Multilinguale Spracherkennung

Kombination akustischer Modelle zur Portierung auf neue Sprachen

Zur Erlangung des akademischen Grades eines
Doktors der Ingenieurwissenschaften
der Fakultät für Informatik
an der Universität Karlsruhe (Technische Hochschule)
vorgelegte

Dissertation

von
Tanja Schultz
aus Oldenburg i/O

Tag der mündlichen Prüfung: **19. Juli 2000**
Betreuer: **Prof. Dr. Alexander Waibel**
Korreferent: **Prof. Dr. Dirk Van Compernolle**

Kurzfassung

Die Entwicklung von Spracherkennern für neue Sprachen ist bislang mit einem sehr hohen Kosten- und Arbeitsaufwand verbunden. Im Rahmen der vorliegenden Arbeit wurden Methoden entwickelt, die diesen Aufwand drastisch reduzieren, ohne daß es zu nennenswerten Leistungseinbußen kommt. Die Methoden basieren auf der Idee, Daten und Informationen aus unterschiedlichen Sprachen, für die bereits Erkenner existieren, auf neue Sprachen zu übertragen. Dazu wurden auf der Basis mehrerer Sprachen multilinguale, d.h. von der Sprache unabhängige Komponenten entwickelt und anschließend mit Hilfe sehr geringer Datenmengen auf neue Sprachen portiert. Um diese Idee zu realisieren, waren im wesentlichen die folgenden Probleme zu lösen:

- *Sammlung und Extraktion relevanter Sprachinformationen aus vielen Sprachen*
- *Entwicklung multilingualer Spracherkennungskomponenten*
- *Konzeption und Implementierung von Portierungsverfahren*

Zur Lösung des ersten Problems wurde das Projekt GlobalPhone initiiert, in dessen Rahmen eine multilinguale Datenbasis in 9 der 12 wichtigsten Weltsprachen entstand. Sie enthält gelesene Sprache von über 1200 Sprechern und hat weltweit großes Interesse hervorgerufen. Auf Basis dieser Daten wurden **monolinguale Spracherkenner in 10 Sprachen** entwickelt, die eine Wortfehlerrate zwischen 10% und 20% erreichen. Diese Ergebnisse bestätigen die Annahme, daß sich die klassischen Algorithmen der Spracherkennung auf alle modellierten Sprachen übertragen lassen. Allerdings zeigen die Erfahrungen dieser Arbeit, daß ein erheblicher Entwicklungsaufwand zur Behandlung sprachenspezifischer Besonderheiten notwendig ist.

Auf der Basis dieser 10 Spracherkenner wurde zur Lösung des zweiten Problems ein globales Phoneminventar entwickelt, das die lautliche Vielfalt aller beteiligten Sprachen abdeckt. Mittels dieses Inventars wurden die akustischen Modelle der Sprachen mit neuen Methoden zu **multilingualen akustischen Modellen** kombiniert. Die Vorteile dieser multilingualen Modelle liegen in der Parametereinsparung und ihrem Nutzen für die Sprachenidentifizierung sowie in ihrer Robustheit bei der Portierung auf neue, bisher nicht gelernte Sprachen.

Zur Lösung des dritten Problems wurden Adaptionsverfahren entwickelt, die eine erfolgreiche **Portierung eines Spracherkenners auf eine neue Sprache** auch dann ermöglichen, wenn nur begrenztes Datenmaterial in der neuen Sprache zur Verfügung steht. Die Portierung umfaßt die Auswahl einer geeigneten Teilmenge des globalen Phoneminventars, die automatische Anpassung vorhandener Aussprachelexika, die Adaption der akustischen Modelle sowie die Spezialisierung gelernter Kontextentscheidungsbäume. Im Zusammenspiel aller Methoden kann ein multilinguales Spracherkennungssystem sehr effizient ohne nennenswerte Leistungsverluste auf eine neue Sprache portiert werden. Die Portierung ist kostensparend, weil sich der Umfang der notwendigen Datensammlung drastisch verringert, und zeitsparend, weil die Adaption mit kleinen Datenmengen schneller durchgeführt werden kann als die komplette Neuentwicklung eines Systems.

Danksagung

Eine Promotion ist niemals die Leistung von nur einer Person. Für meine Arbeit trifft dieser Satz ganz sicher zu, denn ein derart umfangreiches Projekt wie GlobalPhone kann man unmöglich alleine bewältigen.

Als erstes möchte ich mich bei Prof. Dr. Alex Waibel bedanken, der mir die Möglichkeit gegeben hat, ein Projekt dieser Größenordnung zu initiieren und in diesem Rahmen eine Dissertation anzufertigen. Mit seinen sprudelnden Ideen, konstruktiven Anregungen und kritischen Bemerkungen hat er mich stets zu neuen Leistungen angespornt. Prof. Dr. Dirk Van Compernolle möchte ich für die Übernahme des Korreferats danken.

Von allen anderen WegbegleiterInnen und HelferInnen möchte ich als erstes den zahlreichen wissenschaftlichen Hilfskräften meinen Dank aussprechen, die mit großem persönlichen Engagement tolle Arbeit geleistet haben. Sie haben das GlobalPhone-Projekt zum Leben erweckt und aus unserem ohnehin internationalen Institut eine multikulturelle Gemeinschaft gemacht. Die Zusammenarbeit mit ihnen und die gemeinsamen Unternehmungen gaben mir einen Einblick in die verschiedensten Kulturen und sind eine echte Bereicherung für mein Leben. Im einzelnen gilt mein herzlicher Dank Omar Abdallah, Jamal Abu Alwan, Hiroko Akatsu, Giovanni Najera Barquero, Kenan Çarkı, Keal-Chun Cho, Caleb Everett, Raul Ivo Faller, Renato Ferreira, Sanela Habibija, Wajdi Halabi, Evelyn Kimmich, Kyung-Kyu Lee, Natalia und Orest Mikhailiouk, Jae-Ho Park, Martin Sjögren, Sang-Hun Shin, Maho Takeda, Sayoko Takeda, Jing Wang, Tianshi Wei, Jiaxing Weng, Nadia Zouabi, Mutlu Yalçın sowie Olfa Karboul-Zouari und ihrem Mann Mohammed Zouari.

Für das überaus angenehme Arbeitsklima an unserem Institut und die gute Zusammenarbeit möchte ich mich bei allen jetzigen und früheren KollegInnen in Karlsruhe und an der CMU ganz herzlich bedanken. Nennen möchte ich hier: Sondra Ahlén, Markus Baur, Kay Berkling, Finn Dag Buø, Susanne Burger, Matthias Denecke, Michael Finke, Christian Fügen, Jürgen Fritsch, Donna Gates, Marsal Gavalda, Petra Geutner, Hermann Hild, Stefan Jäger, Susanne Kaufmann, Thomas Kemp, Detlef Koll, Alon Lavie, Lori Levin, Stefan Manke, Laura J. Mayfield-Tomokiyo, John Mc Donough, Arthur Mc Nair, Uwe Meier, Florian Metze, Jürgen Reichert, Klaus Ries, Ivica Rogina, Thomas Schaaf, Tilo Sloboda, Hagen Soltau, Rainer Stiefelhagen, Bernhard Suhm, Takashi Tomokiyo, Minh Tue Vo, Ye-Yi Wang, Martin Westphal, Monika Woszczyna, Klaus Zechner und Torsten Zeppenfeld. Meinen Zimmerkollegen und -nachbarn Finn Dag Buø, Detlef Koll und Martin Westphal gebührt mein besonderer Dank, weil sie mich stets mit Rat und Tat fachlich und persönlich unterstützten und klaglos die vielen Unterbrechungen durch all die „Hiwis“ ertrugen.

Ich danke allen, die immer ein offenes Ohr für meine vielen sprachenspezifischen Fragen hatten. In diesem Zusammenhang möchte ich bei Prof. Akira Kurematsu, Michihiro Kitamura, Tomo Nakasuji und Laura J. Mayfield Tomokiyo bedanken, die mich in Fragen über die japanische Sprache berieten, ganz besonders bei Detlef Koll, der mich aktiv unterstützte. Bei Ivica Rogina bedanke ich mich für viele

Hilfestellungen in der kroatischen und englischen Sprache. Mohammed Akbar und Harouna Kabré waren eine große Hilfe für Fragen über die französische Sprache. Gang-Seong Lee danke ich für die zahlreichen Diskussionen über die koreanische Sprache sowie Oh-Wook Kwon und Prof. Young vom ETRI, die uns eine morphologischen Zerlegung des koreanischen GlobalPhone-Korpus zur Verfügung stellten. Bei Petra Geutner bedanke ich mich für die Zusammenarbeit bei der Anwendung ihres HDLA-Ansatzes auf einige Sprachen. Thomas Kemp möchte ich für die Bereitstellung und Hilfe mit den deutschen Daten danken. Hervorheben möchte ich außerdem meine KollegInnen Michael Finke, Hermann Hild, Thomas Kemp, Detlef Koll, Ivica Rogina, Hagen Soltau, Martin Westphal und Monika Woszczyna, die mir durch viele Diskussionen wertvolle Anregungen gaben und Unterstützung zukommen ließen.

Außerdem möchte ich mich bei Kenan Çarkı, Olfa Karboul-Zouari, Daniel Kiecza, Stefan Raschke, Jürgen Reichert, Hagen Soltau und Roald Wolff bedanken, die mit ihren Studien- und Diplomarbeiten wichtige Erkenntnisse und Ergebnisse beigetragen haben.

Die vielen am Projekt beteiligten Personen und der große Umfang der gesammelten Daten haben sehr hohe Anforderungen an all diejenigen MitarbeiterInnen gestellt, die das organisatorische Rückgrat unseres Lehrstuhles bilden. Daher möchte den immer hilfsbereiten Sekretärinnen Ingrid Gemen, Sonja Seitz und Ilona Deger danken. Ein besonderer Dank geht an Silke Dannenmaier, die mich in allen administrativen Problemen stets beraten und tatkräftig unterstützt hat. Für die Betreuung des gesamten Rechnerbetriebes möchte ich mich bei Markus Baur, Frank Dreilich, und Martin Klein bedanken. Sie haben stets einen Platz für meine riesigen Datenmengen gefunden und deren Sicherung gewährleistet.

Einen wertvollen Beitrag bei der Fertigstellung dieser Ausarbeitung leisteten die konstruktiven Anregungen und Kommentaren der zahlreichen KorrekturleserInnen. Mein herzlicher Dank gilt Ute Knapp, Ivica Rogina, Jutta Klein, Hermann Hild, Hans-Peter Eich, Monika Woszczyna, Christoph Becker, Susanne Burger und Daniela Oppermann. Besonders möchte ich mich bei Elmar Nöth bedanken, der mir mit vielen Tips half, die ersten Hürden zu überwinden.

Schließlich möchte ich meiner ganzen Familie und meinen Freunden danken, die mich in meinem Leben begleiten und mir Halt geben. Ein ganz besonderer Dank richtet sich an meine Eltern, die mich in all meinem Tun unterstützen und es mir ermöglichten, mit der Informatik ein zweites Studium durchzuführen.

Heidelberg, 1. Juni 2000 Tanja Schultz

Inhaltsverzeichnis

Kapitel 1

Einführung

Dieser Abschnitt führt in die Anwendungsbereiche mehrsprachiger Systeme ein und definiert den Begriff „Multilinguale Spracherkennung“, wie er in der vorliegenden Arbeit verwendet wird. Es werden die Gründe für die Forderung nach Flexibilität heutiger Erkennungssysteme herausgearbeitet und die Notwendigkeit zur schnellen Portierung auf neue Sprachen motiviert.

Das Ende des 20. Jahrhunderts ist durch eine radikale Veränderung unserer Informations- und Kommunikationsstruktur geprägt. Ein wesentliches Merkmal dieser Veränderung wird durch den Computer bestimmt, der in nahezu alle Bereiche unseres Lebens Einzug gehalten hat. Computergestützte Anwendungen begleiten unseren Geschäfts- und Privatalltag und sind aus Verwaltungs- und Bildungseinrichtungen, Dienstleistungsgewerbe und unserem alltäglichen Leben nicht mehr wegzudenken. Die Sprachtechnologie spielt bei dieser Entwicklung eine sehr große Rolle, denn sie erhöht die Effizienz und Benutzerfreundlichkeit von Mensch-Maschine-Schnittstellen. Daher zeichnet sich, nach einer eher schleppenden Marktentwicklung in den vergangenen 10 Jahren, in jüngerer Zeit ein regelrechter Boom in der Sprachtechnologiebranche ab.

Das zweite deutliche Merkmal einer Veränderung ist die zunehmende Globalisierung, durch die immer mehr Menschen unterschiedlichster Sprachen und Kulturen auf die eine oder andere Weise miteinander interagieren. Zwar existiert mit der am weitesten verbreiteten und meist gesprochenen Sprache, Englisch, zumindest im Computerbereich eine „Lingua franca“, dennoch wird es in absehbarer Zukunft vermehrtes Interesse an vielen verschiedenen Sprachen geben. Nach Prognosen von [CIA98] reduziert sich beispielsweise die klare Dominanz der amerikanischen Internetnutzer von 80% im Jahre 1991 und 55% im Jahr 1998 auf nur noch 40% im Jahr 2000. Die Nachfrage nach Informationen, die nicht in Englisch angeboten werden, wächst also selbst im Internet ständig an.

Die zunehmende Globalisierung hat also keineswegs zu einer Vereinheitlichung von Sprachen oder zu einer Reduzierung der Sprachenvielfalt geführt. Manche Forscher behaupten gar, daß die Zahl der Sprachen auf der Welt zunimmt [Sku88, Mai94] und begründen dies mit dem Ausbau bisher zurückgesetzter Sprachen, dem Vordringen von Fachsprachen und der Differenzierung der Sprachen (etwa Englisch in „British English“ und „American English“) oder mit dem Vorgang ethnischer, sozialer und religiöser Selbstbehauptung *durch Sprache*. Zur Zeit sieht es so aus, als sei die Vielfalt der Sprachen das sichtbarste Zeichen der Verschiedenheit einer sich unaufhaltsam vereinheitlichenden Weltzivilisation.

Die beschriebenen Veränderungen der Kommunikations- und Informationsstruktur, die Globalisierung und die Computerisierung führen zusammengenommen zu einer großen Nachfrage nach *multilingualen* Computerapplikationen, die aller Voraussicht nach in den nächsten Jahren weiter anwachsen wird.

1.1 Anwendungen mehrsprachiger Systeme

Der Begriff „Multilingualität“ ist im Kontext sprachverarbeitender Systeme bisher meist auf die Mehrsprachigkeit von *geschriebenen Texten* beschränkt. Man denkt dabei an textbasierte Übersetzungssysteme oder an intelligente Suchmaschinen, die gewünschte Informationen aus mehrsprachigen Texten ausfiltern. Weitere Verwendungszwecke, die auf der geschriebenen Darstellung basieren, sind mehrsprachige Dokumente wie Lexika und Wörterbücher, die das Arbeiten mit einer Fremdsprache erleichtern.

Aufgrund der Forderung nach effizienten Mensch-Maschine-Schnittstellen und der zunehmenden Mensch-zu-Mensch-Kommunikation über Sprachgrenzen hinweg ist Multilingualität aber gerade in der gesprochenen Form von zunehmender Bedeutung. Dank der Entwicklungen der vergangenen 10 Jahre gibt es bereits zahlreiche sprachgesteuerte Applikationen, die in mehreren Sprachen entwickelt worden sind. Beispiele dafür sind Informationsabfrage- und Datenbanksysteme, wie etwa Zugauskunftssysteme (EVAR [NBM+85], MAIS [MAI98], RAILTEL [BL97], ARISE [ARI98]) und Videoindizierungssysteme (INFORMEDIA [WHW96], VIEW4YOU [Kem99], OLIVE [OLI98], Pop-Eye [PE98]). Ein wichtiger Bereich sind Dialogsysteme, wie das im Rahmen des BMBF geförderte Projekt VERBMOBIL [VER00], das in einem spontansprachlich geführten Mensch-zu-Mensch-Dialog Übersetzungen in den drei Sprachen Deutsch, Englisch und Japanisch anbietet. Im Autobereich entwickeln sich sprachgesteuerte Navigationshilfen zu einem Verkaufsschlager (VODIS [VOD98], SpeechDat-Car [SC98]). Anwendungen zum Erlernen von Fremdsprachen sind ebenfalls zu erwähnen (STiLL [STi99], RECALL [REC98], SPEAK [SPE98a], ILAM [ILA98] , ISLE [ISL98]). Insbesondere in Ländern mit hohen Einwanderungsquoten sind Applikationen gefragt, die ausländische Anrufer nach einer automatischen Sprachenidentifizierung an entsprechende Sprachenexperten weiterleiten. Einsatzgebiete sind öffentliche Einrichtungen wie etwa Notrufzentralen [MBC94], Polizeibehörden (LinguaNet [Lin98]), Verkehrszentralen und Call-Center (IDAS [IDA98], ACCeSS [ACC98]) aller Art. Als Kommunikationshilfe für den Reisenden im Ausland eignen sich tragbare Phrasen- und Wörterbücher, die in ihrer noch futuristischen Variante bis hin zum persönlichen Dolmetscher und Navigator im Taschenformat reichen (TalkingMap [Wes00], MIETTA [MIE98]). Weitere mögliche Einsatzgebiete sind sprachgestützte interaktive Arbeitsplätze für Personen mit Sehschwächen. In der Entwicklung, aber noch etwas von ihrer Marktreife entfernt, sind dagegen Applikationen zur automatischen Synchronisation von Spielfilmen und Fernsehbeiträgen sowie zur vielsprachigen automatischen Verschriftung von Sprachdaten.

Zusammenfassend erstreckt sich der praktische Nutzen mehrsprachiger Applikationen auf die Bereiche Handel (insbesondere internationaler Handel und Geschäftsabwicklungen via Telefon oder Internet), öffentliche Einrichtungen und Gesundheitswesen, Bildung und Erziehung sowie Informationsdienste aller Art. Die dabei zu bewältigenden Probleme lassen sich in die vier Bereiche *Sprachenidentifizierung*, *Spracherkennung*, *Sprachverstehen* und *Sprachübersetzung* einteilen. Die vorliegende Arbeit befaßt sich mit der Spracherkennung und wird die gefundenen Lösungen auch auf die Sprachenidentifizierung anwenden. Das Ziel der hier beschriebenen Forschung ist eine möglichst korrekte textuelle Darstellung der gesprochenen Äußerung. Themen, bei denen die Erfassung der Intention einer Äußerung und daraus abzuleitende Aktionen im Vordergrund stehen, wie es beim Sprachverstehen oder bei der Sprachübersetzung der Fall ist, werden in dieser Arbeit nicht behandelt.

1.2 Multilinguale Spracherkennung

Der Begriff „multilinguales Spracherkennungssystem“ wird innerhalb der Spracherkennungsgemeinde nicht einheitlich verwendet. Multilingualität ist in dieser Gemeinde zu einem wichtigen Schlüsselbegriff geworden. Daher wird häufig das Nebeneinander mehrerer Spracherkenner, die jeweils auf eine einzige Sprache spezialisiert sind, mit diesem Begriff belegt bzw. angepriesen und beworben. Im Sprachgebrauch der vorliegenden Arbeit wird das Nebeneinander mehrerer Sprachen als *mehrsprachige Spracherkennung* oder als *monolinguale Spracherkennung in vielen Sprachen* bezeichnet. Der Begriff *multilinguale Spracherkennung* geht weit über dieses Nebeneinander von Spracherkennern hinaus. Mit multilingualer Spracherkennung wird in dieser Arbeit ein System bezeichnet, das *gleichzeitig* mehrere verschiedene Sprachen verarbeiten kann. Das Konzept des Nebeneinanders ist darin durch das Konzept der Sprachenuniversalität ersetzt. Sprachenuniversalität bedeutet für die Spracherkennung, daß notwendige Wissensquellen und Komponenten von der darunterliegenden Sprache entkoppelt werden und in einen gemeinsamen *universellen* Erkenner integriert werden. Um dieses Ziel zu erreichen, werden zwei Methoden angewandt:

- Gemeinsame Nutzung von Software: Die Trainings- und Evaluationsmodule des Spracherkenners sind abgekoppelt von sprachenspezifischen Informationsquellen. Die sprachenspezifischen Quellen werden statisch dazugebunden oder zur Laufzeit nachgeladen.

- Gemeinsame Nutzung von Daten: Für die Spracherkennung notwendige Wissensquellen wie akustische Modelle, Aussprachewörterbücher und Sprachmodelle werden möglichst *universell*, d.h. unabhängig von der zugrundeliegenden Sprache aufbereitet und durch das gemeinsame Nutzen verschiedensprachiger Daten trainiert.

In heute gängigen kommerziellen Spracherkennungssystemen, beispielsweise von Dragon, IBM, Philips, Lernout+Hauspie oder Lucent wird von der ersten Methode des „Softwareteilens“ Gebrauch gemacht. Die Erkennersysteme werden in vielen Sprachen betrieben und enthalten alle mehr oder minder dasselbe Software-Herzstück. Auch im an dem Institut für Logik, Komplexität und Deduktionssysteme (ILKD) implementierten Spracherkennungssystem *Janus Speech Recognition Toolkit (JRTk)* [FGH+97] wird so vorgegangen.

Die Methode der gemeinsamen Nutzung der Software wird erfolgreich eingesetzt. Die gemeinsame Nutzung von Daten muß allerdings noch auf ihre Durchführbarkeit und Effizienz hin geprüft werden. Dies ist ein wesentlicher Gegenstand dieser Arbeit, die sich mit der Kombination akustischer Modelle anhand eines *universellen* Phoneminventars beschäftigt. Dazu werden die Phoneminventare vieler Sprachen analysiert und deren Gemeinsamkeiten und Unterschiede herausgearbeitet. Auf der Basis des resultierenden universellen Phoneminventars entsteht dann ein multilinguales Spracherkennungssystems, das mehrere Sprachen identifizieren und erkennen kann. Die Leistung des multilingualen Systems wird mit den Leistungen der monolingualen Erkenner verglichen.

Der Nutzen eines multilingualen Systems liegt in der robusteren Schätzung zu lernender Parameter, die durch sprachenübergreifende Verwendung von Daten entstehen, und in kompakteren Systemen durch die gemeinsame Nutzung von Software. Dies spielt vor allem für speicherlimitierte Anwendungen, wie etwa im Auto oder für Client-Server Lösungen mit einfachen Clients, eine wichtige Rolle. Des weiteren können multilinguale Systeme dazu eingesetzt werden, mit a-priori unbekannter Eingabesprache umzugehen. Außerdem sind multilinguale Systeme eine mögliche Alternative zur Erkennung von Sprachen von Nichtmuttersprachlern. Nicht zuletzt ist ein multilinguales System für alle Anwendungen unerläßlich, bei denen es innerhalb einer Äußerung zu Sprachenwechseln kommen kann, die vom Spracherkennungssystem toleriert und möglichst korrekt erkannt werden sollen.

Ein wesentlicher Aspekt der multilingualen akustischen Modellierung ist die Aussicht auf eine gute Anpassungsfähigkeit auf neue Sprachen, die aus dem sprachenübergreifenden Training der akustischen Modelle resultiert.

1.3 Die Forderung nach Flexibilität

Die Forderung nach Flexibilität heutiger Erkennungssysteme resultiert aus dem Problem abzuschätzen, für wieviele und vor allem für welche Sprachen in Zukunft Systeme benötigt werden. Diktiererkenner für große Wortschätze werden von den führenden Anbietern mittlerweile für viele verschiedene Sprachen angeboten. Dragon bietet 7 Sprachen[1], Lernout+Hauspie bietet 8 Sprachen[2], IBM bietet 11 Sprachen[3]. Aber wonach richtet sich das Interesse an den genannten Sprachen? Kann man die Interessenentwicklung absehen und Trendsprachen voraussagen?

Aufgrund wirtschaftlicher Zwänge ist das Interesse an Spracherkennungssystemen immer ein Produkt aus Verbreitung der Sprache und deren wirtschaftlicher oder politischer Bedeutung. Es gibt etwa 25 Sprachen auf der Welt, von denen jede von mehr als 50 Millionen Menschen gesprochen wird. Tatsächlich wurden die ersten Spracherkenner in den drei verbreitesten Sprachen entwickelt. Aber auch weniger verbreitete Sprachen können etwa aus politischen Gründen schnell in das Zentrum des Interesses treten. So gab beispielsweise die amerikanische Forschungsbehörde DARPA[4] im Laufe des Balkankonflikts die Entwicklung eines portablen Übersetzungsgerätes für die Sprache Serbisch/Kroatisch in Auftrag. Das Gerät übersetzt medizinische Ausdrücke ins Englische und unterstützt die medizinische Hilfe der amerikanischen UN-Einheiten im Bosnienkrieg. Aufgrund des Kosovo-Konflikts wurde dieser Prototyp kurzerhand ins Albanische portiert [HBTA99]. Dieses Beispiel zeigt, daß sich schnelle Portierbarkeit in der Praxis als wichtiges und daher als zu optimierendes Merkmal herausgestellt hat.

Während die Portierung von sehr kleinen Prototypen mit relativ geringem Aufwand möglich ist, sind die Entwicklungszeiten für Erkenner mit großen Wortschätzen noch sehr lang und die Kosten sehr hoch. In einer Analyse schätzen Bell Laboratories für die Entwicklung eines Diktiererkenners etwa 9-15 Personenmonate Entwicklungszeit und mehrere 100.000 Euro Entwicklungskosten pro Sprache [GG97]. Neben der reinen Entwicklung des Erkenners entsteht darüber hinaus ein enormer Aufwand zur Sammlung der zum Training notwendigen Sprachaufnahmen und deren Verschriftungen. Um einen Spracherkenner für einen mittelgroßen Wortschatz zu trainieren, muß man von mindestens 10 Stunden auf Wortebene verschriftetem Sprachmaterial ausgehen [LADGA96].

Die Sammlung und Verschriftung einer so großen Menge von Sprachdaten ist aber nicht nur teuer, sondern häufig auch nicht sinnvoll. Aus oben geschilderten Gründen kann sich das Interesse sehr schnell verlagern, so daß die Sammlung umfangreicher Daten den immensen Zeitaufwand nicht lohnt. Die Daten aller Sprachen der Welt zur Sicherheit auf Halde zu produzieren, ist bei ca. 4500 Sprachen (siehe Kapitel 2) ebensowenig sinnvoll, insbesondere deshalb, weil in vielen Fällen der Einsatzbereich eines Systems im Vorfeld nicht klar umrissen ist. Nach heutigem Wissensstand

[1] www.naturalspeech.com: Britisch-Englisch, Französisch, Deutsch, Italienisch, Spanisch, Niederländisch, Schwedisch

[2] www.lhsl.com: amerikanisches und britisches Englisch, Deutsch, Niederländisch, Französisch, Spanisch, Mandarin und Shanghai-Chinesisch

[3] [Kun00] Arabisch, Chinesisch, Deutsch, Englisch (britisches und amerikanisches), Französisch, Hindi, Italienisch, Japanisch, Koreanisch und Spanisch

[4] Defense Advanced Research Projects Agency

müssen brauchbare Systeme auf eine Domäne eingeschränkt werden. Darüber hinaus sind Anwendungen denkbar, bei denen eine repräsentative Datensammlung schwer möglich ist, beispielsweise bei der Erkennung nichtmuttersprachlicher und daher akzentbehafteter Sprache, da sowohl die Muttersprache des Sprechers als auch dessen Vorerfahrung einen Einfluß auf die Aussprache der Nichtmuttersprache haben.

1.4 Schnelle Portierung auf neue Sprachen

Aufgrund der Forderung nach Flexibilität eines Spracherkennungssystems beschäftigt sich die vorliegende Arbeit eingehend mit der Frage, wie der Datenumfang und der Entwicklungsaufwand für neue Sprachen reduziert werden können. Zu diesem Zweck wird untersucht, wie sich die Limitierung von Trainingsdaten auswirkt und welchen Effekt es hat, wenn nur wenig über die neue Sprache bekannt ist bzw. keine Experten verfügbar sind.

Am ILKD sind günstige Voraussetzungen für die Forschung an Erkennungssystemen gegeben, da mit JRTk ein sehr flexibles Spracherkennungstoolkit zur Verfügung steht. Die Entwicklung eines Spracherkenners für große Wortschätze besteht damit aus den Schritten Sammlung und Aufbereitung relevanter Sprachdaten, Festlegung des Phoneminventars, Erstellung eines Aussprachewörterbuchs, Aufbereitung großer Textkorpora zur Berechnung eines Sprachmodells, Training der akustischen Modelle sowie Feinabstimmung des Systems und dessen Evaluation.

Dabei ist die Sammlung der Sprachdaten, wie bereits ausgeführt, der teuerste und zeitraubendste Schritt. Die Idee, die dieser Arbeit zugrundeliegt, besteht darin, den Umfang der benötigten Sprachdaten für eine *neue* Sprache dadurch zu reduzieren, daß Daten bereits bekannter Sprachen mitverwendet werden. Dazu wird der oben beschriebene multilinguale Erkenner auf die neue Sprache angepaßt. Diese Arbeit geht von der Annahme aus, daß sprachenunabhängige Modelle sehr viel robuster auf einen Wechsel der Sprache reagieren und daher mit *kleinen Datenmengen* auf neue Zielsprachen portiert werden können.

Die Vorteile dieser Vorgehensweise bestehen in Folgendem:

- Ein universelles Phoneminventar ermöglicht sowohl die Initialisierung eines Erkennungssystems in einer neuen Sprache als auch die Adaption bestehender Systeme auf neue Sprachen.
- Durch gemeinsame Nutzung von Trainingsmaterial wird die Gesamtmenge der zur robusten Parameterschätzung notwendigen Audiodaten pro Sprache reduziert.
- Der finanzielle und zeitliche Aufwand für die Entwicklung von Erkennern in neuen Sprachen verringert sich.

- Trainieren auf Daten verschiedener Sprachen erhöht die Robustheit der Modelle aufgrund breit gestreuter Modellierung (was nicht notwendigerweise zu einer Verbesserung der Erkennungsraten führen muß).
- Aufgrund der begrenzten Anzahl vom Menschen produzierbarer Laute besteht die Hoffnung, daß ungesehene Laute mit jeder neuen Sprache seltener werden und das Lautinventar bald vollständig abgedeckt wird.

1.5 Gliederung der Arbeit

Der Aufbau der Arbeit entspricht im wesentlichen drei zu lösenden Problemkreisen:

1. Datensammlung und Bau monolingualer Basiserkenner
2. Konzeption und Entwicklung eines multilingualen Erkenners
3. Portierung des multilingualen Erkenners auf neue Sprachen

Kapitel 2 gibt einen Überblick über Merkmale von Sprachen, die für die vorliegende Arbeit relevant sind. Die Einführung in die Grundlagen der multilingualen Spracherkennung in Kapitel 3 definiert alle wesentlichen Begriffe, die zum Verständnis der Arbeit notwendig sind. Gemeinsam mit einem Abriß über die relevanten Arbeiten anderer Gruppen in den Abschnitten 5.2, 6.2 und 7.2 führen die beiden vorherigen Kapitel in das Themengebiet der multilingualen Spracherkennung ein.

Die Kapitel 4 und 5 beschreiben die wesentlichen Arbeiten zur Lösung des ersten Problemkreises. In Kapitel 4 wird das GlobalPhone-Projekt beschrieben, dessen Datensammlung eine notwendige Voraussetzung für diese Arbeit bildet. Kapitel 5 beschreibt die prinzipielle Vorgehensweise beim Bau monolingualer Erkenner. Deren Entwicklung und Vergleich nehmen einen großen Umfang in dieser Arbeit ein, da erst die Analyse der monolingualen Erkenner in vielen Sprachen zeigt, ob bekannte Algorithmen vergleichbare Leistungen in allen Sprachen zeigen und ob es sinnvoll ist, Erkennerkomponenten sprachenuniversell zu erstellen. Kapitel 6 ist der Lösung des zweiten Problemkreises, der Entwicklung multilingualer akustischer Modelle und deren Integration in ein multilinguales Erkennungssystem gewidmet. Der dritte zu lösende Problemkreis umfaßt die Konzeption und die Implementierung von Verfahren zur Portierung des multilingualen Erkenners auf neue Zielsprachen. Damit beschäftigt sich Kapitel 7. In Kapitel 8 wird das GlobalPhone-Demonstrator-System vorgestellt, das im Rahmen dieser Arbeit entstanden ist. Kapitel 9 bildet mit der Zusammenfassung und einem Ausblick den Abschluß der Arbeit.

Kapitel 2

Die Sprachen der Welt

Dieses Kapitel liefert Informationen zu Sprachen, die im Kontext der multilingualen Spracherkennung relevant sind. Neben einer Einführung der wichtigsten Strukturbegriffe wird beschrieben, wieviele Sprachen es gibt, welchen Verbreitungsgrad sie haben und welchen Stellenwert man ihnen beimißt. Die Ausführungen orientieren sich an den Arbeiten von [Cry95, Stö97, PM95]. Abschließend werden die Unterschiede zwischen Sprachen erörtert und die daraus resultierenden Konsequenzen für die multilinguale Spracherkennung abgeleitet.

2.1 Wieviele Sprachen gibt es?

Über die Gesamtzahl der heute auf der Erde gesprochenen Sprachen herrscht keine Einigkeit. Die meisten Nachschlagewerke geben Zahlen zwischen 4000 und 5000 an [Edw95, Stö97, Web92, CMP98], andere Schätzungen schwanken zwischen 3000 und 10.000 [Cry95]. Diese Unsicherheit ist zum einen in der Entwicklungsdynamik begründet, mit der Sprachen entdeckt und erfaßt werden oder aussterben. Es werden nämlich nur solche Sprachen gezählt, die noch aktiv als Muttersprache gesprochen werden. Die Unsicherheit über die Sprachenzahl hat zum anderen aber auch historische, politische, ethnische, religiöse, kulturelle und literarische Hintergründe, die im folgenden erläutert werden.

Entdeckungen
Noch heute werden in neuerforschten Gebieten der Erde auch Völker und damit neue Sprachen entdeckt. Häufiger kommt es allerdings vor, daß zwar die Bewohner einer Region bereits bekannt sind, nicht jedoch deren gesprochene Sprache. Darüber hinaus sind in vielen Ländern die Landessprachen nicht oder nur unvollständig erfaßt [Cry95].

Lebend oder tot

Neben dem Faktor „Entdeckungen“, der die Gesamtzahl der Sprachen erhöht, gibt es einen wichtigen Faktor, der sie reduziert: Im Falle kleiner Sprachgemeinschaften können Sprachen mit erstaunlicher Geschwindigkeit bereits innerhalb einer Generation aussterben. Nach [Cry00] stirbt auf der ganzen Welt im Schnitt alle zwei Wochen eine Sprache aus. Es kann aufgrund politischer oder wirtschaftlicher Erwägungen zur Spaltung oder Auflösung einer Gemeinschaft kommen oder, wie zur Zeit der Kolonialisierung nicht selten, zur Ausrottung oder dem plötzlichen Niedergang einer ganzen Sprachgemeinschaft [Cry95].

Sprache oder Dialekt

Die Existenz einer *Sprache* manifestiert sich in einer zugrundeliegenden Einheit, welche die Sprecher als ihre Sprache identifizieren, und die durch Verwendung einer standardisierten Schriftsprache und eines gemeinsamen literarischen Erbes bestätigt wird [Cry95]. Dagegen wird ein *Dialekt* als regional begrenzte, ursprüngliche und nicht an die Normen der Standardsprache gebundenen Sprachform definiert [DK78] und hängt nach [Web92] von Faktoren wie Erziehung (soziale Schicht, Beruf, Gesellschaftssegment) und dem Grad der Isolation der Sprachgruppe ab. Der Dialekt ist im wesentlichen gesprochene Sprache und unterscheidet sich von der Standardsprache durch stärkere Bildhaftigkeit, viele Augenblicksbildungen und Analogien. Die Standardsprache ist die über Dialekten, Umgangssprachen und Gruppensprachen stehende allgemeinverbindliche genormte Sprachform, die vor allem in der Literatur, im wissenschaftlichen Schrifttum, Presse, Funk und Fernsehen und anderen öffentlichen Bereichen verwendet wird.

In den meisten Fällen lassen sich Sprache und Dialekt unbestreitbar unterscheiden, häufig ist es aber schwierig, klare Grenzlinien zu ziehen. Zur Unterscheidung zwischen Sprache und Dialekt wird das Kriterium der *gegenseitigen Verständlichkeit* angewendet: Wenn sich zwei Menschen sprachlich miteinander verständigen können, ordnet man die von ihnen gesprochene Sprachform derselben Sprache zu. Allerdings gibt es Fälle, in denen man verschiedene Sprachformen derselben Sprache zuschreibt, obwohl sie gegenseitig nicht mehr verständlich sind. Beispielsweise sind alle chinesischen „Dialekte“ zur Sprache Chinesisch zusammengefaßt, obwohl sich ihre Sprecher untereinander nicht mündlich, sondern nur über die gemeinsame Schrift verständigen können.

Noch größere Probleme bereitet das Kriterium der Verständlichkeit bei einem sogenannten Dialektkontinuum, bei dem innerhalb einer geographischen Region eine Kette von Dialekten gesprochen wird. Sprecher eines bestimmten Dialektes sind in der Lage einen anderen Dialekt in unmittelbarer Nachbarschaft zu verstehen, aber andere auf der Kette weiter entfernt liegenden Dialekte können für sie völlig unverständlich sein. In Europa existieren eine ganze Reihe solcher Kontinua: Die Kette von Deutsch (Niederrhein) über Niederländisch nach Flämisch, oder das skandinavische Kontinuum, welches Dialekte des Norwegischen, des Schwedischen und des Dänischen miteinander verbindet. Das nordslawische Kontinuum verknüpft das Slowakische, Tschechische, Ukrainische, Polnische und Russische miteinander. Auf lokaler Ebene läßt sich auf einer solchen Kette kein Punkt bestimmen, an dem eine Sprache aufhört und die nächste anfängt. Trennlinien werden dann anhand anderer Kriterien gezogen, zum Beispiel anhand bestehender Landesgrenzen. So gilt das Plattdeutsche gewöhnlich als deutscher Dialekt, aber Niederländisch, als Variante des Platt- oder Niederdeutschen, wird als eigenständige Sprache bezeichnet. Es kommt auch vor, daß Sprache zum Ausdruck eines Gebietsanspruchs wird. Beispielsweise werden Dialekte, die im Grenzgebiet von Jugoslawien und Bulgarien gesprochen werden, von den Jugoslawen als mazedonische aber von den Bulgaren als bulgarische Dialekte bezeichnet.

Auch die Existenz eines eigenständigen Staates oder die Tatsache, daß ein Dialekt als Amts- oder Verkehrssprache benutzt wird, ist kein zuverlässiges Kriterium zur Abgrenzung von Sprache und Dialekt. Denn dann dürfte das Baskische oder Katalanische nicht als Sprache bezeichnet werden und das Rätoromanische hätte vor dem Zeitpunkt, als man es in der Schweiz zur 4. Amtssprache erhob, als Dialekt bezeichnet werden müssen.

Es gibt viele Fälle, in denen politische, ethnische, religiöse oder andere Gründe eine Aufteilung erzwingen, obwohl es sprachlich keine gibt. Beispiele dafür sind Hindi/Urdu, Bengali/Assamesisch, Serbisch/Kroatisch oder Xhosa/Zulu [Cry95]. Allerdings kommt auch die umgekehrte Situation vor, in der wie das Beispiel Chinesisch zeigt, Sprachformen zu einer Sprache zusammengefaßt werden, obwohl sie nicht zusammengehören.

Sprachnamen

Das Auszählen von Sprachen ist eng verknüpft mit dem Problem der Bezeichnung einer Sprache. Große Weltsprachen sind im allgemeinen unter einem einzigen Sprachnamen bekannt, der sich in andere Sprachen übersetzen läßt wie etwa *Deutsch, German, Tedesco, Nemetsky, Allemand*. Manche Gemeinschaften haben dagegen keinen spezifischen Namen für ihre Sprache, wie etwa die afrikanische Bantu Sprache, in der das Wort Bantu *Volk* bedeutet. Desweiteren gibt es Fälle in denen viele Namen für eine einzige Sprache existieren: Einen, den sich die Gemeinschaft selbst gibt, einen, unter dem sie die Nachbarstämme kennen, einen, den Eroberer der Sprache gaben und einen, den Anthropologen nach der geographischen Region zuteilten. Neben diesen drei genannte Faktoren, die das Sprachenzählen erschweren, kommen noch Mehrdeutigkeiten hinzu, die sich aus unterschiedlichen Schreibweisen ergeben.

Schlußfolgerung

Mit den oben beschriebenen Problemen sind noch längst nicht alle Schwierigkeiten aufgezählt: Wenn sich die Existenz einer Sprache in der Verwendung einer standardisierten Schriftsprache manifestiert, wie geht man mit Mundartdichtung um, und wie soll man mit solchen Sprachen verfahren, die über keine Schrift verfügen. Wie betrachtet man Sondersprachen, wie beispielsweise Rotwelsch, oder Fachsprachen wie die Computersprache oder Pidgin- und Kreolensprachen?

Die angeführten Argumente sollen veranschaulichen, daß sich die Frage nach der Zahl der Sprachen nicht mit einer einzigen Zahl beantworten läßt. Es wird aber trotz alledem immer wieder versucht, möglichst konkrete Zahlen zu nennen. Die in Abbildung 2.1 aufgeführten Sprecherzahlen für die Sprachen der Welt beruhen auf Voegelins Schätzungen in *Classification and Index of the World's Languages* von 1977, in dem über 20.000 Bezeichnungen für Sprachen und Dialekte aufgeführt sind. Danach beträgt die Gesamtzahl der Sprachen etwa 4380. Seit der Veröffentlichung ist die Gesamtzahl der Sprachen wahrscheinlich gesunken, nach Näherungen von [Cry95] liegt sie aber nicht unter 4000.

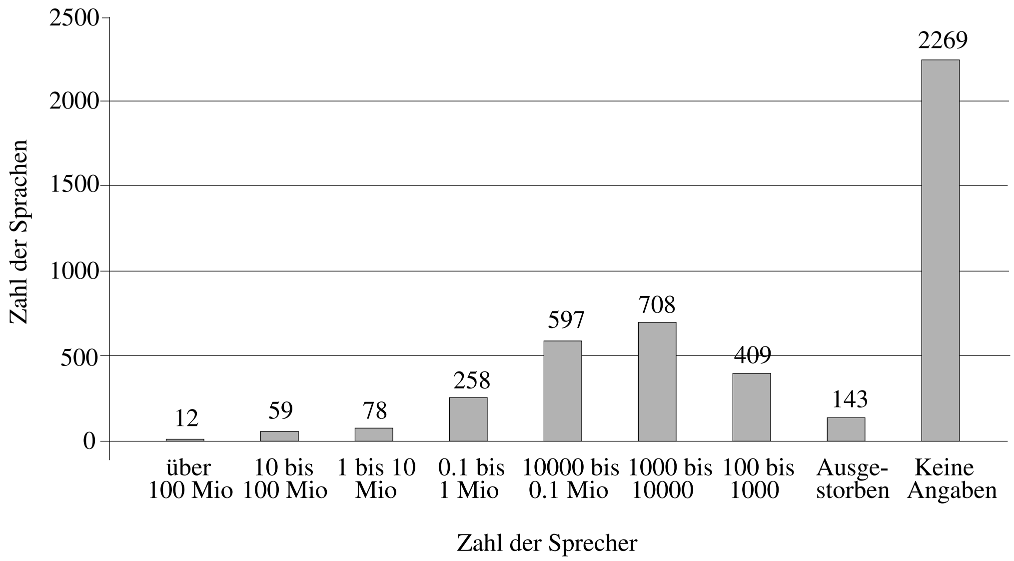


Abbildung 2.1: Geschätzte Zahl der Sprachen und deren Sprecherzahlen

2.2 Verbreitung und Stellenwert von Sprachen

Die Schätzungen der Zahl der Sprecher einer Sprache ist noch komplizierter als die Bestimmung der Gesamtzahl aller Sprachen. Denn die Anzahl der Sprecher kann innerhalb kurzer Zeitspannen großen Schwankungen unterworfen sein. Daher sind hauptsächlich die weltweiten Veränderungen der Bevölkerungszahlen dafür verantwortlich, daß Schätzungen bereits bei ihrer Veröffentlichung veraltet sind. Wegen des hohen Aufwands finden Volkszählungen nur selten statt, daher basieren heutige Schätzungen von Sprecherzahlen auf Datenmaterial, das bereits 1970 erhoben wurde. Damals lag die Gesamtbevölkerung noch unter 4 Milliarden Menschen, Mitte 1987 wurde sie auf 5 Milliarden geschätzt, und für das Jahr 2000 prognostizierte die

UNO damals das 6 milliardste Kind, dessen Geburt am 11. Oktober 1999 gefeiert wurde.

Zuordnungsprobleme ergeben sich in mehrsprachigen Gemeinschaften, in denen viele Menschen bi- oder gar trilingual aufwachsen und sich daher mehreren Muttersprachen zuordnen. Umgekehrt gilt beispielsweise in Ländern wie Indien, daß große Teile der Bevölkerung ihre eigene offizielle Amtssprache Englisch nur mangelhaft beherrschen. Desweiteren kann es bei Volkszählungen vorkommen, daß Menschen Sprachkenntnisse vortäuschen, um kulturelle oder soziale Ziele zu unterstützen, oder daß Regierungen Sprecherzahlen von Minderheitensprachen manipulieren, um deren Bedeutung herunterzuspielen. Diese Liste ließe sich noch endlos weiterführen.

Die Tabelle 2.1 vermittelt einen Überblick über die geschätzte Anzahl von Muttersprachlern und den Sprachraum der bedeutendsten Sprachen der Welt, zusammengestellt nach Angaben aus [Web92].

	Sprache	Sprecher	primärer Sprachraum	Sprachfamilie
1	Mandarin	907 Mio	China	Sino-Tibetan (Sinitisch)
2	Englisch	456 Mio	USA, UK, Canada, Australien	I-E[1] (Germanisch)
3	Hindi	383 Mio	Indien	I-E (Indo-Iranisch)
4	Spanisch	362 Mio	Latein-Amerika, Spanien	I-E (Romanisch)
5	Russisch	293 Mio	Rußland, unabh.russ. Staaten	I-E (Slawisch)
6	Arabisch	208 Mio	Nord-Afrika, mittl. Osten	Afro-Asiatisch (Semitisch)
7	Bengalisch	189 Mio	Bangladesh, Indien	I-E (Indo-Iranisch)
8	Portugiesisch	177 Mio	Brasilien, Portugal, Angola	I-E (Romanisch)
9	Malay-Indo.	148 Mio	Indien, Malaysia, Brunei	Austronesisch (Polyn.)
10	Japanisch	126 Mio	Japan	isolierte Sprache
11	Französisch	123 Mio	F, Canada, Afrika, Schweiz	I-E (Romanisch)
12	Deutsch	119 Mio	D, Österreich, Schweiz	I-E (Germanisch)
13	Urdu	96 Mio	Pakistan, Indien	I-E (Indo-Iranisch)
14	Punjabi	89 Mio	Indien, Pakistan	I-E (Indo-Iranisch)
15	Koreanisch	73 Mio	Korea, China	isolierte Sprache
16	Telugu	71 Mio	Indien	Dravidisch
17	Tamil	67 Mio	Indien, Sri Lanka, Malaysia	Dravidisch
17	Marathi	67 Mio	Indien	I-E (Indo-Iranisch)
19	Kantonesisch	65 Mio	China, Hong Kong	Sino-Tibetan (Sinitisch)
20	Wu/Schanghai	64 Mio	China (Schanghai)	Sino-Tibetan (Sinitisch)
21	Italienisch	63 Mio	Italien, Schweiz	I-E (Romanisch)
22	Vietnamesisch	61 Mio	Vietnam	Austro-Asiatisch (Khmer)
22	Javanesisch	61 Mio	Indonesien	Austronesisch
24	Thai	59 Mio	Thailand	Sino-Tibetan (Thai)
25	Türkisch	57 Mio	Türkei	Altaisch (Turksprache)
...	...	...	...	...
43	Kroatisch	20 Mio	Balkan	I-E (Slawisch)
...	...	...	...	...
85	Schwedisch	9 Mio	Schweden, Finnland	I-E (Germanisch)

Tabelle 2.1: Die wichtigsten Weltsprachen (nach [Web92])

Eine Weltsprache?
Einige Forscher sind der Meinung, daß sich in naher Zukunft eine einzige Sprache als *Weltsprache* durchsetzen wird. Dies könnte sowohl eine natürlich gewachsene Sprache, als auch eine künstliche Sprache sein. Mit Kunstsprachen wie beispielsweise *Esperanto* wurde immer wieder der Versuch unternommen, sich auf eine gemeinsame Minimalkonstruktion zu einigen. Künstliche Sprachen werden in dieser Arbeit nicht betrachtet. Dagegen sind alle natürlichen Sprachen, die als mögliche Anwärter einer Weltsprache gehandelt werden, von besonderer Bedeutung für die Entwicklung zukünftiger Spracherkenner.

In etwa 25% aller Länder hat mehr als eine Sprache offiziellen oder halboffiziellen Status. Orientiert man sich bei der Zählung daher nicht an der Zahl der Muttersprachler, sondern an der Zahl der Menschen, die eine Sprache als Amtsprache sprechen, ergibt sich gegenüber Tabelle 2.1 eine leicht veränderte Rangfolge, die in Tabelle 2.2 für die 20 verbreitetesten Sprachen nach [Cry95] aufgeführt ist. Aus den oben genannten Gründen sind diese Sprecherzahlen in der Regel etwas zu hoch gegriffen, können aber dennoch als guter Indikator für sprachliche Tendenzen herangezogen werden. Danach ist Englisch derzeit die am weitesten verbreitete Sprache. Sie ist Amtssprache in 45 Ländern [Cry95] und die meistgelernte Fremdsprache der Welt [Sku88]. Deshalb gilt Englisch bei vielen als Hauptanwärter für eine Weltverkehrssprache. Zwar spielen in großen Teilen der Welt andere Sprachen eine wichtigere Rolle, etwa Russisch in Osteuropa und Spanisch in Mittel- und Südamerika, doch erreichen diese Sprachen keine so globale Verbreitung wie das Englische. Und obwohl Chinesisch mehr Muttersprachler hat als Englisch, ist sie im Westen zu ungebräuchlich, als daß sie ernsthaft als Anwärter für eine Weltsprache in Betracht käme. Französisch, das vom 17. bis zum Beginn des 20. Jahrhundert als Sprache der Diplomatie galt, ist zwar heute noch sehr weit verbreitet, doch ist deren Gebrauch seit damals rückläufig.

Rang	Sprache	Sprecher	Rang	Sprache	Sprecher
1	Englisch	1400 Mio	11	Japanisch	120 Mio
2	Mandarin	1000 Mio	12	Deutsch	100 Mio
3	Hindi	700 Mio	13	Urdu	85 Mio
4	Spanisch	280 Mio	14	Italienisch	60 Mio
5	Russisch	270 Mio	15	Koreanisch	60 Mio
6	Französisch	220 Mio	16	Vietnamesisch	60 Mio
7	Arabisch	170 Mio	17	Persisch	55 Mio
8	Portugiesisch	160 Mio	18	Tagalog	50 Mio
9	Malaiisch	160 Mio	19	Thai	50 Mio
10	Bengalisch	150 Mio	20	Türkisch	50 Mio

Tabelle 2.2: Die 20 bedeutendsten Amtssprachen der Welt (nach [Cry95])

[1] I-E=Indoeuropäisch

Welche Sprachen sind die wichtigsten?
Der Verbreitungsgrad ist sicherlich eines der wesentlichen Kriterien für die Bedeutung einer Sprache, aber es spielen auch wirtschaftliche Überlegungen eine Rolle. Nach [Wah99] ergibt sich der Stellenwert einer Sprache aus dem Produkt von der Anzahl ihrer Sprecher und dem internationalen Warenaustausch. Tabelle 2.3 zeigt die nach wirtschaftlichen Gesichtspunkten neun stärksten Nationen der Welt [Web92]. Aufgeführt sind die Rangzahlen für das Bruttosozialprodukt (BSP), für den Im- und Export sowie für die Verkaufszahlen von landessprachlichen Büchern.

Land	BSP	Import	Export	Bücher
Amerika	1	1	1	4
Japan	2	3	3	
Deutschland	3	2	2	1
Frankreich	4	4	4	5
Italien	5	6	6	
Großbritannien	6	5	5	2
Russland	7			3
Niederlande		7	7	
Südkorea				6

Tabelle 2.3: Wirtschaftlicher Status der neun stärksten Länder (nach [Web92])

Die DARPA[2] hat in einem Zweistufenplan 10 Sprachen ausgewählt, denen besondere Aufmerksamkeit zuteil wird. Bei der Auswahl legten sie weitere Maßstäbe an, die ihrer Meinung nach die allmähliche Ausbreitung einer Sprache bestimmen: politische und militärische Macht, wirtschaftliche Potenz und Zukunftspotential und religiösen Einfluß. Auf der ersten Stufe steht im DARPA-Plan Englisch aufgrund seiner Verbreitung, Arabisch wegen der Ölvorkommen, Spanisch wegen der Drogenproblematik, Japanisch wegen des Geldes und Chinesisch wegen seines Zukunftspotentials. Auf der zweiten Stufe kommen dann Türkisch, Russisch, Hindi, Koreanisch, Vietnamesisch und Indonesisch [Hov99].

2.3 Klassifikation von Sprachen

Die ersten wissenschaftlichen Versuche, Sprachen zu klassifizieren, wurden gegen Ende des 18. Jahrhunderts unternommen. Man begann, Sprachen systematisch miteinander zu vergleichen und ihre Beziehungen zu ergründen. Im wesentlichen gibt es zwei Ansätze zur Klassifikation von Sprachen: die genetische und die typologische Klassifikation.

[2]Defense Advanced Research Projects Agency

Die *typologische Klassifikation* beruht auf dem Vergleich formaler Ähnlichkeiten zwischen Sprachen. Dabei wird versucht, Sprachen auf der Grundlage ihrer Phonologie, Grammatik und ihres Wortschatzes in strukturelle Typen einzuteilen. Die ersten Typologien wie beispielsweise die von Schlegel (1767-1845) bezogen sich auf morphologische Aspekte, bei denen Sprachen entsprechend ihrer Wortbildungsmuster eingeteilt wurden. Es gibt sechs Möglichkeiten für die Anordnung der Satzglieder Subjekt (S), Verb (V) und Objekt (O). Mehr als drei Viertel aller Sprachen verwenden SVO (u.a. Englisch, Deutsch, Französisch) oder SOV (u.a. Japanisch und Koreanisch), weitere 10-15% konstruieren nach dem Muster VSO (Walisisch, Tonga). Manche Sprachen haben freie Wortstellung. Typologische Fragestellungen sind heute vor allem bei der Suche nach sprachlichen *Universalien* von großem Interesse. Insgesamt ist allerdings fraglich, ob typologische Klassifikationen generell möglich sind, weil es keine reinen Vertreter der einzelnen Sprachtypen gibt. Jede Sprache weist in mehr oder weniger starker Ausprägung Merkmale aller Typen auf. Beispielsweise verwendet das Deutsche für Hauptsätze die Wortstellung SVO, für Nebensätze aber SOV, im Japanischen wird SOV bevorzugt, aber OSV ist auch gebräuchlich [Cry95]. Darüber hinaus hängt die typologische Klassifikation immer von der Beurteilung der zugrundeliegenden Merkmale ab. Berücksichtigt man alle Sprachmerkmale, ergibt sich eine Unzahl möglicher Klassifikationen und damit das Problem der Gewichtung von Kriterien. Die linguistische Theorie steht diesbezüglich noch ganz am Anfang.

Die *genetische Klassifikation* ist historisch orientiert und stützt sich auf die Annahme, daß Sprachen von einem gemeinsamen Vorläufer abstammen. Nach [Bod97] gelten Sprachen als verwandt, „wenn Wortschatz, Struktur und Lautsystem Ähnlichkeiten aufweisen, die auf eine gemeinsame Ursprache hindeuten“. Zur Veranschaulichung von historischen Verwandtschaftsbeziehungen zwischen Sprachen zog Schleicher (1821-1868) das Bild der Sprachfamilie oder des Stammbaumes heran. Die klassische Vorgehensweise der historischen Sprachwissenschaft ist die *komparative Methode*. Dabei wird ausgehend von formalen Ähnlichkeiten und Unterschieden zwischen den Sprachen versucht, einen gemeinsamen Vorläufer zu finden, von dem sämtliche Formen herrühren könnten. Gelingt der Nachweis eines gemeinsamen Vorläufers, gelten die Sprachen als verwandt.

Zur Rekonstruktion der Verwandtschaftsbeziehungen bedient man sich meist zweier Verfahren: Beim Verfahren, das auf der Idee der Lautverschiebung basiert, sucht man nach einen gemeinsamen Vorläufer, von dem sich der Lautbestand der zu vergleichenden Sprachen durch regelmäßige Lautveränderungen ableiten läßt. Dieses Verfahren war sowohl bei den indoeuropäischen als auch bei den austronesischen Sprachen sehr erfolgreich [Kri88]. Bei Sprachpaaren, die weit auseinander liegen, führt diese Methode aber zu Problemen, da Lautverschiebungen nicht immer regelmäßig verlaufen. Beim Verfahren des Massenvergleichs von Greenberg wird die Verwandtschaft anhand von Wortlisten bestimmt. Je mehr Wörter in ihrer lautlichen Form und Bedeutung einander ähneln, umso enger die Verwandtschaft. Vor allem der Einsatz von Computern hat dieses Verfahren vorangetrieben.

Für Eurasien war aufgrund der reichen Schrifttradition die Methode der genetischen Klassifikation sehr erfolgreich, für Sprachen anderer Regionen der Erde hat sie dagegen oft nur hypothetischen Charakter.

Nach Greenberg [Kri84] hatte der *homo sapiens* die erste menschliche Lautsprache. Seine Vorfahren waren aus anatomischen Gründen nicht in der Lage, Lautsprache zu produzieren. Greenberg ist ein Verfechter der Monogenese, d.h. der Annahme, nach der alle Sprachen aus einer gemeinsamen Ursprache hervorgingen, die durch kulturelle Evolution oder göttliche Eingebung entstanden ist. Ein eventuelles Relikt aus dieser einen Ursprache könnte die Folge t-Vokal-k sein, die in mindestens 15 weitverbreiteten Sprachfamilien vorkommt und in den Familien Indoeuropäisch, Austronesisch, Eskimo-Aleutisch, Na-Dene sowie Amerind die Bedeutung *Hand, Finger oder hinweisen* hat. Im Deutschen stammt vermutlich *zeigen* von der indoeuropäischen Wurzel *deik* ab.

Die bedeutendsten Sprachfamilien
Die Abbildung 2.2 zeigt die geographische Verteilung der lebenden, allgemein akzep-

Abbildung 2.2: Geographische Verteilung der Sprachenfamilien (nach [Edw95])

tierten Sprachfamilien der Welt. Dort, wo infolge des Kolonialismus indoeuropäische Sprachen die heimischen Sprachen in eine Minderheitenrolle gedrängt haben, ist statt der heute dominanten die ursprüngliche Sprachfamilie dargestellt. Es werden zwischen 20 und 29 allgemein akzeptierte Großfamilien angegeben, je nachdem ob Japanisch und Koreanisch als eigenständige Großfamilie gezählt werden [Cry95, Gri92], oder der altaischen Familie zugeordnet werden [Edw95], und ob die von Greenberg postulierte amerindische Sprachfamilie akzeptiert wird [Ros91]. Die Tabelle 2.4 vermittelt einen Überblick über die Sprecherzahlen dieser Sprachfamilien nach [Cry95] und [Edw95] sowie einige wichtige Vertreter dieser Familien. Danach sprachen fast die Hälfte der 1981 lebenden Weltbevölkerung eine indoeuropäische Sprache.

Sprachfamilie	Sprecherzahlen	wichtige Vertreter
Indoeuropäisch	2 000 000 000	Englisch, Deutsch, Russisch
Sino-Tibetisch	1 050 000 000	Birmanisch, Chinesisch, Tibetisch
Niger-Kongo	260 000 000	Igbo, Suaheli, Xhosa
Hamito-Semitisch	230 000 000	Arabisch, Haussa, Hebräisch
Austronesisch	200 000 000	Javanesisch, Malayisch, Tagalog
Drawidisch	140 000 000	Tamil, Telugu, Kannada
Japanisch	120 000 000	Isoliert
Altaisch	90 000 000	Aserbaidschanisch, Türkisch, Usbekisch
Austro-Asiatisch	60 000 000	Khmer, Vietnamesisch, Santali
Koreanisch	60 000 000	Isoliert
Thai	50 000 000	Laotisch, Thai, Schan
Nilo-Saharanisch	30 000 000	Dinka, Nande, Nubisch
Amerind	25 000 000	Eskimo, Navajo, Aztekisch
Uralisch	23 000 000	Estnisch, Finnisch, Ungarisch
Kaukasisch	6 000 000	Awarisch, Tschetschenisch, Georgisch
Indo-Pazifisch	3 000 000	Motu, Enga, Medlpa
Khoisan	50 000	Kwadi, Sandawe
Ur-Australisch	50 000	Tiwi, Walmatjari, Pitjantjatjara
Paläosibirisch	25 000	Tschuktschisch, Korjakisch, Giljakisch

Tabelle 2.4: Geschätzte Sprecherzahlen aller Sprachfamilien (nach [Cry95, Edw95])

Die indoeuropäische Sprachfamilie

Auf die indoeuropäische Sprachfamilie soll hier kurz näher eingegangen werden, weil sie die Familie mit den meisten Sprechern ist, und ihr daher bei der Entwicklung eines multilingualen Spracherkenners eine wichtige Rolle zufällt. Die Mehrzahl der im Rahmen von GlobalPhone gesammelten Sprachen gehören dieser Familie an.

Die indoeuropäische Sprachfamilie, früher auch als Indogermanisch bezeichnet, gilt gemeinhin als die am besten erforschte Familie. Sie hat sich über Europa und weite Teile Südasiens ausgebreitet. Ihre Nachfahren sind heute infolge des Kolonialismus auf der ganzen Welt zu finden. Abbildung 2.3 zeigt eine schöne Darstellung des Stammbaums der indoeuropäischen Sprachfamilie.

Erstmals vermutete Jones (1746-1794), daß die zu dieser Zeit nur aus dem europäischen Raum bekannten Sprachen mit dem Sanskrit aus dem indischen Raum einen gemeinsamen Ursprung haben könnten und begründete damit die indoeuropäischen Hypothese. Über die Arbeiten von Bopp, Rask, Grimm, Schleicher und Brugmann gelang dann im 19. Jahrhundert mit Hilfe der komparativen Methode der Existenzbeweis eines gemeinsamen Sprachenvorläufers, dem Proto-Indoeuropäischen.

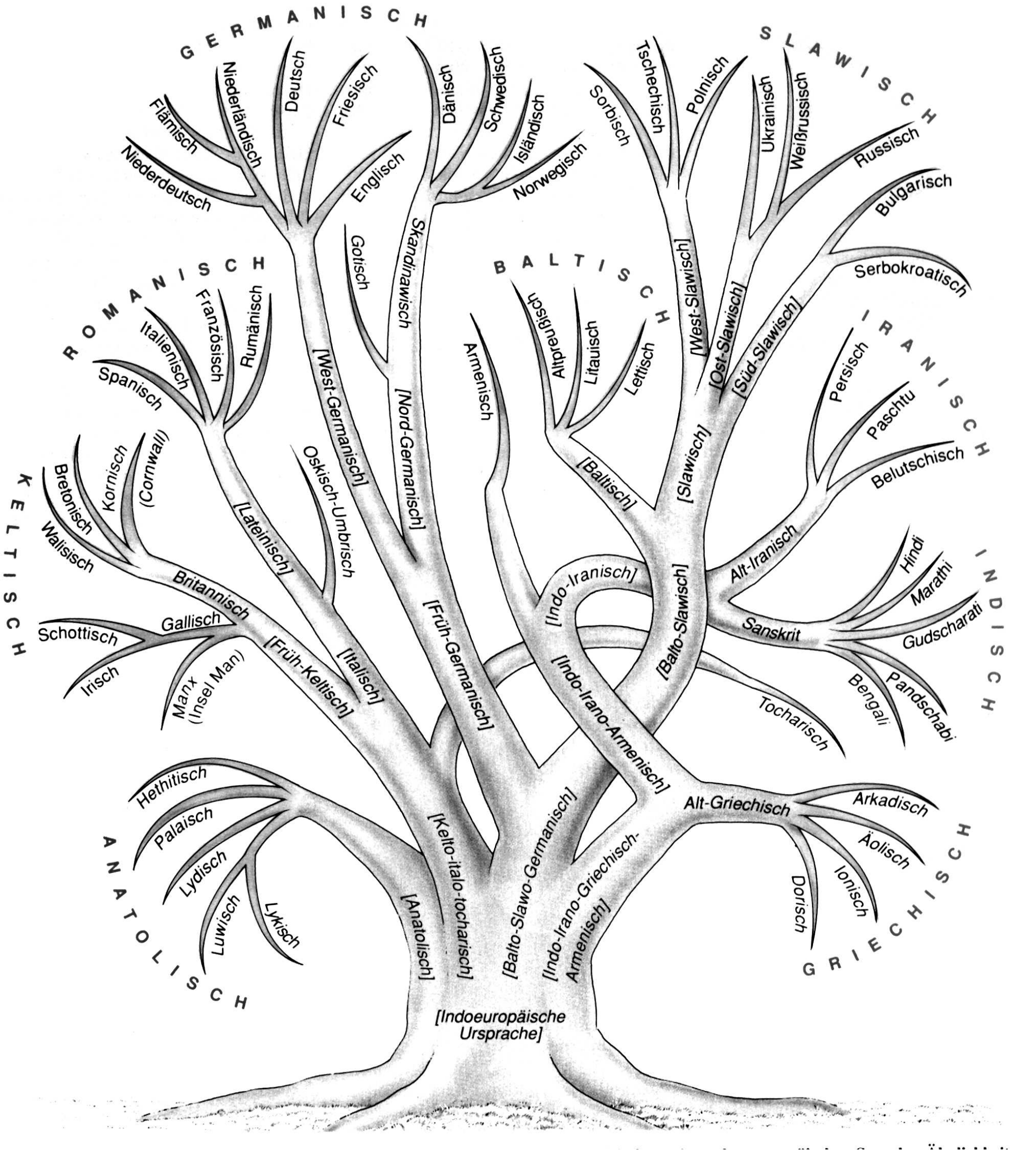


Abbildung 2.3: Stammbaum der indoeuropäischen Sprachenfamilie [GI90]

Aufgrund einer interessanten Untersuchung behaupten [GI90], daß unsere Ursprache vor mehr als 6000 Jahre entstand und ihr Ursprung nicht, wie bisher angenommen, in Europa liegt, sondern in Anatolien. Mit Hilfe der Lautveränderungsgesetze sowie aufgrund neuer Schriftfunde und Entwicklungen in der theoretischen Linguistik, kommen sie zu dem Schluß, daß viele Begriffe der Ursprache eine Natur und Landschaft beschreiben, wie sie nur in Anatolien zu finden ist. Diese Behauptung deckt sich mit den Ergebnissen von [Ren89], der aufgrund archäologischer Funde argumentiert, daß die friedliche Ausdehnung der Ausgangspunkt der Verbreitung der indoeuropäischen Sprachen war, und daß sie mit Ackerbau und Seßhaftigkeit in Anatolien ihren Anfang nahm. Eine der wesentlichen Hypothesen ist, daß die Ausdehnung durch Wanderung in drei große Richtungen voranging. Erstens in den Westen nach Griechenland (Dorische Einwanderung), zweitens nach Osten Richtung Indien über den Iran und nach Ostturkestan, und drittens zunächst nach Osten, dann aber in einem großen Bogen um das Kaspische Meer in das heutige Mitteleuropa hinein. [GI90] gehen davon aus, daß die meisten westlichen Sprachen von diesem letztgenannten Zweig abstammen und während der Wanderung durch den Kontakt mit semitischen und kartwelischen Sprachen viele Wörter entlehnt haben. Als Beweis dafür führen [GI90] das Wort für Traube an: Im Russischen und Italienischen heißt es vino, im Germanischen wein, das auf woi-no oder weo-no zurückgeht. Im Frühsemitischen heißt es wajnu, im Ägyptischen wns, im Kartwelischen wino und im Hethitischen wijana.

Probleme der Klassifikation von Sprachen

Das Beispiel der indoeuropäischen Sprachen veranschaulicht, daß sowohl die typologische als auch die genetische Klassifikation einen gravierenden Mangel haben: Sie lassen die Bedeutung kultureller Verbindungen zwischen Sprachen außer acht. Historisch nicht verwandte Sprachen können sich aufeinander zubewegen oder verwandte Sprachen durch die starke Beeinflussung verschiedener Kulturen voneinander soweit entfernen, daß die Unterschiede stärker zutage treten als die Gemeinsamkeiten. Es ist oftmals nicht zu unterscheiden, ob sich zwei Sprachen ähneln, weil sie gleichen Ursprungs sind oder weil sie sich gegenseitig beeinflußt haben.

Das Wissen über die Verwandtschaftsbeziehungen zwischen Sprachen kann helfen, die Struktur einer Sprache zu verstehen. Zur Konzeption multilingualer Komponenten im Kontext der automatischen Spracherkennung ist es darüber hinaus wichtig zu wissen, worin die Unterschiede und Gemeinsamkeiten von Sprachen liegen. Daher wird im folgenden versucht, Sprachen zu kategorisieren. Die Kategorisierung ermöglicht das Herausarbeiten von Merkmalen, die zwischen Sprachen so ähnlich sind, daß sie in eine sprachenunabhängige Spracherkennungskomponente, wie beispielsweise in ein multilinguales akustisches Modell, zusammengefaßt werden könnten.

Zum Zweck der Kategorisierung von Sprachen wurden in der modernen Linguistik strukturelle Kriterien vorgeschlagen [Cry95]. Beeinflußt von den radikalen Arbeiten

Noam Chomskys, geht man dabei von der Annahme aus, daß sich Sprache durch eine Menge von *linguistischen Regeln* definieren läßt [SW79] und daß diese Regeln geeignete Kriterien zur Kategorisierung von Sprachen sind. Häufig werden die Kriterien in Strukturmodellen dargestellt. Auf der Basis eines solchen Strukturmodells lassen sich Unterschiede und Gemeinsamkeiten von Sprachen auf einer beliebigen Ebene bzw. Ausdrucksform ermitteln. Die Abbildung 2.4 zeigt das Strukturmodell für Sprache, welches der vorliegenden Arbeit zugrundeliegt [Cry95].

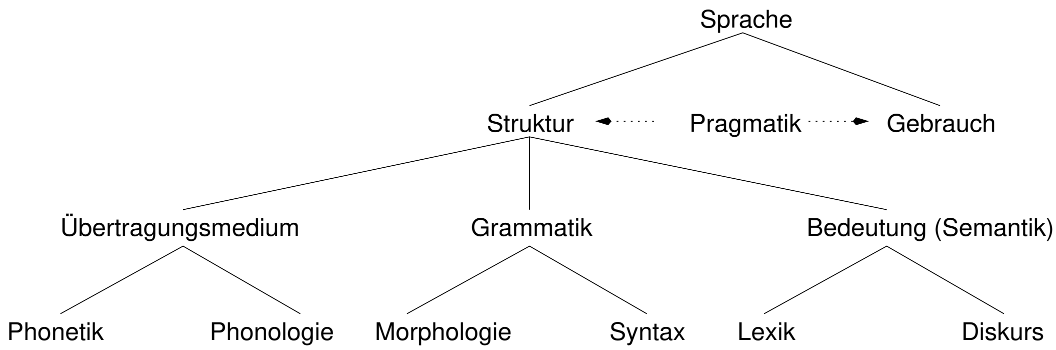


Abbildung 2.4: Sechsschichtiges Strukturmodell für Sprache (nach [Cry95])

Die Analyse des komplexen Phänomens Sprache wird dazu jeweils auf eine Ebene beschränkt, während gewisse Grundannahmen über andere Ebenen bestehen. Beispielsweise hängt die Auswahl und Beschreibung relevanter Laute einer Sprache davon ab, wie sich Wörter aufgrund von Lauten unterscheiden. Hierzu ist Wissen über das Vokabular einer Sprache notwendig, das man zur Übermittlung von Bedeutungsunterschieden benötigt. Der folgende Abschnitt ist zur Beschreibung der Unterschiede zwischen Sprachen an diesem Strukturmodell orientiert.

2.4 Unterschiede zwischen Sprachen

Sprache hat im wesentlichen zwei Ausdrucksformen: Schreiben und Sprechen. Sie werden als unterschiedliche aber gleichberechtigte Formen betrachtet: Sprechen läßt sich durch das Schreiben nicht ersetzen und umgekehrt. Beiden Formen ist jeweils ein eigenständiges Fachgebiet gewidmet und die Forschung beschäftigt sich mit dem Wesen und Ausmaß der zwischen ihnen bestehenden Unterschiede. Im Rahmen eines Diktiererkennungssystems sind beide Formen wichtig. Die gesprochene Sprache ist das Medium, mit dem die Eingabe in den Computer übertragen wird. Die geschriebene Darstellung ist die allgemein übliche Form, in der ein Erkennungssystem das Ergebnis abspeichert. Die Ausgabe geschieht dann je nach Anwendung entweder durch synthetisch generierte Sprache, durch andere sichtbare oder hörbare Aktionen oder wie bei der Diktieranwendung durch die textuelle Darstellung des Gesprochenen. Für die multilinguale Diktieranwendung, die uns im Rahmen dieser Arbeit besonders interessiert, ist dabei die orthographisch korrekte textuelle Darstellung in der landesüblichen Schrift und Schreibrichtung erwünscht. Im folgenden wird daher bei der Diskussion der Gemeinsamkeiten und Unterschiede zwischen Sprachen auf beide Ausdrucksformen Schreiben und Sprechen eingegangen.

2.4.1 Gesprochene Sprache

Auf der Basis des Strukturmodells in Abbildung 2.4 werden die wichtigsten Begriffe und strukturellen Merkmale von gesprochener Sprache beschrieben, anhand derer Sprachen unterschieden werden können. Diese Beschreibung führt von der kleinen zur großen sprachlichen Einheit: *Phonem - Morphem - Wort - Phrase - Teilsatz - Satz.*

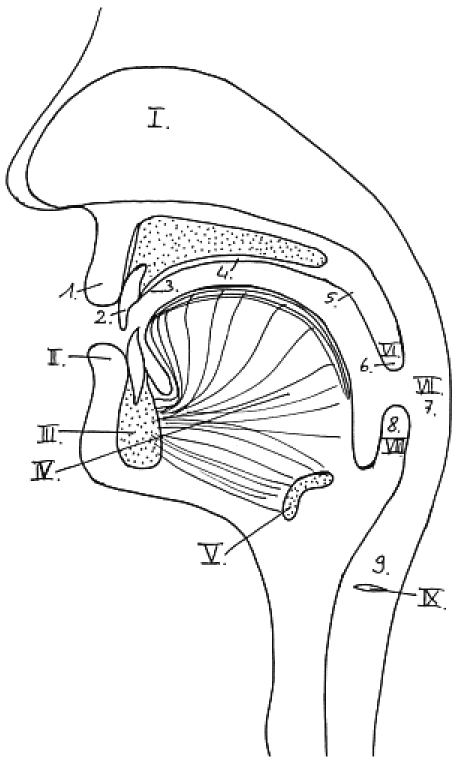
2.4.1.1 Phonetik und Phonologie

Die *Phonetik* erforscht die Gesamtheit der konkreten artikulatorischen, akustischen und auditiven Eigenschaften der möglichen Laute aller Sprachen [Buß90]. Die Laute als das natürliche Medium gesprochener Sprache lassen sich nach ihrer Artikulation im Vokaltrakt, nach ihrer akustischen Übermittlung (Formanten, Frequenz und Intensität) oder nach ihrer auditiven Rezeption klassifizieren. Am gebräuchlichsten sind artikulatorische Beschreibungen, da mit dem Vokaltrakt ein zugänglicher und gut erforschter Bezugspunkt zur Verfügung steht.

Sprachproduktion
Die prinzipielle Produktion von Sprachlauten ist in allen Sprachen gleich, denn sie ist durch die anatomischen Voraussetzungen des Menschen und durch physikalische Gesetze bestimmt. Allerdings gibt es gewisse Mechanismen, die in einer Sprache wichtig sind, während sie in einer anderen Sprache überhaupt nicht in Erscheinung treten. In dem folgenden kurzen Exkurs über Sprachproduktion sollen die wichtigsten Begriffe angeführt werden, die zur Beschreibung von lautlichen Unterschieden zwischen Sprachen im Verlauf der Arbeit immer wieder aufgegriffen werden.

Die Sprachproduktion ist eine Kombination aus einer durch einen Luftstrom verursachten Anregung an der Stimmritze (Glottis) und der Resonanzbildung im Vokaltrakt. Der Vokaltrakt, zu dem Mund-, Rachen- und Nasenraum gehören, wird am einen Ende durch die Stimmbänder und am anderen Ende durch die Lippen und die Nase begrenzt. Bei der Sprachproduktion wird durch Muskelkraft, die den Lungenraum zusammendrückt, ein Luftstrom erzeugt, der über die Luftröhre durch die Stimmritze in den Vokaltrakt gelangt. Ist die Stimmritze verengt, kommt es bedingt durch den Bernoulli-Effekt zu einem annähernd periodischen Öffnen und Schließen der Stimmbänder und damit zu quasiperiodischen Schwingungen[3], dem

[3] Die Zeitdauer zwischen zwei aufeinanderfolgenden Stimmritzenverschlüssen bezeichnet man als Grundperiodendauer T_0 und definiert darüber die Grundfrequenz $F_0 = \frac{1}{T_0}$



Artikulationsorgane sind mit römischen Ziffern gekennzeichnet: I. Nasenhöhle, II. Unterlippe, III. Unterkiefer, IV. Zunge, V. Zungenbein, VI. Zäpfchen, VII. Rachenraum, VIII. Kehldeckel und IX. Stimmritze;

Artikulationsstellen sind mit arabischen Ziffern gekennzeichnet: 1. Lippen (labies), 2. Schneidezähne (dentes), 3. Zahndamm (Alveolardamm), 4. harter, vorderer Gaumen (Palatum), 5. weicher, hinterer Gaumen (Velum), 6. Zäpfchen (Uvula), 7. Rachenwand (Pharyngis), 8. Kehldeckel (Epiglottis), 9. Stimmritze (Glottis).

Abbildung 2.5: Sagittalschnitt des menschlichen Kopfes [Ell97]

stimmhaften Schall. Über die Muskelspannung der Stimmbänder kann die Frequenz der Schwingungen in einem Rahmen von einigen Hertz bis zu mehreren kHz gesteuert werden. Steht die Stimmritze dagegen so weit offen, daß der Bernoulli-Effekt unwirksam bleibt, bewirkt die vorbeiströmende Luft Verwirbelungen, d.h. aperiodisches Rauschen oder stimmlosen Schall.

Die erzeugte stimmhafte oder stimmlose Anregung wird beim Durchlaufen des Vokaltraktes durch die Bewegungen der Artikulationsorgane zu einer Vielzahl unterschiedlicher Laute geformt. Abbildung 2.5 zeigt im Sagittalschnitt des menschlichen Kopfes die Artikulationsorgane und die Artikulationsstellen, die an der Sprachlautmodulation beteiligt sind. Je nach Art der Artikulation und dabei auftretenden Schallenergiemustern unterscheidet man in der Phonetik zwischen *Vokalen* und *Konsonanten.*

Vokale

Vokale werden phonetisch als Laute definiert, bei denen die Luft weitgehend ungehindert durch Mund oder Nase ausströmt. Vokale werden anhand von drei Kriterien beschrieben: der *Zungenposition*, der *Mundöffnung* bzw. Zungenhöhe und der *Lippenrundung.* Man differnziert nach dem Abschnitt des angehobenen Teils der Zunge (vorderer, mittlerer oder hinterer Abschnitt), dem Grad der Zungenhebung in Richtung Gaumen (hoch, mittel, tief) und der Art der Lippenöffnung (Rundung oder Spreizung). Die exakte Position von Zunge oder Gaumen läßt sich kaum messen, auch gibt es aufgrund sprecherspezifischer Mundproportionen keine absoluten Meßwerte. Daher gilt als grobe Regel: Je höher der höchste Punkt des Zungenrückens im Moment der Artikulation liegt, desto geschlossener ist der Vokal; je weiter vorne im Mund der höchste Punkt des Zungenrückens liegt, desto heller ist der Vokal (Beispiel: i ist ein geschlossener heller Vokal, a ein offener dunkler Vokal). Da Vokale von Artikulationsbewegungen an anderen Stellen im Vokaltrakt stark beeinflußt werden, unterscheidet man bei Vokalen zusätzlichen die sekundären Merkmale Nasalierung, Erweiterung des Rachenraumes und Rhotizisierung [Cry95].

In fast allen Sprachen gibt es Vokale, deren auditive Qualität während der gesamten Artikulationsphase konstant bleibt, sogenannte reine Vokale oder *Monophthonge*. Daneben gibt es Vokale, deren Beschaffenheit sich hörbar ändert, man bezeichnet sie als gleitende Vokale. Lassen sich auditiv zwei Phasen unterscheiden, spricht man von *Diphthongen*, bei drei Phasen von *Triphthongen* [Buß90].

Konsonanten

Konsonanten sind in der Phonetik als Laute definiert, die mittels eines Verschlusses oder einer starken Verengung des Vokaltraktes gebildet werden, so daß die Luft nur mit hörbarer Reibung entweichen kann. Konsonanten werden anhand von sechs Kriterien beschrieben: dem Luftstrommechanismus und der -richtung, dem Schwingungszustand der Stimmbänder, der Stellung des Gaumensegels sowie der Artikulationsart und dem -ort im Vokaltrakt.

Der Ursprung und die Richtung des *Luftstromes* bestimmen die Kategorie eines Lautes. Die meisten Konsonanten werden mit egressiver pulmonaler Luft (nach außen strömender Lungenluft) gebildet. Die *Artikulationsart* beschreibt die Weise, auf die der Luftstrom in Rachen- oder Mundhöhle von den Artikulationsorganen gehemmt wird. Dies geschieht im wesentlichen entweder durch vollständigen Verschluß (Plosive=Verschlußlaute, Nasale, Affrikate), durch intermittierenden Verschluß (Vibranten=Schwinglaute, Flaps=geschlagene Laute), oder partiellen Verschluß (Laterale=Seitenlaute, Frikative=Reibelaute). Schwingen die *Stimmbänder* zur Lauterzeugung, spricht man von stimmhaften Konsonanten, ansonsten von stimmlosen. Wird ein Konsonant durch einen Glottisverschluß gebildet, bezeichnet man ihn als einen glottal. Wird das *Gaumensegel* (weicher Gaumen) beim Sprechvorgang abgesenkt, kann Luft durch die Nase ausströmen und erzeugt einen nasalen Konsonanten. Ist das Gaumensegel angehoben und dadurch der Nasenraum verschlossen, entsteht ein oraler Konsonsant, denn die Luft tritt nur durch den Mund aus. Der *Artikulationsort* bezeichnet die Stelle, an der die betreffenden Organe die Verschlußart bilden. Konsonanten werden häufig mit Hilfe zweier Artikulationsorte gebildet, dann unterscheidet man zwischen der primären Artikulation, die den Verschluß bildet, und der sekundären Artikulation, die dem Konsonanten vier verschiedene Färbungen geben kann: eine Labialisierung, Palatalisierung, Velarisierung oder Pharyngalisierung des Konsonanten. Man bezeichnet einen Konsonant als labialisiert, wenn die Lippen während der primären Artikulation gerundet sind; als palatalisiert, wenn die Zunge während der primären Artikulation in eine hohe vordere Stellung gebracht wird. Der palatalisierte Konsonant erhält eine charakteristische j-Färbung, typisch beispielsweise für slawischen Sprachen, in denen sie bedeutungsunterscheidend sind (Beispiel im Russischen *[bratj]=nehmen und [brat]=Bruder*). Wenn die Zunge während der primären Artikulation in eine hintere hohe Stellung gebracht wird, ist der Konsonant velarisiert, er erhält dadurch eine zusätzliche u-Färbung wie im Deutschen das Wort „Kunst“. Wenn der Rachen während der primären Artikulation verengt wird, ist der Konsonant pharyngalisiert, er bekommt eine charakteristische a-Färbung. Die arabische Sprache verfügt über eine Reihe velarisierter und über sehr viele pharyngale Laute.

IPA - das internationale phonetische Alphabet

Die Entwicklung eines gemeinsamen phonetischen Referenzschemas entsprang dem Wunsch vieler Phonetiker, die Ergebnisse ihrer Arbeiten untereinander vergleichen zu können. Im Jahr 1886 trug Otto Jespersen (1860-1943) die Idee eines phonetischen Alphabets vor und 1888 wurde die erste Version des Internationalen phonetischen Alphabets (IPA) veröffentlicht, das von der Phonetiklehrervereinigung API entwickelt worden war. Dieses IPA-Alphabet beruht auf dem Prinzip, daß die artikulierbaren Laute aller Weltsprachen durch das Alphabet beschrieben werden können, und daß jedes Symbol genau einen Laut definiert. IPA ist somit das Ergebnis einer Wiederentdeckung der ursprünglichen Idee des Alphabetes, nämlich die der eins-zu-eins Relation zwischen Graphemen und Phonemen [Van99]. IPA ist aber auch die internationale Bemühung darum, diese Idee gleichzeitig für alle Sprachen der Welt zu lösen. Es nährt damit die Hoffnung auf ein sprachenunabhängiges phonetisches Inventar. In den Augen einiger Forscher ist dies eine illusionäre Hoffnung, denn einerseits belegt jede Sprache nur einen kleinen Teil des artikulatorischen Raumes, was man daran erkennen könnte, daß das IPA Alphabet hunderte von Zeichen benötigt, um alle Sprachen abzudecken. Zweitens gäbe es keinen Experten, der alle Sprachen gleichermaßen beherrscht, so daß die beste Implementierung eines IPA-Alphabets immer aus einem Kompromiß vieler Experten bestehen müßte, die sich untereinander nicht völlig verstehen [Van99].

Selbst wenn das IPA-Alphabet nicht darauf ausgelegt ist, die Ähnlichkeiten zwischen Lauten *verschiedener* Sprachen zu charakterisieren, bietet es derzeit die beste Grundlage Lautsysteme zwischen Sprachen zu vergleichen. Das IPA-Alphabet hat sich in der Literatur durchgesetzt, was einmal darin begründet ist, daß mittlerweile viele Beschreibungen von Lautsystemen auf IPA oder davon abgeleiteten Alphabeten existieren, zweitens existiert derzeit keine bessere Alternative, welche die sprachenunabhängige Abbildung eines Lautes liefert.

Das IPA-Alphabet besteht soweit möglich aus Buchstaben des lateinischen Alphabets; nur wenn es unumgänglich war, wurden neue Symbole eingeführt. Einige dieser Symbole wurden mittlerweile sogar in Schriftsystemen übernommen, die man für schriftlose Sprachen entwickelt hat. Das IPA-Alphabet wurde mehrmals abgeändert und findet heute Verwendung in Wörter- und Lehrbüchern auf der ganzen Welt. In dieser Arbeit wird die 1993 überarbeitete Fassung von [IPA93] verwendet. In der Auslegung des 7-bit ASCII Codes (siehe Abschnitt 2.4.2.3) sind die IPA-Symbole nicht integriert. Daher wurden etliche Fonts und Abbildungstabellen entwickelt, die allerdings nicht standardmäßig installiert sind. Viele behalfen sich daher mit einer mnemonischen Codierung in 7-bit ASCII Zeichen. Das führte schnell zur Verwirrung und Unsicherheit, da eine solche Codierung stark von der muttersprachlichen Sichtweise geprägt ist. Es entstanden einige Arbeiten, die eine computertaugliche 7-bit Codierung entwickelt haben, wie beispielsweise PHONASCII von Allen [All88], SAMPA, das unter dem Esprit-Projekt SAM 1541 entwickelt wurde [SAM98] und WORLDBET Hieronymous [Hie93].

KONSONANTEN

	Bilabial	Labiodent.	Dental	Alveolar	Postalveo.	Retroflex	Palatal	Velar	Uvular	Pharyng.	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k g	q ɢ		ʔ
Nasale	m	ɱ		n		ɳ	ɲ	ŋ	ɴ		
Trill	ʙ			r					ʀ		
Taps oder Flaps				ɾ		ɽ					
Frikative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Laterale Frikative				ɬ ɮ							
Approximant		ʋ		ɹ		ɻ	j	ɰ			
Laterale Approximant				l		ɭ	ʎ	ʟ			
Affrikate											

Treten Symbole paarweise auf, repräsentiert das rechte einen stimmhaften Konsonanten.

Abbildung 2.6: Konsonantenschema (nach IPA - Stand 1993)

Abbildung 2.6 zeigt die IPA-Symbol Tabelle für die Konsonanten. Sie sind in einem Raster angeordnet, das durch die Artikulationsart und den -ort des Lautes aufgespannt wird. Stimmhafte und stimmlose Konsonanten, die in gleicher Art und an demselben Ort entstehen, werden in einer gemeinsamen Zelle angeordnet, der stimmhafte Konsonant rechts, der stimmlose links. Zellen der Tabelle, die in Bereichen liegen, in denen keine Artikulation möglich ist, sind grau schattiert.

Die Vokale sind, wie in Abbildung 2.7 zu sehen, nach drei Gesichtpunkten angeordnet: vertikale und horizontale Zungenposition sowie der Rundung. Es gibt vier Stufen der vertikalen Zungenposition (Geschlossenheit) eines Vokals, drei Stufen der horizontalen Zungenposition (Helligkeit) und zwei Rundungsausprägungen. Die Stufen der Geschlossenheit sind in Zeilen beginnend mit der Kategorie „Geschlossen“ angeordnet, die Helligkeit in Spalten beginnend mit der Kategorie „Vorne/Hell“. Dadurch ergibt sich eine vierstufige, dreiklassige Trapezform, bei der die runden

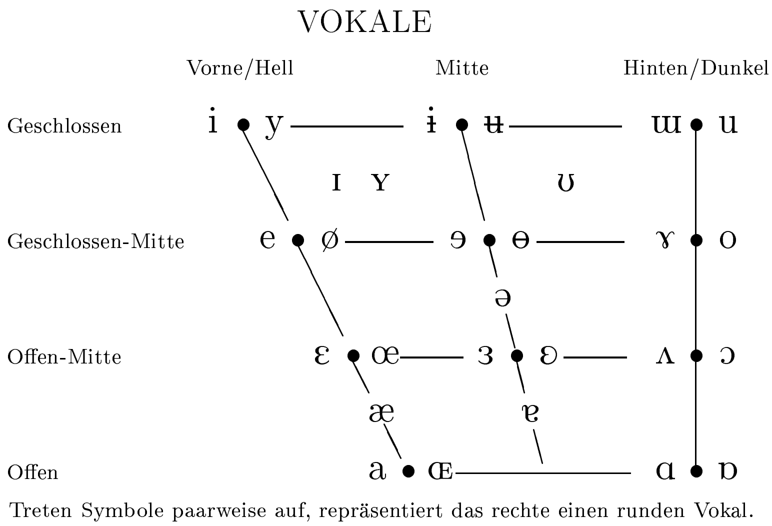


Abbildung 2.7: Vokalschema (nach IPA - Stand 1993)

Vokale rechts von ihrem unrunden Gegenstück stehen. Diese Anordnung geht auf das System der *Kardinalvokale* zurück, welches von dem Phonetiker Daniel Jones (1956) vorgeschlagen wurde. Bei den Kardinalvokalen handelt es sich um ein sprachenunabhängiges Bezugssystem abstrakter Normalvokale die zur Standardisierung der phonetischen Beschreibung von Vokalen herangezogen wurden [Buß90].

Phonologie

Während die in der Phonetik gezogenen Schlüsse über Laute unabhängig von der gesprochenen Sprache immer die gleichen sind, versucht die *Phonologie* die vielen Lautnuancen auf eine endliche Zahl abstrakter lautlicher Einheiten zu reduzieren und in einem sprachenspezifischen System anzuordnen. Für die Phonologie spielt es daher sehr wohl eine Rolle, in welcher Sprache die Laute erzeugt wurden. Insbesondere geht es in der Phonologie um die Aufdeckung der Prinzipien, welche die Verteilung von Lauten in den verschiedenen Sprachen bestimmen und um die Erklärung der auftretenden Unterschiede. Häufig ermittelt man zunächst die phonologische Struktur einer einzelnen Sprache, also die in ihr vorkommenden lautlichen Einheiten und deren Anordnung. Dabei besteht das Ziel darin, diese Einheiten möglichst klein zu wählen, man sucht gewissermaßen nach den Grundbausteinen der lautlichen Substanz einer Sprache. Davon ausgehend werden die Eigenschaften verschiedener Lautsysteme miteinander verglichen und Hypothesen über die Regeln aufgestellt, die der Lautverwendung in bestimmten Sprachgruppen zugrundeliegen (phonologische Universalien) [Cry95].

Phonologische Analysen beruhen auf dem Prinzip, daß bestimmte Laute Bedeutungsveränderungen in einem Wort oder Satz bewirken, andere dagegen nicht. Diese bedeutungsunterscheidende Kraft wird an sogenannten *Minimalpaaren* nachgewiesen. Ein Minimalpaar besteht aus zwei Wörtern verschiedener Bedeutungen, die sich an einer Position durch einen Laut unterscheiden. Verändert sich beim Austausch der Laute die Bedeutung, hat man das kleinste bedeutungsunterscheidende Element gefunden, welches man als *Phonem* bezeichnet (Beispiel Fisch - Tisch).

Phoneme werden durch dieselben phonetischen Symbole dargestellt, allerdings nicht in eckige Klammern [*phon*], sondern in Schrägstriche /*phonem*/ gesetzt, um anzudeuten, daß sie als ein Teil der Sprache gesehen werden, nicht als rein physikalische Laute. Bei dieser phonologischen Bestimmung des Phoneminventars einer Sprache trifft man auf Laute, die keine Bedeutungsunterschiede bewirken, wenn man sie gegeneinander austauscht, sogenannte *Allophone*. So ist es beispielsweise nicht bedeutungsverändernd, wenn man den Laut [*x*] in „Dach" durch den Laut [ç] in „Dich" ersetzt, es entsteht lediglich eine unnatürlich klingende Aussprache. Es handelt sich dabei um zwei verschiedene Varianten eines Phonems, welche in unterschiedlichen stellungsbedingten lautlichen Umgebungen auftreten, sogenannte *stellungsbedingte Varianten*. Es gibt auch *freie Varianten*, die in derselben lautlichen Umgebung auftreten und miteinander vertauscht werden, ohne einen Bedeutungsunterschied hervorzurufen, wie beispielsweise das gerollte und das geschlagene /r/.

Bei der Erforschung einer Sprache ist stets zu bestimmen, welche Laute als Phoneme zählen und welche als Allophone. Dabei muß man auch berücksichtigen, daß Laute in verschiedenen Sprachen unterschiedlich eingesetzt werden können. So kann ein Laut in der einen Sprache als Allophon vorkommen, während er in einer anderen Sprache als selbständiges Phonem fungiert. Beispielsweise handelt es sich bei der retroflexen und palatalen Variante des /*l*/ im Englische „leaf" und „pool" um Allophone desselben Phonems, aber im Russischen um zwei verschiedene Phoneme.

Bei der *phonologischen* Definition (auch linguistische Definition) unterscheidet man Vokale von Konsonanten durch ihre verschiedenartige Verwendung in den Strukturen gesprochener Sprachen. Konsonanten sind Einheiten (C), die am Rande einer Lautfolge auftreten. Vokale (V) dagegen erscheinen im charakteristischen Fall im Zentrum einer Silbe, können aber auch alleine auftreten. In den meisten Fällen entspricht die phonetische Definition eines Lautes im Ergebnis der phonologischen. Es gibt aber auch Fälle, in denen dies nicht so ist, beispielsweise im Englischen bei den Lauten [l], [ɹ], [w] und [j] in Wörtern wie „wet" oder „you". Phonologisch betrachtet müßten sie als Konsonanten gelten, weil sie am Silbenrand auftreten, phonetisch werden sie aber ohne hörbare Reibung artikuliert. [w],[j] ähneln in ihren Schallbildern den Vokalen u, i und werden daher häufig Halbkonsonanten, geräuschlose Dauerlaute (Approximanten) oder Halbvokale genannt. Es wurden viele neue Begriffe vorgeschlagen, um die phonetische Unterscheidung gegen die phonologische Definition abzugrenzen. Da aber das Problem nur bei wenigen Lauten in Erscheinung tritt, haben sich solche Vorschläge nicht durchgesetzt [Cry95].

Koartikulation

Im Unterschied zur orthographischen Wiedergabe durch einzelne Buchstaben sind die den Sprachlauten entsprechenden Lautereignisse keine diskreten Einheiten. Vielmehr ist die Sprachproduktion ein kontinuierlicher Bewegungsablauf artikulatorischer Vorgänge ohne natürliche Einschnitte, wobei sich unsere Artikulationsorgane kontinuierlich über die Zeit der angestrebten Position im Artikulationsapparat annähern. Diese antizipierenden Bewegungsabläufe bei der Artikulation bezeichnet man als *Koartikulation* [Buß90]. Gleicht sich ein Laut dabei dem nachfolgenden Laut (seinem Ziel) an, so handelt es sich phonetisch gesprochen um *regressive* Koartikulation, weist ein Laut dagegen den Einfluß des vorhergehend Lautes auf, spricht man von *progressiver* Koartikulation [Cry95].

Die Koartikulation bezieht sich auf die Sprachproduktion an sich und ist daher unabhängig von der betrachteten Sprache. Phonologisch gesehen kann sich allerdings die Aussprache von Lauten, einzelnen Segmenten oder ganzen Wörtern durch Koartikulationseffekte von Sprache zu Sprache unterscheiden. Dann, wenn Wörter in kontinuierlich gesprochener Sprache miteinander verknüpft werden, können Geschwindigkeit oder Rhythmus bewirken, daß einzelne Segmente sprachenabhängig schwächer artikuliert werden, ganz wegfallen oder eingefügt werden. Beispielsweise fallen im Deutschen häufig alveolare Konsonanten weg, etwa das /t/ als mittlerer Konsonant einer Dreiergruppe wie in „erhältst“. Ein weiteres Beispiel ist der Reduktionsvokal /ə/, der vor Nasalen verschwindet [Koh77]. Man nennt diesen Vorgang, bei dem beim schnellen Sprechen Laute ausgelassen werden, *Elision*. Im Französischen wird häufig ein Laut zwischen zwei Wörter eingefügt, was man als *Liaison* bezeichnet, wie in [lez ami] für „les amis“. In vielen Sprachen gibt es das Phänomen der *Assimilation*, bei der nebeneinanderstehende Laute einander angeglichen werden, wie im deutschen Beispiel [z ɛ ɱ f] für „Senf“.

Für die Leistungsfähigkeit eines Spracherkennungssystems spielt die geeignete Modellierung der Koartikulation eine sehr wichtige Rolle. In fast allen heute gängigen Erkennern wird kontinuierliche Sprache durch eine Abfolge diskreter Phoneme beschrieben (siehe Kapitel 3). Dabei werden Phoneme in Subphoneme zu einem Beginn-, einem Mittel- und einem Endzustand zerlegt, um die dynamischen Phasen eines Phonems zu modellieren. Diese Aufteilung ist phonetisch begründet und sie unterscheidet sich nicht für verschiedene Sprachen. Außerdem werden zur Modellierung eines Phonems dessen benachbarte Phoneme mitbetrachtet. Als einer der ersten führte [Lee88] dazu sogenannte kontextsensitive (oder kontextabhängige) Triphonemmodelle ein, bei deren Modellierung jeweils der linke und rechte Phonemnachbar einbezogen wird. Auf die Arbeiten von [HH91] geht die Idee zurück, kontextabhängige Modelle auf der Ebene von Subphonemen (siehe Abschnitt 3.2.2.2) zu beschreiben. Darüber hinaus wurde die Modellierung von Nachbarschaften über die Wortgrenzen hinaus eingeführt. Heutzutage werden Kontexte in einer theoretisch beliebigen Fensterbreite betrachtet, üblich sind Breiten bis zu 5 Phonemen.

Die Modellierung von Subphonemkontexten ist stark von der betrachteten Sprache abhängig. Phonemnachbarschaften, die in einer Sprache sehr häufig auftreten, können in einer anderen Sprache völlig ohne Belang sein. Darüber hinaus kommt es bei Modellierungen über Wortgrenzen hinweg zu den oben genannten sprachentypischen Elisions-, Liaisons-, oder Assimilationserscheinungen. Für einen multilingualen Erkenner sind diese Unterschiede zu berücksichtigen.

2.4.1.2 Prosodie

Es gibt eine Reihe phonologischer Merkmale, die nicht das einzelne Lautsegment betreffen, sondern weitaus größere Einheiten überspannen, wie Silben, Wörter, Phrasen oder ganze Sätze. Diese Merkmale werden unter dem Begriff Prosodie zusammengefaßt. Im einzelnen sind dies die Basismerkmale Tonhöhe (deren akustisches Korrelat die Grundfrequenz ist), die Lautheit (Energie des Signals), das Sprechtempo (mittlere Phonemdauer) und Merkmale, die eine Veränderung einer oder mehrerer Basismerkmale über die Zeit beschreiben, wie die Intonation, die Akzentuierung, und die zeitliche Strukturierung einer Äußerung durch Gliederung in Phrasen, Rhythmus und Pausen (in Anlehnung an [NBK+97]).

Im Kontext der Prosodie bezeichnet die *Intonation* eine Veränderung der Tonhöhe, die *Betonung* eine Veränderung der Lautheit und die *Akzentuierung* die Veränderung von Tonhöhe in Kombination mit der Lautheit.

Prosodische Merkmale können auf jeder Ebene der automatischen Sprachverarbeitung zur Analyse beitragen. Es werden im folgenden allerdings nur diejenigen Merkmale beschrieben, die im Kontext der Spracherkennung wichtig sind. Die Rolle der Prosodie in der semantischen und pragmatischen Interpretation einer Äußerung wird hier nicht betrachtet. Eine ausführliche Beschreibung darüber findet sich beispielsweise in [Nö91], der konkrete Einsatz und Nutzen prosodischer Information im Sprachübersetzungsprojekt Verbmobil wird in [NBK+97] erläutert.

Auf Wortebene kann das Wissen darüber hilfreich sein, welche Silbe eines Wortes durch Betonung oder Akzentuierung hervorgehoben wird. Der Nutzen hängt allerdings davon ab, ob und wozu der Akzent und die Betonung in einer Sprache verwendet werden.

In sogenannten *Akzentsprachen* werden bestimmte Silben eines Wortes mit einer festgelegten Akzentuierung gesprochen. Beispiele sind Schwedisch, Japanisch und Serbokroatisch. Man unterscheidet Akzentsprachen von solchen Sprachen, in denen die Hervorhebung durch Lautheitsveränderungen geschehen, während die Tonhöhe keinen Einschränkungen unterliegt. In solchen Fällen ist Akzent synonym mit *Betonung*, daher nennt man diese Sprachen *Betonungssprachen*.

In Sprachen mit *festen Betonungsmustern* erscheint die Betonung stets an derselben Position innerhalb eines Wortes. Dies gilt beispielsweise für Tschechisch und Finnisch, in denen die Betonung immer auf der ersten Silbe eines Wortes liegt, im Türkischen auf der letzten und im Polnischen auf der vorletzten Silbe. Dagegen gibt es Sprachen mit *lexikalischen Betonungsmustern*, in denen die Position der zu betonenden Silbe für jedes Wort gesondert festgelegt ist.

In Sprachen, in denen die Muster der lexikalischen Betonung fest vorgegeben sind, wie beispielsweise Deutsch, kann das prosodische Merkmal Akzentuierung oder Betonung die Erkennung von Wörtern unterstützen. In Sprachen, in denen die Position einer Betonung bedeutungsunterscheidend ist (Beispiel Englisch 'permit gegenüber per'mit), kann die Prosodie zur Wortdisambiguierung herangezogen werden (vgl. [Kom96]).

Tonsprachen
Die Intonation ist in vielen Sprachen der Welt nicht nur auf Wort- oder Satzebene bedeutungsunterscheidend, sondern auch auf Phonemebene. Diese Sprachen nennt man *Tonsprachen*. Die distinktiven Tonhöhen heißen Töne oder *Toneme*. Einfache Tonsysteme unterscheiden zwei Toneme (hohe vs. niedrige Tonhöhe), kompliziertere Systeme wie Mandarin Chinesisch haben vier Töne (steigend, fallend, mittel, fallend-steigend), in Thai gibt es fünf, Kanton-Chinesisch hat sogar sechs. Die Intonation kann entweder die Bedeutung eines Wortes oder dessen grammatische Funktion verändern. Im Mandarin-Chinesisch wird durch die Intonation die lexikalische Bedeutung unterschieden (je nach Tonhöhenverlauf des „a“ hat die Einheit „ma“ vier verschiedene Bedeutungen). Im Gegensatz dazu wird in der ostafrikanischen Sprache Twi durch Intonation die Zeitform eines Verbes verändert (ein tiefer Ton signalisiert Präsenz, ein hoher oder hoch-tiefer dagegen Präteritum). Weiterhin gliedert man Tonsprachen in Kontur-Tonsprachen und Register-Tonsprachen. In der ersten Variante kann die Höhe eines Tons wechseln (gleitende Töne) wie es in Thai und im Mandarin Chinesisch der Fall ist, in der zweiten Variante bleibt die Tonhöhe konstant, wie beispielsweise in den Sprachen Zulu und Haussa.

Die Tonhöhe hat einen Einfluß auf die Formanten, die relativen Maxima im Spektralbereich, die charakteristisch für einen Laut sind. Dieses Wissen kann in der Merkmalsextraktion (siehe Abschnitt 3.2.1) eingesetzt werden. Die konkrete Anwendung wird am Beispiel des Mandarin Chinesisch in Kapitel 5 beschrieben.

2.4.1.3 Morphologie und Syntax

Die Morphologie und Syntax bilden als Unterstrukturen der Grammatik das Skelett oder Gerüst einer Sprache. Die *Syntax* bestimmt, wie Wörter angeordnet werden können, um Bedeutungsbeziehungen innerhalb und manchmal zwischen Sätzen aufzuzeigen. Die meisten syntaktischen Studien konzentrieren sich auf die Satzstruktur, da dort die wichtigsten grammatikalischen Beziehungen zum Ausdruck kommen. Die syntaktische Struktur ist für Ethymologen ein Stützpfeiler bei der Typologisierung von Sprachen.

Die *Morphologie* befaßt sich mit der Struktur von Wörtern. Die kleinsten bedeutungstragenden (syntaktische) Einheiten, in die Wörter zerlegt werden können, nennt man *Morpheme*. Morpheme sind weder mit Silben noch mit Wörtern identisch: Ein Morphem kann ein einsilbiges Wort sein, wie „Freund“ oder ein mehrsilbiges Wort sein, wie in „a-ber“; Es können aber auch mehrere Morpheme ein Wort bilden, Beispiel: Unfreundlichkeiten → Un-freund-lich-keit-en

Morphem	Funktion
Un-	Präfix, das eine Verneinung anzeigt
freund-	trägt eine Bedeutung
lich-	zeigt eine Eigenschaft an
keit-	zeigt einen Zustand oder eine Eigenschaft an
en	Suffix, das den Plural anzeigt

Abbildung 2.8: Beispiel einer morphologischen Zerlegung

wie in „freund-lich“, ebenso wie mehrere Morpheme eine Silbe bilden können, wie in „lern-t“.

Haupttypen des Sprachbaus (nach [Stö97])

Die Art und Weise, wie Morphologie und Syntax in der Sprache zusammenwirken, ist ein wichtiger Gesichtspunkt bei der Definition geeigneter Einheiten für die Spracherkennung. Unmittelbar davon betroffen sind das Wachstum des Vokabulars und die Entwicklung von statistischen Sprachmodellen für ein Erkennungssystem. Es sollen daher im folgenden die wichtigsten Begriffe und die unterschiedlichen Typen des Sprachbaus beschrieben werden.

Flexion oder Beugung nennt man die Erscheinung, bei der Wörter einer Sprache in verschiedenen Formen auftreten. Sie dient dazu, den Bedeutungsgehalt eines Wortes unter Beibehaltung der Grundbedeutung abzuwandeln und dessen Beziehungen zu anderen Wörtern innerhalb eines Satzgefüges anzuzeigen. Flektierende Sprachen sind solche, die sich durchgehend oder überwiegend der Flexion zur Bildung von Wortformen bedienen. Beispiele dafür sind das Sanskrit, das Altgriechische, das Latein und in abgeschwächter Form auch das Deutsche. Die Flexion kommt in drei Formen vor: der *Deklination* (Beugung von Substantiven, Adjektiven, Artikeln, Pronomen und Numeralen), der *Konjugation* (Beugung von Verben) sowie der *Komparation* (Steigerung des Adjektivs).

Von *Agglutination* oder Anleimung spricht man, wenn grammatikalische Beziehungen durch Suffixe ausgedrückt werden, die man an den Wortstamm anhängt. Klassische Beispiele sind die Sprachen Finnisch, Ungarisch und Türkisch. In diesen Sprachen tragen alle Suffixe jeweils genau eine Bedeutung und werden in fester Reihenfolge hintereinander an den Wortstamm angefügt. Für jede weitere Bedeutung wird ein weiteres Suffix angehängt wie im Beispiel für die ungarische Sprache „Schiff“ hajo, „Schiffe“ hajo-k, „im Schiff“ hajo-ban, „in den Schiffen“ hajo-k-ban usw. [Bod97].

In diesem Fall spricht man von *Monosemie*. Wenn dagegen ein Suffix mehrere Bedeutungen zugleich trägt, spricht man von *Polysemie*. Als agglutinierende Sprachen bezeichnet man solche, die sich in überwiegender Weise des Mittels der Agglutination bedienen, insbesondere da, wo andere Sprachen das Instrument der Flexion einsetzen.

Isolation nennt man das Prinzip, bei dem Wörter weder durch Flexion abgewandelt noch durch Agglutination verlängert werden. Vielmehr bestehen die Sätze aus einer Folge von unveränderlichen Wörtern, bzw. Elementen. Die Beziehungen untereinander werden durch die Stellung im Satz angezeigt. Wo dies nicht ausreicht, werden grammatische Hilfswörter verwendet (würde man dieses Prinzip ins Deutsche übertragen hieße es beispielsweise ich gehen gestern anstatt ich ging). Beispiele für isolierende Sprache sind die klassische chinesische Sprache und in Ansätzen die englische Sprache.

Bei der *Polysynthese* werden Einheiten durch An- oder Einfügen erweitert. Mit einem Wortstamm werden soviele lexikalische und/oder grammatische Elemente verbunden, daß vieles, was im Deutschen durch einen ganzen Satz ausgedrückt werden muß, dort in ein einziges Wort zusammengepreßt wird (ins Deutsche übertragen hieß es beispielsweise Tag-und-Nacht-Schreibtisch-Hocker). Man findet das Prinzip der Polysynthese vorwiegend bei den Eskimo- und Indianersprachen.

In der Reihe isolierend - polysynthetisch - agglutinierend - flektierend spiegelt sich der Übergang vom *analytischen* zum *synthetischen* Satzbau wider. Analytische Sprachen halten ihre Elemente und vor allem ihre Wurzeln sauber auseinander, im synthetischen Satzbau kommt es zu Formveränderungen. In den meisten Sprachen kommen analytischer und synthetischer Satzbau nebeneinander vor, wie im Deutschen bei der Konjugation mit Hilfsverben (sie tragen - sie werden tragen) neben der Konjugation durch Formveränderung (sie tragen - sie trugen).

Die Morphologie einer Sprache spielt eine große Rolle bei der Entwicklung eines Spracherkennungssystems. Denn aufgrund von Speicher- und Laufzeitlimitierungen sind heutige Spracherkenner in ihrem Vokabularumfang beschränkt. Bei agglutinierenden und stark flektierenden Sprachen hat man aber ein sehr starkes Vokabularwachstum. Die Beschränkung des Erkennervokabulars erzeugt daher in solchen Sprachen eine hohe Quote nicht abgedeckter Wörter, die sogenannte *Out-Of-Vocabulary (OOV)* Rate. Daher ist man bei agglutinierenden und stark flektierenden Sprachen gezwungen, die natürlichen Worteinheiten in kleinere morphembasierte Einheiten zu zerlegen. Dies reduziert die Reichweiten des Sprachmodells (siehe Abschnitt 3.2.3), was man nur teilweise durch eine Erhöhung der Sprachmodellhistorie auffangen kann. Kurze Worteinheiten mit kleinen OOV-Raten wirken sich nicht notwendigerweise positiv auf die Erkennungsresultate aus, da kürzere Worteinheiten eine höhere akustische Verwechselbarkeit haben. Bei Sprachen mit agglutinierenden Sprachbau ist man daher zu besonderen Maßnahmen gezwungen, die in Kapitel 5 erörtert werden.

2.4.1.4 Lexik und Diskurs

Die Semantik gliedert sich in die Teilbereiche Lexik und Diskurs, ihre Grundeinheit bezeichnet man als Lexem oder *lexikalische Einheit*. Die vollständige Auflistung aller lexikalischen Einheiten einer Sprache nennt man Lexikon oder Wörterbuch. Es enthält für jede lexikalische Einheit einen Worteintrag, einen Hinweisen zu deren Schreibweise, zur Aussprache, zum grammatikalischen Rang sowie zur Bedeutung, Geschichte und Verwendung der lexikalischen Einheit in der betreffenden Sprache. Für die Rekonstruktion der Bedeutung einer Wortfolge ist ein solches Lexikon notwendig.

Da man sich in der automatischen Spracherkennung darauf beschränkt, zu einer gesprochen Äußerung diejenige Wortsequenz zu finden (und gegebenenfalls textuell darzustellen), die am wahrscheinlichsten ist, genügt es, die Aussprache und die Schreibweise einer lexikalischen Einheit zu kennen. Im Gegensatz zur Sprachübersetzung oder zum Sprachverstehen wird in der Spracherkennung nicht versucht, die Bedeutung oder Intention des Gesagten zu erfassen. Wenn im Rahmen der Spracherkennung daher von einem Wörterbuch oder Lexikon die Rede ist, dann versteht man darunter eine Auflistung aller als bekannt vorausgesetzten lexikalischen Einheiten einer Sprache, bestehend aus der Schreibweise und der Aussprache dieser Einheit.

2.4.2 Geschriebene Darstellung

Das Training automatischer Spracherkennungssystemen (siehe Kapitel 3) erfordert eine Verschriftung der Sprachdaten, die nach heutigem Stand der Technik manuell erstellt werden muß. Während man in den frühen Phasen der Spracherkennung eine phonemische oder phonetische Kennzeichnung aller gesprochenen Wörter vornahm, beschränkt man sich heutzutage aus Aufwandsgründen meist auf die orthographische Darstellung des Gesprochenen. Für die Verschriftung multilingualer Sprachdaten muß man sich daher mit den sprachenspezifischen Darstellungsformen auseinandersetzen.

Die geschriebene Darstellung von Sprache läßt sich aus zwei Blickwinkeln beschreiben, die im selben Verhältnis zueinander stehen wie Phonetik und Phonologie bei der gesprochenen Form. In der *Graphetik* geht es um die Erforschung der physischen Eigenschaften der Symbole, aus denen Schriftsysteme bestehen. Die *Graphemik* ist dagegen die Wissenschaft von den bedeutungsunterscheidenden Einheiten eines Schriftsystem. Ein Aspekt der Graphetik ist die Vielfalt der graphischen Verfahren, die Sprachen aufweisen. Am auffälligsten ist, daß sich Sprachen hinsichtlich ihrer Schreibrichtung voneinander unterscheiden: von links nach rechts wie im Deutschen, von rechts nach links wie im Arabischen, von oben nach unten wie im klassischen Japanisch oder auch von unten nach oben wie in gewissen Formen des klassischen Griechisch. Es existieren sogar Schriften wie das Bustrophedon bei der in wechselnden Richtungen geschrieben wird [Stö97].

Der Ausdruck *Graphemik* wurde in Analogie zur Phonemik (oder Phonologie) gebildet. Im Fachgebiet der Graphemik wurden viele Begriffe aus der gesprochenen Sprache auf die geschriebene Sprache übertragen. Parallel zum Phonem ist das *Graphem* entstanden, das die kleinste bedeutungsunterscheidende Einheit eines Schriftsystems bezeichnet.

2.4.2.1 Schriftsysteme

Man kann Schriftsysteme zwar auf der Grundlage graphetischer Faktoren wie Größe, Stil, Zeichenanordnung und Schreibrichtung beschreiben, doch erfährt man dabei nichts über die Grapheme des betreffenden Systems. Im Prinzip ließe sich sogar fast jedes Schriftsystem unter Anwendung beliebiger graphetischer Regeln schreiben. Daher teilt man Schrriften in solche Systeme ein, in denen eine klare Beziehung zwischen den Symbolen und den Lauten der entsprechenden Sprache zu erkennen ist - sogenannte *phonologische Systeme* - und solche, in denen es sich nicht so verhält - sogenannte *nicht-phonologische Systeme*.

Abbildung 2.9: Schriftsysteme verschiedener Sprachen (von links nach rechts): Arabisch (arabische Zeichen), Bulgarisch (kyrillische Zeichen), Katalanisch (lateinische Zeichen), Chinesisch (Hanzi), Kroatisch (lateinische Zeichen), Tschechisch (lateinische Zeichen), Englisch (lateinische Zeichen), Griechisch (griechische Zeichen), Hebräisch (hebräische Zeichen), Hindi (Devanagari-Schrift), Italienisch (lateinische Zeichen), Japanisch (Kanji), Koreanisch (Hangul), Rumänisch (lateinische Zeichen), Russisch (kyrillische Zeichen), Serbisch (kyrillische Zeichen), Thai (Thai-Schrift)

Nicht-phonologische Systeme sind eher in der Frühgeschichte der Schrift anzutreffen. Man unterscheidet hier piktographische und ideographische Schriften sowie die Keilschrift, ägyptische Hieroglyphen und die logographischen Schriften.

In der *piktographischen* Schrift geben die Grapheme die Welt in Form von Piktogrammen wieder (z.B. Wellenlinien symbolisieren das Wort Meer). Zur Entzifferung einer solchen Schrift genügt es, die Symbole zu kennen. Diese lassen sich dann unabhängig von der Sprache verbal auf unterschiedlichste Art beschreiben. Piktographische Schriften bilden die ältesten Schriftsysteme und sind auf der ganzen Welt anzutreffen. Fehlender Kontext oder mangelnde Kenntnisse einzelner Zeichen vervielfältigen die Deutungsmöglichkeiten. Daher warten noch heute viele piktographische Schriften auf ihre endgültige Entschlüsselung, wie beim wohl berühmtesten Beispiel, dem Diskus von Phaistos.

Die *ideographische* Schrift gilt als Weiterentwicklung der piktographischen. Ideogramme haben abstrakte oder konventionalisierte Bedeutungen und lassen keine phonologische Abbildung der Realität mehr erkennen. Die meisten Schriften sind jedoch eine Mischform aus piktographischen, ideographischen und sprachlichen Elementen, wie die sumerische, ägyptische und hethitische Schrift.

Bei *logographischen* Schriftsystemen stehen die Grapheme für ganze Wörter, ihre Symbole heißen Logogramme. Bekannte Beispiele sind die chinesische Schrift *Hanzi* und die, aus dieser abgeleiteten, japanische Schrift *Kanji*. Die Bezeichnung Logogramm für ein chinesisches Symbol ist etwas irreführend, denn die chinesische Schrift ist aus einer ideographischen Schrift mit piktographischen Elementen abgeleitet. Daher bezeichnet man ihre Symbole häufig als Ideogramme. Andererseits trifft die Bezeichnung Ideogramm nicht zu, da sich die Schriftzeichen auf sprachliche Einheiten und nicht unmittelbar auf Begriffe oder Dinge beziehen. Darüberhinaus stehen chinesische Zeichen oft für Morpheme, daher müßte die Schrift eher als morphographisch bezeichnet werden, was sich aber bislang noch nicht eingebürgert hat.

Zu den *phonologischen Schriften* zählt man die Silbenschriften und die Alphabetschriften. Bei ersteren entspricht jedes Graphem einer gesprochenen Silbe, in der Regel einem Konsonant-Vokal Paar. Neuere Beispiele sind das Amharische und die japanische *Kana* Schrift. Bei den *Alphabetschriften* hingegen besteht ein direkter Zusammenhang zwischen Graphemen und Phonemen, so daß es sich hier um die ökonomischste und anpassungsfähigste aller Schriftsysteme handelt. Anstatt mehrere tausend Logogramme oder viele dutzende Silben benötigt man hier nur eine relativ geringe Anzahl von Einheiten.

Das früheste bekannte Alphabet ist das nordsemitische, das sich um 1700 v. Chr. in Palästina und Syrien entwickelte. Auf diesem Modell beruht das arabische und das phönizische Alphabet, wobei letzteres wiederum als Muster für die Griechen diente. Das griechische Alphabet war schließlich Vorbild für das etruskische (800 v. Chr.), aus dem sich das frühlateinische und schließlich alle westlichen Alphabete entwickelten.

Die meisten Alphabete bestehen aus 20-30 Symbolen. Je nach Komplexität des Lautsystems sind aber auch kürzere und längere Alphabete möglich. In einem völlig regelmäßigen Alphabet entspricht jedes Graphem genau einem Laut. Darüberhinaus gibt es Alphabete, in denen nur bestimmte Phoneme graphemisch dargestellt werden, etwa die Konsonantenalphabete des Arabischen, in dem die Kennzeichnung der Vokale mittels diakritischer Zeichen freigestellt ist.

Die enge Beziehung zwischen Graphemen und Phonemen ist aus der Sicht der Spracherkennung sehr erwünscht, denn je enger diese Beziehung ist, umso einfacher läßt sich die Aussprache eines Wortes aus ihrer Schreibweise generieren (siehe Abschnitt 2.4.3).

Die meisten heute gebräuchlichen Alphabete erfüllen das Kriterium der Regelmäßigkeit leider nicht, entweder weil das Schriftsystem nicht mit der Ausspracheveränderung Schritt gehalten hat, oder die Sprache ein Alphabet benutzt, das ursprünglich nicht für sie entwickelt wurde. Manche Schriften sind so unpassend für die gesprochene Sprache, daß Schriftreformen eingeführt wurden. Eines der wohl bekanntesten Beispiele ist die von Atatürk 1928 durchgeführte türkische Schriftreform, in der das arabische Alphabet durch das lateinische ersetzt wurde. Das Weglassen der Vokale in der arabischen Schrift hatte im Türkischen zu zahlreichen Mehrdeutigkeiten geführt.

Das Beispiel der vietnamesischen Schriftreformierung, in der infolge der Einflüsse christlicher Missionare vom chinesischen zum lateinischen Schriftsystem übergewechselt wurde, verdeutlicht, daß die Wahl und Anwendung eines Schriftsystems im Grunde willkürlich ist und unabhängig von der Struktur einer Sprache ist. Die Schrift ist ein äußeres Merkmal der Sprache, und das Schriftsystem sagt nichts über den grammatischen Bau oder lautliche Besonderheiten einer Sprache aus [Haa91].

Es können auch innerhalb einer Sprache verschiedene Systeme nebeneinander stehen. Im Chinesischen verwendet man inzwischen neben der traditionellen *Hanzi* Schrift auch das lateinisierte Pinyin-Alphabet. Im Japanischen findet man im Alltag gleich vier verschiedene Schriftsysteme nebeneinander (fünf, sofern man arabische Ziffern als eigenständiges System betrachtet): das lateinische Alphabet z.B. bei Markennamen in der Reklame, die *Kanji* Schrift sowie die Silbenschriften *Hiragana*, zum Ausdruck grammatikalischer Unterschiede, und *Katagana* zur Schreibung von Lehnwörtern aus anderen Sprachen.

2.4.2.2 Segmentierung

In enger Beziehung mit dem Schriftsystem und dem Sprachbau steht die Segmentierung von Zeichenketten in natürliche Einheiten. Sprachen mit ideographischen Schriftsystemen haben häufig keine Segmentierung, wie beispielsweise Chinesisch und Japanisch. Sätze werden als Zeichenketten geschrieben, in denen alle Schriftzeichen ohne Begrenzung aneinandergefügt sind. Eine Zuordnung der Einzelzeichen zu lexikalischen Einheiten kann nur durch eine semantische Interpretation des Geschriebenen geleistet werden. Für die automatische Spracherkennung resultiert daraus die Frage, auf welchen Einheiten die Erkennung basieren soll. Aber auch für Sprachen, die eine natürliche Segmentierung haben, ist diese Frage nicht immer einfach zu beantworten. Die deutsche Sprache liefert beispielsweise neben natürlichen kurzen Einheiten auch sehr lange Phrasen, die durch die nahezu uneingeschränkte Möglichkeit der Kompositabildung entstehen. Die Gruppe der agglutinierenden Sprachen, zu denen beispielsweise Koreanisch und Türkisch zählen, bilden sehr lange Wortphrasen. Würde man Komposita und Wortphrasen als Basiseinheiten eines Spracherkenners verwenden, führte dies zu einem riesigen Vokabularwachstum und zu sehr hohen Out-Of-Vocabulary Raten. Solche Einheiten müssen daher geeignet zerlegt werden. Englisch oder Spanisch sind Beispiele für Sprachen, deren natürliche Worteinheiten meist unverändert als Basiseinheiten für einen Spracherkenner genutzt werden können.

2.4.2.3 Zeichenkodierung, -eingabe und -wiedergabe

In diesem Abschnitt wird die Computerunterstützung der verschiedenen Schriftsysteme beschrieben [Bec84]. Es werden kurz die Probleme erläutert die zur Bewältigung verschiedenersprachiger Texte gelöst werden müssen:

- Zeichenkodierung: Darstellung der Zeichen im Speicher des Computers
- Zeicheneingabe: Eingabe der Zeichen über die Tastatur
- Zeichenwiedergabe: Präsentation der Zeichen auf einem Ausgabemedium.

Zeichenkodierung
Die Notwendigkeit der Codierung von Zeichen ergibt sich aus dem Umstand, daß der Computer Informationen binär abspeichert. Jedem Schriftzeichen muß daher ein binärer Zahlencode zugewiesen werden; die Abbildungsvorschrift bezeichnet man als *Code* oder *Codierungstabelle*. Die Verbreitung von Computern nahm ihren Ursprung im englischsprachigen Raum, wo zur Darstellung der Schrift wenige Zeichen genügen. Die ersten standardisierten Codierungstabellen sehen daher vor, Zeichen mit der kleinsten üblichen Speichereinheit *1 Byte* zu kodieren. Damit lassen sich 256 verschiedene Zeichen darstellen. Der sogenannte ASCII-Code (American Standard Code for Information Interchange) legt nur die ersten 7-bit der Codierung fest. Darin befinden sich die Alphabetzeichen, Interpunktionszeichen, Zahlen, griechische Buchstaben für mathematische Ausdrücke und Linienzeichen.

Mit der zunehmenden Verbreitung des Computers nahm der Bedarf an sprachenspezifischen Sonderzeichen zu. Man erstellte Codierungstabellen für viele Sprachen und definierte diverse auf 8-bit erweiterte Zeichensätze, die die jeweiligen Sonderzeichen enthalten. Die in Tabelle 2.5 aufgelistete ISO Norm 8859-x definiert die Codierungen für gebräuchliche Alphabete. Die Codeseite für westeuropäische Sprachen wird auch häufig als ANSI (American National Standard Insitute) oder Latin-1 bezeichnet.

Die Codierung von Zeichensätzen mit 8-bit ist nur bei Alphabet- und Silbenschriften möglich, die nicht mehr als 256 Zeichen verwenden. Auf ideographische Schriften trifft diese Einschränkung in der Regel nicht zu. Diese Schriften verwenden häufig 10.000 Zeichen und mehr, wie etwa die chinesische Hanzi-Schrift (40.000 bis 60.000 Zeichen, 20.000 im häufigen Gebrauch), und die aus ihr hervorgegangene japanische Kanji-Schrift. Für die koreanische Hangul-Schrift benötigt man etwa 3500 Zeichen. Bei Schriften mit mehr als 256 Zeichen werden 2 Byte zur Codierung verwendet. Auch bei den 2-Byte Schriften gibt es zahlreiche verschiedene Codierungstabellen: für Chinesisch unterscheidet man Guobiao und Big5, für Japanisch gibt es die Shift-JIS Codierung (Japan Industry Standard) sowie die EUC-Codierung und für Koreanisch die Wansung und Johab Codierung KS C-5601.

Die Gründe für die vielen unterschiedlichen Codes sind historisch bedingt. Sie führen immer wieder zu Konvertierungsproblemen beim Austausch von Daten, gerade in

Zeichensatz	Sprachen
8859-1 (Latin 1)	Afrikaans, Katalanisch, Dänisch, Deutsch, Englisch, Faeroese, Finnisch, Französisch, Galizisch, Holländisch, Irisch, Isländisch, Italienisch, Norwegisch, Portugiesisch, Spanisch und Schwedisch
8859-2	Osteuropa, Balkanländer (u.a. Kroatisch)
8859-3	Südosteuropa
8859-4	Skandinavien (größtenteils abgedeckt durch 8859-1)
8859-5	Kyrillisch (u.a. Russisch und Serbisch)
8859-6	Arabisch
8859-7	Griechisch
8859-8	Hebräisch
8859-9 (Latin 5)	wie 8859-1 aber Türkisch statt Isländisch
8859-10 (Latin 6)	Eskimo- und Skandinavische Sprachen

Tabelle 2.5: Zeichensätze nach ISO 8859

der elektronischen Kommunikation sowie beim Transfer zwischen verschiedenen Hardware- und Betriebssystemen und im Internet. Für die notwendigen Konvertierungen sind zahlreiche Abbildungstools, Tabellen und Fonts unterschiedlichster Qualität im Public Domain verfügbar.

Mittlerweile wurde ein neuer Standard für eine globale Zukunft entwickelt, genannt *UNICODE*. Dieser Standard kodiert die gebräuchlichsten Zeichensätze der Welt, wie das lateinische, kyrillische, griechische, hebräische und arabische Alphabet, die siamesische (Thai), mongolische und tibetische Schrift und die asiatischen Schriftzeichen, die unter dem Codenamen CJK (für China, Japan, Korea) geführt werden. UNICODE ermöglicht es, in einem Dokument beliebige Schriften gleichzeitig zu nutzen. Er ermöglicht es sogar, die Schreibrichtung innerhalb eines Dokumentes zu wechseln, was mittels Steuerkommandos realisiert wird. UNICODE wird höchstwahrscheinlich aufgrund seiner Globalität, Flexibilität und Einfachheit die landesspezifischen Codierungen mittelfristig ablösen. Neue Softwareentwicklungen wie Java, Tcl8.0 und Windows NT verwenden bereits UNICODE und werden dessen Verbreitung sicher fördern.

Zeicheneingabe

Die Schriftzeichen der heute gebräuchlichen Alphabetschriften, wie etwa im Deutschen, Englischen, Russischen und Arabischen, passen bequem auf die Tastatur. Durch einfache Softwarelösungen (Abbildungstabellen) kann man die Tastenbelegungen einer Standardtastatur umdefinieren und somit in eine Eingabetastatur für jedes beliebige Alphabet umfunktionieren. Auch die Tastenbelegung von Sonderzeichen, etwa durch Tastenkombinationen, wird z.B. durch Public Domain Software übernommen. Schwieriger wird die Eingabe bei den ideographischen Schriften, wie der Chinesischen Hanzi-Schrift, der Japanischen Kanji- und Koreanischen Hangul-Schrift. Deren Zeichen sind zu zahlreich für eine Tastatur. Früher gab es für die japanische Schrift Schreibmaschinen, die eher kleinen Setzmaschinen glichen. Damit erreichte ein geübter Schreiber aber kaum mehr als 20 Zeichen pro Minute, das sind etwa 10 Seiten pro Tag. Heutzutage greift man zur Eingabe ideographischer Schriften auf andere Methoden zurück:

- Eingabe von Zeichenkennung: Bei dieser Methode wird jedes Zeichen über eine Ziffern- oder Zeichenkennung eingegeben. Dies erfordert einen großen Lernaufwand, führt bei geübten Personen allerdings zu erstaunlich hohen Eingabegeschwindigkeiten.
- Eingabe nach dem Prinzip der phonologischen Konversion: Hierbei wird das Zeichen mittels Lautschrift eingegeben. Anhand von Wörterbüchern generiert der Computer eine geordnete Liste in Frage kommender Zeichen, aus der der Benutzer das richtige auswählt. Die Chinesen benutzen Pinyin als Umschrift und die Japaner die Silbenschrift Hiragana.
- Pen-Eingabe: Mit einem speziellen Schreibgerät werden die gewünschten Zeichen auf ein Tableau geschrieben, mittels Software ausgewertet und in Text übertragen.
- OCR-Eingabe: Durch einen Scanner werden bereits geschriebene Papiervorlagen graphisch an den Computer übertragen. Software zur „Optical Character Recognition (OCR)“ analysiert und zerlegt die Graphik in Einzelschriftzeichen.
- Spracheingabe: Die Eingabe der Zeichen erfolgt durch Sprechen oder Vorlesen einer Textvorlage. Ein Diktiererkenner setzt die gesprochene Sprache in geschriebenen Text um. Mit der Verbesserung der Spracherkennungssysteme wird diese natürliche Eingabemethode in den kommenden Jahren stark an Bedeutung zunehmen.

Zeichenwiedergabe

Handelt es sich um eine einfache Sprache wie Englisch, so gibt es eine eindeutige Entsprechung zwischen der Codezahl und der auf dem Bildschirm oder Papier wiedergegebenen Zeichen. Allerdings müssen dem Computer verschiedene Zeichenformen zur Verfügung stehen, um für die vielen Schriften charakteristisch variable Buchstabenformen dazustellen. Ein Beispiel dafür ist das griechische Sigma σ, welches am Wortende in der Form ς auftritt. Gespeichert werden beide Formen mit derselben Codezahl, bei der Ausgabe überprüft das Wiedergabeprogramm den rechten Kontext des Sigma, bei Leerzeichen wird die Endform, ansonsten die normale Form ausgegeben. Das Arabische ist eine komplizierte Fassung desselben Problems.

Die meisten arabischen Buchstaben haben 4 Formen, je nachdem ob sie isoliert, am Wortanfang, in der Mitte oder am Wortende stehen. Außerdem müssen spezielle Verknüpfungsregeln für Buchstaben beachtet werden. Auch hier wird das Problem durch den Kontext gelöst. Dasselbe gilt für Ligaturen, wie sie in vielen lateinischen Schriften üblich sind. Hindi verfügt über Wörter, die nicht in phonologischer Reihenfolge geschrieben werden, vielmehr erscheint der erste Vokal vor dem ersten Konsonanten. Das Wort hindî wird also eigentlich ihndî geschrieben. In der Thai-Schrift können sich Vokalzeichen aufspalten, um einen Konsonanten zu umschließen.

2.4.3 Beziehung zwischen Orthographie und Aussprache

Graphem-zu-Phonem Relation

Wie bereits in Abschnitt 2.4.2.1 beschrieben, gibt es Schriftsysteme, die grundsätzlich keine Relation zwischen Graphemen der geschriebenen Schrift und Phonemen der gesprochenen Sprache haben. Unter den phonologischen Schriftsystemen, die prinzipiell eine solche Relation aufweisen, gibt es eine große Varianz in der Graphem-zu-Phonem Beziehung. Es gibt neben den Alphabeten, die auf die gesprochene Sprache angepaßt oder zugeschnitten wurden, auch zahlreiche Beispiele, in denen die Modifikationen der Schrift nicht mit der Entwicklung der Aussprache einhergingen. Nur wenige Sprachen vollzogen einen radikalen Schritt, wie im Koreanischen oder Türkischen, bei dem ein neues Schriftsystem entwickelt oder ein bestehendes so angepaßt wurde, daß Grapheme und Phoneme perfekt aufeinander abgestimmt sind. Daher gibt es zwischen den Sprachen große Unterschiede in der Graphem-zu-Phonem Relation. An dem einen Ende der Skala findet man das Kroatische, Türkische und Finnische, bei denen die Buchstaben den Lauten sehr genau entsprechen, am anderen Ende das Englische und Gälische, die beide viele Unregelmäßigkeiten aufweisen. Je weniger Grapheme und Phoneme einander entsprechen, umso mehr Rechtschreiberegeln sind zu erlernen. Außerdem gestaltet es sich dann umso schwieriger, mittels einfacher Regeln aus einem geschriebenen Wort dessen Aussprache abzuleiten, wie es beispielsweise für die Erstellung von Aussprachelexika wünschenswert ist (siehe Abschnitt 5.3.5).

Transformation fremdsprachlicher Namen

Das Training automatischer Spracherkenner erfordert eine Verschriftung der Sprachdaten. Dabei unterscheidet man drei Ebenen der Notation bzw. Etikettierung von Sprachsignalen: eine orthographische Wiedergabe des gesprochenen Textes, eine lexikalische Kennzeichnung der Wörter in phonemisch-kanonischer Form sowie eine Kennzeichnung der faktisch vorliegenden Realisierung in phonologischer Notation [PM92]. Aus Aufwandsgründen gehen die meisten Projekte dazu über, Sprachdaten nur noch auf der ersten Ebene zu verschriften. Für die multilinguale Spracherkennung ergeben sich dabei neben den beschriebenen zeichenbedingten Problemen noch weitere spezifische Probleme: Die zu verschriftende Sprache kann sich insbesondere bei Eigennamen von der zur Verschriftung verwendeten Sprache unterscheiden. Zur Lösung dieses Problems gibt es zwei Konzepte [PM95]: Bei der *Transliteration* werden die im Original erscheinenden Namen von Personen, Orten oder Institutionen in ein anderes Schriftsystem umgesetzt; in der Regel Schriftzeichen für Schriftzeichen. Abzugrenzen ist die Transliteration von der quasiphonemischen *Transkription*, bei der die Laute der Sprache, die verschriftet wird, durch Buchstaben der zur Verschriftung verwendeten Sprache ausgedrückt werden. Beispielsweise erhält man bei der Transliteration des Nachnamens des ehemaligen sowjetischen Staatschefs Russlands die Schreibweise „Gorbacev". Dies resultiert aus der exakten Umsetzung des Russischen in unser Alphabet. Bei der Transkription ergibt sich „Gorbatschow", denn dies entspricht der Aussprache des Namens (nach Duden).

Beide Konzepte haben Nachteile: Die Transkription ist inkonsistent, denn das Ergebnis hängt von der jeweiligen Zielsprache ab (bzw. von der Muttersprache des Transkribierenden), so schreibt man beispielsweise im Deutschen „Tschaikowski", im Englisch „Tchaikovsky" und im Ungarischen „Csajkovszkij". Die Transliteration ist willkürlich, denn in der Zielsprache fehlen oft geeignete Symbole, so daß diakritische Zeichen hinzugefügt werden müssen. Die Willkür bei der Auswahl der Symbole ist dann am augenscheinlichsten, wenn zwischen Ausgangs- und Zielsprache keine klaren lautlichen Entsprechungen bestehen. Derzeit gibt es noch keine international anerkannten Standards, sondern zahlreiche Transliterations- bzw Transkriptionssysteme, die nebeneinander existieren. Daher kommt es zu inkonsistenten Schreibweisen, was zu erheblichen Problemen beim Auffinden von Eigennamen in Nachschlagewerken führen kann. Außerdem erschwert dieser Umstand den Vergleich geschriebener Texte von offiziellen Medien oder allgemein zugänglichen elektronischen Quellen, wie etwa dem Internet. Dies spielt insbesondere beim Bau von Sprachmodellen eine wichtige Rolle (siehe Kapitel 5).

2.5 Konsequenzen für die multilinguale Spracherkennung

Dieser Abschnitt faßt die für die Spracherkennung relevanten Merkmale von Schrift und gesprochener Sprache zusammen und zeigt die Konsequenzen auf, die sich für die Entwicklung und den Vergleich von monolingualen Spracherkennern ergeben. Außerdem resultieren hieraus Anhaltspunkte für den Bau eines multilingualen Erkennungssystems.

Wie bereits in Kapitel 1 beschrieben, ist die computerunterstützte Mensch-zu-Mensch Kommunikation in möglichst vielen Sprachen erwünscht. Aber die in Abschnitt 2.1 aufgezeigte Vielzahl der Sprachen und die damit verbundenen Probleme verdeutlichen, daß es weder sinnvoll noch möglich ist, alle Sprachen der Welt in ein einziges Erkennungssystem zu integrieren. Daher sind als erstes diejenigen Sprachen zu bestimmen, mit denen die multilinguale Spracherkennung betrieben werden soll. Zweckmäßig sind Sprachen mit einem hohen Stellenwert, also solche mit einem großen Verbreitungsgrad und wirtschaftlicher Bedeutung. Nach Auswertung der Analyse in Abschnitt 2.2 sind die 10 wichtigsten davon die Sprachen Englisch, Chinesisch, Hindi, Spanisch, Russisch, Französisch, Arabisch, Japanisch, Koreanisch und Deutsch. Abschnitt 2.4 zeigt die gewaltigen Unterschiede auf, die zwischen verschiedenen Sprachen bestehen.

Insbesondere wird aus dem Abschnitt 2.4.2.1 über Schriftsysteme deutlich, daß nichtphonologische Schriftsysteme mit ihren über 10.000 Zeichen nur mit viel Aufwand zu erlernen sind. Um dennoch eine Fehleranalyse der Ausgabe eines Spracherkenners durchführen zu können, ist es sinnvoll, diese Schriften auf das für uns gebräuchliche lateinische Alphabet zu transformieren.

Eng mit dem Schriftsystem, aber auch mit dem Sprachbau einer Sprache verknüpft, ist das Problem der Segmentierung. Lange Worteinheiten eignen sich nicht für die automatische Spracherkennung, weil daraus das Problem eines großen Vokabularwachstums und damit vieler unbekannter Wörter entsteht. Zu kleine Einheiten schränken dagegen die Reichweite des Sprachmodells ein und neigen zu einer höheren akustischen Verwechselbarkeit. Das Problem der Segmentierung stellt sich daher bei allen Sprachen, die entweder bedingt durch ihr Schriftsystem keine Worteinheiten erkennen lassen oder bedingt durch ihren Sprachbau zu sehr langen Worteinheiten neigen.

Bezüglich der Lautinventare von Sprachen zeigte sich, daß das IPA-Alphabet ein geeignetes Referenzsystem ist, um die Lautsysteme verschiedener Sprachen zu vergleichen und ein *universelles* Phoneminventar aufzustellen, wie es für einen multilingualen Erkenner angestrebt wird. Aus der Sicht der Phonologie wird deutlich, daß bei der Modellierung breiterer Kontexte große Unterschiede zwischen den Sprachen zu erwarten sind. Dies ist eine besondere Herausforderung für den Bau eines multilingualen Erkenners. Im Abschnitt Prosodie wurde deutlich, daß sich Sprachen bezüglich ihrer Tonalität voneinander unterscheiden, was besondere Maßnahmen erfordert.

Eine wesentliche Wissensquelle in der Spracherkennung ist das Aussprachewörterbuch, in dem die Orthographie und die Aussprache aller bekannten lexikalischen Einheiten aufgelistet sind. Der Abschnitt 2.4.3 beschreibt, daß sich aufgrund der unterschiedlich ausgeprägten Graphem-zu-Phonem Beziehung der Aufwand zur Generierung von Aussprachen aus der Schreibweise stark unterscheiden kann. Werden die Aussprachen anhand von Regeln automatisch aus der Orthographie erzeugt, ist das Resultat je nach Sprache von unterschiedlicher Güte.

Kapitel 3

Grundlagen der multilingualen Spracherkennung

In diesem Kapitel wird die „Multilinguale Spracherkennung“ definiert, wie sie im Kontext dieser Arbeit verstanden wird. Daneben werden die wesentlichen Begriffe und Methoden der automatischen Spracherkennung eingeführt, soweit sie für die vorliegende Arbeit relevant sind. Dabei wird besonders auf Aspekte eingegangen, die im Zusammenhang der Multilingualität von Belang sind. Dieser kurze Überblick erhebt keinen Anspruch auf Vollständigkeit. Für eine ausführliche Einführung in die automatische Spracherkennung sei auf Werke verwiesen, wie beispielsweise [WL90, RJ93, ST95], die einen guten Einstieg in die wichtigsten Techniken bieten.

In den letzten Jahrzehnten gab es einen rasanten Fortschritt in der Entwicklung von Spracherkennungssystemen. Die Abbildung 3.1 veranschaulicht diese Entwicklung in Form eines „Urknall-Modells“ (in Anlehnung an [AD99]). Das Modell wird durch die Dimensionen *Sprache*, *Sprechstil*, *Kanal* und *Sprecher* aufgespannt, welche die Anforderungen an die automatische Spracherkennung charakterisieren. Die ersten Forschungsanstrengungen auf dem Gebiet der automatischen Spracherkennung galten der sprecherabhängigen Erkennung von Einzelwörtern aus kleinen Wortschätzen unter ruhigen Aufnahmebedingungen. Erst in den frühen achtziger Jahren konnten Erkennungssysteme kontinuierliche Sprache akzeptieren, sprecherunabhängig arbeiten oder große Wortschätze bewältigen (vgl. [ST95]). Im Blickpunkt heutiger Systementwicklungen stehen unter anderem die Bewältigung von umgangssprachlicher und ungrammatikalischer Sprechweise etwa wie bei CallHome [LDC00, BMM+97] und die Erkennung unter stark geräuschbehafteten Umgebungsbedingungen wie etwa im Auto [WW99, VOD98]. Die Entwicklungen in der Dimension „Sprache/Anwendung“ und „Sprecher“ sind durch Techniken gekennzeichnet, die ein System an neue Gegebenheiten optimal anpassen sollen. „Multilinguale Spracherkennung“, wie sie unten definiert wird, kann als extreme Variante dieser Adaptionsidee betrachtet werden. Die kleinen Markierungen in Abbildung 3.1 veranschaulichen, daß sich die vorliegende Arbeit in der Dimension Sprechstil und Kanal auf einem gesicherten Terrain

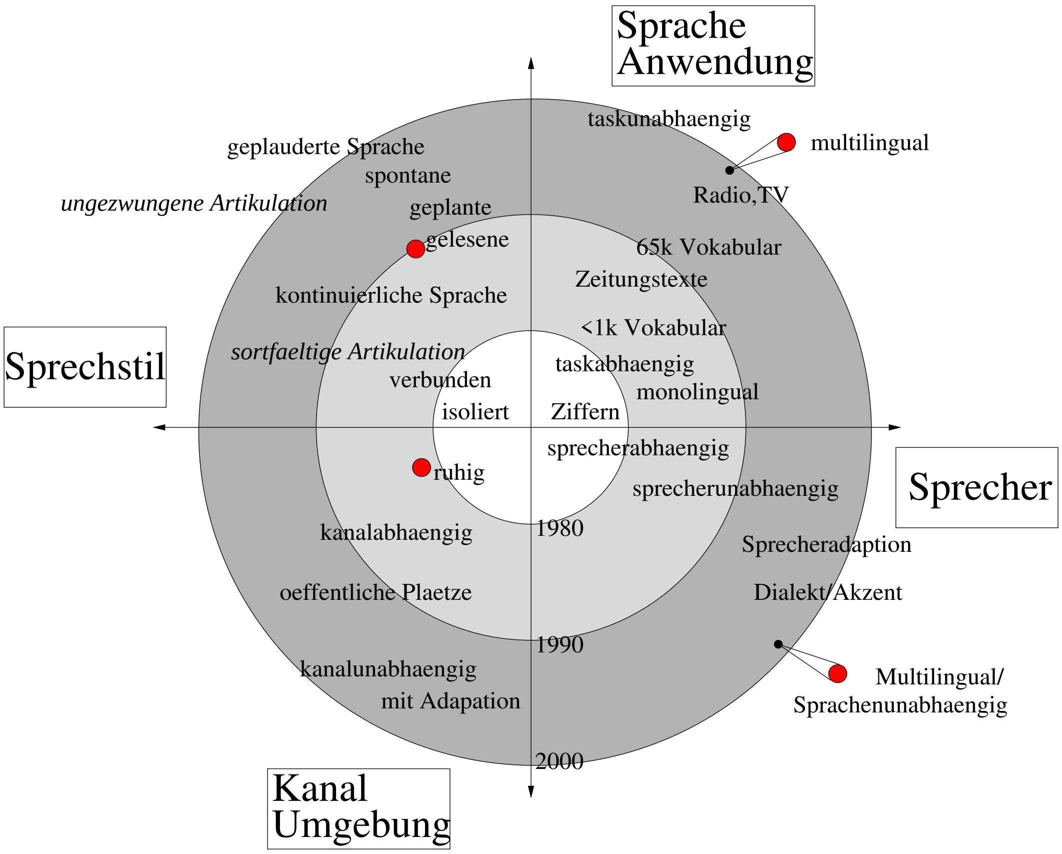


Abbildung 3.1: Urknall-Modell der Spracherkennung (in Anlehnung an [AD99])

bewegt, aber bezüglich der Multilingualität gewissermaßen eine neue Dimension eröffnen.

Das Problem der automatischen Spracherkennung besteht darin, zu einer gesprochenen Äußerung $\mathbf{X}$ diejenige Wortsequenz W^* zu finden, die $\mathbf{X}$ am wahrscheinlichsten produziert:

$$\begin{aligned} W^* &= \underset{W}{\operatorname{argmax}} \; P(W|\mathbf{X}) \\ &= \underset{W}{\operatorname{argmax}} \; \frac{P(\mathbf{X}|W) \cdot P(W)}{P(\mathbf{X})} \\ &= \underset{W}{\operatorname{argmax}} \; P(\mathbf{X}|W) \cdot P(W) \end{aligned} \tag{3.1}$$

Aus dieser *Fundamentalgleichung der Spracherkennung* lassen sich die drei Grundprobleme der Spracherkennung ableiten:

- Akustische Modellierung: Wie berechnet man die bedingte Wahrscheinlichkeit $P(\mathbf{X}|W)$, das Signal $\mathbf{X}$ zu beobachten unter der Annahme, daß die Wortsequenz W gesprochen wurde.

- Sprachmodellierung: Wie berechnet man die a-priori Wahrscheinlichkeit $P(W)$, daß die Wortsequenz W gesprochen wurde.

- Suche: Wie berechnet man auf möglichst effiziente Weise diejenige Wortsequenz W^*, die $P(\mathbf{X}|W) \cdot P(W)$ maximiert.

In Abbildung 3.2 ist die Struktur eines Erkennungssystems veranschaulicht. Das Ziel eines kontinuierlichen Spracherkennungsvorganges ist eine Abbildung der gesprochenen Äußerung $\mathbf{X}$ auf ihre textuelle Darstellung $W^* = w_1 w_2 \dots w_n$. In einer multilingualen Diktieranwendung ist die orthographisch korrekte Darstellung des erkannten Satzes im sprachenspezifischen Schriftsystem gesucht[1]. Wie in Abschnitt 3.2.2 noch erläutert wird, ist zur effizienten Berechnung von $P(\mathbf{X}|W)$ ein mehrstufiger Prozeß notwendig. Dabei wird eine Wortsequenz in einzelne Wörter und diese in kleinere sprachliche Einheiten, etwa Phoneme, zerlegt. Neben dem eigentlichen akustischen Modell in Abbildung 3.2 benötigt man daher als zusätzliche Wissensquelle ein *Aussprachewörterbuch*, in dem alle zu erkennenden Wörter durch die Konkatenation der sprachlichen Untereinheiten repräsentiert sind, und ein *Sprachmodell*, das die Wahrscheinlichkeiten der Konkatenation von Wörtern zu Wortsequenzen angibt. Um überhaupt Sprache in einem automatischen System verarbeiten zu können, ist außerdem eine Signalvorverarbeitung notwendig, die aus dem akustischen Signal geeignete Merkmalsvektoren $\mathbf{X}$ extrahiert. Ein monolinguales Spracherkennungssystem besteht demnach im wesentlichen aus einer Komponente zur Merkmalsextraktion, einem *Akustischen Modell*, einem *Aussprachewörterbuch* und einem *Sprachmodell* sowie dem *Dekodierer*, der unter Ausnutzung dieser Komponenten die beste Wortsequenz W^* findet. Die Komponenten Akustisches Modell, Aussprachewörterbuch und Sprachmodell werden im folgenden unter dem Begriff *Wissensquelle* zusammengefaßt, da sie im allgemeinen die sprachenspezifischen Informationen enthalten.

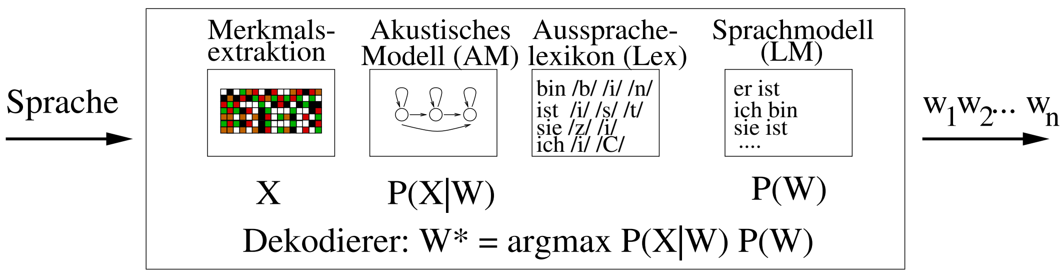


Abbildung 3.2: Automatische Spracherkennung

[1]Es werden hier nicht die Probleme des Sprachverstehens, also der Erfassung der Intention des Gesagten, diskutiert, die für eine Übersetzung von Text in eine andere Sprache gelöst werden müssen

3.1 Definition der multilingualen Spracherkennung

Unter „Multilingualität“ wird in der Literatur im allgemeinen die Mehrsprachigkeit von Spracherkennungsystemen verstanden. Dabei handelt es sich um ein Gesamtsystem, das mehrere Sprachen verarbeiten kann. Solche Gesamtsysteme entstehen beispielsweise durch die Zusammenschaltung mehrerer monolingualer Spracherkennungskomponenten. In den einzelnen Komponenten wird das Problem der Erkennung für jede Sprache separat gelöst, unter Ausnutzung der Wissensquellen, die für jede Sprache separat trainiert wurden. Falls die Eingabesprache in der Anwendungen als nicht bekannt vorausgesetzt wird, sind solche Gesamtsysteme mit einer Sprachenidentifizierungskomponente (LID=Language IDentification) gekoppelt. Die LID identifiziert entweder als vorgeschaltete (front-end) Komponente die Eingabesprache und leitet das Signal an den entsprechenden monolingualen Erkenner weiter. Oder die LID analysiert als nachgeschaltete (back-end) Komponente die Ausgabe aller monolingualen Erkenner und wählt diejenige Sprache, deren Spracherkenner die beste Bewertung der Eingabesprache liefert. Die Abbildung 3.3 veranschaulicht diese Anwendung mit einer nachgeschalteten LID. In solchen Anwendungen kommt in der Regel die gemeinsame Nutzung von Spracherkennungssoftware zum Tragen, d.h. die Extraktion geeigneter Merkmale wird nur einmal durchgeführt. Der eigentliche Dekodierer wird nur einmal implementiert und die benötigten sprachenspezifischen Wissensquellen werden zur Laufzeit dazugebunden.

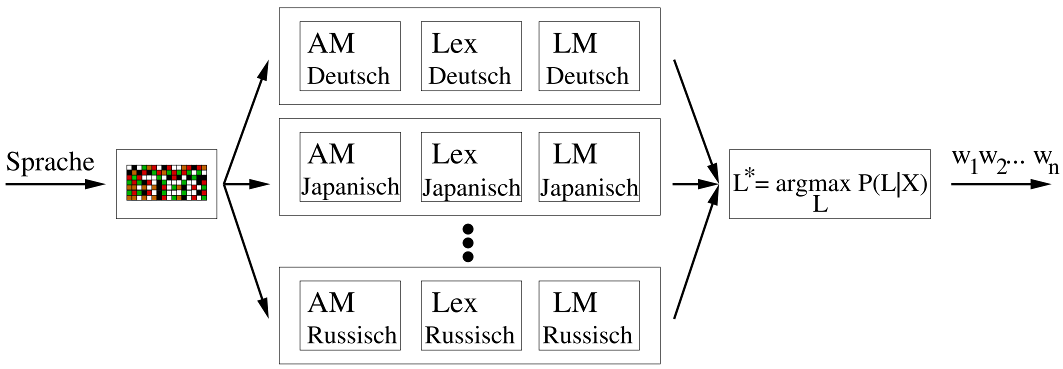


Abbildung 3.3: Mehrsprachige monolinguale Spracherkennung

Der Begriff *multilinguale Spracherkennung* geht in dieser Arbeit über den beschriebenen Ansatz der Mehrsprachigkeit weit hinaus. Als multilinguales Spracherkennungssystem wird hier ein System bezeichnet, in dem mindestens eine der drei Wissensquellen Akustisches Modell, Aussprachewörterbuch oder Sprachmodell von *mehreren Sprachen gemeinsam* genutzt wird. Dabei wird davon ausgegangen, daß die gemeinsame Nutzung einer Wissensquelle deren Training anhand gemeinsamer Daten einschließt. Neben das Konzept der gemeinsamen Nutzung von Software tritt somit zusätzlich das Konzept der gemeinsamen Nutzung von Daten. Eine mit den

Daten vieler Sprachen trainierte Wissensquelle wird als *sprachenunabhängig* oder *multilingual* bezeichnet. Die Abbildung 3.4 zeigt ein multilinguales Spracherkennungssystem, in dem alle drei Wissensquellen multilingual sind.

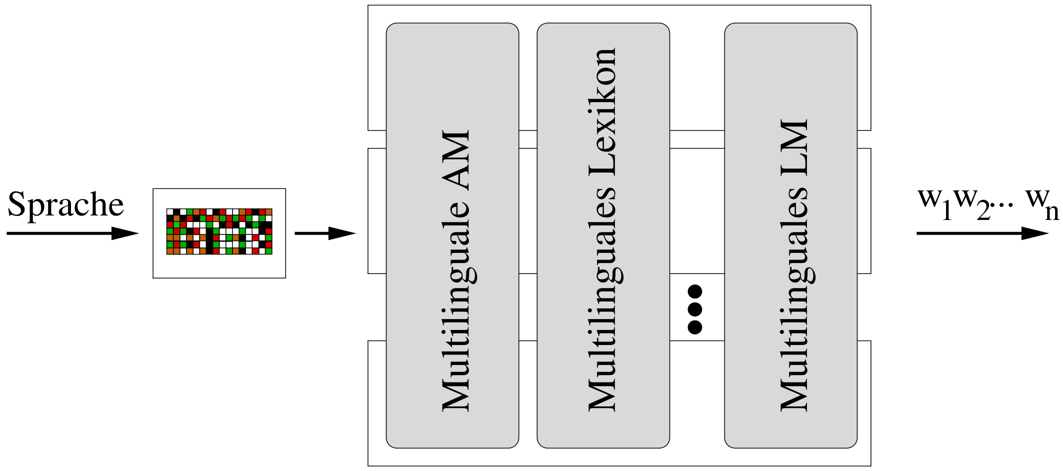


Abbildung 3.4: Multilinguale Spracherkennung

Der Hauptbeweggrund für die multilinguale Modellierung ist, daß durch gemeinsam genutzte Wissensquellen die Gesamtzahl der Systemparameter reduziert werden kann. Man spart den Overhead, der durch viele monolinguale Systeme entsteht. Die Gesamtstruktur wird dadurch schlanker und übersichtlicher und ist leichter zu warten. Dieser Aspekt ist für alle speicherlimitierten Anwendungen, wie beispielsweise im Auto oder in Hand-Helds, sehr wichtig. Ein weiterer interessanter Aspekt ist, daß die Sprachenidentifizierung implizit geschieht, und damit die oben vorgestellten Lösungen einer LID-Komponente entfallen können. Die Motivation für die multilinguale Spracherkennung sind allerdings noch vielfältiger und können danach gegliedert werden, welche der Wissensquellen multilingual realisiert werden:

Multilinguales akustisches Modell Kombiniert man die akustischen Modelle vieler Sprachen miteinander, kann man die Daten aller beteiligten Sprachen nutzen, um die einzelnen Modelle robuster zu schätzen. Sofern mehr Sprachdaten auf ein Modell entfallen, erhofft man sich eine Verbesserung der Erkennungsleistung. Besonders interessant in diesem Zusammenhang ist die Wahl von Phonemen als Modellierungseinheit. Das Lautinventar des Menschen ist beschränkt, daher ist die Annahme berechtigt, daß sich viele Sprachen dieselben oder sehr ähnliche Laute oder Phoneme teilen. Mit zunehmender Anzahl von Sprachen wächst die Wahrscheinlichkeit, das Phoneminventar einer neuen, im Training nicht enthaltenen, Sprache abdecken zu können. Dies ist die Idee der *Portierung auf eine neue Sprache*. Ein multilinguales Phonemset, bestehend aus einer kompakten Menge generalisierungsfähiger Modelle, ist die optimale Ausgangsbasis für einen schnellen Übergang auf eine neue Sprache (vgl. Abschnitt 7).

Multilinguales Sprachmodell Zur Realisierung des Sprachenwechsels innerhalb einer Äußerung werden multilinguale Sprachmodelle benötigt. Das Wechseln der Sprache sogenanntes *Code-Switching* geschieht beispielsweise zur Einbettung anderssprachiger Phrasen (vgl. Abschnitt 6.2.1). Multilinguale Sprachmodelle sind auch zur Sprachenidentifizierung von Interesse, wenn sie zusammen mit mehrsprachigen monolingualen Systemen betrieben werden (vgl. Abschnitt 6.4.1).

Multilinguales Aussprachewörterbuch In der Literatur wird der Term multilinguales Aussprachewörterbuch uneinheitlich verwendet. Zum einen wird damit ein Aussprachewörterbuch bezeichnet, das aus einer Verzahnung monolingualer Aussprachewörterbücher entsteht. Dies kann für Anwendungen sinnvoll sein, in denen die zu sprechenden Einheiten über viele Sprachen hinweg gemeinsamen Gesetzmäßigkeiten gehorchen (vgl. Abschnitt 6.2.2). Zum anderen wird der Begriff für das Problem verwendet, das entsteht, wenn die Aussprachen von Wörtern, die im Phonemset einer Sprache beschrieben sind, auf das Phonemset einer anderen Sprache abgebildet werden sollen.

Die multilinguale Modellierung von Phoneminventaren ist ein Hauptschwerpunkt dieser Arbeit. Es kommen daneben aber auch multilinguale Aussprachewörterbücher im Sinne der Abbildungen von Phonemsets zum Tragen, die für die Portierung auf neue Sprachen notwendig sind. Auf multilinguale Sprachmodelle wird im Zusammenhang mit der Sprachenidentifizierung eingegangen werden. Bevor die Umsetzung dieser Ideen beschrieben wird, erfolgt zunächst eine Einführung in die Techniken, die zur Lösung der drei Grundprobleme der Spracherkennung eingesetzt werden.

3.2 Grundlagen der Spracherkennung

Im folgenden werden die Grundlagen der automatischen Spracherkennung zur Signalvorverarbeitung, akustischen Modellierung, Sprachmodellierung und Suche erläutert. Dabei werden sprachenspezifische Aspekte besonders hervorgehoben.

3.2.1 Signalvorverarbeitung

Das Ziel der Signalvorverarbeitung ist es, eine zur gesprochenen Äußerung zugehörige Schalldruckwelle in ein für die Spracherkennung geeignetes Signal zu überführen. Dazu muß die analoge Schalldruckwelle zunächst *diskretisiert* werden. Der enorme Umfang resultierender Daten zwingt zur Extraktion weniger, relevanter Merkmale. Anschließend werden die ausgewählten Merkmale mit dem Ziel der Dimensionalitätsreduktion oder der Adaption weiter transformiert.

Diskretisierung
Die Schalldruckwelle eines Sprachsignals wird durch ein Mikrophon in ein reellwertiges elektrisches Signal gewandelt. Zur Verarbeitung im Computer wird dieses analoge Signal durch einen Analog-Digital-Wandler in ein digitales Signal transformiert. Dabei wird das reellwertige Signal unter Berücksichtigung des Shannonschen Abtasttheorems abgetastet und quantisiert. Das Shannonsche Abtasttheorem fordert eine Abtastung mit mindestens der doppelten Signalgrundfrequenz, um die vollständige Rekonstruktion des Ursprungssignals zu garantieren. Um alle Sprachlaute akkurat zu repräsentieren, würde man eine Abtastrate von etwas über 20 kHz benötigen [RS78], für die Spracherkennung reicht eine Abtastrate von 16 kHz allerdings völlig aus. Zur Quantisierung der Abtastwerte werden meist 8- oder 16-bit Auflösungen gewählt, d.h. das Sprachsignal wird durch 2^8 bzw. 2^{16} Quantisierungsstufen repräsentiert.

Kurzzeitanalyse
Werden die Sprachdaten mit 16 kHz Abtastrate und 16-bit Auflösung digitalisiert, dann fallen pro Sekunde 32 kByte Daten an. Der nachfolgende Vorverarbeitungsschritt erfolgt daher mit dem Ziel, diese Datenflut zu reduzieren. Es sollen dazu diejenigen Merkmale extrahiert werden, die für die Spracherkennung relevant sind.

Für die Kurzzeitanalyse, bei der man annimmt, daß das Signal über Zeiträume von etwa 10-30 ms stationär ist, werden mit einer geeigneten Fensterfunktion (z.B. Hamming) aufeinanderfolgende Zeitfenster aus dem Signal ausgeschnitten und analysiert. Aus diesen Ausschnitten werden Merkmale extrahiert, die man entweder aus dem Zeitbereich oder aus dem Spektralbereich gewinnt. Fast alle heute gängigen Spracherkenner verwenden Merkmale aus dem Spektralbereich und wenden dabei das gleiche Grundprinzip der Signaltransformation vom Zeit- in den Frequenzbereich an: die Fouriertransformation. Die gängigen Verfahren zur weiteren Reduktion der anfallenden Spektralkoeffizienten basieren auf Melscale-Spektral Koeffizienten oder auf Ceptral Koeffizienten.

Nach der Merkmalsextraktion können weitere Schritte folgen, welche die gewonnenen Merkmalsvektoren transformieren, etwa zur Reduktion der Dimensionalität oder zur Sprecheradaption. In diesem Zusammenhang werden die hier angewendeten Verfahren der Linearen Diskriminanzanalyse und der Vokaltraktlängennormierung erläutert.

Lineare Diskriminanzanalyse (LDA)
Bei der Linearen Diskriminanzanalyse wird eine Hauptachsentransformation auf den extrahierten Merkmalsvektoren durchgeführt. Dabei wird eine LDA-Matrix bestimmt, welche die Merkmalsvektoren so transformiert, daß ihre Koeffizienten dekorreliert und entsprechend ihrer Diskriminierungsfähigkeit sortiert werden, beginnend mit dem erste Koeffizienten als dem mit der größten Varianz. Die LDA berücksichtigt die Klassenzugehörigkeit der Merkmalsvektoren und berechnet die Transformation

so, daß die durchschnittliche Varianz innerhalb der Klassen minimiert und gleichzeitig die Varianz zwischen den Klassen maximiert wird. Merkmalsvektoren, die einer gemeinsamen Klasse angehören, rücken daher quasi näher zusammen, während sich die einzelnen Klassenzentren weiter voneinander entfernen. Die Trennschärfe zwischen den Klassen wird somit erhöht, was die Klassifikationsaufgabe vereinfacht. Durch die Anordnung der Koeffizienten in den Merkmalsvektoren kann mit Hilfe der LDA eine Dimensionalitätsreduktion durchgeführt werden, indem bei der Multiplikation mit der LDA-Matrix die Vektorkoeffizienten niedriger Ordnung abgeschnitten werden.

Vokaltraktlängennormierung (VTLN)

Die Vokaltraktlängennormierung [ZW97] ist ein Adaptionsverfahren, das durch Transformation des Spektralraumes die Auswirkungen unterschiedlich langer Vokaltrakte zu kompensieren versucht. In der Regel haben Männer einen längeren Vokaltrakt als Frauen, was zu einer kürzeren Resonanzfrequenz (erste Formante) führt. Dies ist neben der tieferen Grundfrequenz einer der Gründe, warum Männer eine tiefere Stimme haben. Bei geschlechtsunabhängiger Modellierung würde in den extrahierten Merkmalsvektoren daher neben der sonstigen Sprechervariabilität die zusätzlichen Variation durch lange und kurze Vokaltrakte repräsentiert werden. Das Ziel der VTLN ist es, diese Variation bereits in der Vorverarbeitung zu kompensieren. In der Folge muß nur eine Modell gelernt werden, auf das mehr Daten entfallen und somit robuster geschätzt werden kann.

Zur VTLN wird anhand des Sprachsignals die Vokaltraktlänge l eines Sprecher geschätzt, und so normiert, daß der Durchschnittssprecher eine Vokaltraktlänge von 1.0 erhält. In Abhängigkeit von l wird das Leistungsspektrum des Sprachsignals eines Sprechers durch eine stückweise lineare Abbildung verschoben, so daß es bei einem langen Vokaltrakt angehoben, bei einem kurzen Vokaltrakt gesenkt wird. Wichtig für den Erfolg der VTLN ist die möglichst genaue Schätzung von l. Dazu wird l für jeden Sprecher so bestimmt, daß es die Viterbi-Pfadwahrscheinlichkeiten (siehe nächster Abschnitt) des Sprechers maximiert. Da die Pfadwahrscheinlichkeiten von dem initialen l (im ersten Schritt 1.0) abhängen, muß l durch Iterationen dieses Verfahrens bestimmt werden.

Die Verfahren zur Signalvorverarbeitung sind grundsätzlich unabhängig von der gesprochenen Sprache. Allerdings spielen bei der Auswahl zu extrahierender Merkmale die Eigenschaften einer Sprache sehr wohl eine wichtige Rolle. So haben in Tonsprachen die Verläufe der Grundfrequenz eine bedeutungsunterscheidende Funktion. Es kann für solche Sprachen daher sinnvoll sein, die Information über den Grundfrequenzverlauf direkt in den Merkmalsvektor zu integrieren. In nichttonalen Sprachen sind solche Merkmale nicht relevant. Im Kapitel 5.5.2 wird am Beispiel der chinesischen Sprache der Einfluß verschiedener Merkmalsextraktionen evaluiert.

3.2.2 Akustische Modellierung

Die akustische Modellierung beschäftigt sich mit der Berechnung der bedingten Wahrscheinlichkeit $P(\mathbf{X}|W)$, das Signal $\mathbf{X}$ zu beobachten unter der Annahme, daß die Wortsequenz W gesprochen wurde.

3.2.2.1 Hidden Markov Modelle

Sprache ist eine zeitlich variables, kontinuierliches und komplexes Phänomen. Dies drückt sich beispielsweise darin aus, daß ein Wort durch Koartikulationseffekte, sprecherspezifische Aussprachevarianten oder durch Übertragungseigenschaften des Kanals akustisch sehr unterschiedlich realisiert sein kann. Zur automatischen Erkennung von Sprache benötigt man daher Techniken, mit denen man diese Eigenschaften der Sprache modellieren kann. Die *Hidden Markov Modelle (HMM)* sind die gebräuchlichste Technik in der automatischen Spracherkennung. Die folgende Einführung der HMMs ist an [ST95] und [Rab90] angelehnt.

Es wird zunächst der Fall der Einzelworterkennung angenommen. Gesucht ist das Wort w_i, das für ein gegebenes akustisches Signal $\mathbf{X}$ die a-posteriori Wahrscheinlichkeit maximiert:

$$P(w_i|\mathbf{X}) = \frac{P(\mathbf{X}|w_i) \cdot P(w_i)}{P(\mathbf{X})} \tag{3.2}$$

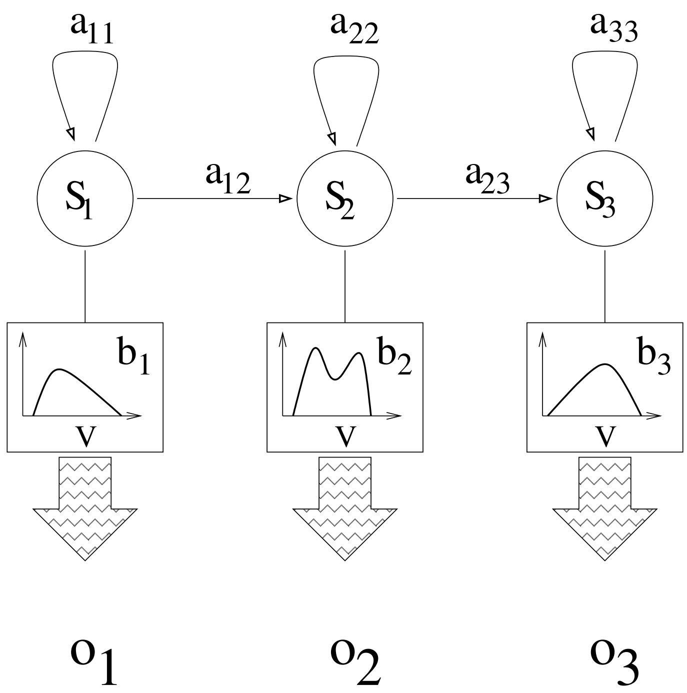
Die Aufgabe des HMM $\boldsymbol{\lambda}_i$ besteht darin, die Verteilungsdichte $P(\mathbf{X}|w_i)$ zu modellieren. $P(\mathbf{X}|\boldsymbol{\lambda}_i)$ wird in einem zweistufigen Zufallsprozeß berechnet, in dem pro Zeitschritt zunächst ein interner HMM-Zustand bestimmt wird und dann die Wahrscheinlichkeit für die Ausgabe eines Merkmalsvektors in diesem Zustand berechnet wird (siehe Abbildung 3.5).

Definition der HMMs
Ein HMM $\boldsymbol{\lambda}$ wird durch die folgenden 5 Komponenten vollständig definiert:

- S ist die Menge aller N HMM-Zustände
  $S := \{S_1, S_2, \ldots, S_N\}$
  Diese Zustände werden in einem diskreten stochastischen Prozeß durchlaufen, der eine Zustandsfolge $\mathbf{q}= q_1 q_2 \ldots q_T, \quad q_t \in S$ entstehen läßt

- $\boldsymbol{\pi}$ ist die Wahrscheinlichkeitsverteilung, die für jeden Zustand S_i die Wahrscheinlichkeit angibt, als erster Zustand q_1 einer Zustandsfolge aufzutreten
  $\pi_i = P(q_1 = S_i), \quad i = 1 \ldots N$

- $\mathbf{A}$ ist die $N \times N$ Matrix von Übergangswahrscheinlichkeiten, welche die bedingte Wahrscheinlichkeit des Wechsels von einem Zustand in den nächsten bestimmt
  $\mathbf{A}= (a_{ij})$, mit $a_{ij} = P(q_t = S_j | q_{t-1} = S_i), \quad i, j = 1 \ldots N$



Gemäß $\boldsymbol{\pi}$ wird für $t = 1$ ein Startzustand $q_1 = S_i$ gewählt. Dann wird gemäß a_{12} aus dem aktuellen Zustand q_1 in den Zustand q_2 übergewechselt. In jedem Zustand S_i wird gemäß der $b_i(v_k)$ das Symbol $o_t = v_k$ ausgewählt. Man bezeichnet diesen Zufallsprozeß als *Hidden* Markov Prozeß, weil für einen Beobachter nur die Folge der ausgegebenen Symbole sichtbar ist, nicht jedoch die Folge der Zustandsänderungen.

Abbildung 3.5: Generierung einer Beobachtungsfolge mittels eines HMM

- V ist die Menge der K beobachtbaren Symbole, aus der in der zweiten Stufe des Zufallsprozesses in jedem Zeitschritt ein Symbol ausgegeben wird
  $V := \{v_1, v_2, \dots, v_K\}$
- **B** ist die Matrix der Emissionswahrscheinlichkeiten, welche die bedingten Wahrscheinlichkeiten dafür angibt, im Zustand S_j die Ausgabe $o_t = v_k$ zu beobachten;
  $\mathbf{B} = (b_j(k))$ mit $b_j(k) = P(o_t = v_k | q_t = S_j), \quad j = 1 \dots N, k = 1 \dots K$

Da man vereinfachend annimmt, daß für den anstehenden Zustandswechsel jeweils nur der unmittelbar vorhergehende Zustand eine Rolle spielt, bezeichnet man $\boldsymbol{\lambda}$ als HMM erster Ordnung. Der Zufallsprozeß ist außerdem stationär, da der absolute Zeitpunkt t unerheblich ist. Abbildung 3.5 veranschaulicht, wie mit dem definierten HMM $\boldsymbol{\lambda}$ eine Observationssequenz $O = o_1 o_2 \dots o_T$ generiert wird.

Man spricht von *diskreten* HMMs, wenn der zugrundeliegende Merkmalsraum diskret ist. In diesem Fall sind die Emissionswahrscheinlichkeiten $b_j(k)$ durch Wahrscheinlichkeitstabellen gegeben, die über den diskreten Symbolen des Ausgabealphabetes definiert sind. Bei *kontinuierlichen* HMMs ist der Merkmalsraum dagegen kontinuierlich. Aus den Wahrscheinlichkeitstabellen werden Wahrscheinlichkeitsdichten $b_j(\boldsymbol{x})$. In die Berechnung der Emissionswahrscheinlichkeiten kann nun auch die Distanz der Beobachtung zur Referenz mit einfließen. Üblicherweise wird $b_j(\boldsymbol{x})$ durch eine Gaußsche Mischverteilung modelliert:

$$b_j(\boldsymbol{x}) = \sum_{l=1}^{L_j} c_{jl} \cdot Gau\beta(\boldsymbol{x} | \mu_{jl}, \Sigma_{jl}) \quad , \quad \sum_{l=1}^{L_j} c_{jl} = 1 \tag{3.3}$$

Wobei L_j die Anzahl der in Zustand S_j eingesetzten Normalverteilungen ist. Die Gauß- oder Normalverteilung $\mathit{Gau\ss}(\boldsymbol{x}|\mu, \Sigma)$ mit dem Mittelwertsvektor μ und der Kovarianzmatrix Σ ist definiert als:

$$\mathit{Gau\ss}(\boldsymbol{x}|\mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^d|\Sigma|}} \cdot e^{\frac{1}{2}(\boldsymbol{x}-\mu)^T \sum_i^{-1}(\boldsymbol{x}-\mu)} \tag{3.4}$$

Die Abbildung 3.6 veranschaulicht eine eindimensionale Gaußsche Mischverteilung mit $L_j = 3$ eingesetzten Normalverteilungen.

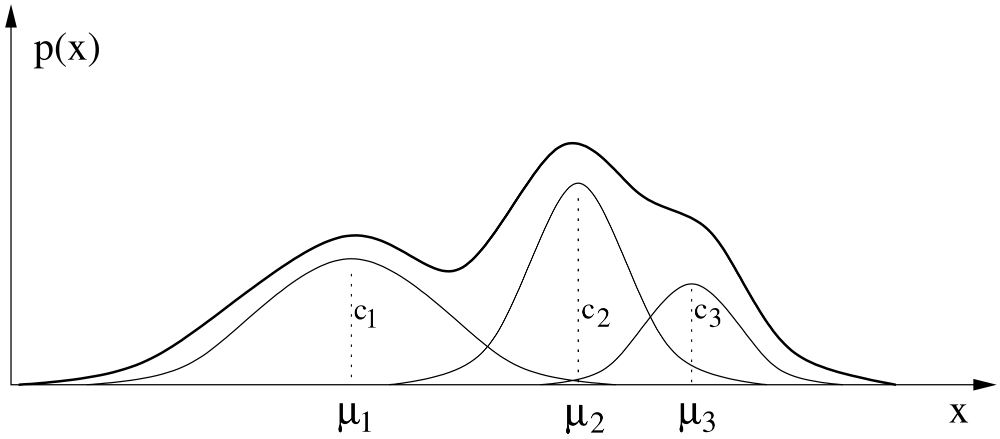


Abbildung 3.6: Eindimensionale Gaußsche Mischverteilung

Die Mittelwertvektoren μ_{jl} und Varianzen Σ_{jl} werden als *Codebook* eines Modells, die Gewichtungskoeffizienten c_{jl} als *Mixturgewichte* bezeichnet.

Bei sogenannten *vollkontinuierlichen* HMMs (CDHMM) besitzt jeder Zustand eine Wahrscheinlichkeitsdichte mit eigenen Mittelwerten, Varianzen und Mixturgewichten. Bei *semikontinuierlichen* HMMs (SCHMM) dagegen versucht man die Vorteile der kontinuierlichen Dichten mit der geringeren Parameterzahl diskreter HMMs zu verbinden. Wie bei den CDHMM besitzt jeder Zustand individuelle Mixturgewichte, die Normalverteilungen werden allerdings global von allen Zuständen gemeinsam genutzt.

Mit dem definierten HMM-Modell lassen sich nun Algorithmen angeben, welche die Fragestellungen der Spracherkennung lösen. Dazu werden zur Erkennung kontinuierlicher Sprache die Satzmodelle aus Sequenzen von Wortmodellen und Wortmodelle gegebenenfalls wiederum aus Sequenzen von kleineren Spracheinheiten zusammengesetzt (siehe Abschnitt 3.2.2.2). Die Erkennung von Satzmodellen wird im Abschnitt 3.3.1 beschrieben, zur Erläuterung der grundlegenden Algorithmen wird in diesem Abschnitt zunächst von Einzelwörtern ausgegangen.

Das Evaluierungsproblem

Wie groß ist die Wahrscheinlichkeit $P(O\boldsymbol{\lambda})$, daß eine bestimmte Beobachtungsfolge

$O = o_1 o_2 \dots o_T$ von einem gegebenen HMM-Modell $\boldsymbol{\lambda}= (\mathbf{A},\mathbf{B},\boldsymbol{\pi})$ erzeugt wurde? Dieses Problem wird mit dem *Vorwärts-Algorithmus* gelöst.

Unter der Annahme, daß eine konkrete Zustandsfolge $\boldsymbol{q} = q_1 q_2 \dots q_T$ gegeben ist, kann die gesuchte Wahrscheinlichkeit $P(O|\boldsymbol{\lambda},\mathbf{q})$ berechnet werden, indem die entsprechenden Übergangs- und Emisionswahrscheinlichkeiten entlang dieser Folge aufmultipliziert werden.

$$P(O|\boldsymbol{\lambda}, \boldsymbol{q}) = \pi_{q_1} b_{q_1}(o_1) \prod_{t=2}^{T} a_{q_{t-1},q_t} b_{q_t}(o_t) \tag{3.5}$$

Die Beobachtungssequenz O kann natürlich durch verschiedene Zustandsfolgen generiert worden sein. Die Wahrscheinlichkeit für O auf irgendeinem aller möglichen Zustandsfolgen $\boldsymbol{q}$ erzeugt worden zu sein, ist

$$P(O|\boldsymbol{\lambda}) = \sum_{q} P(O|\boldsymbol{\lambda}, \boldsymbol{q}) = \sum_{q} \pi_{q_1} b_{q_1}(o_1) \prod_{t=2}^{T} a_{q_{t-1},q_t} b_{q_t}(o_t) \tag{3.6}$$

Eine effiziente Lösung der Gleichung ist der Vorwärts-Algorithmus, der nach dem Prinzip der dynamischen Programmierung verfährt und Vorwärtswahrscheinlichkeiten α als Hilfsgrößen definiert mit $\alpha_t(j) = P(o_1 \dots o_t, q_t = S_j|\boldsymbol{\lambda})$. Dabei gilt:

$$\alpha_t(j) = \begin{cases} \pi_j b_j(o_t) & \text{falls } t = 1 \\ \sum_{i=1}^{N} \alpha_{t-1}(i) a_{ij} b_j(o_t) & \text{falls } t > 1 \end{cases} \tag{3.7}$$

Damit läßt sich $P(O|\boldsymbol{\lambda})$ durch Aufsummieren der Vorwärtswahrscheinlichkeiten berechnen:

$$P(O|\boldsymbol{\lambda}) = \sum_{j=1}^{N} \alpha_T(j) \tag{3.8}$$

Das Dekodierungsproblem

In der Spracherkennung ist man in der Regel nicht an der Gesamtwahrscheinlichkeit $P(O|\boldsymbol{\lambda})$ interessiert, die mit dem Vorwärtsalgorithmus berechnet wird, sondern es soll derjenige Pfad $\boldsymbol{q}^*$ bestimmt werden, der die Beobachtungssequenz O am wahrscheinlichsten generiert hat. Dieses Problem wird mit einer Variante des Vorwärts-Algorithmus, dem *Viterbi-Algorithmus* gelöst.

Gesucht ist also:

$$\boldsymbol{q}^* = \underset{\mathbf{q}\in S^T}{\operatorname{argmax}} \; P(\boldsymbol{q}|O, \boldsymbol{\lambda}) = \underset{\mathbf{q}\in S^T}{\operatorname{argmax}} \; \frac{P(O, \boldsymbol{q}|\boldsymbol{\lambda})}{P(O|\boldsymbol{\lambda})} = \underset{\mathbf{q}\in S^T}{\operatorname{argmax}} \; P(O, \boldsymbol{q}|\boldsymbol{\lambda}) \tag{3.9}$$

Zur Berechnung von $\boldsymbol{q}^*$ wird statt der partiellen Gesamtwahrscheinlichkeit $\alpha_t(j)$ jeweils das Maximum $\theta_t(j)$ berechnet mit:

$$\theta_t(j) = \max_{\mathbf{q}\in S^T} P(o_1 o_2 \dots o_t, q_1 q_2 \dots q_{t-1}, q_t = S_j|\boldsymbol{\lambda}) \tag{3.10}$$

und in der Rekursionsgleichung 3.7 tritt an die Stelle der Summe die Maximumbildung:

$$\theta_t(j) = \begin{cases} \pi_j b_j(o_t) & \text{falls } t = 1 \\ \max_{i=1}^{N} \theta_{t-1}(i) a_{ij} b_j(o_t) & \text{falls } t > 1 \end{cases} \tag{3.11}$$

und es gilt:

$$\max_{i=1}^{N} \theta_T(j) = P^*(O|\boldsymbol{\lambda}) := P(O, \boldsymbol{q}^*|\boldsymbol{\lambda}) \tag{3.12}$$

Das Optimierungsproblem
Die Aufgabe des Optimierungsproblems ist es, zu einem gegebenen Modell $\boldsymbol{\lambda}=(\mathbf{A},\mathbf{B},\boldsymbol{\pi})$ des Wortes w und zur dessen akustischer Realisierung O diejenigen Parameter $\boldsymbol{\lambda}^*$ zu finden, welche die Wahrscheinlichkeit maximieren, O zu produzieren: $\boldsymbol{\lambda}^* = \underset{\lambda}{\operatorname{argmax}} \; P(O|\boldsymbol{\lambda})$.

Zur Beschreibung der Lösung des Optimierungsproblems führt man zunächst analog zu den Vorwärtswahrscheinlichkeiten α die Rückwärtswahrscheinlichkeiten β ein, mit $\beta_t(j) = P(o_{t+1}o_{t+2}\ldots o_T, q_t = S_j|\boldsymbol{\lambda})$. Auch die Rückwärtswahrscheinlichkeiten können induktiv gelöst werden mit:

$$\beta_t(j) = \begin{cases} 1 & \text{falls } t = T \\ \sum_{i=1}^{N} a_{ij} b_j(o_{t+1}) \beta_{t+1}(i) & \text{falls } t < T \end{cases} \tag{3.13}$$

Die Berechnung der Vorwärtswahrscheinlichkeiten stellen den Vorwärtsteil, die Berechnung der Rückwärtswahrscheinlichkeiten der Rückwärtsteil des *Vorwärts-Rückwärts-Algorithmus* dar.

Mithilfe der beiden Hilfsgrößen α und β läßt sich nun zur gegebenen Beobachtungssequenz O und Modell $\boldsymbol{\lambda}$ die Wahrscheinlichkeit $\gamma_t(j) = P(q_t = S_j|O,\boldsymbol{\lambda})$ beschreiben, zum Zeitpunkt t im Zustand S_j zu sein:

$$\gamma_t(j) = \frac{\alpha_t(j)\beta_t(j)}{\sum_{i=1}^{N} \alpha_t(i)\beta_t(i)} \tag{3.14}$$

Der Term γ bestimmt, mit welchem Anteil der Trainingsvektor o_t in die Schätzungen der HMM-Parameter des Zustandes S_j eingeht. Die HMM-Parameter werden nach dem Maximum-Likelihood Kriterium bestimmt, d.h. es werden diejenigen Parameter gesucht, welche die Trainingsstichprobe am besten erklären. Für einfache Dichtefunktionen kann die Maximum-Likelihood-Schätzung analytisch berechnet werden. Für die Lösung des Optimierungsproblems ist allerdings bisher keine analytische Lösung bekannt. Man behilft sich daher mit einem iterativen Schätzverfahren dem EM-Verfahren (Expectation-Modification), equivalent mit dem Baum-Welch-Verfahren [DLR77], das $P(O|\boldsymbol{\lambda})$ lokal maximiert. Mit der in [Rab90] beschriebenen Schätzprozedur werden die HMM-Parameter so modifiziert, daß sie die Trainingsdaten besser modellieren.

Der Umstand, daß der Baum-Welch-Algorithmus zur iterativen Schätzung der HMM-Parameter formuliert wurde, ist ein wesentlicher Grund dafür, daß die HMMs in der akustischen Modellierung zur dominanten Technologie avancierten.

3.2.2.2 Geeignete Modellierungseinheiten für Sprache

Als zu Beginn der automatischen Spracherkennung die Einzelworterkennung mit sehr kleinen Wortschätzen erforscht wurde, war das Wort die bevorzugte Einheit zur Modellierung von Sprache. Für jedes Einzelwort wurden ein oder mehrere Referenzmuster abgespeichert und durch Vergleiche der Referenzen zum Signal $\mathbf{X}$ dasjenige Einzelwort als gesprochen angenommen, dessen Referenzmuster die kleinste Distanz zum Signal hat. Die Vorteile liegen darin, daß das Wort die natürliche Einheit einer Sprache ist und Koartikulationseffekte innerhalb eines Wortes implizit mitmodelliert werden. Dadurch ist die Erkennungsleistung sehr gut. Daher verwenden Einzelworterkenner, wie sie etwa zur Kommandosteuerung eingesetzt werden, auch heute noch das Wort als Modellierungseinheit .

Als die Zahl der zu erkennenden Wörter anwuchs und die Anforderungen durch sprecherunabhängige Systeme anstiegen, war dieses Vorgehen zu speicher- und zeitintensiv. Außerdem wurde mit zunehmender Zahl der Wörter das Trainingsproblem dringlicher. Ganzwortmodelle benötigen zur zuverlässigen Schätzung viele Beispielmuster. Dies macht das gesamte System unflexibel gegen Erweiterungen des Vokabulars, weil für jedes Wort, das zum Erkennervokabular hinzugefügt wird, Trainingsmaterial gesammelt werden muß.

Silben
Um das Flexibilitäts- und Resourcenproblem zu bewältigen, wurde auf eine Zerlegung der Wörter in kleinere Einheiten zurückgegriffen. Die naheliegendste Lösung ist die Zerlegung in Silben. Silben sind linguistisch gut motiviert und ihre Zahl ist in den meisten Sprachen sehr begrenzt. In Silben kann das Betonungsmuster einer Sprache gut erfaßt werden, das in der Regel silbenbasiert ist. Außerdem erfassen Silben Koartikulationseffekte zwischen Phonemen. In vielen Tonsprachen wie beispielsweise Chinesisch konzentriert sich der Tonhöhenverlauf auf die Dauer einer Silbe. Silben sind daher für viele Sprachen sehr gut geeignete Einheiten. In Kapitel 5 wird am Beispiel der chinesischen Sprachen ein solcher Ansatz vorgeführt. Trotzdem haben sich Silben als Modellierungseinheit nicht durchgesetzt, weil es immer noch zuviele Einheiten sind, und damit das Trainingsproblem weiterhin bestehen bleibt.

Phoneme und Subphoneme
Phoneme sind ebenfalls linguistisch gut motiviert, wie bereits in Kapitel 2 ausführlich dargelegt wurde. Für die meisten Sprachen reichen zwischen 30 und 50 Einheiten zur Modellierung aus. Ihre Zahl liegt damit deutlich unter der von Silben. Im Chi-

nesischen beträgt das Verhältnis zwischen Silben und Phonemen beispielsweise 10:1, im Koreanischen dagegen ist das Verhältnis etwa 100:1. Phoneme führen somit zu einem sehr kompakten Inventar von Spracheinheiten und sind daher wesentlich besser zu trainieren als Silben. Zudem sind sie wesentlich flexibler als Wörter, denn mit einem einmal festgelegten Phoneminventar kann jedes neu zu erkennende Wort in das Erkennervokabular aufgenommen werden, ohne daß ein Trainingsmuster für dieses Wort benötigt wird. Die meisten heute gängigen Erkenner unterteilen Phoneme in drei Subphoneme, um die Dynamik innerhalb eines Phonems zu erfassen.

Phoneme als Modellierungseinheit der Sprache haben insbesondere für die multilinguale Spracherkennung eine große Bedeutung. Da das Lautinventar, das der Mensch zu produzieren in der Lage ist, sehr beschränkt ist, kann man davon ausgehen, daß viele Laute von mehreren Sprachen genutzt werden. Demzufolge könnte man durch eine Auswahl vieler Sprachen ein *universelles* Phoneminventar bilden, dessen Einheiten in der Lage sind, alle Wörter einer neuen, nicht in der Auswahl repräsentierten Sprache, zu modellieren. Phoneme erfüllen somit die Eigenschaften der Trainierbarkeit und Flexibilität und bieten die Möglichkeit der Transferierbarkeit auf neue Sprachen.

Polyphone
Den Nachteil der Phoneme gegenüber Silben, nämlich die fehlende Erfassung der Koartikulation, gleicht man in heutigen Systemen durch eine Kontextmodellierung wieder aus. Dazu betrachtet man ein Phonem im Kontext seiner angrenzenden Nachbarphoneme. Bezieht man in die Betrachtung den unmittelbar rechten und linken Nachbarn mit ein, spricht man von *Triphonen.* Dehnt man das Kontextfenster um zwei Phoneme nach links und rechts aus, spricht man von *Quintphonen.* Mit dem Begriff *Polyphon* bezeichnet man ein Phonem im Kontext unbestimmter Breite. Analog zur Modellierung von Subphonemen werden in den heute gängigen Erkennern *Subpolyphone* verwendet. Dabei werden die Subpolyphone eines Phonems nicht in Abhängigkeit der rechts und links angrenzenden Subphoneme modelliert, sondern in Abhängigkeit der Nachbarphoneme.

3.2.2.3 Generalisierte Subpolyphone

Wie in Abschnitt 5.4.3 zu sehen sein wird, kann die Anzahl der kontextabhängigen Modelle sehr groß werden. Je nach Sprache, Kontextbreite und Szenario geht die Modellzahl in die Hundertausende oder gar Millionen. Würde man für jedes der Modelle ein eigenes Codebook bereitstellen, stößt man neben dem Speicheraufwand schnell an die Grenzen der Trainierbarkeit dieser Modelle. Außerdem sind kontextabhängige Modelle mit wachsendem Kontext weniger generalisierungsfähig, d.h. bedingt durch die spezifischen Kontexte, mit denen die Modelle trainiert werden, besteht die Gefahr, daß diese Modelle auf neue Daten nicht mehr gut passen. Um einen möglichst guten Arbeitspunkt zwischen Generalisierungsfähigkeit und Modellgenauigkeit zu erzielen, werden die kontextabhängigen Modelle in den heute gängigen Erkennersystemen durch Ballung geeignet zusammengefaßt.

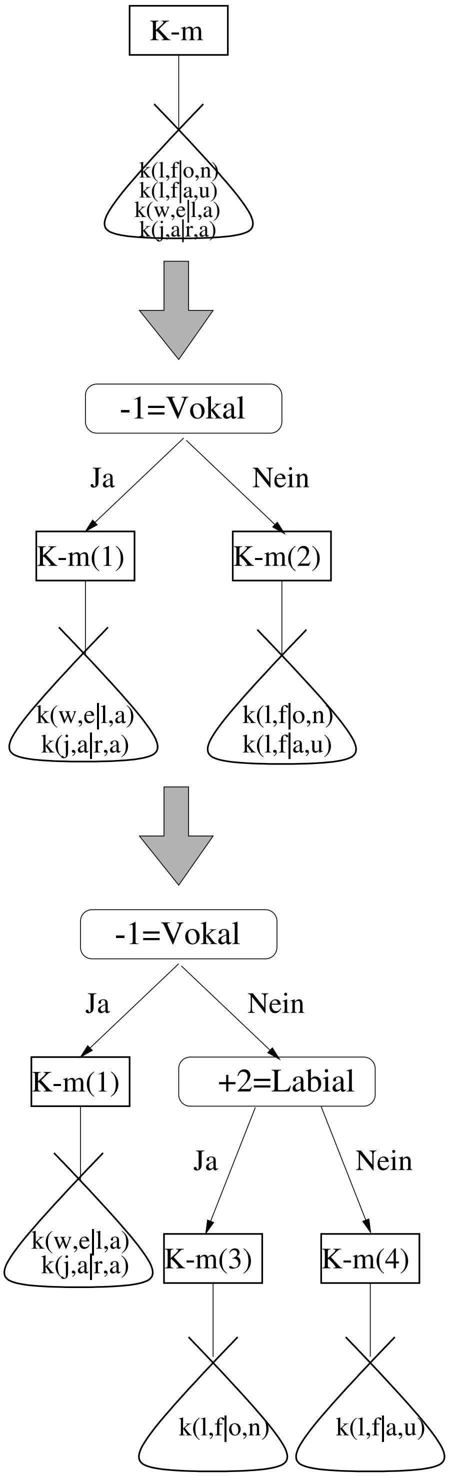
Ballungsalgorithmen lassen sich danach unterscheiden, welche Grundeinheiten zusammengeballt werden, ob es sich um einen agglomerativen (bottom-up) oder divisiven (top-down) Vorgang handelt und welche Distanzmaße zur Ballung verwendet werden. Lee [Lee88] führte als einer der ersten geballte kontextabhängige Modelle ein. Er verwendete als zu ballende Grundeinheiten Triphone und setzte ein agglomeratives Ballungsverfahren ein. Als Distanzmaß verwendete er den Anstieg der Entropie, der durch das Zusammenballen zweier Modelle entsteht.

Die agglomerative Ballung hat zwei wesentliche Nachteile. Erstens wächst die Zahl möglicher Ballungen quadratisch mit der Zahl der Modelle. Für Triphone auf einem eingeschränkten Szenario war dieses Verfahren für Lee noch anwendbar, aber sobald die Menge zu ballender Kontexte sehr groß wird, steigt der Berechnungsaufwand ins Unermeßliche. Zweitens besteht das Problem, daß nur diejenigen Triphone einer geeigneten Klasse zugeordnet werden können, die im Training gesehen wurden. Für ungesehene Kontexte muß man sich durch eine suboptimale Modellierung anhand kontextunabhängiger Restklassenmodelle oder willkürlicher Zuordnungen auf andere Modelle behelfen.

Beide Nachteile werden durch das divisive Ballen behoben. In JRTk Erkenner werden dazu als Grundeinheiten generalisierte Subpolyphone verwendet, wobei der Kontext der zu generalisierenden Subpolyphone bis in die angrenzenden Wörter (cross-word polyphones) reichen kann. Als Distanzmaß wird die Distanz zwischen den Entropien der Mixturgewichteverteilungen der einzelnen Modelle definiert. Aus Aufwandsgründen werden nur gleichartige Zustände desselben Phonems in verschiedenen Kontexten geballt. Das Ballungsverfahren ist Entscheidungsbaum-basiert. Dazu wird ein Fragenkatalog erstellt, der aus phonetisch motivierten Fragen über den Kontext von Phonemen besteht. Aus diesem Katalog wird in jedem Ballungsschritt jeweils diejenige Frage ausgewählt, bei deren Anwendung der Entropieverlust durch Aufteilen des Knotens in zwei Nachfolgerknoten am größten wird. Abbildung 3.7 zeigt die Aufteilung eines Knoten an einem Beispiel.

Da das Entropiemaß nicht geeignet ist, den Ballungsvorgang automatisch abzubrechen, wird vor der Ballung eine Anzahl zu erreichender Ballungsknoten definiert, bei deren Erreichen der Ballungsvorgang endet. Daneben wird in jedem Aufspaltungsschritt die Zahl der möglichen Spaltungen dadurch beschränkt, daß nur Knoten entstehen dürfen, auf die eine festgelegte Zahl von Trainingsdaten entfällt. Eine detaillierte Beschreibung zur Kontextmodellierung in JRTk ist in [Rog97] zu finden.

Kontextabhängige Cross-Word-Triphone haben in der Spracherkennung zu sehr großen Leistungsverbesserungen geführt [Lee88]. Die meisten gängigen Erkenner verwenden heute breitere Kontexte als Triphone. Finke und Rogina untersuchten in [FR97] für mehrere Sprachen und Sprechstile den Effekt breiterer Kontextmodellierungen und stellten auf Englisch eine Verbesserung von 5% durch Septphone statt Triphone auf dem WSJ und von 8% durch Quintphone statt Triphone auf dem



Diese Abbildung veranschaulicht zwei Schritte des divisiven Ballungsprozesses, in dessen Verlauf der erste Teil eines Entscheidungsbaums für das Subphonem K-m entsteht. **Ausgangspunkt:** Zu Beginn des Ballens hängen am Blatt K-m des Entscheidungsbaumes alle Quintphone, die beim Durchsuchen der gesamten Datenbasis als Kontexte von K-m der Breite ±2 gefunden wurden. Hier im Beispiel seien das die 4 Quintphone k(l,f|o,n) k(l,f|a,u) k(w,e|l,a) k(j,a|r,a). **Schritt 1:** Aus einem vorher festgelegten Fragenkatalog (vgl. Abschnitt 5.3.9.3) wird diejenige Frage herausgesucht, die den Knoten K-m in zwei Nachfolgerknoten so aufspaltet, daß der Entropieverlust am größten wird. Hier sei das die Frage danach, ob der linke Kontext von K ein Vokal ist (-1 =Vokal). Entsprechend werden die Kontexte k(l,f|o,n) und k(l,f|a,u) in den NEIN-Nachfolgerknoten und k(w,e|l,a) und k(j,a|r,a) in den JA-Nachfolgerknoten aufgespaltet. **Schritt 2:** Die nächste ausgewählte Frage des NEIN-Knotens ist diejenige, ob der übernächste rechte Kontext ein labialer Konsonant ist ($+2$ = Labial). Der Kontext k(l,f|a,u) kommt in den NEIN-Zweig, der Kontext k(l,f|o,n) kommt in den JA-Zweig. **Dekodierung:** Treten im Dekodierungsvorgang ungesehene Quintphone auf, werden sie durch dasjenige akutische Modell modelliert, das sich durch Traversieren des Entscheidungsbaumes ergibt.

Abbildung 3.7: Entstehung eines Entscheidungsbaumen auf Quintphonen

Switchboard Korpus fest. Der Trend zu einer immer feineren Kontextmodellierung ist aber kontraproduktiv für die multilinguale Spracherkennung. Denn wie in Kapitel 7 gezeigt werden wird, sind die Kontexte eines Phonems sehr stark von der Sprache abhängig, weil sie von den phonotaktischen Regeln und Wörtern der jeweiligen Sprache abhängen. Um auch in der multilingualen Spracherkennung von der Kontextmodellierung profitieren zu können, sind daher neue Methoden notwendig, die im Kapitel 6 und 7 vorgestellt werden.

3.2.3 Sprachmodellierung

Die Aufgabe der Sprachmodellierung besteht darin, die a-priori Wahrscheinlichkeit $P(W)$, für eine gegebene Wortsequenz $W = w_1, w_2, \ldots, w_n$ anzugeben. Diese Wahrscheinlichkeit ist unabhängig vom akustischen Signal $\mathbf{X}$, daher kann ihre Berechnung von der akustischen Modellierung entkoppelt werden. Man unterscheidet zwei Verfahren zur Berechnung von $P(W)$: die linguistisch und die statistisch motivierte Modellierung. Ziel des linguistischen Verfahrens ist es, die syntaktische (und semantische) Struktur von Sprache nachzuempfinden und aus diesem Wissen die Wahrscheinlichkeiten von Wortfolgen abzuleiten. Beim statistischen Verfahren werden dagegen Wahrscheinlichkeiten für Wortübergänge direkt von großen Textkorpora gelernt. In der automatischen Spracherkennung hat sich bislang das statistisch basierte gegenüber dem linguistischen Verfahren als erfolgreicher erwiesen. Zwar ermöglicht das linguistische Verfahren eine Modellierung der syntaktischen Struktur von Sprache, die beim statistischen Verfahren nur eingeschränkt kodiert werden kann, allerdings scheinen die Vorteile der statistischen Variante bisher zu überwiegen, die in der besseren Modellierung semantischer Beziehungen liegen. Neuere Arbeiten, wie beispielsweise die von [JC99], versuchen die Vorteile beider Verfahren miteinander zu kombinieren, so daß sich heute beide Verfahren aufeinander zu bewegen. Da derzeit die statistische Sprachmodellierung noch bei nahezu allen State-of-the-art Erkennern die vorherrschende Technik ist, wird die Beschreibung darauf beschränkt.

N-Gramm-Modelle

Die Wahrscheinlichkeit $P(W)$ einer Wortfolge $W = w_1 w_2 \ldots w_n$ läßt sich angeben als:

$$P(W) = \prod_{i=1}^{n} P(w_i | w_1 w_2 \ldots w_{i-1}) \tag{3.15}$$

Das Problem bei der Berechnung von $P(w_i|w_1 w_2 \ldots w_{i-1})$ liegt in der gewaltigen Anzahl möglicher Wortketten. Bei einer Länge l auf einem Vokabular der Größe $|\mathcal{V}|$ beträgt sie $|\mathcal{V}|^l$. Heutige gängige Vokabulargrößen für Diktieranwendungen liegen im Bereich von 65.000 Wörtern. Die mittlere Länge gesprochener Sätze in Wörtern liegt beim Diktieren je nach Sprache und Segmentierung in der Größenordnung zwischen 10 und 30. Selbst wenn das Resultat von $65000^{10} = 13^{48}$ die eigentliche Zahl vorkommender Wortsequenzen aufgrund semantischer und syntaktischer Einschränkungen überschätzt, verdeutlicht sie, daß sich ein enormes Schätzproblem ergäbe, wenn man die Wahrscheinlichkeiten von 10-Wortketten angeben wollte. Darüber hinaus trifft es für keine Sprache zu, daß im Satzgefüge das i-te Wort von der kompletten Folge aller $i-1$ Vorgängerwörter abhängt. Aus diesen Gründen bildet man die Historie $W_{i-1} = w_1, w_2, \ldots w_{i-1}$ auf eine Äquivalenzklasse ab, die von der Funktion $\phi(h)$ bestimmt wird:

$$P(W) \cong \prod_{i=1}^{n} P(w_i|\phi(W_{i-1})) \tag{3.16}$$

Die Aufgabe der Sprachmodellierung besteht nun darin, eine geeignete Funktion ϕ zu finden und das Schätzproblem von $P(w_i|\phi(W_{i-1}))$ zu lösen. Von [BJM90] wurde die N-Gramm-Äquivalenzklasse $\phi(W_{i-1}) = w_{i-N+1}, w_{i-N+2}, \ldots, w_{i-1}$ vorgeschlagen, in der alle Kontexte in eine gemeinsame Klasse zusammengefaßt werden, sofern deren letzten $N-1$ Wörter identisch sind, d.h. man macht die Annahme, daß das i-te Wort nur von seinen $N-1$ Vorgängern abhängt. In der Praxis hat sich die Einschränkung auf $N=3$ als sehr leistungsfähig und gerade noch handhabbar erwiesen, man spricht dann von 3-Grammen oder *Trigrammen*.

Trigramme werden auf der Basis großer Textkorpora geschätzt, indem die Auftrittshäufigkeiten $\text{Count}_{N-Gramm}$ der entsprechenden Worttupel ausgezählt werden:

$$P(w_i|w_{i-1}, w_{i-2}) = \frac{\text{Count}_{Trigramm}(w_{i-2}, w_{i-1}, w_i)}{\text{Count}_{Bigramm}(w_{i-2}, w_{i-1})} \tag{3.17}$$

Discounting und Backing-Off Modellierung

Trotz der Einschränkung auf Trigramme besteht aufgrund des Datenmangels das Problem der akkuraten Schätzung der Wahrscheinlichkeiten. Für obiges Beispiel ergibt sich bei $l = 3$ eine potentielle Menge von $65000^3 = 275$ Billionen Trigrammen. Es ist offensichtlich, daß man auf einem Trainingskorpus nur einen kleinen Ausschnitt aller möglichen Trigramme beobachtet kann. Um trotzdem eine gute Schätzung für selten oder nie beobachtete Trigramme zu erhalten, wendet man die Methode des *Discounting* [Kat87] und des *Backing-Off* [NEK94] an. Beim Discounting wird von häufig gesehenen Trigrammen ein Teil der Masse aller Wahrscheinlichkeiten abgezogen und gleichmäßig unter den seltenen Trigramme verteilt. Backing-Off wird angewendet, wenn die Zahl gesehener Trigramme zu gering ist, um überhaupt eine Schätzung der Wahrscheinlichkeit vorzunehmen. Dann fällt man auf die unspezifischeren aber zahlreicher beobachteten Bigramme zurück und skaliert deren Wahrscheinlichkeit mit einer Korrekturfunktion *bo*, die garantiert, daß sich die Summe der Trigramm-Wahrscheinlichkeiten zu 1 aufaddieren:

$$P(w_3|w_2, w_1) = P(w_3|w_2) \cdot bo(w_1, w_2) \tag{3.18}$$

Trigramm-Sprachmodelle sind aufgrund ihrer beschränkten Reichweite nicht für alle Sprachen gleich gut geeignet. In der englischen Sprache, auf der die Trigramme erstmals eingesetzt wurden, ist bedingt durch die geringe Flexion, die Wortstellung bedeutungsunterscheidend. Wörter die sich aufeinander beziehen, stehen in der Regel eng beieinander. Dagegen sind Sprachen mit stark flektierendem oder agglutinierendem Sprachbau in ihrer Wortstellung sehr viel freier und damit weniger gut durch Trigramme abzudecken. So reicht beispielsweise im Deutschen die Reichweite von Trigrammen nicht aus, um zusammengehörige Wortformen einander zuzuordnen, wenn deren Bestandteile einen Satzteil umschließen. Die hervorgehobenen Bestandteile des letzten Satzes sollen dies verdeutlichen. So wurden denn auch zahlreiche Alternativen untersucht, wie beispielweise variable N-Gramme [DB95] oder baumbasierte Modelle [BBdSM90]. Alles in allem hat die Trigramm-Sprachmodellierung aber allen Verbesserungsversuchen der letzten 20 Jahre getrotzt [JC99] und ist heute immer noch die in der Spracherkennung dominante Modellierungsart.

Klassenbasierte N-Gramm-Modelle

Wie bereits in Kapitel 2.4.1.3 angedeutet, erzwingen Sprachen mit sehr langen Wörtern oder Wortphrasen eine Zerlegung in kürzere Einheiten. Dadurch wird die Reichweite des N-Gramm-Sprachmodells stark eingeschränkt. Die Möglichkeit der Vorausschau in nachfolgende lexikalische Einheiten nimmt mit der Anzahl der Zerlegungen ab. Dieser Effekt kann wegen des Schätzproblems nicht durch eine Erweiterung der Historie auf $N > 3$ ausgeglichen werden.

Zur Lösung des Problems werden unter anderem Ansätze zur Multigramm-Sprachmodellierung vorgeschlagen, bei denen Einheiten miteinander verschmolzen werden, die die Vorhersage des Nachfolgerwortes nach einem zu definierenden Kriterium maximieren. Eine knappe dazu Einführung findet sich in [RBW96]. Daneben beschäftigen sich Forschungsansätze mit der Verbesserung der Backoff-Strategien. Eine Idee ist die Nutzung weniger spezifischer Äquivalenzklassen, die sich robuster schätzen lassen und daher als Backoff Modell verläßlicher sind. Auf dieser Idee basiert die *klassenbedingte Sprachmodellierung*, bei der zur Schätzung eines Trigramms in Gleichung 3.18 zusätzlich die Wahrscheinlichkeit der Klassenzugehörigkeit c des direkten Vorgängerwortes herangezogen wird (vgl. [MA94]):

$$P(w_3|w_1w_2) := P(w_3|w_1w_2c_2) = P_{c_2}(w_3|w_1w_2) \tag{3.19}$$

Aus einer Trigramm-Modellierung wird so eine Form der 4-Gramm-Modellierung, auf das dieselben Backoff Verfahren angewendet werden können. Der Vorteil der klassenbedingten Sprachmodellierung liegt in der Möglichkeit der robusteren Schätzung der Klassenwahrscheinlichkeit P_c, der Nachteil in der schwachen Aussagekraft des Backoff-Bigramms $P(w_3|c_2)$. Denn w_3 läßt sich nur sehr schlecht aus dem Wissen der Wortklasse c_2 vorhersagen. Zur besseren Modellierung wird die Berechnungsvorschrift aus Gleichung 3.19 abgeändert zu:

$$P(w_3|w_1w_2) := P(w_3|c_3, w_2, w_1) \quad \text{mit } P(c_3) = P(c_3|c_2, w_2, w_1) \tag{3.20}$$

Beim Backoff kann nun statt auf das Bigramm $P(w_3|c_2)$ auf das aussagekräftigere $P(c_3|c_2)$ zurückgegriffen werden. Aus der vorherigen Modellierung mit Worttrigrammen wird so eine verzahnte Modellierung von Wort- und Klassenvorhersage. Die zu lösende Aufgabe besteht nun in der Frage nach der Abbildung der lexikalischen Einheiten w_i auf die Klassen $c_i = c(w_i)$. Die Lösung dieser Abbildung ist sprachenspezifisch, in der japanischen Sprachen bieten sich beispielsweise Klassen von Mora-Silben[2] an [TR97], in agglutinierenden Sprachen wären Suffixklassen angebracht.

Gewichtung von Sprachmodell zum akustischen Modell

In der Suche (vgl. nächsten Abschnitt) werden zur Berechnung der wahrscheinlichsten Wortkette die Emissionswahrscheinlichkeiten der HMMs mit den Wortübergangswahrscheinlichkeiten des Sprachmodells kombiniert. In der Praxis weichen die Mittelwerte und Varianzen der Wahrscheinlichkeiten beider Modelle so stark ab, daß die Suche ohne eine Korrektur von einem der beiden Terme dominiert würde. Diese Korrektur geschieht durch den Parameter z, mit dem das Sprachmodell relativ zum akustischen Modell gewichtet wird.

Schließlich wird noch ein zweiter Parameter, die Wortübergangsstrafe q, eingeführt, welcher die unterschiedliche Länge betrachteter Wortfolgen W normiert, die sonst an keiner Stelle berücksichtigt würde. Die beiden Parameter z und q werden mit einer Kreuzvalidierungsmenge per Hand eingestellt.

$$P(W|\mathbf{X}) = \frac{P(\mathbf{X}|W) \cdot P(W)}{P(\mathbf{X})} \xrightarrow{z,q} \frac{P(\mathbf{X}|W) \cdot P(W)^z \cdot q^{|W|}}{P(\mathbf{X})} \tag{3.21}$$

Perplexität

Die Perplexität ist ein Maß, mit der die Komplexität der Erkennungsaufgabe angegeben wird. Die Perplexität wird aus der Entropie H eines stochastischen Prozesses abgeleitet. H beschreibt den Informationsgehalt, der von einer Quelle generiert wird. Im Fall der Spracherkennung entspricht diese Quelle einem Sprecher, der eine Wortsequenz erzeugt. Die Entropie der Wortsequenz kann dann mittels der Faktorisierung mit N-Grammen approximiert werden zu:

$$H(W) = -\sum P(W) \cdot \log P(W) \approx -\frac{1}{n} \sum_{i=1}^{n} \log P(w_i|w_{i+N-1}, w_{i+N-2}, \ldots, w_{i-1}) := LP \tag{3.22}$$

Die Perplexität ist definiert als $PP = 2^{LP}$ und gibt anschaulich gesprochen den mittleren Verzweigungsgrad eines Sprachmodells an. Die Perplexität nimmt ihren kleinsten Wert $PP = 1$ an, wenn jedes Wort mit Sicherheit vorausgesagt wird. Der größte Wert, den die Perplexität annimmt, entspricht der Größe des Vokabulars

[2] Eine Mora-Silbe besteht aus einer CV-Folge mit Ausnahme von [N], das einzeln stehen kann.

und tritt genau dann ein, wenn alle Worte gleichwahrscheinlich sind, so daß das Sprachmodell keine Information über das jeweils nachfolgende Wort hat.

Die Perplexität wird als Gütemaß herangezogen, wenn man sich zwischen mehreren Sprachmodellen entscheiden muß. Gewünscht ist dasjenige Sprachmodell, dessen Vorhersage der Wirklichkeit am nächsten kommt. Da die „Wirklichkeit“ im Vorfeld nicht bekannt ist, verwendet man einen möglichst repräsentativen Ausschnitt in Form eines Validierungstextes und wählt dasjenige Sprachmodell, das den Validierungstext mit der höchsten Wahrscheinlichkeit vorhersagt. Da ein Validierungstext nicht die gesamte Wirklichkeit repräsentieren kann, sind direkte Vergleiche zwischen Sprachmodellen nur auf identischen Texten sinnvoll. Zum Vergleich verschiedener Wortsegmentierungen ist es allerdings wünschenswert, die Perplexitäten auf verschieden segmentierten Validierungstexten zu vergleichen. Dazu wird die normierte Perplexität $PP^{rel} = PP^{\frac{n'}{n}}$ berechnet [TR97], wobei n die Größe des ursprünglichen Validierungstextes ist, und n' die Größe des resegmentierten Validierungstextes. PP^{rel} ist somit unabhängig von der zugrundeliegenden Segmentierung der Worteinheiten eines Textes.

3.3 Bestimmung der Erkennungsleistung

Zur Evaluation eines Spracherkenners wird eine gesprochene Äußerung dekodiert, was in einer besten Worthypothese und optional einer Menge in Frage kommender Alternativen resultiert. Zur Bewertung dieser Hypothese ist ein Fehlermaß erforderlich. Anhand des mit diesem Fehlermaß gemessenen Fehlers können die Erkennungsresultate verschiedener Systeme miteinander verglichen werden. Der letzte Abschnitt dieses Kapitels faßt die Probleme zusammen, die sich beim Vergleich von Erkennungsergebnissen verschiedener Sprachen ergeben.

3.3.1 Dekodierung kontinuierlich gesprochener Sprache

Das Problem, auf möglichst effiziente Weise diejenige Wortsequenz W^* zu finden, die den Ausdruck $P(\mathbf{X}|W) \cdot P(W)$ maximiert, läßt sich wie beschrieben mit dem Viterbi-Algorithmus lösen. Allerdings muß bei kontinuierlich gesprochener Sprache statt eines Einzelwortes nun eine kontinuierlich gesprochene Wortsequenz W dekodiert werden. Würde man das Maximierungsproblem für W nach dem vorgestellten Prinzip der Einzelworterkennung lösen wollen, dann müßte man für alle denkbaren Wortsequenzen die Produktionswahrscheinlichkeiten berechnen. Das ist in keinem vertretbaren Zeitaufwand zu bewältigen. Wenn andererseits die Segmentierung der gesprochenen Äußerung in Einzelwörtern bekannt wäre, dann würde der Viterbi-Algorithmus in jedem dieser Abschnitte das am besten passende Wort finden und so die wahrscheinlichste Wortsequenz berechnen. Die Erweiterung des

Viterbi-Algorithmus, in der die Segmentierung und die Einzelwortdekodierung gemeinsam gelöst werden, ist der One-Stage-Dynamic-Time-Warping Algorithmus [SC90, Ney90]. Dabei wird statt eines kompletten Satzmodells eine Sequenz beliebig koppelbarer Wortmodelle aufgebaut, wie auf der linken Seite in Abbildung 3.8 für drei Wortmodelle dargestellt. Innerhalb der einzelnen Wortmodelle erfolgt die Berechnung des besten Pfades durch das Wort nach dem Viterbi-Algorithmus. Der Suchprozeß mit den Viterbi-Pfaden innerhalb eines Wortmodelles ist auf der rechten Seite in Abbildung 3.8 illustriert. An den Wortenden kann nun zusätzlich in den Anfangszustand eines neuen Wortmodells gesprungen werden. Da zu jedem Zeitpunkt ein neues Wort beginnen kann, gibt es für jeden Zustand einen möglichen Übergang von einem Wortendezustand zu allen Wortbeginnzuständen. Beim Sprung in ein neues Wort wird in der Regel das beschriebene statistische Sprachmodell mit einberechnet. Die Sequenzen von Einzelwörtern werden auf diese Weise nicht nur durch den akustisch wahrscheinlichsten Pfad von einem zum anderen Wort, sondern auch durch den vorausgegangenen textuellen Kontext gesteuert.

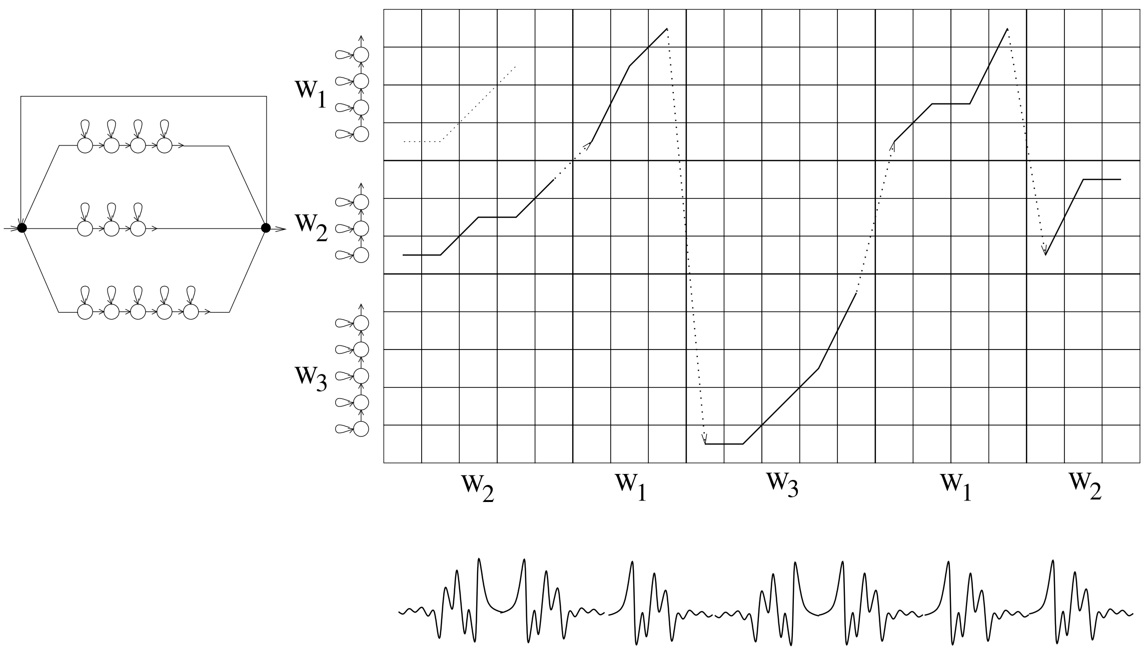


Abbildung 3.8: Verkettung von Wortmodellen (links) und Viterbi-Pfade durch die Suchmatrix (rechts)

Bei der Berechnung des besten Pfades durch die Suchmatrix ist die Bewertung mancher Suchpfade so viel schlechter im Vergleich zu anderen, daß es sehr unwahrscheinlich ist, daß diese Suchpfade als beste Suchpfade enden. Um die Zahl aktiver Suchpfade zu verringern, und damit die Berechnungskosten zu reduzieren, werden die aussichtslosen partielle Suchpfade frühzeitig abgeschnitten. Diesen Vorgang nennt man Pruning (*engl. to prune*). In [Wos98] sind neben einem guten Überblick die in JRTk eingesetzten Pruning-Heuristiken und Techniken ausführlich beschrieben.

3.3.2 N-Besten-Listen, Worthypothesengraphen

Aus der Rückverfolgung des besten Suchpfades entsteht eine Sequenz von Wörtern, die als Hypothese vom Erkenner ausgegeben wird. Werden die N-besten Suchpfade ausgewertet, entsteht eine N-besten Liste von Hypothesen. Diese können sehr kompakt in Form eines Worthypothesengraphen (WHG) dargestellt werden. Die N-besten Listen oder WHGs bieten die Möglichkeit, daraus neue beste Hypothesen zu berechnen, indem beispielsweise höheres Wissen, weitreichende Sprachmodelle oder andere akustische Modelle anwendet werden. Durch die Vorgabe von z, p-Paaren (vgl. Abschnitt 3.2.3) lassen sich darüber hinaus die Parameter des Sprachmodells manuell feineinstellen.

Ebenfalls auf einem WHG basiert der Hypothesis Driven Lexical Adaptation Ansatz (HDLA) von Geutner [GFS97, GFW99]. Dieser Ansatz erlaubt die virtuelle Erweiterung des Erkennervokabulars und bietet somit eine Lösung des OOV-Problems für stark flektierende oder agglutinierende Sprachen. HDLA wurde auf den GlobalPhone-Sprachen Kroatisch und Türkisch erprobt. Der Ansatz und die damit erzielten Resultate werden in Abschnitt 5.5.3 beschrieben.

3.3.3 Messung von Fehlerraten

Um die Fehler der Erkennungsausgabe zu messen, stellt man den Referenzsatz und die erkannte Satzhypothese einander gegenüber und berechnet die minimale Editierdistanz zwischen Referenz und Hypothese. Man unterscheidet die Fehler *Verwechslungen* (N_{sub}), *Auslassungen* (N_{del}) und *Einfügungen* (N_{ins}) und berechnet anhand dieser drei Fehlerklassen das Gütekriterium *Wortfehlerrate (WE)*

$$WE = 100 \cdot \frac{N_{sub} + N_{ins} + N_{del}}{N} \tag{3.23}$$

wobei N die Gesamtzahl der Wörter des Referenzsatzes ist.

Dieses Gütemaß eignet sich gut, um die Erkennungsleistungen innerhalb einer Sprache zu vergleichen. Weil sich die Fehlerrate aber auf Wörter bezieht, ist das Maß nicht immer geeignet, um die Erkennungsleistung zwischen *verschiedenen* Sprachen zu vergleichen (siehe nächster Abschnitt). Außerdem gibt es, wie bereits in Abschnitt 2.4.2.2 beschrieben, in einigen Sprachen überhaupt kein Wortkonzept.

In Sprachen ohne Wortkonzept oder in stark agglutinierenden Sprachen haben sich daher neben dem Wortfehlermaß andere Fehlermaße etabliert, die auf verschieden langen sprachlichen Einheiten wie Silben oder Phoneme basieren. Im folgende sei dieses Konzept am Beispiel der koreanischen Sprache erläutert. Im Koreanischen verwendet man die *Eojeol* Fehlerrate (EE)[3], die „Character" Fehlerrate (CE)[4] und die Phonemfehlerrate (PE). In den folgenden Beispielen zeigt die obere Zeile die Referenz einer Äußerung, die unter Zeile gibt die Hypothese wieder. Zunächst wird die Berechnung der Eojeol Fehlerrate gezeigt:

동무는	언제	아버님에게		편지를	씁니까
동무는		아버님과	어머님에게	편지를	씁니까

Das Eojeol 동무는 wird korrekt erkannt. 언제 ist ein Auslassungsfehler, weil es nicht in der Hypothese erscheint. 아버님에게 wird fälschlicherweise mit 아버님과 verwechselt, ein Subtitutionsfehler also. 어머님에게 ist ein Einfügungsfehler, da es hypothetisiert wird, obwohl es in der Referenz nicht auftaucht. Die verbleibenden zwei Eojeols sind korrekt erkannt worden. Die Zahl N der insgesamt gesprochenen Eojeols beträgt 5, es ergibt sich insgesamt eine Eojeol Fehlerrate von $EE = 100 \cdot \frac{1+1+1}{5}\% = 60\%$.

Um die Character Fehlerrate zu berechnen, bricht man jedes Eojeol in seine Silbenkomponenten auf und berechnet die Fehlerrate auf den sich ergebenden Sequenzen:

동 무 는 언 제 아 버 님   에 게 편 지 를 씁 니 까

동 무 는   아 버 님 과 어 머 님 에 게 편 지 를 씁 니 까

Insgesamt werden 14 Silben richtig erkannt, zwei sind ausgelassen, vier werden fälschlich eingefügt. Damit ergibt sich eine Character Fehlerrate von $CE = 100 \cdot \frac{0+2+4}{16}\% = 37.5\%$.

Zuletzt kann man noch die Phonemerkennungsrate berechnen. Dazu wird von jedem Eojeol/Character die Phonemsequenz aus dem Aussprachelexikon nachgeschlagen. Diese Art der Berechnung der Phonemerkennungsleisung wird hier als gebundene Phonemerkennung bezeichnet weil die Phonemerkennung durch die Vorgabe der Wörter an Wortgrenzen gebunden ist .Im Beispiel resultiert daraus eine Phonem Fehlerrate von: $PE = 100 \cdot \frac{0+4+9}{36}\% = 36.11\%$.

D O NG M U N EU N EO N J E A B EO N I M   E G E . . .

D O NG M U N EU N   A B EO N I M G O AE O M EO N I M E G E . . .

. . . Ph iEO N J I R EU L SS EU M N I GG A

. . . Ph iEO N J I R EU L SS EU M N I GG A

[3]Eojeol = koreanische Wortphrasen
[4]Character = koreanische Silbe

Um die Wortfehlerrate des Koreanischen mit der im Englischen in Bezug zu setzen, hat sich in der Literatur die Approximation $WE \approx \frac{EE+CE}{2}$ etabliert.

Falls die Worteinheiten zweier Sprachen einander nicht entsprechen, müssen die Fehlerraten zueinander in Bezug gesetzt werden. Es gibt allerdings derzeit kaum Sprachpaare, für die eine Approximation der Fehlerraten wie zwischen Koreanisch und Englisch definiert ist. Eine zuverlässige Methode, die Erkennungsraten verschiedener Sprachen miteinander zu vergleichen, bietet die Verwendung einer gemeinsamen Untereinheit als Berechnungseinheit. Dazu eignen sich Fehlerrate auf Phonembasis. Zur Berechnung der Phonemfehlerraten wird entweder, wie oben vorgeführt, jedes hypothetisierte Wort in seine Phonemsequenz zerlegt oder es werden Phonemhypothesen von Erkennern dekodiert, deren Vokabular nicht aus Wörtern sondern aus Phonemen besteht. Diese Erkennung wird als frei laufende Phonemerkennung bezeichnet.

3.3.4 Vergleich zwischen Sprachen

Beim Vergleich der beschriebenen Phonemerkennungsrate sind die Effekte unterschiedlicher Wortlängen zwischen Sprachen bereinigt. Sprachenspezifisch inhärenten Schwierigkeiten einer Sprache schlagen sich nach wie vor in der Hypothese nieder.

Die Dimensionen, die das anfänglich vorgestellte Urknall-Modell in Abbildung 3.1 aufspannen, veranschaulichen die komplexen Anforderungen an einen Spracherkenner. Der Schwierigkeitsgrad einer solchen Anforderung beeinflußt natürlich die Erkennungsresultate. Heutzutage liegen die Leistungen von sprecherabhängigen Einzelworterkennern in ruhigen Umgebungen unter 1% Wortfehlerrate. Dagegen liegen die Erkennungsfehler einer sprecherunabhängigen Erkennung von Umgangssprache, die auf öffentlichen Plätzen aufgenommen wurde, im Bereich von 40%. Kein Forscher käme auf den Gedanken, diese beiden Leistung zueinander in Beziehung zu setzen, weil sich die Systeme in den vier Dimensionen Umgebung, Sprecher, Anwendung und Sprechstil voneinander unterscheiden. Nach Auffassung der Autorin ist Sprache eine weitere Dimension, die den direkten Vergleich von Erkennungsergebnissen erschwert. Selbst wenn die übrigen Dimensionen konstant gehalten werden, unterscheiden sich Sprachen untereinander durch die folgenden Faktoren:

Lautstruktur PHONEMINVENTAR: Einige Sprachen wie etwa Türkisch oder Kroatisch haben ein sehr simples Phoneminventar, andere dagegen sind komplexer wie das Portugiesische oder verwenden Toneme wie das Chinesische (siehe Abschnitt 5.4.2). PHONOLOGIE: Manche Sprachen haben eine Silbenstruktur, die sich sehr klar in der phonetischen und phonologischen Struktur ausdrückt, wie etwa die Mora-Silben im Japanischen. Andere Sprachen wie Deutsch lassen dagegen zahlreiche Konsonantencluster (siehe Abschnitt 5.4.2) zu, die schwerer zu erkennen sind, die Anzahl an Triphonen erhöht (vgl. Abschnitt 5.4.3) und in höheren Phonemperplexitäten (vgl. Abschnitt 5.4.2) resultieren.

Aussprachewörterbuch GRAPHEM-ZU-PHONEM RELATION: Sprachen unterscheiden sich sehr stark hinsichtlich ihrer Graphem-zu-Phonem Beziehung (siehe Abschnitt 5.4.1). In Sprachen, deren Aussprache sehr eng an die Orthographie gebunden ist und keine Ausnahmen zuläßt, ist die Wahrscheinlichkeit höher, daß das Aussprachewörterbuch wenig Fehler enthält. Da das Aussprachewörterbuch eine zentrale Quelle der Spracherkennung ist, wirkt sich dies auf die Erkennungsleistung aus. VERWECHSELBARKEIT: Manche Sprachen haben sehr viele kurze und verwechselbare Funktionswörter, die häufig gesprochen und oft schlecht artikuliert werden. Eine Fehleranalyse von [Rog97] zeigte, daß fast die Hälfte aller Fehler eines englischen Diktiererkenners auf die 6 Funktionswörter *the, a, and, of, that,* und *in* entfallen. Andere Sprachen dagegen haben überwiegend längere, wenig verwechselbare Worteinheiten (siehe Abschnitt 5.4.4). MORPHOLOGIE: flektionsreiche Sprachen, deren Wörter sich nur in leicht zu verwechselnden Endungen unterscheiden, haben trotz ausreichender Wortlänge eine hohe Wortverwechselbarkeit (Beispiel: deutsche Endungen -en -em). HOMOPHONE: Die Anzahl gleichlautender Aussprachen bei unterschiedlicher Wortform ist sehr unterschiedlich. Französisch ist ein Beispiel für eine Sprache, in der sehr viele Homophone vorkommen (*ai, aie, aies, ait, aient, hais, haie, es, est* werden allesamt /ɛ/ gesprochen). Die Art und Weise, wie und ob Homophonfehler gezählt werden hat daher einen signifikanten Einfluß auf den Vergleich.

Vokabular VOKABULARGRÖSSE: Sprachen unterscheiden sich hinsichtlich ihrer Kompaktheit, was bedeutet, daß sie unterschiedlich viele laufende Wörter aber auch unterschiedlich große Vokabularien benötigen, um den identischen Sachverhalt auszudrücken (siehe Abschnitt 5.4.5). VOKABULARWACHSTUM: Bedingt durch den Sprachbau und die Segmentierung sind die Wortlängen sprachenspezifisch sehr unterschiedlich (siehe Abschnitt 5.4.4). Dies resultiert in sehr verschiedenen Vokabulargrößen, aber insbesondere unterschiedlichem Vokabularwachstum. Daraus resultieren für gleichgroße Vokabularien unterschiedlich hohe OOV-Raten (siehe Abschnitt 5.4.4). Ein OOV-Wort kann im Mittel 1.5-2 Folgefehler nach sich ziehen. SCHREIBWEISE: Viele Sprachen unterscheiden keine Groß- und Kleinschreibung, wie etwa Englisch. In anderen Sprachen kann die Schreibweise bedeutungsunterscheidend oder gar ausspracheunterscheidend sein, wie im deutschen Beispiel *weg - Weg*.

Sprachmodell WORTSTELLUNG: In wenig flektierten Sprachen oder solchen mit isoliertem Sprachbau, wird die Beziehung zwischen zwei Wörtern häufig durch die nahe Wortstellung ausgedrückt. Flektionsreiche Sprachen sind in ihrer Wortstellung dagegen wesentlich freier. Bedingt durch die geringe Reichweite der Trigramm-Sprachmodelle sind weniger flektierende Sprachen begünstigt.

Möchte man Sprachen frei von diesen inhärenten Schwierigkeiten ausschließlich auf der akustischen Ebene vergleichen, dann ist die Messung der Phonemfehlerrate auf der Basis eines frei laufenden Phonemerkenners sinnvoll. Mit frei laufendem Phonemerkenner ist ein Erkenner gemeint, der als Vokabular bzw. Aussprachewörterbuch das Phoneminventar verwendet und dessen Sprachmodell gleichverteilte Phonemunigramme enthält.

Um die Leistungen der entwickelten Erkenner in verschiedenen Sprachen zueinander in Beziehung setzen zu können, werden in dieser Arbeit in Kapitel 5 die behandelten Sprachen bezüglich der angeführten Kriterien miteinander verglichen. Neben den Worterkennungsraten werden Phonemerkennungsraten von gebundenen und frei laufenden Phonemerkennern ermittelt.

Kapitel 4

Die GlobalPhone-Datensammlung

Forschung im Bereich multilinguale Spracherkennung ist nicht ohne eine Datenbasis möglich, die ausreichendes Sprach- und Textmaterial von einheitlicher Qualität in vielen Sprachen bereitstellt. Diese Erkenntnis hat sich mittlerweile durchgesetzt, und viele Konsortien kümmern sich heutzutage verstärkt um multilinguale Datensammlungen. Zu Beginn dieser Arbeit gab es noch keine Sammlung von ausreichender Größenordnung. Im Rahmen dieser Arbeit wurde daher das Projekt GlobalPhone *initiiert, um eine eigene Datenbasis zu erstellen. Dieses Kapitel beschreibt die Planung und Durchführung des Projekts und die daraus entstandene* GlobalPhone-*Datenbasis, die die Grundlage dieser Arbeit bildet.*

4.1 Motivation

Im Idealfall stünde zur Forschung im Bereich „Spracherkennung für große Wortschätze unter dem Aspekt der Multilingualität" eine Datenbasis in vielen verschiedenen Sprachen zur Verfügung, die:

- die für die Spracherkennung wichtigsten Sprachen im Sinne von Verbreitungsgrad und wirtschaftlicher Relevanz abdeckt,
- möglichst das komplette lautliche Inventar abdeckt, das der Mensch zur sprachlichen Kommunikation verwendet,
- Personen umfaßt, die einer repräsentativen Auswahl von Muttersprachlern im Sinne von Geschlecht, Alter und Bildung entstammen,
- ausreichend transliteriertes Sprachmaterial enthält, das ein robustes Schätzen akustischer Modelle garantiert (pro Sprache mindestens 10000 Äußerungen mit etwa 100000 kontinuierlich gesprochenen Wörtern ([LADGA96]),

- umfangreiche Texte mit Millionen laufender Wörter zum Trainieren der Sprachmodelle zur Verfügung stellt,
- in möglichst einheitlicher akustischer Qualität vorliegt, um sprachenspezifische Unterschiede erfaßbar machen zu können (gleiche Aufnahmebedingungen, gleiche Umgebungsbedingungen, gleiche Szenarien),
- in allen Sprachen den gleichen Sprechstil aufweist (gelesen, spontan, oder Umgangssprache, im Monolog oder Dialog),
- in semantischer Hinsicht in allen Sprachen äquivalent ist (Vokabular und Domäne).

Zum Zeitpunkt des Beginns dieser Arbeit (1996) gab es nur wenige Datensammlungen, die mehrere Sprachen umfassen. Die umfangreichste Datenbasis, die zur Lösung des Sprachenidentifizierungsproblems entstanden war, ist das *OGI Multilanguage Telephone Speech Corpus* [MCO92], der 1996 bereits 11 Sprachen umfaßte. Da für Sprachenidentifizierung keine Transkriptionen vorausgesetzt werden, wurde nur ein kleiner Anteil der Sammlung transkribiert. Solche Datenbasen sind zur Entwicklung von Erkennern für große Wortschätze ungeeignet. Im Rahmen des deutschen Sprachprojektes VERBMOBIL [VER00] entstand eine sehr umfangreiche, komplett verschriftete Datenbasis. Allerdings umfaßt das Projekt „nur" die drei Sprachen Deutsch, Englisch und Japanisch, was für das im Rahmen dieser Arbeit geplante Forschungsvorhaben zu wenige waren. Neben diesen Datenbasen waren im Jahr 1996 nur monolinguale oder höchstens bilinguale verschriftete Datenbasen verfügbar, wie etwa der ATIS und der *Wall Street Journal* Task in Amerikanischem Englisch von der DARPA, das WSJCAM0 Korpus für Britisches Englisch, dessen französisches Pendant BREF *Le Monde* oder das deutsche PHONDAT-Korpus *Frankfurter Rundschau*. Für Forschungsarbeiten mit dem Ziel des Vergleichs zwischen Sprachen ist eine Zusammenfassung mehrerer solcher Korpora kaum möglich, da die Sammelbedingungen der einzelnen Datenbasen nicht standardisiert sind und daher erheblich voneinander abweichen. Jedes Korpus verfügt über eigene Szenarien und Aufnahmebedingungen.

Während sich mittlerweile die Erkenntnis durchgesetzt hat, daß die Forschung im Bereich multilinguale Sprachtechnologie durch das Fehlen einheitlicher multilingualer Datenbasen stark behindert wird [LADGA96, YADA+97, AD99], gab es bis 1996 keine verschriftete Sprachdatenbasen in genügender Sprachenzahl, ausreichendem Datenumfang und einheitlicher Qualität. Daher entschloß sich die Autorin zur Initiierung des Projekts GlobalPhone, um eigenständig eine multilinguale Datenbasis aufzubauen, die den eigenen Erfordernissen genügt [SWW97, SW98a]. Die Sammlung der GlobalPhone-Datenbasis war damit dem allgemeinen Trend zur Sammlung großer multilingualer Daten voraus. Sie wird heute als Vorbild herangezogen und in Aufzählungen zitiert [AD99].

Mittlerweile bemühen sich zahlreiche international operierende Datenkonsortien darum, große Datensammlung zu koordinieren und zu Forschungszwecken zur Verfügung zu stellen [ML98]. Zu nennen sind hier insbesondere das *Linguistic Data Consortium (LDC)* [LDC00] in den Vereinigten Staaten und die *European Language Resources Association (ELRA)* [ELR98] als Datendistributionsorgan der EU-Language Engineering Initiative. Beispiele für Datensammelprojekte sind Speechdat [Spe98b], welches die Sammlung von Kommandowörtern und isoliert gesprochenen Ziffern und Buchstaben und einigen wenigen gesprochenen Sätzen via Telefon in den 8 europäischen Sprachen Deutsch, Französisch, Italienisch, Britisch-Englisch, Portugiesisch, Spanisch, Schweizer-Französisch und Dänisch umfaßt, und mittlerweile in die vierte Runde geht. Es wurde auf die Sammlung von Autodaten in 10 Sprachen erweitert [SC98]. Im Rahmen des EC-Copernicus Programmes gibt es das Projekt BABEL [BAB99], in dem eine Datenbank für Sprachen aus Zentral- und Osteuropa erstellt werden soll. Gesammelt wird in den 6 Sprachen Bulgarisch, Estnisch, Ungarisch, Polnisch und Rumänisch [Roa97]. Da Multilingualität eines der Kernthemen des fünften Rahmenprogrammes der EU ist, ist die Erstellung großer Datenbasen in vielen Sprachen derzeit eines der vordringlichen Ziele der ELRA. In den Vereinigten Staaten vertreibt das LDC die Daten aus dem Projekt CallHome, in dem Telefonkonversationen in den Sprachen Ägyptisch, Englisch, Mandarin Chinesisch und Spanisch mitgeschnitten werden. Das Hub4 Programm der DARPA umfaßt die Sammlung von Broadcast News Daten in den Sprachen Englisch, Japanisch, Mandarin, und Spanisch. Die Anzahl der Sprachen, der Datenumfang sowie die Szenarien, in denen gesammelt wird, steigt beständig. Das internationale *Coordinating Committee for Speech Databases and Assessment COCOSDA* hat sich mit dem Polyphone Projekt zum Ziel gesetzt, soviele Weltsprachen wie nur möglich in Telefonqualität zu sammeln [BTG94, DBiV+94, MHW+95]. Die Asienerweiterung dieses Kommittees bemüht sich derzeit um den Aufbau von Datenbasen im asiatischen Raum.

4.2 Sprachenauswahl

Bedingt durch die Sprachenvielfalt auf der Welt ist es nicht möglich, alle Sprachen in einer Datenbasis repräsentieren zu wollen. Daher muß zunächst die Zahl der Kandidaten auf ein machbares Maß reduziert werden. Dazu wurde die in Abschnitt 2.2 erarbeiteten Reihung der Sprachen der Welt nach Verbreitungsgrad und Stellenwert herangezogen. Diese Faktoren garantieren die Relevanz und Nützlichkeit eines entstehenden Spracherkennungssysteme. Nimmt man als untere Grenze eine Sprecherzahl von etwa einer Millionen, um für die Spracherkennung interessant zu sein, dann schränkt sich nach den in Abbildung 2.1 präsentierten Zahlen die Auswahl auf etwa 150 Sprachen ein. Im Hinblick auf die Sprachmodellierung sollten ausreichende Textresourcen zur Verfügung stehen, was nach [AD99] auf einige hundert Amtssprachen hochtechnologisierter Länder zutrifft.

Da der Fokus dieser Arbeit auf der Erstellung eines multilingualen Spracherkenners liegt, der leicht auf neue Sprachen portiert werden kann, erfolgte die weitere Auswahl der Sprachen nach dem Kriterium der Diversität. Dazu werden die im Abschnitt 2.4 beschriebenen Strukturkriterien von Sprache herangezogen. Insgesamt sollen die GlobalPhone-Sprachen eine gewisse Varianz der Spracheigenschaften bezüglich phonetischer, phonologischer, prosodischer und grammatischer Merkmale gewährleisten.

Unter Berücksichtigung der genannten Gesichtspunkte und aufgrund einiger äußerer Randbedingungen wurden für die GlobalPhone-Datenbasis 13 Sprachen gesammelt, die in Tabelle 4.1 neben der verwendeten Abkürzung und dem Land, in dem gesammelt wurden, zusammengestellt sind.

Sprache	Abkürzung	Sammelland
Arabisch	AR	Tunesien
Mandarin-Chinesisch	CH	Festland-China
Schanghai-Chinesisch	WU	Festland-China
Deutsch	DE	Deutschland
Japanisch	JA	Japan
Koreanisch	KO	Süd-Korea
Kroatisch	KR	Kroatien und Bosnien
Portugiesisch	PO	Brasilien
Russisch	RU	Weißrußland
Schwedisch	SW	Schweden
Spanisch	SP	Costa Rica
Tamil	TA	Indien
Türkisch	TU	Türkei
Englisch	EN	übernommen aus WSJ
Französisch	FR	übernommen aus BREF

Tabelle 4.1: Die GlobalPhone-Sprachen

Wie der Vergleich mit Tabelle 2.3 zeigt, werden mit den Sprachen Englisch, Deutsch, Japanisch, Französisch und Russisch fünf der sechs volkswirtschaftlich interessantesten Sprachen abgedeckt. Wegen ihres hohen Verbreitungsgrades und ungebrochenen Bevölkerungswachstums sind die Sprachen Hindi, Spanisch und Arabisch relevant, von denen GlobalPhone die letzten beiden abdeckt. Sprachen wie Japanisch, Koreanisch, Chinesisch und Thai sind die Sprachen der Zukunftsmärkte, da sich die asiatischen Staaten zu starken Handelspartnern entwickelt haben. Von diesen 4 Sprachen sind in der GlobalPhone-Datenbasis die ersten drei erfaßt worden.

Durch die marktwirtschaftliche Dynamik in den früheren „Ostblockstaaten“ eröffnen sich neue Möglichkeiten und damit auch sprachliche Herausforderungen für Industrie, Dienstleister, Medien und Behörden. Sprachen wie Tschechisch, Slowakisch, Polnisch, Ungarisch, Bulgarisch, Russisch, Rumänisch, Serbisch, Slowenisch, Kroatisch, Mazedonisch, Weißrussisch, Ukrainisch, Estnisch, Litauisch, Lettisch und andere erhalten dadurch einen ganz neuen Stellenwert. Dies gilt insbesondere für den Standort Deutschland, da sich durch die geographische Nähe schnell große Märkte entwickeln können. In GlobalPhone sind diese Sprachen mit Russisch und Kroatisch (Serbisch) repräsentiert.

Mit den genannten Sprachen deckt die GlobalPhone-Datenbasis neun der 12 wichtigsten Weltsprachen (vgl. Tabelle 2.1) und 11 der 20 wichtigsten Amtssprachen (vgl. Tabelle 2.2) ab. Nach Tabelle 2.4 entspricht das etwa 85% der auf der Welt gesprochenen Sprachfamilien, darunter die Indo-Europäischen Sprachen Deutsch und Schwedisch, die beide wie das Englische germanischen Ursprungs sind sowie Portugiesisch und Spanisch, beide romanisch wie das Französische. Mit den slawischen Sprachen Russisch und Kroatisch ergibt sich ein weiteres nahverwandtes Sprachpaar. Japanisch und Koreanisch sind zwei isolierte Sprachen ohne jegliche Verwandtschaft zu irgendeiner lebenden Sprache. Mandarin und Schanghai sind die beiden häufigsten Dialekte des Chinesischen aus der Sino-Tibetanischen Sprachfamilie. Arabisch gehört als hamitische Sprache zu einer weiteren großen Familie, ebenso wie die Turksprache Türkisch und die dravidische Sprache Tamil. Ein Vergleich mit Tabelle 2.1 zeigt, daß die GlobalPhone-Sprachen von insgesamt 3 der 5 Milliarden Menschen zählenden Weltbevölkerung um 1990 als Muttersprache gesprochen werden.

Die GlobalPhone-Sprachen decken eine sehr breite Palette sprachlicher Merkmale ab, die sich manifestieren in:

- Phonetischen Unterschieden: Das Lautinventar erstreckt sich von sehr einfachen Systemen wie im Türkischen über Inventare mit zahlreichen Diph- und Triphthongen (Koreanische), vielen Nasalen (Portugiesisch) hin zu Laryngallauten (Arabisch) und komplexen Tonsystemen (Mandarin und Schanghai-Chinesisch).

- Phonologischen Unterschieden: Während Sprachen wie Mandarin ein recht einfaches CV oder CVC basiertes Silbenmuster mit einem limitierten Repertoire an Konsonanten und mit charakteristischen Tonhöhenkonturen hat, weisen beispielsweise Deutsch und Englisch eine vielfältige Silbenstruktur auf, mit potentiellen Clustern von bis zu 6 Konsonanten zwischen Vokalkernen aber ohne Tonhöhenkontraste zwischen Silben.

- Prosodischen Unterschieden: GlobalPhone umfaßt mit Mandarin und Schanghai-Chinesisch sowohl zwei Tonsprachen als auch Akzentsprachen wie Japanisch und Betonungssprachen mit entweder festen Betonungsmustern wie Türkisch oder lexikalischen Betonungsmustern wie Deutsch und Englisch.

- Morphologischen Unterschieden: Der Sprachbau der GlobalPhone-Sprachen reicht von isoliert (Chinesisch) über kaum flektiert (Englisch) bis hin zu agglutinierenden Sprachen wie Türkisch und Koreanisch.

4.3 Domänen- und Textauswahl

Zusätzlich zu den oben formulierten Forderungen an die Sprachenauswahl spielten die Faktoren Kosten und Zeit eine wesentliche Rolle beim Entwurf des Korpus. Der aufwendigste und teuerste Teil einer Sammlung von Sprachdaten ist die nachträgliche Verschriftung des gesprochenen Textes. Um möglichst kostengünstig zu bleiben, wurden gelesene Sprachdaten gesammelt, also solche, bei denen die zu sprechenden Texte bereits in elektronischer Form vorliegen.

Sprache	Name der Zeitung URL (Stand Oktober 1999)
Arabisch	Assabah http://www.tunisie.com/Assabah
Chinesisch	Peoples Daily http://www.snweb.com
Deutsch	FAZ, Süddeutsche http://www.faz.de http://www.sueddeutsche.de
Koreanisch	Hankyoreh Daily News http://news.hani.co.kr
Kroatisch	HRT, Nacional, Obzor http://www.hrt.hr/vijesti, www.nacional.hr http://www.tel.hr/hrvatski-obzor
Japanisch	Nikkei Shinbun http://www.nikkeihome.co.jp
Portugiesisch	Folha de São Paulo http://www.uol.com.br/fsp
Russisch	Ogonyok Gaseta und express-chronika http://www.ropnet.ru/ogonyok
Schwedisch	Göteborgs-Posten http://www.gp.se
Spanisch	La Nacion http://www.nacion.co.cr
Tamil	Thinaboomi Tamil Daily http://www.thinaboomi.com
Türkisch	Zaman http://www.zaman.com.tr

Tabelle 4.2: Name und Internet-Quellen der zur Datensammlung verwendeten überregionale landesspezifische Tageszeitungen

Darüber hinaus war es das Ziel, in allen Sprachen Daten über dieselbe Domäne zu sammeln. Als Textquelle wurden im Internet verfügbare überregionale Tageszeitungen der jeweiligen Länder ausgewählt. Die Themengebiete umfassen die Bereiche internationales und nationales Tagesgeschehen sowie Wirtschaftsberichte. Dies gewährleistet den Gebrauch sprachenübergreifender Namen und Bezeichnungen sowie einigermaßen Vergleichbarkeit. Die Domäne ist sehr weit gefaßt und wird so den

Erfordernissen eines Erkenners für große Wortschätze gerecht. Auch kann dadurch das Ziel erreicht werden, möglichst vielfältiges Sprachmaterial zu erhalten, um die Polyphonabdeckung zu maximieren. Die elektronische Verfügbarkeit der Zeitungen garantiert, daß eine nahezu unerschöpfliche Quelle weiterer Textdaten zur Verfügung steht, wie sie zur zuverlässigen Schätzung der Sprachmodelle notwendig sind. Tabelle 4.2 zeigt die zur Sammlung der Sprachdaten verwendeten Tageszeitungen nebst ihren Internet-Quelle.

4.4 Datenerfassung

Es ist nicht einfach, Menschen aus den verschiedensten Teilen der Erde dazu zu bewegen, für Forschungszwecke ihre Sprache zu spenden. Eine Sammlung in Deutschland ist problematisch, weil es nicht einfach ist, viele Muttersprachler für die ausgewählten Sprachen zu finden. Als Möglichkeit bliebe ein Datensammlung via Telefon. Dieses Vorgehen hat aber drei gravierende Nachteile: Erstens sind Telefondaten bandbegrenzt (350 - 3500 Hz), zweitens ist bei Anrufen aus verschiedenen Teilen der Welt mit erheblichen Kanalunterschieden zwischen den Telefonverbindungen zu rechnen. Beide Effekte können unerwünschte Auswirkungen auf die Sprachenvergleiche haben. Drittens erfordert die Sammlung per Telefon eine Werbekampagne größeren Ausmaßes (Funk, Fernsehen, Zeitung, newsgroups), um genügend Teilnehmer zu garantieren. Die Idee war, statt dessen einige ausländische Studierende in Deutschland auszuwählen, und deren persönliche Kontakte im Heimatland zu nutzen, um Sprachdaten vor Ort sammeln zu lassen.

4.4.1 Die Sammlungskampagne

Zur Sammlung der Sprachdaten wurden ausländische Studierende der Universität Karlsruhe engagiert. Sie wurden über die Projektziele informiert, über das Szenario und die zu vermittelnde Sprecherinstruktionen aufgeklärt und im Umgang mit den Aufnahmegeräten geschult. Noch in Deutschland planten und organisierten sie die Sammlung und bearbeiteten die notwendigen Textvorlagen. Anschließend reisten sie im Auftrag des Instituts in ihre Heimatländer und baten dort Freunde, Verwandte und Bekannte für die Wissenschaft Sprache zu spenden.

Das Vorgehen erwies sich für die eigenen Zwecke als sehr gut geeignet. Die Studierenden waren hochmotiviert und sorgsam darauf bedacht, qualitativ hochwertige Arbeit zu liefern. Auf diese Weise wurde eine ausreichend große Zahl von kooperativen Sprachspendern gefunden. Gleichzeitig konnten die Kosten für die Aufenthalte vor Ort gering gehalten werden, weil die Studierenden bei ihren Familien wohnten und die Sprachspender auf eine Aufwandsentschädigung verzichteten.

4.4.2 Die Aufnahmegeräte

Alle Daten wurden mit einem tragbaren DAT-Rekorder Sony TDC-8 und einem Nahsprechmikrofon der Firma Sennheiser HD-440-6 aufgenommen und mit einer Abtastfrequenz von 48 kHz digital auf DAT-Bänder aufgezeichnet. Anschließend wurden sie mit Hilfe einer speziellen Hardware-Karte der Firma MICROWAVE optisch auf einen PC übertragen und dort mit der Software WAVE der Firma Turtle Beach mit 16-bit Auflösung digitalisiert. Zur weiteren Verarbeitung im Spracherkenner wurden die Aufnahmen auf eine Abtastrate von 16 kHz transformiert. Die akustische Uniformität der Daten spielte dabei eine sehr wichtige Rolle - so wurde in früheren Arbeiten der Autorin gezeigt, daß die unterschiedliche Qualität der akustischen Daten einen signifikanten Einfluß auf die Sprachenidentifizierungsleistung hat [SRW96]. Der Vorgang der Digitalisierung und der Reduktion der Abtastfrequenz der Aufnahmen kostete insgesamt zwei Echtzeitfaktoren Rechenleistung.

4.4.3 Die Sprachspender

Die Muttersprachler sollten einen möglichst guten Querschnitt durch die Bevölkerung ergeben. Angestrebt war ein repräsentativer Anteil Frauen und Männer verschiedener Altersstufen (18 - 80 Jahre) und unterschiedlicher Bildungsgrade. In jeder Sprache wurden etwa 100 Muttersprachler gebeten, die vorgelegte Texte zu lesen. Pro Sprecher wurden 15-20 Minuten Sprache aufgezeichnet. Um Lesefehler zu minimieren, war es den Sprechern gestattet, die Texte vorher durchzulesen und bei eventuellen Unklarheiten nachzufragen. Die Sprecher waren angewiesen, am Satzende (in den meisten Sprachen durch ein Satzendezeichen markiert) eine Pause von mindestens einer Sekunde Länge einzulegen.

In einem Sprecherdatenblatt wurden Informationen zu den Aufnahmesitzungen festgehalten. Es wurden soziolinguistische Informationen erfragt, die einen Einfluß auf die Sprachen nehmen können, wie Geschlecht, Alter, Beruf, dialektale Sprachfärbung und Atemwegserkrankungen sowie nach Aufnahmebedingungen wie Raumcharakteristik und Hintergrundgeräuschen.

4.5 Korpusentwicklung

Mit *Korpus* wird im folgenden die Sprach- und Textdatenbasis bezeichnet, die in aufgearbeiteter und validierter Form zum Training und Testen der Spracherkenner bereitsteht.

4.5.1 Das Softwaretool *„mapper“*

Um aus den gesammelten Audiodateien einen brauchbaren Korpus zu generieren, mit dem ein Spracherkennungssystem trainiert und getestet werden kann, wurde das Softwaretool *„mapper“* entwickelt, das auf dem JANUS Recognition Toolkit JRTk aufsetzt. Das Tool bietet eine bedienerfreundliche Oberfläche und integriert alle Komponenten, die zum Umgang mit den Daten notwendig wurden, wie Visualisierung der Audiodaten, Pausendetektion und Zerschneiden von Audiodateien in kleinere Segmente, Romanisieren von Textdaten, Anzeigen und Editieren der Texte, Eingabe und Interpretation der Protokolldateien sowie Aufbereitung und Sicherung der Daten.

4.5.2 Validierung der Sprachdaten

Die aufgezeichneten Sprachdaten durchliefen einen dreiteiligen Validierungsprozeß. Im ersten Schritt werden die Aufnahmeprotokollblätter erfaßt. Diese enthalten alle relevanten Informationen, um die transferierten Audiodaten in satzweise Äußerungen zu zerlegen. Dieser Schritt nimmt pro gesammeltem Sprecher etwa eine Minute Eingabezeit in Anspruch. Im zweiten Schritt werden die aufgezeichneten Daten mittels eines Pausendetektors nach den 1-Sekundenpausen durchsucht, die bei der Aufnahme zwischen zwei zu lesenden Sätzen von den Sprechern eingelegt worden waren (vgl. Abschnitt 4.4.3) und entsprechend segmentiert. Dieser Schritt erfordert keine menschliche Arbeitszeit denn er verläuft voll automatisch in etwa 0.2 Echtzeit. Im dritten Schritt werden die Texte den Audiosegmenten zugeordnet. Eventuelle Segmentierungsfehler, die durch (sehr seltene) lange Pausen innerhalb eines Satzes entstehen können, werden korrigiert. Zur Überprüfung, ob Text und Audio zusammenpassen, wurden alle Daten von denselben Studierenden, die zuvor die Daten gesammelt hatten, probegehört. Falsch gelesene Äußerungen werden dabei entfernt. Kleinere Lesefehler werden auf den Texten korrigiert, gut hörbare Effekte spontaner Sprache wie Häsitationen, Wortabbrüche und Stottern werden in den Texten markiert. Die dazu erforderliche Bearbeitungszeit schwankt je nach Lesequalität der Sprecher erheblich. Gute Sprecher konnten in Echtzeit fertiggestellt werden. Bei sehr schwachen Leseleistungen wurden zur Korrektur bis zu 4 Stunden, als etwa 16 mal Echtzeit investiert. Es gab Fällen, in denen die Sprecher eine derart unzureichende Lesequalität vorwiesen, daß deren Bearbeitung aus Effiziensgründen abgebrochen wurde. Im Schnitt betrug die Fertigstellung eines Sprechers etwa 4 mal Echtzeit.

4.5.3 Zeitaufwand

Als grobe Abschätzung des benötigten Zeitbedarfs für die Erstellung des gesamten GlobalPhone-Korpus wird pro Sprecher von einem mittleren Aufwand ausgegangen:

Aufnahme der Sprache:	15-20 Minuten
Übertragen der Daten auf PC	15-20 Minuten
Reduzieren der Abtastfrequenz	15-20 Minuten
Validieren der Sprachdaten	60 Minuten
Gesamtaufwand pro Sprecher	2 Stunden

Damit ergibt sich als geschätzter Gesamtaufwand für die GlobalPhone-Datensammlung 1200 Sprecher $\times 2$ Stunden = 2400 Arbeitsstunden, wobei die Vorbereitung der Datensammlung, sowie die Dauer und Kosten der Anreise und Aufenthalte vor Ort nicht miteinbezogen wurden. Ebenfalls nicht mitgerechnet sind hier die Arbeitsstunden zur Sammlung und Normalisierung der zusätzlich gesammelten Textdaten und der Aufbereitung der Aussprachewörterbücher.

4.6 Aktueller Stand des GlobalPhone-Korpus

Dieser Abschnitt gibt eine Übersicht über die soziolinguistischen Eigenschaften aller aufgenommenen Sprecher, die aufgezeichneten Sprachdaten und die Textdaten des GlobalPhone-Korpus. In der vorliegenden Arbeit wurden für 12 Sprachen automatische Spracherkennungssysteme trainiert und evaluiert. Für die Sprachen Arabisch, Tamil und Schanghai-Chinesisch sind derzeit noch keine Erkennersysteme verfügbar. Im Januar 2000 wurde die Datenbasis um tschechische Daten erweitert, die allerdings nicht mehr in die Arbeit aufgenommen wurden.

Sprecherstatistik

Die Grafik 4.1 zeigt die Verteilung über das Geschlecht für alle Sprecher der GlobalPhone-Sprachen sowie für Englisch und Französisch. In jeder Sprache waren insgesamt 100 Sprecher geplant, außer in den Sprachen Schanghai-Chinesisch und Tamil, in denen von vorneherein nur einiger Sprecher gesammelt werden sollten. Im Mandarin-Chinesisch, Japanischen und Russischen liegen weit mehr Sprecher vor als geplant. Im Japanischen war eine zweite Sammlungskampagne durchgeführt worden, um das Ungleichgewicht der Geschlechter auszugleichen (siehe unten). Für die chinesische und russische Sprache wurden mehr Sprecher gesammelt, da das Interesse an diesen Sprachen derzeit sehr groß ist.

Es ist zu erkennen, daß bis auf wenige Ausnahmen das Verhältnis zwischen männlichen und weiblichen Sprechern recht ausgeglichen ist. Im Japanischen besteht ein Übergewicht an männlichen Sprechern, was daran liegt, daß die ersten 100 Sprecher im Umfeld einer technischen Universität in Tokio gesammelt wurden, an der keine Frauen studieren. Es wurden daher in einer zweiten Aktion verstärkt japanische Sprecherinnen gesammelt, was das Gleichgewicht nur teilweise wiederherstellte. Im Deutschen ist das Verhältnis noch schlechter. Die deutschen Daten wurden an der Universität Karlsruhe in der Fakultät für Informatik aufgenommen, an der der prozentuale Anteil an Frauen bekanntermaßen sehr gering ist. Im Türkischen ist die Situation genau umgekehrt, hier überwiegen die weiblichen Sprecherinnen. Es war nach Aussagen der Datensammlerin ausgesprochen schwierig, türkische Männer zum Vorlesen von Zeitungstexten zu bewegen.

Die Grafik 4.2 zeigt die Verteilung über die Altersgruppen für alle Sprecher aller Sprachen des GlobalPhone-Korpus. Deutlich sichtbar ist das Übergewicht von

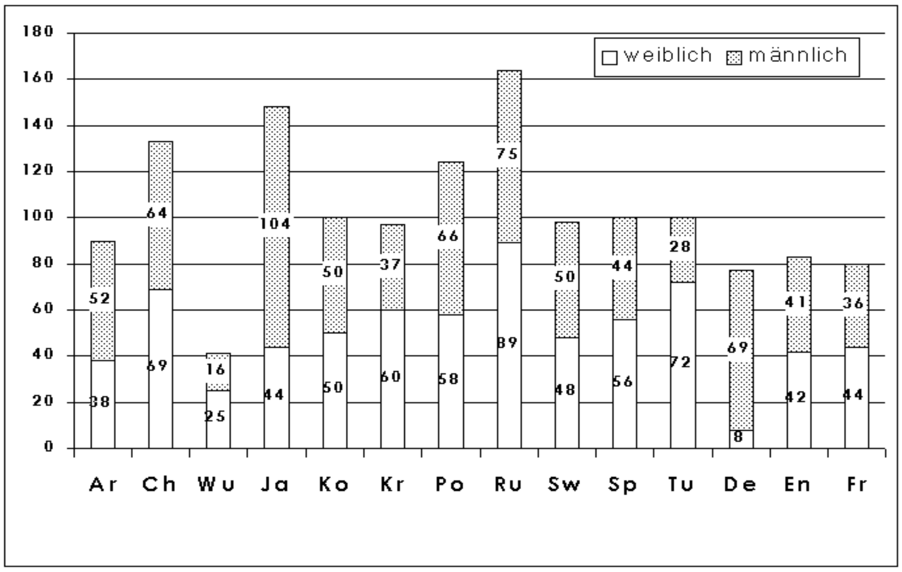


Abbildung 4.1: Geschlecht der Sprecher im GlobalPhone-Korpus

Sprechern aus der Altersgruppe 20–29 Jahre. Das liegt daran, daß aus finanziellen Gründen freiwilligen Sprachspender von den studentischen Datensammler aquiriert wurden. Dabei handelt es sich um Freunde, die naturgemäß ebenfalls aus dem studentischen Umfeld stammen und in etwa dasselbe Alter haben. Die meisten Datensammler berichteten, daß es ausgesprochen schwierig war, ältere, im Berufsleben stehende Personen zum zeitaufwendigen und unentgeltlichen Sprachspenden zu überreden. Insgesamt steht aber dennoch eine breit gestreute Population zur Verfügung.

Statistik über die Aufnahmen und Äußerungen

Tabelle 4.3 gibt einen Überblick über die aufgezeichneten Sprachaufnahmen und textuell erfaßten Äußerungen des GlobalPhone-Korpus. Erklärtes Minimalziel der Sammlung war die Größenordnung von 10 Stunden Sprachmaterial pro Sprache. Diese Zahl garantiert die akkurate Schätzung monolingualer akustischer Modelle für jede Sprache. Die Tabelle 4.3 zeigt, daß dieses Minimalziel in allen Sprachen erreicht wurde. Die Schwankungen zwischen den Sprachen sind dadurch bedingt, daß wie bereits erwähnt einige Sprachen bevorzugt und in mehreren Kampagnen gesammelt wurden (Chinesisch und Japanisch), während andere Sprachen mit niedriger Priorität mitliefen (Schanghai-Chinesisch und Tamil). Insgesamt stehen 270 Stunden Sprachdaten von etwa 1360 Sprechern in 13 Sprachen zur Verfügung. Das entspricht über 110.000 einzelnen Aufnahmen.

Tabelle 4.3 zeigt, daß die mittlere Dauer einer Aufnahme zwischen den Sprachen stark schwankt. Der Durchschnitt über alle Sprachen hinweg liegt bei etwa 8.8 Sekunden. Eine Aufnahme entspricht einer gelesenen Äußerung, d.h. einem geschriebenen Satz in der jeweiligen Online-Zeitung. Wie aus der Tabelle 4.3 zu entnehmen, werden

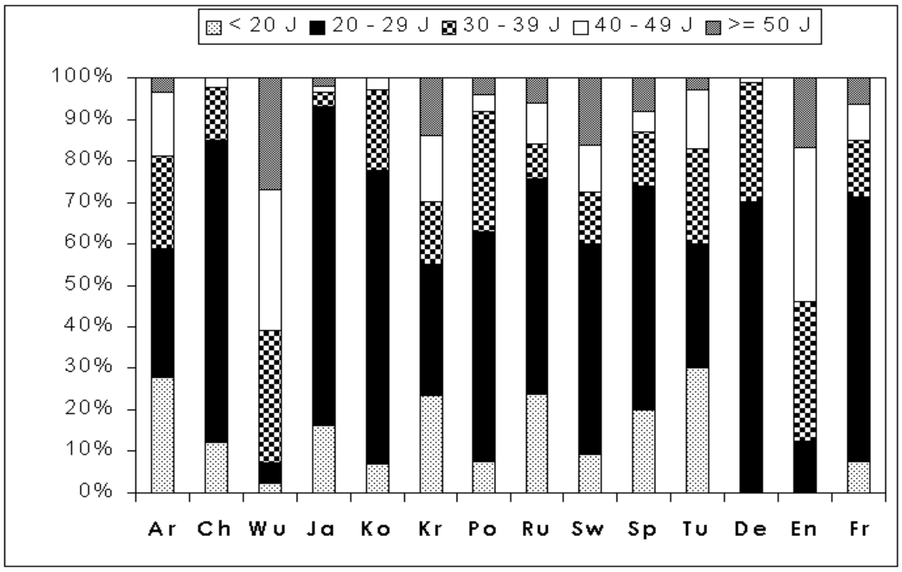


Abbildung 4.2: Altersverteilung der Sprecher im GlobalPhone-Korpus

im Schnitt 18.8 Wörter pro Äußerung gesprochen. Setzt man Dauer der Aufnahme und Länge der Äußerung in Wörtern zueinander in Beziehung, dann ergibt sich über die Sprachen hinweg ein einigermaßen konstantes Verhältnis von etwas mehr als 2 gesprochenen Wörtern pro Sekunde. Auf die sprachenspezifischen Unterschiede wird in Kapitel 5 näher eingegangen.

Aufteilung in Trainings- und Testdaten

In allen Sprachen wurden die Sprecher disjunkt im Verhältnis 80:10:10 auf die Trainings-, Kreuzvalidierungs- und Evaluationsmenge verteilt. Bei der Aufteilung wurde auf Homogenität der Alters- und Geschlechtsverteilung (und der Dialekte, sofern vorhanden) in den einzelnen Sprecheruntermengen Wert gelegt. Darüber hinaus wurde beachtet, daß diejenigen Sprecher, deren Textvorlagen mehrfach verwendet worden waren, in die Trainingsmenge gelegt wurden. Dieser Aspekt war besonders für die türkischen und portugiesischen Daten wesentlich, da in diesen Sprachen wegen eines temporären Mangels an Textvorlagen einige Texte von mehreren Sprechern gelesen wurden. Die textdisjunkte Aufteilung der Sprecher verhindert eine Verfälschung der Aussagen zu Wortfehlerraten und Out-Of-Vocabulary Angaben. Auf den Sprachdaten der Trainingsmenge wurden die akustischen Modelle trainiert. Das Textmaterial der Trainingsdaten wurde zur Sprachmodellierung mitverwendet. Anhand der Kreuzvalidierungsmenge wurden freie Parameter des Spracherkennungssystems, wie beispielsweise z, p eingestellt (vgl. Abschnitt 3.2.3). Die Evaluierungsmenge dient zur Feststellung der Wortfehlerraten.

[1] Englisch (EN) und Französisch (FR) wurden nicht im Rahmen von GlobalPhone gesammelt, da mit dem Wall Street Journal Korpus und dem BREF-le Monde Korpus bereits Daten im gewünschten Format vorliegen.

Sprache	Sprecher	Aufnahmen	Länge pro Aufnahme [s]	Gesamtlänge [Std]	Wörter pro Äußerung	Laufende Wörter
Ch-Mandarin	132	10181	11.0	31.2	25.9	262841
Ch-Schanghai	41	2644	12.9	9.5	-	-
Deutsch	77	10085	6.5	18.3	14.1	151327
Kroatisch	92	4499	12.7	15.9	26.6	119536
Japanisch	144	13067	9.3	33.9	20.7	268248
Koreanisch	100	8107	9.3	21.0	14.6	117263
Portugiesisch	101	10220	9.2	26.0	20.1	208084
Russisch	106	11111	7.2	22.2	15.3	169854
Schwedisch	98	11816	6.6	21.7	15.6	184013
Spanisch	100	6898	11.5	22.1	25.7	171557
Türkisch	100	6950	8.9	22.2	16.2	112672
Englisch[1]	103	7434	7.7	15.9	18.4	137030
Französisch[1]	80	7516	7.0	14.7	17.0	127829
Summe	1364	110528	8.8	269.7	18.8	2030254

Tabelle 4.3: Statistik über die Sprachaufnahmen und Äußerungen des GlobalPhone-Korpus

Kapitel 5

Monolinguale Spracherkennung in 10 Sprachen

Forschung im Bereich multilingualer Spracherkennung basiert auf monolingualen Spracherkennern für unterschiedliche Sprachen. Beim Vergleich zwischen ihnen können hilfreiche Erkenntnisse gewonnen werden. Diese Grundlage zu schaffen durch die Entwicklung und den Vergleich von Spracherkennern in 10 Sprachen ist ein wesentlicher Anteil dieser Arbeit. Das vorliegende Kapitel beschreibt die Vorgehensweise beim Bau der Erkenner, die durch das Streben nach Effizienz und Automatisierung motiviert ist. Anhand der Basissysteme werden die relevanten Eigenschaften von Sprachen herausgearbeitet. Daraus werden Maßnahmen zur sprachenspezifischen Verbesserung der Basissysteme abgeleitet und im letzten Abschnitt des Kapitels exemplarisch an einigen Sprachen beschrieben. Vorangestellt ist ein kurzer Abriß zum Stand der Forschung, der die bisherigen Probleme dieses Forschungszweiges beschreibt.

5.1 Ziele und Kriterien

Das Vorhaben einer Entwicklung von monolingualen Erkennungssysteme in vielen Sprachen legt nahe, diesen Vorgang zu automatisieren. Wesentliche Ziele der „Monolingualen Spracherkennung in vielen Sprachen“ sind daher die Minimierung des Entwicklungsaufwandes von Spracherkennern für einzelne Sprachen und die Untersuchung der Sprachenuniversalität der Spracherkennungssoftware.

Damit einhergehend stellt sich die Frage, ob die mit den klassischen Methoden der Spracherkennung bisher erzielten Resultate sich auf verschiedene Sprachen übertragen lassen. Der Vergleich von Spracherkennern unterschiedlicher Sprachen soll daher Informationen über relevante Eigenschaften von Sprachen herausarbeiten um daraus eventuell Maßnahmen zu sprachenspezifischen Behandlung von Besonderheiten abzuleiten. Die daraus gewonnenen Kenntnisse können zur Entwicklung eines multilingualen Erkenners herangezogen werden.

5.2 Stand der Forschung

Bis in die späten 80er Jahre war das Hauptaugenmerk der Spracherkennung für große Wortschätze auf amerikanisches Englisch gerichtet. Einer der Gründe war die Förderung der HUB-Programme und Benchmarktests durch die DARPA, an denen jährlich renommierte Forschungseinrichtungen gegeneinander antraten. Eines der ersten Forschungsprogramme in Europa, in dem gezielt Spracherkennung und Sprachverstehen in mehreren europäischen Sprachen gefördert wurde, war das 1988 initiierte Esprit-Projekt SUNDIAL (Speech Understanding and Dialog) [SUN99], an dem sich 12 universitäre und industrielle Forschungseinrichtungen aus 4 Ländern beteiligten. 1993 startete die EU das Projekt SQALE [SQA98], in dem die DARPA-Paradigmen auf europäische Sprachen übertragen und ihre Funktionalität überprüft werden sollten. Mit diesen beiden Programmen begann die Forschung sich für die Frage zu interessieren, ob vergleichbare Resultate in verschiedenen Sprachen erzielt werden könnten und ob die bisher aufgestellten Paradigmen sich auf andere Sprachen übertragen lassen.

Forschungsarbeiten können untergliedert werden in solche, die sich mit dem phonetischen Vergleich zwischen Sprachen beschäftigen [CALADG98], und solche, die sich mit der Worterkennung im Kontext von Mensch-Maschine-Schnittstellen beschäftigen, wie etwa Diktiersystemen [LADG95, LADGA96, YADA+97, BBH+95, DAK95, PWY95, WGT+00] oder Auskunftssystemen [ZSP+96, GFG+95], und solche Arbeiten, die die Mensch-zu-Mensch-Kommunikation als Hauptthema haben [BCG+96, BMM+97, VER00, CS99, WSM00].

Phonemerkennung
Arbeiten im Bereich der Phonemerkennung waren die ersten, die sich dem Phänomen des Sprachenvergleiches näherten. Ein Grund dafür ist die Nähe zum Problem der Sprachenidentifizierung, das Anfang der 90er Jahre von großem Interesse war. Ein weiterer Grund ist, daß die Entwicklung von Phonemerkennern, verglichen mit großen Wortschatzsytemen, mit geringerem Aufwand und kleinen Datenmengen zu bewerkstelligen ist. Darüber hinaus haben sich Phonetiker intensiv mit der Ähnlichkeit von Lauten verschiedener Sprachen beschäftigt und die Spracherkennung dadurch beeinflußt. Bisherige Untersuchungen wie etwa [LG93b, CALADG98, CAGADL97, LG93a, Köh96] lassen übereinstimmend den Schluß zu, daß es bereits auf Phonemebene signifikante Unterschiede bezüglich der Erkennung zwischen Sprachen gibt. Bisherige Arbeiten beschäftigten sich allerdings nur mit einer beschränkten Anzahl indoeuropäischer Sprachen. Zwischen diesen Sprachen wurden Unterschiede bezüglich der Phonemverwechslungsraten, der Per-

plexität phonem-basierter N-Gramm-Sprachmodelle und der Anzahl von Polyphonen festgestellt. Die Phonemerkennungsleistungen zeigen auf unterschiedlichen Korpora und Sprachstilen, daß sich die bisher untersuchten Sprachen grob in 3 Klassen einteilen lassen, in denen Spanisch und Italienisch zu den einfachen Sprachen zählen, Deutsch, Französisch und Portugiesisch eine Mittelposition einnehmen und Englisch mit Abstand am schwierigsten zu erkennen ist [Köh96, CALADG98, LG93b, Köh99]. Tendenziell wurde eine Korrespondenz zwischen Größe des Lautinventars und Erkennungsgenauigkeit festgestellt [CALADG98, CAGADL97].

Diktiererkennung auf großen Wortschätzen

Im Dezember 1993 startete die EU das zweijährige Projekt SQALE mit dem Ziel, die Übertragbarkeit der DARPA-Evaluationsparadigmen auf europäische Sprachen zu prüfen [YADA$^+$97]. Es sollte ein Rahmen etabliert werden, um Trainings- und Testmaterial auszutauschen und gemeinsame Protokolle und Mechanismen zu erarbeiten. Außerdem sollten Evaluationen zum Sprachenvergleich geplant, durchgeführt und analysiert werden. An diesem Experiment waren führende Labors beteiligt: Cambridge University als Experte für Britisches und amerikanisches Englisch [PWY95], Philips für Deutsch [DAK95], LIMSI für Französisch [LADG95] [LADGA96] sowie TNO[1] als Experte für Evaluation [SvL95] und als Projektkoordinator. Die Partner entwickelten Erkennersysteme auf gelesenen Daten in den vier Sprachen amerikanisches und britisches Englisch, Französisch und Deutsch und verglichen die Ergebnisse miteinander. Die Untersuchung wurde auf WSJ für Englisch, BREF-Le Monde für Französisch und PHONDAT-Frankfurter Rundschau für Deutsch durchgeführt. In allen Systemen wurden zur akustischen Modellierung phonembasierte kontextabhängige CDHMMs verwendet und zur Sprachmodellierung Trigramme [YADA$^+$97] eingesetzt.

Die einhellige Schlußfolgerung aus diesem Projekt ist, daß für die vier untersuchten Sprachen dieselben Methoden und Algorithmen anwendbar sind. Alle Beteiligten berichteten, daß vergleichbar gute Systeme in den Sprachen erstellt werden konnten. Nach der Festlegung des Phoneminventars und unter Annahme, daß ein Aussprachewörterbuch und genügend Sprachmaterial vorhanden sind, konnte die Trainingsroutine für die initialen Erkenner sprachenunabhängig verlaufen. Allerdings wurde als wichtige Erfahrung festgehalten, daß für alle Systeme nach der initialen Phase mehrere Verbesserungsschritte durchgeführt werden mußten. Trotz der nahen Verwandtschaft der untersuchten Sprachen konnten diese Verbesserungen nur durch sprachenspezifische Lösungen erreicht werden. Infolge der uneinheitlichen Korpora war nicht zu differenzieren, ob gemessene Leistungsunterschiede aus inhärenten Sprachschwierigkeiten resultieren oder auf Datenmängel zurückzuführen sind [DAK95]. Abschließend stellte man die offenen Fragen zur Diskussion, ob man Sprachen auf denselben Vokabulargrößen vergleichen soll, wie man Homophone evaluie-

[1] TNO = TNO Human Factors Research Institute

ren sollte, wie das Problem der Normalisierung der Textdaten inklusive der Versalien gelöst werden soll und ob sich Perplexitäten ohne Normalisierung vergleichen lassen.

In [BBH+95] vergleichen die Autoren Diktiererkenner in den fünf europäischen Sprachen Deutsch, Italienisch, Französisch, Englisch und Spanisch, die mit isolierter Spracheingabe betrieben werden. Die Algorithmen und Methoden sind in allen fünf Sprachen identisch. Das verwendete Sprachmaterial ist nach Aussagen der Autoren in allen Sprachen vergleichbar, so daß sie davon ausgehen, daß gemessene Leistungsunterschiede ihre Ursache in den inhärent in der Sprache verankerten Gegebenheiten haben. Die Schlußfolgerungen bezüglich des Schwierigkeitsgrades für Englisch, Französisch und Deutsch decken sich mit denen aus SQALE. Beim Vergleich der Erkennungsleistungen kommen sie zu dem Schluß, daß Italienisch am leichtesten, dann in abnehmender Folge Englisch, Spanisch und mit einigem Abstand Deutsch und zuletzt Französisch am schwierigsten zu erkennen ist. Ein Vergleich mit den Resultaten der Phonemerkennungsraten aus Abschnitt 5.2 zeigt, daß sich insbesondere für Englisch und Deutsch die Rangfolgen stark verschieben.

Spontansprachliche Mensch-Maschine-Auskunftssysteme

Die Arbeitsgruppe um Zue beschreibt in [ZSP+96] und [GFG+95] das Telefonauskunftssystem VOYAGER und das universelle Informationssystem GALAXY. Derzeit werden die Systeme in den drei Sprachen Englisch, Japanisch und Italienisch betrieben, sollen aber auf Chinesisch und Spanisch erweitert werden. Der Fokus ihrer Forschungsarbeiten liegt darauf, Informationen in Datenbanken so zu speichern, daß sie in möglichst vielen Sprachen abfragbar sind. Das Ziel ist die sprachenunabhängige Informationsabfrage und -speicherung im Hinblick auf den globalen Informationsaustausch. Das gesamte System ist nach dem Prinzip der gemeinsamen Nutzung von Software sprachenunabhängig konzipiert, um die zukünftige Portierung auf neue Sprachen optimal zu unterstützen. Alle sprachenabhängigen Informationen sind in Regelwerken getrennt gespeichert und nicht inhärent im System kodiert. Die Spracherkenner sind derzeit monolingual realisiert, Vergleiche zwischen den Sprachen wurden nicht publiziert.

SQEL[2] ist ein Copernicus-Projekt, in dem ein multilinguales Informationsabfragesystem in einer Zugauskunfts- und Flugbuchungsdomäne entwickelt wird. Es werden die vier Sprachen Deutsch, Slowenisch, Slowakisch und Tschechisch bearbeitet. [HNN98] vergleicht die monolingualen Erkenner dieser vier Sprachen, die alle mit derselben Sprachtechnologie entwickelt wurden. Die Erkennungsergebnisse liegen abgesehen von Tschechisch in vergleichbarem Rahmen. Allerdings liegt auch hier sehr heterogenes Datenmaterial vor.

Spontansprachliche Mensch-Mensch-Übersetzungssysteme

Das Verbmobil Projekt realisiert die Übersetzung spontansprachlicher Mensch-zu-Mensch Dialoge in den drei Sprachen Deutsch, Englisch (Amerikanisches Englisch)

[2] Spoken Queries in European Languages

und Japanisch. An dieser Stelle werden die Erfahrungen auf den drei Sprachen aus der Sicht der Spracherkennung verglichen. Das Projekt bietet allerdings die einmalige Gelegenheit, in weiten Bereichen der zwischenmenschlichen Kommunikation Vergleiche zwischen den drei Kulturen anzustellen. Das ILKD hat im Rahmen von Verbmobil Spracherkennungssysteme in allen drei Sprachen entwickelt [FGH+97, SKW97]. Die Erkennungsleistungen liegen im Bereich von 15% bis 25% Fehlerrate. Japanisch schneidet am besten ab, beim direkten Vergleich ist allerdings zu beachten, daß das japanische Vokabular aufgrund der Segmentierung mit 3000 Einträgen etwa um Faktor 3.5 unter dem des deutschen liegt. Außerdem zeigte sich, daß die japanischen Sprecher einen sehr disziplinierten Sprechstil aufweisen und sich eng an die vorgegebene Domäne halten. Dialoge zwischen deutschen Sprechern sind hingegen von Nebensprechen (Crosstalk) und vielen anderen spontanen Effekten gekennzeichnet, was die Erkennungsaufgabe erschwert. Allerdings profitiert Deutsch von dem fast doppelt so umfangreichen Trainingsmaterial. Englisch zeichnet sich durch extreme Wortverschleifungen und große Dialektvarianz aus. Insgesamt zeigte sich, daß dasselbe Spracherkennungssystem erfolgreich auf alle drei Sprachen angewendet werden kann und daß sich die eingesetzten Techniken in allen Sprachen in Verbesserungen gleicher Größenordnungen niederschlagen.

Umgangssprachliche Mensch-Mensch-Telefondialoge

Umgangssprache ist eine große Herausforderung für die Spracherkennung. Zum Vergleich mit gelesener Sprache wurde 1995 das CallHome-Korpus entwickelt, das umgangssprachliche Telefondialoge zwischen Familienmitgliedern enthält. Bislang wurden die Sprachen Englisch, Mandarin, Japanisch, amerikanisches Spanisch, Ägyptisch und Deutsch gesammelt. Für alle Sprachen ergab sich ein großer Leistungseinbruch im Vergleich zum geplanten Sprechstil, der sowohl durch die Veränderung des Sprechstils, die ungezwungene, ungrammatische Sprechweise als auch durch die geringe Menge verfügbarer Sprach- und Textdaten begründet wird.

[BCG+96] vergleichen die Sprachen Englisch, Spanisch, Mandarin-Chinesisch und Japanisch aus dem CallHome- und dem Ricardo-Korpus und damit Sprachen sehr unterschiedlicher Strukturen. Leider konnten die Autoren aufgrund der schlechten Erkennungsleistung (WE 78% bis 60%) keine Aussage zu Sprachunterschieden machen. Immerhin zeigt der Vergleich der OOV-Raten, der Vokabulare und der Perplexitäten, daß Spanisch und Japanisch aufgrund ihrer reichen Morphologie eine geringere Vokabularabdeckung haben als Englisch und Mandarin. Japanisch hat die niedrigste Perplexität. In [ZC98] wurden Englisch, Spanisch und Ägyptisch miteinander verglichen: Die Unterschiede in den Fehlerraten zwischen den Sprachen waren relativ gering und korrelierten mit der Menge verfügbarer Daten.

In [BMM+97] werden die CallHome-Sprachen Ägyptisch und Spanisch (Puerto Rico, Chile, Kolumbien, Spanien) mit dem BBN-System Byblos verglichen, das die NIST-Evaluation für Spanisch und Ägyptisch gewonnen hat. Das Ergebnis dieser Untersuchung ist, daß dieselbe Spracherkennungstechnik sowohl auf Spanisch als auch auf Arabisch anwendbar ist und daß die Techniken nahezu übereinstimmendes Verhalten auf den bisher untersuchten Sprachen zeigen. Sie weisen experimentell nach, daß die Größenordnungen der Verbesserungen in beiden Sprachen sehr ähnlich verlaufen und bestätigen damit die Kernaussage der SQALE, daß das Prinzip der gemeinsamen Nutzung von Spracherkennungssoftware möglich und sinnvoll ist.

Zusammenfassung
Die verschiedenen Forschungsgruppen kommen bei der Frage nach der Vergleichbarkeit zwischen Sprachen zu weitestgehend übereinstimmenden Ergebnissen. Sprachtechnologien lassen sich im Prinzip auf alle untersuchten Sprachen übertragen [YADA+97, BMM+97, AD99]. Nach der Initialisierung sind die Systeme allerdings noch durch sprachenspezifische Maßnahmen zu verbessern [DAK95]. Die bisherigen Untersuchungen haben gezeigt, daß die Forschung durch das Fehlen multilingualer Datenbasen erheblich behindert wurde [YADA+97, AD99]. Es fehlten multilinguale Datenbasen, die Sprach- und Textdaten für viele verschiedene Sprachen in einheitlicher Qualität und Domäne zur Verfügung stellen. Aufgrund dieser Tatsache sind Vergleiche zwischen Spracherkennern großer Wortschätze bislang sehr selten. Falls vorhanden, sind die Vergleiche auf wenige Sprachen begrenzt und von eingeschränkter Aussagekraft. Vergleiche zwischen den Phoneminventaren mehrerer Sprachen wurden durch den Einsatz von Phonemerkennern gezogen [LG93a, CAGADL97]. Diese Vergleiche waren durch die Forschung im Bereich Sprachenidentifizierung vorangetrieben worden.

5.3 Entwicklung der Basissysteme

Durch die Sammlung der multilingualen Datenbasis GlobalPhone wurde das wesentliche Hindernis zur Entwicklung großer Wortschatzerkenner in vielen Sprachen und deren Vergleich untereinander behoben. Mit der vorliegenden Arbeit werden nun erstmalig Erkenner für 10 verschiedene Sprachen miteinander verglichen, die auf derselben Domäne, im gleichen Sprechstil und in einheitlicher Datenqualität entwickelt wurden. Die 10 Sprachen decken dabei eine breite Varianz von Spracheigenschaften ab (vgl. Kapitel 4).

Die Entwicklung monolingualer Erkennungssysteme in vielen Sprachen, wie sie in diesem Abschnitt beschrieben wird, ist motiviert vom Streben nach einer weitgehenden Automatisierung des gesamten Prozesses. Darüber hinaus sollen die entstehenden monolingualen Erkenner zum Bau eines multilingualen Erkenners verwendet werden, weshalb die Grundstruktur (Vorverarbeitung, HMM-Topologie, Parameterzahl) der Basissysteme für alle Sprachen identisch sein soll.

In diesem Abschnitt werden die Bausteine beschrieben, die zum Bau eines Diktiererkenners benötigt werden. Zum Training und zum Evaluieren aller Erkenner wurde das JRTk-Spracherkennungstoolkit verwendet. Die beschriebenen Bausteine sind daher auf die Behandlung mit JRTk zugeschnitten, werden allerdings in derselben oder ähnlicher Form für die meisten State-of-the-art-Erkenner erforderlich sein.

5.3.1 Training mit JRTk

Es gibt zahlreiche Möglichkeiten, mit JRTk ein Erkennungssystem zu bauen, einige Schritte tauchen allerdings immer wieder auf und sollen im folgenden kurz erläutert werden.

Erstellung der Labels: Die zeitliche Zuordnung von Sprachmerkmalsvektoren zu HMM-Zuständen kann, wie in Kapitel 3 beschrieben, entweder durch den Viterbi- oder den Forward-Backward-Algorithmus erfolgen. Hier wurde ausschließlich der Viterbi-Algorithmus verwendet. Um die zeitaufwendige Berechnung der Zeitzuordnung nicht in jedem Entwicklungschritt wiederholen zu müssen, werden die berechneten Zeitzuordnungen in sogenannten *Label*-Dateien abgespeichert und dienen als Basis für weitere Entwicklungsschritte.

Kmeans-Initialisierung: Auf Basis der gefundenen Zeitzuordnungen werden Beispielvektoren aus der Trainingsstichprobe extrahiert; aus ihnen wird durch den Basic-ISODATA-Ballungsalgorithmus [DH73] für jede definierte Klasse ein eigenes Codebook und Mixturgewichteverteilung initialisiert. Nach diesem Schritt hat man ein voll kontinuierliches HMM auf den definierten Klassen.

Training: Nach der Initialisierung der Modelle werden dem System alle Trainingsdaten präsentiert und die HMM-Parameter mittels des EM-Algorithmus optimiert. Das einmalige Präsentieren der Daten wird als eine *Trainingsiteration* bezeichnet. Im hier verwendeten Training wurden jeweils 4 Iterationen durchgeführt, wobei die Zahl 4 auf Erfahrungswerten basiert.

Einführung der Subpolyphone: Initiale Systeme werden in der Regel auf Subphonemen als Klassen aufgebaut. In diesem Schritt wird die Kontextmodellierung durch Subpolyphone vorbereitet, indem zunächst für jedes Subpolyphon, das in der Trainingsdatenbasis repräsentiert ist, eine eigene Mixturverteilung definiert und trainiert wird. Subpolyphone desselben Subphonems teilen sich in diesem Stadium der Systementwicklung ein Codebook. Dieser Schritt führt also zu einem SCHMM.

Ballung: Mit Hilfe des in Kapitel 3.2.2.3 beschriebenen Ballungsverfahren werden die Subpolyphone zu einer definierten Anzahl Klassen zusammengeballt. Die Klassenzuordnung der Subpolyphone ergibt sich aus der Traversion des entstehenden Kontextentscheidungsbaumes.

Kmeans-Initialisierung und Training der neuen Klassen: Mit der Kmeans-Initialisierung werden die Codebooks und Mixturgewichte der neuen Klassen initialisiert und anschließend mit dem EM-Algorithmus 4 Iterationen trainiert. Der resultierende Erkenner ist wieder ein voll kontinuierliches System, diesmal aber auf Subpolyphon- statt auf Subphonem-Klassen.

Um mit JRTk einen Spracherkenner zu trainieren und zu evaluieren, benötigt man eine Datenbasis, welche die Sprachaufnahmen und die zugehörige Transkription der Äußerungen enthält. Zur Beschreibung der Aussprache eines gesprochenen Wortes muß ein geeignetes Phoneminventar festgelegt werden. Die Aussprachen aller im Training verwendeten und beim Evaluieren zu erkennenden Wörter müssen im Aussprachewörterbuch eingetragen werden. Insbesondere für die Spracherkennung auf großen Wortschätzen wird zur Evaluation ein Sprachmodell benötigt, das auf möglichst großen Textkorpora berechnet wird.

5.3.2 Datenbasis

Mit dem beschriebenen *„mapper“*-Tool (siehe Abschnitt 4.5.1) wurde für alle Sprachen eine *Datenbasis* erstellt, in der die akustische Repräsentation einer Äußerung ihrer textuellen Darstellung zugeordnet ist. Zum Trainieren und Testen der Systeme werden die Sprachdaten verwendet, wie sie in Tabelle 4.3 zusammengefaßt sind.

Bei den GlobalPhone-Textvorlagen handelt es sich, wie in Kapitel 4 erläutert, um Zeitungstexte. Die GlobalPhone-Datenbasis besteht somit aus Verschriftungen auf Wortebene nach dem Prinzip der Transkription. Wie bereits in Abschnitt 2.4.3 ausgeführt, ist es aus Aufwandsgründen nicht möglich, die Sprachdaten phonetisch zu verschriften. Trotzdem enthält die GlobalPhone-Datenbasis mehr Informationen als ein Großteil verfügbarer Datenbasen. Es wurden nämlich hörbare Effekte spontaner Sprache, wie Häsitationen, Wortabbrüche und Stottern in die Verschriftungen eingearbeitet (vgl. Abschnitt 4.5.2, Seite 79). In früheren Arbeiten der Autorin konnte gezeigt werden, daß durch die akustische Modellierung dieser artikulatorischen Geräusche und deren explizite Aufnahme in das Sprachmodell signifikante Verbesserungen der Erkennungsleistung erzielt werden können [SR95].

5.3.3 Romanisierung

Eine große Herausforderung dieser Arbeit war die Frage, wie es ermöglicht werden soll, Spracherkennungssysteme in vielen verschiedenen Sprachen zu erstellen, ohne über muttersprachliche Kenntnisse in diesen Sprachen zu verfügen. Ein neuer Aspekt gegenüber anderen Arbeiten ist damit die Frage, ob es möglich ist, ohne Kenntnisse in einer fremden Sprache ein Erkennungssystem zu bauen. Für den reinen Erkennungsvorgang ist es nach Ansicht der Autorin zumindest in einem frühen Entwicklungszustand nicht notwendig, die gesprochenen Äußerungen zu verstehen. Allerdings muß es zur Analyse und zum Vergleich der Ein- und Ausgabe des Erkenners für die Entwicklerin möglich sein, Zeichenketten zumindest lautlich interpretieren zu können. Auf fremden, nicht phonologischen Schriftsystemen ist ein lautlicher

Vergleich nicht möglich. Das Erlernen des jeweiligen Schriftsystems bedeutet einen großen Aufwand. Selbst Muttersprachler brauchen Jahre zum Erlernen der chinesischen Hanzi-Schrift oder der japanischen Kanji-Schrift. Aus diesem Grund wurde die *Romanisierung* aller fremdsprachlichen Schriftzeichen beschlossen. Der Begriff Romanisierung ist aus dem japanischen Sprachgebrauch übernommen und bezeichnet den Vorgang der Konvertierung in lateinische Schriftzeichen, insbesondere in 7-bit ASCII-Schriftzeichen.

Obwohl die Hauptmotivation der Romanisierung darin bestand, den Trainings- und Erkennungprozeß transparenter zu machen, hilft sie bei der Lösung eines zweiten wichtigen Problems; der automatischen Generierung von Aussprachewörterbüchern. Die Romanisierung eines Wortes, d.h. eine Zeichenkonvertierung nach dem Transkriptionsprinzip (vgl. Abschnitt 2.4.3), liefert nämlich per definitionem die Aussprache dieses Wortes.

Ein Nachteil der Romanisierung, insbesondere für die Diktieranwendung, besteht in der Notwendigkeit einer Rückkonvertierung der lateinischen Schriftzeichen in die entsprechende Schrift der Quellsprache, um das Diktat für Muttersprachler lesbar am Bildschirm darstellen zu können. Insbesondere für die ideographischen Schriften Hanzi und Kanji ist die Romanisierung aber keine eineindeutige Abbildung und kann zu Informationsverlusten führen. Dann kommt es bei der Rückkonvertierung zu Mehrdeutigkeiten, die, wenn überhaupt, nur durch Kontextwissen aufgelöst werden können. Trotz der Problematik des Umkehrungsprozesses, wurde angesichts der notwendigen Transparenz und der erwünschten Zusammenfügung mehrerer Sprachen zu einem monolithischen multilingualen Erkenner dieses Vorgehen gewählt.

Eng mit dem Problem der Romanisierung verknüpft ist die Segmentierung ideographischer Schriften. In Abschnitt 5.5.1.3 wird exemplarisch an den chinesischen Hanzi-Zeichen ein eigener Lösungsansatz zur Romanisierung, deren Umkehrung und zur Segmentierung erläutert.

5.3.4 Auswahl der Phoneminventare

Wie bereits in Kapitel 3 beschrieben, werden Wörter durch die Konkatenation von Phonemen modelliert. Die Basiseinheiten eines Spracherkenners sind somit die einzelnen Phoneme einer Sprache. Das Basisinventar eines Spracherkenners muß aber nicht notwendigerweise dem phonologisch definierten Phoneminventar einer Sprache entsprechen. So kann es beispielsweise geschehen, daß sehr selten repräsentierte Phoneme einer Sprache aus Gründen der robusten Modellschätzung nicht modelliert werden. Daher ist es nicht ganz richtig, vom Phoneminventar eines Spracherkenners zu sprechen. Da sich der Begriff aber allgemein eingebürgert hat, wird er hier weiterhin verwendet werden.

Die Wahl eines geeigneten Phoneminventars sollte sehr sorgfältig getroffen werden. Das entscheidende Kriterium bei der Auswahl eines geeigneten Phoneminventars ist ein guter Arbeitspunkt zwischen Differenzierungsfähigkeit und Generalisierungsfähigkeit der Basiseinheiten. Verwendet man ein zu kleines Inventar, kann es vorkommen, daß zwei verschiedene Wörter der Sprache nicht differenziert werden können. Verwendet man ein zu großes Inventar, leidet die Generalisierungsfähigkeit der Modelle, außerdem besteht die Gefahr von Inkonsistenzen im Aussprachewörterbuch. Die Entscheidung für ein geeignetes Inventar ist selbstverständlich sprachenabhängig. Beispielsweise benötigt man im Deutschen die zwei Phoneme /r/ und /l/ wie das Minimalpaar „Reiter - Leiter" zeigt. Dagegen handelt es sich im Japanischen um Allophone. Würde man im Japanischen /l/ und /r/ getrennt modellieren, dann entfiele weniger Trainingsmaterial auf jedes Modell und es entstünden Unklarheiten im Aussprachewörterbuch.

Zur Festlegung der Phoneminventare für alle GlobalPhone-Sprachen wurden mehrere Werke herangezogen, falls vorhanden insbesondere solche, die von der Spracherkennung motiviert waren. Sie sind im einzelnen in Tabelle 5.1 aufgeführt.

Sprache	verwendete Quellen
Chinesisch	abgeleitet aus [Ter87, Hie93, LM94], Details siehe [Rei97, Rei98]
Deutsch	übernommen aus [Kem99] (SAMPA für Deutsch [SAM98])
Englisch	übernommen aus [Rog97] (Aussprachelexikon der CMU)
Französisch	abgeleitet aus SAMPA für Französisch [SAM98]
Japanisch	abgeleitet aus [Hie93, Ter87, Ara93], Details siehe [SKW97]
Koreanisch	abgeleitet aus [Her94, Cho93], Details siehe [Kie99]
Kroatisch	abgeleitet aus [Bro93], Details siehe [Ras98]
Portugiesisch	abgeleitet aus SAMPA für Portugiesisch [SAM98]
Russisch	abgeleitet aus [Ter87, Hie93]
Spanisch	abgeleitet aus [LM94, Gav96]
Schwedisch	abgeleitet aus [Bon85] und SAMPA für Schwedisch [SAM98]
Türkisch	abgeleitet aus [Kor90, Ter87], Details siehe [Ç98]

Tabelle 5.1: Quellen der Phoneminventare

Die Phoneminventare aller Sprachen wurden in das IPA-Referenzschema eingetragen, um daraus später ein multilinguales Phonemset abzuleiten (vgl. Abschnitt 6.3.1). Die intern im System verwendete Bezeichnung eines Phonemmodells setzt sich aus dem an die IPA-Schreibweise oder das Graphem angelehnten Kürzel und einem 2-Buchstaben-Etikett zusammen, das die Information über die Sprachenzugehörigkeit des Phonems beschreibt. Zur Illustration wird in Tabelle 5.2 der konsonantische Teil des kroatischen Phonemsatzes gezeigt. In einigen Sprachen wurden die ursprünglichen Phoneminventare durch die Analyse zahlreicher Experimente nachträglich verbessert. Die endgültige Version der verwendeten Phoneminventare aller modellierten Sprachen, auf denen die dokumentierten Ergebnisse erzielt wurden, befinden sich in Tabelle 6.1 (S. 160).

	Bilabial	Labiodental	Dental	Alveopalatal	Palatal	Velar
Plosive	p b KR_p KR_b		t d KR_t KR_d			k g KR_k KR_g
Nasale	m KR_m		n KR_n		ɲ KR_nj	
Trill			r KR_r			
Frikative		f v KR_f KR_v	s z KR_s KR_z	s^j z^j KR_S KR_Z		x KR_x
Laterale			l KR_l		ʎ KR_lj	
Approxi-manten					j KR_j	
Affrikate			ts KR_c	tʃ dʒ KR_tS KR_dZ	tʃ^j dʒ^j KR_tj KR_dj	

Tabelle 5.2: Kroatische Konsonanten; IPA (oben) - interne Bezeichnung (unten)

5.3.5 Aussprachewörterbücher

Das Aussprachewörterbuch verknüpft die orthographische Darstellung jeder lexikalischen Einheit mit deren Aussprache, die auf der Basis des definierten Phoneminventars beschrieben wird (vgl. Abschnitt 2.4.1.4). Es ist damit eine zentrale Wissensquelle des Spracherkenners.

Aussprachevarianten
In den meisten Sprachen gibt es festgelegte kanonische Aussprachen einer lexikalischen Einheit. Daneben existiert allerdings eine breite Palette von Aussprachevarianten. Diese reichen von dialektbedingten Verschiebungen lexikalischer Einheiten über sprecherspezifische Varianten bis hin zu situationsbedingten Versprechern. Aus der wortbasierten Verschriftung einer Datenbasis geht nicht hervor, welche Aussprachevariante tatsächlich gesprochen wurde. Wie bereits beschrieben, ist beim Umfang der notwendigen Sprachdaten eine phonetische Umschrift zu aufwendig. Daher behilft man sich damit, daß man im Aussprachewörterbuch die wichtigsten Aussprachevarianten festhält. JRTk erlaubt im Training und in der Erkennung verschiedene Aussprachevarianten eines Wortes. Dazu wird ein Wortgraph aufgebaut, der alle alternativen Aussprachen beachtet.

Beispiele für dialektbedingte Verschiebungen aller lexikalischen Einheiten sind die Zeichenfolge „st“ und „sp“ im Deutschen. In der süddeutschen Variante wird „spitzer Stein“ gesprochen als [ʃ p I ts ɐ ʃ t aI n], in der norddeutschen Variante als [s p I ts ɐ s t aI n]. Ein Beispiel für eine sprecherspezifische Verschiebung ist die Aussprache von Altbundeskanzler Kohl, der das [ç] als [ʃ] spricht, beispielsweise in „Geschichte“. Situationsbedingte Versprecher sind Effekte, die beispielsweise durch Koartikulation entstehen, man versuche sich an: „Blaukraut bleibt Blaukraut und Brautkleid bleibt Brautkleid“.

Es ist aus Effizienzgründen nicht möglich, jede Variante in das Aussprachewörterbuch einzutragen, denn dies würde bedeuten, daß man durch Anhören der Sprachdaten alle Varianten aufsammeln müßte. Außerdem würde dadurch das Aussprachewörterbuch stark aufgebläht, wodurch sich die Verwechselbarkeit erhöht. In GlobalPhone wurde nur Sprache solcher Sprecher aufgezeichnet, die die jeweilige Hochsprache des Landes beherrschen. Trotzdem kommt es bisweilen zu leichten dialektbedingten Färbungen. Ins Aussprachewörterbuch wurden ausschließlich ausgewählte, im Vorfeld bekannte dialektale Verschiebungen eingetragen. Situationsbedingte Versprecher sind in der Datenbasis gesondert markiert (vgl. Abschnitt 4.5.2). Sie wurden auf spezielle Geräuschklassen abgebildet, weshalb Einträge ins Aussprachewörterbuch nicht notwendig sind.

Verfügbare Wörterbücher

Die erforderliche Vorarbeit für die Spracherkennung in 10 Sprachen bestand somit darin, die kanonische Aussprache und, falls vorhanden, wichtige Varianten für jede lexikalische Einheit nebst seiner orthographischen Darstellung bereitzustellen. Zunächst wurde versucht, Basisaussprachewörterbücher aus offiziellen Quellen zu gewinnen. Im diesem Bereich gibt es allerdings derzeit noch große Defizite [AD99], obwohl sich, wie in Abschnitt 4 beschrieben, viele Datenkonsortien um die Sammlung von Daten sehr bemühen. Die LDC [LDC00] stellt Wörterbücher für nur sechs Sprachen aus dem CallHome-Projekt zur Verfügung (Ägyptisch, Englisch, Spanisch, Mandarin, Japanisch, Deutsch), obwohl Sprachdaten in insgesamt etwa 20 Sprachen verfügbar sind. Die ELRA [ELR98] distribuiert Korpora in über 30 Sprachen, verfügt aber nur über größere Lexika in den drei Sprachen Deutsch, Französisch und Italienisch. Für Eigennamen existiert der Onomastica-Korpus, in dem die Aussprachen für 19 Sprachen verfügbar gemacht werden (ELRA-S0022). 1998 wurden von der ELRA Aussprachewörterbücher in 11 europäischen Sprachen versprochen, diese sind aber bis heute nicht verfügbar und wurden mittlerweile aus dem Programm zurückgezogen. Von den beschriebenen Quellen wurde das Spanische Wörterbuch hinzugezogen, allerdings konnte nur ein winziger Bruchteil von Wörtern der GlobalPhone-Domäne damit abgedeckt werden. Für die Sprachen Deutsch und Englisch konnten die in [Kem99] und [Rog97] beschriebenen Wörterbücher übernommen werden. Für das Französische stellte das offizielle LIMSI-Wörterbuch eine Hilfe dar. Für alle anderen Sprachen, nämlich Chinesisch, Kroatisch, Koreanisch, Japanisch, Portugiesisch, Russisch, Schwedisch und Türkisch gab es keinerlei brauchbare Quellen.

Automatische Generierung von Aussprachewörterbüchern
Bedingt durch die GlobalPhone-Domäne und damit den großen Umfang der Vokabularliste pro Sprache kam eine Erstellung der Aussprachewörterbücher von Hand nicht in Frage. Es wurde daher in den genannten Sprachen mit phonologischen Schriften eine automatische Generierung der Aussprachewörterbücher durchgeführt. Dazu wurde in jeder Sprache ein Graphem-zu-Phonem-Tool entwickelt, das im wesentlichen jedes Eingabewort von links nach rechts durchläuft und die Graphemfolge auf eine Folge von Phonemen des entsprechenden Phoneminventars abbildet. Abbildung 5.1 zeigt schematisch die prinzipielle Vorgehensweise bei der automatischen Generierung der Aussprachewörterbücher. Danach wird für jedes Eingabewort geprüft, ob das Wort selbst oder Teile davon in bereits vorhandenen Wissensquellen vorliegen. Unbekannte Graphemsequenzen werden durch die Graphem-zu-Phonem-Regeln ersetzt, falls vorhanden mit bereits bekannten Teilsequenzen zusammengefügt und in das finale Aussprachewörterbuch eingetragen. Die Abfrage von Teilsequenzen wurde für die agglutinierenden Sprachen und solche mit Kompositabildung eingefügt.

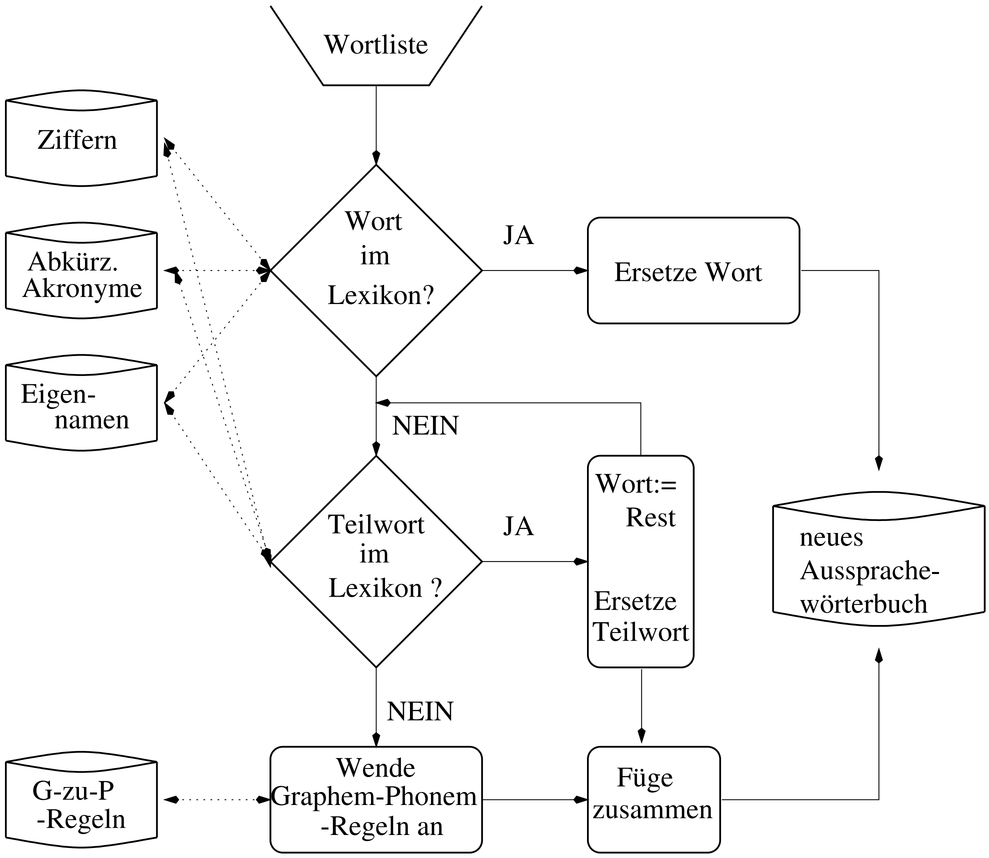


Abbildung 5.1: Automatische Generierung von Aussprachewörterbüchern

Um Sonderfälle handhaben zu können, werden Spezialwörterbücher verwendet. Bedingt durch die Auswahl der Domäne in der GlobalPhone-Datensammlung tauchen nämlich sehr viele fremdländische Eigennamen, Ziffernfolgen, Akronyme und Abkürzungen sowie Maßeinheiten auf. Deren Aussprachen weichen oftmals soweit vom Erwarteten ab, daß sie sich durch Graphem-zu-Phonem Regeln kaum erfassen lassen. Bei fremdländischen Eigennamen bürgern sich in einer Kultur häufig charakteristische aber nicht der Tatsache entsprechende Aussprachen ein. Für Akronyme gibt es zwei verschiedene Bildungsregeln, entweder sie werden als Buchstabensequenz gesprochen, wie ADAC oder als Wort behandelt wie NASA. Abkürzungen und Maßeinheiten haben in der Regel feste Aussprachen. Alle diese Sonderfälle wurden zwar zunächst automatisch generiert, dann aber von Muttersprachlern nachkorrigiert und in die Spezialwörterbücher eingetragen.

Orthografie		IPA	Beispiel (Deutsch)	Romanisierung	Wörterbuch
A	a	ɑ	araba	araba	A R A B A
B	b	b	baba	baba	B A B A
C	c	ʤ	cins (*Dsch*ungel)	cins	C I N S
Ç	ç	ʧ	çek (*Tsch*eche)	**tscH**ek	TSCH E K
D	d	d	dolmak	dolmak	D O L M A K
E	e	e	el	el	E L
F	f	f	fren	fren	F R E N
G	g	g	gemel	gemel	G E M E L
Ğ	ğ	ː	dağ (S*aa*l)	da**g2**	D A
H	h	h	halk (*H*ahn)	halk	H A L K
I	ı	ɨ	ırmak (geh*e*n)	**i2**rmak	I2 R M A K
İ	i	i	ilk	ilk	I L K
J	j	j	jale (Gara*g*e)	jale	J A L E
K	k	k	kara	kara	K A R A
L	l	l	lezzetli	lezzetli	L E Z E T L I
M	m	m	mutlu	mutlu	M U T L U
N	n	n	nazar	nazar	N A Z A R
O	o	o	okul	okul	O K U L
Ö	ö	ø	öğrenci	**o**ˆg2renci	OE R E N C I
P	p	p	para	para	P A R A
R	r	r	raks	raks	R A K S
S	s	s	sonuç (Ma*ss*e)	sonutscH	S O N U TSCH
Ş	ş	ʃ	şöyle (*Sch*aum)	**sscH**oŷle	SSCH OE Y L E
T	t	t	tarih	tarih	T A R I H
U	u	u	uzak	uzak	U Z A K
Ü	ü	y	üzüm	**u**ˆzuˆm	UE Z UE M
V	v	v	vermek (*W*asser)	vermek	V E R M E K
Y	y	j	yara (*j*a)	yara	Y A R A
Z	z	z	zarar (*S*aal)	zarar	Z A R A R
Veraltet		siehe oben	â, î, í, û,	a, i, i, u	

Tabelle 5.3: Türkische Orthographie, Romanisierung und Aussprache

Zur Illustration ist in Tabelle 5.3 ein Auszug aus dem türkischen Aussprachelexikon nebst Romanisierung dargestellt. Man sieht, daß die Beziehung zwischen Graphemen und Phonemen im Türkischen nahezu perfekt ist. Eine Gesamtübersicht der Graphem-zu-Phonem-Relation aller behandelter Sprachen befindet sich in Abschnitt 5.4.1.

5.3.6 Segmentierung

Die Erstellung eines Aussprachewörterbuches setzt die Definition der lexikalischen Einheiten voraus, die zur Spracherkennung verwendet werden sollen. Wie bereits in Abschnitt 2.4.2.2 beschrieben, existieren in den meisten Sprachen natürliche Einheiten, die in der Schrift durch Wortbegrenzungszeichen (meist ein Leerzeichen) von den rechten und linken Nachbarn abgesetzt sind. In den GlobalPhone-Sprachen Chinesisch und Japanisch ist das nicht der Fall. In diesen beiden Sprachen werden die Schriftzeichen ohne Begrenzer aneinandergereiht, so daß aus der schriftlichen Darstellung keine natürlichen Einheiten erkennbar sind. Für die Spracherkennung muß in diesen beiden Sprachen daher eine künstliche Segmentierung vorgenommen werden. Die Implementierung des Segmentierers ist am Beispiel der chinesischen Sprache im Abschnitt 5.5.1.3 beschrieben.

Für Sprachen mit gegebenen natürlichen Einheiten erhebt sich die Frage, ob diese Einheiten zur Spracherkennung geeignet sind. Im Sinne des Vokabularwachstums sind für Sprachen mit zu langen Einheiten Auftrennungen in kleinere, geeignete Untereinheiten sinnvoll oder notwendig. In Abschnitt 5.4.4 wird die durchschnittliche Länge natürlicher Worteinheiten für alle GlobalPhone-Sprachen gezeigt. Die Ergebnisse legen für die Sprachen Koreanisch und Türkisch nahe, kürzere Basiseinheiten zu verwenden. In Abschnitt 5.5.3 werden für beide Sprachen verschiedene Ansätze zur Auffindung geeigneter Basiseinheiten vorgestellt. In den weiteren Betrachtungen in diesem Abschnitt wird davon ausgegangen, daß eine geeignete Segmentierung in allen Sprachen vorliegt.

5.3.7 Statistische Sprachmodelle

Die Verschriftungen der GlobalPhone-Sprachdaten ergaben je nach Sprache bis zu 270.000 Wörter Textumfang. Zur akkuraten Schätzung von Trigramm-Wahrscheinlichkeiten, wie sie im JRTk-Erkenner zum Einsatz kommen, ist das entschieden zu wenig Textmaterial (vgl. Abschnitt 3.2.3). Aus diesem Grund wurden die vorhandenen Daten durch weitere Textdaten ergänzt. Für Deutsch und Englisch wurde das Material aus [Kem99, Rog97] übernommen. Für die anderen Sprachen wurden zusätzliche Textdaten wenn möglich aus denselben Quellen bezogen wie die Textdaten für die Sprachaufnahmen, um möglichst nah an der gesammelten Domäne zu bleiben. Im Laufe der Datensammlung waren jedoch viele Zeitungen, die 1996 noch uneingeschränkten Zugang zu ihren Archiven geboten hatten, dazu übergegangen nur noch einen geringen Ausschnitt auf dem Internet verfügbar zu machen.

Daher mußte in einigen Sprachen auf Textkorpora aus anderen Quellen zurückgegriffen werden. Alle verwendeten Quellen sind in Tabelle 5.4 zusammengestellt. Die Angaben geben einen Überblick über den Umfang und Zeitraum der gesammelten Textdaten in den einzelnen Sprachen.

Sprache	Wörter in [Mio]	Zeitung	Zeitraum	Quelle der Textdaten
Chinesisch	66	Peoples Daily	1991 -1996	http://www.snweb.com
	16.5	XinHua	1994-1996	CDrom LDC95T13 von [LDC00]
Koreanisch	43.3	Chosunilbo	10/95-08/98	http://www.chosun.com
Kroatisch	11	HRT	1995-1996	http://www.hrt.hr/vijesti
		Obzor		http://www.tel.hr/hrvatski-obzor
		Nacional		http:// www.nacional.hr
Japanisch		Nikkei Shinbun		http://www.nikkeihome.co.jp
Portugiesisch		Folha de São Paulo		http://www.uol.com.br/fsp
	0.65	ECI/MCI	1994	Borba/Ramsey European Corpus
	10	AFP	1993-1995	CDrom LDC95T11 von [LDC00]
Spanisch	8	Expansion		CDrom MLCC V1.0 von [ELR98]
	186	Newswire	1993-1995	CDrom LDC95T9 von [LDC00]
Türkisch	1.2	Zaman	1996-1998	http://www.zaman.com.tr
	7.7	Milliyet	1996-1998	http://www.milliyet.com
	0.9	Hürriyet	1996-1998	http://www.huerriyet.com.tr
	4.5	Superhaber	1996-1998	http://www.superonline.com
	0.3	Xn Online	1996-1998	http://www.xn.com.tr
Französisch	4	Le Monde	1994	CDrom BREF-Polyglot von [ELR98]

Tabelle 5.4: Textdaten-Quellen

Vor der Berechnung der Wortwahrscheinlichkeiten wurden die Texte normalisiert, d.h. soweit wie möglich fehlerbereinigt, von Interpunktion befreit und die Großschreibungen am Satzanfang entfernt, sofern sie nicht bedeutungsunterscheidend war. Für alle Sprachen wurde ein Trigramm-Sprachmodell berechnet. Dazu wurden die für die GlobalPhone-Sprachdaten eingeführten Geräuschklassen mit in das Sprachmodell aufgenommen. Frühere Arbeiten der Autorin hatten gezeigt, daß die Modellierung von Geräuschen im Sprachmodell zu Verbesserungen führt [SR95]. In Abschnitt 5.4.4 werden die wichtigen Charakteristika der Sprachmodelle wie OOV-Raten und Perplexitäten sowie die Vokabularwachstumsraten zwischen den Sprachen verglichen.

5.3.8 Initialisierung der Basiserkenner

Das in Abschnitt 5.3.1 beschriebene Verfahren setzt voraus, daß zur Erzeugung der initialen Zeitzuordnungen ein Erkenner zur Verfügung steht. In der Regel verwendet man die Parameter eines existierenden Erkenners, dessen Eigenschaften möglichst nahe an den gegebenen Bedingungen des neuen Erkenners liegen sollten. Kann kein

Erkenner mit den gewünschten akustischen Eigenschaften gefunden werden, sind die initialen Labels erfahrungsgemäß sehr schlecht.

In fast keiner GlobalPhone-Sprache war ein Erkenner zur Erzeugung der initialen Labels gegeben. Am ILKD lagen zum Beginn dieser Arbeit lediglich Spracherkenner in den vier Sprachen Deutsch, Englisch, Japanisch und Spanisch vor. Zur Erstellung der GlobalPhone-Basiserkenner mußten daher zunächst geeignete Verfahren zur Initialisierung entwickelt werden.

In früheren Arbeiten der Autorin war gezeigt worden, daß ein japanischer Erkenner sehr effizient und erfolgreich mit deutschen Modellen initialisiert werden kann [SKW97]. Es lag daher nahe, auch die GlobalPhone-Erkenner mit den Modellen anderer Sprachen zu initialisieren. Dazu wurde auf den in Tabelle 5.5 aufgeführten Erkennern aufgebaut. Alle vier Erkenner waren auf spontansprachliche Daten im Terminabspracheszenario trainiert worden. Bei der Initialisierung der GlobalPhone-Erkenner handelt es sich somit nicht nur um einen Übergang auf eine neue Sprache, sondern auch um einen Übergang von einem spontanen zu einem gelesenen Sprechstil, von Dialog- zu Monologsprachdaten und um einen Übergang von dem Terminabspracheszenario mit mittelgroßem Wortschatz zum Zeitungstextszenario mit sehr großem Wortschatz. Die zu erwartende Qualität der initialen Labels ist daher sehr gering.

Sprache	Vokabular	Phoneme	WE	
			CI	CD
Deutsch	5438	65	36%	14%
Englisch	2601	53	39%	23%
Japanisch	1879	39	24%	10%
Spanisch	3939	47	40%	17%

Tabelle 5.5: Spracherkennungssysteme in spontan gesprochener Sprache [WE in %]

Auf der Basis der in Tabelle 5.5 aufgeführten kontextunabhängigen Erkenner in den vier Sprachen Deutsch, Englisch, Japanisch und Spanisch wurde ein Phonemmodellpool erstellt, der alle 204 sprachenspezifischen Modelle vereint. Dieser Phonempool diente als Ausgangsbasis zur Initialisierung der Parameter aller GlobalPhone-Erkenner [SW97].

QuickBoot

In diesem Abschnitt wird untersucht, welche Verfahren sich am besten eignen, um eine möglichst akkurate und schnelle Initialisierung zu bewerkstelligen. Dazu wurden vier Verfahren erstellt und verglichen.

- **Flat-Start-Initialisierung (Flat)**: Beim Flat-Start-Verfahren werden die Parameter der akustischen Modelle mit Zufallszahlen initialisiert. Der Vorteil liegt darin, daß kein Vorwissen benötigt wird, der Nachteil in der Gefahr, beim Training in ein lokales Optimum zu geraten.

- **Uniforme Initialisierung (Generic)**: Bei diesem Verfahren werden alle akustischen Modelle durch dieselben Parameter initialisiert. Im vorliegenden Fall stammen diese Parameter von einem generischen Modell, das auf einer Mischung aller Phoneme der deutschen Sprache trainiert wurde. Der Vorteil gegenüber dem Flat-Start-Verfahren besteht darin, daß dieses Modell sprachenspezifische Informationen statt Zufallszahlen enthält.

- **Initialisierung mit Modellen aus *einer* anderen Sprache (Boot-1L)**: Bei diesem Verfahren werden die Phonemmodelle einer einzigen Quellsprache (hier Deutsch) auf die Zielsprache übertragen. Die Parameter der Zielphoneme werden durch möglichst ähnliche Modelle der Quellsprache gesetzt. Je größer die Ähnlichkeiten, um so bessere initiale Labels sind zu erwarten.

- **Initialisierung mit Modellen aus *mehreren* anderen Sprachen (Boot-4L)**: Dieses Verfahren unterscheidet sich von dem vorangehenden dadurch, daß die Auswahlmöglichkeit der Quellphoneme erweitert wird. Statt wie bisher nur Deutsch kann nun aus dem 4-sprachigen Phonempool mit den zusätzlichen Sprachen Englisch, Spanisch und Japanisch gewählt werden. Der Vorteil liegt in der differenzierteren Auswahl von Quellphonemen.

Die Experimente zu den vier Initialisierungsmöglichkeiten werden zunächst exemplarisch an der Sprache Kroatisch durchgeführt [Ras98]. Zum Training standen 5 Stunden kroatisches Trainingsmaterial zur Verfügung. Die Abbildungen der Quell- auf die Zielphoneme wurden für die Boot-1L- und die Boot-4L-Methode heuristisch durch das IPA-Referenzschema bestimmt und sind in Tabelle 5.6 beschrieben.

```
                    Der QUICKBOOT-Algorithmus
Schritt 0: Festlegung der Phonemabbildung
Schritt 1: Modellinitialisierung entsprechend Schritt 0
Schritt 2a: Labels erstellen, Kmeans-Initialisierung
Schritt 2b: Training der entstehenden Klassen
Schritt 3: Iteration von Schritt 2
```

Die Ergebnisse nach den jeweiligen Schritten des QUICKBOOT-Algorithmus am Beispiel des kroatischen Erkenners in Tabelle 5.6 zeigen, daß die Initialisierung mit Phonemmodellen einer oder mehrerer Sprachen weit bessere Leistungen erzielt, als mit dem Verfahren des Flat-Start und der uniformen Initialisierung. Ein Vergleich nach weiteren Iterationen zeigt, daß das Flat-Start fast an die beiden Verfahren Boot-1L und Boot-4L heranreicht, allerdings erst nach dreifacher Berechnungszeit.

Ziel	Boot-1L	Boot-4L	Generic	Ziel	Boot-1L	Boot-4L	Generic
a	DE_a	DE_a	DE_Gen	m	DE_m	DE_m	DE_Gen
b	DE_b	DE_b	DE_Gen	n	DE_n	DE_n	DE_Gen
ts	DE_ts	DE_ts	DE_Gen	ɲ	DE_j	SP_ɲ	DE_Gen
d	DE_d	DE_d	DE_Gen	o	DE_ɔ	DE_ɔ	DE_Gen
dʒʲ	DE_ʃ	EN_dʒ	DE_Gen	p	DE_p	DE_p	DE_Gen
dʒ	DE_ʃ	EN_dʒ	DE_Gen	r	DE_r	DE_r	DE_Gen
e	DE_eː	DE_eː	DE_Gen	s	DE_s	DE_s	DE_Gen
f	DE_f	DE_f	DE_Gen	sʲ	DE_ʃ	DE_ʃ	DE_Gen
g	DE_g	DE_g	DE_Gen	t	DE_t	DE_t	DE_Gen
x	DE_h	DE_h	DE_Gen	tʃʲ	DE_tʃ	EN_tʃ	DE_Gen
i	DE_ɪ	DE_ɪ	DE_Gen	tʃ	DE_tʃ	EN_tʃ	DE_Gen
j	DE_j	DE_j	DE_Gen	u	DE_ʊ	DE_ʊ	DE_Gen
k	DE_k	DE_k	DE_Gen	v	DE_v	DE_v	DE_Gen
l	DE_l	DE_l	DE_Gen	z	DE_z	DE_z	DE_Gen
λ	DE_j	DE_j	DE_Gen	zʲ	DE_ʃ	DE_ʃ	DE_Gen

Tabelle 5.6: Abbildung auf kroatische Phoneme für Boot-1L und Boot-4L

Das uniforme Verfahren bleibt weiterhin suboptimal. Die Initialisierung mit den Modellen aus vier Sprachen ist der mit nur einer Sprache leicht überlegen, obwohl nur 5 Modelle ausgetauscht wurden. Offensichtlich lohnt sich die differenziertere Auswahl von Phonemmodellen und wirkt sich positiv auf die Erkennungsleistung aus.

Initialisierung	Boot-4L	Boot-1L	Flat	Generic
Schritt 1	65.8	66.7	96.2	96.8
Schritt 2	63.3	65.8	95.2	96.5
Schritt 3	62.8	63.2	75.1	85.7
1. Iteration Schritt 3			65.6	76.0
2. Iteration Schritt 3			63.5	71.0

Tabelle 5.7: Vergleich unterschiedlicher Initialisierungen des kroatischen Erkenners [WE in %]

In einem anschließenden Experiment überprüften wir, ob die Initialisierung mit Modellen aus mehreren Sprachen auch für andere Sprachen als Kroatisch vergleichbare Ergebnisse erzielt. Die Ergebnisse sind in Abbildung 5.2 für Phonemerkennungsraten in den 3 Sprachen Kroatisch, Türkisch und Chinesisch zu sehen. Für die chinesische Sprache wird deutlich, daß der Phonempool bestehend aus den vier Sprachen Spanisch, Deutsch, Englisch und Japanisch die chinesische Sprache nicht annähernd abdeckt. Daher sind die Resultate direkt nach der Initialisierung der Modelle (Schritt 1) noch sehr schlecht. Trotzdem führt der QUICKBOOT-Algorithmus mit Boot4L-Initialisierung bereits nach dem 3. Schritt zu Ergebnissen, die sogar die beiden anderen Sprachen übertreffen.

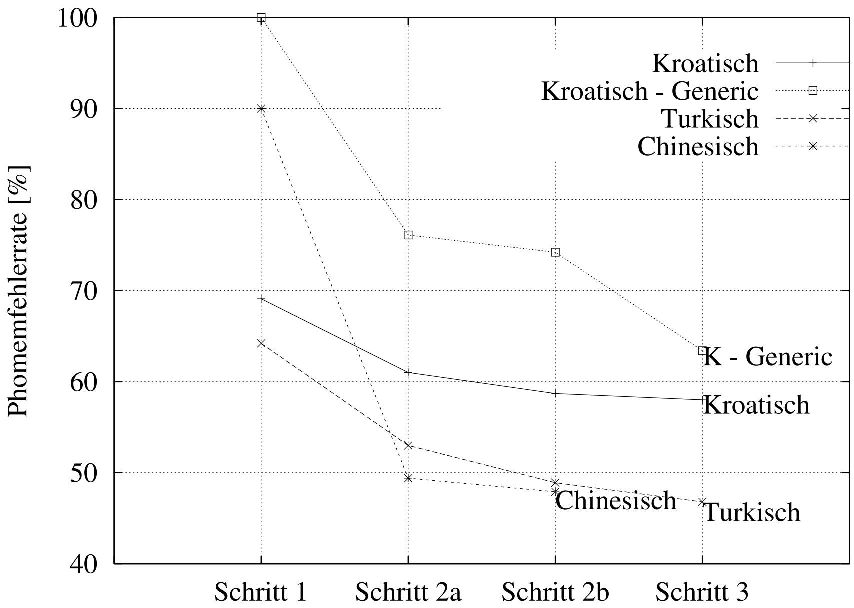


Abbildung 5.2: Phonemfehlerraten mit QUICKBOOT für 3 Sprachen

5.3.9 Weiterentwicklung der Basiserkenner

Das Ziel der Entwicklung von Basiserkennern in allen GlobalPhone-Sprachen war zum einen die Vergleichsmöglichkeit zwischen vielen Sprachen zu bieten, zum anderen aber auch der Bau eines multilingualen Erkennersystems. Beides setzt voraus, daß die Basiserkenner eine einheitliche, sprachenübergreifende Struktur haben. Mit Struktur sind hier die Vorverarbeitung, die HMM-Topologie und die Zahl der modellierten Parameter gemeint. Die initialisierten Erkenner wurden daher mittels der oben beschriebenen Trainingsroutinen in allen 10 Sprachen identisch trainiert. Dazu wurde das Entwicklungsverfahren automatisiert, um den gesamten Vorgang möglichst effizient zu gestalten. Im Anschluß daran wurden die Erkenner analysiert. Dabei wird untersucht, ob und welche sprachenspezifischen Maßnahmen erforderlich sind, um Besonderheiten der Sprache zu modellieren. Diese Maßnahmen werden in Abschnitt 5.5 beschrieben. Im folgenden werden zunächst die Strukturmerkmale der Basiserkenner erläutert, die in allen Sprachen identisch sind.

5.3.9.1 Vorverarbeitung

Die GlobalPhone-Originaldaten wurden, wie in Kapitel 4 beschrieben, in 48 kHz-Qualität mit 16-bit-Auflösung digital aufgezeichnet und in eine Aufnahme pro Äußerung zerlegt. Der erste Schritt der Vorverarbeitung bestand darin, die Abtastrate von 48 kHz auf 16 kHz zu reduzieren, was für die Spracherkennung eine völlig ausreichende Abtastfrequenz darstellt. Anschließend wird von den Abtastwerten der Aufnahme der Mittelwert aller Abtastwerte subtrahiert, um einen eventuellen Offset des A/D-Wandlers auszugleichen. Für die Kurzzeitanalyse werden aus dem Signal jeweils Zeitsegmente von 16 ms ausgeblendet (entspricht 256 Abtastwerten), wobei man annimmt, daß das Signal über diesen Segmenten stationär ist. Das zur Ausblendung verwendete Hamming-Fenster wird dabei mit einem Versatz von 10 ms über das Signal geschoben, so daß sich benachbarte Segmente um jeweils 6 ms überlappen. Anschließend werden auf den je 256 Abtastwerten mittels einer diskreten Fouriertransformation 129 Spektralkoeffizienten berechnet. Auf dem resultierenden Leistungsspektrum wird nun eine Vokaltraktlängennormierung vorgenommen. Dazu wird ein Längenparameter α im Bereich 0.8-1.2 geschätzt und das Spektrum stückweise linear transformiert. Die alle 10 ms anfallenden 129 Koeffizienten werden mit einer mel-skalierten Filterbank auf 30 Dimensionen reduziert. Bei diesem *Mel-Scaling* werden nach gehörphysiologischen Gesichtspunkten jeweils mehrere Frequenzbänder zusammengefaßt, so daß die Frequenzauflösung in niedrigen Frequenzbereichen weniger und in hohen Frequenzbereichen stärker reduziert wird. Die resultierenden 30 Mel-Scale-Koeffizienten werden logarithmiert und durch eine inverse Fouriertransformation in 30 Cepstral-Koeffizienten überführt. Von diesen 30 Koeffizienten werden die ersten 13 weiterverwendet. Mittels der cepstralen Mittelwertsubtraktion werden die Cepstren mittelwertfrei gemacht. Alle 10 ms werden somit Vektoren bestehend aus 13 Koeffizienten berechnet. Diese Vektoren sind stationäre Momentaufnahmen des Sprachsignals und werden daher durch dynamische Merkmale ergänzt. Dazu werden die ersten und zweiten Ableitungen der 13 Cepstren approximiert. Außerdem wird die Nulldurchgangsrate des Signals und die Signalenergie nebst erster und zweiter Ableitung berechnet. Die resultierenden 43 Merkmale werden zu einem Vektor zusammengefaßt. Durch eine LDA-Transformation wird die Dimension auf 32 reduziert.

5.3.9.2 Akustische Modellierung

In allen GlobalPhone-Erkennern wird ein Phonem durch ein in Abbildung 5.3 dargestelltes Links-Rechts-HMM mit 3 Zuständen modelliert. Es sind von jedem Zustand aus nur Übergänge der Reichweite 0 oder 1 erlaubt, d.h. jedes Phonem muß mit einer Mindestlänge von 3 Zuständen durchlaufen werden. Da jeder Zustand einen Frame von 10 ms konsumiert, beträgt die minimale Dauer eines Phonems 30 ms. Die drei Zustände werden unterschieden in einen Beginnzustand, einen mittleren Zustand und einen Endzustand und spiegeln die Dynamik eines Phonems wieder. Die

Zustände werden entsprechend mit den Markierungen *-b* für Beginn, *-m* für Mitte und *-e* für Ende versehen. Man spricht daher von einem Phonem /A/, das in die Subphoneme /A-b,A-m,A-e/ zerfällt.

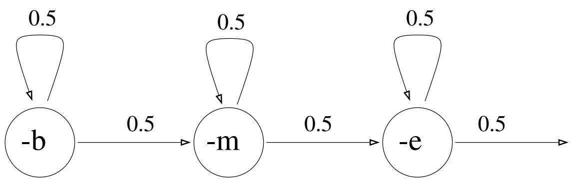


Abbildung 5.3: HMM-Topologie für Phoneme des GlobalPhone-Erkenners

Für das Stille-Modell wird ein 4-Zustands-HMM verwendet wie in Abbildung 5.4 dargestellt. Es wird somit eine Mindestdauer von 40 ms für ein Stille-Modell erzwungen, so daß die Wahrscheinlichkeit einer fehlerhaften Einfügung von Stille verringert wird. Im Gegensatz zum HMM für Phoneme wird beim Stille Modell allerdings nicht zwischen Anlaut, konstanter Phase und Auslaut unterschieden.

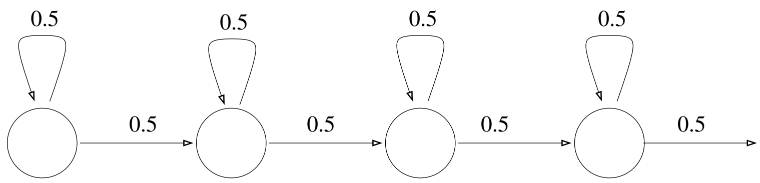


Abbildung 5.4: HMM-Topologie für Stille (SILENCE) des GlobalPhone-Erkenners

Die Zustandsübergangswahrscheinlichkeiten der HMM-Modelle sind uniform auf 0.5 gesetzt und werden nicht trainiert. Die Emissionswahrscheinlichkeiten werden durch Gaußsche Mischverteilungen modelliert. Eine Mischverteilung besteht aus 16 bzw. 32 Normalverteilungen, die entsprechend der oben erläuterten Vorverarbeitung eine Dimension von 32 haben. Da durch die LDA in der Vorverarbeitung die Dimensionen der Merkmalsvektoren im Mittel über alle Klassen dekorreliert werden, wird nicht die vollständigen Kovarianzmatrix sondern nur ihre Hauptdiagonale modelliert, so daß die multivariate Normalverteilung in das Produkt ihrer univariaten Komponenten zerfällt.

Zur *kontextunabhängigen* Modellierung wird jedes Subphonem mit 32 Verteilungen modelliert. Je nach Sprache werden daher 30 bis $100 \times 32 = 960$ bis 3200 Verteilungen geschätzt. Im *kontextabhängigen* Fall wird jedes Subpolyphon mit 16 Gaußverteilungen modelliert. Die Zahl der Subpolyphone wird beim Ballen auf 3000 begrenzt, somit werden pro Sprache $3000 \times 16 = 48000$ Verteilungen geschätzt.

5.3.9.3 Fragenkatalog

Zur Erzeugung des Kontextentscheidungsbaumes wurde für jede Sprache eine Menge von phonetisch bzw. phonologisch motivierten Kontextfragen entwickelt. Im Hinblick auf das multilinguale System sind die Fragenkataloge an der Kategorisierung des IPA-Schemas orientiert. Dies erleichtert eine spätere Zusammenführung der einzelnen Fragenkataloge zu einem multilingualen Fragenkatalog, der alle 10 Sprachen abdeckt. Zur Anschauung zeigt Tabelle 5.8 den Fragenkatalog für die türkische Sprache.

Kontextfrage	Liste der Phoneme
PHONES	@ SIL +QK +hGH A B C TSCH D E F G GJ H I2 I J K L M N O OE P R S SSCH T U UE V Y Z
NOISES	+QK +hGH
CONSONANT	B C TSCH F G GJ H J K L M N P R S SSCH T V Y Z
STOP	P T TSCH K B D C G
STOP-UNVOICED	P T TSCH K
STOP-VOICED	B D C G
FRICATIVE	F S SSCH V J Z
FRI-UNVOICED	F S SSCH
FRI-VOICED	V Z J
NASAL	M N
GLIDE	Y GJ H
BILABIAL	B P M
LABIODENTAL	F V
ALVEODENTAL	D L N R S T Z
PALATOALVEOLAR	C TSCH J SSCH
PALATAL	G K L Y
VELAR	G GJ K L
VOWEL	A E I I2 O OE U UE
VO-BACK	A I2 O U
VO-FRONT	E I OE UE
VO-FRO-UNROUND	E I
VO-FRO-ROUND	OE UE
VO-BAC-UNROUND	A I2
VO-BAC-ROUND	O U
VO-HIGH	I I2 U UE
VO-LOW	A E O OE
VO-LOW-UNROUND	A E
ROUND	O OE U UE
UNROUND	A E I I2

Tabelle 5.8: Phonetische Kontextklassen für das Türkische

5.3.9.4 Zusammenfassung der Systementwicklung

Die Abbildung 5.5 faßt die in diesem Abschnitt beschriebene Entwicklung der monolingualen Basiserkenner schematisch zusammen. Die bei einigen Sprachen notwendigen Maßnahmen zur Behandlung sprachenspezifischer Besonderheiten werden im Abschnitt 5.5 detailliert beschrieben.

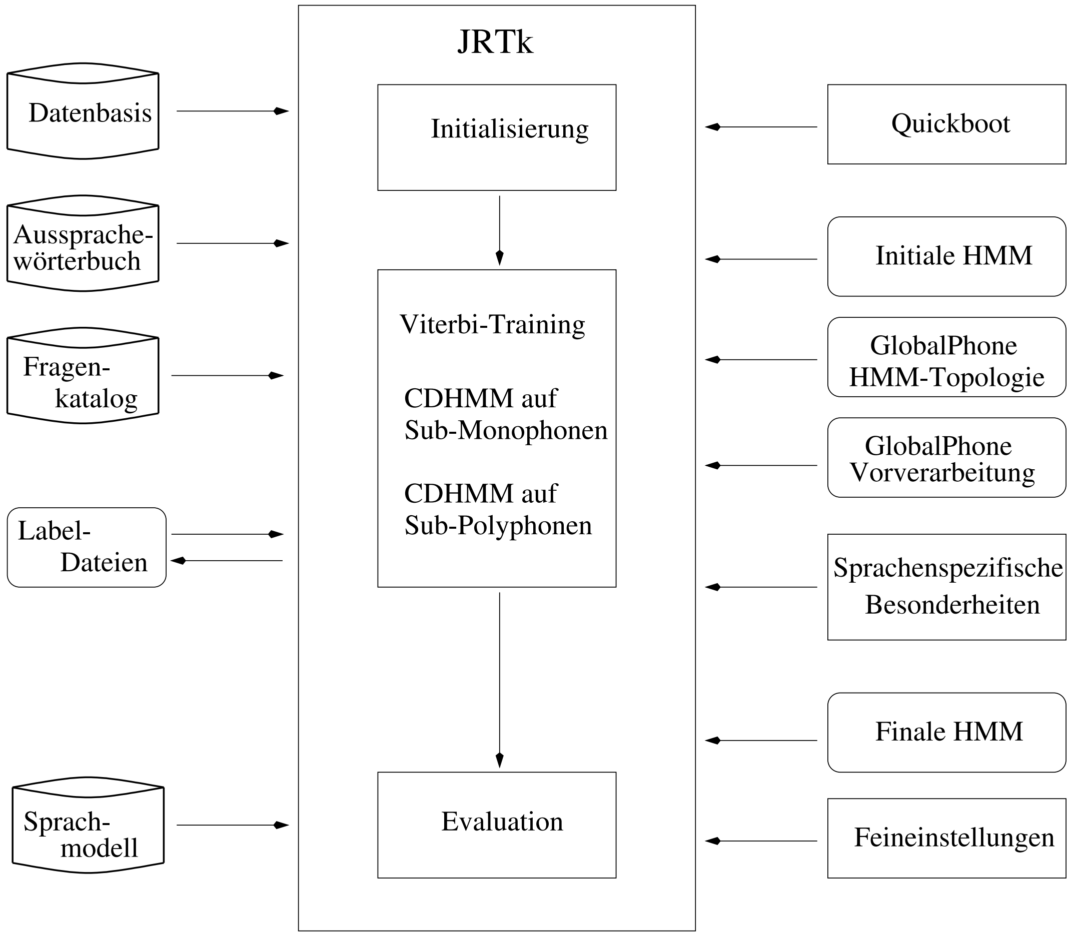


Abbildung 5.5: Systementwicklung der GlobalPhone-Basiserkenner

5.4 Vergleiche zwischen den Sprachen

In diesem Abschnitt werden die GlobalPhone-Sprachen untereinander auf verschiedenen Strukturebenen, wie sie in Kapitel 2 beschrieben wurden, verglichen. Das Ziel ist zum einen, die für die automatische Spracherkennung wesentlichen Aspekte einzelner Sprachen herauszuarbeiten - die Resultate sollen als Kriterien in die sprachenspezifische Sonderbehandlung einfließen. Zum anderen wird das Ziel verfolgt, einen multilingualen Erkenner zu entwerfen. Aus den Ergebnissen der Vergleiche zwischen den Sprachen sollen daher Konsequenzen abgeleitet werden, die bei der Entwicklung eines solchen Systems beachtet werden müssen.

5.4.1 Unterschiede im Schriftsystem

Ein Blick auf die Tabelle 5.9 verdeutlicht die gewaltigen Unterschiede zwischen den GlobalPhone-Sprachen hinsichtlich ihrer Schriften. Die Alphabetschriften kommen mit maximal 100 Zeichen aus (Groß- und Kleinbuchstaben werden getrennt gezählt), während die ideographischen Schriften bis zu 60.000 Zeichen verwenden. Die chinesischen Schriftzeichen wurden von den Japanern in die Kanji-Schrift übernommen, die Koreaner verwenden sie ebenfalls für die Schreibung chinesischer Lehnwörter. Die koreanische Hangul-Schrift ist eine aus nur 40 Zeichen bestehende phonologische Schrift, die allerdings zu 5601 Silbenzeichen (Gulja) kombiniert werden. Für diese drei Schriften werden zur Darstellung 16-bit Zeichen verwendet. Für alle anderen Alphabetschriften genügen 8-bit Codierungen.

Sprache	Schrift	Zeichenzahl
Arabisch	Arabisches Alphabet	100
Chinesisch	Hanzi	60.000
Deutsch	Lateinisches Alphabet + ä,Ä,ö,Ö,ü,Ü,ß	59
Koreanisch	Hangul / Gulja	40 / 5601
	chinesische Zeichen	< 10.000
Kroatisch	Lateinisches Alphabet + ć,Ć,č,Č,š,Š,ž,Ž,đ ,D	62
Japanisch	Hiragana	46
	Katagana	46
	Kanji	6000-7000
Portugiesisch	Lateinisches Alphabet + 2 $\times$ 14 Sonderzeichen	80
Russisch	Kyrillisches Alphabet	57
Schwedisch	Lateinisches Alphabet + ä,Ä,ö,Ö,,å,Å	58
Spanisch	Lateinisches Alphabet + á,Á,é,É,í,Í,ó,Ó,ú,Ú,ü,Ü,ñ,Ñ	66
Türkisch	Lateinisches Alphabet (ohne x) + ı,İ,ö,Ö,ü,Ü,ğ,Ğ	58

Tabelle 5.9: Schriftsysteme und Zeichenanzahlen der GlobalPhone-Sprachen

Graphem-zu-Phonem-Relation

In den ideographischen Schriften Hanzi und Kanji existiert keine Relation zwischen Orthographie und Aussprache. Durch die Romanisierung, die in Abschnitt 5.3.3 beschrieben wird, werden diese Schriften in eine quasiphonologische Schrift transformiert. Ausgehend von der romanisierten Form lassen sich in der japanischen und chinesischen Sprache die Aussprachen mit 50 bzw. 70 einfachen Regeln generieren.

Die koreanische Romanisierung (siehe 5.5.1.1) läßt keine triviale Transformation in Aussprachen zu, weil es sehr viele Assimilationserscheinungen an Silbengrenzen gibt. Unter den restlichen GlobalPhone-Sprachen mit phonologischen Schriften gibt es eine große Varianz in der Beziehung zwischen Graphemen und Phonemen. Aus der (gerundeten) Zahl der Graphem-Phonem-Regeln in Tabelle 5.10, die zur automatischen Generierung des Aussprachelexikons notwendig waren, lassen sich die Sprachen in eine ungefähre Rangfolge einordnen. Danach haben Kroatisch und Türkisch eine nahezu perfekte Graphem-zu-Phonem-Relation. Dann kommen in absteigender Folge Russisch, Spanisch, Portugiesisch, Schwedisch, Französisch, Deutsch und Englisch.

Sprache	Graphem-Phonem Regeln
Kroatisch	30
Türkisch	30
Russisch	50
Spanisch	120
Portugiesisch	130
Schwedisch	250
Romanisiert	
Japanisch	50
Chinesisch	70
Koreanisch	200

Tabelle 5.10: Anzahl benötigter Graphem-zu-Phonem-Regeln

In Zeitungstexten kommen für gewöhnlich viele Ziffern und Zahlen vor. Zur Erzeugung des Aussprachewörterbuchs wurden daher in jeder Sprache für die Zahlwörter und Ordnungszahlen Aussprachegeneratoren geschrieben. In allen untersuchten Sprachen existieren feste Regeln, nach denen Ziffernfolgen ausgesprochen werden. In Deutsch, Englisch, Französisch, Kroatisch, Portugiesisch, Russisch, Schwedisch, Spanisch und Türkisch erfolgt die Zahlenbildung nach dem Zehnersystem. Zur Aussprache werden die einzelnen Bestandteile beginnend von der höchstwertigen Stelle zur niedrigswertigen ausgesprochen. Ausnahme sind die deutsche Sprache, in der die Zehner und Einer vertauscht werden und die französische Sprache, in der die 80 als 4×20 gesprochen wird.

5.4.2 Phonetische Unterschiede

Phoneminventar

Tabelle 5.11 zeigt die Anzahl der Phoneme in 12 GlobalPhone-Sprachen, die zur Modellierung in den Spracherkennern verwendet werden. Es handelt sich also um die unter dem Gesichtspunkt der Spracherkennung geeigneten Basiseinheiten. In

Sprache	Phonemanzahl
Chinesisch	137 Phoneme (Toneme)
Chinesisch	48 Phoneme (keine Toneme)
Deutsch	43 Phoneme
Englisch	43 Phoneme
Französisch	38 Phoneme
Japanisch	31 Phoneme
Koreanisch	41 Phoneme
Kroatisch	30 Phoneme
Portugiesisch	46 Phoneme
Russisch	47 Phoneme
Spanisch	40 Phoneme
Schwedisch	49 Phoneme
Türkisch	29 Phoneme
$\sum$	485 (574) Phoneme

Tabelle 5.11: Phoneminventare für 12 GlobalPhone-Sprachen

jeder Sprache werden zusätzlich noch ein Modell für Stille und zwei Modelle für artikulatorische Geräusche und andere spontane Effekte modelliert. Eine Übersicht über die Schnittmenge der einzelnen Phoneminventare befindet sich in Tabelle 6.1 (siehe S. 160).

Die Größe der Phoneminventare schwankt zwischen 29 für das Türkische und 49 Phonemen für das Schwedische. Die Tabelle 5.11 zeigt, daß zur Modellierung der japanischen, kroatischen und türkischen Sprache sehr kleine Phoneminventare ausreichen. Französisch, Englisch und Deutsch benötigen eine etwas höhere Anzahl von Phonemen und Portugiesisch, Russisch und Schwedisch haben ein noch ausgeklügelteres Phonemsystem. Chinesisch ist eine Tonsprache, die zur Bedeutungsunterscheidung 5 Toneme verwendet. Sollen die Toneme modelliert werden, dann benötigt man im Chinesischen 137 Phoneme. Ohne Toneme läßt sich die Phonemmenge auf 48 reduzieren. In Abschnitt 5.5.2 werden chinesische Erkennungssysteme mit und ohne expliziter Modellierung der prosodischen Merkmale analysiert.

Verhältnis zwischen Konsonanten und Vokalen

Tabelle 5.12 zeigt das Verhältnis zwischen Konsonanten (C) und Vokalen (V) jeweils im Phoneminventar, im Aussprachewörterbuch und im Gesamtkorpus. Laut der UPSID-Studie [Mad84] an 320 untersuchten Sprachen hat eine „typische“ Sprache etwa doppelt soviele Konsonanten wie Vokale im Phoneminventar. Demnach sind Japanisch, Russisch und Türkisch in diesem Sinn Vertreter der typischen Sprache. Kroatisch fällt durch sein kleines Vokalinventar auf. Portugiesisch hat dagegen ein überdurchschnittlich großes Vokalinventar, fast alle Vokale haben nasale und orale Variante. Nach [Cry95] beträgt das Verhältnis zwischen Konsonanten und Vokalen in gesprochener Sprache etwa 60 zu 40. Von allen GlobalPhone-Sprachen findet man dieses Verhältnis nur im Deutschen. Selbst in der Sprache Russisch, die uns Westeuropäern so konsonantenreich erscheint, überschreitet der Quotient zwischen Konsonanten und Vokalen im Korpus nicht den des Deutschen, sondern ist nahe 1 wie in allen anderen GlobalPhone-Sprachen. Obwohl wie bereits festgestellt Japanisch doppelt so viele Konsonanten wie Vokale im Phoneminventar hat, herrscht in der gesprochenen Sprache ein ausgeglichenes Verhältnis. Dies ist durch die Mora-Struktur der Sprache bedingt. Abgesehen von den besprochenen Ausnahmen ist das Konsonant-Vokal-Verhältnis im Phoneminventar, Wörterbuch und gesprochener Sprache recht stabil. Es werden also nicht etwa vokalreichere oder vokalärmere Wörter präferiert, wie man das von langen gegenüber kurzen Wörtern kennt (vgl. Abschnitt 5.4.4).

Language	Phoneminventar			Wörterbuch		Korpus	
	$\sum$	C	V	C	V	C	V
Chinesisch	48	48.9	51.1	55.9	44.1	55.0	45.0
Chinesisch	137	16.8	83.2	55.9	44.1	55.0	45.0
Deutsch	43	51.2	48.8	61.0	39.0	60.5	39.5
Japanisch	31	67.7	32.3	48.2	51.8	51.4	48.6
Kroatisch	30	83.3	16.6	54.8	45.2	52.5	47.5
Koreanisch	41	56.1	43.9	54.9	45.1	54.6	45.4
Portugiesisch	46	45.6	54.3	47.7	52.3	50.1	49.9
Russisch	47	78.7	21.3	56.1	43.9	55.9	44.1
Spanisch	40	60.0	40.0	53.9	46.1	54.0	46.0
Türkisch	29	72.4	27.6	53.5	46.5	53.2	46.8

Tabelle 5.12: Prozentualer Anteil von Konsonanten (C) und Vokalen (V)

Phonemerkennungsraten

In der Tabelle im linken Teil der Abbildung 5.6 sind die Phonemfehlerraten für die 10 GlobalPhone-Erkenner zusammengestellt. Der Vergleich von Erkennern, die auf so vielen verschiedenen Sprachen aber dennoch auf einheitlichen Daten basieren, ist bisher einmalig.

Die Ergebnisse basieren auf Dekodierungen mit einem frei laufenden Phonemerkenner, d.h. das Aussprachewörterbuch besteht aus den jeweiligen Phonemeinträgen. Die akustische Modellierung basiert wie oben beschrieben auf voll kontinuierlichen HMMs mit 32 Gaußschen Mischverteilungen pro Subphonem. Zur Dekodierung sind die Wahrscheinlichkeiten für Phonemübergänge gleichwahrscheinlich gesetzt, so daß die unterschiedlich starken phonotaktischen Einschränkungen (siehe nächster Abschnitt) die Resultate nicht überlagern.

Sprache	Rang	Phoneme	PER
Ch-Mandarin	8	137	45.2
Deutsch	7	43	44.5
Englisch	9	43	46.4
Französisch	2	38	36.1
Japanisch	1	31	33.8
Koreanisch	2	41	36.1
Kroatisch	4	30	36.7
Portugiesisch	10	46	46.8
Spanisch	5	40	43.5
Türkisch	6	29	44.1

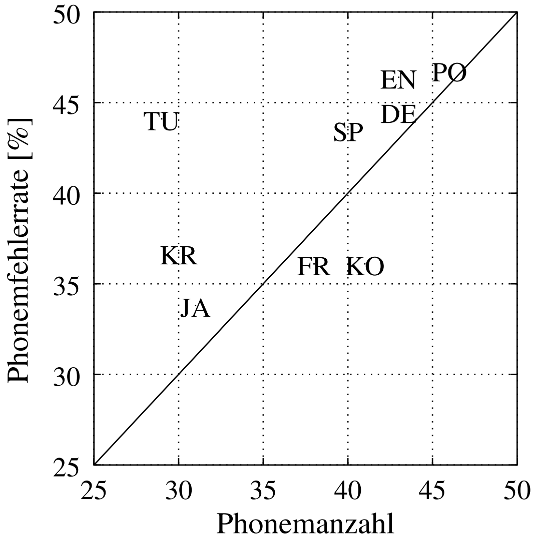


Abbildung 5.6: Phonembasierte Fehlerraten (PER) für 10 GlobalPhone-Sprachen

Aufgrund der Signaleigenschaften würde man erwarten, daß Vokale zuverlässiger zu erkennen sind als Konsonanten. Nach Tabelle 5.12 würde sich also eine Gruppe von einfacheren Sprachen ergeben, in der sich Portugiesisch, Japanisch und Kroatisch befinden. Die Ergebnisse zeigen, daß diese Vermutung auf Japanisch und Kroatisch zutrifft. Umgekehrt müßten konsonantenreiche Sprachen hohe Phonemfehlerraten aufweisen, was für Deutsch zutrifft.

Der dominierendere Faktor scheint aber die Korrelation der Fehlerrate mit der Anzahl modellierter Phoneme zu sein, wie die Abbildung 5.6 verdeutlicht. Das ist dadurch zu erklären, daß mit zunehmender Zahl der Phoneme die Wahrscheinlichkeit für Verwechselungen zunimmt. Für die Sprachen Deutsch, Englisch, Japanisch, Französisch, Spanisch und Portugiesisch steht die Phonemfehlerrate in direktem Verhältnis zur Phonemanzahl. Eine Ausnahme bildet Türkisch, das zwar ein sehr kleines Phonemset hat, aber eine hohe Fehlerrate. Eine Fehleranalyse ergab hier überraschenderweise eine hohe Verwechslung der Vokale [e], [i] und [y], und das obwohl das türkische Vokalsystem sehr einfach ist (vgl. Tabelle 5.3).

Zusammenfassend ergeben sich zwei bzw. drei Leistungsgruppen. Japanisch, Koreanisch, Französisch und Kroatisch zeigen sehr gute phonem-basierte Erkennungsraten. Mit recht großem Abstand folgen Spanisch, Türkisch und Deutsch. Mit etwas geringerem Abstand am schwierigsten zu erkennen sind Chinesisch, Englisch und Portugiesisch. Diese Ergebnisse decken sich mit den Beobachtungen für bekannte Sprachen wie Deutsch, Englisch und Spanisch. Französisch wird als schwieriger, Portugiesisch als einfacher eingestuft [CAGADL97, Köh96]. Für alle anderen Sprachen gibt es derzeit keine Vergleichsmöglichkeiten.

Phonem-Sprachmodelle

Abbildung 5.7 zeigt die N-Gramm-Perplexitäten phonembasierter Sprachmodelle für verschiedene N. Die Perplexitäten wurden jeweils satzweise berechnet, d.h. Wortgrenzen werden ignoriert, wodurch die Einflüsse der Worteigenschaften eliminiert werden. Auffallendes Verhalten zeigt Japanisch, hier fällt die Perplexität von Uni- nach Bigrammen extrem ab. Das ist durch die Mora-Struktur bedingt, die im wesentlichen nur CV-Abfolgen erlaubt. Im Türkischen wird die Vokalharmonie mit wachsender Historie deutlich. Das Harmonieprinzip betrifft die Vokalverteilung innerhalb eines Wortes im Hinblick auf Geschlossenheit und Rundung. Es kommen innerhalb eines türkischen Wortes entweder nur geschlossene oder nur offene Vokale vor. Dieses Prinzip und der geringe Umfang des Phoneminventars führen dazu, daß Türkisch die niedrigste Perplexität hat (für $N \geq 4$). Französisch ist die einzige Sprache, deren Perplexität durch Erhöhung der Historie noch stark verringert werden kann. Diese Resultate werden von [CAGADL97] bestätigt. Brauchbar sind die Erkenntnisse über die Phonemperplexitäten für die Sprachenidentifizierung, in der N-Gramm-Phonemsprachmodelle zum Einsatz kommen (vgl. Abschnitt 6.4.1).

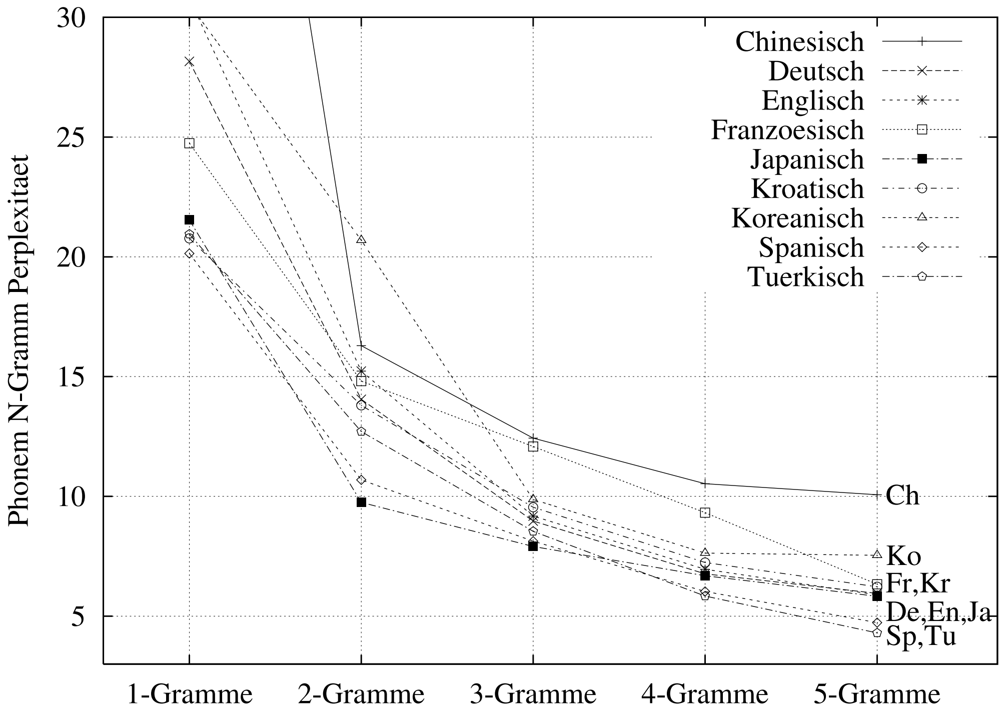


Abbildung 5.7: Phonem N-Gramm-Perplexitäten

5.4.3 Phonologische Unterschiede

Im Dekoder des Erkennersystems kann aufgrund technischer Beschränkungen die Modellierung kontextabhängiger Phonemmodelle derzeit nicht über das erste Phonem des angrenzenden Wortes beziehungsweise über das letzte Phonem des vorhergehenden Wortes hinweg ausgedehnt werden. Enthält ein Korpus eine große Anzahl Wörter, die kürzer sind als die modellierte Kontextbreite, kann daher durch Erweiterung der Kontextbreite kein Gewinn mehr erzielt werden, denn diese Wörter werden bereits als Ganzwortmodelle modelliert. Insgesamt erwartet man, daß der Gewinn, den man durch Verbreitern des Kontextes erzielt, mit der Breite des Kontextes abnimmt.

Ganzwortmodelle

Abbildung 5.8 zeigt für die GlobalPhone-Trainingsdaten den prozentualen Anteil der

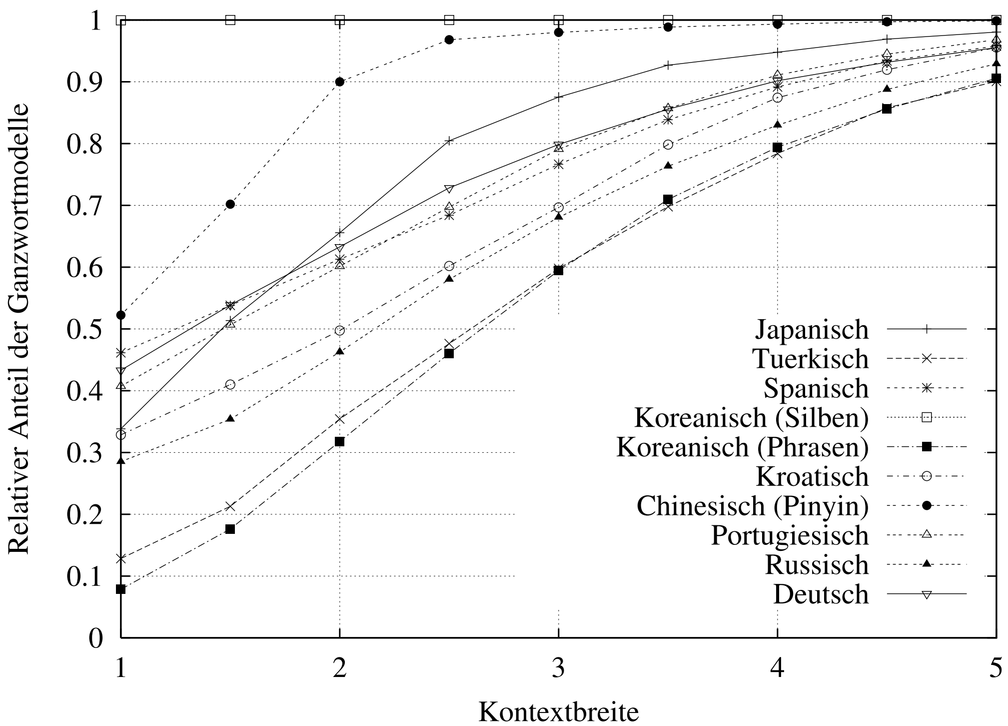


Abbildung 5.8: Relativer Anteil von Ganzwortmodellen für verschiedene Kontextbreiten für die GlobalPhone-Sprachen

Wörter, die bei der Verwendung der Kontextbreiten $1 - 5$ (± 1 =Triphone; ± 2 = Quintphone; ± 3 = Septphone; uws.) als Ganzworte modelliert werden. Wie man sieht, ist die Anzahl der Ganzwortmodelle stark sprachenabhängig. Mit Sicherheit kann man erwarten, daß die Erweiterung der Kontextbreite von 2 auf 3 für die Sprachen Chinesisch und Japanisch weniger Gewinn verspricht, als für Kroatisch oder Russisch. Insbesondere für Sprachen mit sehr langen Einheiten wie Türkisch und koreanischen Phrasen wäre die Modellierung mit Septphonen erwägenswert. Für das Koreanische wurden zwei verschieden segmentierte Korpora zugrundegelegt. Das phrasenbasierte Korpus beruht auf den natürlichen Einheiten, das silbenbasierte Korpus dagegen auf einer Silbenzerlegung. Die Graphik zeigt, welchen Effekt dies auf die Möglichkeit der Modellierung von Polyphonen hat. Mit Triphonen werden bereits 99.8% aller koreanischen Silbeneinheiten als Ganzwortmodelle modelliert, während es bei den natürlichen Einheiten erst 8% sind.

Anzahl der Polyphone

Durch die folgende Analyse der Polyphoneanzahl soll untersucht werden, ob sprachenspezifische Unterschiede in beobachteten Kontexten existieren. Die Abbildung 5.9 zeigt dies für die Kontextbreiten 0 bis ± 6 in 10 GlobalPhone-Sprachen. Da bei dieser Zählung auch nur in das erste Phonem des benachbarten Wortes geblickt wird, ist auch diese Zahl von der Länge der Worteinheiten abhängig.

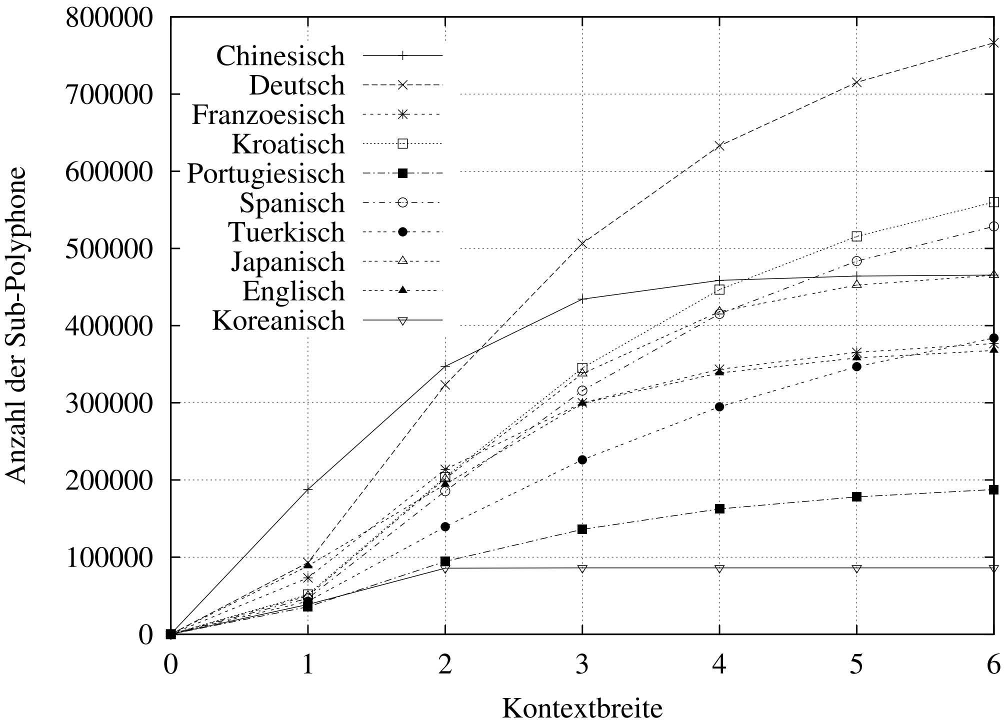


Abbildung 5.9: Anzahl der Polyphone bei verschiedenen Kontextbreiten für 10 Sprachen

Deutsch fällt durch eine außerordentlich große Anzahl Polyphone auf. Diese Beobachtung wird von [DAK95] bestätigt. Mehr Polyphone erfordern eine bessere Generalisierungsfähigkeit der einzelnen Modelle. Wird die Zahl der zu ballenden Subpolyphone bei allen Sprachen auf dieselbe Zahl gesetzt, dann müssen in Sprachen mit vielen Polyphonen mehr Kontexte zusammengeballt werden, als in Sprachen mit wenigen Polyphonen. Auffallend sind die Sprachen Chinesisch und Koreanisch, deren Repertoire an Polyphonen für Kontexte, die breiter sind als ± 3 bzw. ± 2 erschöpft ist. Im Koreanischen ist dies eine unmittelbare Folge der Länge der lexikalischen Einheiten, im Chinesischen spielt dabei die strenge Silbenstruktur CV(C) eine große Rolle.

5.4.4 Unterschiede in der Segmentierung

Bedingt durch das Schriftsystem und den Satzbautyp einer Sprache bestehen gewaltige Unterschiede in der Länge und Anzahl der lexikalischen Einheiten zwischen den einzelnen Sprachen. Schwach flektierende oder isolierende Sprachen haben pro semantische Einheit nur sehr wenige Wortformen, während stark flektierende oder agglutinierende Sprachen zu einem Wortstamm zahlreiche Wortformen bilden können. Dieses Verhalten hat unmittelbaren Einfluß auf das Vokabularwachstum und damit auf die Vokabularabdeckung. In diesem Abschnitt werden diese Kriterien für die GlobalPhone-Sprachen beschrieben und miteinander verglichen.

Wortlänge
Abbildungen 5.10 zeigt die Verteilungen der Wortlängen gemessen in Phonemen für die GlobalPhone-Sprachdaten. Jede Abbildung enthält zwei Kurven, eine zeigt die Verteilung der Wortlängen im Aussprachewörterbuch, die andere die nach den Wortauftrittshäufigkeit gewichteten Wortlängen aller Äußerungen. Insgesamt besteht die Tendenz, im allgemeinen Sprachgebrauch kürzere Wörter häufiger zu verwenden. Lange Wörter werden oftmals abgekürzt oder akronymisiert. Dieses Phänomen ist auf die angeborene Bequemlichkeit des Menschen zurückzuführen und ist als „Zipfs Gesetz“ (oder auch Huffman-Coding) bekannt. Danach gilt $Rang(wort) \times Freq(wort) = const$ bzw. $Länge(wort) \times Freq(wort) = const$. Aus diesem Grund liegen alle Verteilungen der Wortlängen im Korpus links von den Verteilungen im Aussprachewörterbuch. Für spontan gesprochene Äußerungen ist dieses Verhältnis noch extremer.

Grundlage der Berechnungen für die chinesische Sprache sind die romanisierten und segmentierten Pinyineinheiten (siehe Abschnitt 5.5.1.3). Die Transformation führt zu recht kurzen Einheiten mit Schwerpunkten der Verteilung bei 2 und 5 Phonemen. Daraus läßt sich ableiten, daß die zu erwartende OOV-Rate im Chinesischen sehr gering sein wird (siehe Abschnitt 5.4.4).

Die Verteilung über Wortlängen im deutschen Korpus entspricht den Erwartungen: Kurze Wörter sind sehr viel häufiger als lange und die Häufigkeit nimmt mit der Länge ab. Das ausgeprägte Maximum bei 3-phonemischen Wörtern kann durch die bestimmten und unbestimmten Artikel erklärt werden. Die Verteilung im Aussprachewörterbuch hat ein breites Plateau und verläuft sehr flach. Die Länge der Wörter läßt sich durch die Komposita erklären.

Das Auffallendste an der japanischen Verteilung sind die starken Ausprägungen bei gerader Anzahl von Phonemen. Dies gilt sowohl im Aussprachewörterbuch als auch im Korpus. Der Grund liegt in der restriktiven Morastruktur der japanischen Phonologie. Selbst Fremdwörter und Eigennamen werden durch Einfügungen von Vokalen zwischen Konsonantencluster in dieses Schema gepreßt (zum Beispiel wird aus dem deutschen Wort „Post“ im Japanischen „posuto“). Diese restriktive Phonologie erklärt die gute Phonemerkennungsleistung im Japanischen und die geringe Zahl an Polyphonen.

Für die koreanische Sprache wurden die Wortlängenverteilungen für die natürlichen Wortphrasen neben diejenigen auf silbensegmentierten Einheiten aufgetragen. Die enorme Länge der natürlichen Phrasen verdeutlicht, daß mit einer immensen OOV-Rate zu rechnen wäre, wenn man Phrasen als Erkennungseinheiten verwenden würde. Da die koreanische Sprache silbenbasiert aufgebaut und auch geschrieben wird, ist ihre Zerlegung in Silben die naheliegendste Segmentierung. Für die Silben ergeben sich Wortlängen zwischen 1 und 4 Phonemen. Zur Spracherkennung eignen sich solche kurzen Einheiten aufgrund der hohen Verwechslungsraten und der eingeschränkten Reichweite des Sprachmodells nur bedingt, daher werden in Abschnitt 5.5.3.2 Methoden beschrieben, mit sich denen geeignete Einheiten zur Spracherkennung der koreanische Sprache bestimmen lassen.

Die Wortlängenverteilung im kroatischen Lexikon ist wie im portugiesischen sehr symmetrisch mit dem Scheitelpunkt bei 8 bzw. 7 Phonemen. Im Korpus zeigen beide Sprachen eine Dominanz bei 2-Phonem-langen Wörtern. Auch die russischen Lexikoneinträge sind im Schnitt 7 bzw. 8 Phoneme lang und verteilen sich symmetrisch. Damit liegen die Worteinheiten dieser drei Sprachen deutlich über dem Durchschnitt und lassen eine hohe OOV-Rate vermuten.

Im Russischen fällt der große Anteil 1-Phonem-langer Wörtern auf, die 10% des gesamten Korpus ausmachen. Das rührt von der großen Zahl 1-phonemischer Präpositionen wie u, o, k, s, w und Konjunktionen wie i und a her. Geschriebene Texte enthalten viele Nebensatzkonstruktionen, die im Russischen mit „und“ (= i) verknüpft werden.

Die spanische Verteilung wird durch 2-Phonem-lange Wörter dominiert. Fast ein Drittel des gesamten spanischen Textkorpus wird durch die Wörter „de“, „la“, „el“, „que“ und „en“ erzeugt (hier in der Reihenfolge ihrer Häufigkeit). Diese Eigenschaft erschwert die Erkennung der spanischen Sprache, da kurze Wörter leicht zu verwechseln sind.

Die Verteilung für die türkische Sprache verläuft insgesamt etwas flacher aber relativ symmetrisch um den Scheitelpunkt von 5 Phonemen. An der Verteilung über dem Korpus wird sofort erkennbar, daß die türkische Sprache sehr lange Worteinheiten hat. Dies ist ein Resultat ihres agglutinierenden Sprachbaus, in dem die Wortstämme durch fortgesetztes Anleimen von Partikeln verlängert werden. Das läßt eine hohe OOV-Rate erwarten.

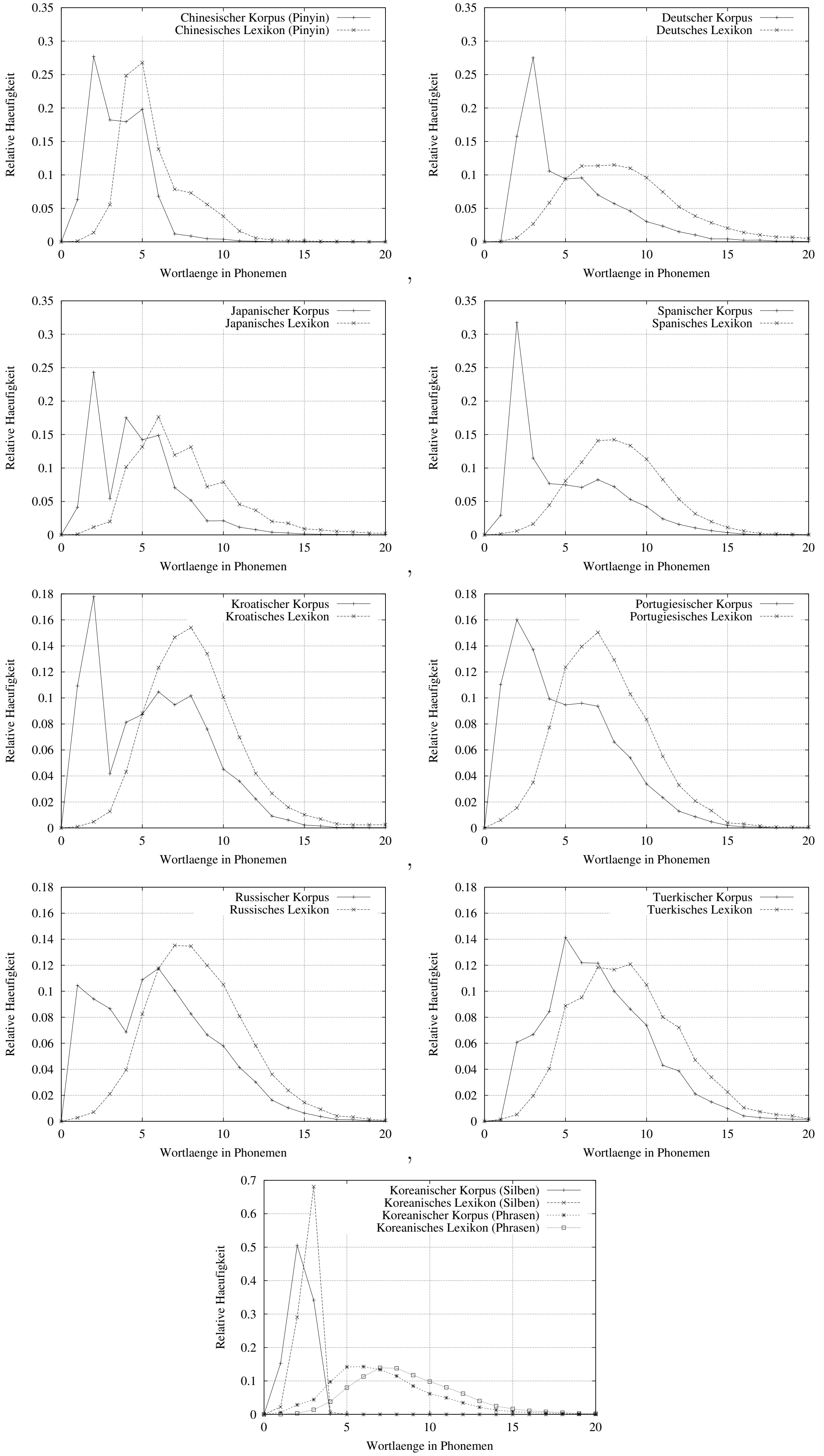


Abbildung 5.10: Relativer Anteil verschiedener Wortlängen

Satzlänge
In Abbildung 5.11 ist die Anzahl gesprochener Wörter zur Dauer der Äußerung für alle gesammelten Sprachen in Beziehung gesetzt. Demnach werden im Mittel etwas mehr als zwei Wörter pro Sekunde geäußert. Türkisch und vor allem Koreanisch weichen von diesem Mittel nach unten ab, Chinesisch nach oben. Das hängt mit der Länge der gesprochenen Einheiten zusammen, die, wie in Abschnitt 5.4.4 deutlich wurde, für Chinesisch sehr kurz, für Türkisch und Koreanisch dagegen sehr lang sind. Chinesische, spanische und kroatische Zeitungen enthalten wesentlich

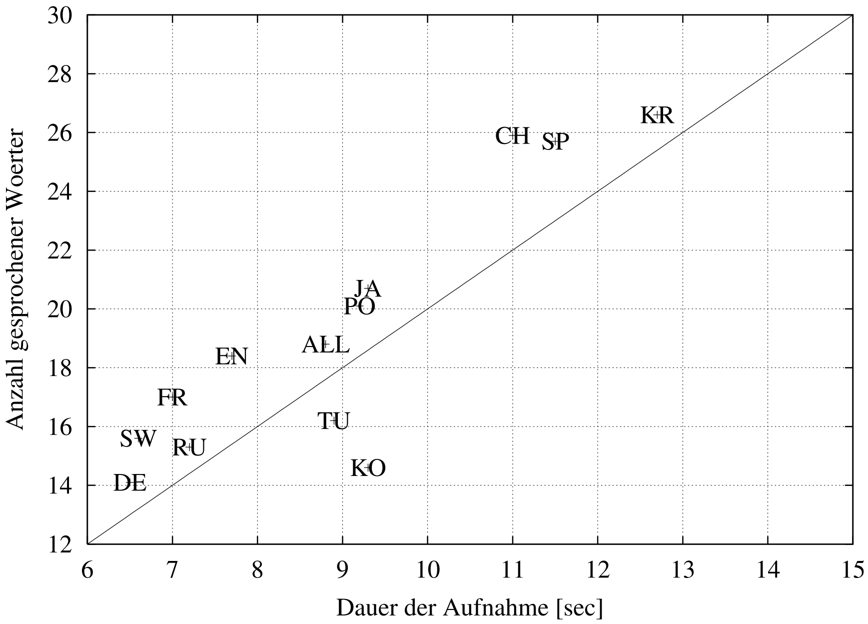


Abbildung 5.11: Äußerungsdauer gegenüber Anzahl gesprochener Wörter

längere geschriebene Sätze. Im Vergleich zu deutschen, russischen, schwedischen oder französischen Sätzen sind sie etwa doppelt so lang. Es ist allerdings offen, ob diese Beobachtung auf sprachenspezifische oder zeitungsspezifische Unterschiede zurückzuführen ist. Im Zusammenhang mit den Untersuchungen zur Kompaktheit im nächsten Abschnitt sieht es aber so aus, als ob es im Deutschen möglich ist, einen Sachverhalt mit weniger Wörtern darzustellen als beispielsweise im Französischen, Portugiesischen oder Spanischen.

Vokabularwachstum und -abdeckung
Aufgrund von Geschwindigkeits- und Speicherlimitierungen muß das zu erkennende Vokabular eines Spracherkennungssystems in der Suche beschränkt werden. Zu erkennende gesprochene Wörter, die nicht im Vokabular vorkommen, sind unbekannte Wörter und verursachen in jedem Fall einen Erkennungsfehler. Tatsächlich entstehen pro unbekanntem Wort im Mittel 1.5 bis 2 zusätzliche Folgefehler. Das Maximum des zu erkennenden Vokabulars in JRTk, wie auch in vielen anderen gängigen Spracher-

kennern, liegt bei 65.536 (65K) Wörtern ($= 2^{16}$; das entspricht der Zahl möglicher Indizes, sofern für die Speicherung 2 Byte verfügbar sind). Daher ist offensichtlich, daß die OOV-Raten einen sehr großen Einfluß auf die Fehlerraten eines Erkenners haben.

Abbildung 5.12 zeigt für alle GlobalPhone-Sprachen das Vokabularwachstum auf den transkribierten Trainingsdaten des GlobalPhone-Korpus. Für das Koreanische sind in der Graphik sowohl silbenbasierte als auch phrasenbasierte Einheiten aufgetragen. Im Chinesischen und Japanischen sind die romanisierten und segmentierten Einheiten zugrunde gelegt. Die Vokabularwachstumsraten auf den Transkriptionen sind teilweise durch sammelbedingte Artefakte überlagert, wie beispielsweise im Portugiesischen. Hier wurden zu Beginn der Datensammlung einige Texte mehrfach gelesen, wodurch das tatsächliche Wachstum des Vokabulars unterschätzt wird. Dasselbe gilt auch für die türkische Datensammlung.

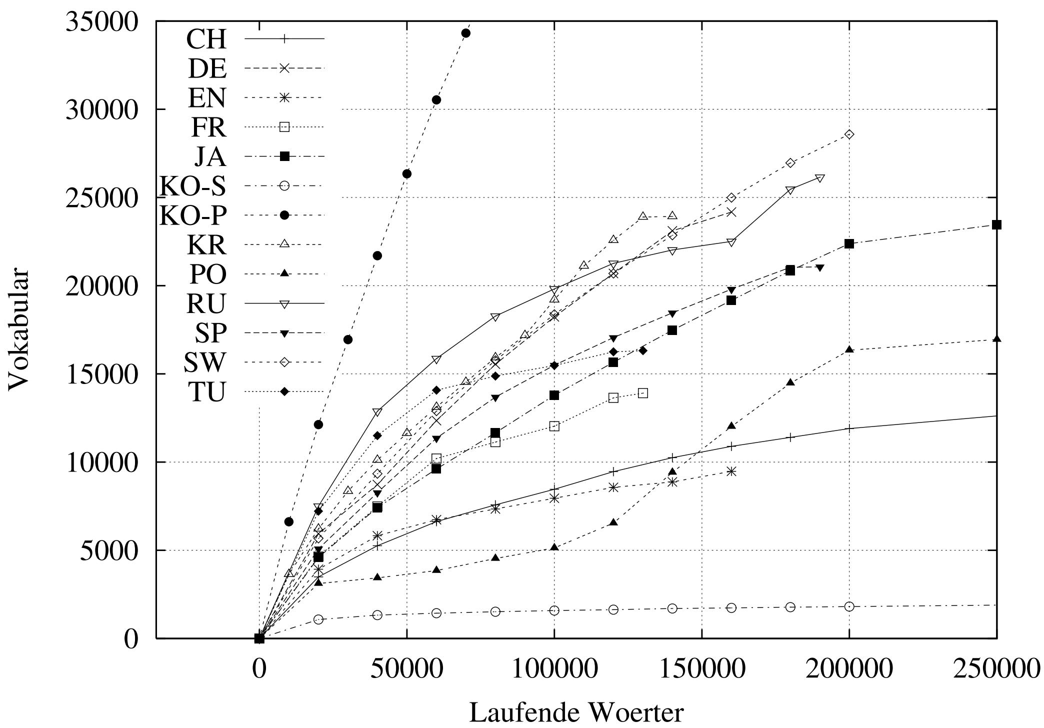


Abbildung 5.12: Vokabularwachstum für 12 GlobalPhone-Sprachen auf dem Transkriptionsmaterial

Die gezeigten Vokabularwachstumsraten resultieren in sehr unterschiedlichen Vokabularabdeckungsraten. Abbildung 5.13 zeigt für die GlobalPhone-Sprachen die Selbstabdeckungsrate (*engl. Selfcoverage*), d.h. die prozentuale Abdeckung des Korpus in Abhängigkeit des Vokabulars. Als Grundlage sind wiederum die transkribierten Textdaten der Audiodaten herangezogen. Aufgrund der geringen Größe dieser

Textdaten sind die Aussagen über das allgemeine Verhalten in den einzelnen Sprachen nur eingeschränkt möglich. Wie bereits in Abschnitt 5.3.7 beschrieben, wurden allerdings in vielen Sprachen zusätzliche Textkorpora großen Umfangs zur zuverlässigeren Schätzung der N-Gramm Wahrscheinlichkeiten gesammelt.

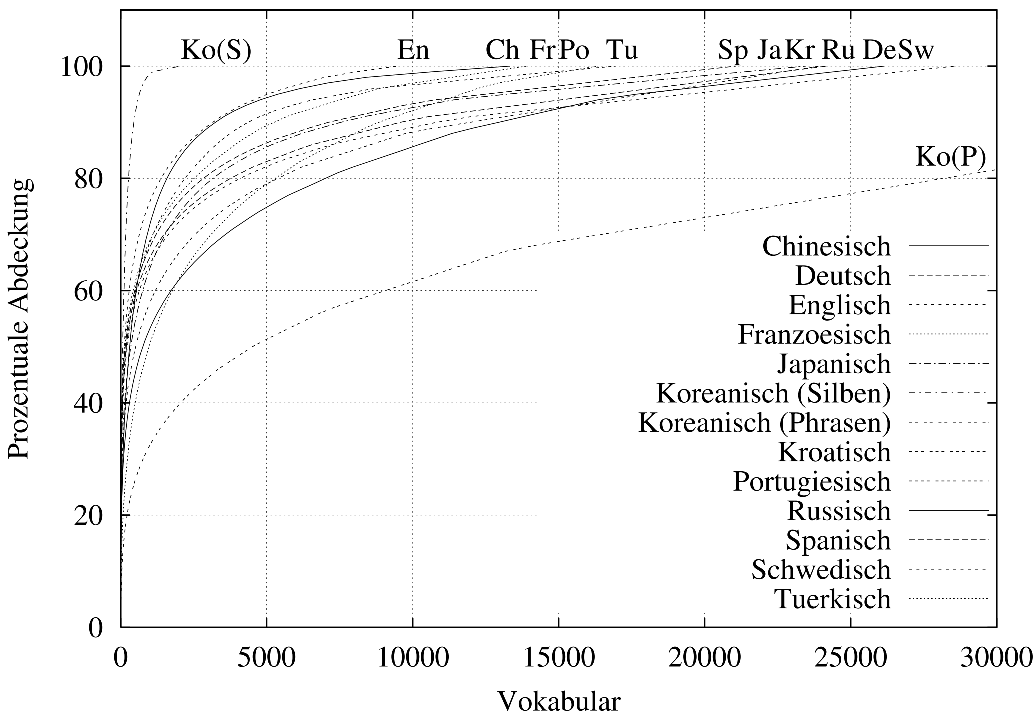


Abbildung 5.13: Vokabularabdeckung für 12 GlobalPhone-Sprachen auf dem Transkriptionsmaterial

Abbildung 5.14 zeigt für die drei Sprachen Chinesisch, Portugiesisch und Türkisch die Vokabularabdeckung gemessen auf großen Korpora, die aussagekräftigere Schlüsse und Vergleiche zulassen. Für die Berechnung der Abdeckungsraten sind ein chinesischer Textkorpus bestehend aus 82 Millionen Pinyineinheiten zugrunde gelegt, ein portugiesischer Korpus von 11 Millionen Wörtern und ein türkischer Korpus mit 15.7 Millionen Wörtern. Abgebildet sind jeweils die Selbstabdeckungsrate des Textkorpus als auch die Abdeckungsrate auf einer Testmenge (*engl. Crosscoverage*). Die Crosscoverage zeigt die OOV-Rate auf dem Testset in Abhängigkeit der Größe des Erkennervokabulars. Die Kurvenverläufe der drei Sprachen sind typische Verläufe für die Segmentierung und den Sprachbautypen, den diese drei Sprachen jeweils verkörpern.

Im Fall der chinesischen Sprache wurde eine Segmentierung in Pinyinsilben vorgenommen (siehe Abschnitt 5.5.1.3), deren Zahl beschränkt ist. Bei einer Vokabulargröße von 10000 Einheiten sind bereits 96.5% aller Einheiten des Testtextes abgedeckt. Bei 20000 Wörtern sind es 99% und bei 58800 sind es 100%. Bei einer

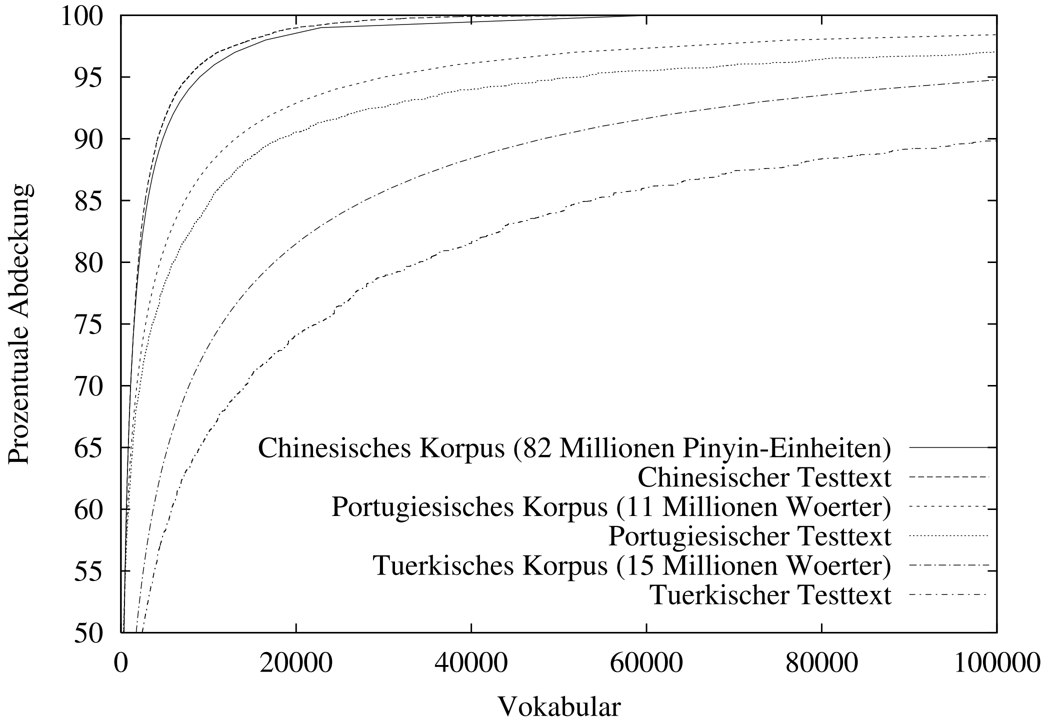


Abbildung 5.14: Vokabularabdeckung am Beispiel Chinesisch, Portugiesisch und Türkisch

Vokabulargröße von 60K gibt es im Chinesischen somit keine unbekannten Einheiten. Ganz anders dagegen die Verhältnisse in der türkischen Sprache. Diese Sprache hat bedingt durch das Prinzip der Agglutination ein riesiges Vokabularwachstum. Daraus resultiert, wie Abbildung eine sehr hohe OOV-Rate. Selbst mit einem Vokabular von 500K beträgt die OOV-Rate im Türkischen noch etwa 5%. Bei einer Größe von 60K Erkennervokabular liegt die OOV-Rate bei fast 15%. Bei 1-2 Folgefehlern pro unbekanntem Wort läge man mit einer solchen OOV-Rate bereits bei 15-30%. Die portugiesische Sprache liegt zwischen diesen beiden Extremen. Als Vertreter der flektierenden Sprachen, hat sie zwar weit mehr Wörter als Chinesisch, aber nicht eine so hohe Wachstumsrate und OOV-Rate wie die agglutinierende Sprache Türkisch. Auf der koreanischen Sprache, deren Wachstumskurven in Abschnitt 5.5.3.2 dargestellt werden, wurde auf den natürlichen Wortphrasen, die ebenfalls durch Agglutination entstehen weit über 30% OOV-Rate gemessen.

In Tabelle 5.13 sind für die GlobalPhone-Sprachen jeweils die OOV-Raten nebst der Größe des Vokabulars auf dem sie gemessen wurden für die GlobalPhone-Sprachen aufgeführt. Für die späteren Erkennungsläufe wurden die OOV-Raten in den Sprachen Deutsch, Französisch, Japanisch, Kroatisch, Portugiesisch, Spanisch und Türkisch kontrolliert, in dem die im Test nicht abgedeckten Wörter in das Erkennervokabular eingetragen und als Unigramme in das Sprachmodell aufgenommen wurden.

Sprache	Vokabular	OOV-Rate
Ch-Mandarin	60K	0%
Deutsch	61K	4.4%
Englisch	64K	0.3%
Französisch	30K	4.7%
Japanisch	22K	3.0%
Koreanisch	64K	0.2%
Kroatisch	31K	13.6%
Portugiesisch	60K	4.3%
Spanisch	30K	5.2%
Türkisch	64K	13.5%

Tabelle 5.13: OOV-Raten für 10 GlobalPhone-Sprachen

5.4.5 Semantische Unterschiede

In diesem Abschnitt wird untersucht, ob sich Sprachen hinsichtlich ihrer Kompaktheit unterscheiden. Mit Kompaktheit wird hier ein Maß dafür bezeichnet, wieviele Wörter zur schriftlichen Darstellung eines Sinngehaltes benötigt werden. Zur Berechnung dieses Maßes braucht man Texte gleichen Informationsgehaltes in mehreren Sprachen. Bei den GlobalPhone-Daten handelt es sich zwar um dieselbe Domäne, aber es kann für keinen Text garantiert werden, daß er eine originalgetreue Übersetzung eines anderen wäre. Texte solcher Art entstehen aber beispielsweise in Organisationen und Behörden, die in mehreren Amtssprachen operieren. Die nachfolgende Untersuchung basiert auf auf Texten, die im Rahmen einer EU-Ankündigung zum europaweiten Spracherkennungsprojekt ESPRIT in 9 EU-Sprachen entstanden und von Elsnet auf CD verfügbar gemacht werden (ECI-CD). Es standen Texte von 20-25K Umfang in den neun Sprachen Dänisch, Deutsch, Englisch, Französisch, Griechisch, Italienisch, Niederländisch, Portugiesisch und Spanisch zur Verfügung. Tabelle 5.14 zeigt für jede Sprache die Summe der Wörter und Zeichen sowie die Größe des verwendeten Vokabulars, die zur Darstellung der Information in den untersuchten Texten benötigt wurden.

Die kompaktesten Sprachen gemessen an der Summe der verwendeten Wörter sind die Sprachen Dänisch und Deutsch. In diesen Sprachen werden die wenigstens Worteinheiten zur Informationsdarstellung benötigt. Das Mittelfeld bilden Italienisch, Englisch und Niederländisch. Dann folgen mit wachsendem Abstand Griechisch, Portugiesisch, Französich und zuletzt Spanisch, das etwa 20% mehr Wörter benötigt, um den gleichen Sachverhalt zu beschreiben wie die Sprache Dänisch. Die Worteinheiten zu zählen wird dem Sachverhalt aber nicht gerecht, denn wie bereits in Abschnitt 5.4.4 festgestellt, variieren die Wortlängen stark und bedingt durch den unterschiedlichen Sprachbau können mehr oder weniger Informationen in ein Wort verpackt werden. Vergleicht man dagegen die Anzahl der Zeichen, ergibt sich ein

Sprache	Rang	#Wörter	Rang	Zeichen	Rang	Vokabular
Dänisch	1	20516	1	144128	6	4428
Deutsch	2	21917	8	173382	7	4656
Englisch	4	23301	2	148502	1	3448
Französisch	7	25047	6	165061	3	4064
Griechisch	6	24692	9	196437	9	4767
Italienisch	3	22723	3	156677	8	4752
Niederländisch	5	24399	7	170901	3	4064
Portugiesisch	8	25048	5	161726	5	4333
Spanisch	9	25382	4	160471	2	3974

Tabelle 5.14: Kompaktheit für neun EU-Sprachen (inklusive Interpunktion)

anderes Bild. Deutsch rutscht als Folge der zahlreichen Komposita von Rang 2 auf 8 ab. Spanisch kommt auf Rang 4 unmittelbar nach Italienisch. Beide Sprachen verwenden viele aber sehr kurze Funktionswörter. Griechisch liegt auf dem letzten Rang, es verwendet nicht nur viele sondern auch lange Wörter.

Der zweite für die Spracherkennung wesentliche Punkt ist die Frage nach der Größe des verwendeten Wortschatzes, der zur Übermittlung der Information verwendet wird. Englisch kommt mit großem Abstand mit den wenigsten Vokabeln aus. Das liegt daran, daß Englisch eine schwach flektierende Sprache ist. Viele Konstruktionen werden daher mit Hilfsverben hergestellt, was wiederum den vierten Rang bei der Kompaktheit erklärt. Niederländisch, Französich und Spanisch bilden die Mittelgruppe. Für Deutsch, Dänisch, Portugiesisch, Italienisch und Griechisch werden sehr viele Vokabulareinträge benötigt, daher sind hier wesentlich höhere OOV-Raten zu erwarten als etwa für Englisch.

5.5 Behandlung sprachenspezifischer Besonderheiten

Die Vergleiche zwischen den GlobalPhone-Sprachen haben gezeigt, daß es Spracheigenschaften gibt, die sich der Standardbehandlung der Basiserkenner entziehen. Die zur Behandlung dieser Merkmale notwendigen Maßnahmen lassen sich untergliedern in:

1. Romanisierung von Sprachen, die ideographische Schriftsysteme haben
2. Segmentierung von Sprachen, die entweder gar keine natürlichen Worteinheiten oder sehr lange Worteinheiten haben
3. Modellierung von Tonsprachen

4. Bestimmung geeigneter lexikalischer Einheiten für agglutinierende Sprachen

Im folgenden werden diese Maßnahmen und die dabei angewendeten Techniken exemplarisch an einigen Sprachen vorgestellt.

5.5.1 Romanisierung und Segmentierung

Wie bereits in Abschnitt 5.3.3 beschrieben, wurden aus Gründen der Transparenz alle Schriften romanisiert. Für die phonologischen Schriften handelt es sich dabei meist um wenige, einfache Abbildungsregeln, für die ideographischen Schriften ist der Vorgang allerdings nicht trivial. Die Probleme der Romanisierung werden hier zusammen mit der Segmentierung betrachtet, weil beide Probleme auf ideographischen Schriften in der Regel gemeinsam gelöst werden.

5.5.1.1 Koreanische Gulja

Die Romanisierung der koreanischen Hangul-Schrift ist eindeutig aber recht aufwendig, da aus dem Hangulalphabet, das 40 Zeichen umfaßt insgesamt 5601 sogenannte *Gulja*-Silbenzeichen komponiert werden können. Die Romanisierung besteht somit aus 5601 Regeln, die in dem online verfügbaren Tool *hcode* [hco98] realisiert sind und sich relativ eng an dem Vorschlag der südkoreanischen Regierung orientieren. Neben der Romanisierung liefert *hcode* eine Zerlegung in Silben, die eindeutig ist, da jedes Gulja-Zeichen genau eine Silbe repräsentiert. Diese Silbenzerlegung ist die Basis für den in Abschnitt 5.5.3.2 beschriebenen Algorithmus zur Bestimmung geeigneter sprachlicher Einheiten.

5.5.1.2 Japanische Kanji

Für die Romanisierung der japanischen Kanji- und Kana-Schriftzeichen existiert das anerkannte Standardtool CHASEN, das zu Forschungszwecken frei verfügbar ist. CHASEN ist ein morphologisches Analysetool für die japanische Sprache [Mat97], das Wörter in ihre morphologischen Bestandteile zerlegt. Als Referenzwerk zur Definition der Wort- und Morphemeinheiten wird dazu der japanische Duden „Daijirin“ verwendet. Neben der Romanisierung wird daher auch die Segmentierung der japanischen Schrift durchgeführt. Im Kontext der Spracherkennung ist diese in der japanischen Sprache sehr wichtig, da im Japanischen Schriftzeichen ohne Zeichenbegrenzer aneinandergefügt werden und daher die Grenzen natürlichsprachlicher Einheiten aus der schriftlichen Darstellung nicht entnommen werden können. Da CHASEN bei Zahlenausdrücken fehlerhafte Ausgaben liefert, wurden diese in einem Vorverarbeitungsschritt gesondert umgeformt.

5.5.1.3 Chinesische Hanzi

Wie die japanische Kanji-Schrift können auch in deren Urvater, der chinesischen Hanzi-Schrift aus der Darstellung keine Rückschlüsse auf Wortgrenzen gezogen werden, da alle Zeichen ohne Begrenzung aneinandergereiht werden. Jedes chinesische Zeichen repräsentiert eine Sprechsilbe. Einheiten, die in der Spracherkennung einsetzbar wären, können entweder durch die Segmentierung in einzelne Schriftzeichen, durch prosodische Informationen [Lyu95] oder durch die Definition semantischer Einheiten gefunden werden.

In der vorliegenden Arbeit wurde zur Romanisierung und Segmentierung der Hanzi-Schrift das *Pinyin*-System zugrundegelegt. Das Pinyin-System ist eine romanisierte Lautschrift für chinesische Schriftzeichen, das von der Regierung der Volkrepublik China eingeführt wurde und in den Schulen Chinas parallel zu den chinesischen Schriftzeichen gelehrt wird. Ein Pinyin entspricht genau einem chinesischen Zeichen, d.h. einer Silbe, und besteht aus lateinischen Buchstaben zur Darstellung der Grundlaute gefolgt von einer Ziffer zur Repräsentation des Tones. Alle gebräuchlichen chinesischen Zeichen können durch 1344 Pinyin ausgedrückt werden.

Im Kontext der Spracherkennung hat die Verwendung der Pinyin mehrere Vorteile [RSW99]:

- Graphem-Phonem-Relation: Das Pinyin-System ist eine Lautschrift und liefert daher die Aussprache eines chinesischen Schriftzeichens.
- Keine OOV-Rate: Verwendet man die 1344 Pinyinsilben als sprachliche Einheiten, dann gibt es keine unbekannten Wörter.
- Komplexität: Pinyineinheiten erlauben daher einen kleinen Suchraum.
- Kompatibilität: Pinyin lassen sich ohne Anpassungen mit den für andere Sprachen entwickten Tools behandeln.
- Modularität: Die getrennte Ausführung der Pinyinkonvertierung ist auch für Text-to-Speech-Anwendungen nützlich.
- Transparenz: Da sich Pinyin eng an der Aussprache orientieren, sind sie für Nichtmuttersprachler einfach zu interpretieren.

Zur Konvertierung der chinesischen Hanzi-Schrift in Pinyin gibt es wie im Koreanischen online verfügbare Tools. Allerdings ist die Komplexität der Aufgabe ungleich höher, da die Konvertierung eines Textes dessen semantische und pragmatische Interpretation erfordert. Ein chinesisches Zeichen kann 20 und mehr verschiedene Bedeutungen haben, die sich in unterschiedlichen Aussprachen und damit unterschiedlichen Pinyin ausdrücken. Insgesamt haben 13% aller chinesischer Zeichen mehr als eine Bedeutung. Eine Analyse verfügbarer Tools ergab für keines eine zufriedenstellende Leistung. Daher wurde im Rahmen dieser Arbeit ein Pinyinkonverter entwickelt [Rei97]. Dieser Konverter zerlegt einen laufenden chinesischen Text in semantische Worteinheiten und konvertiert diese Worteinheiten in Pinyinschreibweise.

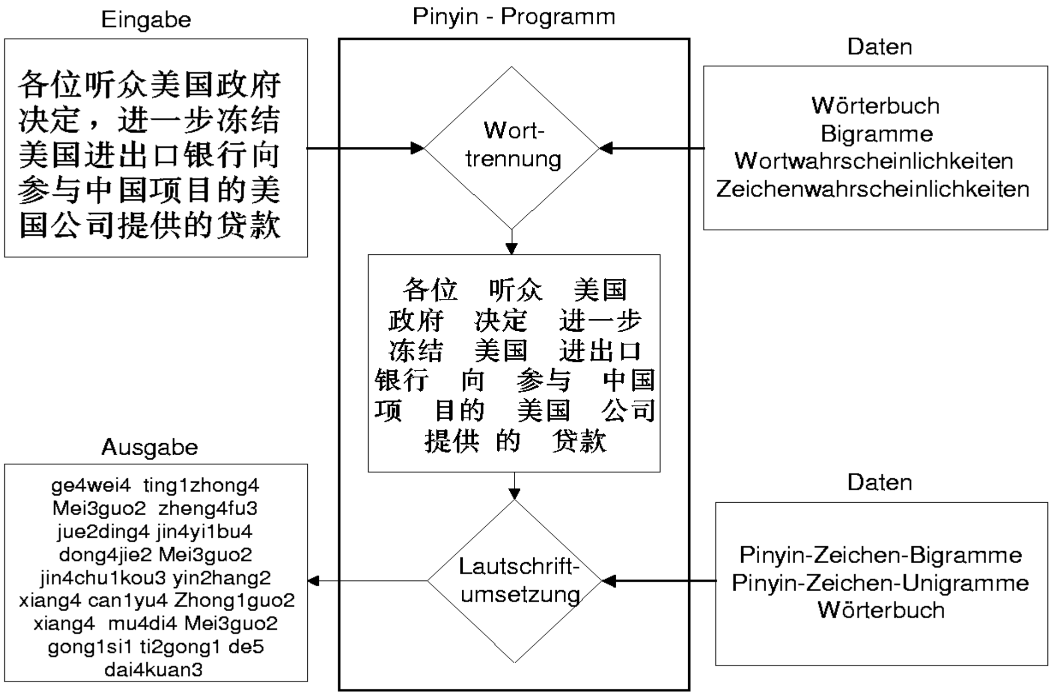


Abbildung 5.15: Der Pinyinkonverter zur Romanisierung und Segmentierung

Eine schematische Darstellung der Transformation eines chinesischen Textes in Pinyineinheiten ist in Abbildung 5.15 zu sehen. Im ersten Schritt wird der laufende Text anhand von Wörterbüchern und Wort- sowie Zeichenbigrammen in semantische Worteinheiten segmentiert, im zweiten Schritt werden diese Worteinheiten anhand von rechtem und linkem Kontext in Pinyin übertragen. Die Länge der resultierenden Worteinheiten variiert zwischen 1 und 10 Silben und beträgt im Mittel etwa 2 Silben, was in etwa einer Worteinheit indoeuropäischer Sprachen entspricht. Die Anwendung von 70 einfachen Graphem-zu-Phonem-Regeln produziert für jede dieser Worteinheiten eine Aussprache.

Mittels einer im Pinyinkonverter eingebauten Validierungsfunktion wurde die Akkuratheit des Pinyinkonverters mit handsegmentierten und handpinyinsierten Texten verglichen. Die Fehleranalyse zeigte, daß sehr viele Fehler durch die verwendeten fehlerbehafteten Wissensquellen bedingt waren. Eine Korrektur dieser Fehlerquellen führte zur finalen Version des Pinyinkonverters. Tabelle 5.15 zeigt die Resultate der Evaluierung in verschiedenen Stadien der Entwicklung.

Das Pinyinumsetzungsprogramm C2T von Tommi Kaikkonen (1992) ist online verfügbar und liefert eine Romanisierung, die in 20.1% fehlerhafte Resultate liefert. Dieses Tool wurde durch eine einfache Worttrennung anhand eines Wörterbuches

System	Segmentierung	Romanisierung
C2T	keine	20.1%
+Wörterbuch	$\approx$ 40%	$\approx$ 20%
+Trennheuristik	$\approx$ 8%	$\approx$ 20%
+Wahrscheinlichkeiten	8%	6%
Pinyinkonverter V1	4.9%	2.1%
Pinyinkonverter Final	3.8%	1.5%

Tabelle 5.15: Segmentierungs- und Romanisierungsfehler des Pinyinkonverters

erweitert. Das Resultat von 40% Worttrennungsfehler zeigt, daß die einfache Quelle eines Wörterbuches zu keiner brauchbaren Segmentierung führt. Wird als einfache Trennheuristik die Regel eingeführt, immer nach dem längsten Wort zu trennen, reduziert sich der Fehler bereits auf etwa 8%. Die Einbeziehung von Wahrscheinlichkeiten bei der Worttrennung und Romanisierung reduziert den Romanisierungsfehler drastisch auf etwa 6%. Der Pinyinkonverter in seiner ersten Version kann diese Resultate deutlich übertreffen und nach Korrektur der Wissensquellen ergibt sich nur noch ein Fehler von 3.8% in der Worttrennung und 1.5% in der Romanisierung.

Rückkonvertierung in chinesische Hanzi
Anschließend wurde der Pinyinkonverter dazu verwendet, anhand großer Textkorpora (siehe Tabelle 5.4) automatisch Regeln zur Rückkonvertierung der Pinyin in chinesische Zeichen zu lernen. Mit über einer halben Millionen Regeln wird eine Rückkonvertierungsfehlerrate von 3.2% erreicht. Zur Angabe der Erkennungsfehler in chinesischen Zeichen müßten diese 3.2% Fehler auf die in Pinyin berechneten Erkennungsfehler addiert werden. Im schlimmsten Fall kann ein Pinyinerkennungsfehler den zur Rückkonvertierung benötigten Kontext verfälschen, so daß noch mehr Fehler entstehen. Die Experimente haben allerdings gezeigt, daß Erkennung und Konvertierung meist dieselben Fehler machen. Die beobachtete Differenz zwischen Pinyinhypothese und Ausgabe in chinesischen Zeichen für den besten Erkenner beträgt 2.6%.

5.5.2 Modellierung von Tonsprachen

Unter allen GlobalPhone-Sprachen sind das Mandarin- und das Schanghai-Chinesisch die einzigen Tonsprachen. Im Mandarin-Chinesisch unterscheidet man 5 unterschiedliche Grundfrequenzverläufe, die bedeutungsunterscheidend sind. Beispielsweise kann das Wort „ma“ durch die entsprechende Modulation des Grundfrequenzverlaufes entweder „Pferd“, „Mutter“ oder „schimpfen“ bedeuten. Für diese Sprache ist es daher wesentlich, daß die Toninformation auf irgendeine Art erfaßt wird. Die chinesische Sprache ist silbenorientiert, wobei eine Silbe in der Regel mit einem Konsonanten beginnt und von einem Vokal gefolgt wird. Vokale werden als Monophthonge, Diphthonge, seltener auch als Triphthonge realisiert. Die Toninformation ist naturgegeben in den Vokalen repräsentiert. Im konventionellen Ansatz werden Silben- und Toninformationen getrennt analysiert und anschließend zusammengeführt, wie beispielsweise in [WSY+95, Lyu95, ACC97]. In neueren Arbeiten ist man dazu übergegangen, die Tonformation in den Merkmalsvektor zu integrieren [CGM+97, ZWL98].

In dieser Arbeit werden zwei Methoden zur Modellierung der Toninformation vorgestellt, die sich beide auf integrierte Verfahren konzentrieren, da man sich damit in der Erkennungsphase die aufwendige Synchronisation der Silben- und Toninformation erspart. Es werden die folgenden Methoden unterschieden:

- Implizite Modellierung durch separate akustische Modelle
- Explizite Modellierung des Grundfrequenzverlaufes

Bei der impliziten Modellierung durch separate akustische Modelle soll der Merkmalsvektor des GlobalPhone-Standarderkenners nicht verändert werden. Die Modellierung der Toneme geschieht auf den in Abschnitt 5.3.9.1 beschriebenen 43 Merkmalskoeffizienten. In diesen Koeffizienten sind bereits Hinweise auf die Grundfrequenz enthalten, da sich die Frequenz der Stimmbandanregung auf die Nulldurchgangsrate und in geringerem Umfang auf die Cepstren auswirkt. Der Vorteil dieses Verfahrens liegt in der Integrierbarkeit und Vergleichbarkeit mit den übrigen GlobalPhone-Erkennern. Bei der expliziten Modellierung des Grundfrequenzverlaufes werden hingegen explizit Merkmale über den Grundfrequenzverlauf in den Merkmalsvektor integriert, um der Toninformation in der chinesischen Sprache entgegenzukommen.

5.5.2.1 Implizite Modellierung durch separate Modelle

Als Basiserkenner wird ein System verwendet, das nach dem in Abschnitt 5.3.9 beschriebenen Verfahren entstand. Der Merkmalsvektor dieses Systems enthält die beschriebenen 43 Koeffizienten, die mit einer LDA auf 24 Koeffizienten reduziert werden. An dieser Stelle sind also keinerlei zusätzliche dedizierte Merkmale zur Modellierung des Grundfrequenzverlaufes enthalten. Die Toninformation wird implizit dadurch modelliert, daß die Vokale in 5 Tonemvarianten unterschieden werden. Für jede dieser Varianten wird ein eigenes Phonem modelliert. Insgesamt werden 137 Phoneme modelliert, von denen 22 Konsonanten sind, 7 Monophthonge und 19 Di- bzw. Triphthonge mit zusammen 113 Tonemvarianten.

Auf der Basis dieser Phoneme wurden zwei Erkennungssysteme entwickelt: Im System CH-SEP wird jede Tonemvariante mit eigenem Codebook und eigenen Mixturgewichten modelliert, im System CH-TAG teilen sich die 5 tonalen Varianten eines

System	Pinyinfehlerrate
CH-TAG	23.3
CH-SEP	24.2

Tabelle 5.16: Implizite Modellierung der Tonsprache Mandarin Chinesisch

Grundvokales jeweils ein Codebook, nur die Mixturgewichte werden getrennt trainiert. In System CH-TAG erhofft man sich durch die gemeinsame Datennutzung ein robusteres Schätzen der Gaußschen Mittelwerte und Kovarianzen. In beiden Systemen werden Quintphone modelliert, die mit dem in Abschnitt 3.2.2.3 beschriebenen Ballungsverfahren auf 1500 Modelle zusammengeballt werden. Zur Ballung des Systems CH-TAG werden in den phonetischen Katalog Fragen nach der Tonalität eines Phonems aufgenommen. Beim Ballungsvorgang entscheiden daher die Daten, ob zwei Tonemvarianten eines Grundvokals gemeinsam oder separat modelliert werden (vgl. dazu Abschnitt 6.3.2.3). Dieser Ansatz realisiert somit eine datengetriebene implizite Modellierung der Toninformation ohne den Vorteil der Parametereinsparung aufgeben zu müssen.

Die Ergebnisse in Tabelle 5.16 zeigen, daß das System CH-TAG gegenüber dem System CH-SEP eine relative Fehlerreduktion von 3% aufweist. Dies wird auf die bessere Ausnutzung der Daten in CH-TAG zurückgeführt. Die Häufigkeiten der unterschiedlichen Töne sind nicht gleichverteilt, so daß manche Toneme nur sehr selten im Sprachmaterial repräsentiert sind. Daher ist die gemeinsame Nutzung von Daten zum Training der Codebooks besonders für gering repräsentierte Modelle gewinnbringend. Die Analyse des Kontextentscheidungsbaumes zeigt, daß die Fragen nach der Tonalität bei der Aufspaltung der Vokalmodelle einen hohen Stellenwert haben. Bei /e/-Vokalen und Diphthongen, die auf /e/ enden, ist die Tonalitätsfrage die wichtigste aller Fragen, bei /a/- und /o/-Vokalen kommen sie an zweiter und dritter Stelle. Bei /i/- und /u/-Vokalen sind die Fragen ab der dritten und vierten Aufspaltungsebene zu finden.

5.5.2.2 Explizite Modellierung der Grundfrequenz

Bei der expliziten Modellierung der Grundfrequenz wird der Merkmalsvektor gegenüber der impliziten Modellierung um Merkmale des Grundfrequenzverlaufes erweitert. Dazu werden in den Merkmalsvektor eines Frames Grundfrequenzinformationen benachbarter Frames integriert. Die absoluten Grundfrequenzwerte sind nicht als Koeffizienten für den Merkmalsvektor geeignet, weil sie stark geschlechts- und sprecherabhängig sind. Dies gilt auch für die Differenz der Grundfrequenzwerte benachbarter Frames. Schubert zeigte in [Sch99], daß der in Oktaven gemessene Abstand benachtbarter Grundfrequenzwerte weniger geschlechtsabhängig ist. Der Merkmalsvektor von Frame k wird daher um die Quotienten aus dem Grundfrequenzwert $F_0(k \pm d)$ der benachbarten d Frames und dem Grundfrequenzwert $F_0(k)$

System	Pinyinfehlerrate
Basissystem	23.2
24 + 8 Nullen	23.3
24 + 8 F_0 Merkmale	22.0

Tabelle 5.17: Explizite Modellierung der Tonsprache Mandarin-Chinesisch

im Frame k erweitert. Dabei soll d so gewählt sein, daß die betrachtete Umgebung groß genug ist, um den Grundfrequenzverlauf innerhalb einer Silbe zu erfassen, aber nicht so groß, daß sie durch Frequenzverläufe der Nachbarsilbe überlagert werden. d wurde empririsch auf $d \in D := \{1, 2, 4, 8\}$ bestimmt, was einer Umgebungsdauer von 160 ms entspricht.

Die naheliegendste Möglichkeit zur Integration der zusätzlichen 8 Grundfrequenzmerkmale bestünde in der Erweiterung des 43-dimensionalen ursprünglichen Merkmalsvektors und anschließender Reduktion der Dimension mittels LDA auf die bisherigen 24 Dimensionen. Bei diesem Vorgehen wäre aber nicht sichergestellt, daß die 8 Grundfrequenzmerkmale im resultierenden 24-dimensionalen Vektor erhalten bleiben und die Leistungsunterschiede nicht auf genau diese 8 Merkmale bezogen werden könnten. Aus diesem Grund wird eine zweistufige LDA angewendet, wobei in der ersten Stufe die 43 ursprünglichen Dimensionen auf 24 Dimensionen reduziert werden und in der zweiten Stufe die 8 Grundfrequenzmerkmale hinzugefügt und mit einer zweiten LDA auf 30 Dimensionen reduziert werden. Die Ergebnisse dieses Experimentes sind in Tabelle 5.17 zu sehen.

Die Pinyinfehlerrate des chinesischen Erkenners konnte durch Hinzufügen der 8 Grundfrequenzmerkmale um 5.2% reduziert werden (23.2% $\rightarrow$ 22.0%). Allerdings wurde dazu gleichzeitig die Dimensionalität von 24 auf 30 erhöht. Um sicherzustellen, daß die Leistungsverbesserung auf die Modellierung der Toninformation und nicht auf die Erhöhung der Dimensionen zurückzuführen ist, wurde in einem weiteren Experiment der Merkmalsvektor um 8 Nullen statt um die 8 Grundfrequenzmerkmale erweitert. Der Vergleich in Tabelle 5.17 zeigt, daß die 5%-ige Verbesserung tatsächlich auf die Modellierung der Toninformation in Verbindung mit der zweistufigen LDA zurückgeführt werden darf.

Explizite Modellierung der Stimmhaftigkeit

Das chinesische Lautinventar enthält für Plosive stimmhafte und stimmlose Varianten und zusätzlich jeweils stärker und schwächer aspirierte Varianten. Daher wurden zusätzlich zu den 8 F_0-Grundfrequenzmerkmalen weitere 9 Merkmale für den Grad der Stimmhaftigkeit des Sprachsignals innerhalb der Frames aus der Umgebung $D := \{0, 1, 2, 4, 8\}$ zum Merkmalsvektor zugefügt. Als Maß für die Stimmhaftigkeit wird die Kreuzkorrelation zweier aufeinanderfolgender Grundperioden in Frame k verwendet (vgl. dazu [Sch99]). Insgesamt werden dem ursprünglichen Merkmals-

System	Pinyinfehlerrate
Basissystem	23.2
24 +8 F_0 Merkmale	22.0
24 +8 F_0 + 9 Stimmhaftigkeit Merkmale	21.4

Tabelle 5.18: Explizite Modellierung der Stimmhaftigkeit für Mandarin Chinesisch

vektor damit 17 weitere Merkmale zugefügt. Wie im obigen Experiment wird eine zweistufige LDA angewendet, um sicherzustellen, daß sich die Effekte auf die Modellierung der Toninformation und der Stimmhaftigkeit beziehen. Die Ergebnisse in Tabelle 5.18 zeigen, daß durch die Modellierung der Stimmhaftigkeit eine weitere Verbesserung erfolgt. Insgesamt ergibt sich gegenüber dem Basissystem eine Verbesserung der Fehlerate um 7.8% in Verbindung mit einer zweistufigen LDA.

Da durch die Merkmale der Stimmhaftigkeit die Leistung des chinesischen Erkenners von 22.0 auf 21.4 verbessert werden konnte, liegt es nahe, diese Merkmale auch auf nichttonale Sprache zu übertragen. Experimente auf der Sprache Deutsch zeigten allerdings keine Verbesserungen [Sch99].

Dimensionalität

In einem abschließenden Experiment wird untersucht, inwieweit sich die Integration der oben beschriebenen Merkmale in unterschiedlich großen Merkmalsräumen auswirkt. Dazu werden die 43 Dimensionen des ursprünglichen Vektors um die 17 Merkmale erweitert und der entstehende Merkmalsraum durch eine einstufige LDA auf 24, 30 und 36 Dimensionen reduziert. Tabelle 5.19 zeigt die Ergebnisse dieser Experimente. Im direkten Vergleich der Dimensionen sinken die Gewinne, die durch die Modellierung der Toninformation und Stimmhaftigkeit erzielt werden, auf 1-4% relative Verbesserung ab. Die in Tabelle 5.17 und 5.18 erreichten Gewinne waren also zumindest zu einem Teil durch die Erhöhung der Dimensionalität gegenüber dem Basissytem bedingt. [ZWG99] hatten bei vergleichbaren Fehlerraten von Verbesserungen derselben Größenordnung ihrer Systeme durch Erfassung der Grundfrequenz und deren erster Ableitung für die CallHome- und die Broadcastdomäne berichtet.

Dimensionen	Pinyinfehlerrate	
	ohne zusätzliche Merkmale	mit 17 zusätzlichen Merkmalen
24	23.2	23.0
30	22.0	21.7
36	22.0	21.1

Tabelle 5.19: Explizite Modellierung der Toninformation bei verschiedenen Dimensionen

Geeignete Modellierungseinheiten für Chinesisch
Das Lautinventar besteht im Bewußtsein eines Chinesen nicht aus Lauten sondern aus Silben. Die chinesische Silbe beginnt in der Regel mit einem Konsonanten und wird gefolgt von einem Vokal, seltener von einer Vokal-Konsonantkomination. Es gibt im Mandarin-Chinesisch 411 Silben mit jeweils 5 verschiedenen Grundfrequnenzverläufen. Tatsächlich sind nicht alle Kombinationen erlaubt, so daß sich das Silbeninventar auf exakt 1338 beschränkt. Jede mündliche chinesische Äußerung ist somit eine Kombination aus diesen wenigen Silben. Aus diesem Grund haben die meisten Silben mehrere Bedeutungen, einige Silben haben bis zu einhundert verschiedene Bedeutungen. Die beschränkte Zahl der Silben legt den Versuch nahe, Silben anstelle von Phonemen als Modellierungseinheiten zu verwenden. Die Modellierung der chinesischen Sprache mit Silben hat zwei Vorteile: Erstens stellen die Silben natürlichere Einheiten dar als Phoneme und sind als Ganzheit vom Grundfrequenzverlauf betroffen, zweitens erlauben Silben eine bessere Modellierung der Koartikulation (vgl. Abschnitt 3.2.2.2).

Zum Aufbau eines silbenbasierten Erkenners wurde für jede im Training vorkommende Silbe ein HMM Modell erstellt. Insgesamt ergaben sich 1269 Modelle. Die Vorverarbeitung und die HMM-Topologie wurden gegenüber der Phonemmodellierung nicht verändert. Die Silbenmodelle wurden mit den einzelnen Modellzuständen der entsprechenden Phoneme initialisiert. Die initialisierten Silbenmodelle durchliefen anschließend drei Zyklen der Trainingsprozedur.

System	Pinyinfehlerrate	
	auf Worte bezogen	auf Zeichen bezogen
kontextunabhängige Phonemmodelle	30.8	22.8
kontextabhängige Phonemmodelle	24.1	15.6
kontextunabhängige Silbenmodelle	29.1	21.3
derzeit bestes System (Phoneme)	20.7	14.5

Tabelle 5.20: Vergleich zwischen silben- und phonembasierten Modellierungseinheiten

Die Ergebnisse in Tabelle 5.20 zeigen, daß man durch die silbenbasierte Modellierung eine Reduktion der Fehlerrate von 1.7% bzw. 1.5% gegenüber der kontextunabhängigen phonembasierten Modellierung erzielt. Der kontextabhängige phonembasierte Erkenner, der auf 3000 Subpolyphonen basiert, ist allerdings signifikant besser als der silbenbasierte Erkenner. Die Verbesserung gegenüber dem kontextunabhängigen phonembasierten Erkenner ist vermutlich in erster Linie auf die bessere Kontextmodellierung zurückzuführen. Allerdings steigt der Speicherbedarf des silbenbasierten Erkenners gegenüber dem kontextunabhängigen phonembasierten Erkenner für die Codebooks um den Faktor 10 und der Zeitbedarf für den Aufbau der JRTk-Suchstruktur um den Faktor 60. Im JRTk-Dekoder wird nämlich für die Berechnung des ersten Viterbi-Suchpasses das Aussprachewörterbuch als Wald von

Bäumen strukturiert. Dieser Wald enthält soviele Bäume wie als Anfangsmodelle aller Wörter des Aussprachewörterbuches auftauchen. Durch die Modellierung von Silben statt Phonemen existiert im Silbenerkenner somit eine wesentlich größere Suchstruktur. Beim Training und Testen eines kontextabhängigen Silbenerkenners wurden denn auch die zumutbaren Grenzwerte für Speicher- und Zeitbedarf überschritten, so daß auf die weitere Entwicklung verzichtet wurde, zumal die zu erwartende Leistungssteigerung bedingt durch die bereits erzielte Kontextmodellierung geringer sein dürfte als beim phonembasierten Erkenner.

5.5.3 Behandlung agglutinierender Sprachen

Während für schwach flektierende Sprachen wie beispielsweise Englisch eine Einschränkung des Suchvokabulars auf 65K nicht mehr als etwa 0.3% OOV-Wörter resultieren, liegen die OOV-Raten für stark flektierende Sprachen wie Kroatisch oder gar agglutinierende Sprachen wie Koreanisch und Türkisch in Bereichen, die eine leistungsfähige Erkennung großer Wortschätze unmöglich machen. Im Prinzip gibt es zwei Möglichkeiten, dieses OOV-Problem zu bekämpfen. Entweder man behält die natürlichen Einheiten dieser Sprachen bei und simuliert im Erkennungsprozeß ein unbeschränktes Erkennervokabular, indem man es zwischen zwei Erkennungsphasen individuell erweitert. Oder man zerlegt die natürlichen Einheiten dieser Sprachen in kürzere Segmente und verwendet letztere als lexikalische Einheiten zur Dekodierung. In der vorliegenden Arbeit finden beide Ansätze Verwendung und werden am Beispiel der Sprachen Kroatisch, Türkisch und Koreanisch aufgezeigt.

5.5.3.1 Hypothesis Driven Lexical Adaptation

Eine Methode, die die erstgenannte Möglichkeit realisiert, ist der „Hypothesis Driven Lexical Adaptation (HDLA)“-Ansatz [Geu99]. Dieser Ansatz basiert auf der Idee, daß unbekannte Wörter bei der Dekodierung durch akustisch ähnlich klingende Wörter ersetzt werden. Insbesondere in stark flektierenden Sprachen ist die Annahme berechtigt, daß es sich bei vielen unbekannten Wörtern um nicht abgedeckte Flexionsformen bekannter Wörter handelt. Daher wird ein zweistufiger Prozeß eingesetzt, in dessen erstem Schritt die aktuelle Äußerung mit einem 65K-Vokabular dekodiert und ein Worthypothesengraph (WHG, siehe Abschnitt 3.3.2) erzeugt wird. Im zweiten Schritt wird aus diesem WHG die Liste aller hypothetisierter Wörter extrahiert und mittels großer Hintergrundkorpora mit akustisch ähnlichen Wörtern zur Größe von 65K aufgefüllt. Auf Basis der entstehenden 65K-Liste werden ein auf die aktuelle Äußerung individualisiertes Aussprachewörterbuch und ein Sprachmodell erstellt. Mit diesen Wissensquellen wird ein zweiter Erkennungslauf auf der Äußerung durchgeführt, in der Hoffnung, daß die Zahl der unbekannten Wörter für die aktuelle Äußerung reduziert werden kann.

Der HDLA-Ansatz eignet sich insbesondere für nicht zeitkritische Anwendungen wie etwa die automatische Transkription von Daten oder die Offline-Diktieranwendung, weil die Individualisierung des Aussprachewörterbuches und des Sprachmodells rechenintensiv ist und für jede Äußerung erneut durchgeführt werden muß.

Der Vorteil des HDLA-Ansatzes gegenüber einem morphembasierten Ansatz liegt darin, daß die natürlichen Einheiten einer Sprache beibehalten werden und damit die Reichweite des Sprachmodells nicht eingeschränkt wird. Außerdem sind keine Algorithmen und Wissensquellen für eine Wortzerlegung notwendig. Eine Einschränkung des Verfahrens ist, daß große Hintergrundkorpora und -lexika benötigt werden. Daher ist der HDLA-Ansatz nur für solche Sprachen sinnvoll, in denen insbesondere die Lexika bereits vorhanden sind oder sich durch automatische Tools generieren lassen.

Der HDLA-Ansatz wurde für die GlobalPhone-Sprachen Kroatisch und Türkisch durchgeführt, für die im Rahmen der Arbeit Graphem-zu-Phonem-Aussprachegeneratoren entwickelt worden waren (vgl. Abschnitt 5.3.5).

Kasus /Numerus	Singular		Plural	
Nominativ	bakşi	der Lehrer	bakşi-nar	die Lehrer
Akkusativ	bakşi-yi	den Lehrer	bakşi-nar-i	die Lehrer
Dativ	bakşi-dur	dem Lehrer	bakşi-nar-tur	den Lehrern
Genitiv	bakşi-n	des Lehrers	bakşi-nar-un	der Lehrer
Ablativ	bakşi-aca	vom Lehrer	bakşi-nar-aca	von den Lehrern
Instrumentalis	bakşi-bar	durch den Lehrer	bakşi-nar-iyar	durch die Lehrer
Komitativ	bakşi-luğa	mit dem Lehrer	bakşi-nar-luğa	mit den Lehrern

Tabelle 5.21: Prinzip der Agglutination beim türkischen Nomen

Türkisch ist das perfekte Beispiel einer agglutinierenden Sprache, d.h. die Flexion geschieht durch fortgesetztes Anhängen von Suffixen. Im Türkischen herrscht dabei das Prinzip der Monosemie, d.h. es werden keine Suffixe verschmolzen sondern in einer festgelegten Reihenfolge aneinandergereiht. Dabei bestimmt der Vokal des zu flektierenden Wortstammes nach den Gesetzen der Vokalharmonie, welche Vokale in den angehängten Suffixen stehen. Das Prinzip des Hintenanhängens wird nur in einem Fall durchbrochen: das Verstärken von Adverbien und Adjektiven geschieht durch Duplizieren und Voranstellen der ersten Silbe (Beispiel: beyaz -weiss, bembeyaz - ganz weiss; cabuk - schnell, carcabuk - sehr schnell). Tabelle 5.21 (aus [CMP98]) impliziert, wie das Agglutinationsprinzip zu einer riesigen Anzahl unbekannter Wörter führen kann, selbst wenn die Grundform eines Wortes im Erkennervokabular enthalten ist.

Die Agglutination macht es möglich, im Türkischen in einem Wort auszudrücken, wozu man in anderen Sprachen ganze Sätze benötigt. Abbildung 5.16 zeigt ein Beispiel und verdeutlicht, wie lang ein türkisches Wort sein kann. Im Beispiel ist das

Osman-lı-laş-tır-ama-yabil-ecek-ler-imiz-den-miş-siniz

Abbildung 5.16: Agglutination im Türkischen

Wort in seine morphologischen Bestandteile zerlegt, die Bedeutung des Wortes ist „verhalte Dich so als wärst Du einer von jenen, von denen wir glauben, daß sie sich nicht zum Osmanen bekehren lassen“.

In Abschnitt 5.4.4 wurde in Abbildung 5.10 bereits herausgearbeitet, daß aus der Agglutination für die türkische Sprache sehr lange Worteinheiten resultieren. In Abschnitt 5.4.5 wurden auch die Folgen für das Vokabularwachstum und die Anteile unbekannter Wörter diskutiert (vgl. Abbildung 5.14).

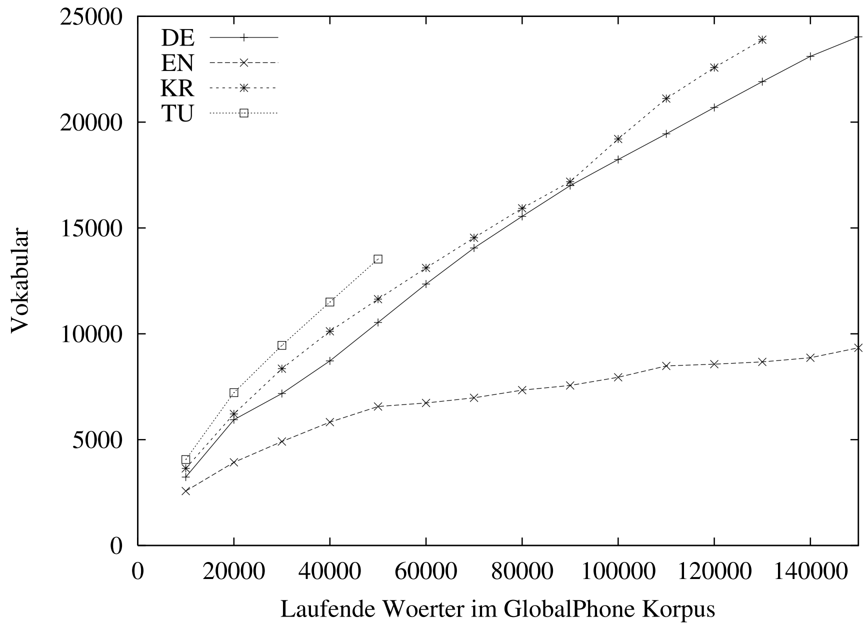


Abbildung 5.17: Vokabularwachstum von Sprachen unterschiedlicher Sprachbautypen

Wie Tabelle 5.22 zeigt, ist die kroatische Sprache ebenfalls sehr formenreich. Sie unterscheidet für Nomen drei Geschlechter und sieben Fälle jeweils im Singular und in zwei Pluralformen sowie fünf Zeitformen für Verben. Die Flexion geschieht wie im Türkischen durch Agglutination, allerdings herrscht Polysemie, d.h. die Suffixe tragen mehrere Bedeutungen. Für die kroatische Sprache ergaben sich in Abschnitt 5.4.4 ebenfalls sehr lange Einheiten, für den Korpus lag die durchschnittliche Länge aber deutlich unter der des Türkischen. Abbildung 5.17 vergleicht noch einmal das

[4] Kroatisch kennt zwei Pluralformen. Die erste Form gilt für Stückzahlen bis 4, die zweite Form für Stückzahlen ab 5.

Kasus /Numerus	Singular		Plural	
Nominativ	zakon	das Gesetz	zakon-a/i[3]	die Gesetze
Vokativ	zakon-e	Gesetz!	zakon-i	Gesetze!
Akkusativ	zakon	das Gesetz	zakon-a/e[4]	die Gesetze
Genitiv	zakon-a	des Gesetzes	zakon-a	der Gesetze
Dativ	zakon-u	dem Gesetz	zakon-ima	den Gesetzen
Instrumentalis	zakon-om	durch das Gesetz	zakon-ima	durch die Gesetze
Lokativ	zakon-u	in dem Gesetz	zakon-ima	in den Gesetzen

Tabelle 5.22: Prinzip der Agglutination beim kroatischen Nomen

Vokabularwachstum von Sprachen unterschiedlicher Sprachbautypen am Beispiel der Sprachen Deutsch, Englisch, Kroatisch und Türkisch. Deutsch liegt bedingt durch die Kompositabildung und reiche Flexion weit über der englischen Sprache. Kroatisch und Türkisch haben verglichen mit den nicht agglutinierenden Sprachen höhere Wachstumsraten. Bedingt durch die Monosemie ist Türkisch im Vergleich dieser Sprachen eindeutiger Spitzenreiter.

System	Vokabular	OOV-Rate
Basissystem Kroatisch1	31K	13.6
HDLA-Adaption Kroatisch1	31K	7.9
Basissystem Kroatisch2	49K	8.7
HDLA-Adaption Kroatisch2	49K	4.8
Basissystem Türkisch	30K	14.9
HDLA-Adaption Türkisch	30K	10.9

Tabelle 5.23: HDLA auf Kroatisch und Türkisch

Die Ergebnisse des HDLA-Ansatzes in Tabelle 5.23 zeigen, daß für die kroatische Sprache die OOV-Rate um 42% für ein 31K-Vokabular und um 45% für ein 49K-Vokabular reduziert werden konnte [GFS97]. Für die türkische Sprache wurde bei 30K eine Reduktion der OOV-Rate um 27% durch HDLA erreicht. Durch die Reduktionen der OOV-Rate konnte die Wortfehlerrate beim 31K-Vokabular um 13% und beim 49K-Vokabular um 11% relativ reduziert werden. Ein großer Anteil der durch unbekannte Wörter bedingten Fehler konnte somit durch den HDLA-Ansatz aufgehoben werden. Die Annahme, daß in agglutinierenden Sprachen die unbekannten Wörter durch andere Flexionsformen desselben Wortstammes ersetzt werden, stellt sich im Fall der kroatischen Sprache als berechtigt heraus und es ist davon auszugehen, daß sich die Reduktion der OOV-Rate in der türkischen Sprache ebenfalls in geringeren Wortfehlerraten auswirken werden. Allerdings bleiben mehr als die Hälfte aller OOV-Wörter weiterhin unbekannt. Dabei handelt es sich einerseits um

Flexionsformen bekannter Wortstämme, die nicht im Hintergrundkorpus zu finden waren, andererseits um neue, unbekannte Wortstämme. Die Experimente auf der kroatischen Sprache wurden auf den GlobalPhone-Daten und zusätzlichen Broadcast News Daten von [GFS97] durchgeführt. Die Experimente in türkischer Sprache wurden in Zusammenarbeit gemacht [ÇGS00].

5.5.3.2 Zerlegung der natürlichen Einheiten

Die zweite Möglichkeit zur Bekämpfung des OOV-Problems ist die Zerlegung der natürlichen Einheiten einer Sprache in kürzere Segmente. Als lexikalische Einheiten im Dekodierprozeß werden dann diese kürzeren Einheiten anstatt der natürlichen Einheiten verwendet. Die Zerlegung kann entweder wissensbasiert beispielsweise durch eine morphologische Analyse der natürlichen Einheiten oder durch eine silbenbasierte Zerlegung oder aber vollständig datengetrieben durchgeführt werden. Die resultierenden Einheiten sollten idealerweise so beschaffen sein, daß sie einerseits eine ausreichende Länge haben, um akustisch nicht zu verwechselbar zu sein und um die Reichweite des Sprachmodells nicht zu sehr zu beschränken, andererseits sollten sie kurz genug sein, um die OOV-Rate gering zu halten.

Der Vorteil der Zerlegung natürlicher Einheiten gegenüber dem HDLA-Ansatz besteht in der Echtzeitfähigkeit. Außerdem benötigt man bei der Zerlegung keine großen Hintergrundkorpora und -lexika. Gegenüber der wissensbasierten Methode hat die datengetriebene Methode den Vorteil, daß kein Vorwissen über das morphologische oder silbenbasierte Konzept der Sprache notwendig ist und sie sich daher leicht auf beliebige andere Sprachen erweitern läßt.

Im Rahmen dieser Arbeit wurde für die türkische Sprache eine silbenbasierte Zerlegung erprobt [Ç98, ÇGS00]. Darüber hinaus wurde für die koreanische Sprache eine datengetriebene Methode zur automatischen Bestimmung geeigneter Einheiten entwickelt und mit einer wissensbasierten morphologischen Dekomposition verglichen [Kie99, KSW99].

Silbenbasierte Zerlegung

Türkisch hat, wie Tabelle 5.21 bereits nahelegte, eine fast perfekte Relation zwischen Morphologie und sprachlicher Funktion. Es liegt daher nahe, das Problem der OOV-Rate im Türkischen durch eine morphologische Zerlegung der natürlichen Einheiten anzugehen. Dazu wird allerdings Expertenwissen in Form eines morphologischen Analysetools benötigt. Nachforschungen im Internet ergaben, daß im Rahmen der „Turkish Natural Language Processing Initiative“ [OB94] ein expertenbasiertes morphologisches Analysetool „Xcorpus“ entwickelt wurde, das auf dem bekannten Analysetool „PC-Kimmo“ von Kimmo Koskennienmi aufsetzt. Freundlicherweise wurde Xcorpus von der oben genannten Initiative zur Verfügung gestellt. Leider stellte sich heraus, daß die morphologische Zerlegung türkischer Texte nicht eindeutig möglich ist, sondern in etwa 10% aller Fälle von Muttersprachlern korrigiert werden muß (vgl. auch [OT96]). Da der Korrekturaufwand für einen 15 Millionen Wörter umfassenden Text zu groß ist, wurde in dieser Arbeit ein neuer Ansatz entwickelt: Die Zerlegung der natürlichen türkischen Einheiten wurde auf Silbenbasis durchgeführt. Als Regelwerk zur Silbentrennung wurde die türkische Version der GNU-Latex-Trennvorschriften verwendet.

Wortebene	yaratılmasında
Silben	ya- ra- tıl- ma- sın- da
$S134$	yara_1 tıl_2 ma_3 sın_3 da_3
$S145$	yaratıl_1 ma_2 sın_3 da_3
$S135$	yara_1 tılma_2 sın_3 da_3

Tabelle 5.24: Beispiel für die Zerlegung eines türkischen Wortes

Die resultierenden Silben wurden anschließend zu längeren Einheiten verschmolzen. Dazu wurden positionsbedingte Silbenklassen $Si_1i_2i_3$ berechnet, in denen aufeinanderfolgende Silben bis zu einer definierten Position i_n zu einer Einheit verbunden werden. Das Silbenklassensystem $S134$ entsteht demnach dadurch, daß alle Anfangssilben (Position 1) mit der direkten Nachfolgersilbe (Position 2) verschmolzen werden und in eine gemeinsame Silbenklasse zusammenfallen. Die Silben, die bei der Zerlegung an 3. Stelle stehen, bleiben unverbunden und bilden eine eigene Klasse, ebenso wie die Silben, die an 4. oder späteren Stellen stehen. Die Bestimmung der Silbenklassen ist heuristisch basiert und orientiert sich an der morphologischen Struktur der Sprache sowie an der Anzahl entstehender Klassenelemente und -vokabulare.

System	Vokabular	Splits	OOV-Rate	WE
Basissystem	30K	1	15.3	34.1
$S134$	14K	1.63	6.0	39.0
$S145$	21K	1.24	7.6	35.7
$S135$	17K	1.45	6.8	37.0

Tabelle 5.25: Silbenbasierte Zerlegung auf Türkisch [WE in %]

Aus Experimenten an einem vorläufigen türkischen System ergaben sich als beste Zerlegungsklassen $S134$, $S145$ und $S135$. Tabelle 5.24 zeigt die resultierenden Silbenklassen der drei Systeme an einem Beispiel. Das Problem der limitierten Reichweite des Sprachmodells wurde durch ein klassenbasiertes 4-Gramm-Modell angegangen (vgl. Abschnitt 3.2.3), wobei als Klassen die Silbenklassen definiert wurden.

Der Vergleich der Ergebnisse der silbenbasierten Systeme mit dem Basissystem zeigt, daß in allen drei Fällen das Erkennervokabular deutlich reduziert und die OOV-Rate mehr als halbiert werden konnte. Im Mittel wurde dazu jedes Wort in 1.2

bis 1.6 Silbenklassen aufgeschnitten. Leider konnte die Reduktion der OOV-Rate nicht in eine Reduktion der Fehlerrate umgesetzt werden. Das System $S145$ kommt immerhin nahe an das Basissystem heran bei einer Reduktion des Erkennervokabulars um 30%, was sich in Laufzeitvorteilen auswirkt. Die Tatsache, daß keine Leistungsverbesserungen erzielt werden, liegt unter anderem an der suboptimalen Zusammenfügung der Silbenklassen zu Wörtern in den Hypothesen. Bislang wurden nämlich ausschließlich die segmentierten Einheiten der besten Hypothese zu Wörtern zusammengefaßt. Aussichtsreicher wäre die Zusammenfügung auf der Basis des Worthypothesengraphen.

Morphembasierte Zerlegung

Koreanisch ist wie Türkisch eine agglutinierende Sprache. An die lexikalischen Stämme können mehr als 400 Suffixe angehängt werden, mit denen grammatikalische Beziehungen wie Tempus, Aspekt, Aktionsart, Modus und Honorativ ausgedrückt werden. Darüber hinaus existieren viele Komposita, die wie im Deutschen ohne Leerzeichen aneinandergefügt sind. Das Resultat ist ein nahezu lineares Vokabularwachstum, wie die Abbildung 5.18 zeigt.

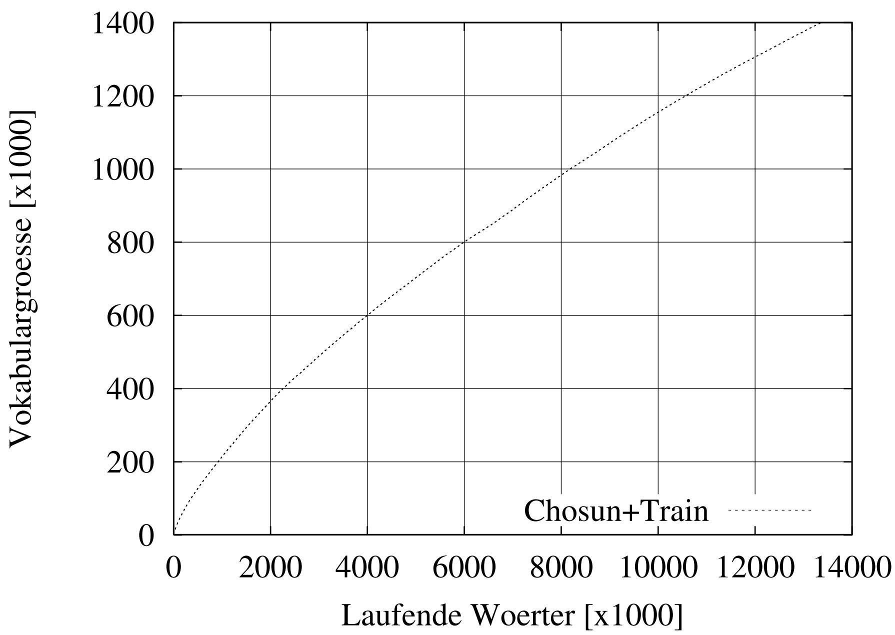


Abbildung 5.18: Vokabularwachstum für Koreanisch

Zur exakten Analyse der OOV-Raten wurde ein großes koreanischer Textkorpus basierend auf der im Internet verfügbaren koreanischen Zeitung *Chosunilbo* (siehe Tabelle 5.4, S. 99) zusammengestellt. Dieses Korpus wurde mit den GlobalPhone-Trainingsdaten zum Textkorpus *Chosun+Train* zusammengefaßt. Zur Entwicklung wurden 15% des *Chosun*-Textkorpus mit der Trainingsmenge zum Textkorpus

PartChosun+Train vereint. Der Textkorpus *Train* referenziert die Textmenge der GlobalPhone-Trainingsdaten. Die Abbildung 5.19 veranschaulicht für die drei genannten Textkorpora die resultierenden OOV-Raten, wenn man die natürlichen Einheiten der koreanischen Sprache zugrundelegt. Man sieht, daß bei einem 65K-Vokabular die Zahl unbekannter Wörter deutlich über 30% liegt. Selbst bei einem potentiellen Erkennervokabular von 1.4 Millionen Wörtern läge die OOV-Rate noch über 11%. Zur automatischen Spracherkennung der koreanischen Sprache ist somit die Zerlegung der Eojeol in kleinere Einheiten unumgänglich.

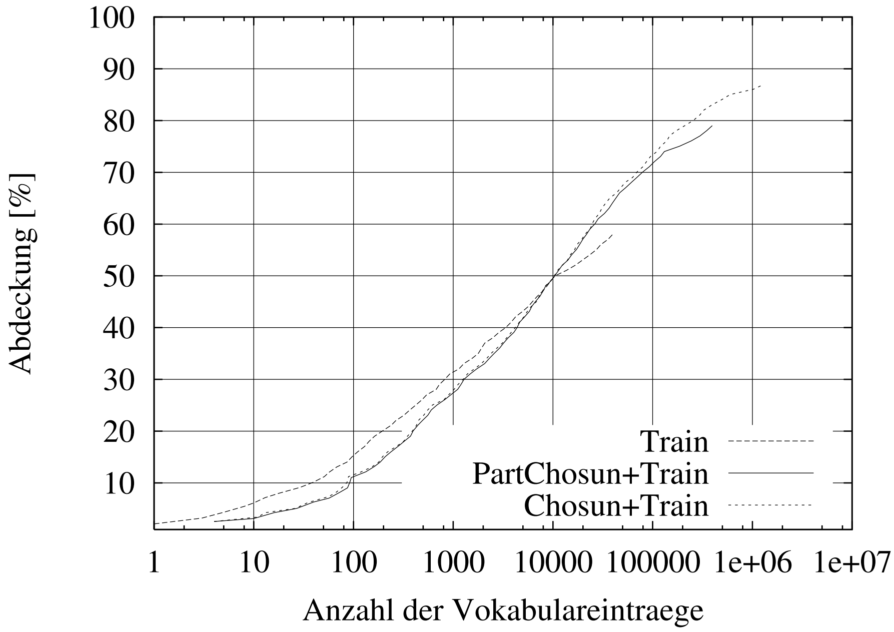


Abbildung 5.19: OOV-Raten für Koreanisch

Während die morphologische Zerlegung im Türkischen mit dem Xcorpus Tool nicht ohne Handkorrekturen durchzuführen war, existiert für die koreanische Sprache ein vollständig automatisches morphologisches Zerlegungstool [Kim96], das vom Electronics and Telecommunications Research Institute (ETRI) in Seoul, Korea auf das GlobalPhone-Korpus angewendet wurde und dessen Ausgabe freundlicherweise zur Verfügung gestellt wurde [Kwo99]. Dabei wird jedes Eojeol in die morphologischen Komponenten zerlegt. Jedes Morphem wird mit einem Part-of-Speech (POS)-Etikett versehen, das angibt, welche grammatikalische Funktion dieses Morphem im vorliegenden Satz hat. Auf der Basis dieser morphologischen Zerlegung wurden zwei Systeme entwickelt: System *MorphTag*, in dem die POS-Etiketten als Teil der lexikalischen Einheit behandelt werden und System *Morph*, in dem die POS-Etikette entfernt wurden. Der Vorteil von *Morph* liegt in einem kompakteren Vokabular, der Vorteil des *MorphTag* in der zu vermutenden höheren Aussagekraft des Sprachmodells durch die zusätzliche Angabe der grammatikalischen Funktion. Die Vokabular-

System	Fehlerrate			OOV	PP
	Eojeol	Silbe	Phonem		
Morph	24.0	13.0	9.4	2.9	143
MorphTag	30.0	16.2	10.7	3.5	486

Tabelle 5.26: Vergleich morphologischer Zerlegungsansätze für Koreanisch

größe der morphembasierten Systeme reduziert sich auf ein Viertel verglichen mit Eojeol und die OOV-Rate bei einem 65K-Vokabular fällt von über 30% bei Eojeols auf etwa 3% bei Morphemen, was als durchaus akzeptable OOV-Rate betrachtet werden kann.

Tabelle 5.26 zeigt die Fehlerraten der morphembasierten Erkennersysteme jeweils gemessen in Eojeoleinheiten, Silbeneinheiten und in Phonemeinheiten. Mit dem System *Morph* werden bessere Erkennungsleistungen erzielt, als mit *MorphTag*. Dieses Resultat ist zum einen durch die niedrigere OOV-Rate von *Morph* bedingt, zum anderen aber auch durch die signifikant höhere Trigramm-Perplexität von *MorphTag*, die sich aus dem größeren Vokabular und der höheren OOV-Rate ergibt. Insgesamt zeigen die Resultate, daß die morphologische Zerlegung zu akzeptablen OOV-Raten und mäßigen Vokabularwachstumsraten führten.

Datengetriebene Zerlegung

Obwohl die morphologische Zerlegung zu sehr guten Erkennungsleistungen führt, bleibt es unbefriedigend, daß zur Zerlegung ein Expertensystem verwendet werden muß, dessen Funktionalität nicht beeinflußt werden kann und zu dessen Erstellung muttersprachliches Wissen und ein großer zeitlicher Aufwand benötigt werden. Im Rahmen dieser Arbeit wurde daher eine Methode implementiert, die rein datengetrieben eine geeignete Zerlegung der Einheiten ermittelt [Kie99]. Gegenüber der wissensbasierten morphologischen Methode hat die datengetriebene Methode den Vorteil, daß keinerlei Expertenwissen notwendig ist, und sie sich daher generell auf andere Sprachen übertragen läßt. Darüber hinaus unterliegt sie keinen Fehlern des morphologischen Zerlegungstools.

Die datengetriebene Methode geht von der Silbenzerlegung der Eojeols aus. Auf dem *Chosun+Train*-Korpus ergibt sich ein Vokabular von etwa 3600 Silben und eine OOV-Rate von 0%. Wie bereits in Abschnitt 5.4.4 beschrieben, eignen sich diese Einheiten aufgrund ihrer akustischen Verwechselbarkeit und der geringen Sprachmodellreichweite jedoch nicht für einen leistungsfähigen Erkenner. Bei der Fehleranalyse des auf Silbeneinheiten basierten Erkenner zeigte sich, daß die häufigsten Erkennungsfehler durch Substitutionsfehler zwischen zwei Silbenpaaren entstehen, deren Übergang VokalSilbe1–Silbengrenze–VokalSilbe2 identisch ist. Beispielsweise haben die Silbenpaare *sin eop (= S I N EO P)* und *si neo (=S I N EO)* denselben Übergang *I N EO*.

System	Fehlerrate			OOV	PP
	Eojeol	Silbe	Phonem		
Datengetrieben	24.6	14.5	9.9	0.2	137
Morphembasiert	24.0	13.0	9.4	2.9	143

Tabelle 5.27: Vergleich zwischen datengetriebener und morphembasierter Zerlegung

Die Idee des datengetriebenen Ansatzes besteht darin, die akustische Verwechselbarkeit dadurch zu verringern, daß diese verwechselbaren Übergänge durch das Verschweißen der Silben eingeschlossen werden. Im genannten Beispiel würden die vier Einheiten *sin, eop, si, neo* durch zwei Einheiten *sineop, sineo* ersetzt. Die Erkennungseinheiten werden dadurch länger, es ergibt sich eine bessere Polyphonausnutzung und die Reichweite des Sprachmodells erhöht sich. Zur Auswahl der zu verschweißenden Silben wurden zunächst alle Übergänge im Korpus gezählt und entsprechend ihrer Auftrittshäufigkeit sortiert. Dann wurde beginnend mit dem häufigsten Übergang dasjenige Silbenpaar miteinander verschweißt, das den aktuellen Übergang am häufigsten produziert. Als Abbruchkriterium dieses Algorithmus diente die 65K-Vokabulargrenze.

Tabelle 5.27 stellt das datengetriebene System dem besten morphembasierten System gegenüber. Die Leistungen beider Systeme sind nahezu äquivalent, was beweist, daß eine rein datengetriebene Bestimmung geeigneter sprachlicher Modellierungseinheiten eine gute Alternative zu Morphemeinheiten darstellt. Das datengetriebene System kann sich mit den Ergebnissen von State-of-the-art-Erkennern anderer Forschungsgruppen auf vergleichbaren Aufgaben durchaus messen. Die datengetriebene Vorgehensweise erlaubt die Systementwicklung ohne dabei Expertenwissen vorauszusetzen und bietet die Möglichkeit der Übertragbarkeit auf andere Sprachen. Dieses Ergebnis ist insbesondere im Kontext der multilingualen Spracherkennung von großer Bedeutung.

5.6 Zusammenfassende Bewertung der monolingualen Spracherkenner

Die multilinguale GlobalPhone-Datenbasis bietet die einmalige Möglichkeit, Spracherkenner für große Wortschätze in vielen unterschiedlichen Sprachen anhand einheitlicher Daten zu entwickeln und zu vergleichen. Insgesamt stehen zum Training der Systeme etwa 220 Stunden Sprache zur Verfügung. Dieses umfangreiche Datenmaterial verteilt sich allerdings auf viele Sprachen, so daß im Mittel etwa 15 Stunden Trainingsmaterial pro Sprache verfügbar ist. Diese Menge liegt um eine Größenordnung unter dem, was zum Training von State-of-the-art Diktiererkennern üblich ist. Deshalb sind für die monolingualen Basiserkenner keine Leistungen zu erwarten, wie sie aus der Literatur für Diktiersysteme bekannt sind.

Der Fokus der Arbeiten, die in diesem Kapitel beschrieben sind, liegt allerdings auch nicht auf der Entwicklung möglichst performanter Systeme, sondern auf der möglichst effizienten Entwicklung von Erkennern für viele verschiedene Sprachen. Dazu wurde die Systementwicklung weitestgehend automatisiert und für 10 Sprachen durchgeführt. Zur Initialisierung der monolingualen Erkenner wurde ein Verfahren entwickelt, das auf bereits trainierte akustische Modelle anderer Sprachen aufsetzt. Die Ergebnisse zeigen, daß die Initialisierung mittels eines mehrsprachigen Phonempools die besten Leistungen erreicht (Abschnitt 5.3.8). Notwendige sprachenspezifische Wissensquellen, wie das Aussprachewörterbuch wurden soweit möglich automatisch generiert. Zu diesem Zweck wurden zahlreiche Graphem-zu-Phonem Tools entwickelt, die die Aussprache eines Wortes aus dessen Schreibweise generieren (Abschnitt 5.3.5). Deren Analyse ergab, daß eine große Varianz in der Relation zwischen Graphemen und Phonemen innerhalb der Sprachen besteht (Abschnitt 5.4.1).

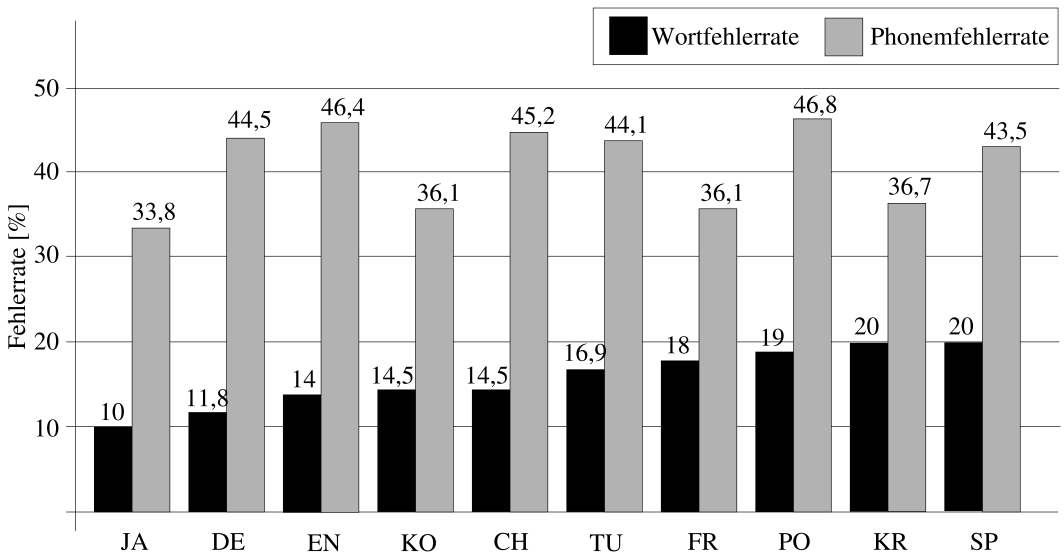


Abbildung 5.20: Fehlerraten für 10 GlobalPhone-Sprachen

Abbildung 5.20 zeigt die Wortfehlerraten der entwickelten Systeme für 10 Sprachen im Vergleich. Der Begriff „Wortfehlerrate" trifft nicht ganz zu, weil nicht in allen untersuchten Sprachen ein Wortkonzept existiert. Die erläuterten Unterschiede im Schriftsystem (Abschnitt 5.4.1) und in der Segmentierung (Abschnitt 5.4.4) bewirken, daß die Fehlerraten nicht in vergleichbaren Einheiten angegeben werden können. Für die künstlich segmentierten Sprachen Chinesisch, Japanisch und Koreanisch (Abschnitt 5.5.1) werden die in der Literatur üblichen Zeichenfehlerraten (Character Error Rate) als Fehlermaß herangezogen. Die unterschiedliche Länge der Grundeinheiten (Abschnitt 5.4.4) in Kombination mit dem morphologischen Bau einer Sprache (Abschnitt 5.4.4) bestimmen das Vokabularwachstum und damit die Zahl der unbekannten Wörter in einem Erkennungssystem. Außerdem sind kurze

Grundeinheiten akustisch leichter verwechselbar und resultieren in einer geringeren Reichweite des Sprachmodells. In Abbildung 5.20 sind daher auch die Phonemfehlerraten eingetragen, die mit frei laufenden Phonemerkennern entstanden (Abschnitt 5.4.2). Sie bieten die Möglichkeit, die Erkennungsraten zwischen verschiedenen Sprachen auf akustischer Ebene miteinander zu vergleichen, ohne daß die Ergebnisse durch die genannten Faktoren überlagert werden.

Die wesentliche Beobachtung im Vergleich der Phonemerkennungsleistungen ist, daß eine Korrelation zwischen der Größe des Phoneminventars und der Phonemerkennungsleistung existiert (Abschnitt 5.4.2). Es bilden sich im wesentlichen zwei Leistungsgruppen mit Japanisch, Französisch, Koreanisch und Kroatisch in der einfacheren Gruppen und mit deutlichem Abstand Türkisch, Spanisch, Deutsch, Chinesisch, Englisch und Portugiesisch. Japanisch scheint akustisch einfach zu erkennen zu sein, was einerseits durch das kompakte Phoneminventar erklärt werden kann, sowie durch die eingeschränkte Phonotaktik aufgrund der Morastruktur, eventuell auch durch den recht hohen Anteil an Vokalen, die einfacher zu erkennen sind als Konsonanten. Kroatisch ist ebenfalls phonetisch recht einfach. Auch hier hat man ein kleines Phonemset und außerdem eine sehr gut Relation zwischen Graphemen und Phonemen. Im anderen Extrem finden sich Portugiesisch, das ein sehr großes Phoneminventar mit vielen oralen und nasalen Vokalen hat, die zu zahlreichen Verwechslungen beitragen. Englisch ist ebenfalls bekannt für seine hohe Verwechselbarkeit. Deutsch ist schwierig, möglicherweise durch das Konsonanten-Vokal Verhältnis, das in dieser Sprache am größten ist (Abschnitt 5.4.2). Außerdem deutet die enorme Anzahl an Polyphonen (Abschnitt 5.4.3) auf eine komplexe deutsche Phonotaktik hin. Chinesisch ist trotz der hohen Zahl an Phonemen nicht signifikant schlechter als Deutsch, was darauf schließen läßt, daß Toneme ausreichend gut disambiguiert werden.

Die Wortfehlerraten liegen in allen Sprachen zwischen 10% und 20% und damit in vergleichbaren Größenordnungen. Dieser Umstand zeigt, daß sich die angewendeten Algorithmen der klassischen Spracherkennung prinzipiell zur Modellierung aller untersuchten Sprachen eignen. Aus Vergleichbarkeitsgründen wurden die Erkenner aller Sprachen nach einheitlichen Gesichtspunkten entwickelt. So ist bei den hier präsentierten Ergebnissen die Signalvorverarbeitung, die HMM-Topologie und die Zahl der Modellparameter identisch. In den Sprachen mit hohen OOV-Raten wurden diese kontrolliert, indem die unbekannten Wörter in das Erkennervokabular aufgenommen und als Unigramme in das Sprachmodell eingetragen wurden.

Im Vergleich haben die einzelnen Entwicklungsschritte in allen Sprachen zu ähnlichen Verbesserungen geführt. Im typischen Verlauf der Systementwicklung entstand nach der Initialisierung eine Fehlerrate von etwa 70%. Durch das Training der Modelle auf der jeweiligen Sprache in mehreren Trainingszyklen wurde eine Fehlerrate von etwa 40% - 50% erzielt, also eine relative Verbesserung von etwa 30% - 40% durch Training auf sprachenspezifischem Material. Durch die Modellierung von Kontexten erreicht man im Schnitt etwa 20% relative Verbesserung in allen Sprachen. Die VTLN erzielt etwa 5-10% relative Verbesserungen in allen Sprachen.

Eine weitere Reduktion der Fehlerraten konnte anhand der Analyse vorläufiger Erkennungsergebnisse durch manuelle Korrekturmaßnahmen erreicht werden. Diese betrafen die Überarbeitung der Aussprachewörterbücher als auch eventuelle Datenkorrekturen. Außerdem wurden die Textkorpora zur robusteren Berechnung der Sprachmodelle erweitert. Die Erfahrungen der Systementwicklung in 10 Sprachen haben ergeben, daß darüber hinaus Maßnahmen zur Behandlung sprachenspezifischer Besonderheiten notwendig waren. Dies betrifft insbesondere die spracheninhärenten Probleme wie die Bestimmung geeigneter Einheiten für nicht segmentierte Sprachen wie Chinesisch und Japanisch, für die agglutinierenden Sprachen wie Koreanisch und Türkisch und für stark flektierende Sprachen wie Kroatisch. Erzielte Teilergebnisse etwa am Beispiel der koreanischen Sprache haben gezeigt, daß zur Bestimmung geeigneter Einheiten auch datengetriebene Methoden erfolgreich sind.

Die ermittelten enormen Unterschiede zwischen Sprachen führen vor Augen, daß der Vergleich der monolingualen Systeme auf der Basis von Wortfehlerraten sehr schwierig ist. Die erzielten Ergebnisse in den einzelnen Sprachen sind sicherlich an der einen oder anderen Stelle durch zusätzlichen Entwicklungsaufwand und vor allem zusätzlichen Daten weiter verbesserbar. Auch werden Sprachen mit enger Graphem-zu-Phonem-Relation durch den automatisierten Aussprachegenerierungsansatz etwas begünstigt.

Insgesamt stehen mit diesen monolingualen Erkennern erstmal große Wortschatzerkenner in vielen Sprachen zur Verfügung und bieten damit eine hervoragende Ausgangsbasis für die Kombination vieler Sprachen in ein multilinguales Erkennungssystem.

Kapitel 6

Multilinguale Spracherkennung

Multilinguale Spracherkennung ist ein sehr aktuelles Forschungsthema, in dem die Begriffe und Ansätze noch nicht einheitlich definiert und beschrieben sind. In diesem Kapitel wird daher zunächst der Forschungsstand aufgearbeitet und strukturiert. Anschließend wird die eigene Arbeit mit dem Schwerpunkt der Kombination akustischer Modelle vorgestellt und evaluiert. Die Beschreibung der Anwendungen resultierender multilingualer Modelle zur Sprachenidentifizierung und Portierung auf neue Sprachen runden dieses Kapitel ab.

6.1 Ziele und Kriterien

Multilinguale Spracherkennung beschäftigt sich mit der sprachenübergreifenden Nutzung von Daten im akustischen Modell, Aussprachewörterbuch oder Sprachmodell (vgl. Abschnitt 3.1). Entsprechend der Wissensquelle, die multilingual gestaltet ist, sind dabei verschiedene Ziele und Kriterien anzulegen. Ein *multilinguales Sprachmodell* wird benötigt, um Code-Switching innerhalb einer Äußerung zu realisieren und kann gemeinsam mit einem *multilingualen Aussprachewörterbuch* zur Sprachenidentifizierung eingesetzt werden.

Der Schwerpunkt dieser Arbeit liegt allerdings auf der Entwicklung von Methoden zur statistischen Modellierung *multilingualer akustischer Modelle*. Die Motivation der multilingualen akustischen Modellierung läßt sich in den folgenden Punkten beschreiben:

- Reduktion der Anzahl zu modellierender Parameter des Gesamtsystems
- Bessere Ausnutzung vorhandener Trainingsdaten vieler Sprachen
- Schaffung einer günstigen Ausgangsbasis zur Modellierung neuer, noch nicht präsentierter Sprachen
- Sprachenidentifizierung

Die Reduktion der Parameterzahl hat einen positiven Einfluß auf das Laufzeitverhalten und ist damit prinzipiell für alle mehrsprachigen Anwendungen von Bedeutung. Außerdem führt die Verringerung der Systemkomplexität zu einer verbesserten Wartbarkeit. Die Möglichkeiten der Sprachenidentifizierung ist ein erwünschtes Nebenprodukt. Das Hauptaugenmerk dieser Arbeit liegt jedoch auf den beiden Aspekten, der besseren Ausnutzung vorhandener Daten, die im Idealfall die Sammlung weiterer Daten überflüssig machen könnte und dem damit verknüpften Aspekt der Portierung akustischer Modelle auf neue Sprachen, in denen gar kein oder nur wenig Datenmaterial vorhanden ist. Die im folgenden beschriebene Modellkombination ist daher auch im Hinblick auf ihren späteren Einsatz in einer multilingualen Anwendung und zur Portierung auf neue Sprachen konzipiert.

Zur Beurteilung der Modellkombinationsmethoden werden mehrere Kriterien berücksichtigt:

- Wortfehlerrate
- Parameterzahl
- Datenausnutzung
- Größe des Phoneminventars

Eines der wichtigsten Kriterium im Kontext der Spracherkennung ist die Wortfehlerrate. Mittels dieses Kriteriums werden die multilingualen mit den monolingualen Modellen verglichen und die Eignung zur Portierung auf neue Sprachen ermittelt. Ein weiteres Kriterium, nach dem die Effizienz der Modelle beurteilt werden kann, ist die Zahl der Parameter, die vom Modell zu lernen sind. Die Ausnutzung von Daten bezieht sich auf den Aspekt, wie sich die Daten auf die einzelnen Modelle verteilen. Die Gesamtgröße des Phoneminventars ist deshalb von Interesse, weil sie Rückschlüsse auf die Systemkomplexität zulassen.

6.2 Stand der Forschung

Der Forschungsbereich „multilinguale Spracherkennung“ ist sehr aktuell und hat gerade erst begonnen sich zu entwickeln. Daher gibt es derzeit noch keinen einheitlichen Fachjargon und es finden sich kaum einführende Darstellungen publizierter Forschungsansätzen. Es wird daher in diesem Abschnitt zunächst versucht, eine Strukturierung der derzeitigen Forschungsansätze vorzunehmen und die wesentlichen Begriffe zu definieren. Dazu sind die aus der Literatur bekannten Forschungsarbeiten bis März 2000 eingeflossen und beziehen auch diejenigen Arbeiten mit ein, die entweder parallel oder aufbauend auf den Publikationen der Autorin entstanden sind. Die Strukturierung orientiert sich daran, welche der in Kapitel 3 ausgeführten Wissensquellen von mehreren Sprachen gemeinsam genutzt werden.

6.2.1 Multilinguale Sprachmodelle

Multilinguale Sprachmodelle werden derzeit zur Sprachenidentifizierung und zur multilingualen Diktiererkennung eingesetzt. Im letztgenannten Bereich ist der Sprachenwechsel innerhalb eines Satzes erwünscht, der nach [Hun97] vorrangig zur Einbettung anderssprachiger Phrasen verwendet wird, wie beispielsweise in Sprache ist nicht einfach, wie man an „time flies like an arrow“ sieht. Dennoch gibt es bisher erst wenige publizierte Arbeiten, die sich mit der Kombination von Sprachmodellen beschäftigen. In den bekannten Arbeiten werden N-Gramm-Modelle verwendet, die nach der Art der Kombinationsmethode unterschieden werden können:

1. Berechnung der N-Gramme auf einem vereinten multilingualen Textkorpus; Sprachenübergänge innerhalb einer Äußerung werden mit speziellen Übergangsstrafen kontrolliert, d.h. erschwert [CDG$^+$97, WRN$^+$98].

2. Berechnung der N-Gramme auf monolingualen Textkorpora; Sprachenübergänge innerhalb einer Äußerung werden durch einen gemeinsamen Backoff-Knoten realisiert [WBNS97, HNN98, AHG$^+$98].

3. Berechnung der N-Gramme auf monolingualen Textkorpora, Verknüpfung durch ein gemeinsames Satzanfangs- und Satzendesymbol; Sprachenübergänge innerhalb einer Äußerung sind nicht möglich [WBNS97, HNN97].

Neti und seine Arbeitsgruppe erstellten in [CDG$^+$97, WRN$^+$98] ein bilinguales klassenbasiertes Trigramm-Backoff-Sprachmodell auf den Sprachen Englisch und Französisch, indem sie die Textkorpora beider Sprachen zu einem Korpus vereinen (Methode 1). Dieses Sprachmodell führte zu großen Leistungsverlusten, die durch einen Backoff auf Unigramme der falschen Sprache entstanden, was mit einem Ungleichgewicht der monolingualen Textmaterialien begründet wird. Zur Lösung des Problems führten sie in das Sprachmodell eine Sprachenübergangsstrafe ein. Diese Maßnahme führte zu Leistungsverbesserungen, konnte aber die Gesamtleistungseinbußen gegenüber zwei monolingualen Sprachmodellen nicht ausgleichen.

In [AHG$^+$98, HNN98] stellt Harbeck ein vierlinguales Sprachmodell auf den Sprachen Slowenisch, Slowakisch, Tschechisch und Deutsch vor, das Code-Switching durch einen gemeinsamen Backoff-Knoten ermöglicht (Methode 2). Der Backoff-Knoten ist als gemeinsame Pause-Kategorie realisiert, so daß der Übergang zwischen zwei Sprachen immer durch den Einschub einer Pause erfolgt. Im Vergleich zum monolingualen Erkenner verliert der multilinguale Erkenner im Schnitt 3.25 Prozentpunkte, wobei die Verluste ungleich auf die Sprachen verteilt sind.

In [WBNS97] vergleicht Weng auf den Sprachen Englisch und Schwedisch die Kombinationsmethoden 2 und 3 miteinander. Die Modellkombination durch ein gemeinsames Satzanfangs- und -endesymbol führte zu Einbußen, die aber durch Ausbalancieren der Textmaterialien eliminiert werden konnten. Dagegen zeigten die Experimente zum Sprachmodell mit gemeinsamen Backoff-Knoten eine dramatische Verschlechterung der Erkennungsraten. Zur Verbesserung der Erkennungsraten wurde eine Nachbewertung der N-Bestenliste durchgeführt, bei der sprachengemischte Hypothesen entsprechend dem Mischungsverhältnis bestraft werden. Diese Maßnahme führte zu kleinen, aber nicht signifikanten Verbesserungen.

Harbeck erstellt in [HNN97] ein bilinguales Bigramm-Sprachmodell auf den Sprachen Slowenisch und Slowakisch nach der Kombinationsmethode 3. Dieses Sprachmodell erlaubt kein Code-Switching, hat aber den Vorteil, daß der Dekodieraufwand in der Sprachenidentifizierung reduziert werden kann. Anhand von Laufzeituntersuchungen weist [HNN97] nach, daß innerhalb der ersten zwei Sekunden alle Suchpfade der nicht hypothetisierten Sprache weggeschnitten werden. Dies verhindert lange Dekodierzeiten, die entstehen, weil einer der beiden Erkenner stets die „falsche“ Sprache dekodiert.

6.2.2 Multilinguale Aussprachewörterbücher

Wie bereits im Abschnitt 3.1 beschrieben, wird der Begriff *multilinguales Aussprachewörterbuch* zum einen für die Verzahnung monolingualer Aussprachewörterbücher verwendet, zum anderen für das Problem die Aussprachen von Wörtern, die im Phonemset einer Sprache beschrieben sind, auf das Phonemset einer anderen Sprache abzubilden.

Bei der Zusammenfügung monolingualer Aussprachewörterbücher kann man im allgemeinen davon ausgehen, daß das resultierende multilinguale Aussprachewörterbuch aus einem Nebeneinander verschiedensprachiger Wörter besteht, die in einen gemeinsamen Rahmen gebracht werden. Daneben sind aber auch Anwendungen denkbar, in denen das Aussprachewörterbuch verzahnt wird, weil die Wörter verschiedener Sprachen als Aussprachevariante eines *gemeinsamen* Wortes realisiert werden. Dies ist dann sinnvoll, wenn die zu sprechenden Einheiten gemeinsamen Gesetzmäßigkeiten gehorchen, wie beispielsweise in der Ziffernerkennung, der Buchstabiererkennung, oder Anwendungen, die in der Hauptsache Eigennamen verwenden, wie Navigationssysteme im Auto [Sho99], Informationssysteme im Bereich Finanzwesen oder Flugauskunft [MPF99] oder im Katasterwesen [ÜSN98].

Das Problem der Abbildung von Aussprachen entsteht bei der Portierung auf neue Sprachen. Dabei müssen die Aussprache der Wörter der Zielsprache möglichst adäquat durch die Phoneme der Quellsprache(n) ausgedrückt werden. Micca [MPF99] stellt zur Lösung des Problems ein Konzept der stationär-transitorischen akustischen Einheiten vor, das auch als kontextabhängige Subwortmodellierung bezeichnet wird [SJR+95]. Die Aussprache eines Wortes wird dabei nicht nur durch die Konkatenation von Phonemen (dem stationären Anteil), sondern auch durch die explizite Formulierung des Übergangs von einem zum anderen Phonem (dem transitorischen Anteil) beschrieben. Die stationären Anteile bleiben sprachenspezifisch, die transitorischen Anteile dagegen können über Sprachen hinweg vermischt werden. Dazu werden die stationären Einheiten heuristisch in wenige Grundklassen eingeteilt und auf dieser Basis Transitionsklassen durch ein Ballungsverfahren [BGM97] gebildet. Die Resultate aus [MPF99] zeigen, daß durch dieses Verfahren Leistungsgewinne erzielt werden können, sofern ausreichend Adaptionsmaterial in der neuen Sprache vorhanden ist. Falls nur wenige oder keine Daten vorhanden sind, mit denen die Transitionsklassen auf die neue Sprache adaptiert werden können, ergibt sich allerdings eine drastische Leisungseinbuße. Dies ist vermutlich die Folge der Tatsache, daß durch die Modellierung der transitorischen Einheiten die Anzahl benötigter Grundmodelle deutlich ansteigt, was viel mehr Trainingsdaten erfordert.

6.2.3 Multilinguale akustische Modelle

Diese Arbeit beschäftigt sich schwerpunktmäßig mit der Kombination akustischer Modelle basierend auf der Modellierung von Wörtern auf Phonembasis. Daher ist der Darstellung relevanter Forschungsarbeiten an dieser Stelle ein breiter Raum gewidmet. Am Ende dieses Abschnittes werden zwei alternative Vorschläge beschrieben, die phonologisch motivierte distinktive Merkmale zur Modellierung verwenden.

6.2.3.1 Phonembasierte Verfahren

Die Idee, akustische Phonemmodelle über Sprachen hinweg gemeinsam zu nutzen, wurde in der Spracherkennung erstmals von Dalsgaard, Andersen und Barry [DA92] formuliert. Sie prägten den Begriff *Polyphoneme* für Phoneme, die einander so ähnlich sind, daß ihre Modelle von mehr als einer Sprache verwendet werden können, und den Begriff *Monophoneme* für solche Phoneme, die für genau eine Sprache spezifisch sind. Das von [ADB93] definierte Poly*phonem* ist nicht mit dem in Abschnitt 3.2.2.2 beschriebenen Poly*phon* zu verwechseln, das ein Phonem im Kontext seiner angrenzenden Nachbarn bezeichnet. Der Ansatz von [DA92] war vom Problem der Sprachenidentifizierung motiviert und wurde zu diesem Zweck als erstes eingesetzt. Daalsgard und Andersen nutzten sowohl Monophoneme als auch Polyphoneme zur Identifizierung [DA94, AD97] und andere übernahmen diese Idee [CAGADL97, KH97]. Wieder andere Forscher konzentrierten sich auf die Monophoneme, mit dem Argument, daß diese sprachdiskriminierende Informationen enthalten [BABC94, ZS95]. Multilinguale Phonemmodelle verbesserten die Leistungen bei der Sprachenidentifizierung [CAGADL97, AD97] und nährten die Hoffnung, daß sich diese Verbesserung auch bei der Spracherkennung einstellen würde.

In diesem Abschnitt werden die bisherigen Arbeiten skizziert, in denen die Kombination akustischer Modelle beschrieben sind. Es werden drei Methoden unterschieden, nach denen monolinguale Lautinventare zu multilingualen Lautinventaren kombiniert werden:

- Heuristische Kombination auf der Basis linguistischen Wissens:
    - Phonetisch/artikulatorisch [DA92, CDG+97, WRN+98, WBNS97]

- IPA-basiert [Köh97, Köh98, Köh99, SW98c] oder Sampa-basiert [AAB+96, AAB+97, ÜSN98]

- Rein datengetriebene Kombination:
  - Bestimmung von Phonemähnlichkeiten auf Basis einer Phonemverwechslungsmatrix [ADB94, DA94, ADB93, Imp99]
  - Bestimmung von Phonemähnlichkeiten auf Basis einer Kombination verschiedener Distanzmaße [BGM97, MPF99]
  - Agglomeratives Ballungsverfahren basierend auf:
    * Likelihood-Distanzen [AD97, Köh99]
    * A-posteriori-Distanzen [CAGADL97]

- Hierarchische Kombination:
  1. Schritt: Heuristische Einteilung der Phoneme in Phonemklassen
  2. Schritt: Datengetriebene Ballung innerhalb der definierten Klassen [Köh99, Köh96, WBNS97, CDG+97, WRN+98, SW98c, SW98b]

Heuristische Kombination
Daalsgard und Andersen bildeten in [DA92] multilinguale Phonemklassen nach auditiven phonetischen Kriterien. Diese Ansatz wurde unter anderem von [CDG+97] und [WRN+98] übernommen. Andere Gruppen ließen sich ebenfalls von phonetischen Kriterien leiten, verwendeten aber zur Phonemklassifizierung entweder grobe Artikulationsklassen [WBNS97] oder Referenzschemata, wie IPA [Köh97, Köh98, Köh99] oder Sampa [AAB+96, AAB+97, ÜSN98].

Bei allen genannten Arbeiten wird nach der Einteilung in Ähnlichkeitsklassen ein gemeinsames Modell für jede Phonemklasse trainiert, indem die Daten aller Sprachen, die ein Phonem in der jeweiligen Klasse haben, zum Training dieses Modells verwendet werden (Prinzip der gemeinsamen Nutzung von Daten). In allen genannten Arbeiten werden auf den multilingualen akustischen Modellen schlechtere Ergebnisse als auf den monolingualen Modellen erzielt, wenn man sie zur Erkennung der Trainingssprachen einsetzt. Eine Ausnahme ist die Anwendung multilingualer Modelle auf akzentbehaftete Sprache, die Übler in [AAB+96, AAB+97, ÜSN98] beschreibt. Im Rahmen des EU-Projekt SPEEDATA hat sie Sprache von uni- und bilingualen Personen aus Südtirol verarbeitet. Infolge der besonderen sprachlichen Situation können dort dialekt- und akzentbehaftete Sprache in Deutsch und Italienisch studiert werden. Diese Situation macht die Mischung von akustischen Modellen über beide Sprachen besonders attraktiv. Durch die heuristische Kombination von 49 italienischen und 62 deutschen Phonemmodellen zu 87 multilingualen Modellen gelang eine Verbesserung der Erkennungsleistungen auf den Sprechern, die eine der

Sprachen nur schlecht beherrschten. Bei Muttersprachlern waren dagegen die monolingualen Modelle besser als die multilingualen, was die Ergebnisse anderer Arbeiten bestätigt.

Datengetriebene Kombination

Andersen präsentierte in [ADB93] den ersten datengetriebenen Ansatz zur Kombination von Phonemmodellen für vier europäische Sprachen. Dazu ermittelte er eine Phonemverwechslungsmatrix zwischen allen Phonemen und berechnete darauf ein Ähnlichkeitsmaß. Die ähnlichsten Phoneme wurden zu Polyphonemen zusammengefaßt. Diese Polyphoneme wurden in Phonemerkennungsexperimenten [ADB93, ADB94, DA94] gegen ihre monolingualen Pendant ausgetauscht, wodurch eine signifikante Verbesserung der Erkennungsleistung erzielt wurde. Die Annahme der Autoren, daß eine zunehmende Robustheit der Modelle diese Verbesserung zustandebrachte, konnte bis heute jedoch nicht bestätigt werden. Tatsächlich stellen sich Verbesserungen durch multilinguale Modelle häufig als Konsequenz aus mehr Daten heraus. In [ADB93] waren zum Training der sprachenspezifischen Modelle nur sechs Minuten Sprache verfügbar, durch die gemeinsame Nutzung von Daten konnten die Polyphonemmodelle mit bis zur vierfacher Datenmenge trainiert werden. Die Resultate beweisen, daß multilinguale Modelle das Problem von untertrainierten Modellen beheben können. Sie beweisen aber nicht, daß die multilinguale Modellierung generell zu besseren Erkennungsleistungen führen, als die monolingualen Modelle.

Imperl [Imp99, IH99] erweiterte das Verfahren von [ADB93] auf kontextabhängige Modelle. Dazu definiert er die Ähnlichkeit zwischen zwei Triphonen verschiedener Sprachen als eine Kombination der Ähnlichkeiten zwischen deren Zentralphonemen und den beiden jeweiligen Monophonen des linken und rechten Kontextes. Der anschließende Ballungsprozeß basiert auf einem Distanzmaß, das als eine gewichtete Summe der beschriebenen Triphonähnlichkeiten definiert ist. Triphone verschiedener Sprachen, deren Ähnlichkeit über einer empirisch bestimmten Schwelle liegen, werden zu gemeinsamen Modellen geballt. Imperl erreichte auf drei Sprachen im besten Fall eine Reduktion der Anzahl der Triphon-Modelle um 40% gegenüber dem monolingualen Fall bei gleichzeitigem relativem Wortfehleranstieg um 9%.

Corredor stellt in [CAGADL97] ein hierarchisches agglomeratives Ballungsverfahren vor, in dem ein unsymmetrisches Ähnlichkeitsmaß zwischen Phonemen verwendet wird, das die a-posteriori Wahrscheinlichkeiten approximiert. Mit diesem Ähnlichkeitsmaß werden 148 sprachenspezifische Phoneme aus vier Sprachen (Englisch, Spanisch, Französisch, Deutsch) zu 83 Klassen geballt, von denen 48 Monophoneme und 35 Polyphoneme sind. Es wurden 7 Polyphoneme gefunden, die sich alle 4 Sprachen teilen: die stimmlosen Plosive /p/, /t/, /k/ sowie die Frikative /s/ und /f/ und die Nasale /n/ und /m/.

Auch Köhler schlägt in [Köh96, Köh99] ein agglomeratives Ballungsverfahren vor, verwendet aber als Ähnlichkeitsmaß zwischen zwei Phonemen deren Likelihood-Distanz und wendet sie auf sechs Sprachen an. Das datengetriebene Verfahren erreicht gegenüber dem heuristischen IPA-basierten Verfahren eine bessere Leistung, bleibt aber hinter der monolingualen Modellierung zurück. Interessanterweise findet Köhler durch das agglomerative Ballen dieselben Polyphoneme wie Corredor [CAGADL97].

Andersen und Daalsgard verwenden in [AD97, DAB98] ebenfalls Likelihood-basierte Distanzen zum datengetriebenen agglomerativen Ballen von Phonemmodellen aus drei Sprachen. In [DAB98] wird die Bedeutung der entstandenen Modellklassen analysiert. Danach können viele Ballungen zwischen Phonemen artikulatorisch-phonetisch begründet werden, was impliziert, daß das Mischen von Phonemen über Sprachen hinweg nicht grundsätzlich falsch ist. Es bleiben allerdings eine Reihe von Phonemklassen, deren Gemeinsamkeiten nicht phonetisch begründbar sind.

In [BGM97] präsentiert Bonaventura ein Distanzmaß, das auf der Kombination von fünf unterschiedlichen Ähnlichkeitsmaßen auf Gaußschen Mischverteilungen basiert. Zwei Phoneme verschiedener Sprachen werden einer gemeinsamen Klasse zugeordnet, sofern deren Distanz unter einen empirisch bestimmten Schwellwert fällt. Das Zusammenlegen der Phoneme geschieht iterativ auf vier Sprachen beginnend mit 133 Modellen. Wie Köhler und Corredor erhält er die vierlingualen Polyphoneme /n/, /m/, /t/, /k/ und /s/. Zusätzlich findet Bonaventura die Vokalklassen /o/, /e/ und /a/. In verschiedenen Erkennungsexperimenten setzt er die gewonnenen Polyphonemmodelle ein, was zu einem großen Leistungsverlust führt, der mit wachsender Anzahl beteiligter Sprachen größer wird. Dagegen zeigt er die Vorteile multilingualer Modelle im Vergleich zu schlecht generalisierten und untertrainierten monolingualen Modellen.

Hierarchische Kombination
Rein datengetriebene agglomerative Kombinationsmethoden haben neben dem hohen Berechnungsaufwand zwei wesentliche Nachteile. Erstens sind die entstehenden Klassen nicht immer einsichtig, und zweitens fehlt eine Abbildungsfunktion von diesen Klassen auf Phoneme, die beispielsweise für die Portierung auf neue Sprachen wünschenswert wäre. Eine hierarchische Kombination kann die genannten Nachteile beheben. Dazu wird ein Zweischritt-Verfahren durchgeführt: Im ersten heuristischen Schritt werden Phonemkategorien festgelegt, im zweiten datengetriebenen Schritt werden innerhalb dieser Phonemkategorien Modelle zusammengeballt.

Weng [WBNS97] experimentiert mit zwei Sprachen. Zur Modellkombination legt sie im ersten Schritt 11 Phonemkategorien nach phonetischen Gesichtspunkten fest und evaluiert zunächst den Effekt der gemeinsamen Datenausnutzung. Dazu vergleicht sie innerhalb einer Phonemkategorie das Mischen der Daten von Phonemen innerhalb der Sprachen (11+11 Modelle) zu Phonemen verschiedener Sprachen (1x11 Modelle). Die Erkennungsleistungseinbußen beim Datenmischen über Sprachen hinweg betragen 5%. Im zweiten Schritt entwickelt Weng auf der Basis der Phonemkategorien ein System, bei dem sich alle Phoneme derselben Phonemkategorie dieselben Codebooks teilen. Durch ein agglomeratives Ballungsverfahren wird bestimmt, welche

HMM-Zustände sich dieselben Mixturgewichte teilen. Die multilingualen Modelle zeigen durchweg schlechtere Erkennungsraten als monolinguale Modelle beim Test auf den Trainingsprachen.

[CDG+97, WRN+98] führten eine Methode des divisiven Ballens durch. Dazu fügten sie neben den Fragen nach phonetischen Kontexten zusätzlich Fragen nach der Sprache eines Modells hinzu. Anhand der Analyse des entstandenen Fragenbaums entwickelten sie zwei Systeme: eines, bei dem Phoneme, an denen Sprachenfragen wurzelnah auftauchten, als monolinguale Phoneme modelliert werden und ein zweites System, bei dem die final geballten Phonemmodelle verwendet werden. Diese beiden Ansätze verglichen sie mit dem oben beschriebenen heuristischen Ansatz. Das Gesamtergebnis zeigt bei allen Kombinationsarten einen Performanzverlust gegenüber den monolingualen Systemen. Die Einbußen liegen beim besten System, das auf den von Sprachenfragen geballten Modelle basiert, bei 5-9% im Vergleich zu den monolingualen Systemen, allerdings enthält diese Fehlerrate den Verlust durch eine Kombination des Sprachmodells.

Köhler definiert in [Köh99] eine hierarchische Kombinationsmethode, bei der im ersten Schritt alle Phoneme anhand des IPA-Schemas zu Phonemklassen zusammengefaßt werden. Im zweiten Schritt wird innerhalb dieser Phonemklassen eine agglomerative Ballung der Gaußschen Mischverteilungen durchgeführt, bei der je zwei hinreichend ähnliche Verteilungen verschiedener Sprachen zu einer Verteilung geballt werden, bis eine definierte Zahl von Verteilungen erreicht ist. Die Mixturgewichte bleiben sprachenspezifisch. Köhler vergleicht diese Methode mit der heuristischen IPA-basierten Modellkombination und mit dem rein datengetriebenen agglomerativen Ballen. Seine Resultate zeigen, daß die heuristische Kombinationsmethode am schlechtesten, die hierarchische Methode am besten abschneidet. Trotzdem sind alle drei Verfahren den monolingualen akustischen Modellen unterlegen.

6.2.3.2 Nicht-phonembasierte Verfahren

Die Leistungsfähigkeit des phonembasierten Ansatzes hat in den letzten 20 Jahren rasant zugenommen und dabei von der wachsenden Zahl von Datenbasen und von neuen Algorithmen profitiert, die zum Modellernen aus Daten entwickelt wurden. Allerdings vertreten immer mehr Forscher die Meinung, daß die phonembasierte Technologie auf einem blinden datengetriebenen Prinzip mit unstrukturierten Basisspracheinheiten fußt, deren Leistung durch die Generalisierungsfähigkeit begrenzt ist [Den97a]. Die Grenzen einer adäquaten Modellierung durch Phoneme werden mit zunehmenden Anforderungen an die Robustheit von Spracherkennungssystemen immer deutlicher. Da man in der klassischen Spracherkennung Sätze bzw. Wörter durch eine sequentielle Abfolge von Phonemen bildet, müssen diese Phonemmodelle eine wachsende Zahl von Parametern durch immer größere Datenmengen lernen. Die Zahl der zu lernenden Modellparameter wird immer größer, die Modelle immer spezifischer und somit die Generalisierung immer kleiner [Den97a, Van99]. Die Probleme werden deutlich, sobald Trainingsdaten und Testsituation nicht mehr übereinstimmen.

Diese fehlende Übereinstimmung manifestiert sich in einer breiten Palette von Faktoren wie wechselnde Sprecher-, Kanal-, Kontext-, oder Sprechstilbedingungen. Die multilinguale Spracherkennung kann als eine extreme Variante einer fehlenden Übereinstimmung zwischen Trainingsdaten und Testdaten aufgefaßt werden. [WTK98] führen die Probleme der phonembasierten multilingualen Spracherkennung nicht nur auf die begrenzte Generalisierungsfähigkeit zurück, sondern darauf, daß Phoneme sprachenkontrastierend definiert und daher inhärent sprachenabhängig sind. Nach ihrer Meinung ergibt sich daher immer das Problem der Abwägung zwischen Modellierungsgenauigkeit und Portierungsfähigkeit.

Die Lösung für dieses Dilemma sehen manche in der Abkehr von der phonembasierten Modellierung. Im Kontext der multilingualen Spracherkennung haben dies Deng [Den97b] und Williams [WTK98] formuliert. Williams schlägt als Spracherkennungseinheiten sub-segmentale Elemente vor, die in allen Sprachen vorkommen und die auf dem linguistischen Modell der „Government Phonologie" basieren. Danach lassen sich alle Laute durch eine Kombination von sieben sogenannten distinktiven Merkmalen bilden. Nach Williams sind diese distinktiven Merkmale zuverlässiger zu erkennen als Phoneme (85-90% Klassifikationsrate auf verschiedenen Sprachen). Ihrer Ansicht nach müßte anhand einer multilingualen Datenbasis nur ein einziger Erkenner trainiert werden, der alle Sprachen zukünftig abdeckt. Ihr Vorschlag sieht vor, die Ausgabe von sieben Merkmalsdetektoren in einem hybriden System als Eingabe in ein segmentbasiertes HMM zu führen. Eine Implementierung des Systems ist allerdings noch nicht realisiert und es konnten bisher keine Ergebnisse präsentiert werden, die den Erfolg dieses Ansatzes nachweisen.

Deng [Den97b] schlägt eine strukturierte Modellierung des Sprachgenerierungsmechanismus nach dem Produktionsmodell von Saltzmann vor. Dabei werden Merkmale definiert, die mit Parametern assoziiert sind, welche die dynamischen Eigenschaften von Vokaltraktvariablen kontrollieren. Dies entspricht im wesentlichen dem Konzept der distinktiven Merkmale. Eine Menge von Regeln legt fest, wie sich diese Merkmale zeitlich überlappen dürfen. Mittels eines endlichen Zustandsautomaten, dessen Zustände eine Merkmalssequenz repräsentieren, werden Wörter oder ganze Sätze produziert. Deng berichtet, daß die Zahl zu modellierender Parameter gegenüber der phonembasierten Modellierung drastisch sinkt. Insbesondere verringere sich der Aufwand des Übergangs von einer zur nächsten Sprache. Nach Deng mußte die Merkmalsmenge, die ursprünglich für Englisch entwickelt worden war, nur geringfügig erweitert werden, um Mandarin-Chinesisch abzudecken. Aufgrund der Beschränktheit der Lautbildungsmöglichkeiten würde nach seiner Annahme bald die Menge aller Merkmale definiert sein, die alle Sprachen abdeckt.

Obwohl die Revision des klassischen Ansatzes einleuchtend klingt, konnte bis dato noch nicht belegt werden, daß ein derart konzipierter Erkenner tatsächlich sprachenunabhängig funktioniert und bei der Portierung auf neue Sprachen die erhoffte Leistung und Effizienz zeigt.

6.2.4 Zusammenfassung

In der multilingualen Spracherkennung hat sich bisher gezeigt, daß es durch die Kombination akustischer Modelle und Sprachmodelle zwar einerseits zu einer enormen Parameterreduktion kommt, die besonders für speicherlimitierte Anwendungen wichtig sind. Andererseits zeigen die Ergebnisse verschiedener Forschungsgruppen, daß die Datenmischung über Sprachen hinweg zu einer Reduktion der Gesamtleistung führt.

Zur Kombination akustischer Modelle wurden die heuristische, die datengetriebene und die hierarchische Methode erprobt. Der Vorteil der ersten Methode liegt in ihrer Einfachheit und der direkten Abbildungsmöglichkeit auf eine neue Sprache. Die Nachteile liegen in der Notwendigkeit von Expertenkenntnissen über Quell- und Zielsprachen. Die datengetriebene Methode beseitigt diese Nachteile, allerdings sind die aus der Ballung resultierenden Klassen oft schwer zu interpretieren. Außerdem wird zur Übertragung auf neue Sprachen Datenmaterial der Zielsprachen benötigt. Aus diesem Grund eignet sich die datengetriebenen Methode nicht zur Portierung auf neue Sprachen, wenn keine Daten in der Zielsprache vorliegen.

Die hierarchische Methode verbindet die Vorteile der heuristischen und datengetriebenen Variante miteinander. Sie ermöglicht die Übertragung von Modellen ohne Daten der Zielsprache, trotzdem sind die Phonemkategorien datengetrieben trainiert. Eine Leistungssteigerung durch die Kombination akustischer Modelle wurde bisher bei akzentbehafteter Sprache [ÜSN98], bei dialektbehafteter Sprache [FGJ98] und bei durch Datenmangel untertrainierten Modellen [ADB93, GG97] beobachtet. Die Kombination von Sprachmodellen ermöglicht den Wechsel von Sprachen innerhalb einer Äußerung und kann vorteilhaft in der Sprachenidentifizierung eingesetzt werden. Die Idee eines nicht-phonembasierten Ansatzes klingt vielversprechend, es wurde aber bis dato noch nicht experimentell bestätigt, ob dieser Ansatz vergleichbare oder bessere Leistungen erbringt als der phonembasierte Ansatz.

6.3 Multilinguale akustische Modellkombination

In diesem Abschnitt werden die Ansätze und Methoden zur statistischen Modellierung multilingualer Modelle dargelegt, die in dieser Arbeit entwickelt wurden. Dabei werden die Ziele und Kriterien berücksichtigt, wie sie im Abschnitt 6.1 definiert worden sind. Nach einer Beschreibung der entwickelten Ansätze werden diese anhand von Experimenten analysiert und evaluiert.

6.3.1 Globales Phonemset

Als Folge der geplanten Nutzung multilingualer Modelle kommt eine rein datengetriebene Kombination akustischer Modelle aus zwei Gründen nicht in Betracht: Erstens besteht dabei grundsätzlich das Problem, daß Laute derselben Sprache in eine gemeinsame Klasse fallen. Dieses Phänomen kommt nicht selten aufgrund von Datenartefakten (vgl. etwa [ADB93, Köh96]) zustande und bewirkt, daß der Phonemsatz einer Sprache kollabieren kann. In diesem Fall ist keine optimale Modellierung der einzelnen Sprache mehr möglich. Zweitens resultieren datengetriebene Verfahren in Klassen, die keine eindeutige Zuordnung zu Phonemen zulassen (vgl.[DAB98]) und daher eine Portierung auf neue Sprachen ohne Datenmaterial unmöglich machen. Da in dieser Arbeit unter anderem Ansätze vorgestellt werden, bei denen keine Daten vorhanden (siehe Kapitel 7) sind, sind die heuristischen und hierarchischen Kombinationsmethoden von besonderer Bedeutung für diese Arbeit.

Als Ausgangsbasis dieser Kombinationsmethoden wird ein Phonemset entwickelt, das einen möglichst breiten Sprachraum abdeckt. In dieser Arbeit wird dabei von der Annahme ausgegangen, daß die artikulatorischen Repräsentationen von Phonemen verschiedener Sprachen hinreichend ähnlich sind, so daß man Phoneme als von der Sprache unabhängig betrachten kann. Auf der Grundlage dieser Annahme kann das sprachenspezifische Phoneminventar von N Sprachen zu einem *universellen Phoneminventar* Υ vereinigt werden: $\Upsilon = \Upsilon_{L_1} \cup \Upsilon_{L_2} \cup \ldots \cup \Upsilon_{L_N}$.

Nach [ADB93] wird zwischen der Menge sprachenunabhängiger *Polyphoneme* Υ_{LI}, die nur solche Phoneme enthält, die in mehr als einer Sprache vorkommen, und N verbleibenden Mengen sprachenabhängiger *Monophoneme* $\Upsilon_{LD_{L_1}}, \ldots \Upsilon_{LD_{L_N}}$ differenziert. Die Menge $\Upsilon_{LD_{L_m}}$ enthält die Phoneme, die nur in Sprache L_m vorkommen. Daher ist $|\Upsilon_{LD_{L_m}}| = 0$, wenn jedes Phonem der Sprache L_m ein Gegenstück in mindestens einer der restlichen $N-1$ Sprachen hat.

Das Poly*phonem* ist nicht mit dem in Abschnitt 3.2.2.2 beschriebenen Poly*phon* zu verwechseln, das ein Phonem im Kontext seiner angrenzenden Nachbarn bezeichnet.

Die Erstellung eines globalen Phonemsets erfordert eine Charakterisierung von Lauten und ein Maß für die Ähnlichkeit zwischen Lauten verschiedener Sprachen. In dieser Arbeit wird dazu als Referenzschema das in Kapitel 2 erläuterte IPA-System [IPA93] verwendet. Die sprachenspezifischen Laute werden damit auf der Basis phonetischen Wissens eingeteilt.

Definition des globalen Phonemsets

Die Definition des *globales Phonemsets* basiert auf den Lauten von 12 GlobalPhone-Sprachen. Alle Laute, die durch dasselbe IPA-Symbol repräsentiert sind, werden einer gemeinsamen Lautklasse im globalen Phonemset zugeordnet. Dazu wird für zwei Laute υ_i, υ_j die Funktion sameIPA(υ_i, υ_j) definiert:

$$\text{sameIPA}(\upsilon_i, \upsilon_j) = \begin{cases} 1, & \exists \upsilon \in \Upsilon_{LI} : \upsilon_i \simeq \upsilon \wedge \upsilon_j \simeq \upsilon \text{ für } i \neq j \\ 0, & \text{sonst} \end{cases}$$

Da die optimale Modellierung jeder individuellen Sprache im Sinne der monolingualen Spracherkennung erhalten bleiben soll, werden die Phoneminventare der besten Erkennersysteme bewahrt, wie sie in Tabelle 5.11 (S. 110) beschrieben sind. Eine

Ausnahme bildet die chinesische Sprache, bei der auf die explizite Tonhöhenmodellierung verzichtet wird, um eine höhere Kompression des Phonemsets zu erreichen.

Tabelle 6.1 zeigt das entstandene globale Phonemset für die 12 Sprachen. In der oberen Hälfte der Tabelle befinden sich alle Polyphoneme, also alle υ_i zu denen ein υ_j existiert mit $\text{sameIPA}(\upsilon_i, \upsilon_j) = 1$. Für jedes Polyphonem ist in der Tabelle die Zahl der Sprachen belegt, die sich dieses Polyphonem teilen. In der unteren Tabellenhälfte sind die Monophoneme jeder Sprache beschrieben, also alle υ_i mit $\forall \upsilon_j, i \neq j : \text{sameIPA}(\upsilon_i, \upsilon_j) = 0$.

Die ursprünglich 485 sprachenspezifischen Phoneme aus allen GlobalPhone-Sprachen werden durch die Abbildungsfunktion sameIPA in 162 Phonemkategorien eingeteilt. Von diesen 162 Phonemkategorien werden 83 Kategorien durch die Daten mehrerer Sprachen trainiert. Die restlichen 79 Phonemkategorien bestehen aus Einzelphonemen, die jeweils nur durch die Daten einer Sprache trainiert werden.

Als Maß für die Ausnutzung der Daten, d.h. der zusätzlichen Menge an Daten, die im Mittel auf ein multilinguales Modell entfallen, wird der *share factor* sf_N für N Sprachen definiert als die Relation zwischen der Summe sprachenspezifischer Phoneme und der Größe des globalen Phonemsets. sf_N gibt die durchschnittliche Anzahl der Sprachen an, die ein Phonem des globalen Phonemsets gemeinsam nutzen:

$$\text{sf}_\text{N} = \frac{\sum_{\text{i}=1}^{\text{N}} |\Upsilon_{\text{L}_\text{i}}|}{|\Upsilon|}, \quad |\Upsilon| = |\Upsilon_{\text{LI}}| + \sum_{\text{i}=1}^{\text{N}} |\Upsilon_{\text{LD}_{\text{L}_\text{i}}}| \tag{6.1}$$

Es gilt die Bedingung $1 \leq \text{sf}_\text{N} \leq \text{N}$. sf_N ist eins, wenn kein einziges Polyphonem existiert, und N, wenn alle N Sprachen dasselbe Phoneminventar haben.

$$\text{sf}_{12} = \frac{|\Upsilon_{\text{ch}}| + |\Upsilon_{\text{de}}| + \ldots + |\Upsilon_{\text{tu}}|}{|\Upsilon|} = \frac{485}{162} = 2.99 \tag{6.2}$$

Für den vorliegenden Fall, mit $N = 12$, erhält man mit 485 sprachenspezifischen Modellen einen Faktor sf_{12} nahe drei. Das bedeutet, daß jedes Phonem des globalen Phonemsets im Mittel von drei Sprachen genutzt wird und entsprechend mit der dreifachen Datenmenge trainiert wird. Um herauszufinden, wie stark sf von den beteiligten Sprachen abhängt, wird sf_k über alle möglichen k-Tupel $(k = 1, \cdots, 12)$ der 12 Sprachen $\binom{12}{k}$ berechnet. Zusätzlich wird der mittlere Quotienten $\text{pm}_\text{k} = |\Upsilon_{\text{LI}}| / \sum_{\text{i}=1}^{\text{N}} |\Upsilon_{\text{LD}_{\text{L}_\text{i}}}|$ zwischen der Anzahl der Polyphoneme und der Anzahl der Monophoneme für alle k-Tupel ermittelt. Die Ergebnisse für sf_k und pm_k sind in Abbildung 6.1 zu sehen.

Genutzt von	#	Modellierte Phoneme (IPA Symbole)	
	83	Polyphoneme: verwendet von $\geq$ 2 Sprachen	
		Konsonanten	Vokale
Allen	4	m,n,s,l	-
11	7	p,b,t,d,k,g,f	-
10	3	-	i,u,e
9	6	ŋ,v,z,j	a,o
8	1	ʃ	-
7	3	r,h,tʃ	-
6	1	-	ɛ
5	9	ɲ,ʒ,x,ts,dʒ	iː,y,ɘ,ɔ
4	4	-	ɨ,ø,ɑ,ei
3	11	ʎ,w,ç	ɪ,uː,eː,œ,oː,æ,ai,aʊ
2	34	pʰ,tʰ,dʲ,kʰ,gʲ,ʁ,ɾ$_r$, θ,ð,sʲ,zʲ,ʂ,ɻ,tsʰ,tʃʲ	ˈi,yː,ɯ,ʊ,ˈe,ɛː,øː,aː,ˈa,ɑː, ˈu,ˈo,aɪ,au,ia,io,eu,oi,oʊ
	79	Monophoneme: gehören zu genau *einer* Sprache	
		Konsonanten	Vokale
CH	15	tʂ,tʰʂ,cç,cçʰ	iʊ,iɛ,ua,uɛ,uɔ,ya,yɛ, iao,uɛi,uai,ioʊ
KR	1	dʒʲ	-
EN	5	ɾ$_d$	ʌ,ɝ,ɔi,ɚ
FR	5	ɥ	ɛ̃,œ̃,ɑ̃,ɔ̃
DE	3	-	ɐ,ʏ,ɔʏ
JA	2	ʔ	ɯː
KO	14	p˺,p’,t˺,t’,k˺,k’, s’,c’ʰ	ie,iɘ,iu,ɨi,oa,uɘ
PO	8	-	ĩ,ũ,ẽ,õ,ɐ̃,ew,ow,aw
RU	15	pʲ,bʲ,tʲ,mʲ,rʲ,vʲ,ʃʲ,ʒʲ,lʲ,ʃtʃ,ʃtʃʲ	ja,jɛ,jɔ,ju
SP	2	β,ɣ	-
SW	9	ʈ,ɖ,ɳ,ɭ,ks	œː,æː,ʉː,ɵ
TU	0	-	-
$\sum$	162	Stille und 2 Geräusche werden von allen Sprachen genutzt	

Tabelle 6.1: Globales Phonemset für 12 Sprachen

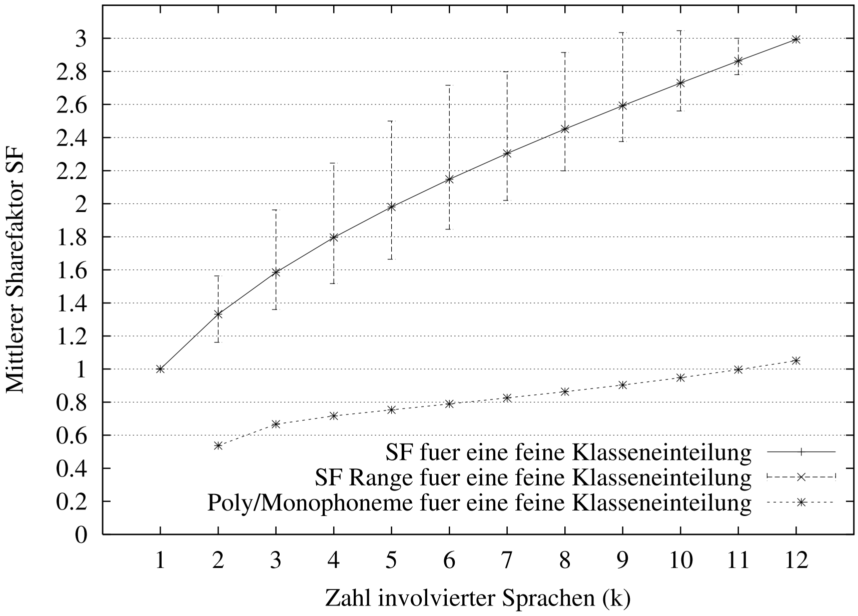


Abbildung 6.1: $\mathrm{sf_k}$ und $\mathrm{pm_k}$ für $\binom{12}{k}$ Sprachtupel für feine Klassengranularität

Die wesentlichen Beobachtungen aus Abbildung 6.1 sind: Erstens wächst $\mathrm{sf_k}$ erwartungsgemäß mit der Zahl k der involvierten Sprachen, allerdings ist das Wachstum eher linear, d.h. daß sich auch mit 12 Sprachen noch keine Sättigung des Phoneminventars einstellt. Zweitens zeigt sich, daß die Spannweite des mittleren Faktors sf enorm groß ist. So hat beispielsweise aus allen 6-Tupeln der $\binom{12}{6} = 924$ Paarungen das Tupel *RU-PO-KO-JA-CH-SW* den geringsten Faktor mit $\mathrm{sf_6} = 1.8$ bei einem Poly- zu Monophonemquotienten von $\mathrm{pm_6} = 0.5$, während das Tupel *KR-JA-FR-DE-TU-SW* das Maximum mit $\mathrm{sf_6} = 2.7$ und $\mathrm{pm_6} = 1.4$ hat.

Insgesamt bleibt pm im Mittel relativ konstant. Die Anzahl der Polyphoneme übersteigt die der Monophoneme erst nachdem 11 Sprachen involviert sind, was die ausbleibende Sättigung unterstreicht. Dagegen hatten die Beobachtungen aus Untersuchungen wie etwa [Köh99, CAGADL97] die Hoffnung geweckt, daß die Zahl der Polyphoneme mit wachsender Sprachenzahl stärker zunimmt als beobachtet. Allerdings basierten die aus der Literatur bekannten Untersuchungen auf homogeneren und deutlich weniger Sprachen.

Die Beobachtungen führen zu der Frage, ob eine grobere Klasseneinteilung der Phoneme den Nutzungsgrad erheblich steigern würde, oder ob das Verhältnis zwischen den Poly- und Monophonemen durch die breite Streuung der Sprachen zustandekommt, die sich in sehr divergenten Phonemen ausdrückt. Es werden daher $\mathrm{sf_k}$ und $\mathrm{pm_k}$ für grobere Phonemklasseneinteilungen berechnet. Die Abbildung 6.2 zeigt das Ergebnis einer Einteilung, welche die Gesamtgröße des globalen Phonemsets um

40% reduziert. Entsprechend hat sich sf im Mittel von 3 auf 5 vergrößert. pm bleibt aber unter 1, selbst nachdem alle 12 Sprachen involviert sind. Durch die grobere Klasseneinteilung wird demzufolge der Anteil an Modellen, die gemeinsame Daten ausnutzen könnten, nicht erhöht. Aus dem höheren Nutzungsfaktor sf kann somit kaum Gewinn erzielt werden, da nach wie vor etwa die Hälfte aller Modelle nur mit Daten einer Sprache trainiert werden.

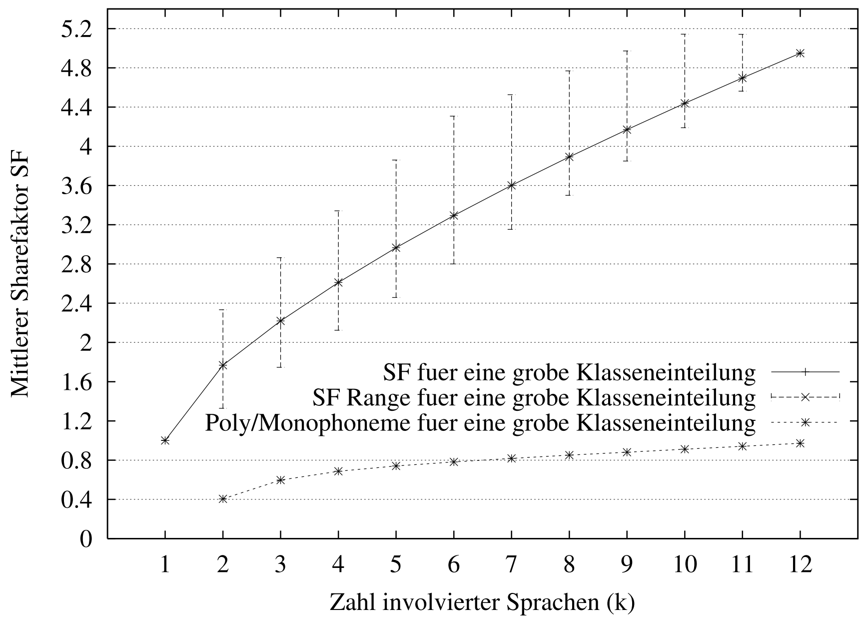


Abbildung 6.2: sf_k und pm_k für $\binom{12}{k}$ Sprachtupel für grobe Klassengranularität

Da bei der groberen Klasseneinteilung die Granularität der Modelle keine optimale monolinguale Modellierung mehr gewährleistet, wird im Hinblick auf eine gute Modellierungsfähigkeit zur multilingualen Anwendung der feineren Klasseneinteilung der Vorzug gegeben.

6.3.2 Multilinguale Kontextmodellierung

Auf der Basis des beschriebenen globalen Phonemsets werden verschiedene Methoden zur Kombination der akustischen Modelle vieler Sprachen untersucht. Bisherige Ansätze in der Literatur blieben auf kontextunabhängige multilinguale Modelle beschränkt. Da für die monolinguale Spracherkennung bewiesen ist, daß die Modellierung breiterer Kontexte einen großen Erkennungsleistungszuwachs zur Folge hat, ist ein Ziel dieser Arbeit, dieses Konzept auf die multilinguale Modellierung zu übertragen. Dazu werden drei verschiedene Methoden zur Kombination multilingualer kontextabhängiger akustischer Modelle eingeführt und anschließend an den resultierenden multilingualen Erkennern evaluiert:

- Die sprachenseparate Kontextmodellierung *ML-SEP*
- Die sprachenvermischte Kontextmodellierung *ML-MIX*
- Die sprachenmarkierte Kontextmodellierung *ML-TAG*

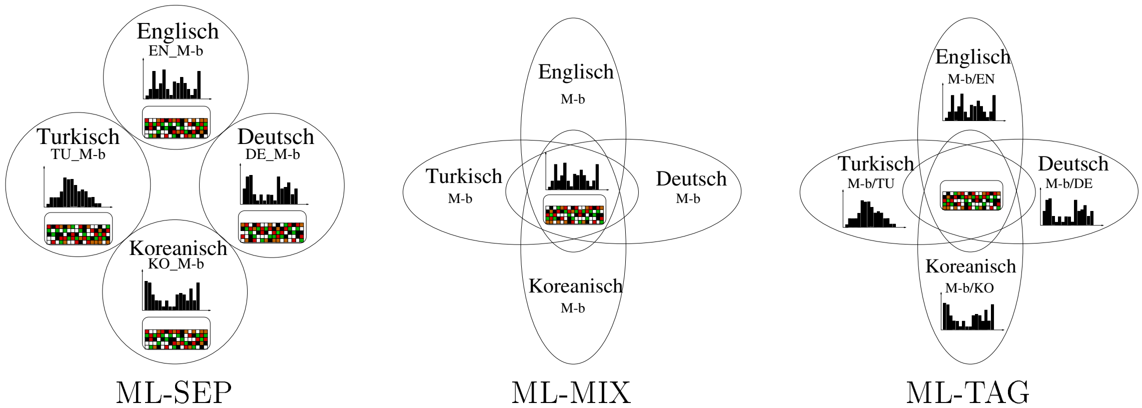


Abbildung 6.3: Methoden zur multilingualen Kontextmodellierung

6.3.2.1 Sprachenseparate Kontextmodellierung ML-SEP

Bei der sprachenseparaten Kontextmodellierung ML-SEP handelt es sich nicht um multilinguale akustische Modelle im eigentlichen Sinn. Es werden nämlich zwei Phoneme verschiedener Sprachen *separat* modelliert, auch dann wenn beide in eine gemeinsame Phonemkategorie fallen. Die multilinguale Komponente bei der Methode ML-SEP ist die Merkmalsextraktion. Denn zur der Berechnung der LDA-Matrix werden die Phonemmodelle aller Sprachen berücksichtigt. Dadurch entsteht ein einziges monolithisches akustisches Modul, das beliebig zwischen den Sprachen hin- und herschalten kann. Gleichzeitig werden die sprachenspezifischen Besonderheiten konserviert. Der Nutzen der sprachenseparaten Modellierung wurde bereits in Kapitel 5 vorgeführt, als mit einem akustischen Modul bestehend aus vier Sprachen die Initialisierung durch QUICKBOOT durchgeführt wurde. Eine weitere Einsatzmöglichkeit ist die Sprachenidentifizierung, wie in Abschnitt 6.4.1 gezeigt werden wird.

Wie bereits in Kapitel 5 erläutert, wird die Wahrscheinlichkeit $p(x|s_i)$, mit der x vom akustischen Modell des Zustands s_i emittiert wird, durch voll kontinuierliche Gaußsche Mischverteilungen berechnet, wobei K_i die Anzahl der im akustischen Modell des Zustands s_i verwendeten Normalverteilungen ist:

$$p(x|s_i) = \sum_{k=1}^{K_i} c_{s_i k} Gau\beta(x|\mu_{s_i k}, \Sigma_{s_i k}) \qquad (6.3)$$

Dann hat die sprachenseparate Kontextmodellierung ML-SEP die Eigenschaft:

$$\text{ML-SEP}: \begin{cases} c_{s_i} \neq c_{s_j} & , \quad \forall i \neq j \\ \mu_{s_i,k} \neq \mu_{s_j,k} , & \forall i \neq j \\ \Sigma_{s_i,k} \neq \Sigma_{s_j,k}, & \forall i \neq j \end{cases}$$

6.3.2.2 Sprachenvermischte Kontextmodellierung ML-MIX

Die wohl naheliegendste Methode, multilinguale akustische Modelle zu realisieren, besteht darin, für jede Phonemkategorie des globalen Phonemsets ein Modell bereitzustellen und dieses Modell mit allen verfügbaren Daten zu trainieren. Sofern die Phonemkategorie zwei Phoneme unterschiedlicher Sprachen enthält, wird das Modell durch einfache Mischung der Daten beider Sprachen trainiert. Diese Form der Modellierung wird daher als ML-MIX bezeichnet. Das Wissen darüber, zu welcher Sprache ein jeweiliges Phonem gehört, wird aufgegeben. Zur kontextabhängigen Modellierung werden die entstandenen multilingualen Subphoneme nach dem Prinzip des entscheidungsbaumbasierten divisiven Ballungsverfahrens auf Basis der Entropie-Distanz (vgl. Abschnitt 3.2.2.3) geballt. Dazu wird ein multilingualer Fragenkatalog erstellt, der die linguistisch motivierten Fragen aller beteiligten Sprachen kombiniert. Beim Ballungsvorgang wird nicht unterschieden, aus welchen Sprachen die Phonemkontexte stammen. Es kann daher vorkommen, daß ein Polyphon in Abhängigkeit der Kontexte aus verschiedenen Sprachen modelliert wird. Der Vorteil der ML-MIX Modellierung liegt darin, daß man neben den Einsparungen durch die multilinguale LDA auch eine Einsparung der Modellparameter μ, Σ hat. Der einfache Ansatz der Datenmischung, der auf den IPA-definierten Phonemkategorien aufsetzt, erlaubt eine sehr einfache Portierung auf neue Sprachen. Zudem wird erwartet, daß durch das Mischen der phonetischen Kontexte verschiedener Sprachen innerhalb eines Polyphons die Robustheit gegenüber den extremen Kontextwechseln bei der Portierung erhöht wird. Diese Mischung phonetischer Kontexte mehrerer Sprachen geschieht allerdings rein heuristisch und beachtet keine statistischen Gegebenheiten oder Ähnlichkeiten zwischen den Kontexten verschiedener Sprachen.

Mit der vorher eingeführten Funktion sameIPA läßt sich die sprachenvermischte Modellierung ML-MIX beschreiben als:

$$\text{ML-MIX}: \begin{cases} c_{s_i} = c_{s_j} & , \quad \forall i,j : \text{sameIPA}(s_i, s_j) = 1 \\ \mu_{s_i,k} = \mu_{s_j,k} , & \forall i,j : \text{sameIPA}(s_i, s_j) = 1 \\ \Sigma_{s_i,k} = \Sigma_{s_j,k}, & \forall i,j : \text{sameIPA}(s_i, s_j) = 1 \end{cases}$$

6.3.2.3 Sprachenmarkierte Kontextmodellierung ML-TAG

Mit der sprachenmarkierten Kontextmodellierung ML-TAG soll die blinde, rein heuristische Mischung von Kontexten unterbunden werden. Dazu wird jedes Phonem mit einer Markierung (*engl.* TAG) versehen, welche die Sprachenzugehörigkeit des

Phonems anzeigt. Die Codebooks aller Phoneme einer Kategorie werden wie beim ML-MIX Verfahren von allen beteiligten Sprachen gemeinsam trainiert, die Mixturverteilungen sind jedoch sprachenspezifisch realisiert. Das Ballungsverfahren, welches bei ML-MIX angewendet wird, wird erweitert, indem Fragen nach der Sprache eines Phonems zum multilingualen Fragenkatalog hinzugefügt werden. Beim Ballungsvorgang entscheiden daher die Daten darüber, ob Fragen nach der Sprache bedeutungsvoller sind als Fragen nach dem phonetischen Kontext. Somit werden Trainingsdaten verschiedener Sprachen nur dann miteinander vermischt, wenn eine hinreichende Ähnlichkeit zwischen den Polyphonemmodellen festgestellt werden kann. Dieser Ansatz realisiert eine datengetriebene multilinguale akustische Modellierung, ohne den Vorteil der Parametereinsparung aufzugeben. Die ML-TAG Methode läßt sich beschreiben als:

$$\text{ML-TAG}: \quad \begin{cases} c_{s_i} \neq c_{s_j} & , \quad \forall i \neq j \\ \mu_{s_i,k} = \mu_{s_j,k} , & \forall i,j : \text{sameIPA}(s_i, s_j) = 1 \\ \Sigma_{s_i,k} = \Sigma_{s_j,k}, & \forall i,j : \text{sameIPA}(s_i, s_j) = 1 \end{cases}$$

Abbildung 6.3 zeigt eine schematische Darstellung der drei erläuterten Modellkombinationsmethoden. Die Mixturgewichte sind in der Abbildung als Histogramm dargestellt und die Codebooks durch eine Anordnung von Merkmalsvektoren in einem gerundeten Rechteck symbolisiert.

6.3.2.4 Analyse des Kontextentscheidungsbaums

Die Analyse der kontextabhängigen multilingualen akustischen Modellierung wird an den fünf Sprachen Japanisch, Koreanisch, Kroatisch, Spanisch und Türkisch durchgeführt. Das Ballungsverfahren beginnt für diese fünf Sprachen mit 650,000 verschiedenen Sub-Quintphonmodellen und wird nach Erreichen von 3000 Sub-Quintphonmodelle beendet. Eine Analyse der beim Ballen verwendeten Fragen soll Aufschluß über die Bedeutung von Sprachenfragen geben, insbesondere interessiert der Vergleich zu sonstigen Kontextfragen. Zu diesem Zweck wird der erzeugte Entscheidungsbaum von ML-TAG herangezogen, der durch den divisiven Ballungsvorgang anhand des multilingualen Fragenkatalogs entstanden ist. Auf der Basis dieses Entscheidungsbaums wird eine entropiebasierte Distanz D_H zwischen den Sprachen berechnet. Dazu wird in jedem Ballungsknoten die Sprachenverteilung der zugehörigen Polyphone ermittelt, wie in Abbildung 6.4 veranschaulicht ist. Die entropiebasierte Distanz D_H der Sprachen ergibt sich aus der Entropie eines Knotens vor seiner Aufspaltung H_{org} und der Entropie der beiden Nachfolgerknoten nach der Aufspaltung H_{yes} und H_{no}:

$$D_H = p(K_{yes}) \cdot H_{yes} + p(K_{no}) \cdot H_{no} - (p(K_{yes}) + p(K_{no})) \cdot H_{org} \qquad (6.4)$$

wobei H_K die Entropie der Verteilung im Knoten $K \in \{org, yes, no\}$ ist

$$H_K = \sum_{i=1}^{l} p_K(i) \cdot \log_2 p_K(i) \text{ und } l = 5 \qquad (6.5)$$

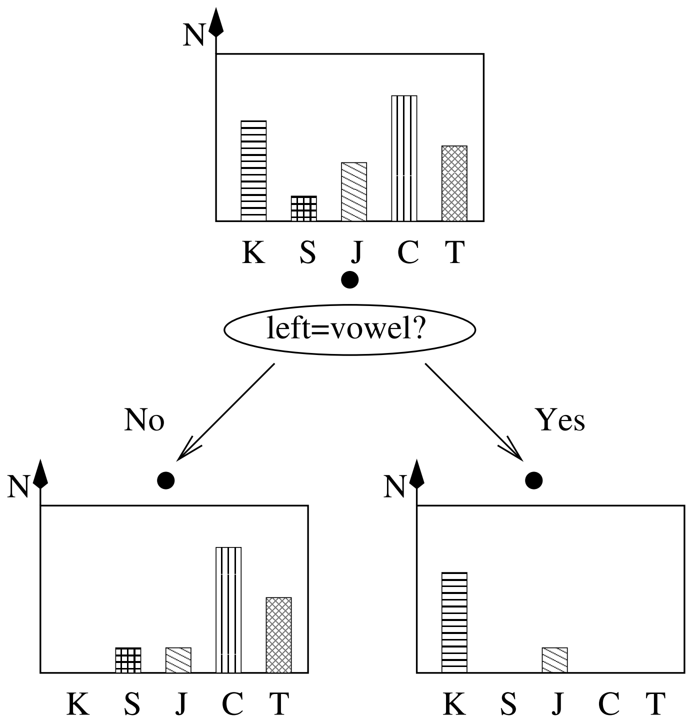


Abbildung 6.4: Ballungsknoten mit Sprachenverteilung

Der Entscheidungsbaum wird nun, am Wurzelknoten beginnend, traversiert und die berechnete Distanz D_H der Sprachen aufsummiert. Abbildung 6.5 zeigt die summierte Distanz aufgetragen über alle Ballungsknoten. Die Kurve *Summe aller Fragen* zeigt den gesamten Entropiegewinn, der durch sämtliche Fragen des Fragenkatalogs erzielt wird, wohingegen die Kurve *Phonetische Kontextfragen* die Distanz aufsummiert, die durch phonetische Kontextfragen entstehen. Der große Abstand zwischen beiden Kurven impliziert, daß ein Großteil der Entropiegewinne auf Fragen nach der Sprachenzugehörigkeit eines Phonems zustandekommen. Insbesondere fällt auf, daß gerade zu Beginn des Aufspaltungsvorgangs Fragen nach der Sprache gestellt werden. Die fünf unteren Kurven zeigen die Beteiligungsraten aufgeteilt nach den einzelnen Sprachen.

Tabelle 6.2 verdeutlicht den Stellenwert, den die Sprachenfragen im Verhältnis zu phonetischen Kontextfragen einnehmen. In dieser Tabelle wird die Häufigkeit, mit der die Sprachfragen während des Ballungsvorgangs gestellt wurden, zusammengefaßt und an vier Abbruchstellen nach jeweils 500, 1000, 1500 und 3000 geballten Polyphonmodellen analysiert.

Bei der Erweiterung des multilingualen Fragenkatalogs sind zu den Fragen nach einer Sprache auch Fragen nach Sprachengruppen hinzugefügt worden. Tabelle 6.2 zeigt, daß die Sprachgruppe Koreanisch+Türkisch (KO+TU) dominiert. Bereits die Analyse von Abbildung 6.5 legt dieses Ergebnis nahe. Offensichtlich unterscheiden sich die koreanischen und türkischen Modelle stark von denen der anderen Sprachen, daher taucht die Frage nach diesen beiden Sprachen insbesondere zu Beginn des Ballungsverfahrens häufig auf. Es folgen die Frage nach Koreanisch und mit einigem Abstand die nach Japanisch und Kroatisch.

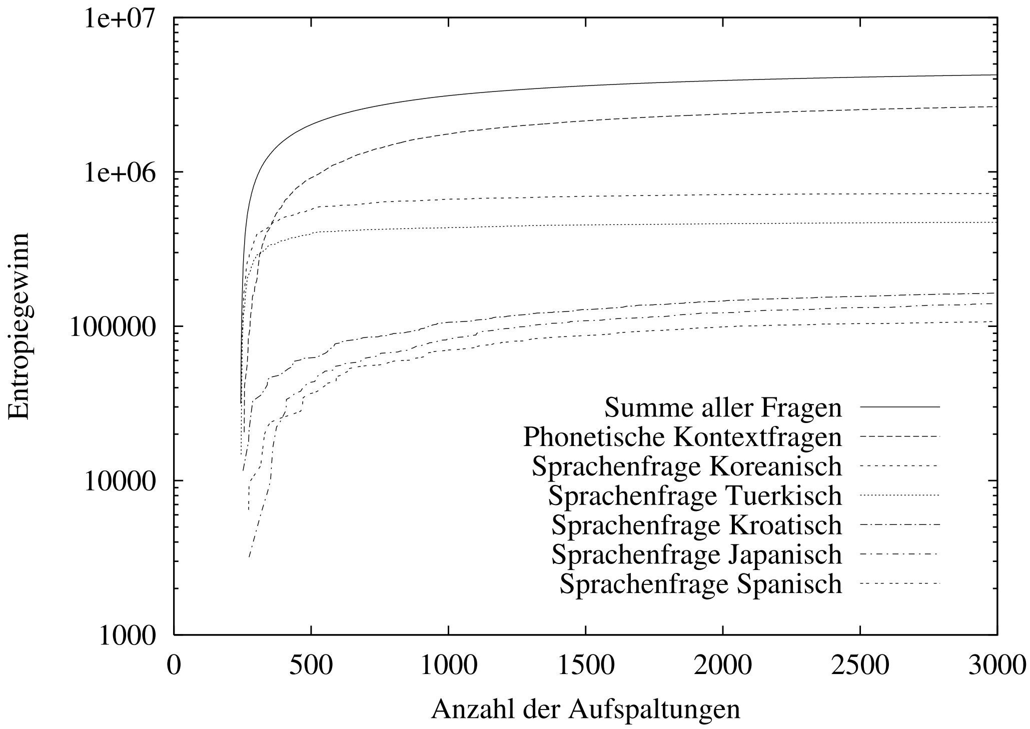


Abbildung 6.5: Analyse der Sprachenfragen

Die Resultate aus Abbildung 6.4 und Tabelle 6.2 zusammengenommen ergeben das Bild, daß Sprachenfragen sehr häufig und besonders zu Beginn des Ballungsprozesses gestellt werden. Es zeigt sich, daß der größte Anteil an Spracheninformation nach etwa 2000 Aufspaltungen ausdifferenziert ist. Ein multilinguales Erkennersystem, das bei 5 Sprachen mehr als 2000 Polyphone modelliert, besteht demnach größtenteils aus monolingualen Modellen. Dieses Resultat ist ein deutlicher Hinweis darauf, daß sich die phonetischen Kontexte zwischen Sprachen signifikant unterscheiden. Für eine gewinnbringende kontextabhängige Modellierung über Sprachen hinweg müssen daher besondere Maßnahmen getroffen werden, die im Kapitel 7 beschrieben werden.

6.3.3 Evaluation der multilingualen Modellierung

Die beschriebenen Methoden zur Kombination akustischer Modelle werden im folgenden evaluiert. Dazu werden die genannten Kriterien Wortfehlerrate, Parameter-

#	500 Modelle	#	1000 Modelle	#	1500 Modelle	#	3000 Modelle
76	KO+TU	92	KO+TU	100	KO+TU	146	word bound
38	KOREAN	54	KOREAN	73	KOREAN	131	back-vow
30	front-vow	48	back-vow	73	back-vow	130	front-vow
27	back-vow	45	front-vow	65	front-vow	128	consonant
23	vowel	38	unvoiced	61	word bound	113	KO+TU
22	unvoiced	37	word bound	53	consonant	98	KOREAN
20	silence	36	vowel	48	unvoiced	97	voiced
19	fric-sibil	32	consonant	48	alveodental	90	vowel
16	word bound	29	silence	46	vowel	88	unvoiced
14	nasal	28	voiced	42	voiced	85	nasal
10	voiced	26	nasal	42	nasal	84	alveodental
10	round	25	frik-sibil	36	silence	79	JAPANESE
10	JAPANESE	24	plos-unvoic	36	plos-unvoic	63	plos-unvoic
10	consonant	23	alveodental	35	frik-sibil	59	frik-sibil
9	plos-unvoic	22	round	32	JAPANESE	59	close-vow
9	open-vow	19	plosive	29	round	56	silence
9	CR+JA+SP	19	JAPANESE	28	plosive	55	round
8	vow-a	16	open-vow	24	CR+SP	54	plosive
8	plosive	16	CR+JA+SP	23	open-vow	47	CROATIAN

Tabelle 6.2: Häufigkeiten von Kontextfragen beim divisiven Ballen

reduktion und Portierbarkeit herangezogen, um die multilingualen Systeme auf ihre Brauchbarkeit hin zu überprüfen und miteinander zu vergleichen.

6.3.3.1 Multilinguale LDA

Zunächst wird im folgenden Experiment der Effekt der sprachenunabhängigen Vorverarbeitung untersucht. Zum Training der multilingualen LDA wurden die Daten der sieben Sprachen Chinesisch, Deutsch, Französisch, Japanisch, Kroatisch, Spanisch und Türkisch verwendet. Gegenüber der jeweiligen monolingualen LDA, in der jeweils nur die sprachenspezifischen Klassen zur Diskriminierung herangezogen werden, müssen in der multilingualen LDA wesentlich mehr Klassen berücksichtigt werden, auch dann, wenn sie in einer bestimmten Sprache nicht enthalten sind. Es kommt dabei zu zwei gegenläufigen Effekten: Einerseits könnte aus der Erhöhung der zu berücksichtigenden Klassen eine Verringerung der Erkennungsleistung resultieren. Andererseits stehen zur Schätzung der LDA-Parameter mehr Sprachdaten zur Verfügung, was auf eine Verbesserung der Erkennungsleistung hoffen läßt.

Die in Tabelle 6.3 dargestellten Erkennungsergebnisse zeigen, daß sich durch die multilinguale LDA nur geringfügige Leistungsverluste gegenüber der monolingualen LDA ergeben. Es kann daher davon ausgegangen werden, daß es durch die sprachenunabhängige Vorverarbeitung zu keinen erheblichen Erkennungsleistungseinbußen kommt. Somit sind die Voraussetzungen für eine sprachenübergreifende einheitliche Vorverarbeitung gegeben. Da die sprachenseparate Modellierung ML-SEP sich

Sprache	Monolinguale LDA	Multilinguale LDA	Δ
Chinesisch	22.8	22.9	0.4%
Deutsch	15.6	16.0	2.5%

Tabelle 6.3: Multilinguale LDA [WE in %]

nur in dieser multilingualen LDA von der monolingualen Modellierung unterscheidet, kann damit davon ausgegangen werden, daß die Ergebnisse der nach der ML-SEP Methode kombinierten Erkenner sich nicht signifikant von denen der monolingualen Erkenner abheben.

6.3.3.2 Vergleich zwischen mono- und multilingualen Modellen

Grundlage der Untersuchung zum Vergleich zwischen der monolingualen und den multilingualen Modellierungen sind mono- und multilinguale Erkenner, die auf den fünf Sprachen Japanisch, Koreanisch, Kroatisch, Spanisch und Türkisch entwickelt wurden. Für die vorgestellten Kombinationsmethoden wurden multilinguale Erkenner nach dem in Abschnitt 5.3.9 beschriebenen Verfahren entwickelt. Alle Erkenner sind in Vorverarbeitung, HMM-Struktur und Trainingsverfahren identisch zu den monolingualen Ausgangssystemen. Die zugrundeliegenden monolingualen Erkenner sind CDHMM Systeme. Jedes monolinguale System wurde auf 1500 Polyphon Modelle geballt, wobei jedes Subpolyphon-Modell 16 Gaußsche Mischverteilungen auf den in Abschnitt 5.3.9.1 beschriebenen Mel-cepstral-Koeffizienten verwendet, die durch eine LDA auf 24 Dimensionen reduziert wurden. Es handelt sich um Systeme, die bereits in einem frühen Stadium entwickelt wurden, daher sind die Erkennungsleistungen geringer als in den finalen Erkennern.

Zusätzlich zu den verschiedenen Kombinationsmethoden wurde mit einer unterschiedlichen Modellanzahl experimentiert. Zur Unterscheidung der verschiedenen Systeme sind sie mit Namen gekennzeichnet, aus denen die Kombinationsmethode sowie die Zahl der Modelle ersichtlich werden. Beispielsweise bezeichnet *ML-TAG7500* das multilinguale System, welches aus der sprachenmarkierten Kombinationsmethode entstand und 7500 Subpolyphone modelliert.

Im folgenden Experiment wird geprüft, wie sich die monolinguale zur multilingualen Modellierung vergleicht. Dazu wurde ein multilingualer Erkenner mittels der Kombinationsmethode ML-TAG gebaut. Dieser Erkenner modelliert 7500 Polyphonmodelle, die Zahl der Systemparameter ist daher bis auf die Einsparung durch die multilinguale LDA identisch mit der Summe der Parameter der fünf monolingualen Systeme. Leistungsunterschiede sind daher nicht in der unterschiedlichen Zahl gelernter Parameter begründet. Der Vergleich der Wortfehlerraten in Tabelle 6.4 zeigt, daß durch die multilinguale Modellierung ML-TAG eine mittlere Leistungseinbuße von 1.66 Prozentpunkten oder 5.8% relative Einbußen ermittelt werden. Diese Einbuße ist aufgrund der Erkenntnisse aus Tabelle 6.3 nicht vollständig durch

Sprache	Monolingual 5 × 1500	ML-TAG7500 1 × 7500	Δ
Kroatisch	26.9	30.2	10.9%
Japanisch	13.0	14.0	7.1%
Koreanisch	47.3	47.7	0.8%
Spanisch	27.6	30.0	8.0%
Türkisch	20.1	21.3	5.6%
Gesamt	26.98	28.64	5.8%

Tabelle 6.4: Vergleich zwischen mono- und multilingualen Modellen [WE in %]

die multilinguale LDA zu erklären, sondern muß durch das Mischen der Daten verschiedener Sprachen entstehen.

Im fünflingualen Fall wurde die Zahl des Phoneminventars von 171 auf 81 reduziert. Der Faktor sf liegt damit bei $\mathrm{sf}_5 = 2.1$. Das Verhältnis pm liegt bei $\mathrm{pm}_5 = \frac{35}{46} = 0.76$. So verdoppelt sich zum Training der Codebooks die Datenausnutzung nahezu. Trotzdem führt die Zunahme der Daten nicht zu einer Leistungsverbesserung gegenüber dem monolingualen Fall. Diese Tatsache wird auf dem Umstand zurückgeführt, daß die Datenmischung in einer generalisierten Modellierung über Sprachen hinweg resultiert, die zu Lasten der Modellgenauigkeit geht. Der Umstand, daß multilinguale Modelle schlechtere Erkennungsraten auf einer der Trainingssprachen im Vergleich zu monolingualen Modellen zur Folge haben, wird auch in anderen Forschungsarbeiten wie etwa [BGM97, CDG+97, Köh98] berichtet.

6.3.3.3 Parameterreduktion

Sprache	ML-TAG7500	ML-TAG3000	Δ
Kroatisch	30.2	31.9	5.3%
Japanisch	14.0	15.0	7.3%
Koreanisch	47.7	49.0	2.6%
Spanisch	30.0	32.4	7.4%
Türkisch	21.3	21.3	0%
Gesamt	28.64	29.92	4.3%

Tabelle 6.5: Reduktion der Parameterzahl für ML-TAG [WE in %]

Mit dem folgenden Experiment wird untersucht, wie sich die Reduktion der Modelle im multilingualen Fall auf die Erkennungsleistung auswirkt. Zu diesem Zweck wurde ein fünflingualer Erkenner ML-TAG3000 entwickelt, der 3000 Subpolyphone statt bisher 7500 Subpolyphone modelliert. Das entspricht einer Reduktion der

Parameterzahl um 40% (mit $3000/(5 \times 1500) = 0.4$). Tabelle 6.5 zeigt die Erkennungsleistungen beider Systeme ML-TAG3000 und ML-TAG7500 im Vergleich.

Der mittlere Leistungsverlust von ML-TAG7500 nach ML-TAG3000 beträgt 1.28 Prozentpunkte oder 4.3% relativ. Wie beim Vergleich der monolingualen zur multilingualen Modellierung unterscheiden sich die Verluste je nach Sprache, was dadurch erklärt werden kann, daß Sprachen, deren Phoneminventare im multilingualen Phonemset besser abgedeckt werden, mit mehr Modellen repräsentiert sind als solche, die viele Monophoneme besitzen. Die mittlere Leistungseinbuße von ML-TAG3000 gegenüber der Leistung der monolingualen Erkenner addiert sich somit auf 2.94 Prozentpunkte oder fast 10% relativ bei einer Einsparung der Parameter um etwa 40%.

6.3.3.4 Vergleich der Kombinationsvarianten

Tabelle 6.6 zeigt die Erkennungsresultate der beiden Methoden ML-MIX und ML-TAG zur multilingualen Kontextmodellierung mit reduzierter Parameterzahl für die simultane Erkennung aller Trainingssprachen im fünflingualen Fall. Der Vergleich der beiden Methoden zeigt, daß die sprachenmarkierte Kontextmodellierung ML-TAG in allen untersuchten Sprachen signifikant bessere Erkennungsleistungen erzielt, als die sprachenvermischte Kontextmodellierung ML-MIX. Im Schnitt ergibt sich eine 5.28 Prozentpunkte bessere Erkennungsleistung für ML-TAG, das sind 15% relative Verbesserung durch die sprachenmarkierte Modellierung. Das Ergebnis läßt darauf schließen, das sich im Hinblick auf die Erkennung von Sprachen auf denen die Modelle trainiert wurden, signifikant bessere Resultate ergeben, wenn man die Information über die Sprachenzugehörigkeit der Modelle bewahrt.

Sprache	ML-TAG3000	ML-MIX3000	Δ
Kroatisch	31.9	35.0	8.8%
Japanisch	15.0	20.0	25%
Koreanisch	49.0	55.0	10.9%
Spanisch	32.4	37.0	12.4%
Türkisch	21.3	29.0	26%
Gesamt	29.92	35.2	15%

Tabelle 6.6: Vergleich der Kombinationsmethoden [WE in %]

6.4 Anwendungen multilingualer akustischer Modelle

In diesem Abschnitt werden zwei wesentliche Anwendungen multilingualer akustischer Modelle beschrieben. Während der Sprachenidentifizierung in der vorliegenden Arbeit als erwünschtes Nebenprodukt nur eine kleine Rolle zukommt, liegt der Fokus auf der Portierung auf neue Sprachen, auf die im Vorgriff auf das nächste Kapitel an dieser Stelle kurz eingegangen werden soll. Dieser Vorgriff dient zur Klärung der Frage, welche der Kombinationsmethoden für die jeweilige Anwendung am besten geeignet sind.

6.4.1 Sprachenidentifizierung

Die Erkennung vieler Sprachen erfordert in diversen Anwendungen die Identifizierung der Eingabesprache. Dies ist Aufgabe der Sprachenidentifizierung (Language IDentification = LID), deren Ziel die möglichst effiziente und korrekte Bestimmung der gesprochenen Sprache ist. Ein typisches Anwendungsbeispiel ist das mehrsprachige Übersetzungssystem Verbmobil. Wie bereits in Abschnitt 3.1 beschrieben wurde, unterscheidet man die nachgeschaltet von der vorgeschalteten LID. Die nachgeschaltete LID verwendet als Entscheidungsgrundlage die Ausgaben von monolingualen Erkennern. Dies hat den Vorteil, daß Wissen auf der Wortebene genutzt werden kann oder zusätzlich geeignete Nachberechnungen auf dem WHG durchgeführt werden können. In früheren Arbeiten der Autorin konnte gezeigt werden, daß dies zu signifikanten Verbesserungen führt [SR95]. Diese Vorgehensweise geht allerdings zu Lasten der Berechnungszeit. Insbesondere wenn viele Eingangssprachen zur Auswahl stehen, muß die Eingabeäußerung von zahlreichen „falschen“ Erkennern dekodiert werden, was hohe Laufzeiten zur Folge hat.

Die vorgeschaltete LID kann dagegen mit einem dedizierten Phonemerkenner erfolgen, der nach der Identifizierung der Sprache die Eingabe auf den entsprechenden monolingualen Erkenner durchschaltet. Insbesondere der Einsatz multilingualer Phonemerkenner ist in diesem Zusammenhang von Interesse. Die Leistungsfähigkeit von Phonemerkennern zur LID wurde bereits mehrfach nachgewiesen [ZB99].

In früheren Experimenten hatte sich gezeigt, daß die sprachenseparaten Phonemmodelle für die Sprachenidentifizierung am besten geeignet sind. Dies läßt sich damit begründen, daß sie sprachenrelevante Informationen besser konservieren, was im Kontext der Identifizierung eine erwünschte Eigenschaft ist. In den folgenden Experimenten wird ein achtlingualer Phonemerkenner mit insgesamt 296 sprachenspezifischen Modellen verwendet. Die Eingabeäußerung wird durch einen frei laufenden Phonemerkenner dekodiert. Die beste hypothetisierte Phonemsequenz wird nachbearbeitet, indem die Sprachenmarkierungen der Phoneme ausgezählt werden. Diejenige Sprache, deren Markierungen am häufigsten in der Hypothese vorkommt, wird als die gesprochene Sprache identifiziert.

Sprache	ohne LM	LM-noSwitch	LM-Switch
Ch-Mandarin	70.3	69.2	70.3
Englisch	100.0	100.0	100.0
Französisch	98.2	100.0	100.0
Deutsch	90.4	76.3	94.7
Japanisch	67.8	69.4	71.1
Kroatisch	52.1	58.9	87.4
Spanisch	77.0	78.9	83.4
Türkisch	62.1	71.0	71.1
Gesamt	77.2	78.0	84.8

Tabelle 6.7: Sprachenidentifizierungsleistung auf 8 GlobalPhone-Sprachen [LID-Rate in %]

Tabelle 6.7 vergleicht die Identifizierungsraten auf kompletten Äußerungen aus acht GlobalPhone-Sprachen. Im ersten Experiment basiert die Identifizierung ausschließlich auf akustischen Entscheidungen, d.h. es wird kein Sprachmodell verwendet (ohne LM). Im zweiten Experiment wird ein Phonemsprachmodell verwendet, das keine Übergänge zwischen Sprachen erlaubt (LM-noSwitch). Im dritten Experiment erlaubt das Sprachmodell explizit den Übergang zwischen Sprachen (LM-Switch). Die beste Identifizierungsrate mit 84.8% ergibt sich für das System, in dem Sprachenübergänge innerhalb einer Äußerung erlaubt sind. Wird in diesem Fall die Dekodierung auf die ersten 3-Sekunden der Äußerung beschränkt, liegen die Resultate bei 81%, was einem relativen Leistungsverlust von 20% entspricht.

Auffallend sind die 100% LID-Raten für Englisch und Französisch. Dies sind die beiden Sprachen, die zwar mit demselben Mikrophon, nicht aber exakt unter den GlobalPhone-Bedingungen gesammelt wurden. Diese Ergebnisse implizieren, daß die Sprachenidentifizierungsleistung durch Kanal- oder sonstige Charakteristika überlagert wird. In früheren Arbeiten der Autorin [ST95] war diese Beobachtung bereits gemacht worden und auch im Rahmen der Verbmobil Anwendung war dies als ein Problem der Sprachenidentifizierung aufgefallen. Gemeinsam mit Kollegen wurde am ILKD jüngst ein konfidenzbasiertes Verfahren entwickelt, das eine weitgehend kanalunabhängige Sprachenidentifizierung ermöglicht [MKS$^+$00].

6.4.2 Erkennung neuer Sprachen: Vorexperimente

Eine wesentliche Anwendung der multilingualen akustischen Modellkombination ist die Nutzung für die Portierung auf neue, im Training nicht präsentierte Sprachen. Im Vorgriff auf das nächste Kapitel zeigen die folgenden Experimente den Nutzen der multilingualen Kontextmodellierung für die Dekodierung einer Sprache auf, die im Training nicht präsentiert wurde. In diesen Vorexperimenten werden die beschriebenen mono- und fünflingualen Erkenner dazu verwendet, deutsch gesprochene Äuße-

Sprache	Wortfehlerrate
Baseline-Erkenner Deutsch	15.8
Monolinguale Erkenner	
Japanisch	78.1
Koreanisch	90.2
Kroatisch	76.2
Spanisch	76.7
Türkisch	74.5
Multilinguale Erkenner	
ML-MIX3000	63.0
ML-TAG3000	69.4
ML-TAG7500	69.1

Tabelle 6.8: Vergleich zwischen mono- und multilingualen Modellen zur Erkennung deutscher Äußerungen [WE in %]

rungen zu erkennen. Im nächsten Kapitel wird der Nutzen multilingualer Modelle für Sprachen demonstriert, die bisher nicht gut erforscht sind und daher nur wenige Quellen bekannt sind, so daß eine Portierung auf diese Sprachen interessant ist.

6.4.2.1 Vergleich zwischen mono- und multilingualen Modellen

Für die folgenden Experimente wurde die Erkennungsaufgabe erleichtert, indem das zu erkennende deutsche Vokabular auf 1625 Wörter beschränkt wurde. Das Vokabular enthält alle Wörter der Testäußerungen. Der deutsche Basiserkenner erreicht auf diesem Testset mit einem 60K-Erkennervokabular unter realen Bedingungen eine Fehlerrate von 15.8%. Tabelle 6.8 vergleicht die Erkennungsleistungen der fünf monolingualen Erkenner mit den Leistungen der multilingualen Erkennern ML-TAG3000 und ML-MIX3000 auf deutschen Daten ohne deutsches Material zum Training präsentiert zu haben.

Die Ergebnisse zeigen, daß die drei multilingualen Erkenner deutlich besser abschneiden als die monolingualen Erkenner. Auffallend ist die große Schwankungsbreite bei den monolingualen Erkennern. Der multilinguale Erkenner ML-MIX ist deutlich besser als der ML-TAG Erkenner. Auch nachdem die Parameterzahl des ML-TAG Erkenners mehr als verdoppelt wurde, ist ML-MIX signifikant besser. Diese Resultate stützen die These, daß die Mischung der Kontexte über Sprachen hinweg die Robustheit der Modelle beim Wechsel auf eine unbekannte Sprache verbessert. Während in den vergangenen Experimenten zur Erkennung bekannter Sprachen die sprachenmarkierte Kontextmodellierung eindeutig bessere Erkennungsleistungen gegenüber der sprachenvermischten Kontextmodellierung zeigte, kehrt sich dieser Sachverhalt bei der Erkennung unbekannter Sprachen um.

6.4.2.2 Multilinguales Aussprachewörterbuch

Zur Erkennung der deutschen Äußerung wird davon ausgegangen, daß bereits ein Aussprachewörterbuch vorliegt. Allerdings wird zur Beschreibung der Aussprache von deutschen Wörtern ein deutsches Phoneminventar verwendet. Zur Dekodierung der deutschen Äußerungen durch einen anderssprachigen Erkenner muß daher der deutsche Phonemsatz des Aussprachewörterbuchs auf den Phonemsatz des dekodierenden Erkenners möglichst passend abgebildet werden.

Sofern keine Trainingsdaten in der zu erkennenden Zielsprache vorhanden sind, kann diese Abbildung nur durch Heuristiken durchgeführt werden. Für die Abbildung eines monolingualen Phonemsatzes auf den deutschen Phonemsatz wurde das IPA-Referenzschema verwendet. Jedes deutsche Phonem wurden dabei durch den jeweils nähsten IPA-Nachbarn der Quellsprache ersetzt.

Bei der Abbildung des multilingualen Phoneminventars auf das deutsche Inventar besteht die Möglichkeit, ein deutsches Phonem durch bis zu fünf IPA-Gegenstücke zu ersetzen, eines von jeder Quellsprache. Zur Abbildung des multilingualen Phoneminventars werden daher zwei Methoden eingeführt: Die parallele (Dict-5L) und die sprachenübergreifende (Dict-ML) Aussprachenmodellierung.

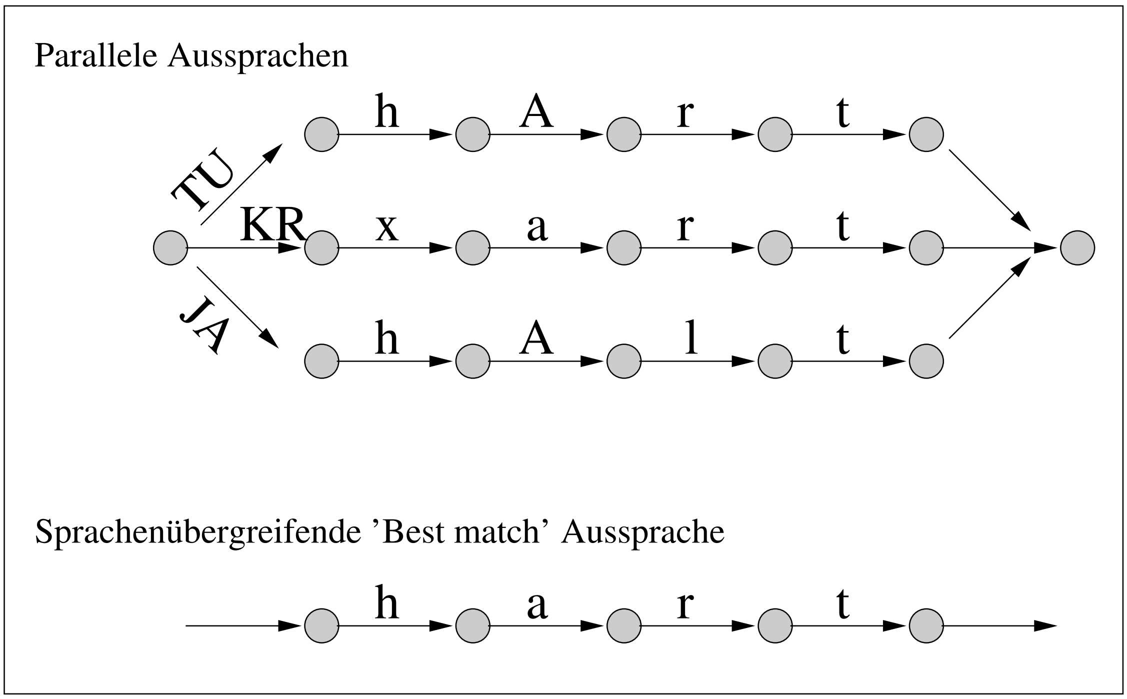


Abbildung 6.6: Parallel und sprachenübergreifende Modellierung von Aussprachen

Die Abbildung 6.6 veranschaulicht beide Methoden und zeigt deren Unterschiede am Beispiel des deutschen Wortes hart. Die Ausspracherepräsentation von hart im

Deutschen ist [h a r t]. Würde man diese Aussprache durch ein türkisches Phonemset ausdrücken, dann wäre die Repräsentation [h ɑ r t], weil es im Türkischen nur das hintere „a“ (ɑ=A) gibt. Im Japanischen gibt es ebenfalls nur das hintere [ɑ], zusätzlich sind r und l Allophone, meist liegt die Aussprache dazwischen. Im vorliegenden japanischen Phonemset ist nur das [l] modelliert, daher ist die Repräsentation von hart im japanischen Phonemset [h ɑ l t]. Im Kroatischen gibt es zwar ein [a] und ein [r] aber kein [h], das durch den nähsten IPA-Nachbarn [x] ersetzt wird.

In der parallelen Aussprachemodellierung wird jede sprachenspezifische Abbildung als Aussprachevariante in das Aussprachewörterbuch eingefügt. Im fünflingualen Erkenner wird also jedes deutsche Wort durch maximal 5 Aussprachevarianten repräsentiert (Dict-5L). In der sprachenübergreifenden Aussprachemodellierung wird jedes deutsche Wort durch genau eine Aussprache repräsentiert, in der jeweils das im IPA-Sinn am besten passende Phonem gesetzt wird (Dict-ML). Im Beispiel wird das deutsche Wort hart durch die tatsächliche Phonemsequenz [h a r t] repräsentiert.

Die Vorteile von Dict-5L liegen in der größeren Flexibilität der akustischen Modellierung. Durch die Modellierung als Aussprachevarianten entscheidet der Erkenner während der Dekodierung, welche der möglichen Varianten am besten zum akustischen Signal paßt. Ein weiterer Vorteil liegt darin, daß die phonotaktischen Gesetzmäßigkeiten nicht verletzt werden. Es können nur Kontexte von Phonemen einer Sprache beobachtet werden. Der Nachteil liegt in der Aufblähung des Aussprachewörterbuches um den Faktor der beteiligten Sprachen. Daher wächst mit steigender Anzahl beteiligter Sprachen die Verwechselbarkeit des Vokabulars an.

Die Methode der sprachenübergreifenden Aussprachemodellierung hält dagegen den Umfang des Aussprachewörterbuches konstant auf der ursprünglichen Größe. Allerdings werden die phonotaktischen Regeln, die beim Training gelernt wurden, verletzt, weil Phonemübergänge entstehen können, die in keiner der Trainingssprachen vorkommen. Dadurch kann es zu einer sehr ungeschickten Modellwahl im Baum kommen.

Sprache	Wörterbuch	Wortfehlerrate	Δ
ML-MIX3000	Dict-ML	66.7	5.5%
ML-MIX3000	Dict-5L	63.0	

Tabelle 6.9: Vergleich der Ausspracheabbildungen [WE in %]

Tabelle 6.9 zeigt die Resultate des Vergleichs der beiden Aussprachemodellierungen. Die Ergebnisse zeigen, daß sich die Flexibilität der Aussprachen in 3.7 Prozentpunkten oder 5.5% Fehlerreduktion niederschlägt.

Sprache	Wortfehlerrate ohne Training	Wortfehlerrate mit Training
Baseline-Erkenner Deutsch	15.8	
Monolinguale Erkenner		
Japanisch	78.1	50.5
Koreanisch	90.2	42.4
Kroatisch	76.2	31.3
Spanisch	76.7	31.9
Türkisch	74.5	28.4
Multilinguale Erkenner		
ML-MIX3000	63.0	27.1
ML-TAG3000	69.4	35.7
ML-TAG7500	69.1	35.4

Tabelle 6.10: Vergleich zwischen mono- und multilingualen Modellen nach dem Training auf 1000 Äußerungen [WE in %]

6.4.2.3 Training mit limitiertem Datenmaterial

Im nächsten Experiment wurde die Frage untersucht, wie sich die Ergebnisse der monolingualen und multilingualen Erkenner durch das Training auf einer begrenzten Menge deutscher Sprachdaten verändern. Dazu wurden die akustischen Modelle durch zwei Iterationen Viterbi-Training auf 1000 deutschen Sätzen von 13 Sprechern auf der deutschen Sprache trainiert.

Die Ergebnisse aus Tabelle 6.10 zeigen für die monolingualen Erkenner eine enorme Spannweite von Verbesserungen. Während sich der türkische Erkenner auf 28.4% Fehlerrate verbessert, bleibt der japanischen Erkenner mit 50.5% deutlich zurück. Eine mögliche Erklärung ist die eingeschränkte Phonotaktik der japanischen Sprache, die keine Konsonantencluster kennt. Die deutsche Sprache ist aber für ihre große Zahl von Konsonantenclustern bekannt. Abgesehen von Japanisch läßt sich eine gewisse Korrelation zwischen Fehlerrate und Größe des Phoneminventars einer Sprache erkennen. Die große Diskrepanz zwischen den Fehlerraten der monolingualen Erkenner läßt darauf schließen, daß es beim Übergang auf eine unbekannte Sprache zu erheblichen Problemen durch unpassende Kontextmodelle kommt. Solche Fehler lassen sich nur durch eine Veränderung der Kontextentscheidungsbäume, nicht jedoch durch Anpassung der bestehenden Kontextmodelle beheben.

Keiner der monolingualen Erkenner übertrifft den multilingualen Erkenner ML-MIX. Dieses Ergebnis bestätigt die Hypothese, daß sprachenvermischte Modelle robuster gegen einen Sprachenwechsel sind, als sprachenspezifische. Die Ergebnisse aus Tabelle 6.8, die bereits nahegelegt hatten, daß die sprachenvermischte Kontextmodellierung zur Portierung auf neue Sprachen besser geeignet sei, werden hier erneut bestätigt.

Sprache	Wörterbuch	Wortfehlerrate	
		ohne Training	mit Training
ML-MIX3000	Dict-ML	66.7	27.1
ML-MIX3000	Dict-5L	63.0	29.2

Tabelle 6.11: Vergleich der Ausspracheabbildungen mit 1000 Äußerungen [WE in %]

Die monolingualen Erkenner erreichen im Schnitt eine Fehlerrate von 36.9%. Damit liegt der multilinguale Erkenner ML-TAG zwar immer noch leicht besser als die mittlere Fehlerrate der monolingualen Erkenner, aber deutlich unter dem besten monolingualen, dem türkischen Erkenner.

Wie der Vergleich der Tabelle 6.9 mit der Tabelle 6.11 zeigt, kehrt sich der positive Effekt der parallelen Aussprachemodellierung nach dem Training um. Ein Grund könnte sein, daß sich die ohnehin limitierten Trainingsdaten nun auf mehr Modelle verteilen, so daß daraus eine weniger gute Anpassung der einzelnen Modelle resultiert.

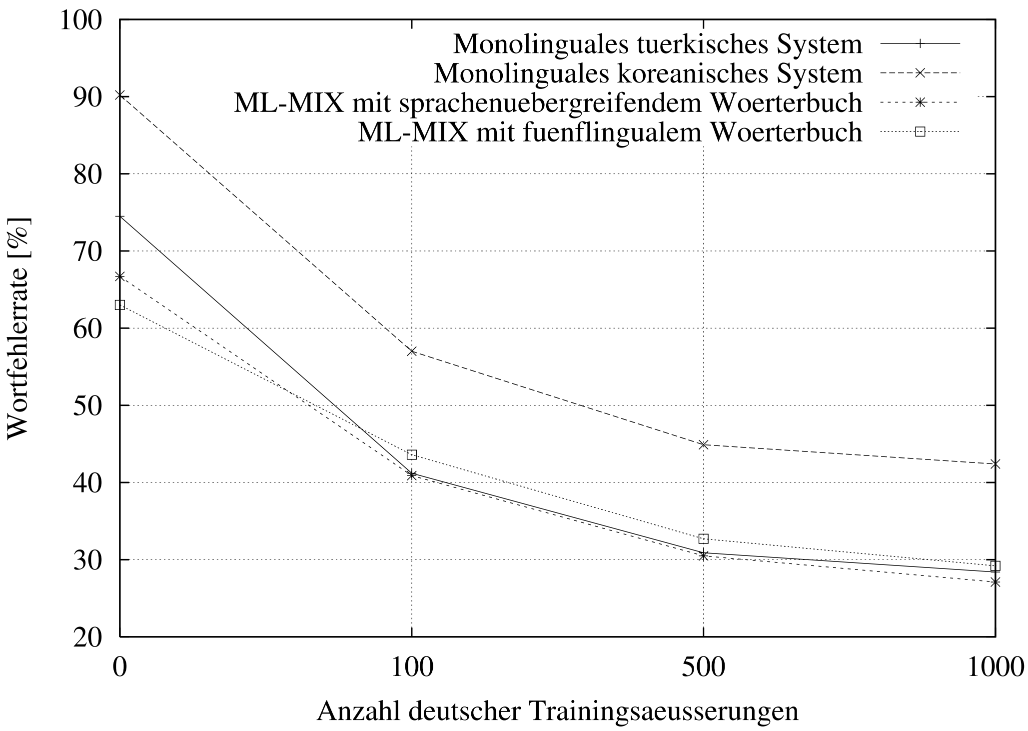


Abbildung 6.7: Effekt verschiedener Trainingsmenge zur Portierung auf Deutsch

Abschließend soll die Leistungsentwicklung in Abhängigkeit der zur Verfügung stehenden Trainingsdaten gemessen werden. Dazu wird auf einer verschieden großen

Zahl deutscher Äußerungen trainiert. Es werden die Leistungen auf Modellen verglichen, die auf keinen Äußerungen, 100 Äußerungen (1500 gesprochene Wörter), 500 Äußerungen (8000 gesprochene Wörter) und 1000 Äußerungen (14000 gesprochene Wörter) gesprochen von einer konstanten Zahl von 13 deutschen Sprechern, trainiert wurden. Die Abbildung 6.7 zeigt die Resultate.

Verglichen werden das zur Portierung auf Deutsch am besten geeignete monolinguale türkische System und das am wenigsten geeignete monolinguale koreanische System mit dem multilingualen ML-MIX System einmal mit paralleler und einmal mit sprachenübergreifender Aussprachemodellierung. Die Ergebnisse zeigen, daß das multilinguale System ML-MIX die Leistungen der monolingualen Systeme übertrifft, sofern man beim Training ein sprachenübergreifendes Aussprachewörterbuch verwendet. Der Verlauf der Leistungsverbesserungen in Abhängigkeit des Trainingsmaterials entspricht den Erwartungen. Es gibt keine signifikanten Unterschiede im Entwicklungsverlauf zwischen mono- und multilingualen Systemen.

6.5 Zusammenfassung

Die in diesem Kapitel beschriebenen Arbeiten konzentrieren sich auf die Entwicklung von Methoden zur multilingualen akustischen Modellierung unter besonderer Berücksichtigung der kontextabhängigen Modellierung. Die Motivation der multilingualen akustischen Modellierung liegt in der sprachenübergreifenden Nutzung von Daten und in der Möglichkeit zur Reduktion der insgesamt zu modellierenden Parameter. Um die Nutzung von Daten vieler Sprachen zu ermöglichen, wurde ein **globales Phonemset auf 12 Sprachen** auf dem IPA-Referenzschema entwickelt. Für 12 Sprachen ergibt sich ein mittlerer Nutzungsfaktor von 3, d.h. im Schnitt wird ein Phonem von drei verschiedenen Sprachen verwendet und kann somit potentiell mit den Daten von drei Sprachen trainiert werden.

Auf Basis des globalen Phonemsets wurden drei Kombinationsmethoden implementiert und evaluiert, die es ermöglichen, akustische Modelle über viele Sprachen zu mischen und dabei kontextabhängig zu modellieren. In der **sprachenseparaten Kontextmodellierung** bleiben die akustischen Modelle der einzelnen Sprachen getrennt. Der multilinguale Aspekt dieser Kombinationsmethode liegt in der Vorverarbeitung, in der zur Berechnung der LDA die Phoneme aller Sprachen als zu diskrimierende Klassen herangezogen werden, und damit eine multilinguale LDA realisieren. Die Ergebnisse zeigen, daß es im Vergleich zur monolingualen LDA durch die multilinguale LDA nur zu geringfügigen 0.4% bis 2.5%-igen relativen Leistungseinbußen kommt und somit eine sprachenübergreifende Vorverarbeitung gerechtfertigt ist.

In der **sprachenvermischten Kontextmodellierung** entstehen multilinguale akustische Modelle dadurch, daß alle Polyphoneme durch die Mischung der Sprachdaten trainiert werden. Das Wissen darüber, zu welchen Sprachen die jeweiligen

Phoneme gehören, wird aufgegeben. Zur Erzeugung kontextabhängiger Modelle wird beim Ballen der Subpolyphone nicht unterschieden, aus welchen Sprachen die Phonemkontexte stammen. In der **sprachenmarkierten Modellierung** dagegen bleibt die Information über die Sprachenzugehörigkeit eines Phonems erhalten. Beim Ballungsvorgang entscheiden die Daten darüber, ob Fragen nach der Sprachenzugehörigkeit bedeutungsvoller sind als Fragen nach dem phonetischen Kontext. Die Analyse des entstandenen Kontextentscheidungsbaumes ergab, daß der größte Anteil an Spracheninformationen nach etwa 2000 Aufspaltungen ausdifferenziert ist. Diese Ergebnisse sind ein deutlicher Hinweis darauf, daß sich die phonetischen Kontexte zwischen Sprachen unterscheiden.

Die multilingualen akustischen Modelle wurden anhand der Kriterien Wortfehlerrate und Parametereinsparung mit den monolingualen Modellen verglichen. Der Vergleich ergibt, daß die sprachenmarkierte Kombinationsmethode 15% bessere Erkennungsleistungen zeigt als die sprachenvermischte Methode, sofern man sie auf die Erkennung der Sprachen verwendet, die im Training präsentiert wurden. Diese Verbesserung wird auf den Erhalt der Spracheninformation in der sprachenmarkierten Kontextmodellierung zurückgeführt. Bei der Evaluation der Parameteranzahl zeigt sich, daß eine Reduktion um 40% auf der sprachenmarkierten Modellierung eine relative Einbuße von 4.3% zur Folge hat.

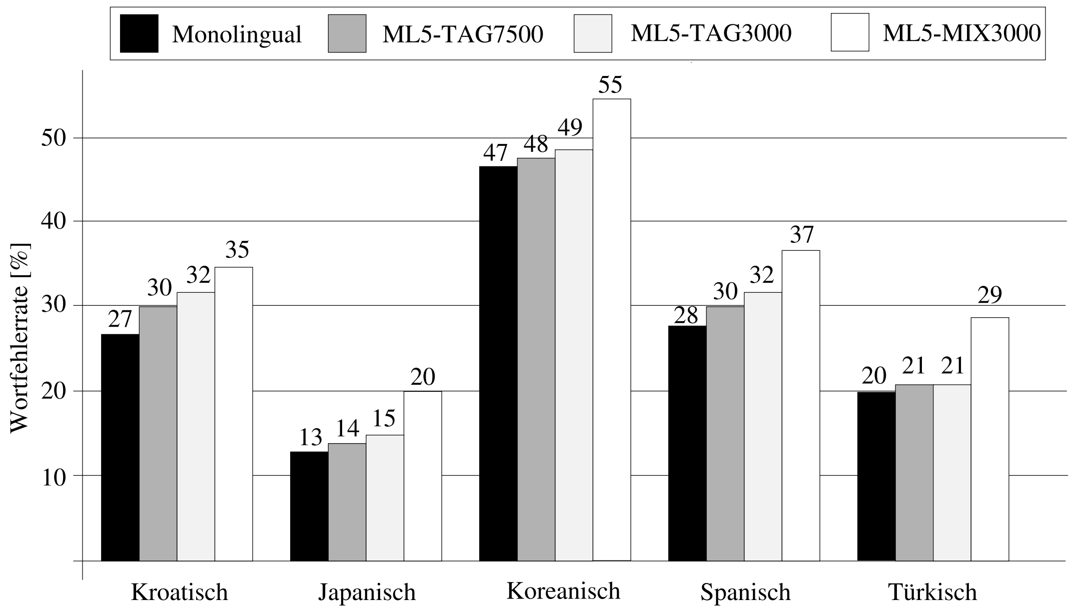


Abbildung 6.8: Vergleich der Kombinationsmethoden [WE in %]

Im Vergleich zu den monolingualen Modellen schneiden die multilingualen Modelle schlechter ab, sofern man sie zur Erkennung der beim Training präsentierten Sprachen heranzieht. Die vorliegenden Experimente ergaben eine relative Leistungseinbuße von 5.8% für die sprachenmarkierte Modellierung. Ein solches Resultat überrascht nicht, da bisherige Erfahrung beispielsweise mit sprecherabhängigen oder kanalabhängigen Erkennungssystemen zeigen, daß spezifische akustische Modelle stets bessere Leistungen erzielen als stärker generalisierende Modelle. Abbildung 6.8 stellt die erzielten Resultate zum Vergleich nochmals einander gegenüber.

Die Erkenntnis, daß multilinguale Modelle den monolingualen Modellen unterlegen sind, gilt unter der Voraussetzung, daß zum Training der monolingualen Modelle umfangreiches Datenmaterial zur Verfügung steht. Diese Voraussetzung ist aber häufig nicht gegeben, entweder weil eine Datensammlung zu aufwendig und deshalb unerwünscht ist oder weil sie ganz unmöglich ist. Liegen keine oder nur limitierte Daten vor, konnte in den Experimente gezeigt werden, daß die multilingualen Modelle, insbesondere die sprachenvermischten Modelle, den monolingualen Modellen überlegen sind. Bei der Anwendung der multilingualen Modelle auf eine neue Sprachen ergaben sich 15.4% bessere Erkennungsleistungen als mit dem besten monolingualen System.

Insgesamt konnte gezeigt werden, daß die multilingualen Modelle einen hohen praktischen Nutzen haben: Sie können zur **Sprachenidentifizierung** eingesetzt werden, in der sich die sprachenseparate Modellierung als vorteilhaft erweist, weil diese Modellierung die sprachenspezifischen Besonderheiten einer Sprache konservieren oder zur **Portierung auf neue Sprachen**, in der die sprachenvermischte Kontextmodellierung die besten Leistungen zeigte und zuletzt zur **simultanen Erkennung mehrerer Sprachen**, wobei der Erhalt der Spracheninformation in der sprachenmarkierten Kontextmodellierung sinnvoll ist. Diese Resultate legen nahe, die Entscheidung für eine der multilingualen Kontextmodellierungsmethoden davon abhängig zu machen, für welchen Zweck die entstehenden Modelle eingesetzt werden sollen.

Diese Ergebnisse und die Resultate aus der Analyse des Kontextentscheidungsbaums lassen den Schluß zu, daß die Erkennungsleistungen bei der Portierung auf neue Sprachen durch eine adäquate Kontextmodellierung weiter verbessert werden können. Die Analyse und ein Vorschlag zur Lösung dieses Problems wird im folgenden Kapitel beschrieben.

Kapitel 7

Portierung auf neue Sprachen

Der Begriff Portierung bezieht sich auf die Übertragung von Wissen, das auf vielen Sprachen gelernt wurde, auf neue, im Training nicht präsentierte Sprachen. Je nachdem, wieviel Datenmaterial in der neuen Sprache verfügbar ist, werden die Techniken Überkreuzsprachlicher Transfer, Bootstrapping und Adaption unterschieden. In diesem Kapitel wird zunächst der Stand der Forschung skizziert. Anschließend werden eigene Lösungen zur Adaption von Kontextmodellen eingeführt, deren Leistungsfähigkeit im experimentellen Abschnitt evaluiert wird. Zur Portierung werden mit Schwedisch und Portugiesisch zwei Sprachen ausgewählt, die bisher wenig erforscht und beachtet sind.

7.1 Ziele und Kriterien

Der Begriff *Portierung* bezieht sich auf den Vorgang der Übertragung eines bestehenden Spracherkenners auf eine neue Sprache. Die Forschung beschäftigt sich mit der Frage, ob sich Informationen aus vielen Sprachen auf eine neue noch nicht präsentierte Sprache transferieren lassen, und wie sich das auf den bisherigen Sprachen erworbene Wissen gewinnbringend einsetzen läßt.

Es sei an dieser Stelle nocheinmal ausdrücklich betont, daß die Untersuchungen zur Portierung eines Spracherkenners in der vorliegenden Arbeit auf die Übertragung der Lautinventare bzw. deren akustischen Modelle konzentrieren wird. Es ist offensichtlich, daß bei der Portierung eines Systems nicht nur die akustischen Modelle, sondern alle sprachenspezifischen Wissensquellen wie Vokabular, Aussprachewörterbuch und Sprachmodelle portiert werden müssen. Hier wird allerdings davon ausgegangen, daß Aussprachewörterbücher und ausreichende Textdaten zur Sprachmodellierung vorhanden sind. Wie bereits in Kapitel 5 beschrieben, wurde die GlobalPhone-Domäne unter dem Gesichtspunkt ausgewählt, daß ausreichendes Textmaterial verfügbar ist. Für die Aussprachewörterbücher wurde in Kapitel 5 für viele Sprachen die automatische Generierung von Aussprachen durch Graphem-zu-Phonem Regeln vorgestellt.

Da der Hauptzeit- und -kostenfaktor bei der Neuentwicklung eines Spracherkenners in einer neuen Sprache die verschrifteten Sprachdaten zum Training der akustischen Modelle sind, verspricht die Aussicht auf die Einschränkung der Sprachdaten den größten Nutzen.

Der Begriff Portierung wird in dieser Arbeit als Überbegriff für drei Kategorien von Übertragung verwendet, die sich nach der Menge der in der neuen Sprachen verfügbaren Daten unterscheiden:

- Bootstrapping: es sind große Datenmengen in der Zielsprache vorhanden
- Adaption: es steht nur sehr limitiertes Datenmaterial zur Verfügung
- Überkreuzsprachlicher Transfer: es existieren überhaupt keine Daten

Entsprechend unterschiedlich sind die mit der Forschungsrichtung verbundenen Probleme. Beim Bootstrapping steht die Frage nach der bestmöglichen Modellinitialisierung im Mittelpunkt. Lösungen zu diesem Problem wurden bereits in Abschnitt 5.3.8 und Abschnitt 6.4 behandelt. Die Forschung im Bereich Adaption konzentriert sich auf die Fragen, wieviele Adaptionsdaten benötigt werden und welche Modelle sich als Ausgangsbasis am besten eignen. Forschungsthemen in diesem Bereich überkreuzsprachlicher Transfer sind, ob die Übertragung zwischen zwei nahverwandten Sprachen besser funktioniert als zwischen nicht verwandten, und ob die Anzahl an Sprachen, mit denen die ursprünglichen akustischen Modelle trainiert sind mit der Leistung in der Zielsprache korrelieren.

7.2 Stand der Forschung

Das Forschungsthema Portierung ist genauso wie die Forschung zur multilingualen Spracherkennung ein sehr aktuelles Forschungsthema, in dem noch keine einheitlichen Begriffe geprägt wurden. Ähnlich wie beim Thema der multilingualen Spracherkennung soll in diesem Kapitel versucht werden, die wesentlichen Begriffe zu definieren und geleistete Forschungsarbeiten zu strukturieren. Dabei wird entsprechend der definierten Einteilung vorgegangen. Es werden dazu auch solche Arbeiten genannt, die auf Vorarbeiten der Autorin aufbauen oder zeitlich parallel stattfanden.

7.2.1 Bootstrapping

Die Ziele des Bootstrappings liegen darin, möglichst effizient einen Erkenner in einer neuen Sprache zu erstellen. Da ausreichend Material in der neuen Sprache vorhanden ist, reduziert sich das Problem auf die Frage nach der besten Initialisierung, d.h. der Berechnung initialer Zeitzuordnungen. Die Modelle, die zur Initialisierung eingesetzt werden, bezeichnet man als *Keim-Modelle (engl. seed models)*. Diese Keim-Modelle fungieren nur als Startbasis, denn nach der Initialisierung wird das System mit viel Daten der Zielsprache trainiert, das Wissen der Keim-Sprachen geht verloren [Zue93], [WKAM94], [CCM98], [OAM+92].

Wheatley et al. analysierte in [WKAM94] ausführlich die Möglichkeiten des Bootstrapping. Die Idee, akustische Modelle der einen Sprache auf eine andere Sprache anzuwenden, schreibt sie Zue zu, der dies in [Zue93] erprobt, allerdings noch nicht systematisch evaluiert hat. Tatsächlich haben wohl viele Arbeitsgruppen parallel diese Technik angewendet, als der Wunsch nach Spracherkennern in vielen Sprachen aufkam.

Am ILKD wurde dieser Ansatz in [OAM+92] bereits mit Erfolg dazu eingesetzt, einen deutschen Erkenner von englischen Modellen zu bootstrappen. In eigenen Arbeiten konnte die Portierung eines deutschen Erkenners auf Japanisch gezeigt werden [SKW97].

In der Untersuchung von [WKAM94] werden zur Initialisierung eines japanischen Erkenners drei Methoden erprobt: die Initialisierung von handerzeugten phonetischen Transkripten, die Initialisierung durch ein „Flat-Start“ Modell (Silence) und die Initialisierung durch englische Keim-Modelle, die heuristisch auf japanische Pendant abgebildet worden waren. Die Flat-Start Modelle benötigten viel mehr Trainingsiterationen, bis sie zu vernünftigen Ergebnissen führten, was durch die in Abschnitt 5.3.8 vorgestellten Ergebnisse bestätigt wird. Die englischen Modelle zeigten signifikant bessere Ergebnisse als die phonetischen Transkripte. Dieses Ergebnis war besonders erfreulich, da der Aufwand der Transkriptionen weit größer ist, als die Nutzung von Daten aus anderen Sprachen. Der Versuch mit geringfügig unterschiedlichen Abbildungen zwischen Quell- und Zielphonem zeigte keine signifikanten Unterschiede. In einem weiteren Versuch wurde die trainierten japanischen Modelle mit den ursprünglichen Keim-Modellen geglättet, in der Hoffnung, daß die sprecherunabhängigen Modelleigenschaften der englischen Modelle auf die japanischen Modelle abfärben. Die Resultate zeigten kleine aber unbedeutende Verbesserungen.

In der Arbeitsgruppe von Zue [GFG+95, ZSP+96] wird ebenfalls vom erfolgreichen Einsatz englischer Keim-Modelle berichtet. Sie initialisierten damit einen italienischen und einen japanischen Erkenner, veröffentlichten allerdings keine konkreten Ergebnisse. Ihrer Erfahrung nach ist es einem bilingualen Entwickler mit der Bootstrapping-Technik möglich, eine Spracherkennungssystem innerhalb von sechs Monaten von einer in die andere Sprache zu transferieren.

Köhler vergleicht in [Köh98] wie Wheatley verschiedene Keim-Modelle zur Initialisierung miteinander. Eine Form der Initialisierung besteht in der Verwendung von phonetisch transkribiertem Material wie bei [WKAM94], in der zweiten Form verwendet er sechslinguale Modelle zur Erzeugung initialer Transkripte und drittens verwendet er die Maximum Aposteriori (MAP) Adaptionstechnik, um diese multilingualen Modelle auf die neue Sprache zu adaptieren. Sind ausreichend Daten vorhanden, ergibt sich, daß die Methode mit phonetisch transkribiertem Material zwar am zeitintensivsten sind, aber am besten funktioniert, die multilingualen Modelle und die adaptierten sind gleich gut. Mit abnehmender Zahl an Datenmaterial zeigt die MAP-Adaption zunehmend Vorteile (siehe unten).

7.2.2 Adaption

Bei der Sprachenadaptionstechnik wird ein Erkenner auf die neue Zielsprache angepaßt, indem sehr limitiertes Trainingsmaterial zum Adaptieren der akustischem Modelle zur Verfügung gestellt wird. Die Forschung konzentriert sich dabei auf zwei Fragen: wieviele Adaptionsdaten werden benötigt, um vernünftige Resultate zu erzielen, und welche akustischen Modelle eignen sich als Ausgangsbasis für eine Adaption [WKAM94, Köh98, SW98b, BBH+99]. Wie man erwarten würde ist die Adaptionsperformanz stark korreliert mit der Menge an Trainingsmaterial mit der die Modelle adaptiert werden. Die Ergebnisse von [WKAM94] zeigen außerdem, daß die Anzahl der Trainingssprecher kritischer ist als die Anzahl von Trainingsäußerungen. Es ist also wichtiger mehr Variabilität zu präsentieren. In eigenen Arbeiten [SW98b] wurde die Frage untersucht, welche akustischen Modelle zur Adaption besonders geeignet sind. Dazu wurde die Effektivität von multilingualen gegenüber monolingualen Modellen untersucht. Aus den Ergebnissen wird geschlußfolgert, daß multilingual Modelle besser funktionieren als monolinguale, was von [Köh98] bestätigt wird.

Bei der Adaption werden Modelle an eine neue Sprache angepaßt, in der nur sehr limitiertes Datenmaterial zur Verfügung steht. In den wenigen bisher vorhandenen Publikationen zu diesem Thema wird Adaption auf eine neue Sprache analog zum Problem der Sprecheradaption betrachtet. Entsprechend befaßt sich die Forschung zunächst mit der Erstellung möglichst generalisierter (sprachenunabhängiger) Modelle. Anschließend werden diese Modelle entweder durch „Preclustering“ (Bestimmung des am besten passenden Subsets) oder durch schnelle Adaptionstechniken wie Bayessche Adaption [Köh98, FML99, NB99] oder die MLLR Transformation[NB99] auf die neue Sprache angepaßt.

In der Arbeit von Bub [BKI97] werden auf Deutsch, Englisch und Spanisch dreilinguale Keim-Modelle durch datengetriebene Modellkombination erstellt mit dem Ziel, diese auf die slowenische Sprache zu portieren. Dazu wird heuristisch eine Abbildung zwischen den Keim-Phonemen und den Phonemen der slowenischen Sprache erstellt. Die akustischen Modelle für Slowenisch werden anhand ihrer dreilingualen Pendant initialisiert und anschließend durch einen Bayesschen Update der Mittelwerte adaptiert. Die Ergebnisse zeigen, daß dreilingualen Modelle sowohl beim überkreuzsprachlichen Transfer als auch bei der Adaptionsmethode bessere Leistungen zeigen, als die jeweiligen monolingualen deutschen, englischen und spanischen Modelle.

[Köh98] verwendet zur Mittelwertsadaption der Gaußschen Mischverteilungen die MAP-Technik und vergleicht sie mit einem Systemstart von phonetischen Transkriptionen. Seine Ergebnisse zeigen, daß MAP bessere Ergebnisse liefert, wenn man von einer limitierten Menge von Adaptionsdaten ausgeht. Stehen mehr als 30 Minuten Daten in der Zielsprache zur Verfügung, werden mit herkömmlichem Viterbi-

Training gleichwertige oder bessere Zahlen erreicht. Fung vergleicht in [FML99] die MAP-Adaption mit dem Transfer von monolingualen Modellen mit datengetriebener Modellabbildung und stellt fest, daß MAP besser funktioniert.

In [WKAM94] wird der Effekt von unterschiedlich vielen Adaptionsdaten für den Transfer von Englisch nach Japanisch untersucht. Dazu wurden Untermengen gebildet, die sowohl in der Anzahl der Trainingsäußerungen als auch in der Anzahl der Trainingssprecher variieren. Die Ergebnisse zeigen, daß für kleine Adaptionsmengen die Erhöhung der Sprecherzahl eine signifikante Verbesserung der Erkennungsleistung zur Folge hat. Die Verdoppelung der Trainingsäußerungen (1000 auf 2000) bringt ebenfalls einen signifikanten Gewinn. Die Erhöhungen des Trainingsmaterials der bisher enthaltenen Trainingssprecher ist allerdings lange nicht so effektiv, wie die Erhöhung der Sprecherzahlen. Zavialogkos war in [ZC98] zu entgegengesetzten Ergebnissen gelangt. Beide Ergebnisse können jedoch durch Artefakte bedingt sein. So räumte [ZC98] ein, daß die Erhöhung des Materials von einem Sprecher auch eine bessere Abdeckung der Dialoge und damit den Kanaleigenschaften, Vokabular und Sprachmodellen gewährleisten.

In [NB99] werden die bayessche Adaptionstechnik MAP mit der transformationsbasierten Technik MLLR verglichen. Bei den Experimenten, in denen englische Modelle auf die Sprache Afrikaans adaptieren werden, kommen sie zu dem Ergebnis, daß MLLR bessere Leistungen als MAP für die Adaption auf neue Sprachen zeigt. Wie viele andere Autoren stellen sie fest, daß eines der Hauptprobleme der Sprachenadaption in der fehlenden Übereinstimmung phonetischer Kontexte begründet ist. Dies ist eines der wesentlichen Unterschiede zum Problem der Sprecheradaption.

7.2.3 Überkreuzsprachlicher Transfer

Mit dem Begriff überkreuzsprachlicher Transfer wird eine Technik bezeichnet, bei der akustische Modelle, die für eine bestimmte Sprache oder Sprachgruppe trainiert wurden, ohne vorherige Manipulation zum Erkennen einer neuen Sprache angewendet werden. Der überkreuzsprachliche Transfer kommt dann zum Einsatz, wenn überhaupt keine Daten in der Zielsprache vorhanden sind. In einem solchen Fall liegt die einzige Möglichkeit in der Schaffung initialer Zuordnungen, um mit unüberwachten Trainingsmaßnahmen eine iterative Verbesserung zu erwirken [ZC98]. Auf diese Weise könnte ein Spracherkennungssystems von Grund auf ohne menschliches Zutun entstehen [CC97, CuC^{+}97].

Die Hauptfragestellungen beim überkreuzsprachlichen Transfer im Kontext vieler Sprachen lassen sich in drei Gruppen einteilen: Erstens, welche Keim-Modelle eignen sich am besten, zweitens gibt es einen Zusammenhang zwischen Transferleistung und der Sprachenähnlichkeiten bzw. der Anzahl von Sprachen auf denen die Keim-Modelle trainiert worden waren, und drittens wie kann die Abbildung auf ein Aussprachewörterbuch realisiert werden.

Zur Klärung der Frage, welche Modelle sich am besten als Keim-Modelle eignen, wendete [CC97] schweiz-französische (Swiss French - Polyphone) und amerikanische (American English - Timit) Modelle auf französisch, britisch, deutsch, spanisch und italienisch gesprochene Zahlen an. Die Ergebnisse zeigen, daß Französisch am besten mit den schweiz-französischen Modellen erkannt wird. [CC97] schließen daraus, daß Modelle deren Sprachen Ähnlichkeiten aufweisen, sich besser als Keim-Modelle eignen. Für das Paar britsch-amerikanisches Englisch kann diese Hypothese in ihren Experimenten allerdings nicht bestätigt werden.

[BKI97] verglichen drei verschiedene sprachenspezifische Keim-Modelle mit multilingualen Keim-Modellen um slowenisch gesprochene Zahlen zu erkennen. Die multilingualen Modelle erreichten dabei die besten Leistungen. [Köh98] hat den überkreuzsprachlichen Transfer auf Deutsch anhand von multilingualen Modelle mit denen von „Flat-Start“ und gebootstrapten Modellen verglichen. Er kommt ebenfalls zu dem Ergebnis, daß die multilingualen Modelle beste Leistungen zeigen. Sie werden erst übertroffen, nachdem die „Flat-Start“ Modelle mit 10 Minuten deutscher Sprache, die gebootstrapten Modelle mit 5 Minuten Sprache trainiert wurden. Auch [GG97] zeigen, daß multilinguale Keim-Modelle zu besseren Ergebnissen führen als monolinguale. Außerdem stellen sie fest, daß die Resultate verbessert werden, wenn mehr Sprachen in die Modelle einfließen. Allerdings räumen sie ein, daß dies vermutlich ein Effekt des Datenzuwachses ist, aufgrund dessen die multilingualen Modelle akkurater geschätzt werden.

Die Frage, wie ein Aussprachewörterbuch für die Zielsprache gefunden werden kann, wurde bereits im vorigen Abschnitt angeschnitten. Die meisten Arbeiten gehen davon aus, daß ein Aussprachelexikon in der Zielsprache vorhanden ist. In diesem Fall führt man beim überkreuzsprachlichen eine heuristische Abbildung von den Quellphonemen zu den Zielphonemen durch. Viele Forschungsgruppen treffen keine Aussagen zu diesem Problem. [BKI97] und [Köh98] haben jeweils eine IPA-Abbildung durchgeführt. [CC97] schlagen eine umfassende Lösung vor, anhand derer ein Aussprachewörterbuch der Zielsprache in universellen Einheiten ausdrücken werden soll. Dazu entwickeln sie einen Algorithmus, der die Basiseinheiten datengetrieben bestimmt. Dieser Algorithmus setzt allerdings das Vorhandensein von auf Phonemebene transkribierten Äußerungen voraus. Durch die Dekodierung der gegebenen Äußerungen mittels eines Phonemerkenners der Quellsprachen werden initiale Aussprachesequenzen bestimmt. Die beste Aussprache wird nach dem Prinzip des genetischen Algorithmus gefunden, indem die Aussprachesequenzen halbiert und zu neuen Sequenzen zusammengesetzt werden. Die vorläufigen Ergebnisse von Schweizer Französisch nach Französisch sind recht vielversprechend [CC97].

7.3 Sprachenadaptive Kontextmodellierung

In den folgenden Abschnitten werden die multilingualen Modelle, die nach den in Kapitel 6 vorgestellten Methoden der sprachenvermischten Kontextmodellierung erstellt wurden, auf neue, im Training nicht repräsentierte Sprachen angewendet. Die Ergebnisse der Experimente aus Abschnitt 6.4.2 haben bereits vermuten lassen, daß es beim Übergang auf neue Sprachen zu fehlenden Kontextübereinstimmungen kommen würde. Im folgenden soll daher zunächst ermittelt werden, wie groß die Überlappungen der phonetischen Kontexte verschiedener Sprachen sind.

7.3.1 Abdeckung phonetischer Kontexte

Um die Überlappung der Kontexte zwischen verschiedenen Sprachen zu messen, wird das Maß des Abdeckungskoeffizienten cc (engl. *coverage coefficient*) eingeführt. Der Koeffizient $\mathrm{cc}(L_T)$, mit der eine Zielsprache L_T durch das Phoneminventar Υ einer zur Verfügung stehenden Quellsprache abgedeckt wird, berechnet sich zu:

$$\mathrm{cc_N(L_T)} = \frac{|\Upsilon_{\mathrm{L_T}} \cap \Upsilon|}{|\Upsilon_{\mathrm{L_T}}|} = 1 - \frac{|\Upsilon_{\mathrm{LD_{L_T}}}|}{|\Upsilon_{\mathrm{L_T}}|} \tag{7.1}$$

Es gilt die Bedingung $0 \leq \mathrm{cc(L_T)} \leq 1$. cc ist Null, wenn kein Phonem der Zielsprache L_T ein Gegenstück im Phonemset der Quellsprachen hat und cc ist eins, wenn jedes Zielsprachenphonem abgedeckt ist. Während der in Kapitel 6 eingeführte share factor sf ein mittleres Maß der gemeinsamen Nutzung der Phoneme eines globalen Phonemsets angibt, mißt der Abdeckungskoeffizient cc den Anteil der Phoneme der Zielsprache L_T, der von Phonemen des globalen Phonemsets abgedeckt wird.

Die Idee der Phonemabdeckung kann sehr einfach auf die Abdeckung von Modellen mit variabler Kontextbreite erweitert werden. Anstelle der Phoneme in Form der Monophone treten nun Triphone, Quintphone und allgemein Polyphone. Darüber hinaus wird zwischen der Abdeckung von Polyphontypen und der von Polyphonvorkommen differenziert. Im letzteren Fall wird zur Berechnung des Abdeckungskoeffizienten eines Polyphons dessen Auftrittshäufigkeit mitgewichtet. Das gewichtete Maß reflektiert die Tatsache, daß es für die Erkennungsleistung von größerer Bedeutung ist, häufig auftretende Polyphone abzudecken als seltene.

Tabelle 7.1 zeigt die Triphonabdeckung im ungewichteten (oberer Eintrag) und gewichteten Fall (unterer Eintrag) für Sprachpaare aus 10 GlobalPhone-Sprachen. Die Ergebnisse zeigen, daß es bereits bei einer Kontextbreite von ± 1 zu sehr geringen Abdeckungsraten kommt.

Um zu untersuchen, wie sich die Abdeckungsraten mit steigender Kontextbreite verhalten, wird cc für einzelne Sprachen, hier exemplarisch an Portugiesisch gezeigt, berechnet. Dazu wird gemessen, wie gut das Lautinventar einer Sprache, hier Portugiesisch mit 46 Phonemen von dem globalen Phonemset des gegebenen Sprachpools (hier 9 Sprachen) abgedeckt wird. In Abbildung 7.1 ist die prozentuale Abdeckung $\mathrm{cc(Po)} \times 100$ gegen die Anzahl der beteiligten Sprachen aufgetragen. Es werden drei Kurven für die Kontextbreiten 0 (Monophone), ± 1 (Triphone) und ± 2 (Quintphone)

B/C	CH	DE	EN	FR	JA	KO	KR	PO	SP	TU
CH	100	0.1 5.3	0.3 6.8	0.1 5.8	0.1 4.2	0.0 5.3	0.1 4.2	0.1 5.4	0.1 5.3	0.2 4.9
DE	0.1 3.9	100	5.5 19.6	19.8 41.6	9.3 19.5	7.2 18.2	18.6 34.9	13.6 28.0	12.9 28.3	12.9 26.1
EN	0.6 5.2	5.4 18.1	100	6.5 18.6	1.8 8.9	3.4 11.6	1.5 7.7	0.9 6.6	1.3 6.6	3.8 9.2
FR	0.1 3.9	29.0 53.3	9.7 16.4	100	10.2 22.7	11.2 28.7	25.8 45.5	18.4 36.4	17.4 41.3	23.1 35.6
JA	0.2 2.5	22.3 33.6	4.5 9.9	16.8 37.4	100	9.8 25.6	16.0 29.2	11.0 27.6	13.6 31.2	25.9 52.5
KO	0.1 4.1	10.3 36.3	4.9 16.1	10.9 35.0	5.8 24.9	100	10.2 38.6	8.0 30.8	9.3 38.4	9.1 26.1
KR	0.2 1.8	39.0 68.8	3.2 5.0	37.0 64.7	14.0 28.2	15.0 34.5	100	31.0 63.0	34.3 61.8	31.5 50.4
PO	0.4 2.3	30.2 57.9	2.0 4.6	28.0 49.5	10.2 26.7	12.5 37.5	32.9 62.5	100	33.5 57.5	19.8 39.9
SP	0.2 2.5	25.4 60.2	2.7 5.6	23.5 60.1	11.2 34.0	12.9 40.1	32.2 64.2	29.7 58.2	100	17.5 41.0
TU	0.8 5.4	29.6 46.0	8.9 18.3	36.3 52.0	24.8 46.1	14.6 33.0	34.4 50.1	20.4 38.6	20.3 39.6	100

Tabelle 7.1: Triphonabdeckung für Sprachpaare aus 10 GlobalPhone-Sprachen

berechnet. Die Berechnung der Abdeckung wurde folgendermaßen durchgeführt: Zuerst wird diejenige Sprache aus dem Sprachenpool ausgewählt, die die höchste Abdeckungsrate auf Portugiesisch erzielt. Diese Sprache wird aus dem Sprachenpool entfernt. Danach wird aus den im Sprachenpool verbliebenen Sprachen diejenige ausgewählt, die gemeinsam mit der entfernten Sprache die höchste Abdeckung auf Portugiesisch erreicht. Die ausgewählte Sprache wird aus dem Pool entfernt und die Prozedur für Sprachentripel, dann für Sprachenquadrupel usw. fortgesetzt. In jedem Schritt wird somit diejenige Sprache ausgewählt, die das Polyphonset maximal komplementiert.

Abbildung 7.1 zeigt, daß, wie erwartet, die Abdeckung für wachsende Kontextbreite dramatisch absinkt. Mit einem Phonempool von 9 Sprachen wird eine Abdeckung auf portugiesischen Monophonen von 91% erreicht, diese sinkt auf 73% für Triphone und auf 47% für Quintphone. Für die Monophonabdeckung genügt der Beitrag der drei wichtigsten Sprachen, um das Maximum der Abdeckungsrate zu erreichen. Bei Triphonen sind es bereits vier Sprachen bevor die Saturierung eintritt. Bei Quintphonen leisten fünf Sprachen einen meßbaren Beitrag zur Abdeckung. Daraus kann

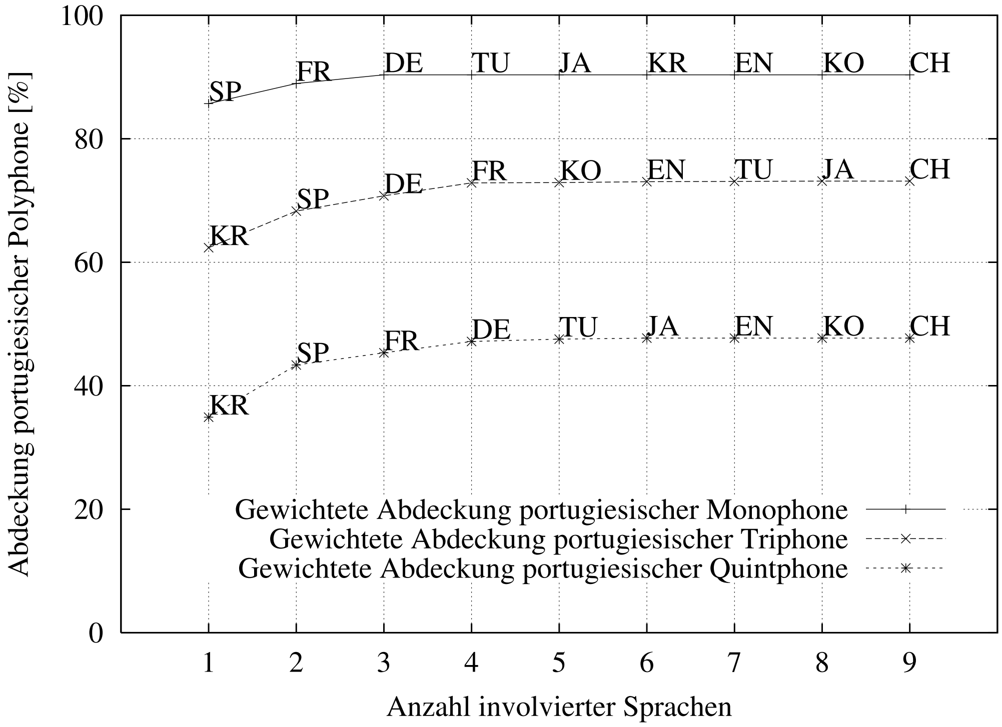


Abbildung 7.1: Abdeckungsrate portugiesischer Polyphone durch 9 Sprachen

gefolgert werden, daß mit wachsender Kontextbreite mehr Sprachen eingesetzt werden sollten.

In den folgenden Experimenten wird untersucht, welche Auswirkung die Entfernung derjenigen Sprachen hat, die zur Abdeckung den Hauptbeitrag leisten. Im Beispiel mit Portugiesisch, werden abwechselnd die Sprachen Spanisch, Deutsch, Kroatisch und Französisch entfernt. Der Beitrag des Spanischen kann fast vollständig durch deutsche plus kroatische Phoneme kompensiert werden. Die gilt ebenfalls für das Entfernen von Deutsch, bzw. Kroatisch. Dies impliziert, daß diese drei Sprachen die gleichen portugiesischen Polyphone abdecken. Dagegen kann der Wegfall der französischen Phoneme nicht kompensiert werden, d.h. Französisch deckt einzigartige Phoneme des Portugiesischen ab. In diesem Fall sind es die nasalen Vokale, die im Portugiesischen sehr häufig auftreten aber in keiner der anderen Sprachen außer Französisch vorhanden sind.

Daraus läßt sich schließen, daß es bei der Auswahl von Sprachen für eine multilinguale Datenbasis wichtig ist, auf dessen Komplementarität zu achten. Diese Eigenschaft ist im Hinblick auf die Portierung auf neue Sprachen vermutlich wichtiger als die reine Anzahl der beteiligten Sprachen. Die Berechnung des Abdeckungskoeffizienten cc trägt dazu bei, dieses komplementäre Sprachenset zu finden.

7.3.2 Polyphone Decision Tree Specialization (PDTS)

Die zweite und wichtigere Schlußfolgerung aus den Ergebnissen der Polyphonabdeckungsraten ist, daß man bei der Modellierung mit kontextabhängigen akustischen Modellen auf neue Sprachen mit sehr niedrigen Überlappungen rechnen muß. Als Folge davon ist zu erwarten, daß ein Kontextentscheidungsbaum, der auf N Sprachen erstellt wurde, nicht sehr gut auf die $N+1$-te Sprache passen wird. Andererseits sollte man nicht auf die Modellierung des Kontextes verzichten, weil prinzipiell ein Gewinn daraus gezogen werden kann. Zur Neuberechnung eines Kontextentscheidungsbaumes ist allerdings eine große Menge Sprachdatenmaterial erforderlich. Um den Vorteil der Kontextmodellierung auch dann nutzen zu können, wenn nur sehr wenig Material in der neuen Sprache zur Verfügung steht, wird in der vorliegenden Arbeit eine Technik eingeführt, die es erlaubt, einen bestehenden Kontextentscheidungsbaum mit limitiertem Material entsprechend anzupassen. Diese Methode wird im folgende als Spezialisierung von Kontextentscheidungsbäumen (engl. *Polyphone Decision Tree Specialisation (PDTS)*) bezeichnet.

Mit der Methode PDTS wird ein Kontextentscheidungsbaum auf die neue Sprache spezialisiert, indem zunächst alle Äste aus dem Baum entfernt werden, die in der Zielsprache nicht vorhanden sind. Im zweiten Schritt wird der beschnittene Baum in einem erneuten Wachstumsprozeß auf die gesehenen Modelle der Zielsprache spezialisiert [SW99, Wol99, SW00].

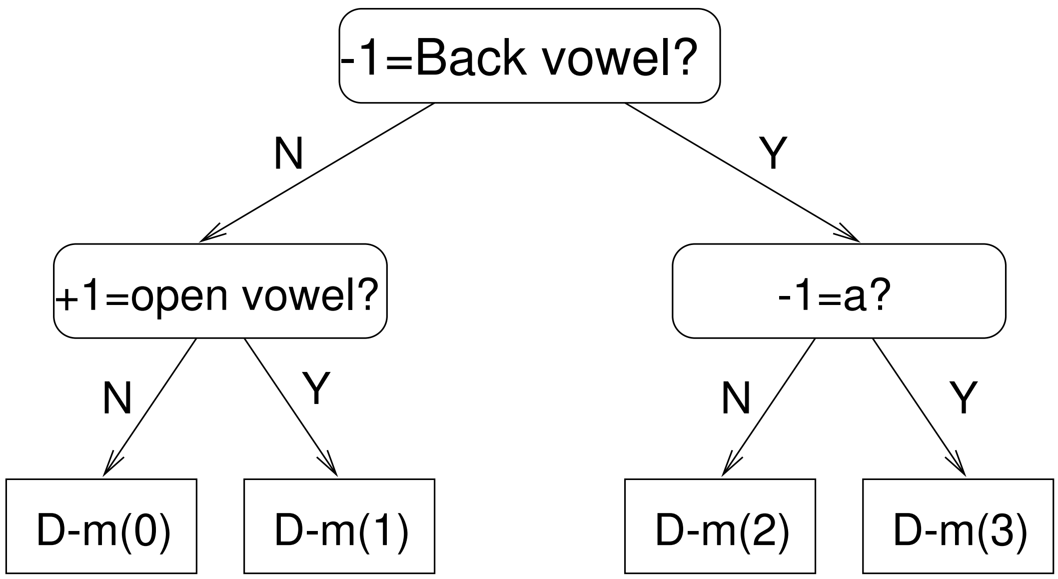


Abbildung 7.2: Entscheidungsbaum **vor** Polyphone Decision Tree Specialization

Abbildung 7.2 zeigt einen Ausschnitt aus einem Entscheidungsbaum für das Subpolyphone /d^j-m/ vor der Anwendung von PDTS. Der ursprüngliche Baum hatte nur drei Aufspaltungen erfahren, die in vier Blatt-Knoten resultieren. Offensichtlich spielt dieses Phonem in den Quellsprachen keine besonders wichtige Rolle. Dagegen kommt dieses Phonem im Portugiesischen sehr häufig vor. Wenn man nun mit diesem Baum portugiesische Sprache erkennen möchte, gelangt man beim Traversieren des Baumes während des Dekodierens zu schlecht modellierten Restklassen, weil die Kontextfragen nicht die portugiesischen Kontextregeln reflektieren. Dies führt zu suboptimalen Erkennungsergebnissen für die portugiesische Sprache.

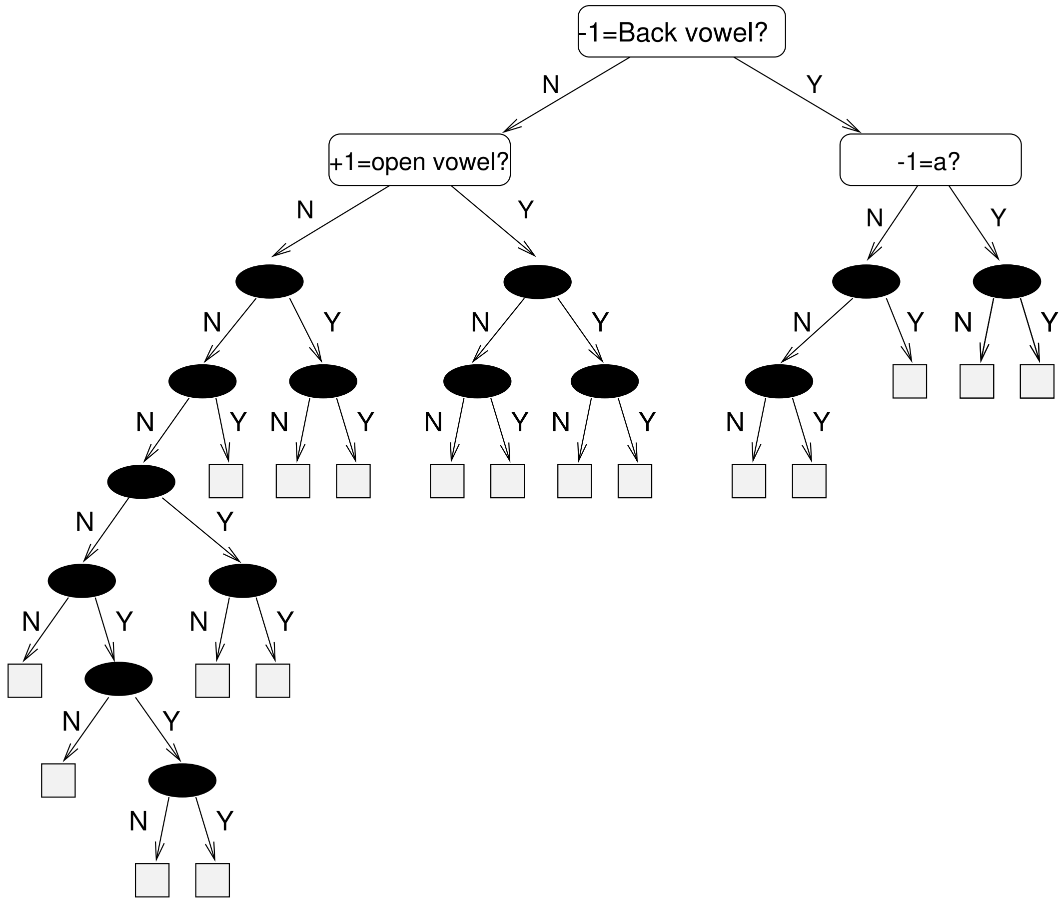


Abbildung 7.3: Entscheidungsbaum **nach** Polyphone Decision Tree Specialization

Abbildung 7.3 zeigt denselben Ursprungsbaum des mittleren Zustandes des Phonems /d^j-m/ nach der Anwendung von PDTS. Der ursprüngliche Entscheidungsbaum wurde durch 14 zusätzliche Fragen in insgesamt 18 Blätter aufgespaltet. Der Wachstumsprozeß des Entscheidungsbaumes wurde abgebrochen, nachdem eine zuvor definierte Anzahl Blattknoten erreicht war. Die Zahl der Blätter hängt daher auch von der vorhandenen Trainingsmenge ab. Der adaptierte Entscheidungsbaum repräsentiert nun passende Kontexte des portugiesischen Phonems /d^j/. Es wird von einem System mit diesem Baum eine verbesserte Erkennungsleistung bei der Anwendung auf portugiesische Sprache erwartet. Diese Hypothese soll in den Experimenten in Abschnitt 7.5 evaluieren werden.

7.4 Überkreuzsprachlicher Transfer und Bootstrapping auf Schwedisch

In diesem Abschnitt wird anhand mono- und multilingualer Systeme auf der Basis von sieben Sprachen der überkreuzsprachliche Transfer und Bootstrapping auf Schwedisch untersucht. Mit Schwedisch wurde bewußt eine Sprache ausgewählt, die im Rahmen der Spracherkennung noch wenig erforscht ist und für die kaum Wissensquellen vorliegen. Das Ausspracheworterbuch wurde durch die in Abschnitt 5.3.5 beschriebene Graphem-zu-Phonem-Konvertierung generiert, wobei Schwedisch mit etwa 250 Graphem-zu-Phonem-Regeln zu den schwierigeren Sprachen zählt. Es sind derzeit noch keine großen Textkorpora gesammelt und aufbereitet worden und die wortbasierten Erkennungsleistungen sind nicht sehr zuverlässig. Das eingesetzte Sprachmodell hat eine Trigramm-Perplexität von 1000 bei einem Vokabular von 24000. Es werden daher die Ergebnisse der Experimente in Phonemerkennungsfehlern gebundener Phonemerkennungsläufe beschrieben. Zum Training der schwedischen Modelle beim Bootstrapping werden etwa 17 Stunden Sprachmaterial verwendet.

Grundlage der Untersuchungen sind mono- und multilinguale Erkenner auf den sieben Sprachen Chinesisch, Deutsch, Französisch, Japanisch, Kroatisch, Spanisch und Türkisch. Dabei handelt es sich um die besten verfügbaren Erkenner, wie sie in Abschnitt 5.6 beschrieben sind.

Der Fokus der Experimente liegt zum einen auf dem Vergleich monolingualer Erkenner zum überkreuzsprachlichen Transfer und zum Bootstrapping. Zum anderen werden verschiedene Methoden zur Phonemabbildung eingeführt und evaluiert. In Abschnitt 6.2.2 wurden bereits zwei Verfahren definiert, allerdings unter der Prämisse, daß keine Trainingsdaten in der neuen Sprachen verfügbar sind. Hier werden nun Methoden vorgestellt, die die vorhandenen Trainingsdaten ausnutzen. Desweiteren wird die Eignung monolingualer kontextabhängiger Phonemmodellen zum überkreuzsprachlichen Transfer untersucht.

7.4.1 Monolingualer überkreuzsprachlicher Transfer

Im folgenden Experiment wird evaluiert, ob sich Ähnlichkeiten oder Sprachverwandtschaften zwischen der Quell- und Zielsprache beim überkreuzsprachlichen Transfer in Erkennungsleistungen wiederspiegeln. Es werden die monolingualen Erkenner der sieben Sprachen auf Schwedisch angewendet, ohne daß deren akustische

Sprache	Überkreuzsprachlicher Transfer
Chinesisch	75.2
Deutsch	64.9
Französisch	69.6
Japanisch	76.0
Kroatisch	59.0
Spanisch	69.6
Türkisch	59.9
$Mittel_{L7}$	67.8

Tabelle 7.2: Monolingualer überkreuzsprachlicher Transfer auf Schwedisch [PE in %]

Modelle auf Schwedisch adaptiert oder trainiert werden. Die Abbildung der Aussprachewörterbücher geschieht über das IPA-Referenzschema. Tabelle 7.2 zeigt für alle sieben Sprachen die Resultate dieses überkreuzsprachlichen Transfers auf Schwedisch.

Die Auswertung zeigt, daß sich im wesentlichen 3 Leistungsgruppen ergeben. Die besten Leistungen beim überkreuzsprachlichen Transfer ergeben der kroatische und der türkische Erkenner, mit einigem Abstand der deutsche Erkenner, mit weiterem Abstand folgt dann die Gruppe Französisch und Spanisch. Weit abgeschlagen liegen Japanisch und Chinesisch. Berücksichtigt man den Verwandtschaftsgrad zwischen Schwedisch und den einzelnen Sprachen, dann müßte der deutsche Erkenner die besten Leistungen zeigen. Deutsch liegt aber erst an dritter Stelle, allerdings vor allen anderen indoeuropäischen Sprachen. Die Spracherkenner mit den besten Leistungen wurden auf Sprachen anderer Sprachfamilien trainiert. Insgesamt läßt sich aus diesen Ergebnissen kein Zusammenhang zwischen Sprachverwandtschaft und Tranfereignung erkennen.

Für Schwedisch wurde durch die in Abschnitt 7.3.1 beschriebene Methode die Phonemabdeckungsrate berechnet. Für Monophone ergab sich als Reihenfolge der Ähnlichkeiten DE-JA-FR-EN-CH-{SP,KR,TU,KO,PO}. Für Triphone ergab sich mit DE-JA-FR-EN-KR eine auf den ersten Positionen identische Reihenfolge. Auch der Zusammenhang mit der Abdeckung der Monophone läßt keinen eindeutigen Zusammenhang zwischen Transferleistung und Ähnlichkeiten erkennen. Ein Zusammenhang mit der Größe des Phoneminventars würde zwar das gute Abschneiden von Kroatisch und Türkisch erklären, aber nicht das von Deutsch. Auch paßt das schlechte Abschneiden von Japanisch nicht zu dieser Hypothese.

Der Vergleich der Ergebnisse des überkreuzsprachlichen Transfers zwischen der am schlechtesten geeigneten Sprache Japanisch (76.0) und der am besten geeigneten Sprache Kroatisch (59.0) zeigen einen enorm großen Unterschied von 17.0 Prozentpunkten, was einem relativen Leistungsunterschied von 22% entspricht. Beim über-

Sprache	Kontext- unabhängig	Kontext- abhängig	Δ
Chinesisch	75.2	76.0	-1.0%
Deutsch	64.9	63.2	2.6%
Französisch	69.6	70.3	-0.7%
Japanisch	76.0	74.1	2.5%
Kroatisch	59.0	58.1	1.5%
Spanisch	69.6	67.1	3.6%
Türkisch	59.9	59.9	0%
$Mittel_{L7}$	67.8	67.0	1.2%

Tabelle 7.3: Kontextunabhängige gegenüber kontextabhängigen Modellen zum überkreuzsprachlichen Transfer auf Schwedisch [PE in %]

kreuzsprachlichen Transfer macht es also einen signifikanten Unterschied, ob die zum Transfer am besten geeignete Sprache bekannt ist.

7.4.2 Eignung monolingualer kontextabhängiger Modelle

Im vorherigen Experiment wurden monolinguale Spracherkenner mit kontextunabhängigen Phonemmodellen eingesetzt. Das folgende Experiment soll klären, ob kontextabhängige Modelle beim überkreuzsprachlichen Transfer eine ähnliche Leistungsverbesserung zeigen wie bei monolingualen Experimenten. Tabelle 7.3 enthält für die sieben Sprachen die Phonemerkennungsleistungen beim überkreuzsprachlichen Transfer auf Schwedisch. Es zeigt sich, daß eine Leistungsverbesserung durch kontextabhängige Modelle bei der Anwendung zum überkreuzsprachlichen Transfer ausbleibt. Kontextabhängige Modelle erzielen im Mittel nur 1.2% bessere Leistungen. Diese Beobachtung wurde bereits für die Erkennung von stark akzentbehafteter Nichtmuttersprache berichtet ([Van99]). Der Grund, daß sich die Gewinne durch Kontextmodellierung nahezu auswaschen, liegt nach der Auffassung der Autorin in der fehlenden Überlappung der Kontexte zwischen verschiedenen Sprachen. Wie Tabelle 7.3 zeigt, gehören die Sprachen Deutsch und Japanisch zu denjenigen Sprachen, bei denen etwas höhere Gewinne durch die Kontextmodellierung beobachtet werden. Diese beiden Sprachen hatten die höchsten Abdeckungsraten mit schwedischen Triphone.

Aus den Beobachtungen ist zu folgern, daß aus einer monolingualen Kontextmodellierung keine Gewinne erwartet werden können, wenn die Modelle ohne vorherige Anpassung zur Erkennung neuer Sprache eingesetzt werden sollen. In Abschnitt 7.5 wird dieses Problem mit der PDTS-Methode angegangen.

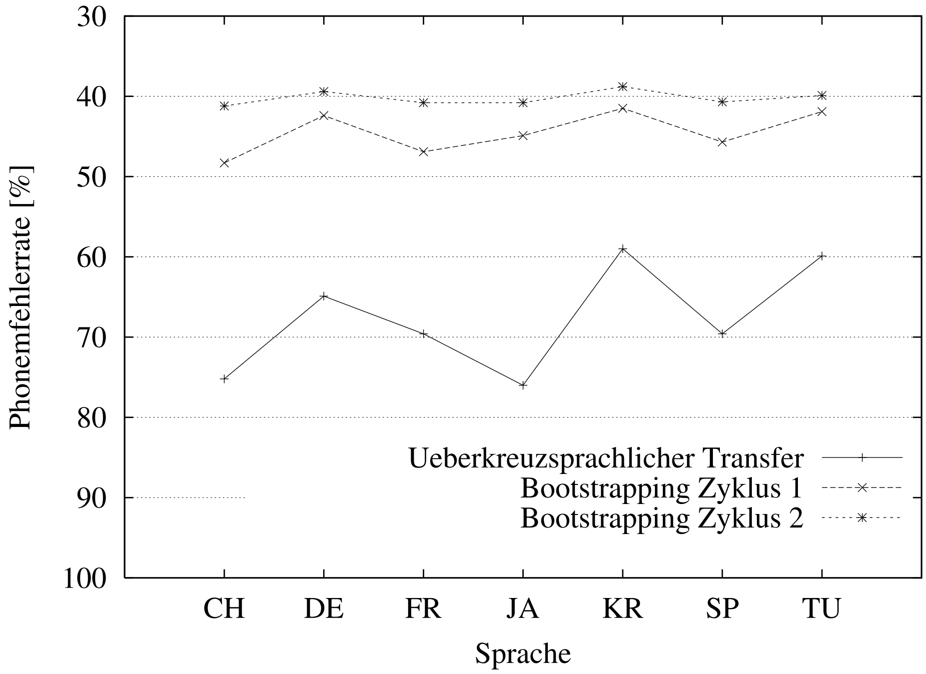


Abbildung 7.4: Monolinguales Bootstrapping auf Schwedisch [PE in %]

7.4.3 Bootstrapping auf Schwedisch

In Abbildung 7.4 sind die Ergebnisse nach einem bzw. zwei Zyklen des QUICKBOOT-Verfahrens dargestellt. Dabei wurden die schwedischen akustischen Modelle nach der Initialisierung mit dem schwedischen Trainingsdatenmaterial trainiert. Aus den abgebildeten Ergebnissen des Bootstrapping wird ersichtlich, daß sich die anfänglich stark variierenden Fehlerraten des überkreuzsprachlichen Transfers durch das Training auf schwedischen Daten zwar einebnen aber nicht ganz nivellieren. Die Spannbreite reduziert sich von 22% auf 5% nach dem ersten Bootstrapping-Zyklus und auf 6% nach dem zweiten Zyklus. Die Differenz der Phonemerkennungsleistungen zwischen dem besten und dem schlechtesten monolingualen Erkenner bleibt konstant auf 2.4 Prozentpunkten in beiden Bootstrapping-Zyklen. Damit kann durch die Auswahl der am besten geeigneten Sprache der Erkennungsfehler im Extremfall somit um 6% relativ gegenüber der schlechtesten Wahl reduziert werden. Insgesamt ist die Auswahl der „richtigen" Sprache beim überkreuzsprachlichen Transfer wesentlich bedeutsamer als beim Bootstrapping.

7.4.4 Multilinguale Phonemabbildungen

Im folgenden sollen zum überkreuzsprachlichen Transfer und zum Bootstrapping multilinguale akustische Modelle eingesetzt werden. Wie die Vorexperimente auf Deutsch in Abschnitt 6.4.2 bereits gezeigt haben, sind für diesen Zweck die sprachenvermischten Modelle am besten geeignet. Im Fall der multilingualen Modelle stehen im Gegensatz zur monolingualen Modellierung nun eine ganze Reihe von Phonemen zur Auswahl, die auf die schwedischen Phoneme abgebildet werden können. Es erhebt sich daher die Frage, wie man die Abbildung der Phoneme geeignet bestimmt. Die Phonemabbildung können dann zu folgenden Zwecken eingesetzt werden:

- Überkreuzsprachlicher Transfer: Abbildung des Ausspracheworterbuchs
- Bootstrapping: Geeignete Initialisierung der Modelle der Zielsprache
- Adaption: Nutzung geeigneter Daten aus mehreren Sprachen.

In bisherigen Experimenten war die Phonemabbbildung unter dem Aspekt betrachtet worden, daß keine Daten der Zielsprache vorhanden sind. Daher kam nur eine heuristische, wissensbasierte Phonemabbildung in Betracht. Beim Bootstrapping steht dagegen ausreichend Trainingsmaterial zur Verfügung, daher bietet sich neben der wissensbasierten Phonemabbildung die Möglichkeit einer automatischen, datengetriebenen Bestimmung der Phonemabbildung. Es werden hier zwei Methoden zur datengetriebenen Bestimmung der multilingualen Phonemabbildung eingeführt und mit der wissensbasierten Abbildungsmethode verglichen.

7.4.4.1 Wissensbasierte Phonemabbildung

In Abschnitt 6.2.2 wurden zwei Methoden eingeführt, mit denen anhand des IPA-Referenzschemas eine wissensbasierte Abbildung der Phoneme vorgenommen wurde. Unter der Voraussetzung, daß Trainingsmaterial in der Zielsprache vorhanden ist, hat sich die sprachenübergreifende Abbildungsmethode (IPA-ML) als leistungsfähiger erwiesen (siehe Tabelle 6.11) als die parallele Aussprachemethode. Daher wurde hier die sprachenübergreifende Methode als Grundlage zum Vergleich der Phonemabbildungsmethoden herangezogen. In Tabelle 7.4 sind die Abbildungen der multilingualen Phoneme auf die schwedischen Phoneme aufgeführt. Die zweite Spalte „IPAMap“ zeigt, welche Phonemabbildungen sich aus der IPA-ML-Abbildung für die schwedische Sprache ergeben. Bis auf die neun mit der Markierung „(-)“ gekennzeichneten Phoneme konnten alle schwedischen Phoneme durch passende IPA-Pendant aus dem siebenlingualen Phoneminventar ersetzt werden.

7.4.4.2 Datengetriebene Phonemabbildung

Für die datengetriebene Methode soll die Tatsache ausgenutzt werden, daß Trainingsmaterial in der Zielsprache vorhanden ist. An dieser Stelle wird vorausgesetzt, daß für eine limitierte Menge von 500 Äußerungen (etwa 1 Stunde Sprachdaten) phonetisch transkribiertes Material zur Verfügung steht. Dies kann entweder durch eine phonetische Transkription erreicht werden, die von Experten durchgeführt wird, oder

Zielphonem	IPAMap	PhonMap	SubPhonMap		
			-b	-m	-e
p	p	p	p-b	p-m	p-e
b	b	b	b-b	b-m	b-e
t	t	t	t-b	t-m	t-e
d	d	d	d-b	d-m	d-e
ʈ	t (-)	t	t-b	t-m	t-e
ɖ	d (-)	d	b-b	d-m	d-e
k	k	k	k-b	k-m	k-e
g	g	g	g-b	g-m	g-e
m	m	m	m-b	m-m	m-e
n	n	n	n-b	n-m	n-e
ɳ	n (-)	n	n-b	n-m	n-e
ŋ	ŋ	ŋ	ei-e	ɴ-m	ŋ-e
r	r	r	r-b	r-m	r-e
f	f	f	f-b	f-m	f-e
v	v	v	v-b	v-m	v-e
s	s	s	s-b	s-m	s-e
ʃ	ʃ	ʃ	ʃ-b	ʃ-m	k-e
ʂ	ʂ	ʃ	s-b	ʃ-m	ʃ-e
ç	ç	x	θ-b	ɯ-m	x-e
h	h	h	h-b	h-m	h-e
j	j	j	ʎ-b	j-m	j-e
l	l	l	l-b	l-m	l-e
ɭ	l (-)	l	l-b	l-m	l-e
ks	x (-)	s	ts-b	s-m	s-e
i	i	e	i:-b	i:-m	e-e
i:	i:	i	ʎ-b	i:-m	i-e
y	y	e:	e:-b	e:-m	i-e
y:	y:	e:	uei-m	e:-m	e:-e
ʉ:	u (-)	ø:	ø:-m	ɯ-m	ɯ-m
u	u	ʊ	ɔ-b	ʊ-b	ʊ-m
u:	u:	u	u-b	ʊ-m	'u-e
e	e	e:	e:-m	e:-m	e-e
e:	e:	e	e:-b	e:-m	e-e
ø	ø	œ	œ-e	œ-m	œ-e
ø:	ø:	œ	ɔ-m	œ-m	œ-e
ə	ə	e	i:-e	uei-m	e-e
ɵ	ə (-)	ɔ	y-m	ɔ-m	ɔ-e
o:	o:	o:	o:-b	o:-m	o:-e
ɛ	ɛ	e	e-b	e-m	e-e
ɛ:	ɛ:	e	e-b	ɛ-m	'e-e
œ	œ	ɐ	ɐ-b	ɐ-m	ɐ-e
œ:	œ (-)	eu	eu-b	eu-b	eu-m
ɔ	ɔ	o:	o:-b	o:-m	o:-e
æ	æ	e	e-b	e-m	ai-m
æ:	æ (-)	'a	œ-b	ɐ-m	'a-b
a	a	ɑ	a:-b	a:-m	'a-m
a:	a:	a:	œ-m	iao-m	au-m
ɑ:	ɑ	ɑ	a:-b	a:-m	'a-e

Tabelle 7.4: Multilinguale Phonemabbildungen auf Schwedisch

durch die automatische Zeitzuordnung mittels eines Phonemerkenners. In den vorliegenden Experimenten wurden die Zeitzuordnungen durch den Viterbi-Algorithmus mittels eines Phonemerkenners automatisch berechnet.

Zur datengetriebenen Bestimmung der Phonemabbildung wird auf 500 Äußerungen eine Viterbi-Zuordnung der Phoneme durch den besten Phonemerkenner ermittelt, der durch das Bootstrapping auf IPA-Abbildungen entstanden ist. Die Ausgabe des Phonemerkenners besteht aus schwedischen Phonemen und wird im folgenden als Referenz bezeichnet. Dieselben 500 Äußerungen werden mit dem multilingualen Phonemerkenner dekodiert, der aus den sprachenvermischten Modellen von sieben Sprachen besteht. Die Ausgabe wird im folgenden als Hypothese bezeichnet. Im Anschluß werden die Referenz und die Hypothese frame-weise verglichen und damit eine Phonemverwechslungsmatrix zwischen den schwedischen Referenzphonemen und den multilingualen hypothetisierten Phonemen berechnet. Diese Matrix enthält für jedes Phonempaar einen Wert, der die Häufigkeit der Verwechslungen der Phoneme dieses Phonempaars angibt. Dieser Wert wird mit der jeweiligen Gesamtauftrittszahl des hypothetisierten Phonems normiert. Aus der entstehenden Matrix wird für jedes Referenzphonem dasjenige hypothetisierte Phonem ausgewählt, dessen normierter Eintrag den höchsten Wert hat. Tabelle 7.4 enthält in der dritten Spalte „PhonMap“ die Ergebnisse für die so entstandene Abbildung der Phoneme.

Das schwedische Phoneminventar ist relativ umfangreich und hat insbesondere viele Vokale. Um zu erreichen, daß diese Vokale besser durch initiale Modelle abgedeckt werden, wird die phonembasierte Abbildung auf Subphoneme erweitert. Dazu wurden die Referenzen und Hypothesen in die jeweiligen Subphonemsequenzen zerlegt und eine neue Phonemverwechslungsmatrix berechnet. Die Ergebnisse der Subphonembasierten Phonemabbildung sind in Tabelle 7.4 in der vierten Spalte „SubPhonMap“ eingetragen.

Diese subphonembasierte Abbildung ist besonders nützlich, wenn ein unbekanntes Phonem ersetzt werden soll, das sich möglicherweise aus mehreren bekannten zusammensetzen läßt. In der heuristischen Abbildungsmethode wurde diese Vorgehensweise für die Initialisierung der chinesischen Modelle eingesetzt. Da es im Chinesischen besonders viele Triphthonge gibt, die in anderen Sprachen nicht vorkommen, wurden diese aus den Subphonemen der jeweiligen Monophthonge zusammengesetzt.

Der Vergleich der resultierenden Abbildungen in Tabelle 7.4 zeigt, daß die schwedischen Konsonanten von allen drei Verfahren weitgehend übereinstimmend abbgebildet werden. Bei den Vokalen gibt es aber große Unterschiede. Die Analyse der Konfusionsmatrix spiegelt dieses Resultat eindeutig wieder. Die Verwechslung des besten Kandidaten mit dem Zielphonem war bei Konsonanten um Größenordnungen häufiger als Verwechslungen mit dem zweitbesten Kandidaten, während bei den Vokalen die Häufigkeiten sehr nahe aneinander lagen. Diese Beobachtungen deutet darauf hin, daß Konsonanten über Sprachen hinweg konstanter sind als Vokale. Bary hat im Kontext von akzentbehafteter Sprache gezeigt, daß sich in Vokalen stärkere Variationen zeigen, als in Konsonanten [BHN89]. In der bekannten Literatur wurden auch eher von Übereinstimmungen zwischen Konsonanten, nicht jedoch zwischen Vokalen berichtet. Möglicherweise ist dies dadurch bedingt, daß Vokale aufgrund der graduellen Zungenbewegungen schwerer zu klassifizieren sind als Konsonanten. Möglicherweise wird das Ergebnis aber auch durch das auffallend große schwedische Vokalinventar hervorgerufen.

Sprache	ÜT	Boot-1	Boot-2
IPAMap	65.8	43.9	40.2
PhonMap	60.9	43.8	39.7
SubphonMap	61.8	42.3	39.5

Tabelle 7.5: Vergleich der Phonemabbildungen [PE in %]

Tabelle 7.5 vergleicht die Phonemfehlerraten der drei Abbildungsmethoden für den überkreuzsprachlichen Transfer und das Bootstrapping miteinander. Es zeigt sich, daß die datengetriebene Bestimmung beim überkreuzsprachlichen Transfer zu signifikant besseren Leistungen führt. Die Abbildung auf Phonemebene ist um 7.4% besser als die heuristische Abbildungsmethode, die Abbildung auf Subphonemebene um 6% besser. Durch das Training auf schwedischen Daten nivellieren sich diese Unterschiede, die datengetriebene Abbildung ist nach einem Zyklus noch um 3.6% besser, nach dem zweiten Zyklus nur noch insignifikant um etwa 1.7% besser als die heuristische Methode. Auffallend ist, daß die anfänglich bessere Methode auf Phonemebene nach dem Training schlechter ist als die Subphonem-Methode.

Sprache	ÜT	Boot-1	Boot-2
$Best_{L7}$	59.0	41.5	38.8
$Mittel_{L7}$	67.8	44.5	40.2
PhonMap	60.9	43.8	39.7
PhonMap128	58.7	36.6	34.3

Tabelle 7.6: Vergleich mono- vs multilinguale Modelle [PE in %]

In den multilingualen Modellen werden die Daten von sieben Sprachen vermischt. Dadurch stehen pro Modell mehr Daten zur Schätzung der Parameter zur Verfügung. Dieser Sachverhalt kann dazu genutzt werden, eine höhere Zahl an Parametern zu modellieren. Bisher wurden die multilingualen Modelle mit der gleichen Zahl an Parametern modelliert wie die monolingualen Modelle. Im vorliegenden Experiment wird die Anzahl der Parameter im Verhältnis der zusätzlichen Daten erhöht. Tabelle 7.6 zeigt das Ergebnis eines multilingualen Phonemerkenners dessen Modelle jeweils mit 128 Gaußschen Mischverteilungen statt der bisher 16 Mischverteilungen modelliert sind ($\text{PhonMap}_1 28$) und vergleicht die Erkennungsleistung mit dem System, das 16 Mischverteilungen modelliert (PhonMap). Beim überkreuzsprachlichen Transfer ergibt sich dadurch eine 3.6% bessere Leistung, beim Bootstrapping ergeben sich 13.6% Verbesserungen, was aber angesichts der zum Training verwendeten umfangreichen schwedischen Daten nicht verwundert.

Abschließend läßt sich sagen, daß die multilingualen Modelle in Kombination mit der subphonembasierten Abbildung im überkreuzsprachlichen Vergleich in allen Fällen bessere Leistungen erzielen als die monolingualen Modelle. Zwar sind die Unterschiede im Vergleich zum besten monolingualen System nicht signifikant, aber im Mittel liefern die multilingualen Modelle eine signifikante Verbesserung von 13.4% gegenüber den monolingualen. Die Schwankungsbreite der monolingualen Systeme ist mit 22% enorm groß. Der Einsatz der monolingualen Modelle zum überkreuzsprachlichen Transfer würde daher a-priori Wissen über die Eignung der einzelnen Sprachen erfordern. Beim Einsatz der multilingualen Modelle ist dieses Wissen nicht notwendig.

7.5 Adaption auf Portugiesisch

In den folgenden Experimenten wird der Nutzen der akustischen Modellkombination im Zusammenspiel mit der PDTS-Methode zur Adaption auf die portugiesische Sprache untersucht. Der in Kapitel 6 beschriebene fünflinguale Erkenner wird unter Verwendung unterschiedlich großer Datenmengen nach Portugiesisch adaptiert. Das verwendete portugiesische Adaptionsmaterial ist in Tabelle 7.7 zusammengestellt. Es werden zwischen 15 und 90 Minuten Sprachdatenmaterial von einer unterschiedlichen Anzahl von Sprechern verwendet. Das geschieht vor dem Hintergrund, daß es in einer Diktieranwendung schwieriger ist, wenige Sätze von vielen Sprecher als viele Sätze von wenigen Sprechern zu erhalten. Dagegen ist zu erwarten, daß die Adaption mit mehr Sprechern zu besseren Leistungen führt. Diese Erwartung soll mit der verschiedenen Zahl von Adaptionssprechern überprüft werden.

Neben den limitierten Sprachdaten aus Tabelle 7.7, bei denen davon ausgegangen wird, daß die zugehörigen Verschriftungen verfügbar sind, wird in den folgenden Experimenten vorausgesetzt, daß auch ein portugiesisches Aussprachewörterbuch und ein Sprachmodell bereitstehen.

Im besten Fall stünde für die Entwicklung eines Spracherkenners ausreichendes Sprachmaterial zur Verfügung. Als Idealziel wird daher die Wortfehlerrate angenommen, die der beste portugiesische Erkenner erreicht, der auf allen verfügbaren portugiesischen Trainingsdaten des GlobalPhone-Korpus, das sind etwa 16.5 Stunden, trainiert wurde. Das Idealziel liegt damit bei einer Wortfehlerrate von 19.0% (vgl. Abbildung 5.20, S. 144).

Die im folgenden beschriebenen Erkennungssysteme werden mit einer Systemidentifizierungsnummer Sn belegt, die als Referenz in der Übersichtsabbildung 7.5

Länge [min]	#Sprecher	Gesprochene Wörter
15	8	2050
25	8	3370
45	8	6283
45	16	6202
90	16	12649
90	32	12082
90	78 (alle)	11476

Tabelle 7.7: Portugiesische Adaptionsdaten

auf Seite 207 verwendet werden, in der alle erzielten Ergebnisse einander gegenübergestellt sind. In der Abbildung werden einige Abkürzungen verwendet. Das Kürzel „ÜT“ bezieht sich auf den überkreuzsprachlichen Vergleich und „Boot“ auf das Bootstrapping-Verfahren. „Vit“ bezeichnet das Training nach dem Viterbi-Algorithmus, „MLLR“ das im nächsten Abschnitt 7.5.2 erläuterte Maximum Likelihood Linear Regression-Verfahren. Als Abkürzung für Kontextentscheidungsbäume wird „ML“ für den fünflingualen Kontextentscheidungsbaum verwendet, „CI“ für eine kontextunabhängige Modellierung, „PO“ für einen Entscheidungsbaum, der ausschließlich auf portugiesischen Daten gelernt wurde und „PDTS“ für einen adaptierten multilingualen Entscheidungsbaum, der durch die PDTS-Methode entstanden ist.

7.5.1 Portierungstechniken

Die ersten portugiesischen Erkenner werden durch den überkreuzsprachlichen Transfer des fünflingualen Erkenners entwickelt. System S1 entsteht aus dem multilingualen kontextabhängigen System ML-MIX3000 (vgl. Abschnitt 6). System S2 entsteht durch den überkreuzsprachlichen Transfer des kontextunabhängigen fünflingualen Erkenners. Mit dem besseren der beiden Erkenner wurden initiale Labels von 15 Minuten portugiesischer Sprache geschrieben. Tabelle 7.8 zeigt den Effekt, den die

System	Portierungstechnik	WE CI	WE CD	Δ	
S2 / S1	Überkreuzsprachlicher Transfer	69.1	72.0	17.4	30.7
S4 / S6	Adaption	57.1	49.9	-	6.8
S3	Bootstrapping	-	46.5		

Tabelle 7.8: Vergleich der Portierungstechniken [WE in %]

Adaption bzw. das Training auf diesen initialen Labels hat, jeweils für die kontextunabhängigen und kontextabhängigen multilingualen akustischen Modelle. Die Ergebnisse der aus dem überkreuzsprachlichen Transfer hervorgegangenen Erkenner S1 und S2 können durch die Adaption auf den initialen Labels um etwa 30% verbessert werden. Die Resultate der adaptierten Erkenner S4 und S6 zeigen, daß sich die positiven Effekte der Kontextmodellierung im multilingualen Fall nicht auswaschen, wie dies im monolingualen Fall beobachtet worden war. Dieses Ergebnis deutet auf eine gewisse Robustheit der multilingualen Modelle gegenüber Kontextwechseln hin. Durch das Bootstrapping- anstelle des Adaptionsverfahren erreicht man eine Leistungssteigerung um 6.8%. Da der wesentliche Unterschied zwischen S3 und S6 im Kontextentscheidungsbaum liegt, der beim Bootstrapping auf den 15 Minuten Sprache gelernt wird, ist zu vermuten, daß eine Adaption der Kontextbäume des adaptierten Systems Gewinne erwarten läßt.

7.5.2 Trainingsmethoden

Unter dem Namen Maximum Likelihood Linear Regression (MLLR) wurde von Legetter und Woodland [LW95] eine Methode eingeführt, die die akustischen Modelle auf einen Testsprecher adaptiert, um durch die resultierenden, besser passenden Modelle die Fehlerrate zu verringern. Die Methode geht davon aus, daß die akustischen Modelle durch Normalverteilungen modelliert werden.

$$L = \sum_{t=1}^{T} (\boldsymbol{x} - (A\boldsymbol{\mu}_s + \mathbf{b}))^T \Sigma_s^{-1} (\boldsymbol{x} - (A\boldsymbol{\mu}_s + \mathbf{b})) \tag{7.2}$$

wobei $\boldsymbol{x}$ der Beobachtungsvektor zum Zeitpunkt t ist, s der diesem Zeitpunkt zugeordnete Zustand und $\boldsymbol{\mu}_s$ der Mittelwertsvektor sowie Σ_s die Kovarianz des Zustands s. Die Transformationsmatrix A und der Verschiebungsvektor $\mathbf{b}$ werden so gewählt, daß die Transformation $\boldsymbol{\mu} \rightarrow A\boldsymbol{\mu} + \mathbf{b}$ die Likelihood L auf dem Adaptionsmaterial maximiert. Die MLLR ist sehr leistungsfähig, weil die Transformation auf *alle* Mittelwertvektoren angewendet wird, nicht nur auf diejenigen, die im Adaptionsmaterial beobachtet wurden.

System	Trainingsmethode	WE	Δ
S5	Viterbi	52.2	+4.4%
S6	MLLR	49.9	

Tabelle 7.9: Vergleich der Trainingsmethoden [WE in %]

Das ursprünglich für die Adaption auf unbekannte Sprecher konzipierte Verfahren wird hier zur Adaption auf unbekannte Sprachen angewendet. Im folgenden Experiment wird die MLLR-Methode mit dem herkömmlichem Viterbi-Training verglichen.

Tabelle 7.9 zeigt das erzielte Ergebnis. Entsprechend den Erwartungen übertrifft MLLR die Leistungen von Viterbi-Training. Im vorliegenden Fall mit 15 Minuten Sprachmaterial ergibt sich eine Verbesserung um 4.4%.

7.5.3 Anwendung von PDTS

Schließlich wird untersucht, wie sich das im Abschnitt 7.3.2 vorgestellte Verfahren zur Spezialisierung gelernter Kontextentscheidungsbäume (PDTS) auf die Erkennungsleistung auswirkt. In Tabelle 7.10 werden die Resultate von PDTS (S10) einerseits mit dem Bootstrapping-Verfahren (S9) verglichen, bei dem der Kontextentscheidungsbaum ausschließlich auf dem Adaptionsmaterial berechnet wird sowie andererseits mit dem System S8, das aus der einfachen MLLR-Adaption der Modelle des bestehenden multilingualen Systems ML-MIX mit multilingualem Kontextentscheidungsbaum ohne PDTS-Adaption entsteht. Tabelle 7.10 enthält die Erkennungsleistungen jeweils nach dem Training bzw. der Adaption mit 15 Minuten bzw. 25 Minuten Sprache. In beiden Fällen ist das Bootstrapping-Verfahren der Modelladaption ohne PDTS überlegen. Stehen mit 25 Minuten mehr Daten zur Portierung zur Verfügung, zeigt das Bootstrapping-Verfahren erwartungsgemäß eine höhere Leistungsdifferenz.

System	Verfahren	15 Min initiale Labels	25 Min gute Labels	Δ	
S6/S8	ML-Baum	49.9	40.6	+6.8	+19.2%
S3/S9	Boot	46.5	32.8	-	+11.9%
S10	PDTS	-	28.9		

Tabelle 7.10: Die PDTS-Methode [WE in %]

Durch die Adaption des multilingualen Kontextentscheidungsbaumes mit der PDTS-Methode kann die Leistung des Bootstrapping-Verfahrens um 11.9% verbessert werden. Damit ist gezeigt, daß durch die PDTS-Methode das Wissen anderer Sprachen erfolgreich auf eine neue Sprache portiert werden kann.

7.5.4 Menge und Qualität der Adaptionsdaten

Im folgenden Experiment wird untersucht, welchen Leistungszuwachs die Verbesserung der Labelqualität bringt. Es werden initiale von guten Labels unterschieden. Die initialen Labels sind mit dem multilingualen System S2 erstellt, das aus dem überkreuzsprachlichen Transfer hervorging. Gute Labels werden als gegeben vorausgesetzt. Sie können entweder manuell von Experten erstellt werden oder durch einen existierenden Erkenner erzeugt werden. In diesem Fall sind sie durch den in Kapitel 5 beschriebenen portugiesischen Erkenner geschrieben worden.

System	Adaptionsdaten	WE	Δ
S6	100 initiale Labels	49.9	+13.2%
S7	100 gute Labels	43.3	

Tabelle 7.11: Vergleich der Qualität der Adaptionsdaten [WE in %]

Der Vergleich der Erkennungsergebnisse in Tabelle 7.11 zeigt, daß sich durch die Verbesserung der Labelqualität eine Steigerung der Erkennungsleistung um 13.2% ergibt.

Zuletzt wird untersucht, wie groß der Leistungszuwachs durch mehr Adaptionsmaterial und mehr Adaptionssprecher ist. Die Verdopplung der Adaptionsmenge ergab eine Verbesserung von 16.6% bzw. 12.5%. Demgegenüber stehen 7.1%, die durch die Verdopplung der Anzahl an Adaptionssprechern erreicht wird. Die weitere Hinzunahme von Sprechern wirkt sich nicht mehr leistungssteigernd aus. Die gleiche Beobachtung wurde von [ZC98] gemacht, der auf monolingualen Erkennersystemen den Einfluß der Sprecherzahl bei der Adaption untersucht hat.

System	Methode	WE	Δ
S10	200 von 8 Sprechern	28.9	
			+16.6
S11	400 von 8 Sprechern	24.1	
			+7.1
S12	400 von 16 Sprechern	22.4	
			+12.5
S14	800 von 16 Sprechern	19.6	

Tabelle 7.12: Vergleich der Menge der Adaptionsdaten [WE in %]

7.6 Zusammenfassung

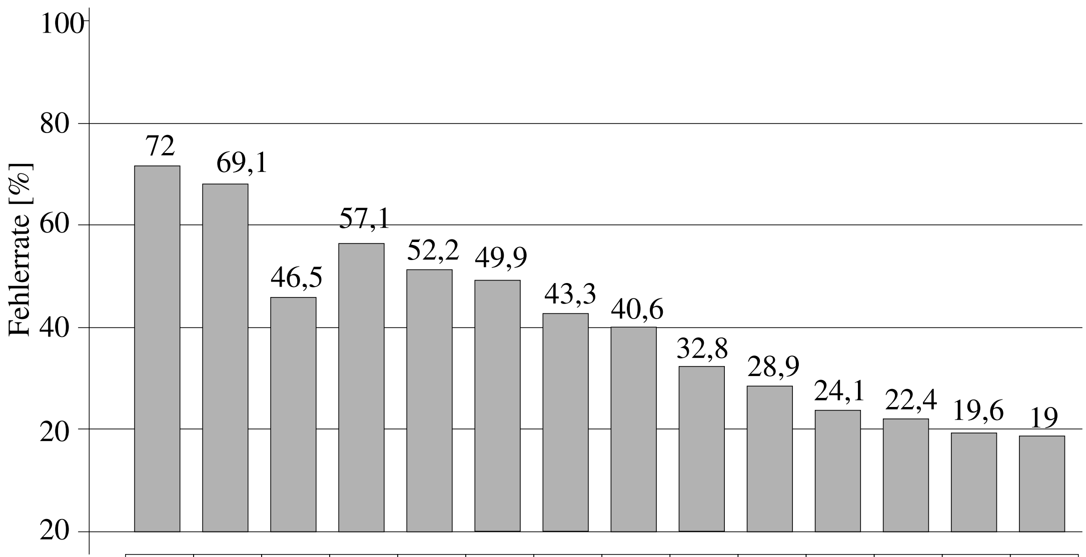
Bislang gibt es erst sehr wenige Forschungsarbeiten zum Thema **Portierung akustischer Modelle auf neue Sprachen**. Das liegt zum einen daran, daß es bisher keine geeigneten Datenbasen gab, anhand derer multilinguale und monolinguale Systeme in vielen Sprachen gebaut werden konnten. Zum anderen sind die Techniken zur Entwicklung multilingualer Modelle noch wenig erforscht. Mit der gesammelten GlobalPhone-Datenbasis, den erstellten monolingualen Systemen in 10 Sprachen und den eingeführten neuen Verfahren zur Kombination multilingualer akustischer Modelle sind mit dieser Arbeit alle Voraussetzungen geschaffen worden, um die Portierung auf neue Sprachen zu untersuchen und in die Praxis umzusetzen. Bei der Untersuchung werden je nach vorhandenem Datenmaterial in der neuen Sprache drei Aspekte der Portierung unterschieden: Der überkreuzsprachliche Transfer, die Adaption und das Bootstrapping-Verfahren.

Am Beispiel der in der Spracherkennung bisher wenig erforschten Sprache Schwedisch wurde untersucht, wie gut sich monolingualer Systeme zum **überkreuzsprachlichen Transfer** eignen. Es stellte sich heraus, daß die Erkennungsraten der sieben monolingualen Spracherkenner auf Schwedisch stark variieren. Je nachdem, von welcher Sprache der überkreuzsprachliche Transfer auf Schwedisch realisiert wurde, ergeben sich bis zu 22% relative Leistungsunterschiede der resultierenden Spracherkenner. Das bedeutet, daß es einen signifikanten Unterschied ausmacht, welche der zur Auswahl stehenden Sprache zur Portierung verwendet wird. Für die multilingualen Modelle zeigte sich, daß sie in jedem Fall besser sind als alle monolingualen Systeme. Im Mittel ergibt sich eine 13.4%-ige Verbesserung der Erkennungsleistung der multilingualen Modelle zum überkreuzsprachlichen Transfer auf die schwedische Sprache. Durch die Anwendung der multilingualen Modelle ist daher kein a-priori Wissen darüber erforderlich, welche Sprache beim Transfer die besten Ergebnisse erzielt.

Zur Bestimmung geeigneter **Abbildungen der multilingualen Phoneme** auf neue Sprachen werden heuristische und datengetriebene Ansätze vorgestellt und miteinander verglichen. Die Phonemabbildungen bieten die Möglichkeit der geeigneten Initialisierung akustischer Modelle, der Auswahl einer geeigneten Teilmenge des globalen Phoneminventars und damit der gemeinsamen Nutzung von Daten sowie der automatischen Anpassung vorhandener Aussprachewörterbücher. Experimente auf der schwedischen Sprache zeigen, daß mit den datengetriebenen Verfahren die Erkennungsfehler bei der Portierung um bis zu 7.4% gegenüber dem heuristischen IPA-Ansatz reduziert werden können.

Bisher sind noch keine Arbeiten bekannt, die bei der Portierung auf neue Sprachen das Problem einer geeigneten Kontextmodellierung betrachten. Da die Modellierung breiter Kontexte im monolingualen Fall zu signifikanten Verbesserungen führt, sollen diese Verbesserungen auch bei der Portierung auf neue Sprachen nicht verloren gehen. Die Limitierung der Daten läßt aber keine Neuberechnung des Kontextentscheidungsbaumes auf der neuen Sprache zu. Um dennoch passende Kontextbäume bereitzustellen, wurde mit **PDTS (Polyphone Decision Tree Specialization)** eine Methode eingeführt, die eine Adaption multilingualer Kontextentscheidungsbäume auch dann ermöglicht, wenn nur wenig Datenmaterial in der neuen Sprache verfügbar ist.

Der große Nutzen der PDTS-Methode wurde in zahlreichen Experimenten zur **Adaption** auf der portugiesischen Sprache gezeigt. Die Abbildung 7.5 faßt die erreichten Wortfehlerraten für alle beschriebenen Experimente zusammen. Das erste System (S1 siehe Abbildung 7.5) entsteht durch den überkreuzsprachlichen Transfer und erreicht 69.1% Wortfehlerrate. Mit diesem System S1 werden initiale Labels von 15 Minuten Sprache geschrieben und zur Adaption verwendet. Die Adaption des kontextunabhängigen Systems ($\rightarrow$ S4) und des kontextabhängigen Systems durch einfaches Viterbitraining ($\rightarrow$ S5) sowie durch MLLR ($\rightarrow$ S6) ergeben signifikante Leistungsgewinne, die bei MLLR mit 27% am höchsten ausfallen. Verwendet man



System	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14
Daten	0		15 Minuten					25 Minuten			45	8 16	90	16,5h
Qualität	-		initiale Labels				gute Labels							
Methode	ÜT		Boot	Mllr	Vit	MLLR			Boot	MLLR				Boot
Baum	ML	CI	PO	CI	ML				PO	PDTS				PO

Abbildung 7.5: Adaption auf Portugiesisch

die Labels zum Bootstrapping des portugiesischen Systems ($\rightarrow$ S3), dann führt dies im direkten Vergleich mit S4 bis S6 zum besten Ergebnis. Dieses zeigt, daß ohne eine Anpassung der Kontextentscheidungsbäume die Bootstrapping-Technik bereits mit 15 Minuten Sprache die beste Lösung bietet. Die erzielte Wortfehlerrate von 46.5% ist allerdings noch sehr unzureichend.

Durch eine Verbesserung der Labelqualität der Labelqualität (S6 $\rightarrow$ S7) und die Verdoppelung des Trainingsmaterials (S7 $\rightarrow$ S8) werden wie zu erwarten signifikante Leistungsgewinne erzielt. Die Erkennungsleistung des Bootstrapping-Verfahrens kann damit von 46.5% auf 32.8% Wortfehlerrate (S3 $\rightarrow$ S9) gesteigert werden.

Damit sind die Möglichkeiten der bisher gängigen Methoden zur Portierung allerdings ausgeschöpft. Durch das neuartige PDTS-Verfahren wird nun statt einer Neuberechnung der Kontextentscheidungsbäume der bisherige multilinguale Entscheidungsbaum adaptiert. Im Vergleich zur Bootstrapping-Technik kommt es dadurch zu einer signifikanten Verbesserung der Erkennungsleistung um fast 12% (S9 $\rightarrow$ S10). Dieses Ergebnis beweist, daß durch PDTS das Wissen aus mehreren Sprachen zur Portierung auf neue Sprachen gewinnbringend eingesetzt werden kann.

Eine weitere Verdoppelung der Adaptionsdaten resultiert in einer Steigerung auf 24.1% Wortfehlerrate (S10 $\rightarrow$ S11). Durch Adaption auf mehr Sprechern bei gleichbleibender Sprachdatenmenge kann die Adaptionsleistung weiter verbessert werden

(S11 $\rightarrow$ S12). Schließlich kann mit 90 Minuten Sprachdaten unter Anwendung der PDTS-Methode ($\rightarrow$ S13) eine Wortfehlerrate von 19.6% erreicht werden. Das beste portugiesische System (S14) erzielt mit 19.0% Wortfehlerrate nur 3% bessere Erkennungsleistungen, obwohl es mit insgesamt 16.5 Stunden portugiesischem Sprachmaterial trainiert wurde. Während das Training dieses Erkenners mehrere Tage Rechenzeit beansprucht, verläuft der Portierungsvorgang mit PDTS vollautomatisch und benötigt nur 3 bis 5 Stunden Rechenzeit auf einer SUN Ultra-2 mit einem 300 MHz-Prozessor.

Kapitel 8

Der GlobalPhone-Demonstrator

Auf der Basis der monolingualen GlobalPhone-Erkenner und der kombinierten multilingualen akustischen Modelle ist im Rahmen dieser Arbeit ein Demonstrationssystem für multilinguale Applikationen implementiert worden [Ras00]. Das System akzeptiert derzeit gelesene Äußerungen in den neun Sprachen Chinesisch, Deutsch, Englisch, Französisch, Japanisch, Koreanisch, Kroatisch, Spanisch und Türkisch. Die gesprochene Sprache wird vom System automatisch identifiziert, die Eingabeäußerung dekodiert und der erkannte Text im entsprechenden Schriftsystem dargestellt. Die Systemstruktur ist modular aufgebaut, so daß das Demonstrationssystem jederzeit um neue Sprachen erweitert werden kann.

Der GlobalPhone-Demonstrator kann in drei Systemmodi betrieben werden:

1. Dedizierte Sprachenidentifizierung: Das System wird mit einer beliebigen der neun Sprachen konfrontiert und identifiziert sie. Die identifizierte Sprache wird dem Benutzer durch das Symbol einer Landesflagge angezeigt.

2. Monolinguale Diktiererkennung: Durch Anklicken des Symbols der Landesflagge wählt der Benutzer einen Diktiererkenner in einer der neun möglichen Sprachen aus. Der ausgewählte Erkenner dekodiert die Eingabeäußerung, die in der spezifizierten Sprache erwartet wird, und gibt den erkannten Text aus.

3. Multilinguale Diktiererkennung: Das System wird mit einer beliebigen der neun Sprachen konfrontiert und entscheidet automatisch, welcher der neun Erkenner für die Eingabe zuständig ist. Die Eingabe wird mit dem zuständigen Erkenner dekodiert und ausgegeben.

Implementierung

JRTk ist eine objektorientierte, in *C* implementierte Entwicklungsumgebung. Ein Spracherkennungssystem wird durch aufeinander aufbauende Objekte und deren Kommunikation realisiert. Die Handhabung der verwendeten Objekte geschieht über einen integrierten Tcl/TK-Interpreter. Dadurch wird eine sehr flexible und interaktive Programmierung ermöglicht. Die graphische Benutzeroberfläche des GlobalPhone-Demonstrators, die in Abbildung 8.1 gezeigt ist, wurde in JAVA realisiert. Die Kommunikation der JAVA-Oberfläche mit JRTk geschieht über die umgeleiteten Ein- und Ausgabeströme von JRTk. Von der JAVA-Oberfläche aus können wichtige Parameter zur Kontrolle der elementaren Objekte und Datenstrukturen der Erkenner eingestellt werden, die in einzelne Module gekapselt sind.

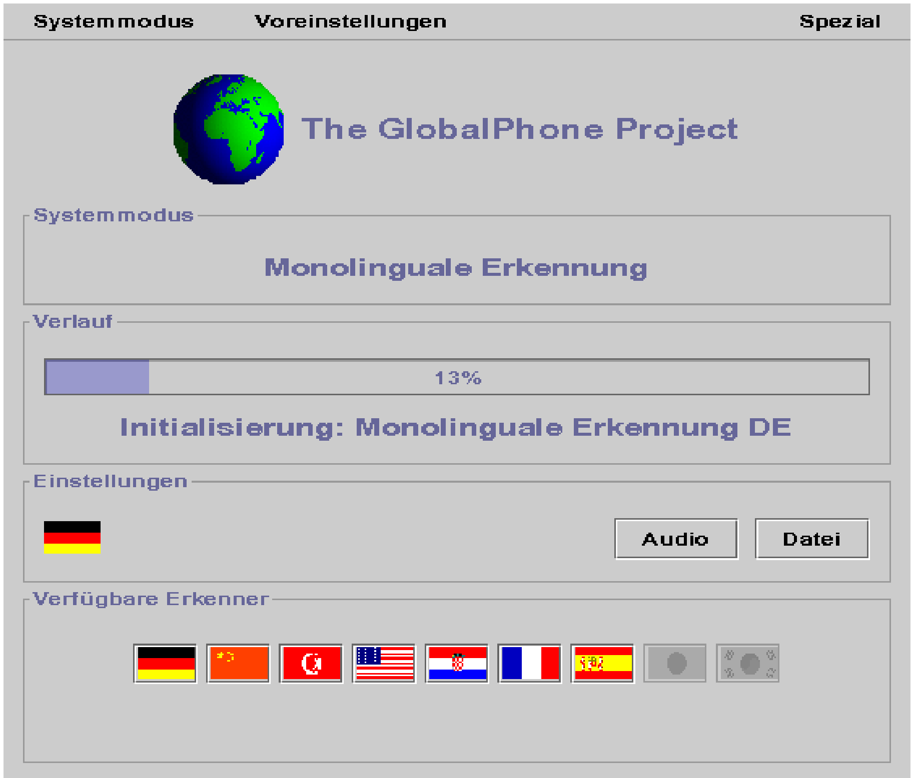


Abbildung 8.1: Graphische Benutzeroberfläche des GlobalPhone-Demonstrators

Internationalisierung

Einer der Gründe für die Wahl von JAVA war dessen Unterstützung von UNICODE zur Zeichendarstellung. Dadurch können problemlos die verschiedenen 8-bit und 16-bit-kodierten Schriftsysteme dargestellt werden. Insbesondere im Kontext multilingualer Applikationen wie Diktiererkennung ist dies ein wichtiger Bestandteil. Ein weiterer Standard, die JAVA-Internationalization (I18N), ermöglicht die multilinguale Beschriftung aller Komponenten der Oberfläche und deren Speicherung in sogenannten `Properties`-Dateien. Wählt der Anwender eine der neun Sprachen aus, dann können alle Beschriftungen der Benutzeroberfläche in diese Sprache umgesetzt und im zugehörigen Schriftsystem dargestellt werden. Dies ermöglicht die adäquate Bedienung des Demonstrators für Anwender ohne fremdsprachliche Kenntnisse.

Sprachenidentifizierung

Bereits in Abschnitt 6.4.1 wurden die möglichen Verfahren zur Sprachenidentifizierung und deren Vor- und Nachteile beschrieben. Von der nachgeschalteten LID wird in der Demonstration aus Laufzeitgründen kein Gebrauch gemacht. Die vorgeschaltete LID wird insbesondere zur dedizierten Sprachenidentifizierung verwendet. Dazu wird der in Abschnitt 6.4.1 beschriebene Ansatz mit sprachenseparaten Modellen verwendet. Im GlobalPhone-Demonstrator wurde als drittes Verfahren eine neue Lösung implementiert, die die Leistungsvorteile der nachgeschalteten LID mit den Geschwindigkeitsvorteilen der vorgeschalteten LID verbindet. Diese Lösung baut auf der Kommunikation zwischen allen beteiligten monolingualen Erkennern auf. Im Dekodierprozeß der Eingabe werden so sukzessive nichtzuständige Erkenner deaktiviert, wodurch gegenüber der nachgeschalteten LID Rechenzeit eingespart wird.

Kommunikationstruktur

Zur Realisierung der Kommunikation zwischen mehreren monolingualen Spracherkennungssystemen wurde ein übergeordnetes Suchobjekt eingeführt, das die Steuerung der einzelnen Suchobjekte der beteiligten Erkenner ermöglicht. Aufgabe dieser „Metasuche“ ist es, die Suchobjekte aller Sprachen mit der gleichen Eingabe zu starten, deren Ausgaben in definierten Intervallen miteinander zu vergleichen und die Dekodiervorgänge der nicht zuständigen Erkenner zu deaktivieren. Zu diesem Zweck kann der Vorgang der Dekodierung zu beliebigen Zeitpunkten unterbrochen und wieder aufgesetzt werden, ohne daß bereits dekodierte Segmente verloren gehen. Zu jedem Vergleichszeitpunkt werden diejenigen Erkenner deaktiviert, deren Ausgabebewertungen von der Bewertung des besten Erkenners weiter entfernt als ein definierter Schwellwert liegen.

Durch dieses hierarchische Ausschlußverfahren können schlecht passende Erkenner bereits sehr schnell abgeschaltet werden, was das Laufzeitverhalten signifikant gegenüber der nachgeschalteten LID verbessert. Dagegen verhindert man bei schwer zu identifizierenden Eingabeäußerungen eine verfrühte und dadurch möglicherweise fehlerhafte Entscheidung. Bei Entscheidungsunsicherheit werden automatisch entsprechend längere Segmente von den aktuell in Frage kommenden Erkennern dekodiert. Mit dieser kommunikationsbasierten hierarchischen Lösung konnte die Sprachenidentifizierungsrate gegenüber der vorgeschalteten LID um etwa 30% verbessert werden. Im Vergleich zur vorgeschalteten LID ist allerdings ein höherer Speicher- und Laufzeitaufwand erforderlich.

Der GlobalPhone-Demonstrator bietet eine hervorragende Ausgangsbasis für zukünftige Forschungsarbeiten, in denen beispielsweise die Auswirkungen multilingualer Aussprachewörterbücher und Sprachmodelle durch die Nutzung der Daten von vielen Sprachen exploriert werden können. Derartige Wissensquellen würden dann beispielsweise das Code-Switching erlauben. Der Demonstrator ist auf diese Form der multilingualen Erkennung bereits vorbereitet.

Kapitel 9

Zusammenfassung

Die Entwicklung eines Spracherkenners ist bisher mit einem sehr hohen Kosten- und Arbeitsaufwand verbunden. Der wesentliche Grund dafür sind die zum Training der akustischen Modelle notwendigen umfangreichen Datenbasen. In dieser Arbeit werden Methoden vorgestellt, die den Entwicklungsaufwand reduzieren. Dazu werden verschiedene Aspekte der multilingualen Spracherkennung bearbeitet. Zuerst wird eine umfangreiche multilinguale Datenbasis erstellt (Kapitel 4). Auf diesen Daten werden monolinguale Erkennungssysteme in 10 Sprachen entwickelt und relevante Sprachinformationen extrahiert (Kapitel 5). Davon ausgehend werden Techniken zur Kombination multilingualer akustischer Modelle entworfen und evaluiert (Kapitel 6). Die multilingualen Modelle bilden die Ausgangsbasis zur schnellen Portierung eines Spracherkenners auf neue Sprachen, ohne dazu umfangreiches Datenmaterial in diesen Sprachen vorauszusetzen (Kapitel 7).

9.1 Die wichtigsten Ergebnisse und Beiträge

Gegliedert nach den drei zu lösenden Problemen *Sammlung und Extraktion relevanter Sprachinformationen aus vielen Sprachen*, *Entwicklung multilingualer Spracherkennungskomponenten*, *Konzeption und Implementierung von Portierungsverfahren* werden im folgenden die wichtigsten Ergebnisse und Beiträge dieser Arbeit zusammengefaßt. Detaillierte Beschreibungen der im einzelnen erzielten Ergebnisse befinden sich in den Zusammenfassungen der jeweiligen Abschnitte 4.6, 5.6, 6.5 und 7.6.

Multilinguale Datenbasis. Zur Lösung des ersten Problems wurde zunächst das multikulturelle Projekt GlobalPhone initiiert, in dem eine multilinguale Datenbasis großer Wortschätze in dreizehn Sprachen gesammelt wird. Die GlobalPhone-Datenbasis deckt 9 der 12 wichtigsten Weltsprachen ab und enthält transkribierte Sprache von über 1200 Sprechern. Die Sammlung dieser Datenbasis ist dem heutigen

Trend zur Sammlung multilingualer Datenbasen voraus gewesen und hat weltweit sehr großes Interesse hervorgerufen.

Monolinguale Spracherkennung in 10 Sprachen. Eine große Herausforderung dieser Arbeit war die systematische Entwicklung von Spracherkennern in vielen verschiedenen Sprachen nach einheitlichen Gesichtspunkten. Die Tatsache, daß dies in 10 verschiedenen Sprachen für große Wortschatzerkenner realisiert wurde, ist nach Kenntnis der Literatur einmalig.

Im Vordergrund der Systementwicklung steht daher die Frage nach Effizienz und Automatisierung des gesamten Entwicklungsprozesses. Es wurde eine Technik zur schnellen Initialisierung von Spracherkennern eingeführt. Dabei zeigte sich, daß die Initialisierung mittels eines Pools von Phonemmodellen aus mehreren Sprachen anderen Verfahren überlegen ist. Die Automatisierung schließt auch die Erstellung notwendiger Wissensquellen wie Ausspracheworterbücher und Sprachmodelle mit ein. Für fast alle Sprachen wurden Graphem-zu-Phonem-Tools zur Aussprachegenerierung entwickelt, die den Einsatz neuer Techniken wie HDLA ermöglichten. Erzielte Lösungen, wie die datengetriebene Bestimmung geeigneter sprachlicher Modellierungseinheiten für die koreanische Sprache, haben gezeigt, daß leistungsäquivalente Systeme auch ohne den Einsatz von muttersprachlichen Experten gebaut werden können.

Auf der Basis der automatisch generierten Wissensquellen und etwa 15 Stunden Audiodaten pro Sprache wurden monolinguale Spracherkennungssysteme in 10 Sprachen entwickelt. Sie erreichen eine Phonemfehlerrate zwischen 34% und 47% sowie eine Wortfehlerrate zwischen 10% und 20%. Diese Ergebnisse bestätigen die Annahme, daß sich die verwendeten klassischen Algorithmen der Spracherkennung auf alle modellierten Sprachen übertragen lassen. Allerdings zeigen die Erfahrungen dieser Arbeit, daß ein erheblicher Entwicklungsaufwand zur Behandlung sprachenspezifischer Besonderheiten notwendig ist. Dazu wurden diverse Algorithmen und Techniken entwickelt, die von der Romanisierung und Segmentierung ideographischer Schriftsysteme im Japanischen und Chinesischen über die Modellierung der Tonalsprache Chinesisch bis hin zur Behandlung agglutinierender und stark flektierender Sprachen wie Koreanisch, Türkisch und Kroatisch reichen.

Multilinguale Spracherkennung. Aufbauend auf den monolingualen Basissystemen wurden Verfahren implementiert und evaluiert, die die akustischen Modelle vieler Sprachen zu multilingualen akustischen Modellen kombinieren. Dazu wurde auf zwölf Sprachen ein universelles Phoneminventar entwickelt, das die lautliche Vielfalt aller beteiligten Sprachen widerspiegelt. Die Kombination zu multilingualen Modellen wurde auf die kontextabhängige Modellierung erweitert. In der sprachenseparaten Kontextmodellierung werden dazu die Daten verschiedener Sprachen separat gehalten. In der sprachenvermischten Modellierung werden die Kontextmodelle aller beteiligten Sprachen durch Mischen der Daten gemeinsam trainiert, wodurch das sprachenspezifische Wissen verloren geht. In der sprachenmarkierten Modellierung entscheiden die Daten, welche Kontexte über Sprachen hinweg gemeinsam modelliert werden können und welche separat bleiben sollten. Die Analyse des beim Ballen der Modelle entstandenen Kontextentscheidungsbaumes ergab, daß bei 5 Sprachen der größte Anteil an Spracheninformationen nach etwa 2000 Aufspaltungen ausdifferenziert ist. Die resultierenden Erkenner werden zur Lösung von drei Probleme verwendet: Zur simultanen multilingualen Erkennung, zur Sprachenidentifizierung und zur Portierung auf neue Sprachen.

Die Ergebnisse legen nahe, daß je nach Anwendungsgebiet unterschiedliche Kombinationsmethoden verwendet werden sollten. Zur multilingualen Erkennung zeigen die sprachenmarkierten Modelle die besten Leistungen. Dies wird auf den Erhalt der Spracheninformation in den Kontextmodellen zurückgeführt. Im Mittel kommt es zu einer Verbesserung von 15% gegenüber der sprachenvermischten Modellierung. Gegenüber der auf die jeweilige Sprache spezialisierten monolingualen Erkenner kommt es nur zu einer Leistungseinbuße von 5.8% Prozent bzw. 1.66 Prozentpunkten. Dieses Ergebnis spiegelt die allgemeine Erfahrung wieder, daß spezifische Modelle zu besseren Leistungen führen als generalisierte. Zur Sprachenidentifizierung sind die sprachenseparaten Modelle besonders geeignet, da es bei dieser Aufgabe auf die maximale Separation von Sprachen ankommt. Es werden mit diesen Modellen Sprachenidentifizierungsraten von 84% auf 8 Sprachen erreicht. Zur Portierung auf neue Sprachen erzielen die sprachenvermischten Modelle die besten Resultate. Dies resultiert aus der robusteren Modellierung gegenüber fehlenden Kontextüberlappungen.

Portierung auf neue Sprachen. Die sprachenvermischten multilingualen Kontextmodelle bilden die Ausgangsbasis für die Experimente zur Portierung eines multilingualen Erkenners auf Sprachen, die bisher in der Spracherkennung wenig erforscht wurden. Je nach dem Anteil, den die Daten der neuen Sprache am gesamten Trainingsmaterial ausmachen, werden Verfahren zum überkreuzsprachlichen Transfer, zur Adaption und zum Bootstrapping unterschieden.

Zur Bestimmung geeigneter Abbildungen der multilingualen Phoneme auf neue Sprachen werden heuristische und datengetriebene Ansätze vorgestellt und miteinander verglichen. Die Phonemabbildungen bieten die Möglichkeit der geeigneten Initialisierung akustischer Modelle, der Auswahl einer geeigneten Teilmenge des globalen Phoneminventars und damit der gemeinsamen Nutzung von Daten sowie der automatischen Anpassung vorhandener Aussprachewörterbücher. Experimente auf der schwedischen Sprache zeigen, daß mit den datengetriebenen Verfahren die Erkennungsfehler bei der Portierung um bis zu 7.4% gegenüber dem heuristischen Ansatz reduziert werden können.

Mit der Polyphone Decision Tree Specialization Methode (PDTS) wird eine neues Verfahren zur Spezialisierung gelernter Kontextentscheidungsbäume vorgestellt. Damit können multilinguale kontextabhängige Modelle mit sehr wenig Sprachmaterial auf neue Sprachen adaptiert werden. Experimente auf portugiesischer Sprache demonstrieren, daß ein mit 90 Minuten Sprachmaterial mittels der PDTS-Methode portierter Erkenner eine Wortfehlerrate von 19.6% liefert. Das beste portugiesische System erreicht nur 3% bessere Erkennungsleistungen, obwohl es mit insgesamt 16.5 Stunden Sprachmaterial trainiert wurde. Während das Training dieses Erkenners mehrere Tage Rechenzeit beansprucht, verläuft der Portierungsvorgang mit PDTS vollautomatisch und benötigt nur 3 bis 5 Stunden Rechenzeit auf einer SUN Ultra-2 mit einem 300 MHz-Prozessor.

In dieser Arbeit konnte gezeigt werden, daß multilinguale akustische Modelle eine effiziente Portierung von Erkennungssystemen auf neue Sprachen ermöglichen. Der Einsatz multilingualer Modelle spart Kosten, weil durch das Ausnutzen der Daten anderer Sprachen der Umfang notwendiger Trainingsdaten in der neuen Sprache drastisch verringert werden kann und spart Entwicklungszeit, weil eine Adaption mit kleinen Datenmengen schneller durchgeführt werden kann als die komplette Neuentwicklung eines Systems.

9.2 Ausblick

An dieser Stelle sollen einige interessante Aspekte der multilingualen Spracherkennung aufgegriffen werden, die von dieser Arbeit profitieren können.

Nahezu alle gängigen und erfolgreichen Spracherkenner stützen sich auf die phonembasierte Modellierung mit HMMs. Die rasante Leistungsentwicklung der letzten 20 Jahre hat diesem Konzept recht gegeben. Daher haben sich die Methoden der vorliegenden Arbeit auf diese Modellierungsart konzentriert. Allerdings zeigen sich mit zunehmenden Anforderungen an die Robustheit von Spracherkennungssystemen Grenzen der adäquaten Modellierung von Sprache durch Phoneme. Phonemmodelle werden anhand von immer mehr Daten immer stärker auf die Erkennungsaufgabe spezialisiert, wodurch die Generalisierungsfähigkeit der Modelle abnimmt. Dieses Problem betrifft auch die multilinguale akustische Modellierung, zumal nach Meinung einiger Forscher Phoneme kontrastierend definiert und sprachenabhängig sind. Eine interessante Möglichkeit bietet daher ein nicht phonembasierter Ansatz, in dem distinktive Merkmale nach dem Vorbild der Sprachproduktionsmechanismen phonologisch definiert werden. Sprache wird dann nicht mehr durch Phoneme modelliert, sondern durch die zeitliche Überlappung distinktiver Merkmale beschrieben. Da bei allen Menschen die Produktion von Sprache gleich ist, besteht die berechtigte Hoffnung, daß sich alle Sprachen durch eine sehr beschränkte Menge von distinktiven Merkmalen beschreiben lassen. Voraussetzung wäre allerdings, daß sich die merkmalsbasierte Modellierung als leistungsfähig und der phonembasierten Modellierung zumindest als ebenbürtig erweist.

In der vorliegenden Arbeit wurde gezeigt, daß sich Wissen aus vielen Sprachen erfolgreich auf neue Sprachen übertragen läßt. Die Arbeit konzentrierte sich auf akustische Modelle, weil die zu deren Training notwendigen Sprachdaten den kostenintensivsten Teil einer Datensammlung ausmachen und bot erste Lösungen zur Übertragung vorhandener Aussprachewörterbücher. In Zukunft könnte aber auch die Portierung auf Sprachen erwünscht sein, für die es keine oder nur sehr limitierte Audio- und Textquellen gibt. In solchen Fällen erfordert eine vollständig automatisierte Portierung eines Spracherkenners darüber hinaus Methoden zum automatischen Erlernen von Aussprachen und zur Erstellung zuverlässiger Sprachmodelle mit limitiertem Textmaterial. Für Alphabetschriften wären beispielsweise generalisierte Graphem-zu-Phonem-Abbildungsregeln denkbar, die aus den Aussprachewörterbüchern bereits vorhandener Sprachen gelernt werden könnten. Diese Methode könnte gänzlich ohne Sprachdaten der Zielsprache auskommen. Für auf Wortgrenzen markierte Sprachdaten könnte man auf der Basis multilingualer Phonemsets Aussprachen erlernen.

In der Sprachmodellierung besteht bisher die Notwendigkeit riesiger Textkorpora, die nur in einigen hundert aller 4500 Sprachen vorhanden sind. Das Datenproblem könnte umgangen werden, wenn Sprachmodelle in einen multilingualen, generischen Teil und einen sprachenspezifischen, regelbasierten Teil aufgetrennt würden. Der multilinguale generische Teil enthält Wissen über sprachenunabhängige Konzepte (Beispiel Interlingua) und kann auf vorhandenen Massendaten vieler Sprachen robust geschätzt werden. Der sprachenspezifische regelbasierte Teil würde speziell für die Zielsprache erstellt und benötigte nur wenige Textmuster. Auf diese Weise würde die Portierung eines Spracherkenners nahezu automatisch erfolgen und auch für solche Sprachen durchführbar sein, in denen nicht viele Sprach- und Textdaten verfügbar sind.

Abbildung 9.1: Mit GlobalPhone wäre Hägar das nicht passiert

Literaturverzeichnis

[AAB+96] **U. Ackermann, B. Angelini, F. Brugnara, M. Frederico, D. Giuliani, R. Gretter, G. Lazzari und H. Niemann**. Speedata: Multilingual Spoken Data-entry. In *Proc. International Conference on Spoken Language Processing*, S. 2211–2214, Philadelphia, PA, Oktober 1996. IEEE.

[AAB+97] **U. Ackermann, B. Angelini, F. Brugnara, M. Frederico, D. Giuliani, R. Gretter und H. Niemann**. Speedata: A Prototype for Multilingual Spoken Data-entry. In *Proc. EUROSPEECH, European Conference on Speech Communication and Technology*, S. 355–358, Rhodos, Griechenland, September 1997. ESCA.

[ACC97] **Y. P. Alfred, L. Chan und P. Ching**. Automatic Recognition of Continuous Cantonese Speech with Very Large Vocabulary. In *Proc. EUROSPEECH, European Conference on Speech Communication and Technology*, Band 3, S. 1551–1554, Rhodos, Griechenland, September 1997. ESCA.

[ACC98] **ACCeSS**. *Automated Call Center Through Speech Understanding System (EU-Projekt)*. Internet, http://www.wcl.ee.upatras.gr/access/access.html, September 1998.

[AD97] **O. Andersen und P. Dalsgaard**. Language-identification based on Cross-Language Acoustic models and Optimised Information Combination. In *Proc. EUROSPEECH, European Conference on Speech Communication and Technology*, S. 67–70, Rhodos, Griechenland, September 1997. ESCA.

[AD99] **M. Adda-Decker**. Towards Multilingual Interoperability in Automatic Speech Recognition. In *Proc. ESCA-NATO Tutorial and Research Workshop on Multi-lingual Interoperability in Speech Technology*, Seite (keine Seitennummerierung), Leusden, Niederlande, September 1999. ESCA.

[ADB93] **O. Andersen, P. Dalsgaard und W. Barry**. Data-Driven Identification of Poly- and Monophonemes for four European Lanugages. In *Proc. EUROSPEECH, European Conference on Speech Communication and Technology*, S. 759–762, Berlin, September 1993. ESCA.

[ADB94] **O. Andersen, P. Dalsgaard und W. Barry**. On the Use of Data-Driven Clustering Techniques for Language Identification of Poly- and Mono-phonemes for four European Languages. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Band 1, S. 121–124, Adelaide, 1994. IEEE.

[AHG+98] **M. Aretoulaki, S. Harbeck, F. Gallwitz, E. Nöth, H. Niemann, J. Ivanecky, I. Ipsic, N. Pavesic und V. Matousek**. SQEL: A Multilingual and Multifunctional Dialogue System. In *Proc. International Conference on Spoken Language Processing*, Sydney, November 1998.

[All88] **G. D. Allen**. The PHONASCII System. *Journal of the International Phonetic Association*, 18:9 – 25, 1988.

[Ara93] **T. Arai**. Automatic Language Identification using Sequential Information of Phonemes. Technical report, Department of Computer Science and Engineering, Oregon Graduate Institute of Science & Technology, August 1993.

[ARI98] **ARISE**. *Automatic Railway Information Systems for Europe (EU-Projekt)*. Internet, http://www.vecsys.fr/tap/arise_eng.html, September 1998.

[BAB99] **BABEL**. *EC Copernicus (EU-Projekt)*. Internet, http://midwich.reading.ac.uk/research/speechlab/BABEL, September 1999.

[BABC94] **K. Berkling, T. Arai, E. Barnard und R. Cole**. Analysis of Phoneme-based Features for Language Identification. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Band 1, S. 289–292, Adelaide, Oktober 1994. IEEE.

[BBdSM90] **L. Bahl, P. Brown, P. de Souza und R. Mercer**. A Tree-based Statistical Language Model for Natural Language Speech Recognition. In *Readings in Speech Recognition*, S. 507–514. Morgan Kaufman Publishers, San Mateo, CA, 1990.

[BBH+95] **J. Barnett, P. Bamberg, M. Held, J. Huerta, L. Manganaro und A. Weiss**. Comparative Performance in Large-Vocabulary Isolated-Word Recognition in Five European Languages. In *Proc. EUROSPEECH, European Conference on Speech Communication and Technology*, S. 189–192, Madrid, September 1995. ESCA.

[BBH+99] **B. Byrne, P. Beyerlein, J. M. Huerta, S. Khudanpur, B. Marthi, J. Morgan, N. Peterek, J. Picone und W. Wang**. Towards Language Independent Acoustic Modeling. Technical report, Johns Hopkins University, 1999. The 1999 Language Engineering Workshop.

[BCG+96] **J. Barnett, A. Corrada, G. Gao, L. Gillick, Y. Ito, S. Lowe, L. Manganaro und B. Peskin**. Multilingual Speech Recognition at Dragon Systems. In *Proc. International Conference on Spoken Language Processing*, S. 2191 – 2194, Philadelphia, PA, Oktober 1996. IEEE.

[Bec84] **J. Becker**. Mehrsprachige Textverarbeitung. *Spektrum der Wissenschaft*, S. 42–54, September 1984.

[BGM97] **P. Bonaventura, F. Gallocchio und G. Micca**. Multilingual Speech Recognition for Flexible Vocabularies. In *Proc. EUROSPEECH, European Conference on Speech Communication and Technology*, S. 355–358, Rhodos, Griechenland, September 1997. ESCA.

[BHN89] **W. J. Bary, C. E. Hoequist und F. J. Nolan**. An approach to the problem of regional accent in automatic speech recognition. *Computer Speech and Language*, 3:355–366, 1989.

[BJM90] **L. Bahl, F. Jelinek und R. Mercer**. A Maximum Likelihood Approach to Continuous Speech Recognition. In *Readings in Speech Recognition*, S. 308–319. Morgan Kaufman Publishers, San Mateo, CA, 1990.

[BKI97] **U. Bub, J. Köhler und B. Imperl**. In-Service Adaptation of Multilingual Hidden-Markov-Models. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, S. 1451–1454, München, April 1997. IEEE.

[BL97] **R. Billi und L. Lamel**. RailTel: Railway Telephone Information Services. *Speech Communication*, 23:63–65, 1997.

[BMM+97] **J. Billa, K. Ma, J. W. McDonough, G. Zavaliagkos, D. R. Miller, K. N. Ross und A. El-Jaroudi**. Multilingual Speech Recognition: The 1996 Byblos Callhome System. In *Proc. EUROSPEECH, European Conference on Speech Communication and Technology*, S. 363–366, Rhodos, Griechenland, September 1997. ESCA.

[Bod97] **F. Bodmer**. *Die Sprachen der Welt*. Parkland, 1997.

[Bon85] **M. Bonner**. *Kompaktgrammatik Schwedisch*. Klett Verlag, 1985.

[Bro93] **W. Browne**. *The Slavonic Languages*. 1993.

[BTG94] **J. Bernstein, K. Taussig und J. Godfrey**. Macrophone: an American English telephone speech corpus for the Polyphone Project. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Band 1, S. 81–84, Adelaide, 1994. IEEE.

[Buß90] **H. Bußmann**. *Lexikon der Sprachwissenschaft*. Alfred Kröner Verlag, Stuttgart, 1990.

[CAGADL97] **C. Corredor-Ardoy, J. Gauvain, M. Adda-Decker und L. Lamel**. Language Identification with Language-independent Acoustic Models. In *Proc. EUROSPEECH, European Conference on Speech Communication and Technology*, S. 55–58, Rhodos, Griechenland, September 1997. ESCA.

[CALADG98] **C. Corredor-Ardoy, L. Lamel, M. Adda-Decker und J. Gauvain**. Multilingual Phone Recognition of Spontaneous Telephone Speech. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, S. 413–416, Seattle, WA, 1998. IEEE.

[CC97] **A. Constantinescu und G. Chollet**. On Cross-Language Experiments and Data-Driven Units for ALISP (Automatic Language Independent Speech Processing. In *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, S. 606–613, St. Barbara, CA, Dezember 1997. IEEE.

[Ç98] **K. Çarkı**. Entwicklung eines türkischen Spracherkennungssystems für große Vokabulare. Diplomarbeit, Betreuerin Tanja Schultz, Universität Karlsruhe, Institut für Logik, Komplexität und Deduktionssysteme, September 1998.

[CCM98] **M. Cettolo, A. Corazza und R. D. Mori**. Language Portability of a Speech Understanding System. *Computer Speech and Language*, 12:1 – 21, 1998.

[CDG+97] **P. Cohen, S. Dharanipragada, J. Gros, M. Monkowski, C. Neti, S. Roukos und T. Ward**. Towards a Universal Speech Recognizer for Multiple Languages. In *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, S. 591–598, St. Barbara, CA, Dezember 1997. IEEE.

[CGM+97] **C. Chen, R. Gopinath, M. Monkowski, M. Picheny und K. Shen**. New Methods in Continuous Mandarin Speech Recognition. In *Proc. EUROSPEECH, European Conference on Speech Communication and Technology*, Band 3, S. 1543–1546, Rhodos, Griechenland, September 1997. ESCA.

[ÇGS00] **K. Çarkı, P. Geutner und T. Schultz**. Turkish LVCSR: Towards better Speech Recognition for Agglutinative Languages. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Istanbul, Turkei, Juni 2000. IEEE.

[Cho93] **R. Choe**. *Phonemvarianzregeln im Deutschen und Koreanischen.* Kümmerle Verlag, Göppingen, 1993.

[CIA98] *Computer Industry Almanac Inc.* 1998.

[CMP98] **B. Comrie, S. Matthews und M. Polinski** (Hrsg.). *Bildatlas der Sprachen.* Autoren- und Medienservice (AMS), Weltbild Verlag, Augsburg, 1998.

[Cry95] **D. Crystal**. *Die Cambridge Enzyklopädie der Sprache.* Campus, 1995.

[Cry00] **D. Crystal**. Death Sentence. *Spotlight, Das aktuelle Magazin in Englisch*, S. 14–20, März 2000.

[CS99] **C-STAR**. *Consortium for Speech Translation Advanced Research.* Internet, http://www.c-star.org, September 1999.

[CuC+97] **G. Chollet, J. Černocký, A. Constantinesu, S. Deligne und F. Bimbot**. Towards ALISP: A proposal for Automatic Language Independent Speech Processing. In *NATO Summer School.* Nato, Jersey, 1997.

[DA92] **P. Dalsgaard und O. Andersen**. Identification of Mono- and Polyphonemes using acoustic-phonetic Features derived by a self-organising Neural Network. In *Proc. International Conference on Spoken Language Processing*, S. 547–550, Banff, Alberta, Canada, Oktober 1992. IEEE.

[DA94] **P. Dalsgaard und O. Andersen**. Application of Inter-Language Phoneme Similarities for Language Identification. In *Proc. International Conference on Spoken Language Processing*, S. 1903–1906, Yokohama, Japan, Oktober 1994. IEEE.

[DAB98] **P. Dalsgaard, O. Andersen und W. Barry**. Cross-language merged Speech units and their Descriptive Phonetic Correlates. In *Proc. International Conference on Spoken Language Processing*, Sydney, November 1998.

[DAK95] **C. Dugast, X. Aubert und R. Kneser**. The Philips Large-Vocabulary Recognition System for American English, French and German. In *Proc. EUROSPEECH, European Conference on Speech Communication and Technology*, S. 197–200, Madrid, September 1995. ESCA.

[DB95] **S. Deligne und F. Bimbot**. Language Modeling by Variable Lengh Sequences: Theoretical Formulation and Evaluation of Multigrams. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, S. 169–172, Detroit, MI, Mai 9-12 1995. IEEE.

[DBiV+94] **M. Damhuis, T. Boogaart, C. in't Veld, M. Versterijlen, W. Schelvis, L. Bos und L. Boves**. Creation and Analysis of the Dutch Polyphone Corpus. In *Proc. International Conference on Spoken Language Processing*, Yokohama, Japan, September 1994. IEEE.

[Den97a] **L. Deng**. A Dynamic Feature-Based Approach to Speech Modeling and Recognition. In *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, S. 107–114, St. Barbara, CA, Dezember 1997. IEEE.

[Den97b] **L. Deng**. Integrated-multilingual speech recognition using universal phonological features in a production model. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, S. 1007–1010, München, April 1997. IEEE.

[DH73] **R. Duda und P. Hart**. *Pattern Classification and Scene Analysis*. John Wiley & Sons, New York, Chichester, Brisbane, Toronto, Singapore, 1973.

[DK78] **W. Digel und G. Kwiatkowski** (Hrsg.). *Meyers Neues Lexikon in 8 Bd.* Bibliographisches Institut AG, Mannheim, 1978.

[DLR77] **A. Dempster, N. Laird und D. Rubin**. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society*, 39(1):1–22, 1977.

[Edw95] **J. Edwards**. *Multilingualism*. Penguin Books, 1995.

[Ell97] **T. Ellbogen**. *Phonetik Seminar*. Internet, http:///www.phonetik.uni-muenchen.de/MUSE/Seminare/PHON_Einf/anatomie, 1997.

[ELR98] **ELRA**. *The European Language Resources Association.* Internet, http://www.icp.grenet.fr/ELRA/home.html, September 1998.

[FGH+97] **M. Finke, P. Geutner, H. Hild, T. Kemp, K. Ries und M. Westphal**. The Karlsruhe-Verbmobil Speech Recognition Engine. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, S. 83–86, München, April 1997. IEEE.

[FGJ98] **V. Fischer, Y. Gao und E. Janke**. Speaker-Independent Upfront Dialect Adaptation in Large Vocabulary Continuous Speech Recognizer. In *Proc. International Conference on Spoken Language Processing*, Sydney, November 1998.

[FML99] **P. Fung, C. Y. Ma und W. K. Liu**. MAP-Based Cross-Language Adaptation Augmented by Linguistic Knowldedge from English to Chinese. In *Proc. EUROSPEECH, European Conference on Speech Communication and Technology*, Band 2, S. 871–874, Budapest, Ungarn, September 1999. ESCA.

[FR97] **M. Finke und I. Rogina**. Wide Context Acoustic Modeling in Read vs. Spontaneous Speech. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, S. 1743–1746, München, April 1997. IEEE.

[Gav96] **M. Gavalda**. Carnegie Mellon University. Persönliches Gespräch im Rahmen eines Gastaufenthaltes an der CMU, 1996.

[Geu99] **P. Geutner**. *Adaptive Vocabularies in Large Conversational Speech Recognition.* Dissertation, Universität Karlsruhe, Institut für Logik, Komplexität und Deduktionssysteme, 1999.

[GFG+95] **J. Glass, G. Flammia, D. Goodine, M. Phillips, J. Polifroni, S. Sakai, S. Seneff und V. Zue**. Multi-lingual Spoken Language Understanding in the MIT Voyager System. *Speech Communication*, 17:1–18, 1995.

[GFS97] **P. Geutner, M. Finke und P. Scheytt**. Hypothesis Driven Lexical Adaptation for Transcribing Multilingual Broadcast News. Technical Report CMU-LTI-97-155, Language Technologies Institute, Carnegie Mellon University, 1997.

[GFW99] **P. Geutner, M. Finke und A. Waibel**. Selection Criteria for Hypothesis Driven Lexical Adaptation. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Seattle, WA, 1999. IEEE.

[GG97] **S. Gokcen und J. Gokcen**. A Multilingual Phoneme and Model Set: Towards a universal base for Automatic Speech Recognition. In *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, S. 599–603, St. Barbara, CA, Dezember 1997. IEEE.

[GI90] **T. W. Gamkrelidse und W. W. Iwanow**. Die Frühgeschichte der indoeuropäischen Sprachen. *Spektrum der Wissenschaft*, S. 130–137, Mai 1990.

[Gri92] **B. F. Grimes**. *Ethnologue: Languages of the World.* Summer Institute of Linguistics, http://www.sil.org/Ethnologue, 1992.

[Haa91] **H. Haarmann**. *Universalgeschichte der Schrift.* Campus Verlag, Frankfurt/Main; New York, 2 edition, 1991.

[HBTA99] **M. Hunt, P. Bamberg, J. Tucker und S. Anderson**. A Military Operational Automatic Interpreting System. In *Proc. ESCA-NATO Tutorial and Research Workshop on Multi-lingual Interoperability in Speech Technology*, S. 75–78, Leusden, Niederlande, September 1999. ESCA.

[hco98] **hcode**. *Hangul Conversion Tool.* Internet, http://ftp.kaist.ac.kr/hangul/code/hcode, November 1998.

[Her94] **W. Herrmann**. *Lehrbuch der modernen koreanischen Sprache.* Helmut Buske Verlag, 1994.

[HH91] **M.-Y. Hwang und X. Huang**. Shared-Distribution Hidden Markov Models for Speech Recognition. Technical report, CMU-CS91-124, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, 15213, USA, April, 1991.

[Hie93] **J. L. Hieronymus**. ASCII Phonetic Symbols for the World's Languages: Worldbet. *Journal of the International Phonetic Association*, 1993.

[HNN97] **S. Harbeck, E. Nöth und H. Niemann**. Multilingual Speech Recognition. In *Proc. SQEL, 2nd Workshop on Multi-Lingual Information Retrieval Dialogs*, S. 9–15, Plzeň, Tschechien, April 1997. University of West Bohemia.

[HNN98] **S. Harbeck, E. Nöth und H. Niemann**. Multilingual Speech Recognition in the Context of Multilingual Information Retrieval Dialogues. In *Proc. of the first Workshop on Text, Speech and Dialogue TSD*, S. 375–380, Brno, Tschechien, September 1998. Masaryk University.

[Hov99] **E. Hovy**. University of Southern California. Persönliches Gespräch am Rand des EAGLES Workshop in Pisa, 1999.

[Hun97] **M. J. Hunt**. Practical large-vocabulary Speech Recognition in a Multilingual environment. *Speech Communication*, 23:297–305, 1997.

[IDA98] **IDAS**. *Interactive telephony-based directory assistance services (EU-Projekt).* Internet, http://www.tik.ee.ethz.ch/˜ idas, September 1998.

[IH99] **B. Imperl und B. Horvat**. The Clustering Algorithm for the Definition of Multilingual Set of Context Dependent Speech Models. In *Proc. EUROSPEECH, European Conference on Speech Communication and Technology*, Band 2, S. 887–890, Budapest, Ungarn, September 1999. ESCA.

[ILA98] **ILAM**. *Interactive Language Aquisitation through Multimedia Stimulated Conversation and Pronunciation (EU-Projekt).* Internet, http://www2.echo.lu/langeng/projects/ilam/index.html, September 1998.

[Imp99] **B. Imperl**. Clustering of Context Dependent Speech Units for Multilingual Speech Recognition. In *Proc. ESCA-NATO Tutorial and Research Workshop on Multi-lingual Interoperability in Speech Technology*, S. 17–22, Leusden, Niederlande, September 1999. ESCA.

[IPA93] **IPA**. The International Phonetic Association (revised to 1993) - IPA Chart. *Journal of the International Phonetic Association*, 1(23), 1993.

[ISL98] **ISLE**. *Interactive Spoken Language Education (EU-Projekt).* Internet, http://nats-www.informatik.uni-hamburg.de/˜ isle/index.html, September 1998.

[JC99] **F. Jelinek und C. Chelba**. Putting Language into Language Modeling. In *Proceedings of the Eurospeech 1999*, Seite Keynote 1, Budapest, Ungarn, September 1999. ESCA.

[Kat87] **S. Katz**. Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 35, 1987.

[Kem99] **T. Kemp**. *Ein automatisches Indexierungssystem für Fernsehnachrichten.* Dissertation, Universität Karlsruhe, Institut für Logik, Komplexität und Deduktionssysteme, 1999.

[KH97] **H. Kwan und K. Hirose**. Use of Recurrent Network for Unknown Language Rejection in Language Identification Systems. In *Proc. EUROSPEECH, European Conference on Speech Communication and Technology*, Band 1, S. 63–67, Rhodos, Griechenland, September 1997. ESCA.

[Kie99] **D. Kiecza**. Datengetriebene Bestimmung von Vokabulareinheiten für die koreanische Spracherkennung auf großen Wortschätzen. Diplomarbeit, Betreuerin Tanja Schultz, Universität Karlsruhe, Institut für Logik, Komplexität und Deduktionssysteme, Oktober 1999.

[Kim96] **J.-H. Kim**. *Lexical Disambiguation with Error-driven Learning.* Dissertation, Computer Science Department, Advanced Institute of Science and Technology, Korea, 1996.

[Koh77] **K. Kohler**. *Einführung in die Phonetik des Deutschen.* Erich Schmidt Verlag, Berlin, 1977.

[Köh96] **J. Köhler**. Multi-lingual Phoneme Recognition Exploiting Acoustic-phonetic Similarities of Sounds. In *Proc. International Conference on Spoken Language Processing*, S. 2195–2198, Philadelphia, PA, 1996. IEEE.

[Köh97] **J. Köhler**. Multilingual Phone Modelling for Telephone Speech. In *Proc. SQEL, 2nd Workshop on Multi-Lingual Information Retrieval Dialogs*, S. 16–19, Plzeň, Tschechien, April 1997. University of West Bohemia.

[Köh98] **J. Köhler**. Language Adaptation of Multilingual Phone Models For Vocabulary Independent Speech Recognition Tasks. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, S. 417–420, Seattle, WA, 1998. IEEE.

[Köh99] **J. Köhler**. Comparing three Methods to Create Multilingual Phone Models for Vocabulary Independent Speech Recognition Tasks. In *Proc. ESCA-NATO Tutorial and Research Workshop on Multi-lingual Interoperability in Speech Technology*, S. 79–84, Leusden, Niederlande, September 1999. ESCA.

[Kom96] **R. Kompe**. *Prosody in Speech Understanding Systems.* Dissertation, Universität Erlangen-Nürnberg, 1996.

[Kor90] **J. Kornfilt**. Turkish and the Turkic Languages. In *The Worlds Major Languages ([Com90])*, S. 619–645. Oxford University Press, London, 1990.

[Kri84] **M. Krifka**. Neue Theorien der Lautverschiebung. *Spektrum der Wissenschaft*, S. 31–32, Oktober 1984.

[Kri88] **M. Krifka**. Die Sprachfamilien Amerikas and die Ursprache der Menschheit. *Spektrum der Wissenschaft*, S. 40–41, Januar 1988.

[KSW99] **D. Kiecza, T. Schultz und A. Waibel**. Data-Driven Determination of Appropriate Dictionary Units for Korean LVCSR. In *Proc. International Conference on Speech Processing (ICSP'99)*, S. 323–327, Seoul, Korea, August 1999. Acoustic Society of Korea.

[Kun00] **S. Kunzmann**. IBM - Europäische Sprachlabors. Persönliches Gespräch, 2000.

[Kwo99] **O.-W. Kwon**. Korean Large Vocabulary Continuous Speech Recognition Using Morpheme-Based Units. Technischer Bericht, Institut für Logik, Komplexität und Deduktionssysteme, Universität Karlsruhe, Dezember 1999.

[LADG95] **L. Lamel, M. Adda-Decker und J. Gauvain**. Issues in Large Vocabulary Multilingual Speech Recognition. In *Proc. EUROSPEECH, European Conference on Speech Communication and Technology*, S. 185–189, Madrid, September 1995. ESCA.

[LADGA96] **L. Lamel, M. Adda-Decker, J. Gauvain und G. Adda**. Spoken Language Processing in a Multilingual Context. In *Proc. International Conference on Spoken Language Processing*, S. 2203–2206, Philadelphia, PA, Oktober 1996. IEEE.

[LDC00] **LDC**. *The Linguistic Data Consortium.* Internet, http://www.ldc.upenn.edu, Januar 2000.

[Lee88] **K. F. Lee**. *Large-Vocabulary Speaker-Independent Continuous Speech Recognition: The SPHINX System.* Dissertation, Carnegie Mellon University, CMU-CS-88-148, Pittsburgh, PA, 1988.

[LG93a] **L. Lamel und J. Gauvain**. High Performance Speaker Independent Phone Recognition Using CDHMM. In *Proc. EUROSPEECH, European Conference on Speech Communication and Technology*, Berlin, 1993. ESCA.

[LG93b] **L. Lamel und J. Gauvain**. Identifying Non-linguistic Speech Features. In *Proc. EUROSPEECH, European Conference on Speech Communication and Technology*, S. 23–30, Berlin, 1993. ESCA.

[Lin98] **LinguaNet**. *Communication Through the Language Barrier (EU-Projekt).* Internet, http://www.cbs.dk/projects/linguanet, September 1998.

[LM94] **T. Lander und S. T. Metzler**. The CSLU Labeling Guide. Technical report, Center for Spoken Language Understanding, Oregon Graduate Institute, 1994.

[LW95] **C. Leggetter und P. Woodland**. Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models. *Computer Speech and Language*, 9:171 – 185, 1995.

[Lyu95] **R.-Y. Lyu**. Golden Mandarin (III) - A User-Adaptive Prosodic-Segment-Based Mandarin Dictation Machine for Chinese Language with Very Large Vocabulary. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, S. 57–60, Detroit, MI, Mai 9-12 1995. IEEE.

[MA94] **J. Miller und F. Alleva**. Evaluation of a Language Model using a Clustered Model Backoff. In *Proc. International Conference on Spoken Language Processing*, Yokohama, Japan, Oktober 1994. IEEE.

[Mad84] **I. Maddieson**. *Patterns of Sounds.* Cambridge: C.U.P., 1984.

[Mai94] **H. Maier**. *Die vielen Sprachen und die Eine Welt.* Robert Bosch Stiftung, 1994.

[MAI98] **MAIS**. *Multilingual Automatic Inquiry Systems (EU-Projekt).* Internet, http://www.linglink.lu/le/projects/mais/index.html, September 1998.

[Mat97] **Y. Matsumoto**. Japanese Morphological Analysis System: CHASEN. Technical Report NAIST-IS-TR97007, Nara Institute of Science and Technology, 1997.

[MBC94] **Y. Muthusamy, E. Barnard und R. Cole**. Reviewing Automatic Language Identification. *IEEE Signal Processing Magazin*, Vol. 11 Nr. 4:33–41, Oktober 1994.

[MCO92] **Y. Muthusamy, R. Cole und B. Oshika**. The OGI multi-language telephone speech corpus. In *Proc. International Conference on Spoken Language Processing*, S. 895–898, Banff, Alberta, Canada, Oktober 1992. IEEE.

[MHW+95] **Y. Muthusamy, E. Holliman, B. Wheatley, J. Picone und J. Godfrey**. Voice across Hispanic America: A Telephone Speech Corpus of American Spanish. In *IEEE*, S. 85–88, 1995.

[MIE98] **MIETTA**. *Multilingual Information Extraction for Tourism and Travel Assistance (EU-Projekt).* Internet, http://pikas.inf.tu-dresden.de/aktivitaeten/ki97-98/FF/mietta.html, September 1998.

[MKS+00] **F. Metze, T. Kemp, T. Schaaf, T. Schultz und H. Soltau**. Confidence Measure based Language Identification. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Istanbul, Turkei, Juni 2000. IEEE.

[ML98] **J. Mariani und L. Lamel**. An Overview of EU Programs related to Conversational/Interactive Systems. In *Proc. DARPA Workshop on Broadcast News Transcription and Understanding*, S. 247–253, Lansdowne, VA, Februar 1998.

[MPF99] **G. Micca, E. Palme und A. Frasca**. Multilingual Vocabularies in Automatic Speech Recognition. In *Proc. ESCA-NATO Tutorial and Research Workshop on Multi-lingual Interoperability in Speech Technology*, S. 65–68, Leusden, Niederlande, September 1999. ESCA.

[NB99] **C. Nieuwoudt und E. Botha**. Adaptation of Acoustic Models for Multilingual Recogntion. In *Proc. EUROSPEECH, European Conference on Speech Communication and Technology*, Band 2, S. 907–910, Budapest, Ungarn, September 1999. ESCA.

[NBK+97] **E. Nöth, A. Batliner, A. Kießling, R. Kompe und H. Niemann**. Prosodische Information: Begriffsbestimmung und Nutzen für das Sprachverstehen. In *Mustererkennung*, S. 37–52. Informatik aktuell, Springer, Heidelberg, 1997.

[NBM+85] **H. Niemann, A. Brietzmann, R. Mühlfeld, P. Regel und G. Schukat**. The Speech Understanding and Dialog System EVAR. In **R. D. Mori und C. Suen** (Hrsg.), *New Systems and Architectures for Automatic Speech Recognition and Synthesis*, NATO ASI Series, S. 271–302. Springer Verlag, 1985.

[NEK94] **H. Ney, U. Essen und R. Kneser**. On Structuring Probabilistic Dependences in Stochastic Language Modelling. *Computer Speech and Language*, 8(1):1–38, 1994.

[Ney90] **H. Ney**. The Use of a One-stage Dynamic Programming Algorithm for Connected Word Recognition. In *Readings in Speech Recognition*, S. 188–196. Morgan Kaufman Publishers, San Mateo, CA, 1990.

[Nö91] **E. Nöth**. *Prosodische Information in der automatischen Spracherkennung - Berechnung und Anwendung.* Niemeyer, Tübingen, 1991.

[OAM+92] **L. Osterholtz, C. Augustine, A. McNair, I. Rogina, H. Saito, T. Sloboda, J. Tebelskis, A. Waibel und M. Woszczyna**. Testing Generality in JANUS: a Multi-lingual Speech Tranlslation System. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 1992.

[OB94] **K. Oflazar und H. Bozşahin**. Turkish Natural Language Processing Initiative: An Overview. In *TAINN94 (3rd Turkish AI and Neural Network Conference)*, 1994.

[OLI98] **OLIVE**. *A Multi-lingual Indexing Tool for Broadcast Material based on Speech Recognition (EU-Projekt).* Internet, http://twentyone.tpd.tno.nl/Olive, September 1998.

[OT96] **K. Oflazar und G. Tür**. Combining Hand-crafted Rules and Unsubervised Learning in Constraint-based Morphological Disambiguation. In *ACL96 (Proceedings of ACL Conference on Empirical Methods in Natural Language Processing*, Philadelphia, PA, Mai 1996.

[PE98] **Pop-Eye**. *A Multilingual Continuous Video Disclosing Tool, Based on Subtitle Indexing and Partial Translation (EU-Projekt).* Internet, http://www.pop-eye.tros.com, September 1998.

[PM92] **B. Pompino-Marschall**. *PhonDat, Verbundvorhaben zum Aufbau eine Sprachsignaldatenbank für gesprochenes Deutsch.* FIPKM, Ludwig-Maximilian Universität München, Juli 1992.

[PM95] **B. Pompino-Marschall**. *Einführung in die Phonetik.* de Gruyter Studienbuch, Berlin, 1995.

[PWY95] **D. Pye, P. Woodland und S. Young**. Large Vocabulary Multilingual Speech Recognition using HTK. In *Proc. EUROSPEECH, European Conference on Speech Communication and Technology*, S. 181–184, Madrid, September 1995. ESCA.

[Rab90] **L. Rabiner**. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. In *Readings in Speech Recognition ([WL90])*, S. 267–296. Morgan Kaufman Publishers, San Mateo, CA, 1990.

[Ras98] **S. Raschke**. Automatische Generierung eines Ausspracheworterbuches und Initialisierung eines Erkenners für die kroatische Sprache. Studienarbeit, Betreuerin Tanja Schultz, Universität Karlsruhe, Institut für Logik, Komplexität und Deduktionssysteme, Januar 1998.

[Ras00] **S. Raschke**. Multilinguale Spracherkennung durch Kommunikation monolingualer Systeme - Implementierung und Visualisierung. Diplomarbeit, Betreuerin Tanja Schultz, Universität Karlsruhe, Institut für Logik, Komplexität und Deduktionssysteme, März 2000.

[RBW96] **K. Ries, F. D. Buø und A. Waibel**. Class Phrase Models for Language Modeling. In *Proc. International Conference on Spoken Language Processing*, Philadelphia, PA, Oktober 1996. IEEE.

[REC98] **RECALL**. *Repairing Errors in Computer-Assisted Language Learning (EU-Projekt).* Internet, http://www.infj.ulst.ac.uk:80/˜ recall/, September 1998.

[Rei97] **J. Reichert**. Lautschriftumsetzung und Worttrennung der chinesischen Schriftsprache. Studienarbeit, Betreuerin Tanja Schultz, Universität Karlsruhe, Institut für Logik, Komplexität und Deduktionssysteme, April 1997.

[Rei98] **J. Reichert**. Spracherkennung im Chinesischen. Diplomarbeit, Betreuerin Tanja Schultz, Universität Karlsruhe, Institut für Logik, Komplexität und Deduktionssysteme, Dezember 1998.

[Ren89] **C. Renfrew**. Der Ursprung der indoeuropäischen Sprachfamilie. *Spektrum der Wissenschaft*, S. 114–122, Dezember 1989.

[RJ93] **L. Rabiner und B. Juang** (Hrsg.). *Fundamentals of Speech Recognition*. Signal Processing. Prentice Hall, Englewood Cliffs, 1993.

[Roa97] **P. Roach**. The Babel Project - Speech databases from Central & Eastern Europe. In **K. Choukri** (Hrsg.), *The ELRA Newsletter*, Seite 9, Paris, October 1997. ELRA/ELDA.

[Rog97] **I. Rogina**. *Parameterraumoptimierung für Diktiersysteme mit unbeschränktem Vokabular*. Dissertation, Universität Karlsruhe, Institut für Logik, Komplexität und Deduktionssysteme, 1997.

[Ros91] **P. E. Ross**. Streit um Wörter. *Spektrum der Wissenschaft*, S. 92–101, Juni 1991.

[RS78] **L. Rabiner und R. Schafer** (Hrsg.). *Digital Processing of Speech Signals*. Prentice Hall, Englewood Cliffs, 1978.

[RSW99] **J. Reichert, T. Schultz und A. Waibel**. Mandarin Large Vocabulary Speech Recognition using the GlobalPhone Database. In *Proceedings of the Eurospeech 1999*, S. 815–818, Budapest, Ungarn, September 1999. ESCA.

[SAM98] **SAMPA**. *Speech Assesssment Methods Phonetic Alphabet*. Internet, http://www.phon.ucl.ac.uk/home/sampa/home.htm, September 1998.

[SC90] **H. Sakoe und S. Chiba**. Two-Level DP-Matching - A Dynamic Programming-Based Pattern Matching Algorithm for Connected Word Recognition. In *Readings in Speech Recognition ([WL90])*, S. 180–187. Morgan Kaufman Publishers, San Mateo, CA, 1990.

[SC98] **SpeechDat-CAR**. *Speech Databases for Voice Driven Teleservices and Control in Automotive Environments (EU-Projekt)*. Internet, http://www.speechdat.com/SP-CAR, September 1998.

[Sch99] **K. Schubert**. Grundfrequenzverfolgung und deren Anwendung in der Spracherkennung. Diplomarbeit, Betreuer Martin Westphal, Universität Karlsruhe, Institut für Logik, Komplexität und Deduktionssysteme, September 1999.

[Sho99] **M. Shozakai**. Speech Interface VLSI for Car Applications. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Seattle, WA, 1999. IEEE.

[SJR+95] **J. Salavadera, C. Jacobsen, M. Rahim, I. Zeljkovic und J. Wilson**. Multi-lingual Connected Digits Recognition. In *Proc. EUROSPEECH, European Conference on Speech Communication and Technology*, S. 2119–2122, Madrid, 1995. ESCA.

[Sku88] **S. Skudlik**. Die Kinder Babylons. In *Gerechtfertigte Vielfalt. Zur Sprache in den Geisteswissenschaften*. Els Oksaar, 1988.

[SKW97] **T. Schultz, D. Koll und A. Waibel**. Japanese LVCSR on the Spontaneous Scheduling Task with JANUS-3. In *Proc. EUROSPEECH, European Conference on Speech Communication and Technology*, Band 1, S. 367–370, Rhodos, Griechenland, September 1997. ESCA.

[SPE98a] **SPEAK**. *Supported Prototype Easy-access Authoring Keys (EU-Projekt).* Internet, http://www2.echo.lu/langeng/eu/le3/speak/speak.html, September 1998.

[Spe98b] **SpeechDat**. *Speech Databases (EU-Projekt).* Internet, http://www.cis.uni-muenchen.de/hot/speechdat.html, September 1998.

[SQA98] **SQALE**. *Speech Recognition quality assessment for linguistic engineering (EU-Projekt).* Internet, http://www.tno.nl/instit/tm/index.html, September 1998.

[SR95] **T. Schultz und I. Rogina**. Acoustic and Language Modeling of Human and Nonhuman Noises for Human-to-Human Spontaneous Speech Recognition. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Band 1, S. 293–296, Detroit, MI, Mai 1995. IEEE.

[SRW96] **T. Schultz, I. Rogina und A. Waibel**. LVCSR-based language identification. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Band 2, S. 781–784, Atlanta, GA, Mai 7-10 1996. IEEE.

[ST95] **E. Schukat-Talamazzini**. *Automatische Spracherkennung.* Vieweg Verlag, Braunschweig, 1995.

[STi99] **STiLL**. *(EU-Projekt).* Internet, http://www2.echo.lu/langeng/eu/le3/still-/still.html, September 1999.

[Stö97] **J. Störig**. *Abenteuer Sprache - Ein Streifzug durch die Sprachen der Welt.* Humboldt Verlag, 1997.

[SUN99] **SUNDIAL**. *Speech Understanding and Dialogue (EU Esprit-Projekt).* Internet, http://www-uk.research.ec.org/esp-syn/text/2218.html, November 1999.

[SvL95] **H. Steeneken und D. van Leeuwen**. Multilingual Assessment of Speaker Independent Large Vocabulary Speech Recognition Systems: the SQALE Project. In *Proc. EUROSPEECH, European Conference on Speech Communication and Technology*, S. 1271–1274, Madrid, September 1995. ESCA.

[SW79] **N. Smith und D. Wilson**. *Modern Linguistics. The Results of Chomsky's Revolution.* John Spiers, Harvester Press, Brighton, Sussex, 1979.

[SW97] **T. Schultz und A. Waibel**. Fast Bootstrapping of LVCSR Systems with Multilingual Phoneme Sets. In *Proc. EUROSPEECH, European Conference on Speech Communication and Technology*, Band 1, S. 371–373, Rhodos, Griechenland, September 1997. ESCA.

[SW98a] **T. Schultz und A. Waibel**. Das Projekt GlobalPhone: Multilinguale Spracherkennung. In *Proc. of the 4th Conference on Natural Language Processing KONVENS-98*, S. 179–189, Bonn, Oktober 1998. Peter Lang Verlag.

[SW98b] **T. Schultz und A. Waibel**. Language Independent and Language Adaptive Large Vocabulary Speech Recognition. In *Proc. International Conference on Spoken Language Processing*, Band 5, S. 1819–1822, Sydney, November 1998.

[SW98c] **T. Schultz und A. Waibel**. Multilingual and Crosslingual Speech Recognition. In *Proc. DARPA Workshop on Broadcast News Transcription and Understanding*, S. 259–262, Lansdowne, VA, Februar 1998.

[SW99] **T. Schultz und A. Waibel**. Language adaptive LVCSR through Polyphone Decision Tree Specialization. In *Workshop on Multi-lingual Interoperability in Speech Technology (MIST '99)*, S. 85–90, Leusden, Niederlande, September 1999. NATO.

[SW00] **T. Schultz und A. Waibel**. Polyphone Decision Tree Specialization for Language Adaptation. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Istanbul, Turkei, Juni 2000. IEEE.

[SWW97] **T. Schultz, M. Westphal und A. Waibel**. The GlobalPhone Project: Multilingual LVCSR with Janus-3. In *Proc. SQEL, 2nd Workshop on Multi-Lingual Information Retrieval Dialogs*, S. 20–27, Plzeň, Tschechien, April 1997. University of West Bohemia.

[Ter87] **E. Ternes**. *Einführung in die Phonologie.* Wissenschaftliche Buchgesellschaft, Darmstadt, 1987.

[TR97] **L. J. M. Tomokiyo und K. Ries**. What makes a word: Learning base units in Japanese for Speech Recognition. In *Proceedings of the Workshop on Natural Language Learning*, 1997.

[ÜSN98] **U. Übler, M. Schüßler und H. Niemann**. Bilingual and Dialectal Adaptation and Retraining. In *Proc. International Conference on Spoken Language Processing*, Sydney, November 1998.

[Van99] **D. VanCompernolle**. Speech Recognition by Goats, Wolves, Sheeps and Non-Natives. In *Proc. ESCA-NATO Tutorial and Research Workshop on Multi-lingual Interoperability in Speech Technology*, S. 3–9, Leusden, Niederlande, September 1999. ESCA.

[VER00] **VERBMOBIL**. *Verbmobil (BMBF Verbundprojekt).* Internet, http://verbmobil.dfki.de, Februar 2000.

[VOD98] **VODIS**. *Advanced Speech Technologies for Voice Operated Driver Information Systems (EU-Projekt)*. Internet, http://isl.ira.uka.de/VODIS, September 1998.

[Wah99] **W. Wahlster**. Verbmobil. Mündliche Äußerung während seines Vortrages auf dem C-Star Workshop in Schwetzingen, September 1999.

[WBNS97] **F. Weng, H. Bratt, L. Neumeyer und A. Stolke**. A Study of Multilingual Speech Recognition. In *Proc. EUROSPEECH, European Conference on Speech Communication and Technology*, S. 359–362, Rhodos, Griechenland, September 1997. ESCA.

[Web92] **Webster** (Hrsg.). *Webster's New Encyclopedic Dictionary*. Black, Dog & Leventhal, 1992.

[Wes00] **M. Westphal**. TalkingMap - ein tragbares Touristeninformationssystem. Kooperation mit dem European Media Lab (Heidelberg), als Teil von DeepMap, 2000.

[WGT+00] **A. Waibel, P. Geutner, L. M. Tomokiyo, T. Schultz und M. Woszczyna**. Multilinguality in Speech and Spoken Language Systems. *IEEE Transactions*, 2000. Erscheint voraussichtlich im August 2000.

[WHW96] **H. Wactar, A. Hauptmann und M. Witbrock**. Informedia: News on Demand Experiments in Speech Recognition. In *Proc. of the ARPA SLT workshop*, 1996.

[WKAM94] **B. Wheatley, K. Kondo, W. Anderson und Y. Muthusamy**. An Evaluation of Cross-language Adaptation For Rapid HMM Development in a new language. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, S. 237–240, Adelaide, 1994. IEEE.

[WL90] **A. Waibel und K. F. Lee** (Hrsg.). *Readings in Speech Recognition*. Morgan Kaufman Publishers, San Mateo, CA, 1990.

[Wol99] **R. Wolff**. Adaption von Kontextentscheidungsbäumen auf neue Sprachen. Studienarbeit, Betreuerin Tanja Schultz, Universität Karlsruhe, Institut für Logik, Komplexität und Deduktionssysteme, Juni 1999.

[Wos98] **M. Woszczyna**. *Fast Speaker Independent Large Vocabulary Continuous Speech Recognition*. Dissertation, Universität Karlsruhe, Institut für Logik, Komplexität und Deduktionssysteme, Februar 1998.

[WRN+98] **T. Ward, S. Roukos, C. Neti, M. Epstein und S. Dharanipragada**. Towards Speech Understanding across multiple Languages. In *Proc. International Conference on Spoken Language Processing*, Sydney, November 1998. IEEE.

[WSM00] **A. Waibel, T. Schultz und F. Metze**. *The Verbmobil Book*, chapter Multilingual Speech Recognition. Springer Verlag, Heidelberg, 2000. Erscheint voraussichtlich im Juli 2000.

[WSY$^+$95] **H. Wang, J. Shen, Y. Yang, C. Tseng und L. Lee**. Complete Recognition of Continuous Mandarin Speech for Chinese Language with Very Large Vocabulary but limited Training Data. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, S. 61–64, Detroit, MI, Mai 9-12 1995. IEEE.

[WTK98] **G. Williams, M. Terry und J. Kaye**. Phonological Elements as a Basis for Language-independent ASR. In *Proc. International Conference on Spoken Language Processing*, Sydney, November 1998.

[WW99] **M. Westphal und A. Waibel**. Towards Spontaneous Speech Recognition for Onboard Car Navigation and Information Systems. In *Proc. EUROSPEECH, European Conference on Speech Communication and Technology*, Budapest, Ungarn, September 1999. ESCA.

[YADA$^+$97] **S. Young, M. Adda-Decker, X. Aubert, C. Dugast, J.-L. Gauvain, D. Kershaw, L. Lamel, D. Leeuwen, D. Pye, A. Robinson, H. Steeneken und P. Woodland**. Multilingual large vocabulary speech recognition: the European SQALE project. *Computer Speech and Language*, 11:73 – 89, 1997.

[ZB99] **M. Zissman und K. M. Berkling**. Automatic Language Identification. In *Proc. ESCA-NATO Tutorial and Research Workshop on Multi-lingual Interoperability in Speech Technology*, S. 93–101, Leusden, Niederlande, September 1999. ESCA.

[ZC98] **G. Zavaliagkos und T. Colthurst**. Utilizing Untranscribed Training Data to Improve Performance. In *Proc. DARPA Workshop on Broadcast News Transcription and Understanding*, S. 301–305, Lansdowne, VA, Februar 1998.

[ZS95] **M. Zissman und E. Singer**. Language Identification using Phoneme Recognition and Phonotactic Language Modeling. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Band 5, S. 3503–3506, Detroit, MI, May 1995. IEEE.

[ZSP$^+$96] **V. Zue, S. Seneff, J. Polofroni, H. Meng und J. Glass**. Multilingual Human-Computer Interactions: From Information Access to Language Learning. In *Proc. International Conference on Spoken Language Processing*, S. 2207 – 2210, Philadelphia, PA, Oktober 1996. IEEE.

[Zue93] **V. Zue**. Multilingual Research in the Spoken Language Systems Group. In *Proc. Speech Research Symposium SRS XIII*, Baltimore, Maryland, Juni 1993.

[ZW97] **P. Zhan und M. Westphal**. Speaker Normalization Based on Frequency Warping. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, S. 1039–1042, München, April 1997. IEEE.

[ZWG99] **P. Zhan, S. Wegmann und L. Gillik**. Improvements to Dragon Systems' 1998 Mandarin Broadcast News Transcription System. In *Proc. DARPA*

Workshop on Broadcast News Transcription and Understanding, Herndon, VA, März 1999.

[ZWL98] **P. Zhan, S. Wegmann und S. Lowe**. Dragon Systems' 1997 Mandarin Broadcast News System. In *Proc. DARPA Workshop on Broadcast News Transcription and Understanding*, Lansdowne, VA, Februar 1998.

Abbildungsverzeichnis

Tabellenverzeichnis

Index