungen von gesprochenen Äußerungen, die von 33 Versuchspersonen während einer Wizard-of-Oz-Simulation innerhalb der NASGRA-Domäne gesammelt wurden. Die Prozentrate für die korrekte Konvertierung einer semantischen Gliederung in die entsprechende Intention (d.h. applikationsspezifischer und maschineninterpretierbarer Code) beträgt 97,3% [2][6].

8.3 Sprachproduktion

Die Sprachproduktion wurde mit 307 existierenden semantischen Gliederungen innerhalb der NASGRA-Domäne getestet. Daraus wurden Wortketten in den vier Zielsprachen Deutsch, Englisch, Französisch und Slowenisch erzeugt. Wir baten verschiedene Versuchspersonen, die Semantik (verständlich oder unverständlich) und die Syntax (korrekt, unüblich, falsch) der optimierten Wortketten W_{opt} zu beurteilen.

Tab. 1: Ergebnisse der Sprachproduktion [6]

Zielsprache	Semantik	Syntax	
	verständlich	Korrekt	unüblich
Deutsch	95,9 %	84,4 %	5,2 %
Englisch	94,8 %	81,1 %	10,4 %
Französisch	93,8 %	82,4 %	9,8 %
Slowenisch	88,3 %	82,1 %	8,5 %

Diese Prozentraten mit einer durchschnittlich korrekten Semantik von 93,2 % und einer durchschnittlich nicht-falschen Syntax von 91,0 % bestätigen, daß der semantikbasierte Übersetzungsansatz eine Alternative zu wesentlich komplexeren Ansätzen darstellt, sofern kurze und grammatikalisch korrekte Sätze innerhalb einer umgrenzten Domäne übersetzt werden sollen.

Literatur

[1] J. Earley: An Efficient Context-Free Parsing Algorithm, Comm. of the ACM, vol. 13 (1970), no. 2, S. 94-102

[2] M. Ebersberger, J. Müller, H. Stahl: A Compiler-Interpreter-System for Decoding the User's Intention within a Speech Understanding Application, Tagungsband KI-96 (Dresden, 1996), Lecture Notes in Artificial Intelligence 1137, Springer, S. 61-65

[3] C. Fischer, P. Havel, G. Schmidt, J. Müller, H. Stahl, M. Lang: Kommandierung eines Serviceroboters mit natürlicher, gesprochener Sprache, in G. Schmidt, F. Freyberger (Hrsg.): Tagungsband „Autonome Mobile Systeme 1996“ (München), Springer „Informatik aktuell“, 1996, S. 248-261

[4] C. Krapichler, M. Haubner, A. Lösch, K. Englmeier: A Human-Machine Interface for Medical Image Analysis and Visualization in Virtual Environments, Tagungsband „1997 International Conference on Acoustics, Speech, and Signal Processing“ (München), S. 2613-2616

[5] M. Lang, H. Stahl: Spracherkennung für einen ergonomischen Mensch-Maschine-Dialog, Zeitschrift „mikroelektronik“, Heft 2/1994, S. 78-82

[6] J. Müller: Die semantische Gliederung zur Repräsentation des Bedeutungsinhalts innerhalb sprachverstehender Systeme, Dissertation, Herbert Utz Verlag Wissenschaft, München, 1997

[7] H. Stahl: Konsistente Integration stochastischer Wissensquellen zur semantischen Decodierung gesprochener Äußerungen, Dissertation, Herbert Utz Verlag Wissenschaft, München, 1997

[8] A.J. Viterbi: Error Bounds for Convolutional Codes and an Asymptotical Optimal Decoding Algorithm, IEEE Trans. Information Theory, vol. 61 (1973), S. 268-278

[9] W. Wahlster: Verbmobil - Translation of Face-to-Face Dialogs, Tagungsband Eurospeech 1993 (Berlin), Addenum „Opening and Plenary Sessions“, S. 29-38

[10] E. Zwicker: Psychoakustik, Springer-Verlag, Berlin, 1982