BMBF

Verbmobil
Verbundvorhaben

Abschlußbericht VM-TP 3.3
Lauteinheit deutsch

Gudrun Flach

TU Dresden

Report 102
30.1.96

30.1.96

Gudrun Flach

Institut für Technische Akustik
Fakultät Elektrotechnik
Technische Universität Dresden
Mommsenstr. 13
01062 Dresden

Tel.: (0351) 463 - 4283
Fax: (0351) 463 - 7091
e-mail: `flach@eakws1.et.tu-dresden.de`

Gehört zum Antragsabschnitt: 3.3

Die vorliegende Arbeit wurde im Rahmen des Verbundvorhabens Verbmobil vom Bundesministerium für Bildung, Wissenschaft, Forschung und Technologie (BMBF) unter dem Förderkennzeichen 01 IV 102 K2 gefördert. Die Verantwortung für den Inhalt dieser Arbeit liegt bei dem Autor.

Inhaltsverzeichnis

Abbildungsverzeichnis

1 Einleitung

1.1 Anliegen des Teilprojektes "Lauteinheit deutsch"

Das Teilprojekt "Lauteinheit deutsch" hat die Auswahl geeigneter Lauteinheiten für das Deutsche zum Ziel, wobei eine optimale Erkennungsrate das wesentlichste Entscheidungskriterium darstellt. In dem Arbeitspaket werden alternativ zwei Typen von Lauteinheiten in Hinblick auf ihre Eignung zur Erkennung untersucht:

- frei gewählte kontextabhängige Einheiten
- silbenorientierte Lauteinheiten

Ein weiterer wesentlicher Schwerpunkt des Arbeitspaketes besteht in der Erstellung und Wartung eines Aussprachewörterbuchs speziell für die Anwendung in der akustischen Erkennung.

kontextabhängige Lauteinheiten Wortbeschreibungen bauen auf Modellen für Wortuntereinheiten auf. Die Auswahl erfolgt mit phonologischen, phonetischen, insbesondere aber auch anhand technischer Kriterien. Die phonetischen und technischen Kriterien sind sehr eng verzahnt, da darin die Ähnlichkeit von Lauten, die Ausprägung der Laute in verschiedenen Kontexten und Hinweise für sinnvolle Auftrennungen von Modellen enthalten sind. Ein rein technisches Problem ist die Trainierbarkeit der Modelle mit einer gegeben Stichprobe.

Kontextabhängige Lautmodelle in Form von Diphonen und Triphonen werden zur Zeit favorisiert. Als Kontext werden dabei Lautkategorien benutzt, bei deren Auswahl wieder phonetische und technische Gesichtspunkte eine Rolle spielen.

silbenorientierte Lauteinheiten Besondere Vorteile für die akustisch-phonetische Analyse verspricht die Nutzung der Silbenstruktur. Bei diesem Ansatz erfolgt die akustische Modellierung mit Hidden-Markov-Modellen für ganze Phonemfolgen (Konsonantencluster und Vokalklassen) unter Berücksichtigung ihrer Position in der Silbe.

Aussprachewörterbuch Für die gewählten Repräsentationseinheiten muß ein domänenspezifisches Aussprachelexikon erstellt werden. Wesentlich dabei ist die Auswahl einer Darstellungsform für spezielle, nur in Spontansprache vorkommende sprachliche Besonderheiten (Häsitationen, Wortabbrüche und Nichtwörter). Weiterhin muß innerhalb des Aussprachewörterbuchs das Problem von Aussprachevarianten und Wortverschmelzungen mit erfaßt werden. Dazu wird Sprachmaterial untersucht und es werden Generierungsregeln abgeleitet.

1.2 Inhalt des Berichtes

In dem vorliegenden Bericht werden zunächst Grundlagen aus der Phonetik und Phonologie aufgeführt, auf deren Basis Untersuchungen zum Thema durchgeführt wurden. Weiterhin werden Vorverarbeitungs- und Erkennungsstrategien zur Zuordnung des physikalischen Signals zu symbolischen Beschreibungen erwähnt. Diese beiden Abschnitte dienen der Schaffung eines Überblicks über die Problematik. In weiteren Punkten werden die Untersuchungen zu frei gewählten kontextabhängigen Lauteinheiten und zu den silbenorientierten Lauteinheiten vorgestellt.

Einen wesentlichen Schwerpunkt im Rahmen des Arbeitspaketes stellten die Arbeiten zur Generieung und Wartung des Aussprachewörterbuches dar. In dem dazu gehörigen Abschnitt des Berichtes werden diese Arbeiten vorgestellt. Diese umfassen:

- eine Beschreibung der verwendeten Korpora
- Untersuchungen zu Aussprachevarianten und Wortverschmelzungen
- die Ableitung von Regeln zur automatischen Erzeugung von Aussprachevarianten und Wortverschmelzungen
- eine Beschreibung der zur Wörterbuchgenerierung und -wartung entwickelten Tools

Weiterhin enthalten ist eine Beschreibung der Arbeiten zur Generierung von Hintergrundwissen für Evaluationsexperimente für Vorverarbeitungsstrategien mit dem Hidden-Markov-Model-Tool-Kit ([ht93]).

Im letzten Abschnitt werden experimentelle Untersuchungen zu Aussprachevarianten und Wortverschmelzungen vorgestellt.

2 Phonetisch-phonologische Grundlagen

Bei der Untersuchung der humanen Sprachverarbeitung zeigt sich, daß Wörter und Laute als kleinste Einheiten von kognitiver Relevanz sind. Sie sind für den Hörer und Sprecher abstrahierbare Einheiten bei der Perzeption und Produktion einer lautsprachlichen Äußerung. Sprachverarbeitende Systeme sollten daher ebenfalls in der Lage sein, diese Einheiten aus dem Redestrom zu abstrahieren, wobei das Ausgangssignal (Schalldruck-Zeit-Funktion) jedoch eine weitgehend kontinuierliche und demnach schwer zu segmentierende Form aufweist. Die Betrachtung phonetisch-phonologischer Grundlagen zur Sprachbeschreibung soll mögliche Ansätze zur Festlegung und Beschreibung möglicher Lauteinheiten aufzeigen.

2.1 Phonetische Beschreibung realer Sprache

Die gesprochenen Wörter einer Sprache sind durch den Hörer in Phoneme segmentierbar. Als Phoneme werden die kleinsten bedeutungs**unterscheidenden** Abschnitte der Lautsprache bezeichnet. Dabei ergibt sich ein endliches Inventar von 40 bis 80 Einheiten, mit denen jedem Wort einer Sprache eine einzige (eindeutige) Repräsentation zugeordnet werden kann.

Phoneme sind rein theoretische Größen, die der Festlegung der lexikalischen Elemente lautsprachlicher Äußerungen dienen. Die Basis dafür bildet eine eindeutige Zuordnungsvorschrift von Phonemen und Graphemen. Die so erzeugten Phonemfolgen müssen im Anschluß in konkrete Phonemrealisierungen transformiert werden. Diese werden aus der phonologischen Wortbeschreibung und dem damit gegebenem Kontext abgeleitet, z. B. ergibt sich eine unerschiedliche Aspiration des /p/-Phonems in Abhängigkeit von seiner Stellung im Wort und dem umgebenden Kontext. Die so entstehende Repräsentationsform wird als allophonische Ebene bezeichnet. Gegenüber der phonologischen Wortbeschreibung erzielt man eine der realen Sprachproduktion nähere Beschreibung. Diese Ebene ist jedoch weiterhin theoretisch, denn die Worttranskriptionen beider Ebenen (allophonisch-phonetische(enge) und phonologische(weite)) können automatisch ineinander überführt werden.

Phoneme haben nicht nur einen Bezug zu Graphemen, sie können auch durch Merkmale ausgedrückt werden, die Angaben zu ihrer artikulatorischen Realisierung beschreiben. Diese Merkmale werden als distinktive Merkmale bezeichnet [ch68]. Die distinktiven Merkmale werden für jedes Phonem als *vorhanden, nicht vorhanden* bzw. *nicht relevant* gekennzeichnet. Phonologische (distinktive) Merkmale zerlegen immer ganze Phonemklassen in Teilklassen. Das Gesamtinventar distinktiver Merkmale ermöglicht die eindeutige Beschreibung jedes Phonems einer Sprache.

Rein phonetische Beschreibung von Äußerungen. Zur Untersuchung des Zusammenhangs zwischen einer regelbasiert erzeugten allophonisch-phonetischen

Beschreibung von Sprachäußerungen und realer Sprache werden durch phonetisch kompetente Beobachter enge phonetische Transkriptionen (z. B. mit API-Inventar) erstellt, innerhalb derer alle akustisch unterscheidbaren Varianten extra gekennzeichnet werden. Das Ziel dabei besteht in einer unmittelbaren, d. h. nicht durch inhaltliche Verarbeitung modifizierten Beschreibung sprachlicher Äußerungen. Das bedeutet, daß in diesen Beschreibungen nur die durch das Ohr wahrnehmbaren und differenzierbaren Bestandteile enthalten sein dürfen. Dabei ist zu beobachten, daß reale Lautereignisse bei eingeschränktem Kontext nicht in jedem Fall in Wortfolgen aufgelöst werden können. Durch den eingeschränkten Kontext ist eine inhaltliche Verarbeitung nicht erfolgreich möglich, so daß die Abbildung des akustischen Ereignisses auf eine Symbolfolge (Wortfolge) nicht realisiert werden kann. Wird der Kontext vervollständigt, ist jedoch eine vollständige Abbildung der gesamten Äußerung möglich. Wir finden hier den Nachweis einer engen Wechselwirkung zwischen Hören und Verarbeiten. Sprecher und Hörer sind in der Lage, die Wörter auf ihre Zitierformen zurückzuführen. Die produzierten und perzipierten reduzierten Formen sind somit durch komplexe wissensbasierte Verarbeitung auf symbolische Repräsentationen abbildbar. Diese Tatsache muß beim Entwurf automatischer Spracherkennungssysteme, insbesondere bei der Entwicklung der Akustik-Phonetik-Orthographie-Ebene speziell berücksichtigt werden.

Bei der Untersuchung dieser symbolphonetischen Wortrepräsentationen beobachtet man systematische Veränderungen der zugrundeliegenden kanonischen Ausgangsform . Es treten zwei Typen von Veränderungen auf:

- die Verstärkung (/'zi:b@n/ → /'zi:bEn/, /'zi:'bEn/ bzw. /zi'bEn/) und
- die Abschwächung (/'zi:b@n/ → /'zi:bm/ bzw. /'zi:m/)

Phonetische Abschwächungsprozesse lassen sich durch Regeln präzise beschreiben, die die Auswirkung von Assimilation, Koartikulation, Reduktion und Elision im konkreten Lautkontext beschreiben. Bei der Abschwächung treten drei verschieden Stufen auf:

1. Die lexikalische Differenzierung wird auf phonetisch minimale Weise ausgedrückt.
2. Die lexikalische Differenzierung wird unmöglich (phonetisch gleiche Realisierung lexikalisch unterschiedlicher Wörter).
3. Von der phonetischen Oberfläche verschwinden ganze Wörter. Diese müssen dann aus dem Kontext rekonstruiert werden.

Ein derartiges Regelsystem zur Ableitung möglicher abgeschwächter Formen für das Deutsche ist in [ko90] beschrieben. Diese Regeln kann man zwar auf Wortebene anwenden, jedoch ist zur Ableitung natürlichsprachlicher Phonation die Berücksichtigung der syntaktischen und semantischen Kategorien und der konkreten Sprechsituation (mit dem Szenario weitgehend festgelegt) erforderlich. Die

Phänomene der Verstärkung sind noch enger an die Kategorien Syntax, Semantik und Sprechstil gebunden.

Mit Hilfe phonetisch-phonologischer Beschreibungen und und lautmodifizierender Regelwerke ist man in der Lage, der natürlichen Artikulation gut angepaßte phonetische Beschreibungen zu erzeugen. Problematisch dabei ist jedoch, daß die erzielten Ergebnisse nach wie vor theoretische Konstrukte sind. Das wesentliche Problem der automatischen Spracherkennung, die Beschreibung des Zusammenhanges zwischen den Signalen, die vom Nervensystem verarbeitet werden und dem kognitiven Resultat dieser Verarbeitung, also der Kategorie der subjektiv resultierenden Wahrnehmung, wie in [vm90] formuliert, kann damit noch nicht gelöst werden. Dieser Zusammenhang muß momentan im Rahmen eines Lernprozesses auf der Grundlage leistungsfähiger phonetischer Beschreibungen modelliert werden.

2.2 Generierung phonetischen Wissens für Sprachverarbeitungssysteme

Zur Erstellung geeigneten phonetisch-phonologischen Hintergrundwissens für die Sprachsignalverarbeitung müssen nach [vm90] folgende Grundsätze beachtet werden:

1. Die Modellentwicklung muß sich an der phonetisch gegebenen Realität orientieren.

2. Die tatsächliche Realisierung der Wörter durch menschliche Kommunikatoren muß untersucht werden.

3. Der Zusammenhang zwischen symbolphonetischer und signalphonetischer Beschreibung muß untersucht werden, um eine richtige Zuordnung zu ermöglichen.

4. Signalphonetische Parameter müssen kognitiv durch Auswertung geeigneter Wissensquellen bewertbar sein.

5. Der Erwerb des nötigen Wissens wird durch Erfassung und systematische Auswertung von regional gesprochenem Hochdeutsch unterstützt.

6. Es müssen leistungsfähige Segmentations-, Transkriptions- und Auswertesysteme zur Verfügung stehen.

7. Neben der Erfassung von Sprachmaterial müssen auch sprechakttypischer Aspekte unterschiedlicher Szenarien und deren Auswirkung auf resultierende phonetische Laut- und Wortformen untersucht werden.

8. Im Vorfeld muß gelesenes Sprachmaterial untersucht werden. Das Ziel dabei ist die Herstellung einer Verbindung zu Analyseergebnissen von Laborsprache.

9. Im zweiten Schritt muß Spontansprache erfaßt und untersucht werden

10. Das Ziel ist die Erweiterung bestehender Regelsysteme. Syntax, Semantik, Pragmatik und Stil sollten mit einbezogen werden für Prosodieeinsatz und Generierung von Aussprachevarianten aus kanonischen Formen.

2.3 Phonologische Lautbeschreibung

Allgemeines Phoneme als kleinste Einheit der lautlichen Beschreibung werden als Mengen von Merkmalen beschrieben, die sich an der artikulatorischen Phonetik orientieren. Die Grundlage dafür bilden die Parameter Artikulationsort und Artikulationsart. Koartikulationsphänomene lassen sich durch Betrachtung der Einzellautartikulation erklären.

Man unterscheidet bei der Lautbildung Artikulationsort und Artikulationsart (siehe Tabelle 1 und 2).

Bezeichnung des Artikulationsortes	beteiligte Artikulatoren
bilabial	Lippen
labiodental	Lippen und Zähne
dental	Zähne
alveolar	Zahntaschen
palatoalveolar	Gaumen
palatal	Gaumen
velar	Gaumensegel, weicher Gaumen
uvular	Zäpfchen
pharyngal	Rachen
laryngal	Kehlkopf

Tabelle 1: Liste der Artikulationsorte

Auf dieser Basis sind auch Koartikulationsphänomene beschreibbar, indem man die Artikulation der Einzellaute und deren mögliche Wechselwirkungen betrachtet.

Phonotaxe Hinsichtlich der zulässigen Lautfolgen existieren sprachspezifische Einschränkungen, die in der Phonotaxe beschrieben sind. Phonotaktische Einschränkungen beziehen sich meist auf die sprachliche Struktur der Silbe. Jede Silbe besteht aus einem Gipfel, der ein betonter Laut sein muß; die Elemente rechts und links vom Silbengipfel sind entsprechend steigender Sonorität angeordnet. Dazu wird die unter den Phonemen geltende Sonoritätshierarchie ausgewertet. Die Sonorität eines Lautes ist mit seiner Artikulationsart verknüpft. Sie ist bei den Plosiven am geringsten und nimmt entsprechend der in der Tabelle 2 verzeichneten Reihenfolge bis zu den Vokalen zu. (Dabei sind jedoch sprachspezifische Ausnahmen zu beachten.) Weitere Resriktionen für den Silbenaufbau

resultieren aus der zulässigen Länge und aus sprachspezifischen Einschränkunkungen bezüglich zugelassener Folgen innerhalb der Konsonantenbereiche.
Die formale Silbendefinition umfaßt demnach :

- den Anfangsrand (**Onset**)
- den Gipfel (**Peak**) und
- den Endrand (**Koda**).

Dabei können Peak und Koda zum **Reim** der Silbe zusammengefaßt werden.

regelhafte Ausspracheveränderungen Innerhalb der Phonologie werden ebenfalls Ausspracheveränderungen betrachtet. Hier stehen die regelhaften Phonemvarianten, die Allophone, im Mittelpunkt. Dabei werden neben phonologischen Phänomenen auch morhophonemisch bedingte Abweichungen betrachtet, die durch Flexion und Derivation bedingte Veränderungen beschreiben. Diese sind hier jedoch im Falle einer orthographischen Repräsentation nicht von Interesse. Die hier zu betrachtenden lautverändernden Prozesse sind Assimilationen, Dissimilationen und Tilgungen. Wir sprechen von :

- **Assimilation**, wenn Merkmale von Segmenten durch Merkmale benachbarter Segmente überschrieben werden
- **Dissimilation**, wenn gleiche Merkmale in benachbarten Segmenten werden so verändert werden, daß ein Kontrast entsteht.

2.3.1 Linguistische Wortmodellierung

Linguistische Wortmodelle haben 2 Strukturebenen. Diese Strukturebenen umfassen die phonemische und die morphemische Ebene. Die Darstellung erfolgt im einfachen Fall als Phonem- und Morphemfolgen, im weiterentwickelten Fall in Form phonologischer bzw. morphologischer Strukturen. Eine phonologische Struktur ist dabei eine Wortrepräsentation in Form phonologischer Ereignisse unterschiedlicher relativer Ausdehnung (vgl. Tab. 3). Morhophologische Strukturen werden über die wortsyntaktische morphologische Kombinatorik sowie über Prinzipien der linearen und simultanen Verbindung von Morphemkombinationen definiert.
Die Grundlage der linguistischen Wortmodellierung bildet die Tatsache, das Wörter in diskrete Segmente unterteilt werden können, obwohl sie durch ein kontinuierliches physikalisches Signal repräsentiert werden. Den Wortbildungssegmenten werden distinktive Merkmale zugeordnet, die als Ereignisse interpretiert werden. Die Wortrepräsentation erfolgt somit in Form von Ereignisstrukturen, in denen die Einzelereignisse mit Hilfe von Präzedenz-, Überlappungs- und Teil-von-Operatoren verknüpft werden. Die Merkmale benachbarter Segmente beeinflussen sich gegenseitig und mit Hilfe von Generalisierungs-, Unifikations- und

Spezialisierungsoperationen werden die Merkmalsstrukturen zur Repräsentation von realer Sprache genutzt.

In die Wortrepräsentationen durch Ereignisstrukturen können noch weitere Strukturebenen eingearbeitet werden, z. B. die Silbenstruktur des betrachten Wortes bzw. der betrachteten Wortfolge. Dabei wird eine formale Silbendefinition genutzt:

S-Sequenz	$\rightarrow$	Silbe | Silbe S-Sequenz
Silbe	$\rightarrow$	Ansatz Reim
Ansatz	$\rightarrow$	$C^n, 0 \leq n \leq 3$
Reim	$\rightarrow$	Gipfel Koda
Gipfel	$\rightarrow$	V
Koda	$\rightarrow$	$C^m, 0 \leq m \leq 5$

Merkmalsstrukturen und Silbenstrukturen für sprachliche Äußerungen können verknüpft werden. Diese Repräsentation (vgl. Abbildung 1) kann z. B. für eine lexikalische Strukturierung einer kontinuierlichen Eingangslautfolge eingesetzt werden, unter der Voraussetzung, daß für die verzeichneten Merkmale geeignete Detektoren zu Verfügung stehen, die deren Repräsentanz im Signal ermitteln können.

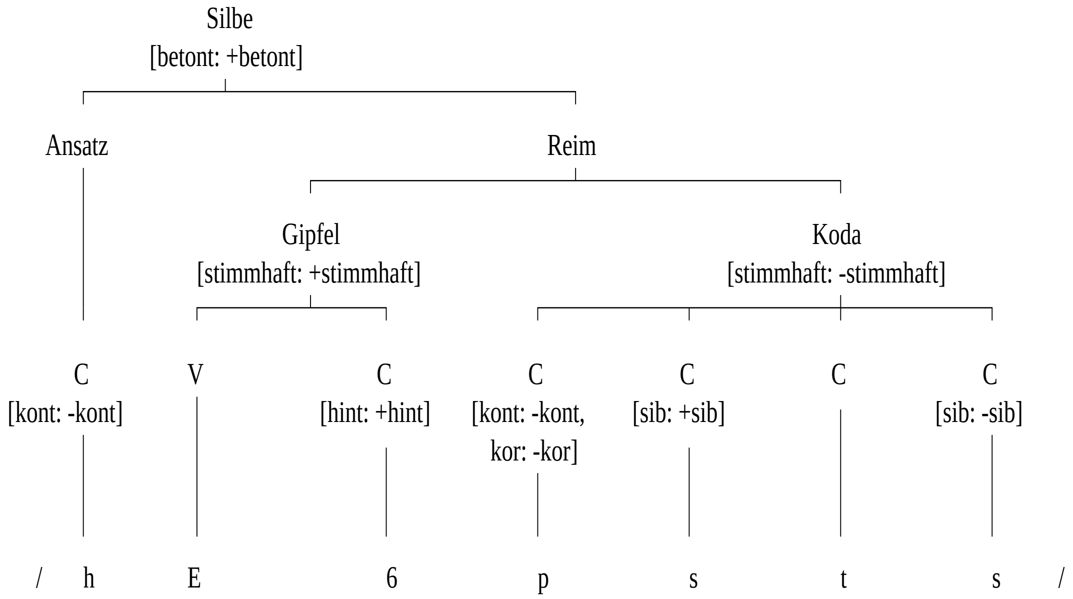


Abbildung 1: Kombination von Merkmalsstruktur und Silbenstruktur

Bisher verwendete Wortbeschreibungen für die Spracherkennung und Sprachsynthese bestehen aus flachen (d. h. linearen) Modellen, deren Elemente durch verschiedene stochastische Verfahren mit Signalparametern und Zeitbereichen verbunden werden.

2.4 Interpretation und Segmentierung von Sprachsignalen

Die wahrgenommenen Schallsignale werden gehört und anschließend in eine symbolische Beschreibung umgewandelt. Dieser Prozeß verläuft nicht als streng gerichteter Prozeß mit harten Entscheidungen. Wenn keine eindeutige Entscheidung möglich ist, müssen nachfolgende Stufen mit mehreren parallelen Hypothesen arbeiten. In integrierten Verarbeitungssystemen erfolgt zudem eine Wechselwirkung aller Ebenen zur Entscheidungsfindung, dabei kommt es zu einer asynchronen Verarbeitung der Eingangsiformation. Diese Erscheinung wird in der humanen Sprachverarbeitung als *Prinzip der verzögerten Bindung* bezeichnet.

Bei der Interpretation und Segmentierung von Sprachsignalen müssen bei der Analyse Invarianzen aus dem Sprachsignal extrahiert werden. Diese Invarianzen werden zur Ableitung robuster Lauthypothesen und Segmentgrenzen verwendet. Bei der Interpretation von Sprachsignalen muß weiterhin beachtet werden, daß die enthaltene Information nicht gleichmäßig über das Signal verteilt ist. Aus der humanen Sprachverarbeitung sind folgende Ankerpunkte für den Verstehensprozeß bekannt:

- betonte Silben
- Wortanfänge und -enden
- Zeitpunkte, an denen sich die spektrale Gestalt des Sprachsignals signifikant ändert.

Segmentierung und suprasegmentale Merkmale Ziel der Segmentierung ist die Zerlegung des Sprachsignals in homogene Bereiche, für die anschließend Merkmale berechnet werden, mit denen die Segmente markiert werden können. Dabei ist eine Segmentierung in mehreren Ebenen zu erwarten, da die einzelnen Merkmale nicht in die gleichen zeitlichen Strukturen abgebildet werden. Es werden merkmalabhängige unterschiedliche Segmentlängen auftreten. Darin kommt die Vielzahl gleichlaufender Prozesse bei der Spracherzeugung zum Ausdruck. Neben den segmentalen Merkmalen sind auch suprasegmentale Merkmale zu finden, die Sprachcharakteristika über Einzelsegmente hinweg beschreiben. Segmentale und suprasegmentale Merkmale (z. B. Segmentlänge und Intonation) werden durch Syntax, Semantik und Sprechstil verändert. Wesentliche Fortschritte bei der Modellbildung könnten durch die Auswertung dieser Größen erzielt werden. Damit wird jedoch die Ebene der Wörter oder Silben als Modellbaustein verlassen, was eine gravierende Zunahme des Modellinventars besonders für die Spracherkennung zur Folge hätte.

Durch die Markierung der Segmente mit speziellen Merkmalen werden für diese Abschnitte phonetische Hypothesen generiert. Die zulässigen Merkmale werden in Form eines geeigneten phonetischen Inventars definiert. Dieses Inventar ist ein Bestandteil des akustisch-phonetischen Wissens des Systems, für das ein leistungsfähiger Repräsentationsformalismus entwickelt werden muß.

akustisch-phonetisches Lexikon Eine Form der Darstellung des akustisch-phonetischen Wissens ist das akustisch-phonetisches Lexikon. Dieses kann Informationen unterschiedlichen Typs enthalten, die durch spezielle Darstellungen repräsentiert werden:

- akustisch → Spektralmuster
- phonemisch → Phoneme
- phonologisch → Wörter

Für alle enthaltenen Repräsentationsformen muß ein speziell optimierter Zugriffsalgorithmus zur Verfügung stehen, der mit mehrdimensionalen Hypothesen arbeiten kann und den optimalen Entscheidungszeitpunkt realisiert. Bei der Entwicklung der akustisch-phonetischen Wissenskomponente müssen somit (nach [vm90]) folgende Probleme bearbeitet werden.

- Wie kann die im Signal enthaltene Information so aufbereitet und repräsentiert werden, daß sie mit Informationen aus anderen Verarbeitungskomponenten sinnvoll kombiniert werden kann?
- Wie kann eine Strategie aussehen, die die Interaktion der akustisch-phonetischen Analyse mit den anderen Verarbeitungskomponenten des Systems steuert?

2.5 Wortuntereinheiten

Bei der Verarbeitung fließend gesprochener Sprache können nicht Sätze oder Wörter als Erkennungseinheiten genutzt werden. Der Grund dafür liegt in der zu großen Anzahl, in der Tatsache, daß im allgemeinen keine Wortpausen detektiert werden können und daß zuviel Trainingsmaterial erforderlich wäre. Daher werden Wortuntereinheiten verwendet. Je kleiner die Wortuntereinheiten gewählt werden, um so häufiger sind sie in einem Trainingskorpus enthalten. Darüberhinaus ist ein auf Wortuntereinheiten basierendes Referenzwissen leichter erweiterbar bzw. an neue Domänen zu adaptieren.

mögliche Wortuntereinheiten

- **Silben:** Sie weisen eine feste Konsonant-Vokal-Struktur auf und beinhalten kontextuelle Information. Die Silbenanzahl im Deutschen beträgt nach [ru84] ca. 5000, im Englischen nach [ai88] ca 10.000.
- **Halbsilben:** Halbsilben entstehen aus Silben durch eine Aufteilung in der Silbenmitte. Es werden initiale und finale Halbsilben unterschieden (im Deutschen je ca. 1000). Halbsilben beinhalten die stärksten Koartikulationseffekte. Die Koartikulation zwischen aufeinanderfolgenden Halbsilben ist relativ schwach.

- **Konsonantcluster und Vokale:** Eine weitere Reduktion des Beschreibungsinventars ergibt sich durch die Aufteilung von Silben und Halbsilben in initale und finale Konsonantcluster und Vokale. In [sc84] werden 47 Anfangskonsonantencluster, 20 Vokale und 103 bzw. 50 Endkonsonantencluster für die 8000 bzw. die 1001 häufigsten deutschen Wörter angegeben.

- **Laute:** Bei der Verwendung von Lauten als Wortuntereinheiten steht dem Vorteil der geringen Klassenanzahl der Nachteil des Verlustes der Kontextinformation gegenüber. Für Laute können die in der statistischen Erkennung notwendigen klassenspezifischen Parameter aus im Vergleich zu anderen Einheiten relativ kleinenStichproben geschätzt werden. Bei der Verwendung von Lauten können Aussprachevarianten auf der Wortebene gut erfaßt werden.

- **Diphone:** Bei Diphonen werden Lautübergänge mit in einer Einheit erfaßt, damit sind wesentliche Kontexteinflüsse im Modell enthalten. Das Diphoninventar einer Sprache wird mit 1000-2000 angesetzt.

- **Triphone:** Bei den Triphonen wird der rechte und linke Kontext eines Lautes im Modell erfaßt. Problematisch ist die unterschiedliche Verteilung der möglichen Triphone, daraus könnten Probleme beim Anlernen der Modelle entstehen. Diese können durch den Einsatz verallgemeinerter Triphone [le89] umgangen werden.

- **phonetische Kategorien:** Grobe phonetische Kategorien, deren Anzahl noch geringer als die Lautanzahl ist, können für eine Präselektion vom Wortkandidaten [fi88] oder als Grundstufe einer hierarchischen Segmentierung verwendet werden ([di77], [lo80]).

Artikulationsart	Charakteristik	Beispiel
Plosive	Verschlußlaute, Stops	*[t],[d]*
Frikative	Reibelaute, Spiranten	*[f],[s]*
Affrikate	sich öffnender Verschluß	*[dS]*
Nasale	Verschluß mit geöffnetem Velum)	*[m]*
Liquide	weniger Behinderung des Luft- -stroms als bei Frikativen	*[l], [r]*
Halbvokale		*[j]*
Vokale		*[a], [u]*
	Kieferstellung	hoch, mittel, tief
	Zungenstellung	vorn, zentral, hinten
	Lippenstellung	rund, gespannt nicht-rund

Tabelle 2: Liste der Artikulationsarten

<table>
<tr><td>kontinuierlich:</td><td>-kont</td><td>+kont</td><td>-kont</td><td>+kont</td><td>-kont</td><td>+ kont</td></tr>
<tr><td>stimmhaft:</td><td>-st</td><td>+st</td><td colspan="4">-st</td></tr>
<tr><td>sibilant:</td><td colspan="3">-sib</td><td>+sib</td><td>-sib</td><td>+sib</td></tr>
<tr><td>koronal:</td><td colspan="3">-kor</td><td colspan="3">+kor</td></tr>
<tr><td></td><td>h</td><td>E 6</td><td>p</td><td>s</td><td>t</td><td>s</td></tr>
</table>

Tabelle 3: Beispiel für eine phonologische Struktur

3 Lauteinheiten und Spracherkennung

Bei der akustischen Spracherkennung muß das Problem der Zuordnung eines kontinuierlichen physikalischen Signals zu einer Symbolfolge gelöst werden. Auf die dabei auftretenden Schwierigkeiten wurden bereits im Abschnitt 2.1 hingewiesen. Der Ablauf der akustischen Erkennung ist im Schema 2 abgebildet. Mit Hilfe spezieller Merkmaldetektoren werden aus dem akustischen Signal relevante Merkmale extrahiert. Die am meisten verbreiteten und leistungsfähigsten Verfahren nutzen dabei spektrale Eigenschaften des Sprachsignals. Es werden Spektral- bzw. Cepstralkoeffizienten für definierte Zeitfenster und Spektralbänder bestimmt. Diese bilden eine das Sprachsignal beschreibende Merkmalvektorfolge. Auf diese Merkmalvektorfolge werden im Anschluß höhere (im Sinne der Verarbeitungshierarchie) Einheiten abgebildet. Diese Abbildung erfolgt meist in einem mehrstufigen Prozeß, wobei zunächst Bewertungen für Lauteinheiten über der Merkmalvektorfolge berechnet werden und danach Bewertungen für Worthypothesen über der Menge von Lauteinheithypothesen. Die Bewertungsberechnung wird meist durch die Berechnung von Abständen zu systeminternen numerischen bzw. symbolischen Referenzmustern realisiert. Der Output eines solchen akustischen Erkenners ist damit eine Menge von Worthypothesen, die entsprechend ihrer Ähnlichkeit zum akustischen Signal zeitlich fixiert und bewertet sind. Diese Hypothesenmenge wird als Wortgraph bezeichnet und bildet die Eingangsinformation für die nachfolgenden Verarbeitungsstufen, die eine Interpretion und Transformation des ursprünglichen Signals realisieren. Daraus ist ersichtlich, daß die höheren Verarbeitungsebenen in bisher realisierten Systemen nur bis zur Wortebene greifen können und andererseits die akustischen Erkennungssysteme eine Entscheidung für Worthypothesen treffen müssen.

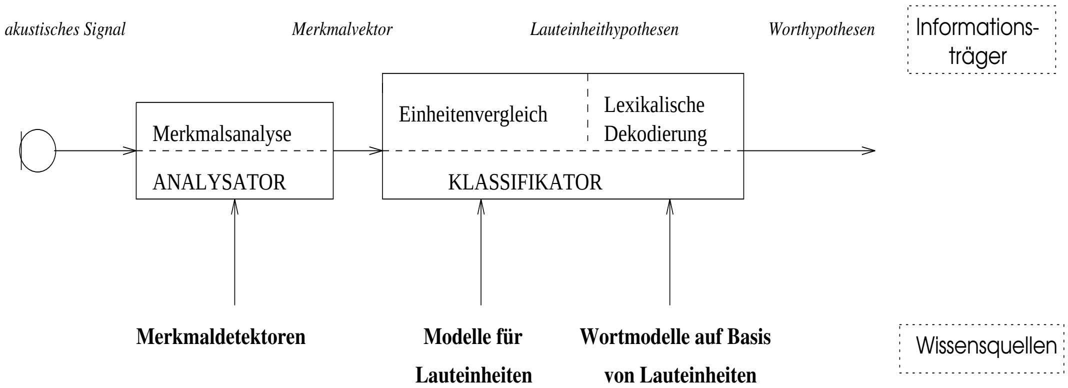


Abbildung 2: Schema der akustischen Spracherkennung

3.1 Probleme bei der Interpretation von Sprachsignalen

Wie schon erwähnt, besteht das Problem der akustischen Erkennung in der Merkmalsextraktion und der Merkmalsvektorklassifikation. Besondere Schwierigkeiten resultieren dabei aus folgenden Eigenschaften des zu verarbeitenden Eingangssignals:

- Die Schalldruck-Zeit-Funktion fließend gesprochener Sprache enthält keine sicheren Markierungen für Wortgrenzen.

- Einzelne Sprachbausteine werden in fließender akzentuierter Sprache verändert. Die Koartikulation finden wir in jedem Fall auf der Lautebene, aber auch innerhalb größerer Segmente bei spezieller Akzentuierung im Satzkontext. Hier sind syntaktische, semantische und sprechstilabhängige Faktoren zu berücksichtigen.

- Gesprochene Sprache ist sehr variantenreich. Neben den inter- und intraindividuellen Varianten kommen noch Veränderungen z. B. beim Flüstern, Singen oder bei Heiserkeit vor, die in den Referenzmodellen mit erfaßt werden müßten.

- Bei gesprochener Sprache treten mehrere Informationsebenen gleichzeitig auf. So sind neben Syntax- und Semantikeinflüssen auf die Intonation auch noch Auswirkungen des Geschlechts, der Identität und der seelischen Verfassung des Sprechers zu beobachten.

3.2 Bildung höherer Einheiten

Zur Abbildung höherer Einheiten auf die Merkmalvektorfolge kommen stochastische, wissensbasierte oder konnektionistische Verfahren zum Einsatz. Mit Hilfe dieser Verfahren wird die Ähnlichkeit von Abschnitten der aktuellen Merkmalvektorfolge zu internen Referenzmustern der gesuchten höheren Einheiten berechnet. Die Entscheidung über die Darstellung eines Abschnitts der Merkmalvektorfolge durch ein oder mehrere Elemente der Referenzelementemenge wird anhand der Ähnlichkeit getroffen.

3.2.1 Stochastische Verfahren

Das am weitesten verbreitete Verfahren der stochstischen Klassifikation sind die Hidden-Markov-Modelle (HMM). Die Referenzmuster werden hierbei durch Referenzmodelle in Form endlicher stochastischer Automaten repräsentiert. Bei der Suche nach einem geeigneten Referenzmuster für einen Abschnitt der Merkmalvektorfolge werden diese Modelle durchlaufen und die dabei von ihnen erzeugte Ausgabe wird mit dem Testmuster (Merkmalvektorfolgenabschnitt) verglichen. Entschieden wird für das (bzw. die) Modell(e), das die ähnlichste Folge produziert hat. Die Grundlage der stochastischen Verfahren bildet die Bayes'sche Entscheidungstheorie:

$$P(W \mid A) = \frac{P(W)P(A \mid W)}{P(A)}$$

mit	P(W)	Sprachmodell
	P(A|W)	akustisches Modell
	P(A)	Wahrscheinlichkeit des Sprachsignals

Das akustische Modell des Systems besteht aus einem HMM für jede zu erkennende akustische Einheit. Dieses Modell besteht aus N Knoten (denen eine feste Zeit zugeordnet wird) und gerichteten und Kanten zwischen diesen Knoten. Die Kanten sind mit einer Übergangswahrscheinlichkeit gekennzeichnet, wobei die Summe der Übergangswahrscheinlichkeiten für alle von einem Knoten wegführenden Kanten 1 ist. Beim Durchlaufen einer Kante emittiert das Modell Wahrscheinlichkeiten für das Auftreten der Symbole (diskrete HMM) bzw. der Werte (kontinuierliche HMM)des Definitionsbereiches.

Modellierungseinheiten für stochastische Verfahren. Die kleinsten sprachlichen Einheiten, die primär vom System erkannt werden sollen, stellen die Modellierungseinheiten dar. Generell wird zwischen *akustisch motivierten* und *phonetisch motivierten* Einheiten unterschieden. Bei der Auswahl der Modellierungseinheiten muß darauf geachtet werden, daß sie hinreichend gut modelliert werden können. Dabei ist zu beachten, daß sehr kleine Einheiten zwar häufig in einem Trainingskorpus auftreten, dafür aber stärker durch den umgebenden Kontext beeinflußt werden. Als mögliche Modellierungseinheiten kommen in Betracht:

kontextunabhängige Phoneme

Bei der Verwendung kontextunabhängiger Phoneme hat man den Vorteil eines sehr kleinen Inventars (ca. 50), jedoch den Nachteil des kontextbedingten Variantenreichtums. Für diese Modellierungseinheiten erfolgt die Worträpresentation im Lexikon durch Folgen oder Netze von Phonemmodellen. Bei der Verwendung zyklischer Modelle ([me87])könnte auf das Lexikon verzichtet werden, was jedoch ein beträchtliches Anwachsen möglicher Hypothesenmengen zur Folge hätte.

Diphone

Ein Diphon ist eine phonetische Einheit zwischen zwei aufeinanderfolgenden Lautmitten. Bei steigender Anzahl gegenüber kontextunabhängigen Phonemen bieten sie den Vorteil der Modellierbarkeit von Lautübergängen ([cr73]).

kontextabhängige Phoneme

Bei kontextabhängigen Phonemmodellen wird der linke und rechte Kontext mit im Modell erfaßt. Die so entstehenden Triphone (Anzahl rein kombinatorisch 50^3; jedoch durch phonetische Regeln einschränkbar) sind jedoch u. U. sehr schwer zu modellieren, da sie eine sehr variable Verteilung in den Trainingskorpora aufweisen.

wortabhängige Phoneme

Für diese Modellierungseinheiten werden bestimmte Phonemmodelle extra im Wortkontext trainiert. Dieses Verfahren kann für ausgewählte Wortklassen, die in hoher Frequenz im Korpus auftreten genutzt werden. Im Spracherkennungssystem SPHINX wurde dieses Verfahren für Funktionswörter eingesetzt ([le88]).

Silben und Halbsilben

Silben stellen eine akustisch-perzeptiv wichtige Einheit bei der Sprachverarbeitung dar, da sie bei noch handhabbarer Anzahl Kontexteffekte weitgehend beinhalten. Das Aufwandsproblem kann reduziert werden, wenn Halbsilben (Unterteilung in anlautende und auslautende Hälfte) verwendet werden ([ru89]).

Segmente

Segmente haben keine unmittelbare Entsprechung, die durch die Phonetik-Phonologie beschrieben wird. Sie werden nach rein technischen Gesichtspunkten ausgewählt. Die Segmentierung erfolgt z. B. nach dem Maximum-Likelihood-Prinzip. Die Segmentgrenzen werden bei starken lokalen Diskontinuitäten im Signal gesetzt.

Wörter

Bei der Modellierung ganzer Wörter treten eine Reihe von Problemen auf. Zunächst müssen relativ große Modelle vorgesehen werden (50 Zustände bei Zeitfenstern von 10ms für ein 500ms langes Wort), reduzierte Modelle wurden in [ra85] untersucht. Um diese Modelle gut zu trainieren, muß der Trainingskorpus genügend viele Varianten für jedes Wort enthalten. Weiterhin muß für jedes neue Wort in der Erkennerdomäne ein erneutes Training realisiert werden. Genügend gut trainierte Wortmodelle weisen jedoch eine hohe Leistungsfähigkeit auf, und sind für Erkennungsaufgaben mit sehr geringem, konstanten Wortschatz oder als Ergänzung für spezielle Wörter anwendbar.

3.2.2 Wissensbasierte Verfahren

Wissensbasierte Verfahren werden zur Interpretation der vom Objekt abgenommenen Merkmale eingesetzt. Das Ziel besteht in einer abstrakten Beschreibung des Objektes. Dazu muß

- eine Wissensbasis (Regeln)
- eine Faktenbasis und
- eine Kontrollkomponente

zur Verfügung stehen. Wissensbasierte Verfahren sind dem humanen Perzeptionsverhalten nachempfunden. Sie realisieren eine kontextabhängige Merkmalinterpretation, wobei sowohl binäre Merkmale (Merkmal vorhanden vs. Merkmal nicht vorhanden) als auch qualitative Merkmale verarbeitet werden. Beim Verarbeitungsprozeß werden mehrere Hypothesen parallel betrachtet, die sich gegenseitig unterstützen, sich widersprechen oder einander ausschließen können. Die Wertigkeit einzelner Merkmale kann unterschiedlich sein, besonders sicher detektierte Merkmale können zum Ausschluß anderer Hypothesen führen.

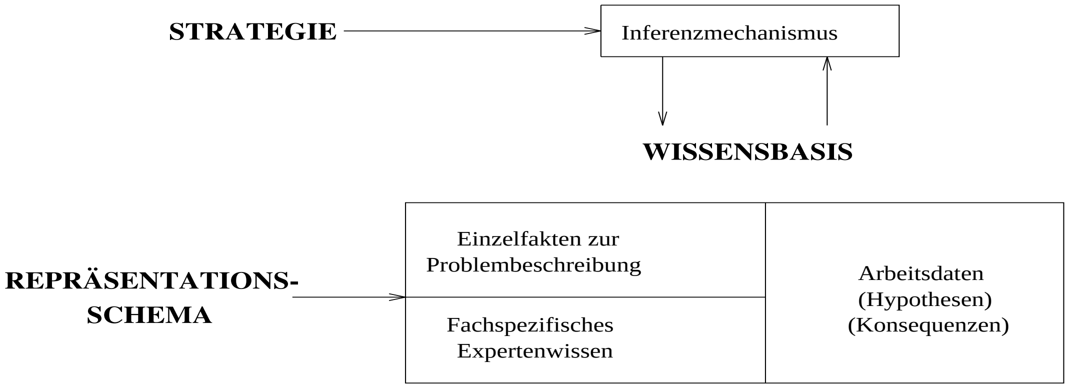


Abbildung 3: Grundstruktur eines wissensbasierten Systems nach [sa90]

Bei wissensbasierten Verfahren in Spracherkennungssystemen findet man meist eine Kombination von Sprachsignalverarbeitung und Verarbeitung natürlicher Sprache. Problematisch ist jedoch, daß nur durch ein geschlossenes Modell des Gesamtprozesses und eine leistungsfähige Kontroll- und Steuerstrategie die inherente Mehrdeutigkeit von Teilhypothesen auf mehreren Ebenen zu beherrschen ist. Wissensbasierte Systeme in der Erkennung und Verarbeitung natürlicher, gesprochener Sprache sind bisher nur in Ansätzen realisiert, obwohl sie durch Modularisierung eine gute Möglichkeit zur Implementierung neuen Wissens in Teilgebieten ermöglichen.

3.2.3 Konnektionistische Verfahren

Konnektionistische Verfahren bzw. Neuronale Netze sind aufgrund ihrer Adaptionsfähigkeit besonders für Mustererkennungsprobleme geeignet. Ein neuronales Netz besteht aus mehreren Neuronen, die untereinander und mit der Außenwelt auf bestimmte, in der *Konnektionsmatrix* beschriebene Art verbunden sind. Intern können die Neuronen in Schichten angeordnet sein (hierarchische Netze). Die einzelenen Neuronen sind durch Aktivierungszustände, die diskret oder kontinuierlich sein können und durch Aktivierungsfunktionen charakterisiert. Als Aktivierungsfunktionen kommen deterministische oder stochastische Funktionen in Betracht. Weiterhin werden für Neuronale Netz Lernregeln definiert. Diese beschreiben die Anfangsgewichte der einzelnen Neuronen und die Veränderung der

Gewichte bei der Verarbeitung der Lernmuster. Für den Lernvorgang, der überwacht oder nicht überwacht ablaufen kann, wird ein geeignetes Abbruchkriterium definiert.

Man unterscheidet zwischen mehreren Netzgrundtypen (Perceptron, Thermodynamische Modelle, Hopfield-Netze und Mermalskarten), die sich durch Modellierungsgüte und Verarbeitungszeit unterscheiden. In der Sprachsignalerkennung werden meist Peceptrons für verschiedene Modellierungseinheiten eingesetzt. Die hier zu findenden Modellierungseinheiten entsprechen den unter 3.2.1 erwähnten.

Für einen effizienten Einsatz in der Spracherkennung muß das Problem der Behandlung von Zeitstrukturen noch extra betrachtet werden, da herkömmliche Neuronale Netze nur für statische Muster optimal sind. Folgende Verfahren kommen dabei zum Einsatz:

- *Fixed Length Time Compression*

  Dabei wird das längste Wort als Referenz verwendet und kürzere Wörter werden durch Pausen auf dessen Dauer verlängert. Für jeden Merkmalvektor (evtl. quantisiert) des längsten Wortes steht ein Eingangsknoten zur Verfügung.

- *Time Delay Neural Network*

  Dieser Ansatz ([wa87]) beschreibt ein multi-layer-Perzeptron mit vier Schichten und 16 Eingangsneuronen. An diese wird für jede Zeitscheibe je ein Cepstralkoeffizient angelegt. Die Neuronen sind je über drei Synapsen mit den Neuronen der darunterliegenden Schicht verknüpft. Diese Synapsen repräsentieren jeweils den Koeffizienten der aktuellen und der beiden vorangegangenen Zeitscheiben. Daurch wird die zeitliche Abhängigkeit der Cepstralkoeffizientvektoren über mehrere Zeitscheiben erfaßt.

4 Kontextabhängige Lauteinheiten

Bei der Festlegung einer geeigneten Erkennungseinheit für gesprochene Sprache ergeben sich durch die Kontinuität des Signals Probleme. Als Erkennungseinheit soll der Teil des Signals verstanden werden, der als eigene Einheit klassifiziert wird. An diese Einheiten werden im Sinne guter Erkennungsergebnisse folgende Ansprüche gestellt:

- geringe Klassenanzahl der gewählten Einheiten
- gute Segmentierbarkeit des Signals entsprechend der gewählten Einheiten
- geringe Koartikulation zwischen den einzelnen Einheiten

Diese Forderungen werden nicht gleich gut von möglichen Erkennungseinheiten (z. B. Phoneme, Silben, Wörter) erfüllt. Meist muß ein Kompromiß eingegangen werden. Im Folgenden werden die grundlegend unterschiedlichen Erkennungseinheitenklassen, die *kontextabhängigen* Lauteinheiten und die *silbenorientierten* Lauteinheiten vorgestellt.

4.1 Auswahl von Lauteinheiten

Unsere Untersuchungen gehen zunächst davon aus, daß die natürliche Hierarchie der Spracheinheiten zur Segmentierung in Lauteinheiten genutzt wird. Das bedeutet, es werden Signalabschnitte, die den Sprachlauten entsprechen, als Wortuntereinheiten gewählt. Die eigentliche Problematik besteht nun darin, diese Signalabschnitte im Sprachsignal sicher zu detektieren und zu klassifizieren.

Bei Spracherkennungssystemen für sehr große Wortschätze wird vorgeschlagen, die Erkennung mehrstufig zu realisieren. Es wird zunächst eine Groberkennung mit wenigen Phonemklassen vorgenommen, um aus dem Gesamtwortschatz eine Untermenge der möglichen Erkennungsergebnisse zu extrahieren. In [hu84] werden dazu 6 Phonemklassen verwendet:

- Vokale
- Explosivlaute
- Nasale
- Liquide
- stimmlose Frikative
- stimmhafte Frikative

Bei der Anwendung auf ein 20.000 englische Wörter umfassendes Lexikon entstanden Wortgruppen (= Wörter, die durch die gleiche Phonemklassenfolge beschrieben werden) mit durchschnittlich 25 Wörtern. Der maximale Umfang einer

Wortgruppe lag bei 200 Wörtern, ein Drittel der Wörter wurde durch die Phonenklassenfolge eindeutig beschrieben.

Diese phonetischen Klassen haben den Vorteil, daß sie im Signal schnell und relativ sicher zu detektieren sind. Sie ermöglichen aber keine eindeutige Worterkennung, so daß die entstehenden Wortgruppen im Rahmen einer feineren zweiten Erkennerstufe noch diskriminiert werden müssen.

Bei der Auswahl geeigneter Lauteinheiten könnte dieses Konzept genutzt werden, um ein Lautgruppeninventar zu definieren, das bei guter Detektion eine genügend gute Diskrimination des Wortschatzes ermöglicht. Im Folgenden werden Untersuchungen dazu vorgestellt. Die angegebenen Phonemgruppeninventare wurden durch die Auswertung von Verwechslungsmatrizen gewonnen ([we94c]).

4.1.1 Vorverarbeitung des Signals und Klassifizierung von Signelabschnitten

Die Segmentierung des Signals erfolgt durch ein Vorverarbeitungssystem mit modularer hierarchischer Grundstruktur. Die Signalvorverarbeitung wird durch eine 8-Kanal-Filterbank (Akustikprozessor) realisiert. Für 2.25 ms lange Signalabschnitte wird ein Merkmalvektor erzeugt, anschließend erfolgt eine Glättung über 4 Merkmalvektoren. Es folgt eine Vektortransformation mit tabellierten Dichtefunktionen, bei der für jeden Vektor eine Zugehörigkeit zu 64 vordefinierten Phonemklassen ermittelt wird. Diese 64-dimensionalen Zugehörigkeitsvektoren werden anschließend durch eine Matrixmultiplikation auf n-dimensionale Vektoren abgebildet, die die Zugehörigkeit zu n Phonemgruppen repräsentieren (vgl. Phonemgruppendefinitionen in Anhang A und Anhang B).

4.1.2 Segmentierung des Signals

Aus diesen zeitdiskreten Zugehörigkeitsvektoren werden symbolische Phonemhypothesen gebildet. Eine symbolische Hypothese charakterisiert einen Signalabschnitt, der durch einen Start- und Endzeitpunkt, einen Namen (Label zur Lautklassenkennzeichnung) und eine Bewertung, die die Ausprägung der Hypothese beschreibt, gekennzeichnet ist. Die symbolischen Hypothesen werden durch Auswertung der Phonemgruppenzugehörigkeiten ermittelt. Wird im Signal über einen Mindestzeitraum die gleiche Phonemgruppe mit einer definierten Mindestbewertung detektiert, wird für diesen Zeitraum eine symbolische Hypothese generiert, deren Bewertung dem Mittelwert der Einzelbewertungen entspricht.

4.2 Beurteilung der Lauteinheiten

Eine Möglichkeit der Beurteilung der gewählten Lauteinheiten besteht in der Untersuchung der durch sie gewährleisteten Wortdiskrimination innerhalb eines konkreten Wortschatzes. Als Wortuntereinheiten wurden Gruppen über der Menge der Sprachlautbeschreibungen, die in dem zur Verfügung stehenden feingelabelten Datenmaterial (PHONDAT 2, vgl. [pd92]) enthalten sind, gebildet. Es wurden zwei Gruppenbildungen untersucht (siehe Tabellen in Anhang A und Anhang B). Diese Gruppenbildungen ermöglichen über einem Korpus ausgewählter spontansprachlicher Dialoge (vgl. Transliterationen in cdrom_v1.tar auf dem Verbmobil-Server; 2438 Vollformen) die in den Tabellen 4 und 5 angegebenen Wortdiskriminationen. Darunter soll verstanden werden, wieviele Gruppen von n nicht mehr unterscheidbaren Wörtern über dem Korpus durch Anwendung einer Labelgruppierung entstehen.

Gruppengröße	n=1	n=2	n=3	n=4
Anzahl	2246	65	14	5

Tabelle 4: Wortdiskrimination für Labelgruppenbildung I

Gruppengröße	n=1	n=2	n=3	n=4	n=5	n=6	n=7	n=8	n=9
Anzahl	1975	129	23	18	7	1	2	0	1

Tabelle 5: Wortdiskrimination für Labelgruppenbildung II

Die Ergebnisse bestätigen die Erwartung, daß mit einem Beschreibungsinventar von 29 Labelgruppen noch eine fast vollständige Diskrimination des Korpus möglich ist. Ca. 92% der Wörter des Wortschatzes werden noch eindeutig beschrieben. Die in den Zweiergruppen enthaltenen, nicht mehr unterscheidbaren Wörter, sind größtenteils phonetisch identisch, d. h. es handelt sich um Groß- und Kleinschreibung eines Wortes oder um verschiedene gleichlautende orthographische Varianten (Homophone) wie "Gayst"(als Eigenname) und Geist oder Müller und Mueller. Weitere Zusammenfassungen entstehen durch die Einbeziehung von Aussprachevarianten, wodurch z. B. "sind" und "Sinn" gleichlautend werden.

Bei Verwendung eines Beschreibungsinventars von 11 Labelgruppen ergibt sich noch eine gute Diskrimination (ca. 81% des Wortschatzes wird eindeutig beschrieben) Die entstehenden Gruppen nicht mehr unterscheidbarer Wörter könnten u. U. zu Problemen bei der weiteren Verarbeitung führen, da durch die grobe Lautgruppierung **und** die Einbeziehung von Aussprachevarianten z. B. folgende Gruppe entsteht:

Standard	Variante	Transkription	Grobbeschreibung
fange	fang	faN	*AN
haben	ham	ham	*AN

Die bisher verwendeten Beschreibungen für Lauteinheiten in Sprachlautgröße können auch zur Modellierung größerer Signalabschnitte (Halbsilben, Silben) genutzt werden. Dabei können weitere kontextabhängige Beschreibungen, wie z.

B. spezielle Lautübergänge und eine Silbenkernmarkierung einbezogen werden. Untersuchungen dazu wurden durchgeführt und sind in [je93] beschrieben.

4.2.1 Korpusuntersuchungen

Als Zusatzinformationen, die zur Wichtung der bei der Erkennung entstehenden Hypothesenmengen von Bedeutung sind, können Korpusbesonderheiten herangezogen werden, z. B. welche Wörter als Bestandteile in anderen enthalten sind. Die daraus resultierende lexikalische Ambiguität führt zur einer Hypothesenanhäufung, die durch Wissen über die Wortverdeckung wieder abgebaut werden kann. Nach [fr90] kann als allgemeine Information ermittelt werden, welche Anzahl Wörter in anderen an unterschiedlichen Positionen enthalten ist (siehe Tabelle 6). Speziell kann auch für jedes Wort die Liste seiner "Wortbestandteile" abgefordert werden.

Die Tabelle zeigt den Aufbau des oben erwähnten Korpus bezüglich der enthalten Wörter einer bestimmten Länge von n Phonemen und die im Korpus auftretende Wortverdeckung. Dabei ist zeilenweise für die Wörter der Länge n angegeben (absolut und prozentual), wie viele von ihnen in anderen, längeren Wörtern am Anfang, in der Mitte bzw. am Ende vorkommen. Die Wortüberdeckung im Korpus muß bei der Auflösung von mehrdeutigen Hypothesen mit berücksichtigt werden.

WL[a]	ANZ[b]	ANFANG	Anteil in %	MITTE	Anteil in %	HINTEN	Anteil in %
0	1	1	1.000	0	0.000	0	0.000
1	13	10	0.769	12	0.923	11	0.846
2	66	56	0.848	61	0.924	50	0.758
3	146	108	0.740	89	0.610	56	0.384
4	266	132	0.496	78	0.293	71	0.267
5	307	87	0.283	44	0.143	56	0.182
6	343	68	0.198	19	0.055	60	0.175
7	233	36	0.155	12	0.052	32	0.137
8	220	37	0.168	2	0.009	19	0.086
9	220	25	0.114	1	0.005	11	0.050
10	155	26	0.168	1	0.006	8	0.052
11	127	11	0.087	1	0.008	3	0.024
12	95	8	0.084	1	0.011	2	0.021
13	59	1	0.017	0	0.000	0	0.000
14	48	3	0.062	0	0.000	0	0.000
15	31	2	0.065	0	0.000	0	0.000
16	26	7	0.269	0	0.000	0	0.000
17	17	5	0.294	0	0.000	1	0.059
18	20	3	0.150	0	0.000	0	0.000
19	22	6	0.273	0	0.000	0	0.000
20	14	2	0.143	0	0.000	0	0.000
21	6	1	0.167	0	0.000	0	0.000
22	2	1	0.500	0	0.000	0	0.000
23	1	0	0.000	0	0.000	0	0.000

[a] Wortlänge in Phonemen
[b] Wortanzahl der Länge n

Tabelle 6: Wortüberdeckungen im Korpus “Spontane Dialoge”

5 Silbenorientierte Lauteinheiten

In Abschnitt 4 wurden die Ansprüche an Erkennungseinheiten formuliert. Nach den Ausführungen zu den *kontextabhängigen* Lauteinheiten werden im Folgenden die *silbenorientierten* Lauteinheiten vorgestellt.

5.1 Erkennungseinheit Sprechsilbe

Bei der automatischen Spracherkennung spielt die Wortebene eine sehr wichtige Rolle, da die Wörter sowohl eine grammatische als auch eine semantische Funktion als Bedeutungsträger haben. Wörter sind jedoch neben ihrer großen Anzahl auch durch ihre schlechte Detektierbarkeit nicht gut als Basiserkennungseinheiten für ein Spracherkennungssystem geeignet. Bei der Untersuchung gesprochener Sprache ergibt sich jedoch die *Sprechsilbe* als eine im Kommunikationsakt relativ sicher identifizierbare Einheit (prosodische Einheit). Sie ist in der Sprachäußerung separierbar, ist dem Sprecher intuitiv vertraut und bietet auch eine Verbindung zur Linguistik als Baustein für die Wörter der verwendeten Sprache. Die Eignung der Silbe als Erkennungseinheit wird durch folgende Erkenntnisse unterstützt:

- **Artikulatorisch**, bei der Sprachproduktion spielt die Sprechsilbe eine wesentliche Rolle. Die Grundgeste beim Sprechen besteht im Öffnen und Schließen bzw. Verengen des Artikulationstraktes. Die dabei entstehenden Intervalle können den einzelnen Silben zugeordnet werden. Die Silbe ist damit durch ein Ansteigen und Abfallen der Sonorität gekennzeichnet, was ihre Detektion in dem kontinuierlichen Sprachsignal ermöglicht. Die Bedeutung der Sprechsilbe als artikulatorische Einheit kommt auch bei einer Betrachtung der Entwicklung des Sprechens beim Menschen zum Ausdruck. So werden zunächst Silbeninventare zur Artikulation benutzt, die Auflösung in darunterliegende Segmente erfolgt erst zu einem späteren Zeitpunkt.

- **Perzeptiv** wird die Sprechsilbe als Gestalt, als Ganzes wahrgenommen und ist mehr als die Summe ihrer Teile. Die Silbe existiert vor den isolierten Teilen, diese sind nur nach Silbenwahrnehmung durch Abstraktion als Laute zu gewinnen [or80]. Auch TILLMANN verweist darauf, daß eine Silbe nicht als Summe, Gruppe oder Sequenz von Lauten aufgefaßt werden kann, sondern ein sprachliches Phänomen eigener Art ist, das vom Einzellaut nicht gefaßt werden kann [ti64]. Die Ursache für die geschlossene Silbenwahrnehmung liegt in der Koartikulation und Steuerung bei der Produktion von lautsprachlichen Äußerungen. Die Lautstellungen werden nicht nacheinander eingenommen, die Bewegungen der Artikulationsorgane sind teilweise simultan und ineinander verschränkt [me33]. In [st79] wird die Vorstellung entwickelt, daß eine Silbensegmentierung eher in einer peripheren Stufe der Wahrnehmung erfolgt, Während die Segmentierung in einzelne Phoneme den höheren Verarbeitungsprozessen des Gehirns zuzuschreiben ist und damit letztlich ein Ergebnis der Klassifizierung ist.

- **Praktisch realisierbar** ist die Detektion der Silbe in dem kontinuierlichen Sprachsignal aufgrund des durch sie verkörperten Sonoritätsanstiegs. Dadurch kann die psychoakustische Empfindungsgröße Lautheit zur Segmentierung genutzt werden. Weiterhin ist der vokalische Kern aufgrund seiner spektralen Besonderheiten gut zu erkennen. Silbengrenzen sind durch relative Minima der Lautheit markiert, die exakte Entscheidung über die Lage der Silbengrenze kann unter Einbezug der Klassifikationsergebnisse getroffen werden.

5.2 Modellierung von Sprechsilben

Sprechsilben können als geschlossene Einheiten modelliert werden, bzw. sie können noch in disjunkte kleinere Einheiten zerlegt werden. Die Entscheidung darüber wird meist vom verwendeten Erkennungsverfahren abhängig gemacht. Für stochstische Erkennungsverfahren stellt die Gesamtanzahl von Sprechsilben meist noch ein zu großes Inventar dar, da ein umfangreicher Lernprozeß erforderlich ist, für den genügend Trainingsmaterial zur Verfügung stehen muß. Für Erkennungsverfahren, die die Silbenmodelle als Verknüpfungsvorschriften für Beschreibungen kleinerer Signalabschnitte verwenden, werden nur formale Modelle benötigt, so daß die Anzahl kein Problem darstellt (vgl. Abbildung 4 und 5).

5.3 Silbenmodelle für stochastische Erkennungssysteme

Für stochastische Erkennungssysteme werden die Modelle für die Erkennungseinheiten als Hidden-Markov-Modelle, bestehend aus einer Menge von Zuständen, Übergängen zwischen den Zuständen, sowie Schhätzungen für Übergangs- und Emissionswahrscheinlichkeiten realisiert. Die Anzahl der Zustände repräsentiert die Struktur der zu modellierenden Einheit (die Zustandsanzahl könnte z. B. entsprechend der mittleren Frameanzahl der Modelleinheiten gewählt werden), die zugelassenen Übergänge repräsentieren Zustandsabfolgen, sie können damit zur Modellierung von Reduktionen innerhalb eines Modell dienen. Um leistungsfähige Modelle für Sprechsilben trainieren zu können, müßten im Trainingsmaterial alle möglichen Sprechsilbenvarianten mit ihren realen Auftretenswahrscheinlichkeiten enthalten sein.

Das Silbeninventar im Deutschen. Bei der Abschätzung des Silbeninventars einer Sprache kann entweder die Zählmethode angewendet werden (z. B. die Kaeding-Zählung für das Deutsche) oder es können Silbenteile betrachtet und verknüpft werden. Nach der Kaeding-Zählung treten in den 7995 hochfrequenten deutschen Wortformen 2614 Sprechsilbentypen auf. Bei einer Zerlegung der Silben ergeben sich:

- 50 verschiedene Silbenansätze
- 20 verschiedene Gipfel
- 160 verschiedene Kodas

Das ergibt $50 * 20 * 160 = 160.000$ Silben. Aus phonotaktischen Einschränkungen für die Kombination von Gipfel und Koda resultiert jedoch eine geringere Anzahl. Ca. 8000 Silben sind in existierenden Wörtern lexikalisiert. Problematisch bei der Abschätzung und Modellierung des Sprechsilbeninventars sind die durch Koartikulation und Steuerung bedingten Realisierungsvarianten. Sie führen zum Beispiel zu Silben, die keinen Vokal mehr enthalten, sondern silbische Konsonanten. Die Aufteilung eines Wortes in Sprechsilben ist im realsprachlichen Fall ebenfalls teilweise problematisch und muß außer von der orthographisch-phonologischen Seite auch von der signalanalytischen Seite betrachtet werden.

Aus der Betrachtung dieses Intentars wird deutlich, daß sowohl umfangmäßig als auch inhaltlich der Aufwand zum Training von Ganzsilbenmodellen zu hoch ist. Hinzu kommt, daß reale Trainingskorpora nicht die erforderlichen Datenmengen enthalten. Einen Ausweg aus dieser Situation bietet die Aufspaltung der Modelle in Silbenteilmodelle.

5.3.1 Silbenteilmodelle

In [ru91] wird dazu folgender Ansatz vorgestellt. Die Silben werden in ihre perzeptiv und produktiv wesentlichsten Bestandteile zerlegt:

- die Anfangskonsonantenfolge
- den Vokal bzw. Diphtong
- das Rudiment
- den Suffix

Durch die Aufteilung der Endkonsonantenfolge in die Bestandteile Rudiment und Suffix sind die Probleme, die aus einer möglichen Reduktion der Endkonsonantenfolge resultieren, besser modellierbar. Da für Rudiment und Suffix auch die leere Kette zulässig ist, ergibt sich ein festes Raster der Erkennungseinheiten:

```
...AKF VOK RUD SUF AKF VOK RUD SUF AKF VOK RUD SUF ...
```

Für die Anfangskonsonantenfolgen, Rudimente und Suffixe im Deutschen wird das in den Tabellen 7, 8 und 9 enthaltene Inventar angegeben (in SAMPA-Notation).
Der so definierte Silbenaufbau ist geeignet zur Beschreibung von Normsprechsilben. Für bestimmte spontansprachliche Besonderheiten, wie den Ausfall des

b	C	d	f	g	h	j	k	l	m	n	p	r	s	S	t	v	x	Z
											pf							
											pfl							
											pfr							
bl			fl	gl			kl				pl			Sl				
														Sm				
				gn			kn							Sn				
														Sp				
														Spl				
														Spr				
br		dr	fr	gr			kr				pr			Sr	tr			
															ts			
															tsv			
														St				
														Str				
									kv					Sv				

```
   Zusaetzlich: Q (Glottislaut)
                / (leere AKF)
```

Tabelle 7: Inventar der Anfangskonsonantenfolgen

Schwa-Lautes und das vokalisierte [*r*] im Silbenauslaut [*er*] muß die Definition noch erweitert werden, um die entstehenden neuen Konsonantenkombinationen, die silbische Konsonanten enthalten, beschreiben zu können. Diese Erweiterung betrifft die Definition der Vokale, sie wird von den Vokalen und Diphtongen ausgehend um die möglichen silbischen Konsonanten erweitert:

```
VOK = { Vokale, vokalisiertes r, Diphtonge, /m./,/n./,/l./,/r./ }
```

Mit dieser Erweiterung ist gewährleistet, daß die oben erwähnte Teilsilbenabfolge in der akustischen Realisierung enthalten ist. Zur Synchronisation wird bei der Erkennung die Detektion des Silbenkerns eingesetzt. Zwischen zwei Silbenkernen wird die im Bild 6 repräsentierte Abfolge von Silbenteilen zur Verkettung der Modelle bei der Erkennung durch den Viterbi-Algorithmus eingehalten. Dabei wird im Hauptpfad die orthographienahe Aussprache durchlaufen, der Nebenpfad repräsentiert das Vorkommen eines silbischen Konsonanten.

Mit dieser Definition steht ein leistungsfähiges Modell zur Verfügung, das mit einem relativ kleinen Inventar eine große Vielfalt an Realisierungsmöglichkeiten von Sprechsilben beschreibt.

```
------------------------
C  k  l  m  n  p  r  N  x
lC lk    lm ln lp
nC
   Nk
rC rk rl rm rn rp
------------------------
Zusatz: \(leeres Rudiment)
```

Tabelle 8: Inventar der Rudimente

```
------------------
f   s   S   t
fs          ts
fst     Sst tst
            tS
            tSst
            tSt
ft  st  St
fts sts
------------------
Zusatz: -(leeres Suffix)
```

Tabelle 9: Inventar der Suffixe

5.4 Silbenmodelle als Verknüpfungsvorschriften

Beim Einsatz der Silbenmodelle als Verknüpfungsvorschriften (vgl. [we94a]) für bewertete Hypothesen unterhalb der Silbenebene kommt es darauf an, möglichst alle Realisierungsvarianten zu erfassen. Die Modelle werden in Form von Folgen oder Netzen bereitgestellt und dienen der Reduktion der Hypothesenanzahl, die auf der Ebene der eigentlichen Erkennungseinheiten erzeugt wurde. Schematisch ist der Einsatz der Modelle im Bild 7 für Wortmodelle dargestellt.

Es zu sehen, daß nur für die Signalabschnitte eine Worthypothese generiert wird, für die im Referenzwissen ein Modell enthalten ist, das genau durch eine reale Hypothesenfolge erfüllt wird. Im Falle geringfügiger Abweichungen von Hypothesenfolge und Modell wird für den gesamten Abschnitt keine Worthypothese erzeugt. Bei der Verknüpfung der Hypothesen für die Erkennungseinheiten werden zeitliche Randbedingungen gesetzt, so daß auch nicht unmittelbar aneinandergrenzende Hypothesen verbunden werden können. Überlappungsbereiche und Lücken sind zulässig.

Da der Einsatz von Wortmodellen zur Verknüpfung der Hypothesen den Nachteil hat, daß bei einer Inventarerweiterung wieder neue Modelle generiert werden müssen, sollen Silbenmodelle zum Einsatz kommen. Die Verwendung von Silben hat zum einen den Vorteil eines begrenzten Inventars und zum anderen sind innerhalb der Sprechsilben die wesentlichsten Koartikulationseffekte erfaßt. Die Silbenmodelle sollen aus einem Text, der die Erkennerdomäne abdeckt, automatisch generiert werden. Dazu wird aus dem eingegebenen Text wird eine Wortliste erzeugt. Zu dieser orthographischen Repräsentation wird die phonetische generiert und in Sprechsilben untergliedert. Für die gefundenen Sprechsilben werden kontextabhängige Realisierungsvarianten generiert.

5.5 Automatische Sprechsilbenzerlegung

Die automatische Sprechsilbenzerlegung fügt in den von der Graphem-Phonem-Umsetzung generierten Phonemfolgen die Sprechsilbengrenzen ein. Die Generierung der Sprechsilbengrenzen erfolgt nach einem Regelwerk, wobei unterschiedlich aufgebaute Eingangsdaten verarbeitet werden können. Die Eingabedaten sind Phonemfolgen, in denen entweder keine, teilweise oder vollständige prosodische Informationen enthalten sind. Das zur Sprechsilbentrennung erarbeitete Regelwerk basiert im Wesentlichen auf Silbentrennregeln aus [du83] und [or80], wobei speziell [du83] stark an den graphemischen Silbenbegriff angelehnt ist und in Hinblick auf die vorliegende Zielstellung einiger Ergänzungen bedurfte. Beim Entwurf des Regelwerkes wurde auf von einander unabhängige, als Teilmodule abarbeitbare Trennregeln orientiert. Die Module unterligen einer logischen Abarbeitungsreihenfolge, wobei einzelne Module aus dieser Reihenfolge gestrichen werden können. Der vorliegende Aufbau der automatischen Sprechsilbenzerlegung ist in Abbildung 8 dargestellt. Im Ausnahmelexikon sind vollständige bzw. teilweise Phonemkettenzerlegungen repräsentiert, die durch die nachfolgenden Regeln nicht abgedeckt werden. Durch die lineare Regelanwendung bedingt, wird zuerst auf das Vorkommen dieser Ausnahmen getestet. Anschließend kann im Falle eines einzelnen Konsonaten zwischen zwei Vokalkernen, wobei der erste Vokal ein Kurzvokal ist, dieser Konsonant verdoppelt werden und somit auf die beiden entstehenden Sprechsilben aufgeteilt werden. Der Präfixtrenner sucht in der vorliegenden Phonemkette nach in einer speziellen Liste verzeichneten Präfixen und markiert diese durch Sprechsilbengrenzen. Der Sonderindikatortrenner (Nach- und Vorindikator) wurde zur besseren Unterscheidung zwischen `| t - Fugenlaut s |` und `| ts |` als Repräsentant des graphischen Zeichens [*z*] implementiert. Die nach Abarbeitung dieser Module vorliegende Phonemkette wird in Konsonant-Vokal-Darstellung umgesetzt. Über dieser K-V-Struktur kommen, nachdem die Vokalkerne markiert wurden, die folgenden Trennregeln zur Anwendung:

- bei mindestens einem Konsonanten zwischen zwei Vokalkernen wird vor dem letzten Konsonant getrennt
- zwischen Vokalkernen wird getrennt (Diphtonge sind monophonematisch)

Hier wird auch das Problem der schwa-Elision mit erfaßt, indem bei der Konsonant-Vokal-Strukturbildung Schwachtonsilben ohne Vokalkern das Symbol KT (T...tonhaft) zugewiesen bekommen, wobei im K-V-Trenner das Symbol T dem Symbol V gleichgestellt ist. Als Ergebnis liegt für jedes Wort der Wortliste eine Phonemkette mit markierten Sprechsilbengrenzen vor. Diese Ketten werden zur Erzeugung einer Sprechsilbenliste genutzt, wobei bestimmte Kontextinformationen (Start- , Mittel- und Endsilbe) mit verzeichnet werden.

Die erwähnte Sprechsilbenliste umfaßt bei der Generierung nur die aus der Graphem-Phonem-Umsetzung resultierenden Normsprechsilben (- wegen der zu erwartenden Datenmengen werden nur kanonische Transkriptionen generiert).

Zur Erzeugung leistungsfähiger Referenzmodelle sollen diese Normsprechsilben durch die zu erwartenden Aussprachevarianten ergänzt werden. Zur Ermittlung von repräsentativen Abweichungen von der Standardaussprache wurde reales Sprachmaterial untersucht. Der verwendete Korpus war das feingelabelte Sprachmaterial der Domäne IC-Auskunft [pd92]. Der Umfang betrug 194 Wortformen in 64 Sätzen. Es wurden die Realisierungen von 11 Sprechern und zwei kanonische Transkriptionen betrachtet. Das Gesamtmaterial umfaßte 7514 Realisierungen wobei die 194 Wortformen in 875 Varianten auftraten. Die beobachteten Abweichungen von der Standardlautung wurden klassifiziert und nach Häufigkeit sortiert [fl94]. Anhand der Untersuchungsergebnisse für reale Sprache wurden folgende Ausspracheveränderungen modelliert:

- 'schwa'-Elision am Silbenauslaut vor m, n und l
- Nasalassimilation (nach 'schwa'-Elision)
- Plosivlautreduktion an Silbengrenzen
- Varianten bei Vokal - r - Verbindungen

Erkennungsexperimente mit Sprechsilbenmodellen sind in [fl95a] beschrieben.

5.6 Verkettung von Silbenhypothesen

Zur Ermittlung des Erkennungsergebnisses müssen die als Zwischenrepräsentation ermittelten Silbenhypothesen zu Wörtern und Sätzen bzw. Phrasen verknüpft werden. Bei der optimalen Verknüpfung der Hypothesen sind zeitliche Einschränkungen, die zulässige Lücken und Überdeckungen zwischen einzelnen Hypothesen beschreiben und inhaltliche Aspekte, die zulässige Silbenfolgen im Rahmen der betrachteten Domäne betreffen, zu beachten.

5.6.1 Zeitaspekt der Silbenverknüpfung

Der Zeitaspekt der Silbenhypothesenverknüpfung betrifft die zeitliche Aufeinanderfolge von Silbenhypothesen, die zu größeren Einheiten verbunden werden können. Die Silbenhypothesen sind über dem Signal mit festen Start- und Endzeitpunkten angeordnet. Eine Verknüpfung zweier oder mehrerer Silbenhypothesen zu einer Worthypothese ist nur dann zulässig, wenn die zwischen den Hypothesen auftretenden Lücken bzw. Überlappungen unterhalb einstellbarer Schwellen liegen. Innerhalb der damit definierten Verknüpfungsbereiche werden die Verknüpfungen gleich bewertet.

Zur Optimierung dieser zeitlichen Verkettung wurden die tatsächlich auftretenden Lücken und Überlappungen untersucht. Dazu wurden die Anordnungen der nach Formel (1) relativ zu ihrer korrekten Position im Satz ausgewählten Silbenhypothesen untersucht.

$$\forall i(k_i = argmin_k(\mid tr_i - ts_k \mid + \mid te_k - tr_{i+1} \mid)) \quad mit 1 \leq i \leq n, 1 \leq k \leq m \quad (1)$$

mit	tr_i, tr_{i+1}	Referenzzeit für Silbenstart und -ende der Referenzsilbe RS_i
	ts_k, te_k	Start- bzw. Endzeitpunkte der Silbenhypothese H_k
	n	Anzahl der Referenzsilben
	m	Anzahl der Silbenhypothesen

Die Ergebnisse [al95] zeigen, das die Übergänge eine Gauß-ähnliche Verteilung um den Nullpunkt aufweisen. Das bedeutet, daß die Silbenhypothesen dicht um den Idealfall des nahtlosen Übergangs zueinander streuen.

Die aus der Untersuchung gewonnenen Erkenntnisse wurden genutzt, um die bei der Verknüpfung der Hypothesen auftretenden Lücken und Überlappungen nicht nur über einen Schwellwert zu bewerten, sondern einen der konkreten Länge entsprechenden Faktor nach (2) zu ermitteln und in die Ermittlung der Worthypothesenbewertung einfließen lassen zu können.

$$f(t_{diff}) = \exp -(t_{diff}^2/2\sigma^2) \tag{2}$$

mit t_{diff} - Differenzzeit zwischen zwei Silben.

5.6.2 Inhaltsaspekt der Silbenverknüpfung

Neben der Einschränkung der Verknüpfungsmöglichkeiten durch die Beachtung zeitlicher Aspekte spielt der Inhaltsaspekt der Verknüpfung ebenfalls eine wichtige Rolle bei der Generierung von Silbenfolgen. Die gewonnenen Silbenhypothesen können nur entsprechend zulässiger sprachlicher Strukturen miteinander verknüpft werden. Diese Strukturen werden durch Sprachmodelle beschrieben, wobei regelbasierte und statistische Sprachmodelle zum Einsatz kommen. Der Einfluß des Sprachmodells ist in der stochastischen Entscheidungstheorie durch die enthaltene a-priori-Wahrscheinlichkeit der Wortfolge W $P(W)$ vgl. Formel (3) realisiert.

$$\hat{W} = argmax_W P(A \mid W)P(W) \tag{3}$$

Die a-priori-Wahrscheinlichkeit einer Wortfolge kann durch eine Grammatik (regelbasierter Ansatz) oder durch ein statistisches Sprachmodell beschrieben werden. Für die hier betrachtete Realisierung wurde ein statistisches Sprachmodell, das die Wahrscheinlichkeit des Auftretens einer Silbenfolge approximiert, verwendet.

Bei der Generierung des Sprachmodells wurden zwei Varianten untersucht. Zum einen wurden aus einem vorliegenden Korpus die Silbenpaarhäufigkeiten ermittelt und zum anderen wurde versucht, über den Silben Äquivalenzklassen zu bilden und diese zur Ermittlung von Silbenfolgenwahrscheinlichkeiten einzusetzen.

relative Wortpaarhäufigkeiten Statistische Sprachmodelle approximieren die Wahrscheinlichkeit von Wortfolgen. Da zur Abschätzung der Wahrscheinlichkeit längerer Folgen sehr große Stichproben erforderlich sind, beschränkt man sich meist auf die Betrachtung von Wortpaaren (Bigram-Modell) oder Worttripeln (Trigram-Modell). Eine häufig verwendete Methode zur Ermittlung der Bigram-Wahrscheinlichkeit $P(w_i \mid w_{i-1})$ besteht in der Ermittlung der relativen Wortpaarhäufigkeiten aus einem Textkorpus (siehe z.B. [ue94]):

$$P(w_i \mid w_{i-1}) = h(w_i \mid w_{i-1}) = N(w_{i-1}, w_i)/(N(w) - 1) \tag{4}$$

wobei $N(w_{i-1}, w_i)$ die Anzahl der gesehenen Wortpaare und $N(w)$ die Anzahl der gesehenen Wörter ist. Für alle im Trainingskorpus nicht gesehenen Wortpaare ist die Bigram-Wahrscheinlichkeit Null. Für diesen Fall müssen spezielle Abschätzungen getroffen werden.

Die Sprachmodelle werden im Allgemeinen auf Wortbasis erzeugt. Für die hier geschilderte Anwendung wurde ein Sprachmodell erzeugt, das die relative Silbenpaarhäufigkeit beschreibt.

Äquivalenzklassen Neben den Bi- und Trigram-Modellen auf Basis konkreter Wörter können Sprachmodelle auch auf der Basis von Äquivalenzklassen gebildet werden. Dabei wird ein Wort einer oder mehreren Klassen zugeordnet. Diese Äquivalenzklassen können zum Beispiel durch Wortkategorien (auch 'parts-of-speech', Abk. POS) repräsentiert sein. Die Bildung von Äquivalenzklassen unterliegt zwei Erfordernissen ([je90]):

- Die Aufteilung des Wortschatzes in Klassen muß fein genug sein, um eine ausreichende **Genauigkeit** bei der Vorhersage des Nachfolgers zu gewährleisten.

- Die Anzahl der zu einer Klasse gehörenden Wörter muß groß genug sin, um eine gewisse **Sicherheit** der Vorhersage (Aussagekraft) zu garantieren.

Um ein Silben-Bigram-Modell auf der Basis von Äquivalenzklassen zu erstellen, wurden zwei Herangehensweisen für die Definition von Silbenkategorien getestet.

- syntaktisch-semantische Kategorien (part-of-speech)

  Es wurde ein Grundinventar von 32 syntaktisch-semantischen Wortkategorien ([al95]) um die Einteilung nach der Stellung der Silbe im Wort erweitert. Die vier Silbenpositionen sind:

  - Vorsilbe
  - Endsilbe
  - Zentralsilbe
  - Einsilber

Damit erhält man 128 Silbenkategorien. Eine Zuordnung einer Silbe zu mehr als einer Kategorie ist durch diese Einteilung möglich, so daß (4) zur Berechnung der Wahrscheinlichkeit verwendet wird:

$$P(s_r \mid s_q) = \sum_{c_r} h_0(s_r \mid c_r) \sum_{c_q} h_1(c_r \mid c_q) h_0(c_q \mid s_q) \tag{5}$$

mit:

mit:	s_r	Silbe r
	s_q	Silbe q
	c_r	Kategorie r
	c_r	Kategorie q

- Konsonant-Vokal-Klassen (KV-Klassen)

  Um eine teilweise automatische Gewinnung der Silbenkategorien zu erreichen, wurde ein zweiter Ansatz verfolgt. In [or80] ist eine Einteilung der Sprechsilben in Konsonant-Vokal-Folgen vorgenommen worden. Im Deutschen können danach die Sprechsilben in 24 KV-Klassen eingeteilt werden. Anhand des zur Verfügung stehenden Datenmaterials konnten 23 gegenüber [or80] leicht veränderte Klassen definiert werden [al95].

  Ein Nachteil dieses Ansatzes besteht darin, daß kein syntaktisch-semantischer Zusammenhang zwischen KV-Klassen vorausgesetzt werden kann.

Experimente zur Worthypothesenbildung auf der Basis der unter Zeit- und Inhaltsaspekt kombinierten Silbenhypothesen sind in [fl95b] dargestellt.

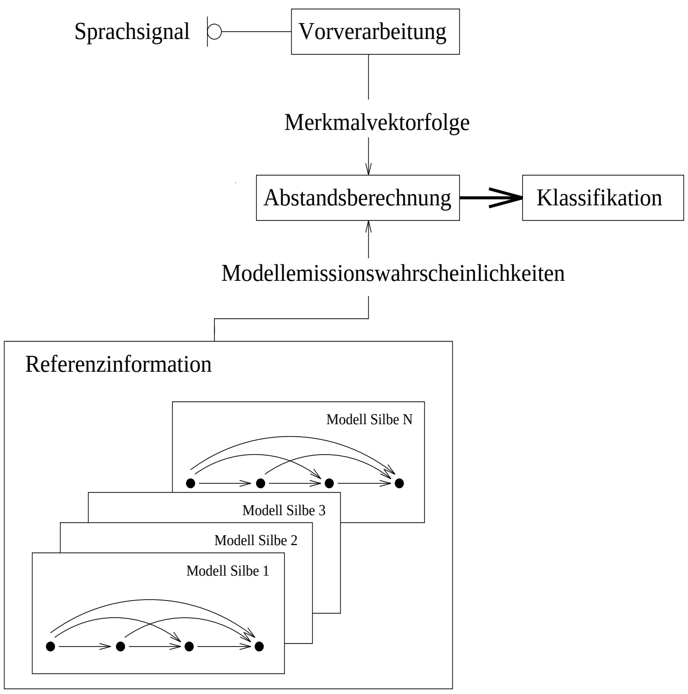


Abbildung 4: Silbenmodelle in einem stochastischen Erkennungssystem

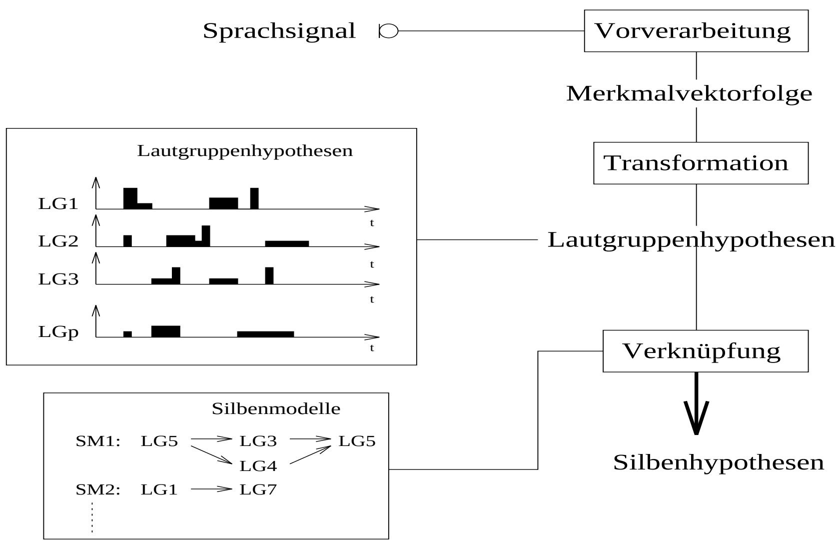


Abbildung 5: Silbenmodelle als Verknüpfungsvorschrift

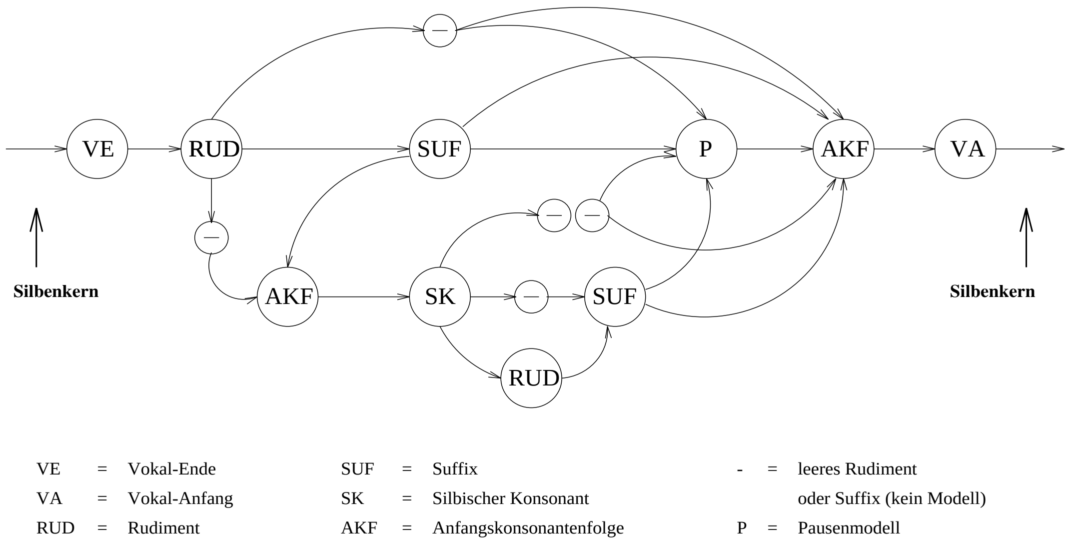
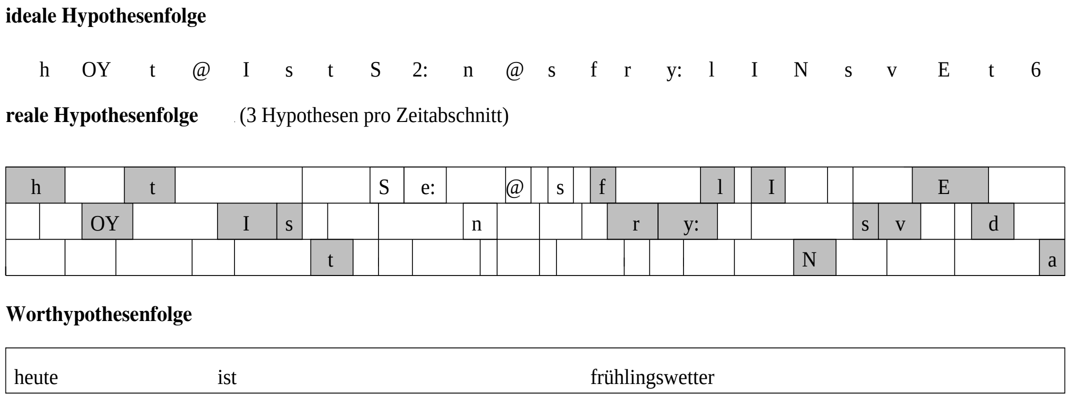


Abbildung 6: Phonotaktische Mikrosyntax für die Silbenteilerkennung nach [ru91]



orthographisches Wort	Modelle	erkanntes Modell
heute	h OY t @	
	h OY t	●
ist	Ist	●
schönes	S 2: n @ s	
Frühlingswetter	f r y: l I N s v E t 6	
	f r y: l I N s v E d 6	
	f r y: l I N s v E d a	●

Abbildung 7: Verknüpfung von Lauthypothesen

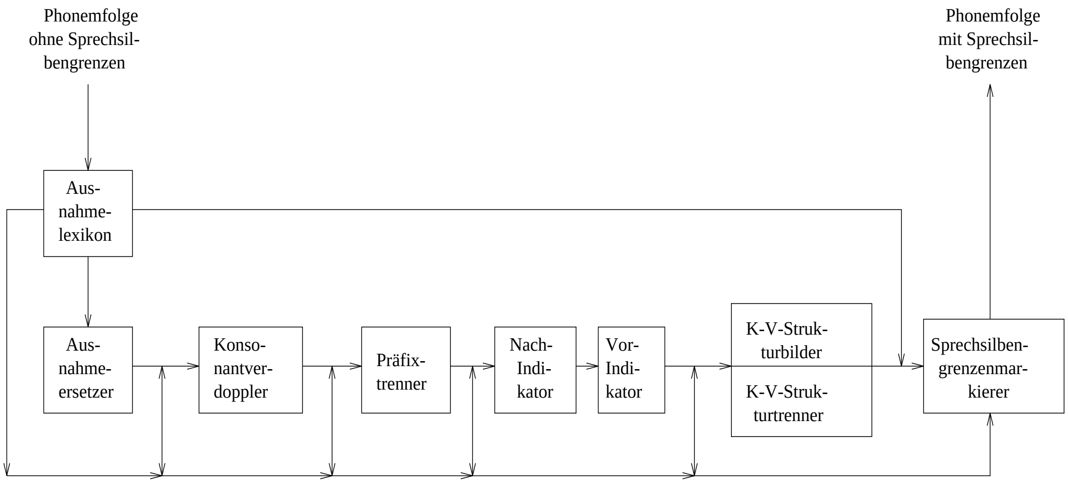


Abbildung 8: Automatische Sprechsilbensegmentierung nach [jo93]

6 Aussprachelexikon

6.1 Einsatz des Aussprachewörterbuchs in Spracherkennungssystemen

6.1.1 Aussprachlexika für spezielle Korpora

Aussprachelexika enthalten eine Zuordnung von orthographischer und phonetischer Repräsentation von Wörtern. Es gibt allgemeine Lexika, die die Normaussprache für einen umfangreichen Grundwortschatz enthalten (z. B. [aw82] und [du84]). Für Spracherkennungsaufgaben werden meist spezielle Lexika eingesetzt, die neben der Normaussprache auch möglichst akustiknahe Beschreibungen der zu erkennenden Wörter enthalten. Neben der Optimierung des Aufwandes kann dadurch auch auf die diskursbedingten Besonderheiten besser eingegangen werden.

Wesentliche Kriterien, die beim Aufbau von Aussprachewörterbüchern zu berücksichtigen sind, ergeben sich aus der vorliegenden Erkennungsaufgabe. Somit ist besondere Aufmerksamkeit zu richten auf:

die zugelassene Wortschatzgröße

Bei einer Erkennungsaufgabe mit kleinem Wortschatz kann man davon ausgehen, daß die Wortfrequenz relativ hoch ist, was eine detaillierte Modellierung ermöglicht. Hinzu kommt, das kleine Wortschätze dem Nutzer vollständig bekannt sein müssen, somit sind auch starrere syntaktische Strukturen anzunehmen, was wiederum zu einer speziellen Art der Artikulation führt. Man kann davon ausgehen, daß Erkennungsaufgaben mit kleineren Wortschätzen eine geringere Vielfalt an Aussprachevarianten und Verschmelzungen aufweisen, daß jedoch extremere Formen auftreten können, da der Nutzer quasi in einer künstlichen Kommunikationssituation agiert.

Bei Erkennungsaufgaben mit sehr großen Wortschätzen können wir von einer natürlichen Kommunikationssituation ausgehen. Wir beobachten hier sehr unterschiedliche Wortfrequenzen, auch tritt das Problem des unbekannten Wortes auf, da dem Nutzer keinerlei Beschränkungen vorgegeben sind. Spontansprachliche Phänomene treten wie innerhalb der humanen Kommunikation auf. Die verwendeten syntaktisch-semantisch-pragmatischen Strukturen sind sehr vielfältig und entsprechen nicht mehr den Definitionen der Korrektheit bzw. Wohlgeformtheit.

die behandelte Domäne

Die von der Erkennungsaufgabe behandelte Domäne hat ebenfalls Einfluß auf die zu erwartende Wortformenvielfalt und den vom Nutzer gewählten Sprachtyp und Sprechstil. Eine genau definierte, formale Domäne motivert den Nutzer zu einer korrekten und durchdachten Sprechweise, was wiederum Einfluß auf die enthaltenen spontansprachlichen Erscheinungen und den Wortschatzumfang hat. In diesen Bereich fallen die sprachgesteuerten Auskunfts- und Buchungssysteme. Auch kann durch eine vom System ausgehende Dialogsteuerung der Sprechstil des Nutzers beeinflußt werden.

Weitgefaßte Domänen, die auch emotionale Bereiche mit erfassen, sind apriori sehr schwer bezüglich der vom Nutzer verwendeten Sprache einzugrenzen. Neben einem sehr großen Wortschatz und umfangreichen spontansprachlichen Bestandteilen beobachten wir noch die Erscheinung der semantischen Kreativität, die die Interpretation erschwert.

die Art der betrachteten Sprache

Bei Spracherkennungssystemen können vier grundsätzliche Spracharten unterschieden werden:

- Einzelworte
- gelesene Sprache
- fest definierte, kleine Wortschätze
- Spontansprache

Einzelworte sind durch deutliche Artikulation und definierte Pausen am Wortanfang und am Wortende gekennzeichnet. Bei gelesener Sprache finden wir ebenfalls noch deutliche Artikulation, teilweise sogar Überartikulation, d. h. sehr orthographienahe Aussprache vor. Aussprachevarianten auf Wortebene treten ebenfalls auf, Wortverschmelzungen sind kaum zu beobachten. Definierte Wortpausen treten nicht mehr auf. Bei fest definierten, kleinen Wortschätzen (meist Kommandosprachen) wird ebenfalls deutlich artikuliert. Spontansprachliche Phänome sind an bestimmte Wörter bzw. Wortverbindungen, die durch Wortschatz- und Syntaxdefinition festgelegt sind, gebunden. Bei Spontansprache ergeben sich keine Einschänkungen hinsichtlich der Wortwahl und der verwendeten syntaktisch-semantischen Strukturen. Daraus resuliert eine Sprache, die durch variierende Artikulationssorgfalt und variables Sprechtempo gekennzeichnet ist. Wir finden das gesamte Spektrum der spontansprachlichen Phänomene (vgl 6.1.2) vor.

den potentiellen Nutzer des Systems

Die Besonderheiten der Sprache und des Wortschatzes, die ein Sprachverarbeitungssystem handhaben muß, werden auch stark von Nutzer bestimmt. Ein geschulter Nutzer, der viel Erfahrungen im Umgang mit derartigen Systemen und Wissen über im Hintergrund ablaufende Prozesse hat, produziert eine leichter zu verarbeitende Sprache, da er seine Kommunikationsabsicht möglichst effizient realisieren möchte. Ein ungeübter und unsicherer Nutzer weiß nicht, welche Art von Sprache gut geeignet ist. Ein spielerischer Nutzer wird versuchen, die Grenzen des Systems auszuloten.

6.1.2 Beschreibung spontansprachlicher Besonderheiten

Spontane Sprache weist spezielle Besonderheiten auf, die im Referenzwissen eines Systems zur Verarbeitung dieses Sprachtyps mit enthalten sein müssen. Für folgende Erscheinungen müssen spezielle Beschreibungen im Referenzwissen enthalten sein:

- Häsitationen
- Wortabbrüche
- Nichtwörter
- Aussprachevariationen innerhalb von Wörtern
- Wortverschmelzungen
- Störungen

Bei **Häsitationen** handelt es sich um Einfügungen in die Wortfolge, die eindeutig kein Wort sind (z. B. *äh, ähm, hm*) und unterschiedliche Bedeutung in der Sprachäußerung haben können. Sie werden zum Teil verwendet, um Denkpausen zu füllen - in diesem Fall würde ihnen keine Bedeutung zukommen. In anderen Fällen werden sie jedoch mit einer Kommunikationsabsicht verwendet, sie können zur Zustimmung, zum Ausdruck des Anzweifelns oder oder zur Überleitung eingesetzt werden. Je nach Einsatz werden sie mit unterschiedlichen prosodischen Parametern erzeugt und müßten auch durch einsatzspezifische Modelle beschrieben werden. In jedem Fall können sie nicht bei der der akustischen Erkennung ausgeblendet werden, für sie sind entsprechende Modelle bzw. Einträge im Aussprachewörterbuch vorzusehen.

Die Modellierung von **Wortabbrüchen** stellt ebenfalls ein Problem dar. Da in spontaner Sprache jedes Wort an jeder beliebigen Stelle abgebrochen werden kann, müßte die Modellierung so realisiert werden, daß die Wort- oder Silbenmodelle, die zur Verknüpfung der trainierten Basiseinheiten verwendet werden, an jeder beliebigen Stelle verlassen werden können. Das heißt aber, daß an jeder Stelle erneut alle Verknüpfungsmodelle gestartet werden müßten. Daraus resuliert ein sehr großes Anwachsen der Hypothesen über einen bestimmten Signalabschnitt. Hinzu kommt, daß die abgebrochenen Wörter keine Funktion im Satz haben, sie entweder im Anschluß wiederholt oder völlig ersetzt werden. Wortabbrüche müßten aus diesem Grund nicht als extra Einträge im Aussprachewörterbuch enthalten sein. Für Erkennungsexperimente innerhalb einer abgeschlossenen Sprachdatenmenge sind die darin enthaltenen Wortabbrüche jedoch von Interesse. Die in den VERBMOBIL-Dialogtransliterationen zur Terminabsprache enthaltenen Wortabbrüche sind daher als Einträge im Aussprachewörterbuch enthalten.

Bei **Nichtwörtern** handelt es sich um stückweise Veränderungen innerhalb von Wörtern, die diese zu zwar u. U. erkennbaren aber nicht lexikalisierten

Wörten machen. Nichtwörter können wegen ihres zufälligen Erzeugungsprozesses nicht im Aussprachewörterbuch erfaßt werden. Die in den VERBMOBIL-Dialogtransliterationen zur Terminabsprache enthaltenen Nichtwörter sind aus den oben genannten Gründen in das vorliegende Aussprachewörterbuch aufgenommen worden.

Die **Aussprachevariationen innerhalb von Wörtern** führen zu zum Teil erheblichen Abweichungen von der Normrealisierung, so daß eine Abbildung der Basiselementehypothesen auf ihre in der Normrealisierung verzeichnete Abfolge nicht mehr möglich ist. Da ein Großteil der Aussprachevariationen auf Koartikulations- und Steuerungsphänomene zurückgeführt werden kann, sind sie vom orthographischen Text ausgehend, regelbasiert modellierbar. Die so erzeugbaren Varianten können entweder für alle Formen oder für besonders hochfrequente Formen in das Aussprachewörterbuch aufgenommen werden.

Wortverschmelzungen treten in Spontansprache in Abhängigkeit vom Sprechstil auf. Sie betreffen besonders Sprachabschnitte, in denen wenig semantische Information enthalten ist. Sie führen jedoch teilweise zu Erscheinungen an der akustischen Oberfläche, die eine Rückführung auf die Zitierformen der Wörter unmöglich machen. Da aber eine vollständige Wortfolge für eine richtige Interpretation durch höhere Verabeitungsstufen erforderlich ist, müssen die phonetischen Repräsentationen der Wortverschmelzungen zusammen mit ihren orthographisch korrekten Auflösungen im Aussprachewörterbuch enthalten sein. Sie können regelbasiert für bestimmte Abfolgen ausgewählter syntaktischer Wortkategorien erzeugt werden.

Störungen treten in Spontansprache meist in Form außersprachlicher Umweltgräusche auf. Sie können somit im Rahmen der akustischen Erkennung ausgeblendet werden. Sie werden durch spezielle akustische Modelle beschrieben die keinen Bezug zu einer orthographischen Repräsentation haben. Die Störmodellerzeugung kann durch die Erfassung und Auswertung realer Störungen [ma94] unterstützt werden.

6.1.3 Regeln für Aussprachevarianten und Wortverschmelzungen

Für die vom Text ausgehende Generierung von Aussprachevarianten und Wortverschmelzungen werden aus der Koartikulation und Steuerung ableitbare Regeln zusammengestellt. Laut [ko95] werden *Koartikulations- und Steuerungserscheinungen als Prozesse interpretiert, was nur unter der Voraussetzung des Ansatzes von Ausgangsformen möglich ist. Das bedeutet, daß auch hier die phonetische Form des Einzelwortes wieder zugrundegelegt wird, daß also aus dem phonetisch kodifizierten Lexikoneintrag Veränderungen im Satzzusammenhang über eine Reihe von Regeln hergeleitet werden.*

Aussprachevariationen sind auf Koartikulation, worunter das Prinzip der wechselnden Koordination der artikulatorischen Parameter gefaßt wird, und auf die Steuerung, die die Artikulationsreduktion beschreibt, zurückzuführen. Zwei Komplexe von Artikulations- und Steuerungsphänomenen im Deutschen, die Assimilationen und Elisionen einerseits und die schwachen Formen andererseits bilden die Grundlage für die Ableitung von Variations- und Verschmelzungsregeln.

Assimilation und Elision Unter Assimilation verstehen wir die Angleichung benachbarter Segmente im Sprachsignal. Bei dieser Erscheinung der Angleichung unterscheidet man bezüglich der Richtung die regressive (der nachfolgende Laut verändert den vorangegangenen) und die progressive (der vorangehende Laut verändert den nachfolgenden) Assimilation. Weiterhin ist von Bedeutung, welcher artikulatorische Parameter von der Assimilation betroffen ist und welche zeitliche Extension die Angleichung hat. Die Elision beschreibt den Segmentausfall. Hier ist von Bedeutung, welcher Art das elidierte Segment ist. Assimilationen und Elisionen können weiterhin nach ihrer Position in linguistischen Einheiten, nach ihrer phonetisch-segmentellen Umgebung, der vorhandenen Akzentuierung und der morphologischen und syntaktischen Sruktur des Kontextes unterschieden werden. Weiterhin kann ihre Gültigkeit entsprechend der stilistischen Sprachgestaltung unterschiedlich sein. Im folgenden werden einige Assimilations- und Elisionsregeln nach [ko95] mit Beispielen angegeben:

- @-Elision

  Ausfall des Reduktionsvokals /@/ vor einem Nasal nach dem akzentuierten Vokal

$$/le:b@nt/ \rightarrow /le:bnt/$$

- regressive Assimilation des Artikulationsortes

  regressive Anpassung des Artikulationsortes in der Folge Plosiv-Nasal bzw. Labial-Dorsal

$$/anbInd@n/ \rightarrow /ambInd@n/$$

- progressive Assimilation des Artikulationsortes

  progressive Angleichung des Artikulationsortes in den Folgen Labial,Dorsal-apikaler Nasal und Frikativ-Nasal

$$/e:b@n/ \rightarrow /e:bn/ \rightarrow /e:bm/$$

$$/ru:f@n/ \rightarrow /ru:fn/ \rightarrow /ru:fm/$$

- t-Elision

  In der Mitte von Dreierkonsonantengruppen oder nach Frikativen und nach /*n*/ in Senkungssilben entfällt das /*t*/.

$$/glants/ \rightarrow /glans/$$

$$/vIrtSaftUnt.../ \rightarrow /vIrtSafUnt.../$$

- regressive Assimilation der Artikulationsart

  Ein wortfinales /s/ wird an ein wortinitiales /S/ assimiliert.

$$/dasSIf/ \rightarrow /daSSif/$$

- progressive Assimilation der Stimmlosigkeit

  Stimmhafte Plosive oder /z/ werden nach stimmlosen Plosiven oder stimmlosen Frikativen ebenfalls stimmlos realisiert.

$$/daszElb@/ \rightarrow /dassElb@/$$

- regressive Assimilation der Nasalität

  Bei stimmhaften Plosiven nach Nasalen tritt eine Assimilation der Nasalität auf.

$$/tsUmbaISpi:l/ \rightarrow /tsUmmaISpi:l/$$

- progressive Assimilation der Nasalität

  Stimmhafte Plosive vor Nasalen übernehmen die Nasalität.

$$/zIgna:l/ \rightarrow /zINna:l/$$

- Geminatenreduktion

  Geminaten in finaler Stellung werden reduziert. Bei Zwischenwortgeminaten wird zum zweiten Element reduziert.

$$/kOm@n/ \rightarrow /kOmn/ \rightarrow /kOmm/ \rightarrow /kOm/$$

$$/anmo:nta:g@n/ \rightarrow /ammo:nta:g@n/ \rightarrow /amo:nta:gen/$$

- Sonorisierung

  Intervokalische stimmlose Plosive und Frikative können stimmhaft werden.

$$/dasmUsICmax@n/ \rightarrow /dasmUzICmax@n/$$

Schwache Formen Es gibt eine Klasse von Funktionswörtern, die bei Taktsenkung (d. h. wenn sie nicht akzentuiert sind) wie unbetonte Wortsilben behandelt werden. Zu dieser Klasse zählen Pronomina, Artikel, Formverben, Präpositionen, Konjunktionen und Adverbien. Auf diese schwachen Formen sind Steuerungsregeln anwendbar die Schwachtonreduktionen bewirken. Auf diese so entstandenen Formen sind die im Vorhergehenden erwähnten Assimilations- und Elisionsregeln, zum Teil wiederholt, anwendbar.

Weitere Veränderungen an den schwachen Formen können abgeleitet werden durch:

- den phonetischen Kontext

  In einem bestimmten phonetischen Kontext können weitere Reduktionen auftreten, z. B. gilt nach Dauerlauten:
  /Ist/ $\rightarrow$ */is/* $\rightarrow$ */s/*
  /e:r s g@kOm@n/ vs. */kUrt Is g@kOm@n/*

- den syntaktischen Kontext

  In der Äußerung *Woher kennst 1 du 2 meinen Chef 3?* kann an den mit 1,2 und 3 markierten Stellen die Konjunktion *denn* eingefügt werden. Während diese Wort an den Positionen 1 und 2 die schwache Form */n/* bilden kann, ist dies an Position 3 aufgrund des syntaktischen Kontextes unmöglich.

- semantische Faktoren

  Eine potentiell schwache Form kann durch semantische Faktoren so gekennzeichnet werden, daß keine Reduktion möglich ist, es kann sogar zu einer besonderen Hervorhebung kommen:
  Er geht zur (zu der) Hütte.
  Er geht zu der Hütte, über die wir schon gesprochen hatten.
  Er geht zu d e r Hütte.

- stilistische Faktoren

  Wachsendes Sprechtempo, abnehmende Formalität und zunehmende Dialektfärbung ist mit einer zunehmenden Veränderung der Sprache bezüglich des Anteils schwacher Formen verbunden

6.2 Funktionen zur Erzeugung von Referenzwissen aus Texten

Einen wesentlichen Anteil an der Arbeit zum Schwerpunkt "Aussprachewörterbuch" bildete die Entwicklung von Hilfsmitteln, mit denen aus Eingangstexten, die in vielfältiger Form Einen wesentlichen Anteil an der Arbeit zum Schwerpunkt "Aussprachewörterbuch" bildete die Entwicklung von Hilfsmitteln, mit denen aus Eingangstexten, die in vielfältiger Form vorlagen, die für ein Experimentiersystem zur Sprachsignalverarbeitung erforderliche Referenzinformation möglichst effizient erzeugt werden kann. Obwohl das eigentliche Kernstück des Aussprachewörterbuchs die Zuordnung von orthographischer und phonetischer Beschreibung von Erkennungseinheiten (meist auf Wortebene) ist, war für das verwendete Experimentiersystem ([we94a]) ein reichhaltigeres Referenzwissen erforderlich, das aus dem Aussprachewörterbuch aktuell generierbar sein muß. So wurden neben der orthographischen und der phonetischen Ebene auch noch die Silben-, Phonem- und Merkmalebene untersucht. Es wurden eine Reihe von Funktionen entwickelt, die die wortbezogene und die silben- und lautbezogene Verarbeitung von beliebigen Eingangstexten realisieren.

6.2.1 Wortorientierte Verarbeitungsfunktionen

Die Abbildung 9 zeigt die im Rahmen der wortorientierten Verarbeitung generierbaren Repräsentationen. Aus dem in beliebiger Form vorliegenden Text wird zunächst eine Wortliste[1] erzeugt. Diese Wortliste kann mit implizitem Filter generiert werden, eine andere Möglichkeit besteht in der Verwendung eines extra Textfilters, durch welches zunächst ein normierter Text erzeugt wird. Dieser normierte Text wird so erzeugt, daß in ihm alle akustisch repräsentierten Äußerungsbestandteile enthalten sind, d. h. es werden auch Häsitationen, Wortabbrüche und Nichtwörter repräsentiert, die damit auch Bestandteile der Wortliste sind. Aus der Wortliste wird ein Aussprachewörterbuch generiert, in das Einträge aus einem oder mehreren bereits existierenden Wörterbüchern übernommen werden. Eventuell fehlende Einträge werden interaktiv ergänzt. Die Erzeugung des Aussprachewörterbuchs könnte auch durch automatische Graphem-Phonem-Umsetzung realisiert werden, was jedoch mit zunehmender Anzahl natürlichsprachlicher Phänomene sehr hohe Anforderungen an diesen Umsetzalgorithmus stellt. Das Aussprachewörterbuch kann weiterhin

- mit anderen **kombiniert** werden
- in ascii-Format **konvertiert** werden
- zur Erzeugung einer ascii-Wortliste genutzt werden. Diese Funktion ist für die Kombination des Experimentiersystems mit externen Modulen (z. B. Aachener Sprachmodell) wichtig.
- zur Erzeugung einer phonetischen Wortliste genutzt werden
- nach verschiedenden Kriterien sortiert werden (Dudensortierung, Phonemsortierung, Merkmalsortierung).

Die phonetische Wortliste, die aus dem Aussprachewörterbuch erzeugt wird, bildet die Grundlage zur Erstellung von Wortmodellen, z. B. in Form von

[1] Das Format der angegebenen Datenstrukturen ist dabei ein in der AG Sprache des ITA der TU Dresden entwickeltes internes Dateiformat (vgl. [ru92]). Für den Datenexport stehen Konvertierungsprogramme *interneDateinorm* $\rightarrow$ *ascii* zur Verfügung.

Merkmal- oder Grobeigenschaftsfolgen. Weiterhin wird die phonetische Wortliste zur Generierung des Sprechsilbeninventars des betrachteten Wortschatzes verwendet. Dazu wird ein regelbasierter Sprechsilbentrennalgorithmus eingesetzt ([jo93] vgl. auch 5.5).

Spezielle Korpusuntersuchungen, die die Trennbarkeit der Wörter des Vokabulars betreffen, werden ebenfalls auf der Basis der phonetischen Wortliste durchgeführt. Dabei wird untersucht, inwieweit einzelne Wörter des Wortschatzes als Bestandteile anderer Wörter auftreten. Damit können sich überdeckende Hypothesen erfaßt werden.

Die realisierten Funktionen der wortorientierten Verarbeitung von Eingangstexten werden im Folgenden kurz beschrieben.

$\boxed{\text{dial\_konv}}$

Erzeugen eines normierten Textes
Aus einem Text in beliebiger Form wird ein normierter Text erzeugt. Alle im Originaltext auszublendenden bzw. zu ersetzenden Zeichen sind in einer extra zu definierenden Liste erfaßt. Verarbeitet werden Texte im ascii-Format. Die Dateinamen von Quell- und Zieltext, sowie der Listenname müssen übergeben werden.

$\boxed{\text{creatwl}}$

Erzeugen einer Wortliste
Erzeugung einer Wortliste für einen gefilterten orthographischen Text. Für jedes Wort wird die Vorkommenshäufigkeit im Text angegeben. Der Text muß als ascii-File vorliegen, der Name wird abgefragt.

$\boxed{\text{exwl}}$

Erweitern einer Wortliste
Erweiterung einer bestehenden Wortliste für einen weiteren gefilterten orthographischen Text. Die Vorkommenshäufigkeit wird aktualisiert. Der Text muß als ascii-File vorliegen, der Name wird abgefragt.

$\boxed{\text{txt\_filt}}$

Erzeugen einer Wortliste mit vorgeschaltetem Textfilter
Erzeugung einer Wortliste für einen ungefilterten orthographischen Text. Der Text wird zuerst gefiltert (entsprechend der im PHONDAT-Manual definierten Sonderzeichenliste), die separierten Wörter werden im Anschluß in eine leere Wortliste einsortiert. Die Vorkommenshäufigkeit wird ermittelt. Der Text muß als ascii-File vorliegen, der Name wird abgefragt.

$\boxed{\text{ext\_filt}}$

Erweitern einer Wortliste mit vorgeschaltetem Textfilter
Erzeugung einer Wortliste für einen ungefilterten orthographischen Text. Der Text wird zuerst gefiltert (entsprechend der im PHONDAT-Manual definierten

Sonderzeichenliste), die separierten Wörter werden im Anschluß in eine existierende Wortliste einsortiert. Die Vorkommenshäufigkeit wird aktualisiert. Der Text muß als ascii-File vorliegen, der Name wird abgefragt.

ascwl

Erzeugen einer ascii-Wortliste
Die angegebene Wortliste im DNorm 3.0 Format wird in ein ascii-File konvertiert. Der Filename wird abgefragt.

asdnwl

Konvertieren einer ascii-Wortliste in DNorm
Die angegebene Wortliste im ascii Format wird in ein DNorm-File konvertiert. Der Filename wird abgefragt.

creatpl

Erzeugen eines Ausspracheswörterbuchs 1
Für eine vorliegende Wortliste wird ein Aussprachewörterbuch angelegt. Dazu werden aus einem vorhandenen allgemeinen Aussprachewörterbuch die zu der Wortliste passenden Einträge für die orthographische und die phonetische Wortdarstellung entnommen. Einträge der Wortliste, die im allgemeinen Aussprachewörterbuch nicht enthalten sind, werden mit leerer phonetischer Darstellung repräsentiert.

fpl

Erzeugen eines Aussprachewörterbuchs 2
Ein mit creatpl angelegtes Aussprachewörterbuch wird aufgefüllt.Dazu werden aus einem vorhandenen allgemeinen Aussprachewörterbuch die zu der orthographischen Repräsentation passenden Einträge für die orthographische und die phonetische Wortdarstellung entnommen. Einträge des aufzufüllenden Aussprachewörtebuchs, die im allgemeinen Aussprachewörterbuch nicht enthalten sind, werden mit leerer phonetischer Darstellung repräsentiert.

fipl

Erzeugen eines Aussprachewörterbuchs 3
Ein mit creatpl angelegtes Aussprachewörterbuch wird vollständig aufgefüllt. Dazu werden werden interaktiv die noch offenen phonetischen Repräsentationen ergänzt. Die Bearbeitung kann durch Eingabe von [*] unterbrochen werden.

pl_edit

Editieren eines Aussprachewörterbuchs
Ein in DNorm 3.0 Format vorliegendes Aussprachewörterbuch kann bearbeitet werden. Mögliche Funktionen, die interaktiv ausgewählt werden, sind

- Einfügen von Worteinträgen (z.B. Aussprachevarianten)
- Streichen von Worteinträgen
- Korrekturen von Worteinträgen

combi_lex

Kombination von Aussprachwörterbüchern
Zwei angegegebene Aussprachewörterbücher werden zu einem neuen kombiniert, wobei die Einträge beider alphabetisch sortiert ausgegeben werden.

asdnpl

Konvertieren des Aussprachewörterbuchs $ascii \rightarrow DNorm$
Ein als ascii-File vorliegendes Aussprachwörterbuch wird in DNorm 3.0 konvertiert. Der Dateiname des ascii-Aussprachewörterbuchs wird abgefragt.

dnaspl

Konvertieren des Aussprachewörterbuchs $DNorm \rightarrow ascii$
Ein als DNorm 3.0-File vorliegendes Aussprachwörterbuch wird in ein ascii-File konvertiert. Der Dateiname des ascii-Aussprachewörterbuchs wird abgefragt.

wlpl

Erzeugen einer ascii-Wortliste für ein Aussprachewörterbuch
Für das mit pl spezifizierte Aussprachewörterbuch im DNorm 3.0 Format wird eine ascii-Wortliste erzeugt. (Diese wird für den Einsatz des Aachener Sprachmodells benötigt). Der Filename wird abgefragt.

phwl

Erzeugen einer phonetischen Wortliste
Aus einem Aussprachewörterbuch wird eine phonetische Wortliste erzeugt. Dazu werden die phonetischen Beschreibungen extrahiert, und aus diesen werden die Sprechsilbengrenzen gestrichen. Diese Repräsentationen werden ausgegeben.

mix

Buchstabensortierung des Wörterbuchs
Es wird ein nicht nach Groß- und Kleinbuchstaben getrenntes, alphabetisch sortiertes Wörterbuch erzeugt (Dudensortierung).

$\boxed{\text{unisort}}$

Phonem- oder Merkmal-Sortierung des Wörterbuchs
Ein vorliegendes Aussprachewörterbuch mit Merkmal- bzw. Grobeigenschaftsrepräsentation wird orthographisch, phonetisch oder entsprechend der Grobeigenschaften sortiert. Es wird nach ascii-Code sortiert. Der Sortiermodus wird abgefragt.

$\boxed{\text{creatwm}}$

Erzeugen von Wortmodellen
Für die in einer phonetischen Wortliste enthaltenen Wörter werden Wortmodelle erzeugt. Dazu wird das Wort in Modellelemente zerlegt (z. B. Phonemlabel), die durch laufende Nummern codiert werden. Das Wortmodell enthält das Wort als Modellelementefolge, die Anzahl der Modellelemente und die Nummernfolge.

$\boxed{\text{siltren}}$

Silbentrennung
Für eine Liste phonetischer Wörter wird eine regelbasierte automatische Silbentrennung erzeugt.

$\boxed{\text{sort\_len}}$

Sortierung nach Wortlängen
Eine phonetische Wortliste wird nach Wortlängen sortiert. Am Anfang steht die größte Wortlänge.

$\boxed{\text{winw}}$

Wort-in-Wort-Darstellung
Es wird eine gegebene (orthographische oder phonetische) Wortliste mit dem Ziel untersucht, für jedes Wort alle Teilwörter, die ebenfalls in der Liste enthalten sind, herauszusuchen.Wird mit einer orthographischen Wortliste gearbeitet, kann die entsprechende Spalte des Aussprachewörterbuchs genutzt werden.

$\boxed{\text{winws}}$

Wort-in-Wort-Darstellung mit Sprechsilbengrenzen
Es wird eine gegebene phonetische Wortliste mit dem Ziel untersucht, für jedes Wort alle Teilwörter, die ebenfalls in der Liste enthalten sind, herauszusuchen.Im Unterschied zu winw werden jedoch die Sprechsilbengrenzen mit einbezogen.

$\boxed{\text{statist}}$

Wort-in-Wort-Statistik
Für ein gegebenes Aussprachewörterbuch wird die Wort-in-Wort-Statistik angegeben. Ausgehend von einer längensortierten Wortliste wird die Anzahl vorkom-

mender Wörter der Länge n und der Prozentsatz dieser Wörter, die am Anfang, innerhalb und am Ende eines anderen Wortes vorkommen, angegeben.

6.2.2 Silben- und labelorientierte Verarbeitungsfunktionen

creatsl

Erzeugen einer Silbenliste
Aus einem Aussprachewörterbuch mit eingetragenen Sprechsilbengrenzen wird eine Silbenliste erstellt. Die Vorkommenshäufigkeit einer Silbe innerhalb des Aussprachewörterbuchs wird ermittelt.

creatsm

Erzeugen von Silbenmodellen
Für die in einer Silbenliste enthaltenen Silben werden Silbenmodelle erzeugt. Dazu wird die Silbe in Modellelemente zerlegt (z. B. Phonemlabel), die durch laufende Nummern codiert werden. Das Silbenmodell enthält die Silbe als Modellelementefolge, die Anzahl der Modellelement und die Nummernfolge.

wl_syl

Erzeugen einer Wortliste für Silben
Aus dem angegebenen Aussprachewörterbuch werden für eine interaktiv eingegebene Silbe (phonetische Form ohne Silbengrenzen) alle die Wörter aufgelistet, in denen diese Silbe enthalten ist.

asdnlab

Konvertieren einer Labelliste $ascii \rightarrow DNorm$
Eine als ascii-File vorliegendes Labelliste wird in DNorm 3.0 konvertiert. Der Dateiname der ascii-Labelliste wird abgefragt.

6.2.3 Satzorientierte Verarbeitungsfunktionen

setnum

Sätze als Wortnummernfolgen
Die Sätze einer gegebenen Satzliste werden entsprechend dem angegebenen Aussprachewörterbuch in Wortnummernfolgen kodiert. Diese Funktion kann ebenfalls zur Nummernkodierung von Sätzen als Silbenfolgen eingesetzt werden. Die Nummern entsprechen der Reihenfolge in der Wort- bzw. Silbenliste. Die Wort- bzw. Silbenliste kann von beliebiger Struktur sein, wobei in der ersten Komponente das zu kodierende Element in orthographischer Form (bei Wörtern) bzw. in Lautschrift (bei Silben) stehen muß.

lab_einz

Vereinzelung von Labelstrings
Zur Untersuchung von Aussprachevarianten werden Originallabelfiles mehrstufig verarbeitet. Zunächst wird aus einem auf Framebasis vorliegenden Labelstring ein reduzierter Labelstring erzeugt, der jede Labelkennzeichnung nur einmal enthält.

$\boxed{\text{w\_paus}}$

Einfügen von Pausen in Labelstrings
Die zweite Verarbeitungsstufe nach der Labelvereinzellung ist das interaktive Einfügen von Wortpausen in den Einzellabelstring. Dazu wird der Labelstring für die betrachtete Sprachäußerung einzeln dargeboten, im Bedarfsfall kann eine Wortpause markiert werden.

$\boxed{\text{label\_file}}$

Erzeugen eines Labelfiles
Für einen als Labelfolge vorliegenden Satz wird ein Labelfile erzeugt. Dazu wird die Labelfolge in Einzellabel zerlegt und für jedes Label ein Datensatz in der Zieldatei angelegt. Am Satzanfang, zwischen den Wörtern und am Satzende wird eine Pause eingefügt.

$\boxed{\text{asdnset}}$

Konvertieren einer Satzliste $ascii \rightarrow DNorm$
Eine als ascii-File vorliegende Satzliste wird in die DNorm-darstellung konvertiert. Die Konvertierungen sind größtenteils erfordrlich, um dateinormbasierte Dienstprogramme anwenden zu können.

$\boxed{\text{setsyl}}$

Darstellung eines Satzes als Silbenfolge
Für die Experimente zur Silbenverknüpfung werden als Referenzinformation Sätze als Silbenfolgen benötigt. Satzlisten in DNorm-Repräsentation können mit dieser Funktion in Silbenfolgen konvertiert werden. Neben der Satzliste ist ein Aussprachewörterbuch, das die phonetische Repräsentation mit eingefügten Silbengrenzen enthält, erforderlich.

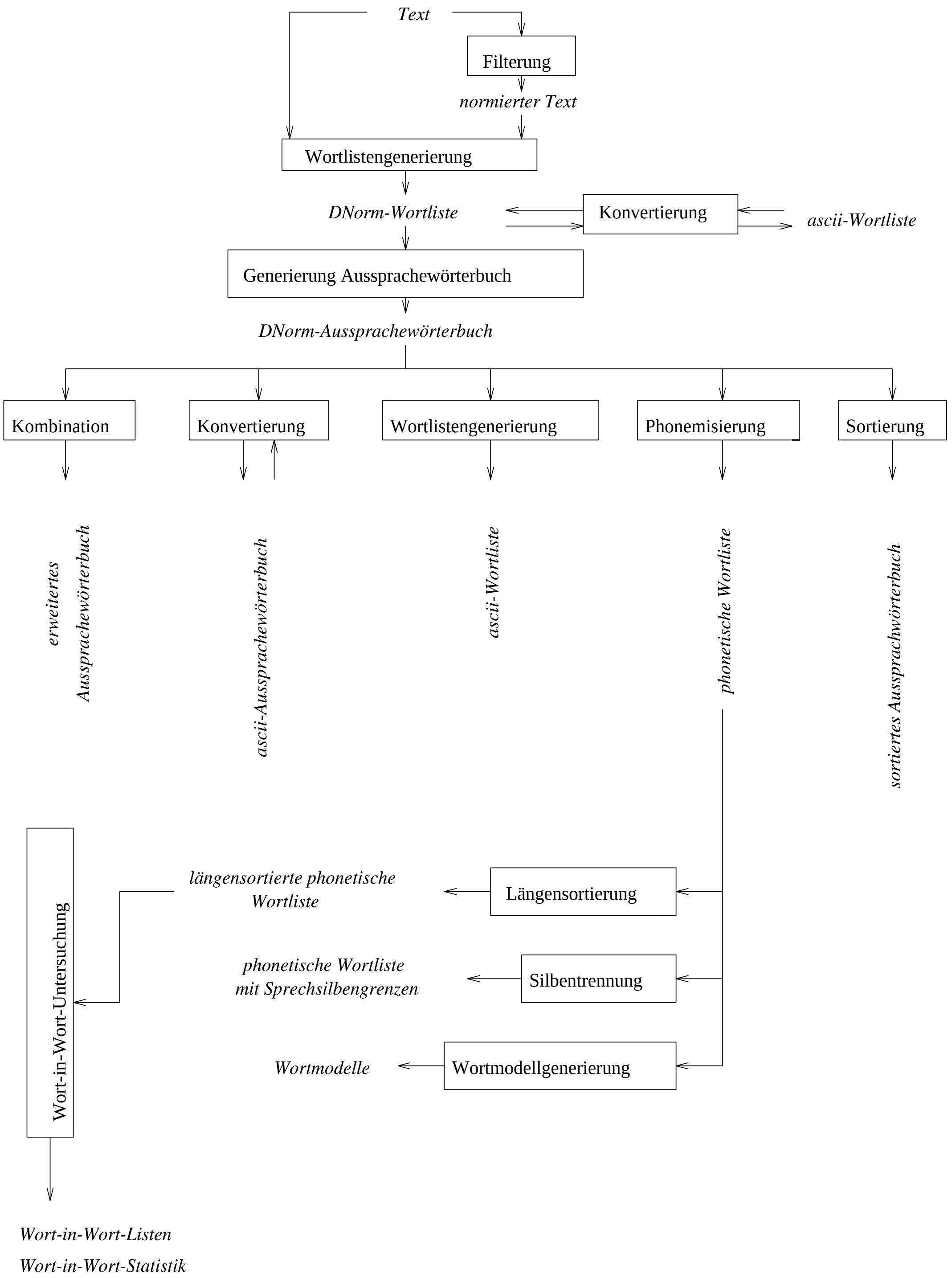


Abbildung 9: Überblick über die Module zur Textbearbeitung

7 Aufbereitung und Untersuchung von realem Sprachmaterial

Für die Entwicklung der in 5 beschriebenen Silbenmodelle zur Hypothesenverknüpfung und zur Erstellung eines für reale Sprache relevanten Aussprachewörterbuchs wurden Sprachmaterialuntersuchungen durchgeführt. Dazu war Material erforderlich, das von Phonetikexperten akustiknah, d. h. möglichst ohne Verarbeitung, mit Lautlabeln versehen wurde. Zu Beginn der Arbeiten stand dafür das in [pd92] beschriebene Material zur Verfügung. Die Untersuchung der Aussprachevarianten wurde auf der Basis dieses Materials durchgeführt. Die Ergebnisse konnten an dem jetzt zur Verfügung stehenden spontansprachlichen Material verifiziert werden. Die Untersuchungen zu Wortverschmelzungen wurden auf der Basis der Transliterationen der spontansprachlichen VERBMOBIL-Dialoge vorgenommen.

7.1 Untersuchung von Aussprachevarianten

Zur Ermittlung von repräsentativen Abweichungen von der Standardaussprache wurde reales Sprachmaterial untersucht. Der verwendete Korpus war das feingelabelte Sprachmaterial der Domäne IC-Auskunft [pd92]. Der Umfang betrug 185 Wortformen in 64 Sätzen. Es wurden die Realisierungen von 11 Sprechern und zwei kanonische Transkriptionen nach [aw82] und [du84] betrachtet. Das Gesamtmaterial umfaßte 7514 Realisierungen wobei die 185 Wortformen in 875 Varianten auftraten. Die beobachteten Abweichungen von der Standardlautung wurden klassifiziert und nach Häufigkeit sortiert. Dabei ergab sich folgende Beschreibung der Abweichungen:

1 Ausfall des Glottis-Lautes [QInt6slti: - Int6slti:]

2 'schwa'-Elision [gu:t@n - gu:tn]

3 Varianten bei Vokal - r -Verbindungen [dys@ldOrf - dys@ldO6f]

4 Änderung der Vokaldauer [E - @, I - i:,i]

5 Änderung der Vokalqualität [I - @, E: - e:]

6 stimmhaft - stimmlos -Änderung [v - f]

7 Nasalassimilation [ge:b@n - ge:bm]

8 Lautverschmelzung bei gleichem Silbenaus- und -Anlaut [ts|ts - ts]

9 orthographische Lautveränderungen [IC -Ik, UN - UNk]

10 Laut- und Silbenelision [IC - C, StUnd@n - StUn]

Die Verteilung der Varianten ist sehr unterschiedlich. Auf die Varianten 1 und 2 entfallen 50.7%, auf 3 5.2 %, auf 4 1.8 %, auf 5 2,4 % und auf 5 und 6 jeweils 0.8 %. Alle weiteren Abweichungen traten nur vereinzelt auf.

7.1.1 Formalisierung der Aussprachevarianten

Die beobachteten Aussprachvarianten werden mit dem Ziel ausgewertet, eine domänenunabhängige Spezifikation von Lautelisionen und Lautsubstitutionen aufzustellen. Dazu wird zunächst die Übereinstimmung der Beobachtungen mit den durch phonologische Regeln [ra91] beschriebenen möglichen Aussprachevariationen überprüft.

Anwendung phonologischer Regeln Phonologische Regeln beschreiben das Wissen über die Verknüpfung von Phonemen zu möglichen Sequenzen und beinhalten damit auch Beschreibungen möglicher Aussprachevariationen. Die erfaßten Aussprachevarianten werden unterteilt nach [ra91]:

- Tilgung von Segmenten (Elision)
- Hinzufügung von Segmenten (Epenthese)
- Veränderung von Segmenten in ihren Merkmalen (Assimilation)
- Umstellung von Segmenten (Metathese).

Obwohl in dem untersuchten Text alle Variationsmöglichkeiten vorkamen, wurden zunächst nur die häufigsten mittels phonologischer Regeln modelliert. Besonderes Augenmerk wurde dabei auf Lautelisionen gerichtet. Diese sind nur über explizite Variantenbeschreibungen zu modellieren, da in diesem Fall eine Lautgruppierung mit „kein Laut" nicht möglich ist.
Phonologische Regeln haben die folgende allgemeine Form:

$$A \longrightarrow B \qquad | \qquad X\_Y$$

Die eigentliche Regel lautet demnach *ersetze A durch B, wenn A im Kontext X, Y auftritt.* Der Kontext kann dabei nur einseitig oder garnicht spezifiziert sein. Anhand der Untersuchungsergebnisse für reale Sprache wurden folgende Ausspracheveränderungen modelliert:

- 'schwa'-Elision am Silbenauslaut vor m, n und l
- Nasalassimilation (nach 'schwa'-Elision)
- Plosivlautreduktion an Silbengrenzen
- Varianten bei Vokal - r - Verbindungen

Die potentiellen Ausspracheveränderungen wurden dabei jeweils als zusätzliche Variante mit in das Wörterbuch aufgenommen. Der daraus resultierende Zuwachs des Wörterbuchumfangs betrug für die betrachtete Domäne IC - Auskunft 63 % (Erweiterung von 918 Einträgen auf 1494 Einträge).

Lautgruppierung Zur Vermeidung eines weiteren Anwachsens des Lexikonumfangs erfolgt die Beschreibung der durch Lautsubstitution bedingten Varianten durch Lautgruppierung. Das Ziel dabei besteht in der geeigneten Zusammenfassung des ursprünglichen Phonembeschreibungsinventars von 64 Symbolen zu Gruppen, die jeweils ähnliche Laute beinhalten. Die Entscheidung über die Lautähnlichkeit wurde hier anhand der beobachteten Lautverwechslungen in realer Sprache getroffen.Es wurden folgende drei verschiedene Lautgruppierungen aufgestellt:

- **Gruppierung 1:** Zusammenfassung der Vokallängen und Abbildung der Vokal - r - Verbindungen auf den Vokal (29 Phonemgruppen)
- **Gruppierung 2:** wie 1 und Zusammenfassung stimmhaft - stimmlos - Unterscheidungen (25 Phonemgruppen)
- **Gruppierung 3:** wie 2 sowie Nasalgruppierung und Zuordnung von C zur Gruppe (g, k)und Zuordnung von S zur Gruppe (z, s)(21 Phonemgruppen)

Die Lautgruppierung stellt eine Verunschärfung der Referenzmodelle dar, die neben der Erhöhung der Robustheit ein Anwachsen der Worthypothesen für einen Signalabschnitt zur Folge haben kann, da Wörter, die sich nur durch Laute unterscheiden, die zu einer Gruppe zusammengefaßt wurden, die gleiche Modellrepräsentation bei unterschiedlichen Orthographien aufweisen. Für eine Phonemgruppierung muß daher untersucht werden, welche Wortgruppierungen über einem Wortinventar dadurch hervorgerufen werden. Tabelle 10 zeigt die entstehende Wortgruppenbildung für das um die Aussprachvarianten erweiterte Wörterbuch für die drei erwähnten Labelgruppierungen.

	Elemente pro Gruppe				
	1	2	3	4	5
Labelgruppierung 1	1421	32	3	-	-
Labelgruppierung 2	1411	37	3	-	-
Labelgruppierung 3	1388	45	2	-	2

Tabelle 10: Beispiel für Wortgruppierungen über verschiedenen Wortinventaren

Ergebnisse von Erkennungsexperimenten mit und ohne Modellierung von Aussprachevarianten wurden durchgeführt und sind in [fl94] und [fl95c] dargestellt.

7.2 Untersuchung von Wortverschmelzungen

Für die Untersuchungen zur Modellierung von Wortverschmelzungen wurden die VERBMOBIL-Dialogtransliterationen verwendet. Zum Untersuchungszeitpunkt standen sechs CD's mit deutschsprachigen Dialogen zur Verfügung. Das Material umfaßte 459 spontansprachliche Dialoge.

Zur Vorbereitung der Untersuchungen wurden die Dialoge zunächst in einzelne Turns zerlegt. Die Einzelturns wurden entsprechend dem akustischen Sprachdatenmaterial mit korrespondierenden Dateinamen gekennzeichnet. Alle nichtsprachakustischen Zusatzinformationen wurden herausgefiltert. In dieser Form stehen die Turntransliterationen als Referenzinformation für Erkennungsexperimente zur Verfügung. In diesen Darstellungen sind alle in den Transliterationen aufgezeichneten spontansprachlichen Erscheinungen erhalten geblieben. Sie enthalten Aussprachevarianten, Wortabbrüche, Wortverschmelzungen, Häsitationen und Nichtwörter. Aus diesem Material wurde durch satzweise Verarbeitung eine Wortliste erstellt, die durch die kanonischen phonetischen Transkriptionen ergänzt wurde. Diese Wortliste umfaßt 7171 Einträge. Innerhalb dieser Wörter bzw. Einträge wurden 206 verschiedene Wortverschmelzungen gezählt, die im Gesamtmaterial 1033 mal auftraten. Diese Wortverschmelzungen wurden in ihre Bestandteile aufgelöst und analysiert (vgl. dazu Anhang D).

7.2.1 Beobachtete Verschmelzungstypen

Entsprechend den zu Wortverschmelzungen führenden schwachen Typen und den auf sie anwendbaren Koartikulations- und Steuerungsregeln (siehe Abschnitt 6.1.3) wurden die Bestandteile der aufgelösten Wortverschmelzungen kategorisiert und nach Typen sortiert. Dabei konnten folgende Gruppen gebildet werden:

- **TYP 1:** (Verb/Hilfsverb) + Pronomen

Beispiel:	*erinnersch*	brauchen wir
	brauchma	erinnere ich
	hobs	habe es

- **TYP 2:** (Verb/Hilfsverb) + Pronomen + (Personalpronomen
Temporaladverb
Konjunktion
Artikel
Reflexivpronomen
Modaladverb)

Beispiel:	*könntmas*	könnten wir es
	hoffmamol	hoffen wir einmal

- **TYP 3:** Präposition + Artikel

Beispiel:	*aufs*	auf das
	ausm	aus dem
	fürs	für das

- **TYP 4:** Konjunktion + Hilfsverb + (Artikel
Personalpronomen)

Beispiel:	*asos*	also ist (das/es)

- **TYP 5:** Hilfsverb + Artikel

Beispiel:	*ischtn*	ist ein
	isa	ist ein
	isn	ist ein

- **TYP 6:** Interrogativpronomen + (Verb) + (Pronomen)

Beispiel:	*wises*	wie ist es
	wis	wie es / wie ist
	wers	wer ist / wer es

- **TYP 7:** Kausaladverb + Artikel

Beispiel: *wegem* wegen dem

- **TYP 8:** Präposition + Artikel

Beispiel: *vomir* von mir

- **TYP 9:** Artikel + Adverb

Beispiel: *abißl* ein bißchen

- **TYP 10:** Adverb + Adverb

Beispiel:	*aunich*	auch nicht
	auma	auch einmal
	aunoch	auch noch

- **TYP 11:** Konjunktion + Interrogativpronomen

  Beispiel: *awie* aber wie

- **TYP 12:** Hilfsverb + Präposition + Artikel

  Beispiel: *bim* bin mit dem

- **TYP 13:** Pronomen + (Verb/hilfsverb)

  Beispiel: *mimissn* wir müssen

- **TYP 14:** Pronomen + Adverb

  Beispiel: *cheute* ich heute

- **TYP 15:** Temporaladverb + Artikel

  Beispiel: *mahn* (ein)mal einen

- **TYP 16:** Verkürzung

  Beispiel: *drumrum* darumherum

- **TYP 17:** Lautanpassung und Geminatenreduktion

Beispiel:	*gudaß*	gut daß
	gudann	gut dann

7.2.2 Vorschlag zur Modellierung von Wortverschmelzungen

Bei der Modellierung von Wortverschmelzungen muß domänenspezifisch vorgegangen werden. Zum einen ist der vom Sprecher gewählte Sprechstil stark von der vorliegenden Domäne abhängig, zum anderen haben die Untersuchungen gezeigt, daß sich die häufiger auftretenden Verschmelzungen auf bestimmte Sprachabschnitte konzentrieren. Zum Teil nehmen sie als "Wortkonglomerate" die Funktion von Füllwörtern ein.

Die Wortverschmelzungen sind zum großen Teil nur auflösbar, wenn sie durch spezielle Modelle bzw. Verknüpfungsregeln beschrieben werden. Da über ihre syntaktisch-semantische Bedeutung erst nach ihrer Auflösung und Verarbeitung entschieden werden kann, muß zu den Spezialmodellen noch die vollständige orthographische Repräsentation zur Verfügung stehen.
Für die Modellierung wird folgendes Vorgehen vorgeschlagen:

- Nach Analyse der Domäne werden aus der zugörigen kategorisierten Wortliste die Wörter bzw. Wortkombinationen ausgewählt, die schwache Formen bilden können. Auf diese Einzelwörter oder Wortgruppen werden dann die vom weiteren Kontext unabhängigen Koartikulations- und Steuerungsregeln stufenweise angewendet. Die entstehenden Formen werden jeweils als Ergänzung in das Aussprachewörterbuch aufgenommen.

- Da die Generierung der Verschmelzungen viele neue Formen in das Aussprachelexikon einfügt, muß ein geeigneter Steuerungsmechanismus zur Verfügung stehen. Hier könnte die Frequenz der betrachteten Wörter bzw. Wortverbindungen über einem großen Korpus ermittelt und einbezogen werden. Weiterhin kann unter den angewendeten Regeln eine Hierarchie definiert werden, oder sie könnten mit einer Bewertung versehen werden.

- Eine entscheidende Größe bei der Generieriung der Wortverschmelzungen ist der dadurch bedingte Zuwachs des Wörterbuchumfangs. In Wechselwirkung mit den in Punkt 2 erwähnten Steuerungsmöglichkeiten liefert der angestrebte maximale Umfang des Aussprachewörterbuchs das Abbruchkriterium bei der regelbasierten, domänenspezifischen Generierung von Wortverschmelzungen.

7.3 Erstellung von HTK-Datenbasen für Evaluierungstests

Es bestand die Aufgabe, mehrere verschiedene, im VERBMOBIL eingesetzte Vorverarbeitungsalgorithmen zu evaluieren. Dazu sollte das kommerziell verfügbare Hidden-Markov-Models-Toolkit eingesetzt werden. Für diese Erkennungsexperimente mit dem HTK wurden spezielle Referenzdaten benötigt, die mit Hilfe von im Rahmen des Teilprojektes 3.3 erarbeiteten Funktionen generiert wurden. Die erwähnten Referenzdaten umfassen:

- die zu verarbeitenden Sätze (Lernphase, Testphase) in orthographischer Form

- eine Liste der in den Sätzen enthaltenen Wörter
- ein Ausspracheweörterbuch für die in den Sätzen enthaltenen Wörter
- eine Wortpaargrammatik über den Input-Sätzen

Im Folgenden werden die in Abbildung 10 gezeigten Softwarekomponenten zur Generierung dieser Informationen aus Texten vorgestellt.

7.3.1 Zerlegen der Dialogtransliterationen

Für die VERBMOBIL-Dialoge zur Terminabsprache liegen jeweils dialogweise vollständige Transliterationsfiles vor. Da die Zeitfunktionen jeweils nur einen Turn umfassen, müssen die Transliterationsfiles ebenfalls in einzelne Turns zergliedert werden. Dabei muß auf Übereinstimmung der Dateinamen geachtet werden, um eine automatisierte Verarbeitung zu unterstützen. Zur Zerlegung der Transliterationen steht die Funktion `trl_split` zur Verfügung. Diese Funktion liest ein Transliterationsfile, bestimmt aus der Sprecheridentifikation den Namen der Ausgabedatei für den aktuellen Turn und gibt die gelesene Datei turnweise in das angegebene Zielverzeichnis aus. Der Aufruf erfolgt mit folgendem Kommando:

```
use :  trl_split [-cx] <in_dat> <out_dat>
    ==>  Zerlegung von Transliterationsfiles in Turns
          [-cx] CD-Nummer
          <in_dat> Standard-Dialog-Transliteration
          <out_dat> Verzeichnis fuer Turns
```

Bei der Erzeugung der Dateinamen müssen folgende Besonderheiten berücksichtigt werden:

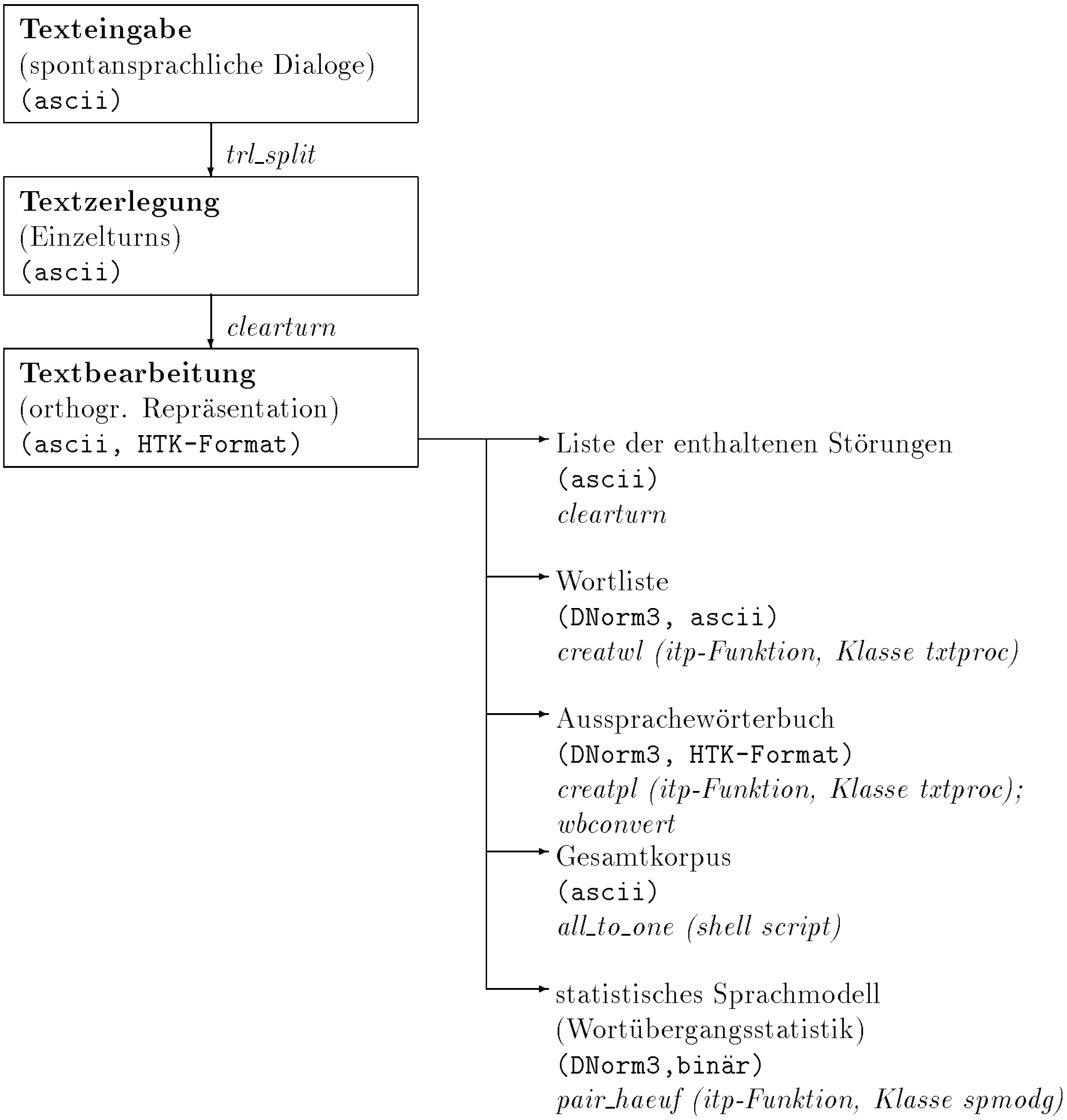
- CD1
  - trl-Files g***a.trl

    Numerierung der Turns beginnt mit 000, ein Dialog pro File, Extension der Zeitfunktionen: .a16
  - trl-Files M***D.trl

    Numerierung der Turns beginnt mit 000, mehrere Dialoge pro File, Extension der Zeitfunktion repräsentiert Dialog: .a16, .b16, .c16 usw.
  - trl-Files N***K.trl

    Numerierung der Turns beginnt mit 001, ein Dialog pro File, Extension der Zeitfunktionen: .a16
- CD2
  - trl-Files G***A.TRL

    Numerierung der Turns beginnt mit 000, ein Dialog pro File, Extension der Zeitfunktionen: .a16



fett ... Bearbeitungsstufe
roman... Ergebnis
`type` ... Datenformat
ital ... Funktion

Abbildung 10: Module zur HTK-Datengenerierung und deren Wechselwirkung

  - trl-Files M***D.TRL
    Numerierung der Turns beginnt mit 000, ein Dialog pro File, Extension der Zeitfunktion repräsentiert Dialoginhalt: .a16, .b16, .c16, .d16 und muß per Hand editiert werden.
  - trl-Files N***K.TRL
    Numerierung der Turns beginnt mit 000, ein Dialog pro File, Extension der Zeitfunktionen: .a16

- CD3
  - trl-Files G***A.TRL
    Numerierung der Turns beginnt mit 000, ein Dialog pro File, Extension der Zeitfunktionen: für ersten Sprecher .a16, für zweiten Sprecher .b16
  - trl-Files M***D.TRL
    Numerierung der Turns beginnt mit 000, ein Dialog pro File, Extension der Zeitfunktion: .a16

Als Ergebnis der Zerlegung liegen die einzelnen Turns als Files vor. Es ist darauf zu achten, daß die Transliterationen standardgemäß vorgenommen wurden. Es wurde beobachtet, daß Fehler auftreten, wenn am Zeilenanfang statt der laut Standard geforderten Leerzeichen Tabulatoren verwendet werden. Die Originaltransliterationen wurden entsprechend korrigiert. Die Generierung der Filenamen erfolgt, soweit es automatisch möglich ist, entsprechend der Namen der Zeitfunktionsfiles. Bei der zweiten CD sind für die mit M beginnenden Transliterationen Korrekturen von Hand erforderlich, da die Wahl der Zeitfunktionsdateinamen dialoginhaltsspezifisch erfolgte.

Die vereinzelten Turns werden weiterverarbeitet mit dem Ziel, eine reine akustiknahe graphemische Repräsentation der Äußerung zu erhalten.

7.3.2 Bereinigen der Dialogturntransliterationen

Die vorliegenden Beschreibungsfiles der Turns enthalten noch die bei der Verschriftung zulässigen Zusatzinformationen:

```
ja <A> , also f"ur den eint"agigen , wenn wir den als
<:<Lachen> erstes:> erledigen wollten , quasi , w"are mir
ganz recht Montag der achte November .
```

Die bereinigten Dateien sollen nur noch die im Aussprachewörterbuch enthaltenen Wörter enthalten, wobei auch Häsitationen, Wortabbrüche, Verschmelzungen, Aussprachevarianten und Einzellautrealisierungen enthalten sein können. Das Repräsentationsinventar wird bei der Definition des Aussprachewörterbuchs festgelegt.

```
ja   also f"ur den eint"agigen  wenn wir den als  erstes
erledigen wollten  quasi  w"are mir ganz recht Montag der
achte November
```

Die Bereinigung der Dialogturntransliterationen erfolgt in einem mehrstufigen Prozeß. Zunächst wird aus der gesamten Datenbasis eine Liste nichtlexikalischer Zeichenketten erzeugt, die als Zusatzinformationen bei der Transliteration verwendet wurden. Im nächsten Schritt werden spezielle nichtlexikalisierte Strings aus den Transliterationen entfernt. Im dritten und letzten Schritt werden die akustiknahen graphemische Repräsentationen erzeugt, indem alle noch vorhandenen Zusatzzeichen, einschließlich der Satzzeichen, entfernt werden. Diese Schritte werden durch die Funktion `filt_nonlex` mit verschiedenen Parametern realisiert. Der Aufruf erfolgt mit dem Kommando:

```
use :  filt_nonlex [-g] <infile> <listfile>
use :  filt_nonlex [-e] <listfile> <infile> <outfile>
use :  filt_nonlex [-k] <infile> <outfile>
-g  :  generate list of nonlexical strings
-e  :  eliminate special nonlexical strings
-k  :  complete orthographic sentences
```

Die aus dem CD-Inventar erzeugten Listen zur Streichung artikulatorischer bzw. nichtartikulatorischer Störungen sowie Sonderzeichen stehen extern zur Verfügung. Die Dialogturns sind damit als ascii-Files, die nur die akustiknahe graphemische Repräsentationen enthalten, verfügbar. Für Ihre Anwendung im HTK müssen sie in ein spezielles Format konvertiert werden. Diese Konvertierung wird durch die Funktion `wbconvert` realisiert. Der Aufruf erfolgt mit dem Kommando:

```
use :  wbconvert [-l] [-s] [-p] <in_dat> <out_dat> <help_dat>
-l  :  Konvertieren des Ausssprachewoerterbuchs
       <in_dat> DNorm-Ausssprachewoerterbuch
       <out_dat> HTK-Ausssprachewoerterbuch
       <help_dat> aktuelle Labeltabelle
-s  :  Konvertieren von Saetzen in HTK-Format
       <in_dat> Satz in ascii-Repraesentation
       <out_dat> Satz in HTK-Repraesentation
-p  :  Darstellung von Saetzen als Phonemlabelstrings
       <in_dat> Satz in  HTK-Repraesentation
       <out_dat> Satz als Phonemlabelstring
       <help_dat> HTK-Ausssprachewoerterbuch
```

7.3.3 Erzeugen einer aktuellen Wortliste

Die von Zusatzinformationen bereinigten Dialogturntransliterationen werden zur Bildung einer aktuellen Wortliste verwendet. An dieser Stelle können auch beliebige andere Texte, die jedoch nur die im Ausspracheswörterbuch zu repräsentierenden Sprachbestandteile (Wörter, Nichtwörter, Aussprachevarianten, Häsitationen, Abbrüche, Einzellautrealisierungen) enthalten dürfen. Der bzw. die Texte werden einzeln verarbeitet, wobei eine Wortliste mit Gesamtvorkommenshäufigkeit im Text erstellt wird. Als Ergebnis liegt eine aktuelle Wortliste als DNorm3-File vor, die die Wörter als Strings einschließlich ihrer Vorkommenshäufigkeit im Text enthält.

7.3.4 Erzeugen eines aktuellen Aussprachewörterbuchs

Die erzeugte Wortliste wird zur Bildung eines domänenspezifischen Aussprachewörterbuchs verwendet. Aus einem vorhandenen Aussprachewörterbuch in DNorm3 werden die in der Wortliste enthaltenen Wörter herausgesucht und mit ihrer phonetischen Transkription in das aktuelle Aussprachewörterbuch übernommen. Wörter, die in der aktuellen Wortliste, aber nicht in dem allgemeinen Aussprachewöerbuch enthalten sind, werden ohne phonetische Transkription in das aktuelle Aussprachewörterbuch übernommen. Anschließend wird das domänenspezifische Aussprachewörterbuch in ein ascii-File konvertiert und mit einem Standardeditor interaktiv aufgefüllt. Das vervollständigte Aussprachewörterbuch wird anschließend wieder in DNorm3 rückkonvertiert. Das so erzeugte Aussprachewörtebuch ist in DNorm3 repräsentiert und muß für die HTK-Anwendung noch speziell konvertiert werden. dazu dient die Funktion `wbconvert` , die oben bereits in ihrer Anwendung vorgestellt wurde.

7.3.5 Erzeugen von Labelfiles für die einzelnen Dialogturns

Die Datenverarbeitung im HTK erfordert für die im Lernprozeß zu verarbeitenden Daten bzw. zur Berechnung einer Phonemerkennungsrate die Darstellung der Sprachäußerungen als Labelfiles. Die Darstellung beinhaltet den Startzeitpunkt, den Endzeitpunkt und das zugehörige Phonemlabel als Folge für die gesamte betrachtete Äußerung. Da kein hand- bzw. automatisch gelabeltes Material zur Verfügung steht, wird diese Kette aus der orthographischen Form erzeugt. Bei diesem Vorgehen muß vorausgesetzt werden, daß alle aus der akustiknahen graphemischen Repräsentation abgeleiteteten Phoneme in der dadurch ebenfalls festgelegten Reihenfolge vorkommen. Bei dieser Annahme wird weiterhin von der konkreten Phonemdauer abstrahiert, so daß die Startzeit und Endzeit nicht relevant sind und mit dummy-Werten belegt werden können. Die Erzeugung der Labelfiles wird durch die Funktion `wbconvert` realisiert. Der Funktionsaufruf ist bereits weiter oben erwähnt. Dabei werden die orthographisch repräsentierten Sätze mit Hilfe des Aussprachewörterbuches in Phonemfolgen konvertiert. Diese Phonemfolgen werden im Anschluß auf die für die Anwendung speziell definierte Labelgruppenmenge projiziert und mit fiktiver Start- und Endzeit versehen.

7.3.6 Erzeugen eines aktuellen Sprachmodells

Zur Verbesserung der Erkennerleistung werden Sprachmodelle eingesetzt, die Informationen über in der betrachteten Domäne zulässige Wortfolgen enthalten.

Sprachmodelle können sehr unterschiedlich realisiert werden. Neben einer expliziten Grammatikdefinition für die jeweilige Domäne kommen auch statistische Sprachmodelle zum Einsatz, die eine Aussage über die Wahrscheinlichkeit einer bestimmten Wortfolge ermöglichen. Im hier vorliegenden Fall wird das Sprachmodell als Wortübergangsstatistik realisiert. Dabei wird über einem möglichst großem Korpus die relative Häufigkeit bestimmt, mit der Wort i auf Wort j folgt. Dieses Maß wird bei der Bildung von Wortketten oder Wortgittern zur Bewertung möglicher Pfade verwendet. Für nicht beobachtete Wortpaare wird ein Standardwert angenommen, der die Verknüpfung dieses Paares zwar ermöglicht, ihm jedoch eine entsprechende geringe Bewertung zuweist. Zur Erzeugung der Wortübergangsstatistik werden Sätze der betrachteten Domäne verarbeitet. In einer Matrix werden die beobachteten Wortübergänge gezählt und anschließend wird zeilenweise normiert, so daß für jedes Wortpaar der Domäne ein Wert für seine Beobachtungshäufigkeit gefunden wird. Die resultierende Matrix der Wortpaarhäufigkeiten kann als DNorm3-File bzw. als Binärfile - speziell für die Anwendung im HTK - ausgegeben werden.

Literatur

[ai88] Ainsworth, W.A. *Speech Recognition by Machine.* Peter Peregrinus Ltd., London, 1988.

[aw82] *Großes Wörterbuch der deutschen Aussprache.* Leipzig: Bibliographisches Institut. 1982.

[al95] Altendorf, H.*Optimale Verkettung von Silbenhypothesen.* Diplomarbeit. TU Dresden. 1995.

[be94] Berlin, V. *Verarbeitung symbolischer Hypothesen.*Diplomarbeit.TU Dresden.1994.

[ch68] Chomsky, N., M. Halle. *The Sound Pattern of English.* Harper and Row. New York. 1968.

[cr73] Cravero, M., R. Pieraccini und F. Rainieri. *Definition and Evaluation of Phonetic Units for Speech Recognition by Hidden Markov Models.* Proc.ICASSP'86. Tokyo. 1986.

[di77] Dixon, N.R., H.F. Silverman. *The 1976 modular acoustic processor (MAP).* IEEE Trans. on Acoustics, Speech and Signal Processing. ASSP-25(5). 1977.

[du83] *Der Große Duden, Wörterbuch und Leitfaden der deutschen Rechtschreibung.* Autorenkollektiv. Leipzig. 1983.

[du84] *Der Duden in 10 Bänden Bd. 6: Aussprachewörterbuch.* Mannheim, Wien, Zürich. Bibliographisches Institut. 1984.

[fi88] Fissore, L. et all. *Strategies for Lexical Access to Very Large Vocabularies.* Speech Communication. 7(4). 1988.

[fl94] Flach, G. *Beschreibung von Aussprachevarianten.* in: Fellbaum, K. (Hrsg.) 5. Konferenz Elektronische Sprachsignalverarbeitung. Studientexte zur Sprachkommunikation. Berlin. 1994.

[fl95a] Flach, G. *Automatische Silbenzerlegung und Modellierung.* in: Fortschritte der Akustik Teil II. DAGA95. Saarbrücken. 1995.

[fl95b] Flach, G. Altendorf, H. *Optimale Verkettung von Silbenhypothesen.* in: Hoffmann, R., Ose, R. (Hrsg.). 6. Konferenz Elektronische Sprachsignalverarbeitung. Studientexte zur Sprachkommunikation. Wolfenbüttel. Sept. 1995.

[fl95c] Flach, G. *Modelling Pronunciation Variability for Special Domains.* Proceedings EUROSPEECH'95. Madrid. 1995.

[fr90] Frauenfelder, U. Peeters, G. *Lexical Segmentation in TRACE: An Exercise in Simulation.* in: Altmann, G. T. M. Cognitive Models of Speech Processing. MIT Press. 1990.

[ht93] Young, S.J., Woodland, P.C., Byrne, W.J. *HTK: Hidden Markov Model Toolkit V1.5 - User Manual.* Cambridge. University Engeneering Department Speech Group & Entropic Research Labs Inc. 1993.

[hu84] Huttenlocher, D.P., Zue, V.W. *A model of lexical access from partial phonetic information.* Proceedings IEEE. ICASSP. San Diego. 1984.

[je90] Jelinek, F. *Self-Organized Language Modeling for Speech recognition.* in WaibelA., Lee, K.-F. (Hrsg). Readings in Speech recognition. New York. 1990.

[je93] Jelitto, J. *Exprimente zur Lautübergangsdetektion.* internes Arbeitspapier. TU Dresden. 1993.

[jo93] Jokisch, O. *Automatische Sprechsilbenzerlegung.* Großer Beleg. internes Arbeitspapier. TU Dresden. 1993.

[ko90] Kohler, K.J. *Segmental reduction in connected speech in german: phonological facts and phonetic explanations.* Dordrecht, Boston, London. Kluwer Academic Publishers. 1990.

[ko95] Kohler, K. *Einführung in die Phonetik des Deutschen.* Berlin. Schmidt. 1995.

[le88] Lee, K.-F. *Large-Vocabulary Speaker-Independent Continuos speech Recognition: The SPHINX System.* Report. CMU. April 1988.

[le89] Lee., K.-F. *Automatic Speech Recognition, The Development of the SPHINX System.* Kluwer Academic Publishers. Boston, Dordrecht, London. 1989.

[lo80] Lowerre, B., Reddy, D.R. *The HARPY Speech Understanding System.* in: Lee, W.A.(Hrsg). Trends in Speech Recognition. Prentice-Hall Inc. Englewood Cliffs. New Jersey. 1980.

[ma94] Marijczuk, P. *Erstellung einer Geräuschdatenbank.* Diplomarbeit. TU Dresden. 1995.

[me33] Menzerath, P., de Lacerda. *Koartikulation, Steuerung und Lautabgrenzung.* Berlin, Bonn. 1933.

[me87] Merialdo, B. *Speech Recognition with Very Large Size Dictionary.* Proceedings ICASSP'87. Dallas. April 1987.

[or80] Ortmann, W.D. *Sprechsilben im Deutschen.* Goethe-Institut. München. 1980.

[pd92] *Aufbau einer Signaldatenbank für gsprochenes Deutsch.* PHONDAT- Verbundvorhaben der Universitäten Kiel, Braunschweig, München. 2. Zwischenbericht. 1992

[ra85] Rabiner, L. R., Levinson, S. E. *A Speaker-Independent, Syntax-Directed, Connected Word Recognition system Based on Hidden Markov Models and Level Building.* IEEE Trans. Acoustics, Speech and Signal Processing. 33(3). 1985.

[ra91] Ramers, K.-H, Vater, H. *Einführung in die Phonologie.* Hürth-Efferen. Gabel Verlag. 1991.

[ru92] Rudolph, T. *Sprachdatenverwaltung mit DNORM 3.0.*Verbundprojekt ASL. ASL-Süd-Bericht 13-92/TUD. TU Dresden. Dezember 1992.

[ru84] Ruske, G. *Halbsilben als Verarbeitungseinheiten bei der automatischen Spracherkennung.* Sprache und Datenverarbeitung. 8(1/2). 1984.

[ru88] Ruske, G. *Automatische Spracherkennung: Methoden der Klassifikation und Merkmalsextraktion.* Oldenbourg-Verlag. München-Wien. 1988.

[ru89] Ruske, G. *Gehörbezogene automatische Spracherkennung.* Informationstechnik. 31(5). 1989.

[ru91] Ruske, G., Weigel, W. *Explizite Wissensdarstellung für die silbenorientierte Analyse fließender Sprache.* 1991.

[sa90] Sagerer, G. *Automatisches Verstehen gesprochener Sprache.* Mannheim, Wien, Zürich. BI-Wissenschafts-Verlag. 1990.

[sc84] Schotola, T. *On the use of demisyllables in automatic word recognition.* Speech Communication. 3. 1984.

[st79] Studdert-Kennedy, M. *Speech Perception.* Status Report on Speech Research. SR-59/60. 1979.

[ti64] Tillmann, H.G. *Das phonetische Silbenproblen. Eine theoretische Untersuchung.* Dissertation. Bonn. 1964.

[ue94] Ueberla, J.P. *Analyzing Weakness of Language Models.* in: Trost, H. (Hrsg.) KONVENS '94. Wien, Berlin. 1994.

[vm90] Autorenkollektiv. *Verbmobil-Nordverbund-Studie.* 1990.

[wa87] Waibel, A. et al. *Phoneme Recognition using Time-Delay Neural Networks.* Technischer Bericht. ATR. Oktober 1987.

[we94a] Westendorf, C.-M. *Generierung von Worthypothesen und Satzerkennung - Verfahren und Programme.* ITA. TU Dresden. 1994.

[we94b] Westendorf, C.-M. *VSPN V1.0 Programmsystem zur Verarbeitung von Vektorfolgen.* Programmier- und Referenzhandbuch. ITA. TU Dresden. 1994.

[we94c] Westendorf, C.-M. *Datenanalyse mit itp und dana - Beispiele für Verfahren, Programme und Resultate.* ITA. TU Dresden. 1994.

A Phonemklasseninventar 1 (11 Klassen)

Gruppenbezeichnung	enthaltene Label (nach PHONDAT 2)
A (a-ähnlicher Vokal)	a:, a, 6, a:6, a6
E (e-ähnlicher Vokal)	e:, e, E:, E, 2:, 2, 9, @, e:6, E:6, E6
I (i-ähnlicher Vokal)	i:, i, I, y:, y, Y, i:6, I6, Y6, Y:6
U (dunkler Vokal)	o:, o, O, u:,u, U, o:6, O6, u:6, U6, U:, O:
. (stimmloser Plosiv)	p, k, t
- (stimmhafter Plosiv)	b, g, d
N (Nasal)	m, n, N
# (l und r)	l, r, R
* (stimmloser Frikativ)	s, S, f, x, C, h
$ (stimmhafter Frikativ)	z, v, j
+ (Pause)	Q, .

B Phonemklasseninventar 2 (29 Klassen)

Gruppenbezeichnung	enthaltene Label (nach PHONDAT 2)
a	a:, a, a:6, a6
e	e:, e, @, e:6
E	E:,E, E:6, E6
i	i:, i, I, i:6, I6
o	o:, o, O, o:6, O6, O:
u	u:, u, U, u:6, U6, U:
y	y:, y, Y, Y:6, Y6
2	2:, 2, 9
6	6
p	p
b	b
k	k
g	g
t	t
d	d
m	m
n	n, N
l	l
r	r, R
s	s
z	z
S	S
v	v
f	f
x	x
C	C
h	h
j	j
.	Q, .

C Aussprachevarianten im PHONDAT-II Material

Norm und Aussprachevarianten	Häufigkeit	Regel
UNg@fE:6		
Ung@fE:6	4	
UN@fe:6	1	
UNg@fe:6	2	
Un@fe:6	1	
QUnt		
Unt	36	
QUnt	31	
QU	1	
Un	13	
QUn	14	
Und	3	
QUnd	4	
n	2	
QUnt6brEC@n		
QUnt6brEC@n	1	
Unt6bRECn	5	
Unt6brECn	5	
Unt6brECN	1	
Qa:		
a:	2	
Qa:	11	
Qa:b6		
a:b6	12	
Qa:b6	9	
Qa:v6	2	
O6	1	a:$\rightarrow$ O b$\rightarrow$
a:v6	1	
Qa:b@nt		
a:b@nt	6	
Qa:b@nt	6	
Qa:bnt	8	
Qa:md	1	b@n$\rightarrow$ m
a:bnd	3	

Norm und Aussprachevarianten	Häufigkeit	Regel
Qa:mt	4	b@n→ m
Qa:bmt	4	b@n→ bm
Qa:bm	3	b@n→ bm
a:md	2	b@n→ m
abmnd	1	b@n→ bmn
a:bm	4	b@n→ bm
a:bnt	9	
a:bmt	8	b@n→ bm
a:mt	3	b@n→ m
a:bn	2	b@n→ bn
a:m	5	b@n→ m
a:bmnt	1	b@n→ bmn
a:bmd	1	b@n→ bm
a:m	1	b@n→ m
Qa:bmnt	1	b@n→ bmn
Qa:b@nts		
a:b@nts	2	
Qa:b@nts	2	
a:bns	4	
Qa:mts	2	b@n→ m
Qa:ms	2	b@n→ m
a:bnts	3	
Qa:bns	1	
Qa:bnts	3	
a:mts	3	b@n→ m
Qa:bms	2	b@n→ bm
a:ms	1	
a:bmts	1	b@n→ bm
Qa:x@n	2	
a:x@n	3	
Qa:x@n	2	
a:xn	4	
a:xN	2	
Qa:xn	2	
QaIn		
aIn	29	
QaIn	7	
QE	1	aI→ E
E	1	aI→ E

Norm und Aussprachevarianten	Häufigkeit	Regel
En	1	aI→ E
QaIn@		
aIn@	138	
QaIn@	36	
aIn	4	
a@	1	aI→ a
n@	2	aI→
an@	6	aI→ a
E	1	aI→ E
nE	1	aI→
QaI@	2	
En@	2	aI→ E
aI@	2	
QaIn@m		
aIn@m	3	
QaIn@m	2	
aIm	6	n@m→ m
aInm	1	
QaIm	1	n@m→ m
QaIn@n		
aIn@n	4	
QaIn@n	3	
aIn	16	n@n→ n
QEn	1	aI→ E n@n→ n
QaIn	2	n@n→ n
QaSaf@nbU6k		
QaSaf@nbU6k	1	
QaSafnbU:k	1	
QaSafnbU6k	2	
aSafnbUk	1	
aSafnbU6k	4	
QaSafmbU6k	2	f@n→ fm
aSafmbU6k	1	
QaUf		
aUf	7	
QaUf	3	
aUv	2	f→ v
QOf	1	aU→ O

Norm und Aussprachevarianten	Häufigkeit	Regel
QaUf@nthalt		
QaUf@nthalt	1	
aUfnthald	2	
aUfnthalt	9	
QaUksbU6k		
QaUksbU6k	5	
aUksbU:k	1	
aUksbUk	1	U6→ U
aUksbU6k	4	
aUksbOk	1	U6→ O
QaUskUnft		
aUskUnft	9	
QaUskUnft	2	
aUskUnf	2	
Qalzo:		
alzo	2	
Qalzo:	18	
Qalso:	2	
alzo:	1	
azo:	1	
Qaz	1	o:→
also:	1	z→ s
Qam		
am	87	
Qam	29	
Om	1	a→ O
QankOm@n		
ankOm@n	5	
QankOm@n	3	
ankOm	8	m@n→ m
QankOmn	2	
ankOmn	4	
aNkOm	2	m@n→ m
QankOm	2	m@n→ m
QankUnftstsaIt		
QankUnftstsaIt	5	

Norm und Aussprachevarianten	Häufigkeit	Regel
QankUnstsaIt	1	ft→
QankUftsaIt	1	ft→ f
QankUnftsaIt	2	ft→ f
QaNkUnfsaIt	1	ft→ f ts→s
QankUnftsaI	1	ft→ f
ankUnftsaIt	1	
Qap		
ap	21	
Qap	16	
ab	2	
Qapfa:6tstsaIt		
Qapfa:6tstsaIt	1	
Qapfa:6tstsaId	1	
apfa:6tsaIt	5	ts→
apfatsaIt	1	a:6→ a
apfa:6tstsaIt	2	ts→
Qapfa:6tsaI	1	ts→
Qapfa:tstsaIt	1	a:6→ a:
Qaxt		
axt	23	
Qaxt	25	
Qaxd	1	
axd	2	
ax	1	
b@n2:tIg@n		
b@n2:tIg@n	1	
b@n2:tIg@		
b@n2:tIg@	10	
b@n2tig@	1	2:→ 2 I→ i
b@n2:tIgE	1	@→ E
bIt@		
bIt@	13	
ba:d@n		
ba:d@n	4	

Norm und Aussprachevarianten	Häufigkeit	Regel
ba:dn	21	
ba:n	1	
ba:nf6bIndUN		
ba:nf6bIndUN	5	
ba:nf6bInUN	3	
ba:nf6bInUNk	4	N→ Nk
ba:nf6bIndUNk	1	N→ Nk
baI		
baI	13	
brOYCt@		
brOYCt@	18	
bROYCt@	18	
brOYCt	3	
braUx		
braUx	7	
bRaUx	5	
braUx@		
braUx@	21	
bRaUx@	18	
braUxt		
braUxt	3	
bRaUxt	1	
bRaUxd	5	
braUxd	4	
dEn		
dEn	9	
En	4	
dEs		
dEs	11	
d@s	2	E→ @
dIrEkt		
dIrEkt	3	

Norm und Aussprachevarianten	Häufigkeit	Regel
dIREkt	5	
diREkt	1	I→ i r→ R
di:rEkt	4	I→ i:
dO6t		
dO6t	7	
dO:t	1	O6→ O:
dO6d	1	
dOt	1	O6→ O
dO6	1	
dORt	1	O6→ OR
dO6tmUnt		
dOrtmUnt	5	O6→ Or
dO6tmUnt	23	
dO6tmUnd	2	
dOtmUnt	1	O6→ O
dO6dmUnd	2	
dO6dmUnt	4	
dO6tmUn	2	
dYs@ldO6f		
dYs@ldO6f	3	
dYs@ldO:f	1	O6→ O:
dYsldO6f	7	
dYsldOf	1	O6→ O
da:		
da:	24	
a:	2	
da6f		
da6f	10	
da:f	1	
da6v	1	f→ v
dan		
dan	51	
da	1	
das		

Norm und Aussprachevarianten	Häufigkeit	Regel
das	26	
de:m		
de:m	36	
dem	1	
dm	1	e:→
d@m	1	e:→ @
de:n		
de:n	37	
d@n	2	e:→ @
de:6		
de:r	5	e:6→ e:r
de:6	58	
e:6	1	
dE6	1	e:6→ E6
detsEmb6		
detsEmb6	3	
dEtsEmb6	3	e→ E
@tsEmb6	1	e→ @
dItsEmb6	1	e→ I
dEtsEm6	4	b→
d@tsEmv6	1	e→ @
di:		
di:	25	
ti:	1	
Qe:6		
e:6	9	
Qe:6	4	
Qe:6st@n		
Qe:6st@n	1	
Qe:6stn	5	
Qe:6sdn	1	
e:6sdn	1	
e:6stn	3	
e:6sn	1	

Norm und Aussprachevarianten	Häufigkeit	Regel
f6bIndUN		
f6bIndUN	19	
f6bInUN	18	
f6bIndUNk	1	N→ Nk
f6bInUNk	1	N→ Nk
f6bIndUN@n		
f6bIndUN@n	4	
f6bIndUN	2	N@n→ N
f6bInUN	4	N@n→ N
f6bIndUNn	1	
f6bInUN@	1	
f6bInUNn	1	
fE:6t		
fE:rt	2	E:6→ E:6
fE:6t	11	
fE:6d	5	
fe:6t	3	E:6→ e:6
fEd	1	E:6→ E
fe:6d	2	E:6→ e:6
f6d	2	E:6→ E
fI6tse:n		
fI6tse:n	7	
fI6tsen	1	
fI6ts@n	4	e:→ @
fOn		
fOn	204	
fn	3	O→
f@n	1	O→ @
fa:r@n		
fa:r@n	34	
fa:n	82	a:r→ a:
fa:R@n	28	a:r→ a:R
fa:rn	1	
faRn	1	a:r→ aR
fa:Rn	10	a:r→ a:R

Norm und Aussprachevarianten	Häufigkeit	Regel
fa:6t		
fa:rt	2	a:6→ a:r
fa:6t	19	
fa:t	4	a:6→ a:
fa:6d	1	
faI6ta:k		
faI6ta:k	12	
faI6tag	1	
fe:lt		
fe:lt	9	
fe:ld	2	
fe:l	2	
fi:rUnttsvantsIC		
fi:rUnttsvantsIC	2	
fi6UntsvantsIC	1	i:r→ i6
fi@UntsvantsIC	1	i:r→ i@
fi:6UnsvansIk	1	i:r→ i:6 ts→ s
fi:@ntsvantsIC	1	i:r→ i:
fI6UntsvantsIC	1	i:r→ I6
fi:6ntsvantsIC	3	i:r→ i:6
fi:6UntsvansIC	1	i:r→ i:6
fi:6UntsvantsIC	1	i:r→ i:6
fi:6@ntsvantsIk	1	i:r→ i:6
filaICt		
filaICt	2	
fIlaICt	22	i→ I
flaICt	1	i→
f@laICt	1	i→ @
fo:6		
fo:r	2	o:6→ o:r
fo:6	24	
fo:6mIta:k		
fo:rmIta:k	5	o:6→ o:r
fo:6mIta:k	41	
fo:6mIt6g	1	a:→ 6

Norm und Aussprachevarianten	Häufigkeit	Regel
fomItag	2	o:6→ o
fo:6mItak	4	
fo:6mItax	1	k→ x
fOmItak	1	o:6→ O
fomItak	1	o:6→ o
fOmIta:k	4	o:6→ O:
fomIta:k	2	o:6→ o
fO:mItak	1	o:6→ O:
fo:6mIta:ks		
fo:6mIta:ks	10	
fO:mIta:ks	1	o:6→ O:
fOmIta:ks	1	o:6→ O
fraIta:k		
fraIta:k	6	
fRaIta:g	1	
fRaItak	2	
fraItak	1	
fRaIta:k	2	
fraIta:	1	k→
fraNkfU6t		
fraNkfUrt	7	
fraNkfU6t	35	
fRaNfU6t	2	k→
fRaNkfU6t	28	
fraNfU6t	8	k→
fRaNfU6d	2	k→
fraNfU6	1	k→
fRaNkfUt	5	U6→ U
fRaNfUt	2	k→ U6→ U
fRaNkfU6Rt	1	U6→ U6R
fry:		
fry:	29	
fRy:	21	
fRy	2	y:→ y
fy:6		
fy:r	4	y:6→ y:r

Norm und Aussprachevarianten	Häufigkeit	Regel
fy:6	48	
gE6n		
gE6n	10	
gE6nE	1	$\rightarrow$ E
gE6n@		
gE6n@	10	
gE6n	2	
gi:pt		
gIpt	3	
gi:pt	19	
gipt	2	
gi:bt	1	
gibt	1	
ge:b@n		
ge:b@n	2	
ge:bm	7	b@n$\rightarrow$ bm
gebn	1	
ge:m	1	b@n$\rightarrow$ m
ge:bn	2	
ge:g@n		
ge:g@n	6	
ge:gN	22	
ge:gn	7	
ge:N	4	g@n$\rightarrow$ N
ge:t		
ge:t	82	
ge:d	15	
ge:b	1	t$\rightarrow$ b
ged	3	
get	3	
glaIs		
glaIs	13	
grEntsba:nho:f		
grEntsba:nho:f	4	

Norm und Aussprachevarianten	Häufigkeit	Regel
gREnsba:nho:f	1	ts→ s
gREntsba:no:f	1	h→
gREntsba:nhof	1	
grEntsba:nOf	1	h→ o:→ O
gREntsba:nho:f	3	
grEnsba:nho:f	2	ts→ s
gu:t		
gu:t	13	
gu:t@n		
gu:t@n	7	
gutn	2	u:→ u
gu:dn	4	
gudn	3	u:→ u
gu:tn	19	
gUdn	2	u:→ U
gUtn	1	u:→ U
gun	1	u:→ u
hEt@		
hEt@	23	
hEd@	2	
Et@	1	h→
hIn		
hIn	13	
hInd@laN		
hInd@laN	31	
hIn@laN	8	
hInd@laNk	5	N→ Nk
hIn@laNk	4	N→ Nk
hInd@laNg	3	N→ Ng
hn@laN	1	
hOYt@		
hOYt@	98	
hOYd@	1	
hOYtEO	1	@→ EO
hOYte	2	@→ e

Norm und Aussprachevarianten	Häufigkeit	Regel
hOYt	2	
hambU6k		
hambUrk	11	
hambU6k	109	
hambU:k	4	
hambUk	2	U6$\rightarrow$ U
xambU6k	2	h$\rightarrow$ x
hamU6k	4	b$\rightarrow$
hambU6g	5	
hamU6	1	b$\rightarrow$ k$\rightarrow$
hambU6Q	1	k$\rightarrow$ Q
hambo:k	1	U6$\rightarrow$ o:
hambU6	1	k$\rightarrow$
hambUg	1	U6$\rightarrow$ U
hambo:g	1	U6$\rightarrow$ o: k$\rightarrow$ g
hano:f6		
hano:f6	21	
xano:f6	1	h$\rightarrow$ x
hanof6	2	
hano:v6	2	f$\rightarrow$ v
Qi:tse:		
Qi:tse:	2	
i:tse:	9	
itse	1	
jEtst		
jEtst	9	
jEtsd	1	
jEts	3	
ja:		
ja:	77	
ja	1	
je:d@nfals		
je:d@nfals	2	
je:dnfas	1	
je:nfals	5	

Norm und Aussprachevarianten	Häufigkeit	Regel
jednfals	1	
je:nfas	1	
je:dnfals	3	
k9ln		
k9ln	64	
k9n	1	
k9n@n		
k9n@n	2	
k9n	11	n@n→ n
k9nt@n		
k9nt@n	2	
k9n	4	t@n→
k9ndn	1	
k9ntn	3	
k9nQn	2	t@→ Q
kOm@n		
kOm@n	17	
kOm	16	m@n→ m
kOmn	6	
kan		
kan	90	
ka	14	
ko:blEnts		
ko:blEnts	14	
ko:blEns	9	
koblEns	2	
kOblEnts	1	o:→ O
lEtst@		
lEtst@	10	
lEts@	3	
laUf@		
laUf@	13	
li:g@va:g@n		

Norm und Aussprachevarianten	Häufigkeit	Regel
li:g@va:g@n	2	
li:g@va:N	3	g@n→ N
lig@va:gN	1	
li:g@va:gn	3	
li:g@va:gN	3	
li:j@va:gN	1	
lo:sfa:r@n		
lo:sfa:r@n	3	
lo:sfa:n	6	a:r→ a:
lo:sfa:6n	1	a:r→ a:6
lo:sfa:R@n	3	
m2:klICst		
m2:klICst	26	
m9klIS	1	2:→ 9 Cst→ S
m2klIC	1	2:→ 2 st→
m2klICs	1	2:→ 2
m2:klICs	4	
m2:klIC	2	st→
m2:klIS	1	Cst→ S
m2:glICs	1	
m2:klICSt	1	
m2:klIst	1	
m2klICst	1	2:→ 2
m2:kdICst	1	
m2:glICst	1	
m2:klIsd	1	
m2klICsd	1	2:→ 2
m2:klICsd	7	
m2:glICsd	1	
m9Ct		
m9Ct	13	
m9Ct@		
m9Ct@	124	
m9Cd@	1	
m9Ct	5	

Norm und Aussprachevarianten	Häufigkeit	Regel
mInd@st@ns		
mInd@st@ns	2	
mIn@Sdn	1	s→
mIn@sns	1	
mIn@sn	3	s→
mIn@stn	1	s→
mInd@stns	4	
mIn@stns	1	
mIt		
mIt	35	
mId	40	
mI	1	
mIQ	1	t→Q
mO6g@n		
mOrg@n	8	O6→ Or
mO6g@n	9	
mO:N	3	O6→ O: g@n→ N
mO6gN	44	
mOgN	6	O6→ O
mO6	1	g@n→
mO6gn	12	
mO6N	15	g@n→ N
mO:gN	4	O6→ O:
mO6Ng	1	g@n→ N → g
mO6Rgn	1	
mO6g@ns		
mOrg@ns	3	O6→ Or
mO6g@ns	3	
mOgns	1	O6→ O
mO6Ns	9	g@n→ N
mOgNs	3	O6→ O
mO6gns	4	
mO6gNs	14	
mO6gnts	1	s→ ts
mO6g@s	1	
mUs		
mUs	52	

Norm und Aussprachevarianten	Häufigkeit	Regel
mYnC@n		
mYnC@n	52	
mYnCn	51	
mYnCN	1	
man		
man	64	
mam	1	
manhaIm		
manhaIm	13	
me:6		
me:6	11	
m6	1	e:6→ 6
mi:6		
mi:r	3	i:6→ i:r
mi:6	36	
mo:nta:k		
mo:nta:k	20	
mo:nt6k	1	a:→ 6
montag	2	
mo:nta:g	2	
mo:ntax	1	k→ x
nE:Cst@		
nE:Cst@	4	
nE:kst@	3	
nEks@	1	
nECst@	1	
ne:Cs@	2	
ne:st@	2	
nE:Cst@n		
nE:Cst@n	2	
nE:kStn	1	
nE:ksdn	1	
neCsn	1	E:→ e
ne:Cstn	3	
nE:Cstn	1	

Norm und Aussprachevarianten	Häufigkeit	Regel
ne:Csn	2	
nE:kstn	1	
ne:kstn	1	
nICt		
nICt	40	
nICd	5	
nIC	7	
nOYn		
nOYn	65	
nOx		
nOx	102	
nO	2	x→
nY6nbE6k		
nY6nbE6k	9	
nY6nbE6g	2	
Y6nbE6k	1	
na		
na	13	
na:x		
na:x	592	
na:	5	x→
Na:x	1	
na:xmIta:k		
na:xmIta:k	11	
naxmItak	1	
naxmIta:k	1	
naIn		
naIn	13	
naja:		
naja:	9	
naja	1	
na:ja	1	a→ a:

Norm und Aussprachevarianten	Häufigkeit	Regel
aja:	1	
na:ja:	1	a→ a:
naxt		
naxt	13	
ne:m@		
ne:m@	5	
nem	1	
ne:m	6	
Qo:d6		
o:d6	26	
Qo:d6	11	
d6	1	o:→
o:l6	1	
Qo:n@		
o:n@	21	
Qo:n@	30	
on@	1	
Qo:ke:		
Qo:ke:	5	
Oke:	2	o:→ O
oke	1	
o:ke:	3	
QOke:	1	o:→ O
past		
past	12	
pas	1	
re:g@nsbUrk		
re:g@nsbUrk	12	
re:gNsbU6g	7	Ur→ U6
regNsbok	1	Ur→ o
re:gNsbU6k	26	Ur→ U6
re:gnsbU6k	18	Ur→ U6
re:NsbU6k	2	g@n→ N Ur→ U6
re:NsbU6g	4	g@n→ N Ur→ U6

Norm und Aussprachevarianten	Häufigkeit	Regel
regNsbU6g	3	Ur$\rightarrow$ U6
regNsbog	2	Ur$\rightarrow$ o
re:gNsbU6Q	1	Ur$\rightarrow$ U6 k$\rightarrow$ Q
re:gNsbOk	1	Ur$\rightarrow$ O
re:gNsbUk	1	Ur$\rightarrow$ U
ro:m		
ro:m	7	
Ro:m	5	
Rom	1	
ta:g@n		
ta:g@n	2	
ta:gN	9	
ta:N	1	g@n$\rightarrow$ N
ta:gn	1	
ta:k		
ta:k	26	
tse:n		
tse:n	48	
tsen	4	
tsu:k		
tsu:k	108	
tsuk	7	u:$\rightarrow$ u
tsug	9	u:$\rightarrow$ u
su:k	5	ts$\rightarrow$ s
tsu:N	1	k$\rightarrow$ N
tsu:g	7	
sug	4	ts$\rightarrow$ s u:$\rightarrow$ u
suk	2	ts$\rightarrow$ s u:$\rightarrow$ u
tsu:kf6bIndUN		
tsu:kf6bIndUN	45	
tsu:kf6bInUN	46	
sukf6bInUN	6	ts$\rightarrow$ s u:$\rightarrow$ u
tsu:kf6bIndUNk	4	N$\rightarrow$ Nk
tsukf6bInUN	14	u:$\rightarrow$ u
tsukf6bInUNg	1	u:$\rightarrow$ u N$\rightarrow$ Ng

Norm und Aussprachevarianten	Häufigkeit	Regel
tsu:kfE6bInUN	2	6→ E6
tsu:gf6bInUN	1	
su:kf6bInUN	1	ts→ s
tsu:kf6bInUNk	7	N→ Nk
tsukf6bIndUN	1	u:→ u
tsu:f6bInUN	1	k→
tsugf6bInUN	1	u:→ u
tsUrYk		
tsUrYk	9	
soRYk	1	ts→ s
sUrYk	2	ts→ s
tsvIS@n		
tsvIS@n	3	
tsvISn	8	
svISn	2	ts→ s
tsvIS@ndU6C		
tsvIS@ndU6C	1	
tsvISndUIC	2	U6→ UI
tsvISndU6C	9	
tsva:6		
tsva:r	3	a:6→ a:r
tsva:6	26	
tsva:	6	a:6→ a:
tsv6	1	a:6→ 6
tsfa	1	a:6→ a
svO	1	ts→ s a:6→ O
sva	1	ts→ s a:6→ a
tsvaI		
tsvaI	9	
svaI	4	ts→ s
tsvaIUnttsvantsIC		
tsvaIUnttsvantsIC	4	
tsvaI@ntsvansIk	2	
tsvaIUntsvantsIC	6	
tsvaInsvanzIC	1	ts→ s ts→ z
tsvaIntsvantsIk	2	

Norm und Aussprachevarianten	Häufigkeit	Regel
tsvaIntsvansIC	2	ts→ s
tsvaInsvansIC	2	ts→ s ts→ s
tsvaIntsv@ntIC	1	a→ @ ts→ t
tsvaI@nsvantsIC	2	ts→ s
tsvaI@ntsvansIC	2	ts→ s
svaInsvansIC	1	ts→ s ts→ s ts→ s
tsvaIntsvantsIC	1	
tsvaIUnttsvantsICst@n		
tsvaIUnttsvantsICst@n	2	
tsvaIUntsvansIkStn	1	ts→ s
tsvaIUntsfantsIgsdn	1	
svaIntsvansICsdn	1	ts→ s ts→ s
tsvaI@nsvantsIkstn	1	
tsvaIUntsvantsICstn	1	
tsvaI@ntsvansICsn	1	ts→ s
tsvaInsvantsICstn	1	ts→ s
tsvaI@ntsvantsICstn	2	
tsvaIntsvantsICstn	1	
tsvaI@ntsvantsIstn	1	
tsy:g@		
tsy:g@	11	
syg@	1	ts→ s y:→ y
sy:g@	1	ts→ s
Qu:6		
u:r	10	u:6→ u:r
u:6	87	
Qu:6	33	
vElC@m		
vElC@m	4	
vEICm	2	
vECm	3	
vElCm	4	
vEn		
vEn	25	
vE	1	

Norm und Aussprachevarianten	Häufigkeit	Regel
vIl		
vIl	51	
fIl	1	
van		
van	101	
vaN	3	
ve:nIC		
ve:nIC	9	
venIC	2	
ve:nIk	2	
vi:		
vi:	13	
vi:fi:l		
vi:fi:l	12	
vi:fi:	1	
vo:baI		
vo:baI	13	
Qy:b6		
y:b6	45	
Qy:b6	16	
y:v6	4	
Qy:b6haUpt		
y:b6haUpt	9	
Qy:b6haUpt	2	
QybahaUpd	1	6→ a
Yb6haUpt	1	y:→ Y
Qy:b6nE:Cst@n		
Qy:b6nE:Cst@n	1	
Qy:b6nE:kStn	1	
y:b6nE:ksdn	1	
Qyb6nE:stn	1	y:→ y
y:b6ne:Cstn	2	
y:b6nE:Cstn	2	

Norm und Aussprachevarianten	Häufigkeit	Regel
Qy:b6nE:Csn	1	
Qy:b6nE:kstn	1	
Qy:b6ne:Cstn	1	
y:b6ne:Cst	1	
zIC6		
zIC6	12	
sIC6	1	z→ s
zOlt@		
zOlt@	12	
zOnst		
zOnst	6	
zOnsd	2	
sOnst	2	z→ s
zOns	2	
zOntst	1	s→ st
zOnta:k		
zOnta:k	8	
sOnta:k	1	z→ s
zOnta:g	1	
zOntak	2	
zOnta:x	1	k→ x
za:g@n		
za:g@n	2	
sa:gN	1	z→ s
za:gN	8	
za:N	1	g@n→ N
za:gn	1	
zaIn		
zaIn	28	
saIn	11	z→ s
zamsta:k		
zamsta:k	39	
samSta:k	4	z→ s
zamst6k	3	a:→ 6

Norm und Aussprachevarianten	Häufigkeit	Regel
zamstag	2	
zamstak	14	
zamsta:C	1	k$\rightarrow$ C
zamstax	1	k$\rightarrow$ x
samStak	1	z$\rightarrow$ s
ze:6		
ze:6	7	
se:6	5	z$\rightarrow$ s
zi:		
zi:	24	
si:	2	z$\rightarrow$ s
zi:b@n		
zi:b@n	3	
si:bm	2	z$\rightarrow$ s
zibn	1	
zi:bm	6	
zi:bn	1	
zo:		
zo:	20	
so:	6	z$\rightarrow$ s

D Wortverschmelzungen im VM-Dialogmaterial

Transliteration	Transkription	Häufigkeit	Orthographie	Feinkategorie
äddich	E dIC	1	hätte_ich	1
abam	a: bam	1	aber_am	2
abaswü	a bas vy:	1	aber_es_würde	3
abißl	a bIsl	1	ein_bißchen	62
aisis	a IsIs	1	aber_ist_es	4
asos	a zo:s	2	aber_so_ist	5
aufm	aUfm	5	auf_dem	6
aufs	aUfs	2	auf_das	6
auma	aU ma	1	auch_einmal	7
aunich	aU nIC	2	auch_nicht	63
aunoch	aU nOx	1	auch_noch	8
ausm	aUsm	3	aus_dem	6
awie	a vi:	2	aber_wie	9
biich	bi: IC	1	bin_ich	1
bim	bIm	1	bin_mit_dem	10
bleima	blaI ma	2	bleiben_wir	1
brauchma	braUx ma	1	brauchen_wir	1
chönnt	C9nt	1	ich_könnte	11
cheute	COY t@	1	ich_heute	12
dürfma	dY6f ma	1	dürfen_wir	1
daschtimmt	da StImt	1	das_stimmt	13
davleich	da flaIC	1	da_vielleicht	14
desez	dE z@s	1	das_ist_es	15
desis	dE zIs	1	das_ist	16
dessis	dE zIs	1	das_ist	16
dortreffn	dO6 trEfn	1	dort_treffen	17
dreffmauns	drEf ma Uns	1	treffen_wir_uns	18
drumrum	drUm rUm	1	darum_herum	19
erinnersch	6 I n6 S	1	erinnere_ich	20
esein	Es aIn	1	ist_ein	21
fänds	fEnts	3	fände_es	20
fürn	fy:6n	4	für_den	6
			für_ein	6
			für_einen	6
fürs	fy:6s	3	für_das	6
fangma	faN ma	2	fangen_wir	20
fliengma	fli:N ma	1	fliegen_wir	20
gansu	gan zu:	1	kannst_du	1
gehma	ge: ma:	2	gehen_wir	20
großnganzn	gro:s n gan tsn	1	großen_und_ganzen	22

gudaß	gu: das	1	gut_daß	23
gudann	gu: dan	1	gut_dann	23
hälsdn	hEls dn	1	hälst_du_denn	24
hälste	hEls t@	4	hälst_du	1
hättma	hEt ma	1	hätten_wir	1
hätts	hEts	1	hätte_es	1
hättste	hEts t@	1	hättest_du	1
habenen	ha: b@ n@n	1	haben_einen	21
haich	ha: IC	1	habe_ich	1
haldma	hald ma	3	halten_wir	1
			halte_einmal	25
haltma	halt ma	1	halten_wir	1
			halte_einmal	25
hamaba	ha ma: ba	1	haben_aber	27
hamir	ha mi:6	1	haben_wir	1
hamma	ha ma	14	haben_wir	1
hamno	ham no	1	haben_noch	27
hamsich	ham zIC	1	haben_sich	28
hamwa	ham va	5	haben_wir	1
hasch	haS	1	habe_ich	1
hatsen	hat s@n	1	hat_sie_ein	29
			hat_sie_einen	29
hobs	hops	1	habe_es	1
hoffmamol	hOf ma mo:l	1	hoffen_wir_einmal	30
imeim	I maIm	1	in_meinem	31
imeine	i: maI n@	5	ich_meine	32
inem	in Em	2	in_einem	6
iner	i n6	1	in_einer	6
innennäxn	I n@ nE: ksn	2	in_den_nächsten	33
irgenwama	I6 g@n va ma	1	irgend_wann_einmal	34
isa	Is a	1	ist_ein	21
isaau	Is a aU	1	ist_er_auch	24
isan	Is an	1	ist_ein	21
ischlech	i SlEC	1	ist_schlecht	35
ischlecht	i SlECt	1	ist_schlecht	35
ischn	I Sn	1	ist_ein	21
ischo	I So:	1	ist_schon	36
ischtn	IS tn	1	ist_ein	21
isdes	Is dEs	2	ist_das	21
isma	Is ma	1	ist_mir	37
isn	I sn	9	ist_ein	21
iss	Is	1	ist_es	21
isser	I s6	1	ist_er	1
isses	I s@s	23	ist_es	1

issn	I sn	3	ist_ein	21
käms	kE:ms	1	käme_es	1
köma	k9 ma	3	können_wir	1
kömir	k9 mi:6	1	können_wir	1
kömma	k9 ma	27	können_wir	1
kömmas	k9 mas	1	können_wir_es	38
kömmes	k9 m@s	1	können_wir_es	38
kömmir	k9 mi:6	1	können_wir	1
kömwa	k9mva	2	können_wir	1
kömwir	k9m vi:6	1	können_wir	1
könndes	k9n dEs	1	könnte_es	1
könndmer	k9nd m6	1	könnten_wir	1
könnmer	k9n m6	1	können_wir	1
könns	k9ns	2	können_sie	1
könnSe	k9n se	2	können_sie	1
könntma	k9nt ma	8	könnten_wir	1
könntmas	k9nt mas	1	könnten_wir_es	38
könntmer	k9nt m6	1	könnten_wir	1
könntnwa	k9n tn va	1	könnten_wir	1
könnts	k9nts	2	könnte_es	1
könnwa	k9n va	4	können_wir	1
könnwas	k9n vas	1	können_wir_es	38
könnwer	k9n v6	2	können_wir	1
kamma	ka ma	2	kann_man	39
kammaama	ka ma a ma	1	kann_man_einmal	40
kannste	kans t@	1	kannst_du	1
kannsues	kan su: Es	1	kannst_du_es	38
komma	kO ma	2	kommen_wir	1
kommse	kOm s@	1	kommen_sie	1
kommwa	kOm va	1	kommen_wir	1
laßma	las ma	1	lassen_wir	1
lassmama	las ma ma	1	lassen_wir_einmal	41
müßma	mYs ma	5	müssen_wir	1
müßtma	mYst ma	1	müßten_wir	1
müßtmades	mYst ma des	1	müßten_wir_das	42
müßtmas	mYst mas	1	müßten_wir_das	42
			müßten_wir_es	38
müßtmer	mYst m6	1	müßten_wir	1
müßtns	mYs tns	1	müßten_es	1
			müßten_das	43
müßts	mYsts	1	müßte_es	1
müssmas	mYs mas	1	müssen_wir_das	42
müssmer	mYs m6	2	müssen_wir	1
müstma	mYst ma	2	müßten_wir	1

machemer	ma x@ m6	1	machen_wir	1
machmades	max ma dEs	1	machen_wir_das	42
machmas	max mas	7	machen_wir_das	42
			machen_wir_es	38
machmasenn	max mas En	1	machen_wir_es_denn	44
machmir	max mi:6	1	machen_wir	1
mahma	ma: ma	1	machen_wir	1
mahmama	ma: ma ma	1	machen_wir_einmal	41
mahn	ma:n	4	mal_einen	45
mama	ma ma	3	machen_wir	1
mamades	ma ma dEs	1	machen_wir_das	42
mamas	ma mas	1	machen_wir_das	42
			machen_wir_es	38
mamaso	ma ma zo:	2	machen_wir_so	46
mibm	mIbm	2	mit_dem	6
midem	mI de:m	1	mit_dem	6
mider	m Id6	1	mit_der	6
mim	mIm	15	mit_dem	6
			mit_einem	6
mimeiner	mI maI n6	1	mit_meiner	31
mimissn	mI mI sn	1	wir_müssen	11
mirs	mi:6s	2	mir_ist	47
mitm	mItm	4	mit_dem	6
			mit_einem	6
mitn	mItn	1	mit_den	6
nachm	na:xm	1	nach_dem	6
nehma	ne: ma	15	nehmen_wir	20
nehmwa	ne:m va	8	nehmen_wir	20
nehmwer	ne:m v6	2	nehmen_wir	20
sähs	ze:s	6	sähe_es	20
sags	za:ks	6	sage_es	20
sama	za: ma	3	sagen_wir	20
samama	za: ma ma	2	sagen_wir_einmal	48
sangma	za:N ma	6	sagen_wir	20
sangmer	za:N m6	1	sagen_wir	20
sangwama	za:N va ma	1	sagen_wir_einmal	48
sangweama	za:N ve a ma	1	sagen_wir_einmal	48
schauma	SaU ma	2	schauen_wir	20
schaunwa	SaUn va	1	schauen_wir	20
schlags	Sla:ks	1	schlage_es	20
sehmwa	ze:m va	1	sehen_wir	20
sGott	sgOt	1	Grüß_Gott	49
siehste	zi:s t@	1	siehst_du	20
siehts	zi:ts	1	siehst_es	20

siehtsn	zi: tsn	2	siehst_es_denn	50
simma	zI ma	1	sind_wir	1
sollma	zOl ma	1	sollen_wir	1
sollmas	zOl mas	2	sollen_wir_das	42
			sollen_wir_es	38
sollnmia	zOln mi:6	1	sollen_wir	1
sollns	zOlns	1	sollen_es	1
solltma	zOlt ma	6	sollten_wir	1
trags	tra:ks	3	trage_es	20
treffma	trEf ma	8	treffen_wir	20
treffmer	trEf m6	1	treffen_wir	20
treffnma	trEfn ma	1	treffen_wir	20
treffwer	trEf v6	1	treffen_wir	20
verschiebmades	f6 Si:p ma dEs	1	verschieben_wir_das	51
versuchmas	f6 zu:x mas	1	versuchen_wir_es	52
			versuchen_wir_das	20
vomir	fO mi:6	1	von_mir	31
vorm	fo:6m	4	vor_dem	6
währenem	vE: r@ n@m	1	während_einem	53
wärne	vE:6 n@	2	wäre_eine	21
wärs	vE:6s	546	wäre_es	1
wüich	vy: IC	1	würde_ich	1
würdi	vY6 di:	1	würde_ich	1
würds	vY6ts	47	würde_es	1
würdsn	vy6tsn	1	würde_es_ein	29
			würde_es_einen	29
			würde_es_den	29
würs	vy:6s	1	würde_es	1
wegem	ve: g@m	1	wegen_dem	6
weii	vaI i:	1	weil_ich	54
weiich	vaI IC	2	weil_ich	54
wema	vE ma	3	wenn_wir	54
wemas	vE mas	1	wenn_wir_es	55
			wenn_wir_das	56
wemma	vE ma	9	wenn_wir	54
wemmas	vE mas	1	wenn_wir_es	55
			wenn_wir_das	56
wemmer	vE m6	1	wenn_wir	54
wemwer	vEm v6	1	wenn_wir	54
wenicht	vE nICt	1	wenn_nicht	64
wennir	vE ni:6	1	wenn_ihr	54
wenns	vEns	1	wenn_es	54
			wenn_das	57
werdma	vE6t ma	1	werden_wir	1

werds	vE6ts	1	werde_es	1
werma	vE6 ma	3	werden_wir	1
			werde_einmal	25
werns	vE6ns	1	werden_es	1
			werden_das	43
wers	vE6s	1	wer_es	58
			wer_das	59
wis	vi:s	1	wie_es	58
			wie_das	59
wises	vi: z@s	1	wie_ist_das	60
			wie_ist_es	61
wollma	vOl ma	2	wollen_wir	1
wommanal	vO ma na:l	1	wollen_wir_einmal	41

E Feinkategorien für Wortverschmelzungstypen

Kategrie	enthalteneWortklassen
1	Hilfsverb_Personalpronomen
2	Konjunktion_Präposition
3	Konjunktion_Personalpronomen_Hilfsverb
4	Konjunktion_Hilfsverb_Personalpronomen
5	Konjunktion_Modaladverb_Hilfsverb
6	Präposition_Artikel
7	Adverb_Temporaladverb
8	Adverb_Konjunktion
9	Konjunktion_Interrogativadverb
10	Hilfsverb_Präposition_Artikel
11	Personalpronomen_Hilfsverb
12	Personalpronomen_Temporaladverb
13	Demonstrativpronomen_Verb
14	Konjunktion_Modaladverb
15	Demonstrativpronomen_Hilfsverb_Personalpronomen
17	Präposition_Verb
18	Verb_Personalpronomen_Reflexivpronomen
19	Kausaladverb_Präposition
20	Verb_Personalpronomen
21	Hilfsverb_Artikel
22	Adverb_Konjunktion_Adverb
24	Hilfsverb_Personalpronomen_Konjunktion
25	Hilfsverb_Temporaladverb
26	Hilfsverb_Temporaladverb_Konjunktion
27	Hilfsverb_Konjunktion
28	Hilfsverb_Reflexivpronomen
29	Hilfsverb_Personalpronomen_Artikel
30	Hilfsverb_Personalpronomen
31	Präposition_Possesivpronomen
32	Personalpronomen_Verb
33	Präposition_Artikel_Adjektiv
34	Temporaladverb_Temporaladverb
35	Hilfsverb_Adverb
36	Hilfsverb_ Temporaladverb
37	Hilfsverb_Possesivpronomen
38	Hilfsverb_Personalpronomen_Personalpronomen
39	Hilfsverb_Indefinitpronomen
40	Hilfsverb_Indefinitpronomen_Temporaladverb
41	Hilfsverb_Personalpronomen_Temporaladverb
42	Hilfsverb_Personalpronomen_Demonstrativpronomen

43	Hilfsverb_Demonstrativpronomen
44	Hilfsverb_Personalpronomen_Personalpronomen_Konjunktion
45	Temporaladverb_Artikel
46	Hilfsverb_Personalpronomen_Modaladverb
47	Possesivpronomen_Hilfsverb
48	Verb_Personalpronomen_Temporaladverb
49	Grußformel
50	Verb_Personalpronomen_Konjunktion
51	Verb_Personalpronomen_Demonstrativpronomen
52	Verb_Personalpronomen_Personalpronomen
53	Konjunktion_Artikel
54	Konjunktion_Personalpronomen
55	Konjunktion_Personalpronomen_Personalpronomen
56	Konjunktion_Personalpronomen_Demonstrativpronomen
57	Konjunktion_Artikel
58	Interrogativadverb_Personalpronomen
59	Interrogativadverb_Demonstrativpronomen
60	Interrogativadverb_Hilfsverb_Demonstrativpronomen
61	Interrogativadverb_Hilfsverb_Personalpronomen
62	Artikel_Indefinitpronomen
63	Adverb_Adverb
64	Konjunktion_Adverb

F Gruppenbildung über Feinkategorien der Wortverschmelzungen

Verschmelzungstypen (grob)		Verschmelzungstypen (fein)	Anzahl
Hilfsverb_Pronomen_		1,28,30,37,39,43	74
	_Konjunktion	24	2
	_Artikel	29	2
	_Pronomen	38,42	17
	_Adverb	40,41,46	57
	_Pron._Konj.(101)	44	1
Adverb_Artikel(45)		45	1
Verb_Pronomen_		20	25
	_Pronomen	18,51,52	3
	_Adverb	48	3
	_Konjunk.(32)	50	1
Präposition_Artikel_		6	19
	_Adjektiv(20)	33	1
Konjunktion_Pronomen_		54	8
	_Pronomen	55,56	4
	_Hilfsverb(13)	3	1
Hilfsverb_Artikel(11)		21	11
Hilfsverb_Adverb_(5)		25,35,36	5
Konjunktion_Hilfsverb_Pronomen(4)		4	4
Adverb_Pronomen(4)		58,59	4
Pronomen_Hilfsverb_		11,47	3
	_Pronomen(4)	15	1
Konjunktion_Adverb_		9,14,64	3
	_Hilfsverb(4)	5	1
Präposition_Pronomen(3)		31	3
Adverb_Adverb(3)		7,34,63	3

Hilfsverb_Konjunktion(2)	27	2
Konjunktion_Artikel(2)	53,57	2
Adverb_Konjunktion_	8	1
_Adverb(2)	22	1
Adverb_Hilfsverb_Pronomen(2)	60,61	2
Pronomen_Verb(2)	13,32	2
Hilfsverb_Präposition_Artikel(1)	10	1
Konjunktion_Präposition(1)	2	1
Präposition_Verb(1)	17	1
Adverb_Präposition(1)	19	1
Pronomen_Adverb(1)	12	1
Artikel_Pronomen(1)	62	1
Grußformel(1)	49	1