DIE SCHNITTSTELLE ZWISCHEN SPRACHSIGNALVERARBEITUNG UND SYMBOLISCHER SPRACHVERARBEITUNG

Gudrun Flach
Technische Universität Dresden

1. Einleitung

Bei der Erkennung sprachlicher Einheiten von mehr als einem Wort Länge und angenommenem inhaltlichen Zusammenhang werden in Spracherkennungssysteme meist zusätzliche Informationen über den Sprachaufbau eingefügt. Somit wird die Sprachsignalverarbeitung, die als Eingangsgröße physikalische Sprachsignalparameter und als Ausgangsgröße eine bzw. mehrere Symbolhypothesen hat, durch die symbolische Sprachverarbeitung, die aus den Symbolhypothesen eine Satz- bzw. Phraseninterpretation erzeugt, ergänzt. Diese beiden Grundverarbeitungsabschnitte werden auch als Bottom-End (Sprachsignalverarbeitung)- und Top-End (symbolische Sprachverarbeitung)- Verarbeitung bezeichnet (siehe Bild 1).

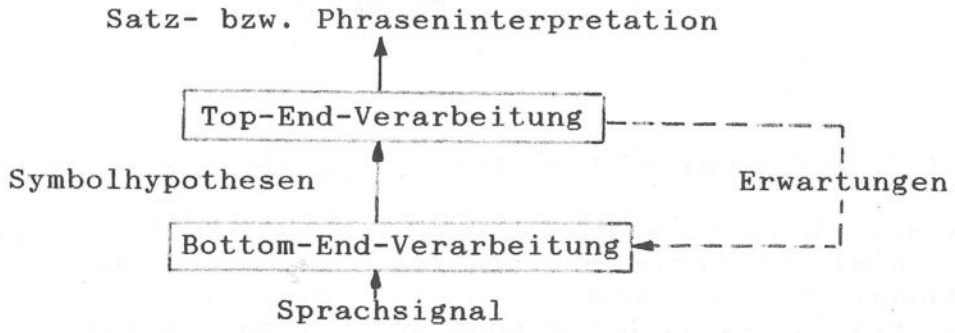

Bild 1: Grundaufbau eines Spracherkennungssystems

Bei der Gestaltung der Schnittstelle zwischen den beiden grundlegenden Verarbeitungsabschnitten sind zwei Problemkreise zu untersuchen, die wesentlichen Einfluß auf solche Leistungsparameter wie Verarbeitungsleistung und Steueraufwand des Systems haben. Zum einen betrifft das die Qualität der Übergabesymbole zwischen Sprachsignalverarbeitung und symbolischer Verarbeitung (z. B. Allophone, Phoneme, Halbsilben, Silben, Wörter) und zum anderen die Art und Weise der Wechselwirkung beider Komponenten. Diese Wechselwirkung reicht von der unidirektionalen Übergabe "von unten nach oben" bis hin zur Erstellung präziser Erwartungen auf der Basis einer Sprachbeschreibung zur Steuerung der Sprachsignalverarbeitung.
Anliegen des Vortrages ist es, diese Schnittstelle aus der Sicht heute existierender Systeme und in Hinblick auf künftige Forderungen an den Leistungsumfang von Spracherkennungssystemen zu beschreiben.

2. Sprachsignalverarbeitung und Sprachverarbeitung

Ein Spracherkennungssystem soll aus den beiden Verarbeitungsstufen Sprachsignalverarbeitung und Sprachverarbeitung bestehen. Im Folgenden soll kurz dargelegt werden, welcher Leistungsumfang von

beiden Stufen erbracht werden soll. Die Problematik der Schnittstelle zwischen beiden wird hierbei schon offenbar, soll jedoch zunächst noch zurückgestellt werden.

2.1. Leistungsumfang der Sprachsignalverarbeitung

Wie schon in der Einleitung erwähnt wurde, wird der Leistungsumfang von dem Sprachsignal bis zu einer symbolischen Darstellung bemessen. Dazu kann die in Bild 2 dargestellte, an den klassischen Aufbau eines Erkenners angelehnte Anordnung zugrundegelegt werden, wobei nochmals darauf verwiesen werden soll, daß in dem vorliegenden Beitrag der Begriff "Erkenner" im Sinne eines Spracherkenners insgesamt weiter gefaßt werden soll (==> Erkennen im Sinne von Verstehen).

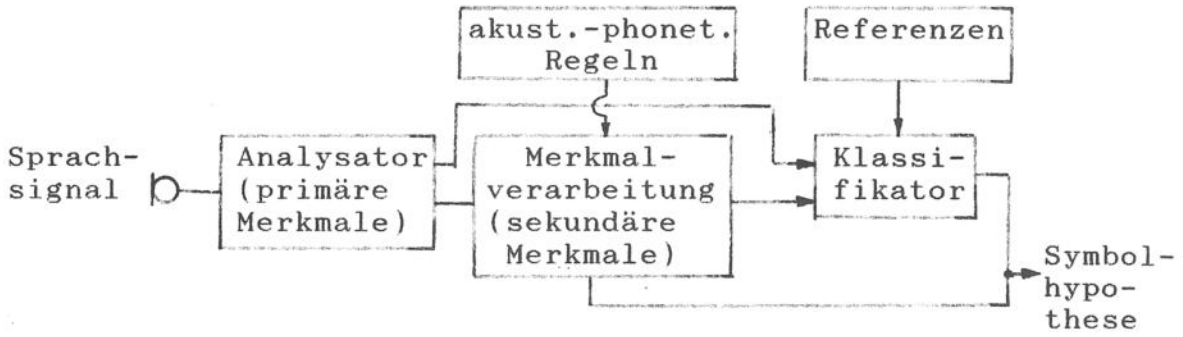

Bild 2: Aufbau eines Sprachsignalverarbeitungssystems

Das Sprachsignal wird nach Abtastung und Digitalisierung mit verschiedenen Verfahren analysiert (Spektralanalyse, Nulldurchgansanalyse, Untersuchung des Energieverlaufs bzw. des Hüllkurvenverlaufs usw.). Die gewonnenen Merkmale werden entweder direkt zur Klassifikation durch Vergleich mit Referenzmerkmalen geleitet oder durch Anwendung spezieller akustisch-phonetischer Regeln zu sekundären Merkmalen verarbeitet. Diese verarbeiteten Merkmale dienen entweder wieder dem Klassifikator als Eingangsgröße oder stellen bereits selbst Symbolhypothesen dar.
Die Sprachsignalverarbeitung ist damit durch folgenden Leistungsumfang beschrieben:

Eingangsinformation: Sprachsignal
Ausgangsinformation: Symbolhypothesen
Wissensquellen:
- deklarativ: akustisch-phonetische Regeln
  Referenzmuster
- prozedural: Transformationsalgorithmen
  Klassifikationsalgorithmen

2.2. Leistungsumfang der symbolischen Sprachverarbeitung

Die symbolische Sprachverarbeitung muß nun für diese von der Signalverarbeitung erzeugten Symbolhypothesen eine Satz- bzw. Phraseninterpretation erstellen. Vorbilder für diesen Verarbeitungsabschnitt finden sich in der automatischen Sprachverarbeitung, wo lexikalisches, syntaktisches und semantisches Wissen

untersucht und geeignet dargestellt zur Verfügung gestellt wird. Jedoch treten bei dem Versuch der Übernahme zwei Probleme auf. Zum einen hat die automatische Sprachverarbeitung den Vorteil, ihre Eingangsinformation auf lexikalisch korrekte Wortfolgen beschränken zu können, zum anderen wird als Ausgangsinformation oft "nur" die Entscheidung gefällt, ob ein verarbeiteter Satz Satz der betrachteten Sprache ist. Es ist offensichtlich, daß hier die Eingangsgröße "symbolische (d.h. möglicherweise nicht wortweise) Beschreibung von Spontansprache" (meist syntaktisch nicht korrekt) zu großen Problemen führen wird.
Somit wird der geforderte Leistungsumfang der symbolischen Sprachverarbeitung im vorliegenden Fall beschrieben durch:

Eingangsinformation: Symbolhypothesen
Ausgangsinformation: Phraseninterpretation
Wissensquellen:
deklarativ: an Symbole angepaßtes Wörterbuch
an Phrasen angepaßte Syntax- und Semantikregeln
prozedural: robuste Wortparsing- und Phrasenparsingalgorithmen

3. Beschreibung von Spracherkennungsaufgaben

Nachdem im Punkt 2.2. allgemeine Forderungen an die symbolische Sprachverarbeitung angeführt wurden, soll im Folgenden durch eine Beschreibung verschiedener Typen von Spracherkennungsaufgaben gezeigt werden, daß die symbolische Sprachverarbeitung in starkem Maße aufgabenabhängig gestaltet werden kann bezüglich ihres konkreten Leistungsumfangs. Das wiederum kann bei der Gestaltung der Schnittstelle zu unterschiedlich aufwendigen Realisierungen verwendet werden.

3.1. Künstliche Sprache

Spracherkennungsaufgaben, bei denen aufgabenbedingt das Vokabular und der Sprachaufbau auf eine künstliche Sprache zurückgeführt werden können, zeichnen sich durch folgende Besonderheiten aus:

- Das Vokabular ist meist überschaubar klein.
- Der Satzaufbau entspricht einer definierten, meist sehr eingeschränkten Syntax.
- Meist ergibt sich die Seantik fast vollständig aus der syntaktischen Struktur.
- Die Sprachen sind meist nicht ohne Wortpausen sprechbar.

Beispiele für derartige Sprachen sind Kommandosprachen bis hin zu Programmiersprachen.
Bei der Schnittstellengestaltung für derartige Sprachen sind zwei Vorteile auszunutzen, zum einen die relativ leichte Bestimmung der Wortpausen und zum anderen die strenge Syntax, die günstig zur Einschränkung des Suchraumes der Sprachsignalverarbeitung verwendet werden kann.

3.2. Natürliche Sprache

Eine Erkennungsaufgabe für natürliche Sprache sei dadurch gekennzeichnet, daß der Sprecher gut artikulierte und fehlerfreie natürliche Sprache produziert, wie z. B. beim Vorlesen. Die entstehende Sprache ist dann gekennzeichnet durch

- ein sehr großes Vokabular mit bekannter Wort-Bedeutungs-Zuordnung
- grammmatikalisch richtigen, u. U. jedoch sehr komplizierten Satzbau
- komplizierte semantische Strukturen (über Satzgrenzen reichend)
- eine Satzintonation, die der Interpunktion entspricht und Bedeutungsinhalte widerspiegelt
- die Möglichkeit fließender Sprechweise

Die Schnittstellengestaltung für Sprachen dieser Art kann die vorausgesetzte gute Artikulation und den richtigen Satzbau zur Rückwirkung der symbolischen Verarbeitung auf die Sprachsignalverarbeitung nutzen. Für die symbolische Verarbeitung steht die Forderung nach sehr leistungsfähigen Sprachmodellen, da u. U. sehr lange Sprachabschnitte strukturell untersucht werden müssen.

3.3. Spontansprache

Die Probleme, die bei dieser Sprachart auftreten seien an einem Beispiel demonstriert (geübter Sprecher, bewußte Spontansprache): "Das is richtig - wir haben also sehr viel Glück gehabt - wir habm ein -- halbes Jahr sechs Monate habm wir intensiv gearbeitet -- und wir hatten das Glück gehabt äh daß äh jemand von der -- Hansa -- Schallplattenproduktion zu uns runter kam -- diese Band gehört hat äh 'hm und meinte ich find das so toll -- komm wir machen 'ne Langspielplatte ne."

"-" ... kurze Pause
"--"... lange Pause

Diese Art von Sprache ist ihrerseits durch folgende Phänomene gekennzeichnet:

- Sprechpausen sind kaum identisch mit Interpunktion.
- Es treten sprachlich bedeutungslose Einfügungen auf (äh,'hm, ne)
- Es kommt zur Wiederholung gleichen Inhalts (hier mit anderen Worten: halbes Jahr - sechs Monate).
- Die grammatischen Strukturen sind im Rahmen der Gesamtäußerung nicht exakt(z. B. Zeitverwendung).
- Die Wortpausen sind durch die natürliche Sprechweise nicht explizit.

Bei der Verarbeitung dieser Art Sprache sind vor allem beim Aufbau des sprachlichen Hintergrundwissens spezielle Ansätze erforderlich. So ist es nicht notwendig, für die gesamte Sprachäußerung eine zusammenhängende Struktur zu erzeugen, es ist hier ausreichend, zusammengehörende Wörter zu gruppieren (Phrasenbildung) und diese in einen losen inhaltlichen Zusammenhang zu bringen. Auf diese Weise kann eine Verarbeitung organisiert werden, die nicht jeden Signalabschnitt linguistisch interpretiert.
Die Wechselwirkung zwischen Sprachsignalverarbeitung und Sprachverarbeitung ist hier durch den losen syntaktischen Aufbau der Äußerungen und die verschliffene Sprechweise sehr kompliziert. Große Bedeutung kommt der Gestaltung des Wörterbuches als Element der Schnittstelle zu. Wortfolgen können nur teilweise "vorhergesagt" werden. Pausen stellen keine eindeutige Unterstützung beim Aufbau einer Struktur dar.

4. Möglichkeiten der Schnittstellengestaltung

Bei der Betrachtung der Spracherkennungsaufgaben wurden schon Bemerkungen zu prinzipiell möglichen Wechselwirkungen der Bottom-End- und der Top-End-Verarbeitung gemacht. Es werden prinzipiell zwei Arten der Wechselwirkung innerhalb dieser Schnittstelle betrachtet, die im Folgenden kurz vorgestellt werden sollen.

4.1. Rückwirkungsfreie Kombination von Bottom-End- und Top-End-Verarbeitung

Die rückwirkungsfreie Kombination ist dadurch gekennzeichnet, daß die sprachsignalverarbeitung ergänzt wird durch eine Verarbeitungsstufe, die ihre Ausgangsinformation (Symbolhypothesen) weiter verarbeitet mit dem Ziel einer Interpretation, ohne ihrerseits die Sprachsignalverarbeitung zu beeinflussen, z. B. durch eigene Hypothesen oder gezielte Rückfragen. Bild 3 zeigt den prinzipiellen Aufbau eines solchen Systems.

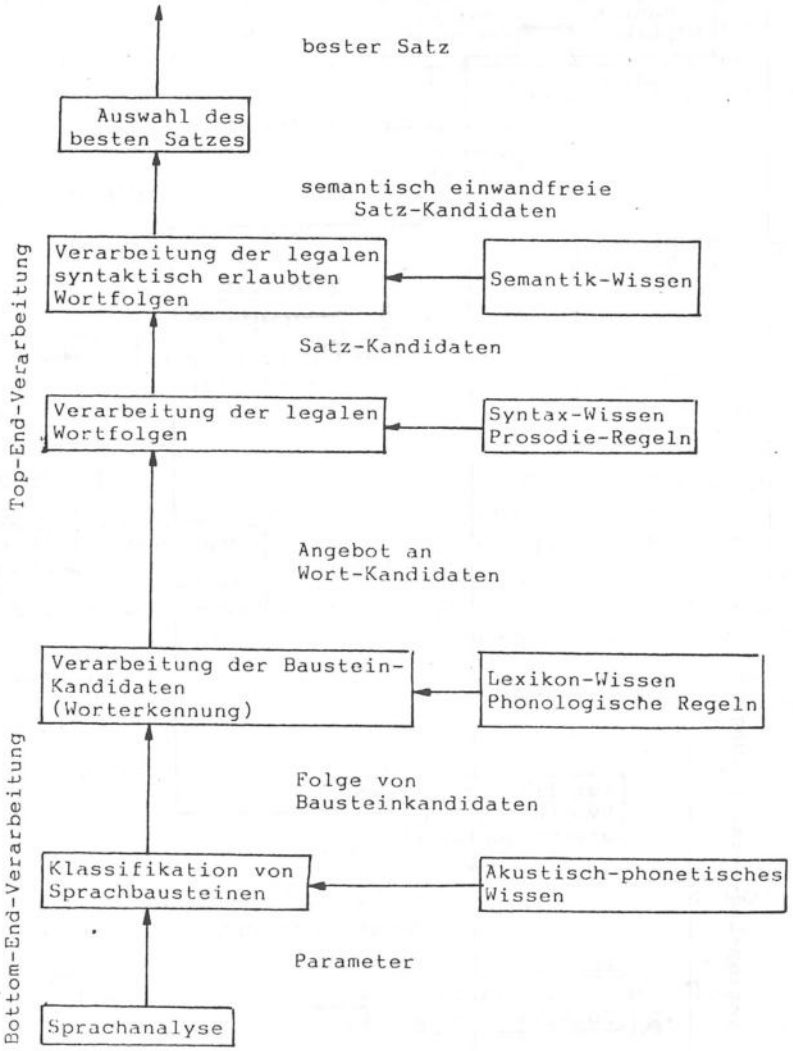

Bild 3:Beispiel für ein rückwirkungsfreies Spracherkennungssystem nach /1/

Die Schnittstelle zwischen beiden Verarbeitungsabschnitten wurde hier in die Wortebene gelegt, was aus der Sicht der Top-End-Verarbeitung gerechtfertigt erscheint. Im Interesse der Universalität und Vergleichbarkeit mit rückwirkenden Systemen (siehe

4.2.) erscheint es jedoch günstiger, die Bottom-End-Verarbeitung nicht bis zur Wortebene zu rechnen, sondern ihren Output auf eine beliebige, linguistisch relevante Symbolebene zu fixieren. Die lexikalische Komponente wird damit direkter Bestandteil der Schnittstelle und kann durch universelle Gestaltung die Verwendung unterschiedlicher Übergabesymbole ermöglichen (siehe Abschnitt 5).

4.2. Rückwirkende Kombination von Bottom-End- und Top-End-Verarbeitung

Die rückwirkende Kombination von Sprachsignalverarbeitung und Sprachverarbeitung ist dadurch gekennzeichnet, daß die "höhere" Ebene durch linguistisch motivierte Erwartungen an den Aufbau einer sprachlichen Äußerung eine Unterstützung der Sprachsignalverarbeitung bieten kann.

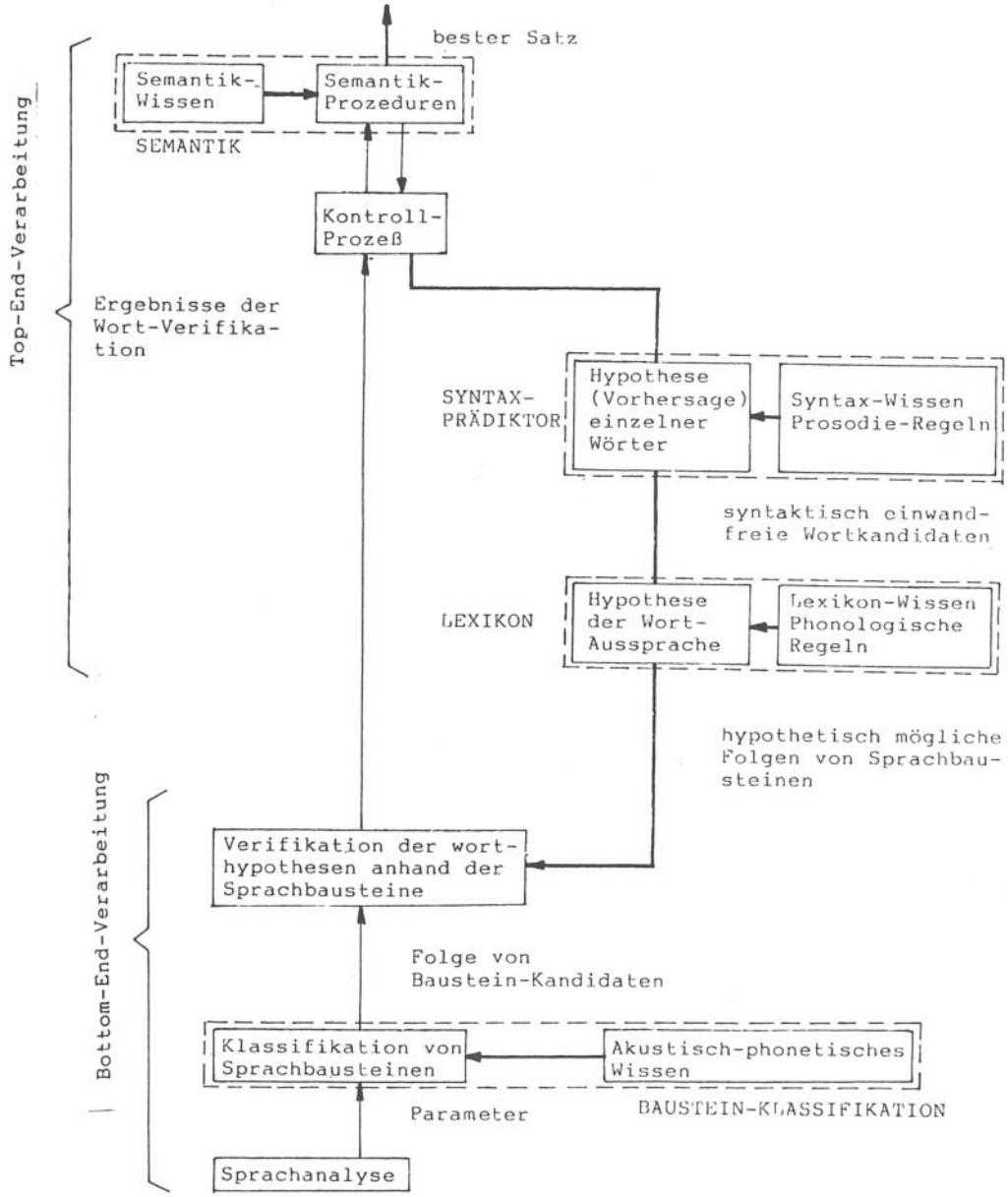

Bild 4: Beispiel für ein rückwirkendes Spracherkennungssystem nach /1/

Diese Unterstützung reicht von der Einschränkung des Suchraumes bis hin zu einer speziellen, lautspezifischen Analyseauswahl. Dieses Vorgehen erscheint aus der Untersuchung humaner Sprachverarbeitungsprinzipien naheliegend, da die menschliche Verarbeitungsleistung stark durch linguistisch motivierte Einschränkungen beeinflußt wird /2/. Im Bild 4 ist ein rückwirkendes Spracherkennungssystem dargestellt.
Bei dem hier vorgeschlagenen Konzept verfügt die Bottom-End-Verarbeitung über keine autonome lexikalische Komponente, d. h. sie ist direkt auf Hypothesen "von oben" angewiesen. Eine wie in 4.1. bereits vorgeschlagene Herauslösung des Lexikons würde auch hier mehr Universalität ermöglichen, indem auch "nur von unten" Worthypothesen erzeugt werden können. Die Wirkungsweise von oben würde dadurch nicht eingeschränkt, da die Wortebene als Eingangs- bzw. Ausgangsinformation der Top-End-Verarbeitung erhalten bleiben würde. Die Herauslösung des Lexikons würde auch im Fall eines rückwirkenden Spracherkennungssystems die Verwendung verschiedener Formen der Bottom-End-Verarbeitung bezüglich der von ihr erzeugten Ausgangssymbole ermöglichen. Die lexikalische Komponente eines solchen Systems bestimmt durch ihren Aufbau, wie "tief" durch Vorhersage von oben in den Prozeß der Sprachsignalverarbeitung eingegriffen werden kann.

5. Beschreibung von Übergabesymbolen

Die Überlegungen zur Gestaltung der Schnittstelle zwischen der Sprachsignalverarbeitung und der symbolischen Sprachverarbeitung umfassen die Art und Weise der Wechselwirkung und die Gestalt der Übergabesymbole. Im Abschnitt 4 wurden die Vorteile der gesonderten Betrachtung der lexikalischen Komponente als wesentliches Element der Schnittstelle dargestellt. Nun soll die Form der Übergabesymbole und ihr Einfluß auf die Gesamtkonzeption betrachtet werden. Der jetzt zugrundegelegte Systemaufbau ist in Bild 5 dargestellt. Der Aufbau der Übergabesymbole wird durch zwei Faktoren bestimmt. Zunächst sei nur die Richtung Bottom-End-Verarbeitung --> universelle Schnittstelle betrachtet. Hier werden linguistisch relevante Symbole erwartet, die nach Durchlauf der Regelanwendung und Klassifikation (siehe 2.1.) vorliegen, da diese Verarbeitungsstufen in der Erstanwendung dem Lexikon nicht zur Verfügung stehen. Die andere Richtung der Wechselwirkung kann jedoch anders gestaltet werden. Dies wiederum hängt vom Wörterbuchaufbau ab, wie in Bild 5 angedeutet. Das Lexikon könnte im umfangreichsten Fall so aufgebaut sein, daß in einer Richtung die Zuordnung Symbolhypothesenfolge --> Wort und Übergabe an die Top-End-Verarbeitung erfolgt und in der anderen Richtung die Verarbeitung Worthypothese --> Symbolhypothesenfolge --> Merkmalfolge --> Parameterfolge, wobei mit abnehmender Komplexität auch die Möglichkeit der Einflußnahme abnehmen muß.
Das bedeutet, die Steuerung von oben muß in jeder Ebene zunehmend "empfehlenden Charakter" haben, da die auf Parameterebene vorhandene Unsicherheit (durch Realisierungsvielfalt) zugunsten des erwarteten Ergebnisses ausgenutzt wird.

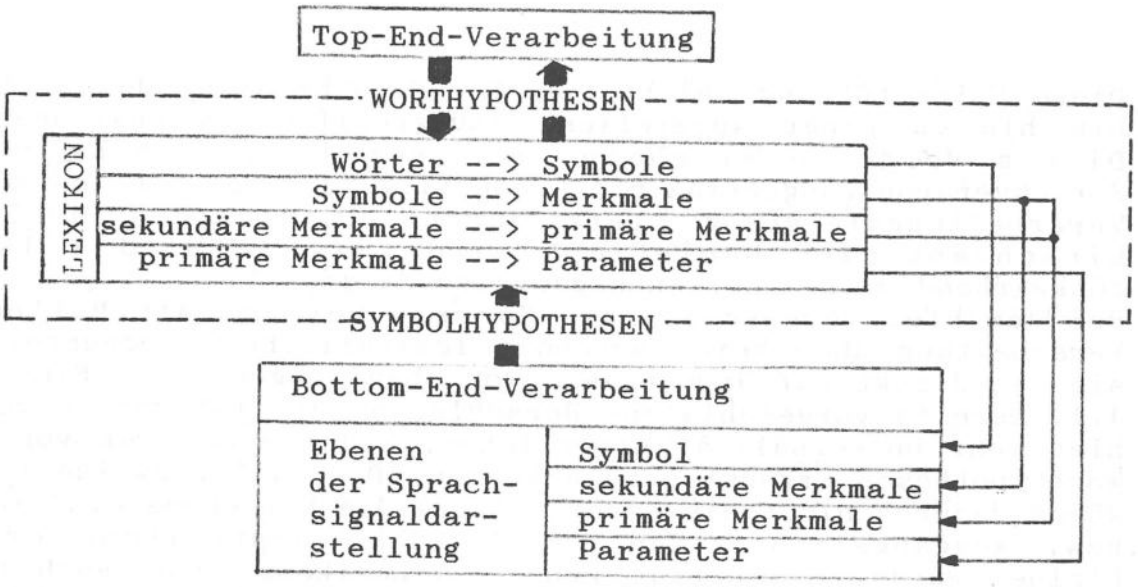

Bild 5: Die universelle Schnittstelle in einem Spracherkennungssystem mit Rückwirkung

Die Form der Übergabesymbole wird systemabhängig davon bestimmt, für welche sprachlichen Einheiten Regeln und Referenzmuster günstig bereitgestellt werden können. Sie reichen von Phonemen bis hin zu Wörtern, wobei der geringen Anzahl bei sehr kleinen Übergabesymbolen (Basiselementen) die komplizierte Ermittelbarkeit und starke Kontextabhängigkeit gegenübersteht. Existierende Spracherkennungssysteme verwenden z. B. Phoneme, Halbsilben, Silben oder Wörter, wobei jedoch zu bemerken ist, daß die Entscheidung für eine Symbolart meist ohne Betrachtung von Alternativen erfolgt. Zur generellen Verbesserung der Leistungsfähigkeit von Spracherkennungssystemen sollten dazu grundlegende Untersuchungen durchgeführt werden.

6. Zusammenfassung

Spracherkennungssyteme,die Äußerungen von mehr als einem Wort Länge verarbeiten sollen, bestehen meist aus einer Kombination von Sprachsignalverarbeitung und symbolischer Sprachverarbeitung. Das Zusammenspiel beider Komponenten kann verschieden realisiert werden und beeinflußt erheblich sowohl den Gesamtaufwand als auch das Gesamtergebnis. Maximale Leistungsfähigkeit ist bei rückwirkender Zusammenarbeit der beiden Komponenten und einer sorgfältigen Abstimmung der lexikalischen Komponente auf den Leistungsumfang der Sprachsignalverarbeitung zu erwarten.

7. Literatur

/1/ Sickert, K. Automatische Spracheingabe und Sprachausgabe Verlag Markt u. Technik, Haar bei München, 1983

/2/ Marslen-Wilson, W. D. Aspects of human speech understanding in: Fallside, F. ; Woods, W. (eds.) Computer Speech Processing, Prentice Hall International, Englewood Cliffs, 1987