Entwicklung eines türkischen Spracherkennungssystems für große Vokabulare

Diplomarbeit

Kenan Çarkı

Betreuung:

Prof. A. Waibel

Dipl.-Inform. Tanja Schultz

Institut für Logik, Komplexität und Deduktionssysteme
-Lehrstuhl Prof. A. Waibel-
Universität Karlsruhe

Karlsruhe, den 30. September 1998

Zusammenfassung

Im Rahmen dieser Arbeit wurde ein türkisches Spracherkennungssystem entwickelt. Die türkischen Daten enstammen dem GlobalPhone Projekt, welches die Erstellung mehrerer fremdsprachiger Datenbasen und eines multilingualen Erkenners zum Ziel hat. Es wurden dazu gelesene Artikel aus den Gebieten Innen-, Außenpolitik und Wirtschaft ausgewählt. Dadurch ist das Vokabular größer als bei einem eingeschränktem Task. Dieses hier entstandene Diktiersystem wurde mit Hilfe des Janus Speech Recognition Toolkits entwickelt. Ein spezielles Hauptmerkmal der türkischen Sprache ist, daß sie aufgrund der Eigenschaft der Agglutination, viele verschiedene Flexionen eines Wortes einbringt. Dieses hat ein starkes Vokabularwachstum zur Folge und damit auch einen hohen Anteil an unbekannten Wörtern in der Testmenge. In weiterführenden Ansätzen wurde das Problem dieser hohen 'Out-of-vocabulary'-Rate näher betrachtet und versucht durch Reduktion derselben mittels silbenbasierter Wortzerlegungen eine Steigerung der Erkennungsleistung zu erreichen. Die Wortfehlerrate des besten Systems liegt bei 34,1% mit einer OOV-Rate von 15,3%.

Inhaltsverzeichnis

Tabellenverzeichnis

Abbildungsverzeichnis

Kapitel 1

Einleitung

In der vorliegenden Arbeit wird die Entwicklung eines türkischen Spracherkennungssystems für gelesene Zeitungstexte beschrieben. Die Arbeit ist Teil des GlobalPhone Projektes an der Universität Karlsruhe, am Institut für Logik, Komplexität und Deduktionssysteme. Dieses Projekt kommt dem Wunsch nach multilingual sprachverarbeitenden Systemen nach, weil heutzutage immer mehr Menschen verschiedenster Herkunft und Sprache enger zusammenarbeiten. So ist es ein Ziel des GlobalPhone Projektes multilinguale Spracherkenner für weit verbreitete Sprachen zu erstellen. Hierzu wurde eine Datensammlungsaktion für diverse Sprachen gestartet [19]. Durch die Tatsache, daß ich die türkische Sprache spreche, wurde durch mich angeregt auch türkische Daten zu sammeln. So wurde durch die Datensammlung der Grundstein für ein türkisches Spracherkennungssystem gelegt. Beim Sprachspenden wurden Zeitungstexte vorgelesen, also diktiert, welches sich von spontaner Sprache durch sorgfältigeres Sprechen und grammatische Korrektheit unterscheidet und daher einfacher zu erkennen sein sollte. Das entwickelte Spracherkennungssystem ist somit ein Diktiersystem. Zur eigentlichen Entwicklung des Erkenners wurde das Janus Speech Recognition Toolkit [1] verwendet.

Die türkische Sprache hat bedingt durch ihre Eigenschaft der Agglutination, welches wörtlich 'anleimen' bedeutet, und sich auf die Konkatenation der Suffixe bezieht ein starkes Vokabularwachstum. Diese Suffixe haben jeweils nur eine Bedeutung, die eindeutig Auskunft beispielsweise nur über die Person (Sing./Plural) oder den Kasus geben. Man nennt diese Eigenschaft der Suffixe Monosemie[1]. Durch die Agglutination wächst das Vokabular stark, weil von vielen Worten verschiedenste Flexionen einfließen. Dies hat zur Folge, daß es eine hohe Out-of-vocabulary Rate[2] gibt und dementsprechend viele Wörter nicht erkannt werden können.

[1] das Vorhandensein nur einer Bedeutung zu einem Wort
[2] OOV-Rate: Prozentsatz der unbekannten Worte in der Testmenge

Im folgenden wird eine Übersicht der Kapitel dieser Arbeit gegeben.

Die türkische Sprache

Dieses Kapitel führt in die Grundlagen der türkischen Sprache ein und legt spezielle Eigenschaften dieser dar. Es werden dabei das Schriftsystem, die Phonetik, morphologische Eigenschaften und die Syntax vorgestellt.

Die Daten

Es wird in diesem Abschnitt auf die Datensammlung, die anschließende Aufbereitung und die Entwicklung einer Romanisierung beschrieben. Zudem wird das erste Sprachmodell vorgestellt, der Ausbau des Sprachmodellkorpus auf ca. 16 Mio. Worte und die verschiedenen Ausbesserungen an der Datenbasis erläutert.

Die Erkennungssysteme

In diesem Kapitel wird die Entwicklung der Spracherkennungssysteme beschrieben, von der ersten Initialisierung bis zum kontextabhängigen System.

Silbenbasierte Wortzerlegungsansätze

Hier werden weiterführende Ansätze zur Behebung der hohen OOV-Rate betrachtet. Es bot sich zur Reduktion der unbekannten Worte (OOV-Worte) eine zerlegungsbasierte Erkennung an. Wir haben im konkreten Fall Silben betrachtet. Dazu werden durch Zusammenfassen von Silben zu größeren Einheiten verschiedene Wortzerlegungen gebildet und für sieben der Möglichkeiten kontextunabhängige Systeme entwickelt. Für die drei besten Systeme werden dann zudem kontextabhängige Systeme entwickelt und die Ergebnisse miteinander verglichen.

Danksagung

Danken möchte ich an dieser Stelle:
Tanja Schultz, meiner Betreuerin, für die gute Zusammenarbeit und die vielen hilfreichen Anregungen, welche sie mir während der ganzen Zeit gab. Dong Hoon Van Uytsel für die entscheidende Idee zur Generierung der Silbentrennung. Klaus Ries für die Diskussionen

und die hilfreichen Tips bei der Entwicklung der zerlegungsbasierten Systeme. Petra Geutner für die Experimente zur Reduktion der OOV-Rate mittels HDLA[3]. Uwe Meier, daß er die Datensammlung mit mir gemeinsam angestoßen hat und damit erst diese Arbeit ermöglichte. Thomas Schaaf für das Korrekturlesen und die anregenden Diskussionen, welche ich vor allem in der Endphase der Arbeit mit ihm führen konnte. Mutlu Yalçın für das Erstellen der Datensammlung, für die Transliteration der Daten und daß sie mir mit ihren türkischen Fachkenntnissen hilfreich zur Seite stand. Und dann ihr als meiner Freundin, für ihre Unterstützung und dafür, daß sie stets an mich glaubte. Meinen Eltern Hidir und Saadet Çarkı und meinen Geschwistern Inan und Bircan, die mich während des gesamten Studiums, stets unterstützten und auf die ich immer zählen konnte.

[3]Hypothesis Driven Lexical Adaptation

Kapitel 2

Die türkische Sprache

Um ein Spracherkennungssystem für eine neue Sprache zu entwickeln, muß man sich zunächst mit der Sprache selbst vertraut machen. Für die Spracherkennung ist es wichtig, zu wissen, wie das Schriftsystem beschaffen ist, welchen Umfang die Phonemmenge besitzt und inwieweit es besondere Eigenschaften gibt, die man beachten sollte.

2.1 Die Sprache

Historie und Verbreitungsgebiet

Die türkische Sprache gehört zu der Familie der altaischen Sprachen, welche sich in drei Zweige aufteilt:

- Mandschu und tungusische Sprachen
- Mongolisch
- Turk- oder Türksprachen

Durch die Gemeinsamkeiten der drei Gruppen ist eine gemeinsame Abstammung sehr wahrscheinlich. Wobei wir schon gleich bei dem Problem der Einteilung in Sprachfamilien sind. Man muß hierbei vorsichtig vorgehen, und trifft dabei unter Umständen auf verschiedene Fachmeinungen. Früher wurde diese Familie beispielsweise häufig mit den uralischen Sprachen in eine uralisch-altaische Familie zusammengefaßt.

Das Verbreitungsgebiet der Turksprachen (auch früher turk-tatarisch genannt) erstreckt sich vom Westen beginnend in Thrakien bis tief nach Persien und Sibirien hinein. In dieser Arbeit handelt es sich, wenn hier von *Türkisch* die Rede ist, immer um das *Türkeitürkisch*, welches den größten Verbreitungsgrad hat. Heutzutage wird es, wie der Name schon sagt, in der Türkei gesprochen, welche aktuell ca. 65 Millionen Einwohner hat. Türkisch ist nur eine von vielen Einzelsprachen dieses Zweiges. So wird beispielsweise in fünf von sechs ehemaligen Sowjetrepubliken mit mehrheitlich islamischer Bevölkerung, eine Sprache gesprochen, die zur Familie der Turksprachen zählt. Das sind Aserbaidschanisch, Kasachisch, Kirgisisch, Usbekisch und Turkmenisch [22, 23].

2.2 Das Schriftsystem

Das ursprüngliche Schriftsystem war zu Zeiten des Osmanischen Reiches arabisch. Nach der Gründung der Türkischen Republik 1923, wurde 1928 ein neues Schriftsystem mit dem lateinischen Alphabet eingeführt.
In Tab. 2.1 sind alle 29 Buchstaben des Alphabets eingetragen. Es sind auch einige Aussprachebeispiele der Buchstaben angegeben. Dies wurde jedoch nur bei den Buchstaben gemacht, deren Aussprache sich von der im Deutschen unterscheidet.

2.3 Die Phonetik

Im Türkischen gibt es keine Diphthonge[1], gelegentlich gibt es Doppelvokale (z.B. m*aa*lesef). Sie werden so ausgesprochen, daß zwischen ihnen kein Stimmabsatz hörbar wird, und sie wie eine Verlängerung klingen. Bei Doppelkonsonanten hingegen wird jeder Bestandteil ausgesprochen (z.B. a*nn*e, man spreche an-ne) [17].

Phonemmenge

Das IPA [5] Konsonanten-Raster der 21 türkischen Konsonanten ist in Tab. 2.2 zu sehen [13]. Eine ausführlichere Darstellung ist in [9] zu finden.
Die 8 türkischen Vokale sind in Tab. 2.3 in das IPA Vokal-Raster eingetragen.

[1] aus zwei Vokalen gebildeter Laut, Doppellaut, wie z.B. au, eu im Deutschen

Tabelle 2.1: Türkisches Alphabet mit Aussprachebeispielen

Buchstaben	Türkisch	Deutsch
A a	araba	
B b	baba	
C c	cins	*Dsch*ungel
Ç ç	çek	*Tsch*eche
D d	dolmak	
E e	el	
F f	fren	
G g	güneş	
Ğ ğ	dağ, eğer	S*aa*l, *j*a
H h	halk	*H*ahn
I ı	ırmak	geh*e*n
İ i	ilk	
J j	jale	Gara*g*e
K k	kara	
L l	lezzetli	
M m	mutlu	
N n	nazar	
O o	okul	
Ö ö	öğrenci	
P p	para	
R r	radyo	
S s	sonuç	Ma*ss*e, Kla*ss*e
Ş ş	şöyle	*Sch*aum
T t	tarih	
U u	uzak	
Ü ü	üzüm	
V v	vermek	*W*asser
Y y	yara	*j*a
Z z	zarar	*S*aal,*S*eife

Tabelle 2.2: IPA Konsonanten-Raster im Türkischen

Konsonanten		Bilabial	Labiodent.	Dental-Alv.	Palatoalv.	Palatal	Velar	Glottal
Plosiv	stimmlos	p		t	ç		k	
	stimmhaft	b		d	c		g	
Frikativ	stimmlos		f	s	ş			
	stimmhaft		v	z	j			
Nasal		m		n				
Liquid	lateral			l				
	nonlateral			r				
Glide						y	ğ	h

Tabelle 2.3: IPA Vokal-Raster im Türkischen

	Rund		Unrund	
	Vorne	Hinten	Vorne	Hinten
Tief	ü	u	i	ı
Hoch	ö	o	e	a

2.4 Morphologie und Syntax

Eigenschaften der türkischen Sprache im Vergleich zum Deutschen

Agglutination Wörtlich bedeutet agglutinierend (lat. Ursprung) *anleimend*, das heißt in dieser Sprache wird zur Flexion des Wortes, der Mechanismus des Aneinanderhängens angewendet. Türkisch kann man agglutinierend nennen, weil das 'Anleimen' im starken Maße angewendet wird. Dies läßt sich an dem Beispielword 'ev'(Haus) verdeutlichen, 'evim' (mein Haus), 'evler' (Häuser), 'evlerim' (meine Häuser). Jede dieser Endungen hat nur *eine* Bedeutung, wobei hier '*ler*' für den Plural steht und '*im*' das Possessivsuffix ist, damit wird jede Funktion durch ein eigenes Suffix ausgedrückt. Ein besonders langes Wort ist in Tab. 2.4 [13] zu sehen.

Tabelle 2.4: Beispiel für Agglutination

Osmanlılaştıramayabileceklerimizdenmişsiniz
aufgeteilt in Morpheme:
Osman lı laş tır ama yabil ecek ler imiz den miş siniz

Die Bedeutung lautet engl:"(behaving) as if you were of those whom we might consider not converting into an Ottoman".
Türkisch hat bedingt durch die Eigenschaft der Agglutination eine hohe Vokabularwachstumsrate. Dies hat zur Folge, daß bei einem Test des Erkenners mit unbekannten Texten relativ viele Wörter nicht im Vokabular auftauchen. Dies sind die Out-of-Vocabulary Worte. Die Rate beträgt bei uns bis zu $16,8\%$.

Vokalharmonie Türkisch hält sich an das Prinzip der Vokalharmonie. Dies bedeutet, daß sich Vokale der Suffixe an den Stamm anpassen. Die Pluralbildung für die Wörter 'ev' (Haus) und 'çocuk' (Kind) lautet 'ev*ler*' (die Häuser) beziehungsweise 'çocuk*lar*'.

Kasus

Es gibt sieben Kasus im Türkischen, welche zur Übersicht in Tab. 2.5 dargestellt sind [4]. Dies sind die aus dem Deutschen bekannten vier Kasus Nominativ, Genitiv, Dativ und Akkusativ. Im Türkischen kommen zusätzlich Ablativ (woher?), Lokativ (wo?) und Komitativ/Instrumentalis (mit wem?/womit?). Zur Vertiefung empfehlen sich [13] und [10].

Tabelle 2.5: Die türkischen Kasus

Kasus	Beispiele	Bedeutung(deutsch)
Nominativ	ev, kapı	das Haus, die Tür
Genitiv	evin, kapının	des Hauses, der Tür
Dativ	eve, kapıya, çocuğa	dem Haus, der Tür, dem Kind
Akkusativ	evi, kapıyı	das Haus, die Tür
Ablativ	evden, kapıdan	von zu Hause, von der Tür
Lokativ	evde, kapıda	zu Hause, an der Tür
Komitativ/Instrumentalis	evle, kapıyla, çocukla	mit dem Haus, mit der Tür, mit dem Kind

Aufbau von Zahlen

Die Bildung der türkischen Zahlen erfolgt nach einer einfachen Systematik. Von der höchsten Zehnerstelle beginnend, werden die einzelnen Zahlen aneinandergehängt. Beispielsweise wird 1784 als 'Eintausend Siebenhundert Achtzig Vier' gelesen. Ausnahmen von dieser Regel gibt es nicht.

Kapitel 3

Die Daten

Bevor man mit dem Bau des Spracherkenners beginnen kann, ist es notwendig eine entsprechende Datenbasis zu erstellen. Dazu muß gesprochene Sprache, in diesem Fall türkische, aufgenommen werden. Mit diesen dann gesammelten Daten kann dann die Entwicklungsarbeit begonnen werden. Doch zunächst wird auf die Datensammlung selbst eingegangen, welche im Rahmen des GlobalPhone Projektes erfolgte. Die Daten sind nicht nur auf ein Themengebiet begrenzt, sondern umfassen die Bereiche Wirtschaft, Innen- und Außenpolitik.

3.1 Die Datensammlung

Woher sollte man die Audioaufnahmen erhalten? Um eine Beeinflußung der Sprache durch einen längeren Aufenthalt im Ausland, in dem Fall Deutschland, zu vermeiden, wurde die Entscheidung getroffen, die Daten am besten vor Ort in der Türkei zu sammeln. Zu diesem Zweck wurde eine wissenschaftliche Hilfskraft und Türkisch Muttersprachlerin eingestellt. Sie sollte in der Türkei die Daten sammeln. Zunächst mußten jedoch Texte gefunden werden, welche auch von den Spendern vorgelesen werden konnten.

Die Texte

Zu Beginn gestaltete sich die Textsuche schwierig. Die Möglichkeit Zeitungen vor Ort einfach zu kaufen und vorlesen zu lassen, schied aus, da der hohe Aufwand einer kompletten Transkription vermieden werden sollte. Es mußte eine Möglichkeit gefunden werden, Texte in elektronischer Form zu erhalten, dazu bot sich sinnvollerweise das WWW [1] an. Es begann die Suche nach einer türkischen Tageszeitung im WWW, jedoch waren die

[1] World Wide Web

größten türkischen Zeitungshäuser zu diesem Zeitpunkt (Mai 1996) noch in der Vorbereitungsphase der Internetauftritte. Andere hatten noch keine umfangreichen Archive erstellt, die nutzbar waren. Letztlich fand sich nur eine Zeitung, die genügend Daten bot: ZAMAN [2]. Der erste Gedanke, die Texte hier in Karlsruhe auszudrucken und dann nach İstanbul mitzunehmen, gestaltete sich als schwierig, da es nicht möglich war, Zeitungsartikel mit den türkischen Sonderzeichen zu erhalten. Ein erster E-mail Kontakt mit dem Webmaster der ZAMAN brachte aber die Zusage, die Artikel vor Ort auf Diskette als Write-Format-Dokumente [3] mit türkischen Fonts zu erhalten, andere große Tageszeitungen waren dazu beispielsweise nicht bereit.

Die Sprecher

Nach dem Drucken der Textvorlagen mußten nun 100 Personen gefunden werden, die bereit waren, ca. 15 Minuten lang Zeitungstexte vorzulesen.
Es wurde dabei darauf geachtet, daß das Sprecheralter zwischen 18 und 80 lag, und eine annähernde Gleichverteilung auch in Hinsicht auf das Geschlecht erfolgte. Dieses stellte jedoch ein Problem dar, weil Männer, vor allem über 35, weniger Interesse hatten, und daher die Datenbasis viele Frauen enthält (siehe Tab. 3.1).
Die Verwandten der wissenschaftlichen Hilfskraft unterstützten sie in İstanbul tatkräftig. Sie spendeten selbst Sprache, und 'überredeten' Nachbarn, Mitarbeiter, Kollegen, Schulkameraden, Kommilitonen, und quasi alle Leute dazu, ihren Teil durch eine Sprachspende zum Gelingen der Datensammlung beizutragen. So wurden auch beispielsweise in der Kantine einer Firma, während der Mittagspause Aufnahmen gemacht, wofür die Arbeiter einen Teil ihrer Pause opferten.

Der Zeitaufwand

Die Datensammlung in İstanbul dauerte vom 01.06.-12.06.1996. In diesen 11 Tagen hatte man mit vielen technischen Schwierigkeiten zu kämpfen. Die Texte mußten nochmals aufgearbeitet werden. Um die Quellen für die Sprachspender leicht lesbar zu machen, wurden die den Textfluß störenden Konstruktionen beseitigt, wie z.B Tabellen, Diagramme etc. Dies kostete mehrere Tage Zeit. Alleine das Ausdrucken war nur möglich, weil privat ein PC-Arbeitsplatz zur Verfügung gestellt wurde. Ohne die Hilfe des privaten Umfeldes, wäre der Erfolg der Datensammlung nie möglich geworden.

[2] Website: www.zaman.com. Zaman ist eine mittelgroße, fundamentalistische, türkische Tageszeitung
[3] Microsoft Write-Format

Die gesamte Datensammlung dauerte mit Vorbereitung, Aufnahmen und Reise 130 Stunden, wobei hier aber nicht die Zeit eingerechnet ist, welche dritte Personen beispielsweise mit dem Ausdrucken verbrachten.

Technisches

Die Aufnahmen wurden mit einem tragbaren DAT-Recorder Sony TDC-8 und einem Nahsprechmikrofon Sennheiser HD-440-6 gemacht. Ursprünglich mit 48 kHz 16 Bit mono aufgezeichnet wurden die Daten in 16 kHz 16 Bit mono umgesampelt.

Fakten zur Datenbasis

Eine Übersicht der wichtigsten Fakten ist in Tab. 3.1 zusammengefaßt. Zudem ist zu sagen, daß oftmals Aufnahmen mit Hintergrundgeräuschen, Gesprächen und damit sehr unruhigem Szenario vorlagen. Weiterhin sollte ebenfalls die Tatsache genannt werden, daß viele der Sprachspender die Spende als sehr stressig einstuften. Dies wirkte sich natürlich auch auf die Art des Lesens aus.

Tabelle 3.1: Datenbasis: Fakten

Sprecheranzahl	100
männlich	28
weiblich	72
Dialekte: hochtürkisch	92
Dialekte: anatolisch	8
Krank, Schnupfen oder Allergie	12
Raucher	42
Durchschnittsalter	27,5 J.
Alterspanne	13-53 J.
# Äußerungen ingesamt	6872
Laufende Worte	112 K
Vokabulargröße	16 K
Gesamtdauer	17h
Mittlere Länge der Äußerungen	8,9 sec
Gigabyte Daten (trl+audio)	2 GB

Hinweise und Anregungen

Vorgabe für das Szenario der Aufnahmen war es, einen ruhigen Hintergrund zu haben. Dies ist für saubere Aufnahmen wichtig und auch um die Konzentration der Sprecher beim Lesen nicht zu stören. Da aber kein spezieller Datensammelraum für die Sprachspenden zur Verfügung stand, sondern die Aufnahmen immer vor Ort bei den Sprachspendern vorgenommen wurden, konnte nicht immer gewährleistet werden, daß es ruhig war. Für die meisten Sprachspender war es eine ungewöhnliche Situation laut vorlesen zu müssen, zudem mit der Anweisung, bei fehlerhaftem Vorlesen den gesamten Satz zu wiederholen. Auch Personen, die nach eigenen Angaben sehr viel Zeitung lesen, stuften diese 15 Minuten als sehr stressig ein. Ein größeres Problem war jedoch, daß zu wenige Texte vorhanden waren, und daher einige mehrmals von verschiedenen Sprechern gelesen wurden (siehe Kap.3.3). Für zukünftige Datensammlungen ist es mittlerweile problemlos möglich eine entsprechend große Anzahl an Texten auszudrucken und diese dann in die Türkei mitzunehmen. Es würde sich auch anbieten, bereits vor der eigentlichen Reise über entsprechende Kontakte im Vorfeld freiwillige Sprachspender zu suchen. Die Belohnung der Sprachspender mit 'Give-Aways' z.B. im Form von GlobalPhone Kugelschreibern hat sich auch gut bewährt.

3.2 Datenaufbereitung

Die auf DAT-Tapes gesammelten Daten wurden auf einen Rechner übertragen und dort validiert. Das bedeutete, daß die gelesenen Sätze mit den Originalsätzen verglichen werden mußten und bei Fehlern dementsprechend eine Korrektur an der Referenz vorgenommen wurde.

Erstellung der Transliterationen

Das Testhören erforderte einen hohen Zeitaufwand. Für jeden Sprecher, der jeweils ca. 15 Minuten gesprochenen Text lieferte, sind vier Stunden Arbeitseinsatz zu rechnen. Es wurde dabei jeder Sprecher nur einmal angehört, also kein 2. Durchgang gemacht. Die Transliterationen wurden weiterhin um Markierungen für artikulatorische Geräusche ergänzt.

Validierung der Transliterationen

Nach dem einmaligen Hören der Sprecherdaten waren jedoch immer noch eine Vielzahl von Fehlern in den Datensätzen. So mußten noch mehrere Checks folgen, weil das Bearbeitungstool einige Fehler produzierte, wie z.B. falsches Abschneiden der Transliteration,

zu kurze Audiodateien etc. Andere Schwierigkeiten gab es mit fehlerhafter Rechtschreibung in den Zeitungsartikeln. Dadurch bedingt, wurde es nötig, mehrmals die Daten testzuhören und zu bereinigen.
Wie in Tab. 3.2 zu sehen, wurden in zwei größeren Stufen fast 1000 Äußerungen testgehört, wobei oftmals ein mehrmaliges Hineinhören nötig war, 4-5 mal war die Regel.

Tabelle 3.2: Verbesserungen der Transliterationen

Test 1.Stufe	
# Testgehörte Äußerungen	206
Test 2.Stufe	
# Testgehörte Äußerungen	> 700
# Änderungen an Dict	> 50

Bereinigung der Groß-/Kleinschreibung

Bei den ersten Tests des Erkenners zeigte sich, daß einige Fehler aus unterschiedlichen Schreibweisen resultierten. Dazu wurde ein Algorithmus zur Vereinheitlichung der Groß- und Kleinschreibung entwickelt.

Es werden nur diejenigen Wörter betrachtet, die einen großen Anfangsbuchstaben haben. Von diesen ausgehend werden Mutationen gebildet, um festzustellen, ob es von diesen Wörtern andere Schreibweisen gibt. So wird davon ausgegangen, daß es nur drei Varianten geben kann:
komplett klein geschrieben, erster Buchstabe groß und komplett groß geschrieben, z.B. in Überschriften. Es wird dann von diesen Varianten, wenn vorhanden, diejenige ausgewählt, die statistisch am häufigsten auftritt.

Skripten

Im Laufe der Arbeit wurden Skripten für die erwähnte Bereinigung und verschiedenste Aufgaben geschrieben. Um diesen Aufwand zu schätzen, wurden die Codezeilen aufsummiert. Die Anzahl der LOC (engl. lines of code) bewegt sich in der Größenordnung von 20000 Zeilen Code in diversen Programmiersprachen, wie Tcl, Perl und Shell-Programmierung.

Tabelle 3.3: Übersicht der entstandenen Datenbasen

Datenbasis	Kommentar
dbase-I	erstmals mit allen 100 Sprechern
dbase-II	Korrekturen an Transliterationen
dbase-LI	Bereinigung Groß-/Kleinschreibung
dbase-LII	Korrektur von weiteren Fehlern
dbase-LIII	Abschluß der Verbesserungen

Überblick Datenbasis

Im Laufe dieser Arbeit wurden die Transliterationen immer weiter verbessert und bei größeren Fehlerkorrekturen eine neue Datenbasis erzeugt. In der Übersicht sind diese in Tab. 3.3 abgebildet und jeweils mit einem kurzen Kommentar versehen.

3.3 Trainings- und Testdaten

Aufteilung in Training- und Testmenge

Für die Entwicklung eines Spracherkennungssystems benötigt man eine Trainingsmenge und eine Testmenge. Bei der Testmenge unterscheidet man nochmals in Entwicklungstestmenge und Evaluierungstestmenge.
Mit der Entwicklungstestmenge (engl. development testset) stellt man für den Test erkennerspezifische Parameter ein. Das zweite Testset ist das Evaluationset, um die endgültige Wortfehlerrate zu ermitteln.

Die drei Mengen sollten text- und sprecherdisjunkt sein. Bei der türkischen Datensammlung stellte dies ein Problem dar, da einige Texte mehrmals von verschiedenen Sprechern gelesen wurden. Dies machte es unmöglich die Mengen textdisjunkt zu trennen, ohne vorher gewisse Einschränkungen zu machen.

Folgende Lösung wurde angewandt. Es wurde festgestellt, daß einige Texte nur jeweils von einem einzigen Sprecher vorgelesen worden waren. Diese Sprecher bildeten die Basis unserer Testmengen. Man mußte jedoch beachten, daß bei diesen Sprechern nicht der komplette Datensatz benutzt wurde, sondern nur diese “Unikat”-Texte.

Im konkreten Fall waren es 22 “Unikat”-Texte und somit 22 Sprecher, die für die Testmengen in Betracht kamen. Jeweils die Hälfte kam in eine der Testmengen. Die restlichen 78 Sprecher bildeten die Trainingsmenge.

Trainingsdaten

Die Trainingsdaten umfassen 78 Sprecher. Genauere Angaben sieht man in Tab. 3.4.

Tabelle 3.4: Fakten zu Trainingsdaten

#Sprecher	#Sätze	#Worte	#Vokabular	Gesamtdauer	mittl. Dauer
78	5418	86624	12641	13h 4min 38sec	8,69sec

Testdaten

Es wurden verschiedene Testmengen erstellt, welche in der Übersicht Tab. 3.5 angegeben sind. Da Tests mit vielen Äußerungen sehr zeitintensiv sind, wurden kleinere Testmengen gebildet. Hierbei wurde darauf Wert gelegt, daß möglichst alle Sprecher beachtet wurden, und die Auswahl der Äußerungen nicht einseitig war. Mit den größeren Testmengen wurden vielversprechende Systeme getestet. Die kleineren dienten für erste Experimente. D100[4] ist eine Untermenge von D240. D11 und D14 haben nur eine gemeinsame Äußerung und sind Teilmengen von D240. E124[5] ist disjunkt zu den mit D bezeichneten Mengen. Die OOV-Rate OOV-1 wurde mit den ca. 30000 Worten aus Vokabular Vokab-T25M (S.26 Tab. 4.4), welches alle Trainingswörter und die 25000 häufigsten Wörter des größten Sprachmodellkorpus enthält, berechnet. Man erhlt nur 30000 Worte, weil etliche Worte in beiden Listen vorkommen. Für die OOV-Rate OOV-2 wurde mit dem Trainingsvokabular Vokab-T (S.26 Tab. 4.4), ca. 13000 Worte umfassend, gerechnet.

Die relativ hohe OOV-Rate sieht man auch in der Abbildung 3.1 in der die Self- und Crosscoverage[6] des gesamten Sprachmodell Korpus (15,67 Mio. Worte) abgebildet. Bei einer Vokabulargröße von ca. 60000 Worten hat man nur eine Crosscoverage von 85% und

[4] D steht für Development, die nachfolgenden Ziffern für die Anzahl der Äußerungen

[5] E = Evaluation

[6] Damit ist die Vokabularabdeckungsrate des Textkorpus gemeint. Im ersten Fall mit dem Korpusvokabular und dem Korpus selbst, und im zweiten Fall mit einem Testkorpus.

Tabelle 3.5: Fakten zu Testdaten

Testset	#Sprecher	#Sätze	#Worte	#Vokabular	Gesamtdauer	OOV-1	OOV-2
D11	11	11	224	205	2min 16sec	11,1%	20,1 %
D14	11	14	216	204	2min 10sec	13,2%	28,2 %
D100	11	100	1788	1161	17min 43sec	15,3%	27,0%
D240	11	240	4125	1997	40min 20sec	16,8%	28,1%
E124	10	124	2439	1350	21min 56sec	11,5%	23,2%

somit eine OOV-Rate von 15%. Diese 60K Worte werden später bei den Spracherkennungssystemen von Interesse sein, da dies auch an die Obergrenze des Vokabulars für die Tests stößt. Eine hohe OOV-Rate drückt aber die Performanz wesentlich. Jeder Prozentpunkt der OOV-Rate vermindert die Erkennungsleistung erfahrungsgemäß um 1-2%, weil das unbekannte Wort nicht erkannt wird und damit auch die ersten nachfolgenden Worte als Folgefehler falsch erkannt werden.

3.4 Entwicklung einer Romanisierung

Die türkischen Daten liegen in ISO 8859-9 Kodierung vor. Um die Transliterationen für Janus zu nutzen, erfolgte eine Umwandlung in 7-bit ASCII Zeichen. Es mußten die Sonderzeichen umgesetzt werden. Die Abbildung ist Tab. 3.6 zu entnehmen. Buchstaben, die nicht im Türkischen existieren, wie W, Q, X, w, q, x müssen trotzdem beachtet werden, da sie in Fremdwörtern auftauchen. Diese werden nicht romanisiert, sondern beibehalten, wie natürlich auch die anderen Nichtsonderzeichen. Es ist hierbei zu beachten, daß ein großes I2 für ein İ steht und ein kleines i2 für ein ı. In der anschließenden Romanisierung wurde dieser Sonderfall natürlich beachtet.

Im Rahmen der Romanisierung wurden Sonderzeichen, welche nicht für die Spracherkennung erforderlich sind, entfernt, wie Kommas, Punkte und andere Satzzeichen. Es ist noch darauf hinzuweisen, daß im Rahmen der türkischen Sprachreform in den 20er Jahren dieses Jahrhunderts anfangs übernommene arabische Zeichen, in Form von Vokalen mit Akzenten, nicht mehr gängig sind. In unseren Zeitungsquellen dennoch auftauchende Diakritiken wurden in ihre aktuelle Form überführt.

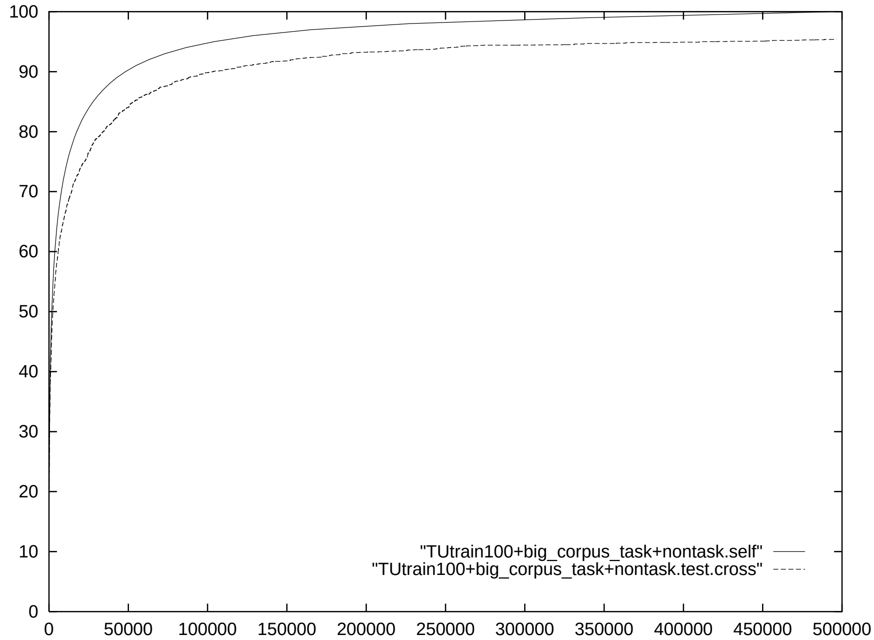

Abbildung 3.1: Self- und Cross-Coverage des 15,7 Mio Korpus

Tabelle 3.6: Romanisierungsabbildung

Türkische Darstellung	Romanisierung
Ç	TscH
Ğ	G2
Ö	O^
Ü	U^
İ	I2
Ş	Ssch
ç	tscH
ğ	g2
ö	o^
ü	u^
ı	i2
ş	sscH

Tabelle 3.7: Umsetzung veralteter Buchstaben

Alte Schreibweise	Umsetzung
â	a
î	i
í	i
û	u

3.5 Sprachmodell

Bei wahrscheinlichkeitsbasierter Spracherkennung, die hier angewandt wird, gibt es zwei wesentliche Komponenten, die zur Berechnung der Hypothese beitragen. Zum einen wäre da die akustische und zum anderen die sprachliche Komponente. Zur Vertiefung verweist man am besten auf [21, 24, 26]. Nun wird aber nachdem bereits auf die Trainingsdaten des Erkenners eingangen wurde, also dem Training der akustischen Komponente, auf die sprachliche eingegangen. Ein Sprachmodell liefert einem die Wahrscheinlichkeit für das folgende Wort. Dazu wird die Historie betrachtet. Es empfiehlt sich [24] zu konsultieren, um näheres zu erfahren. Verschiedene Sprachmodelle (engl. language model, Abkürzung LM) wurden mit dem Tool *ngrammodel* [18] erstellt. Es wurden zu Beginn die vorhandenen Texte aus der Trainingsmenge der Datensammlung genommen.

Perplexität und OOV auf der Trainingsmenge

Die ersten Ergebnisse ergaben auf 13k Vokabular eine Perplexität[7] von 130 und eine OOV-Rate von 14%. Wobei hier noch keine Bereinigung der Groß- und Kleinschreibung erfolgt war.

Verbesserung des Sprachmodells

Das ursprüngliche Sprachmodell auf dem GlobalPhone Korpus erstellt, wurde schrittweise erweitert, um eine zuverlässigere Schätzung der Wahrscheinlichkeiten zu erhalten. Daher wurden Bestrebungen angestellt, den Korpus für die Sprachmodellierung auszubauen.
Es gab zwei Stufen, in der ersten wurde als Korpusziel 1 Mio. Worte gesetzt, in der 2. Ausbaustufe nahmen wir das anspruchsvolle Ziel von 10 Mio. Worten ins Visier.

[7] Verzweigungsgrad, genaueres in [26]

Tabelle 3.8: Sprachmodellkorpus: Daten und Fakten

Quelle	Adresse	Themengebiete	# Worte
Zaman	http://www.zaman.com	task	1,02 Mio.
		nontask	1,20 Mio.
Milliyet	http://www.milliyet.com	task	3,92 Mio.
		nontask	3,81 Mio.
Hürriyet	http://www.huerriyet.com.tr	task	0,35 Mio.
		nontask	0,58 Mio.
Superhaber	http://www.superonline.com/superhaber	task	1,39 Mio.
		nontask	3,09 Mio.
Xn Online	http://www.xn.com.tr	task	0,11 Mio.
		nontask	0,16 Mio.

Tabelle 3.9: Übersicht: Task und Nontask Daten des Sprachmodells

Themengebiete	# Worte
task: Wirtschaft, Innen- und Außenpolitik	6,79 Mio.
nontask: sonstiges	8,84 Mio.
Summe	15,63 Mio.

Zusammenstellung des erweiterten Sprachmodellkorpus

Als mögliche Quellen kamen Texte aus dem Internet in Frage. Der Bestand von Zeitungswebsites war seit Beginn der Datensammlung stetig angewachsen und bot eine Auswahl von verschiedenen Verlagen.
Tasknah bedeutet, daß die Inhalte sich im Bereich der Trainingsdaten bewegen. Dies sind wie schon zu Beginn erwähnt: Innen-, und Außenpolitik, sowie Wirtschaftsthemen. In Tab. 3.8 sind die Zeitungsquellen, die Website und eine Angabe über die Anzahl laufender Worter, unterschieden nach task und nontask angegeben.
Am Ende standen als tasknahe Daten 6,79 Mio. Worte zur Verfügung und als taskfremde, bzw. nontask Daten 8,84 Mio. Wörter. Insgesamt umfaßte nun der Sprachmodellkorpus 15,63 Mio. Worte (Tab. 3.9). Für die Berechnung der Sprachmodelle wurde der Trainingskorpus einfach gewichtet eingebracht.

Endzustand der Sprachmodellierung

Mit diesen neuen Korpusdaten wurden zwei Sprachmodelle erstellt, eines welches die tasknahen Daten enthielt (LM-Task) und ein zweites zusätzlich noch mit nontask Daten (LM-Task+Nontask). Die Tab. 3.10 zeigt alle für die Tests relevanten Sprachmodelle. Mit den Bezeichnungen 'trainMost', 'taskMost' bzw. 'nontaskMost' sind die häufigsten Wörter des Trainingskorpus, des tasknahen Korpus und des taskfremden Korupus gemeint. Bei den größeren Sprachmodellen soll als Anmerkung noch die Perplexität (PP) und die OOV-Rate genannt werden. LM-T-Z hat eine PP von 43,09 und eine OOV-Rate von 13,09 %. LM-Task besitzt eine PP von 284,3 und LM-Task+Nontask 285,6 mit derselben OOV-Rate von 14,12 %.

Tabelle 3.10: Übersicht Sprachmodelle

Bezeichnung	Korpus	Vokabular
LM	train	train
LM-OOV-0	train	dev+trainMost
LM-T-Z	train+ 1 Mio. Zaman	train
LM-Task	train+Task	train+taskMost
LM-Task+Nontask	train+Task+Nontask	train+task/nontaskMost

Kapitel 4

Die Erkennungssysteme

In diesem Kapitel werden wir die auf die Entwicklung der ersten Spracherkennungssysteme eingehen und verschiedene Systeme für die türkische Sprache erstellen.

4.1 Phonemsystem

Die grundlegende Schwierigkeit bei der Entwicklung eines neuen Erkenners besteht darin, daß man dieses System sinnvoll initialisieren möchte. Mit Initialisierung ist das Festlegen der Initialgewichte gemeint. Der Begriff Gewichte bezieht sich dabei auf Kodebücher und Verteilungen.
Dieses erste System wurde mit den Gewichten eines multilingualen Erkenners für die Sprachen Deutsch, Spanisch, Englisch, Japanisch initialisiert. Dies erforderte eine Abbildung des türkischen Phonemsatzes auf die Phoneme des multilingualen Erkenners, welche akustisch am nächsten lagen (siehe Tab. 4.1).

Entwicklung eines Phonemsystems

Ein Fakt im Türkischen ist, daß alles so gelesen wird, wie es geschrieben ist. Nach einer Einführung zur Aussprache der diakritischen Zeichen im Türkischen (beispielsweise in Tab. 2.1), sollte man in der Lage sein, korrekt vorlesen zu können. Aus diesem Grund hält sich das gewählte Phonemset weitgehend an das türkische Alphabet. Es wurde anhand von [9] in Kap. 2.3 bereits die türkische Phonemmenge vorgestellt, indem es in das IPA-Raster nach Konsonanten und Vokalen getrennt eingetragen wurde. Nun wurde ein Phonemsystem für das Erkennungssystem festgelegt. Dieses Phonemsystem muß in einem rechnerkompatiblem Format sein.
Es wurde die Konvention eingeführt jeweils TU_ vor das entsprechende Phonem zu setzen.

Tabelle 4.1: Initialisierung der türkischen Phoneme

Türkische Phoneme	Multiling.
TU_A	DE_A
TU_B	DE_B
TU_C	EN_JH
TU_CH	DE_TSCH
TU_D	DE_D
TU_E	DE_E
TU_F	DE_F
TU_G	DE_G
TU_GJ	DE_J
TU_H	DE_H
TU_I2	DE_E2
TU_I	DE_I
TU_J	EN_JH
TU_K	DE_K
TU_L	DE_L
TU_M	DE_M
TU_N	DE_N
TU_O	DE_O
TU_OE	DE_OE
TU_P	DE_P
TU_R	DE_R
TU_S	DE_S
TU_SH	DE_SCH
TU_T	DE_T
TU_U	DE_U
TU_UE	DE_UE
TU_V	DE_V
TU_Y	DE_J
TU_Z	DE_Z
TU_+QK	DE_+QK
TU_+gHg	DE_+gHg
SIL	SIL

Dies diente zur Vereinfachung der Integrationsarbeit in einen multilingualen Erkenner. Die letzten drei Phoneme wurden für Stille, Fragmente[1] und 'human noise' [2] eingeführt. So ergab sich eine Menge von 32 Phonemen, mit denen wir weitergearbeitet haben. Das türkische Phonemsystem ist in Tab. 4.1 zusammengefaßt.

Erstellung eines Aussprachewörterbuches

Zu einer der arbeitsintensivsten, aber auch wichtigsten Aufgaben gehört die Erstellung eines Aussprachewörterbuches oder auch englisch 'Dictionary' genannt. Dieses muß zu jedem Worteintrag des Trainingskorpus die phonetische Umsetzung enthalten.
Was im Türkischen zu einem starken Wachstum des Wörterbuches führte, ist die Tatsache, daß Türkisch eine agglutinierende Sprache ist, und somit zur Folge hat, daß es viele ähnliche Wörter gibt, die sich nur in der Endung/Suffix unterscheiden.
Zur Erzeugung des Wörterbuches wurde die 'Graphem-to-Phonem' Methode gewählt, weil sich im Türkischen die Aussprache an das geschriebene Wort hält. Es ist dazu eine Abbildung eines Graphems auf das entsprechende Phonem zu erzeugen. Die Abbildung der Grapheme in Phoneme ist in Tab. 4.2 eingetragen. Dabei ist die Umsetzung der Zahlen jedoch nicht abgebildet. Es ist dazu anzumerken, daß Zahlen manuell umgesetzt wurden. Schwierigkeiten bereitet die Umsetzung der Zahlen insofern, daß zwar die Bildung der Zahlen nach dem in Kap. 2.4 vorgestellten Schema geschieht, jedoch Ordnungszahlen und Datumsangaben, sowie reelle Zahlen für Ausnahmen sorgen und damit alle manuell umzusetzen sind. Bei Datumsangaben wurden dabei die verschiedenen Aussprachevarianten berücksichtigt. Einige Sonderzeichen sind ebenfalls zu beachten, wie beispielsweise Prozentzeichen, Dollarsymbol etc. Bei Akronymen wurden zudem noch Aussprachevarianten automatisch eingefügt, da diese buchstabiert oder gelesen werden konnten. Auftretende Fremdwörter mußten jedoch manuell mit Aussprachevarianten versehen werden, weil nicht jeder Sprachspender die entsprechende Fremdsprache beherrschte, und daher diese Worte einfach auf 'Türkisch' aussprach.
Im Rahmen dieser Arbeit wurden verschiedene Wörterbücher erzeugt. Die für die Tests relevanten Wörterbücher werden in Tab. 4.3 vorgestellt. Durch eine Erweiterung des aus dem GlobalPhone Korpus erzeugten TUdict-I um die 25000 häufigsten Wörter des größten Korpus (siehe Kap. 3.5) gelangt man zu TUdict-II. Zu den beiden Wörterbüchern werden auch entsprechende Vokabularlisten angegeben, welche keine Testwörter enthalten, da diese das für den Test bekannte Vokabularinventar darstellen. Vokab-T enthält ausschließlich die Wörter des Trainings, wohingegen Vokab-T25M zusätzlich die 25000 häufigsten Wörter des größten Sprachmodellkorpus enthält. Diese sind in Tab. 4.3 und 4.4 angegeben.

[1] Wörter, die nicht vollständig gelesen wurden, bezeichnet man als Wortfragmente
[2] Für Häsitationen und Intrjektionen

Tabelle 4.2: Graphem zu Phonem Abbildung

Graphem	Phonem
I2	TU_I2
O^	TU_OE
o^	TU_OE
SscH	TU_SH
sscH	TU_SH
TscH	TU_CH
tscH	TU_CH
U^	TU_UE
u^	TU_UE
W	TU_W
a	TU_A
ag2	TU_A TU_A
og2	TU_O TU_O
ug2	TU_U TU_U
b	TU_B
c	TU_C
d	TU_D
e	TU_E
f	TU_F
g	TU_G
g2	TU_GJ
h	TU_H
i2	TU_I2
i	TU_I
j	TU_J
k	TU_K
l	TU_L
m	TU_M
n	TU_N
o	TU_O
p	TU_P
r	TU_R
s	TU_S
t	TU_T
u	TU_U
v	TU_V
w	TU_V
y	TU_Y
z	TU_Z

Tabelle 4.3: Wörterbücher

Bezeichnung	Einträge
TUdict-I	14433
TUdict-II	31330

Tabelle 4.4: Vokabular

Bezeichnung	Einträge
Vokab-T	12641
Vokab-T25M	30107

4.2 Bootstrapping

Um die Integration in das multilinguale Erkennungssystem zu erleichtern, wurde darauf geachtet, daß dieselbe Vorverarbeitung des Sprachsignals einsetzt, wie auch dieselbe Modellierung des Parameterraumes [20].

Da wir einen Erkenner für eine neue Sprache entwickeln werden, haben wir keine Parametereinstellungen, das heißt insbesondere keine Kodebücher und Verteilungen. Aus diesem Grunde haben wir den Erkenner mit den Gewichten des multilingualen Erkenners initialisiert.
Mit diesen Initialgewichten haben wir erste Labels berechnet, das heißt eine Zuordnung von Audioaufnahme zu der entsprechenden Transkription gemacht. Mit diesen Labels konnten wir dann ein erstes System und die nächsten Schritte zum Training des Systems angehen.

4.3 Kontextunabhängiges System

Der erste Ansatz führt zum kontextunabhängigen System (engl. context independent, CI-System), welcher bei der phonetischen Modellierung keinen Unterschied zwischen Phonemen in verschiedenen Kontexten macht.

Tabelle 4.5: Experimente kontextunabhängiger Systeme

System	dbase	LM	Vokab.	Testset	WE
TS1	dbase-LI	LM	TUdict-T	D14	51,5 %
TS1	dbase-LI	LM-OOV-0	TUdict-T	D11	42,6 %
TS1	dbase-LI	LM	TUdict-T	D100	61,1 %

Ergebnisse der Experimente

Für 14 Äußerungen hat sich folgendes Resultat nach vier Trainingsiterationen ergeben (Tab. 4.5). Es wurde hier mit einer OOV-Rate von 0 % (OOV-O) getestet, weil alle Wörter bekannt waren. Mit WE ist die Wortfehlerrate bzw. engl. 'Word Error Rate'[3] gemeint.
Das Ergebnis mit dem Language Model (LM-OOV-0), welches die Testwörter als Unigramme enthält ist um einige Prozente besser.

4.4 Kontextabhängiges System

Von dem bestehenden CI-System wurde darauf aufbauend der nächste Schritt gemacht, um zu einem kontextabhängigen System (engl. context dependent, CD-System) zu kommen.

Kontextbetrachtungen bei der Phonemmodellierung

Wenn, wie zuvor in Kap. 4.3, Phoneme ohne ihren Kontext betrachtet werden, erfolgt keine Differenzierung der Phoneme in Abhängigkeit ihrer Umgebung. Die Aussprache eines Phonems variiert jedoch je nach Nachbarphonemen und durch Beachtung dieser Nachbarn erhält man eine feinere Modellierung der Phoneme. Aus diesem Grund werden wir nun eine tiefgehendere Unterscheidung der akustischen Einheiten anwenden. Eine nötige Klassifizierung der Phoneme wurde anhand des eigenen Wissens als Muttersprachler und mit zu Hilfenahme von Fachliteratur aus dem Gebiet der türkischen Sprachwissenschaften [9] erstellt.

[3] $WE = \frac{\#Deletions + \#Insertions + \#Substitutions}{\#TotalRef.Length} * 100\%$

Entwicklung von phonetischen Fragen

In Tab. 4.6 ist die nun entwickelte Unterteilung der verschiedenen phonetischen Fragen dargestellt. Auf der linken Seite sind die Fragen, rechts die zugeordneten Phoneme zu sehen.

Polyphone

Polyphone sind markierte Phoneme, die das Zentralphonem darstellen, mit der Information, welches der linke und der rechte Kontext dazu ist. Es wurden zwei Systemzweige untersucht, die sich in der Breite des zu betrachtenden Kontextes unterschieden. Zum einen wurden Triphone mit Kontext +/- 1 und Quintphone mit Kontext +/- 2 betrachtet. Eine Einschränkung unter Janus ist, daß nur *ein* Phonem aus dem nächsten Wort mit beachtet werden darf. Zur Ermittlung der Polyphone wird die Trainingsmenge analysiert und alle gefundenen Polyphone werden als Baumstruktur gespeichert. Durch phonetische Fragen werden diese nach dem Kontext geballt.

Experimente mit kontextabhängigen Systemen

Der Vergleich in Tab. 4.8 zwischen dem Triphone (Systemzweig mit Bezeichnung b) und Quintphone System (Systemzweig mit Bezeichnung a) zeigt, daß anfangs das Quintphone System in der Performanz um 1-2 Prozentpunkte besser ist, als das Triphone System.
Die Systeme TS3a(Quintphone) und TS3b(Triphone) sind die ersten kontextabhängigen Systeme mit Gewichten, die acht Iterationen weitertrainiert wurden. Der nächste Schritt erforderte das Schreiben neuer Label, da es einige Verbesserungen an der Datenbasis gab. Dann wurde vier Iterationen mit TS4a/b trainiert, und zusätzlich mit größeren LMs getestet. Es wurde stets mit TUdict-I getestet (OOV-O).
Die folgenden Tests in Tab. 4.9 wurden mit dem Vokabular Vokab-T25M und dem Wörterbuch TUdict-II erzielt. Dies sind alle Test mit einer OOV-Rate zwischen 11,1% und 15,8%. Als Datenbasis wurde auch dbase-LIII benutzt. Bei der Benutzung des größeren Sprachmodells, welches auch nontask Daten enthält, wurde ein Ansteigen der Wortfehlerrate festgestellt. Dies kann von zwei Faktoren abhängig sein, zum einen, daß der Trainingskorpus nur einfach gewichtet eingebracht wurde und mit seinen ca. 90000 Worten in der Gesamtzahl von ca. 16 Mio. Worten nicht mehr so relevant ist und zum anderen, daß die nontask Daten auch nicht sehr hilfreich sind. In Abb. 4.1 ist die schrittweise Verringerung der Wortfehlerrate (WE) dargestellt.
In Tab. 4.10 ist weiteres System getestet worden. Ein kontextabhängiges System, wel-

Tabelle 4.6: Phonetische Fragen für das Türkische

PHONES	@ SIL TU_+QK TU_+hGH TU_A TU_B TU_C TU_CH TU_D TU_E TU_F TU_G TU_GJ TU_H TU_I2 TU_I TU_J TU_K TU_L TU_M TU_N TU_O TU_OE TU_P TU_R TU_S TU_SH TU_T TU_U TU_UE TU_V TU_Y TU_Z
NOISES	TU_+QK TU_+hGH
CONSONANT	TU_B TU_C TU_CH TU_F TU_G TU_GJ TU_H TU_J TU_K TU_L TU_M TU_N TU_P TU_R TU_S TU_SH TU_T TU_V TU_Y TU_Z
STOP	TU_P TU_T TU_CH TU_K TU_B TU_D TU_C TU_G
STOP-UNVOICED	TU_P TU_T TU_CH TU_K
STOP-VOICED	TU_B TU_D TU_C TU_G
FRICATIVE	TU_F TU_S TU_SH TU_V TU_J TU_Z
FRI-UNVOICED	TU_F TU_S TU_SH
FRI-VOICED	TU_V TU_Z TU_J
NASAL	TU_M TU_N
GLIDE	TU_Y TU_GJ TU_H
BILABIAL	TU_B TU_P TU_M
LABIODENTAL	TU_F TU_V
ALVEODENTAL	TU_D TU_L TU_N TU_R TU_S TU_T TU_Z
PALATOAlVEOLAR	TU_C TU_CH TU_J TU_SH
PALATAL	TU_G TU_K TU_L TU_Y
VELAR	TU_G TU_GJ TU_K TU_L
VOWEL	TU_A TU_E TU_I TU_I2 TU_O TU_OE TU_U TU_UE
VO-BACK	TU_A TU_I2 TU_O TU_U
VO-FRONT	TU_E TU_I TU_OE TU_UE
VO-FRO-UNROUND	TU_E TU_I
VO-FRO-ROUND	TU_OE TU_UE
VO-BAC-UNROUND	TU_A TU_I2
VO-BAC-ROUND	TU_O TU_U
VO-HIGH	TU_I TU_I2 TU_U TU_UE
VO-LOW	TU_A TU_E TU_O TU_OE
VO-LOW-UNROUND	TU_A TU_E
ROUND	TU_O TU_OE TU_U TU_UE
UNROUND	TU_A TU_E TU_I TU_I2

Tabelle 4.7: Anzahl der Polyphone im Trainingskorpus

# Triphones	42925
# Quintphones	140125

Tabelle 4.8: Experimente kontextabhängige Systeme

System	#Modelle	dbase	LM	Testset	WE
Quintphonsysteme					
TS3a	1500	dbase-LIII	LM-T-Z	D14	38 %
TS3a	1500	dbase-LIII	LM-T-Z	D11	40,4 %
Triphonsysteme					
TS3b	1500	dbase-LIII	LM-T-Z	D14	39,8 %
TS3b	1500	dbase-LIII	LM-T-Z	D11	41,7 %
Quintphonsysteme					
TS4a	1500	dbase-LIII	LM-Task	D11	29,6 %
TS4a	1500	dbase-LIII	LM-Task+Nontask	D11	30,5 %
TS4a	1500	dbase-LIII	LM-Task	E124	35,6 %
TS4a	1500	dbase-LIII	LM-Task	D100	36,2 %
Triphonsysteme					
TS4b	1500	dbase-LIII	LM-Task	D11	31,8 %
TS4b	1500	dbase-LIII	LM-Task+Nontask	D11	30,9 %
TS4b	1500	dbase-LIII	LM-Task	E124	35,6 %
TS4b	1500	dbase-LIII	LM-Task	D100	35,4 %

Tabelle 4.9: Experimente-2 kontextabhängiger Systeme

System	#Modelle	dbase	LM	Testset	OOV-Rate	WE
Quintphonsysteme						
TS5a	1500	dbase-LIII	LMtask	D11	11,1 %	27,4 %
TS5a	1500	dbase-LIII	LMtask	D100	15,8 %	34,1 %
TS5a-AV	1500	dbase-LIII	LMtask	D100	15,8 %	34,1 %
TS5a-AV	1500	dbase-LIII	LMtask+nontask	D100	15,8 %	34,6 %
Triphonsysteme						
TS5b	1500	dbase-LIII	LMtask	D11	11,1 %	28,3 %
TS5b-AV	1500	dbase-LIII	LMtask+nontask	D100	15,8 %	36,3 %

Tabelle 4.10: Experimente-3 kontextabhängiger Systeme

System	#Modelle	dbase	LM	Testset	OOV-Rate	WE
S12	3000	dbase-LIII	LMtask	D11	11,1 %	31,8 %
S12-ohne VTLN-Test	3000	dbase-LIII	LMtask	D11	11,1 %	32,3 %

ches 3000 Modelle hat und zudem mit Vokaltraktlängennormierung (VTLN) arbeitet. Dieses System wurde mit einem erweiterten phonetischen Fragenkatalog und VTLN trainiert. Von einem größeren Test mußte aus Zeitgründen abgesehen werden. Ein Grund für die schlechte Performanz könnte sein, daß in diesem Fragenkatalog, in der Frage nach 'stimmhaften' Phonemen nur die Konsonanten beachtet wurden, jedoch scheint dies unwahrscheinlich, da VTLN nie Performanzverluste bringt. Eine weitere Untersuchung dieses Sachverhaltes war zeitlich jedoch leider nicht möglich.

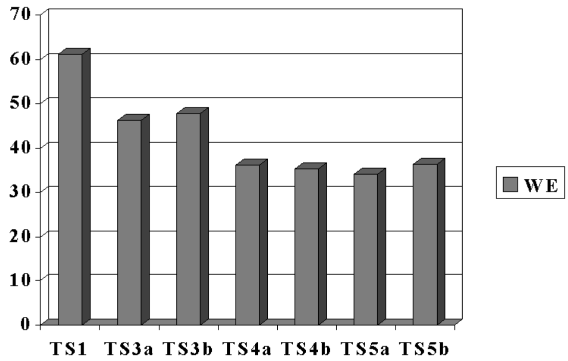

Abbildung 4.1: Vergleich der Systeme

Kapitel 5

Silbenbasierte Wortzerlegungsansätze

Die Motivation für die Betrachtung anderer Ansätze liegt in der Eigenschaft des Türkischen, viele neue Wörter in einer Testmenge einzubringen. Bedingt durch die Eigenschaft der Agglutination gibt es neue Wortbildungen durch Konkatenation anderer Suffixe. Wenn man nun die Betrachtungsebene von Worten auf kleinere Einheiten verlegt, sollte sich auch die OOV-Rate reduzieren lassen.

5.1 Vorüberlegungen

Um die morphologischen Bestandteile im Türkischen zu erhalten, wurde eine Möglichkeit der Zerlegung von Worten gesucht. Bemühungen in diese Richtung wurden auch in der Türkei gemacht. Anfang 1994 wurde eine 'Turkish Natural Language Processing Initative' gestartet, welche es sich zur Aufgabe gemacht hat, Vorarbeit für Forschung im Bereich Sprachverarbeitung und -erkennung zu leisten[14]. Im Rahmen dieser Initative enstand auch ein Tool zur Zerlegung türkischer Texte in ihre grammatikalischen, morphologischen Partikel [15]. Dieses Tool 'Xcorpus' ist jedoch interaktiv ausgelegt. In einem von zehn Fällen ist statistisch gesehen keine eindeutige Zerlegung möglich [16]. Dies liegt daran, daß versucht wird die korrekte, grammatikalische Zerlegung zu finden, was nicht immer möglich ist, da verschiedene Flexionen eines Wortes dieselbe Schreibweise besitzen können. Zudem liefert es im erfolgreichen Fall zwar die grammatikalisch korrekten Suffixe, welche jedoch aufgrund der Vokalharmonie in einer allgemeinen Form gehalten sind und somit nicht unmittelbar für eine Wortzerlegung zu nutzen sind. Das Tool baut auf PC-Kimmo, dem von Kimmo Koskenniemi entwickelten 'Two Level Processor for Morphological Analysis' auf.
Wegen der oben genannte Schwierigkeiten konnte 'Xcorpus' nicht eingesetzt werden. Es mußte eine andere Zerlegungsmöglichkeit gefunden werden. Hier kam eine andere Idee

gelegen, welche eine Trennung der Wörter in Silben ermöglichte [1]. Durch das Textsatzprogramm Latex[2] sollte der Textkorpus zerlegt werden, indem Latex mit entsprechender Vorlagenformatierung angehalten wurde jedes Wort zu trennen. Dafür wurde ein türkisches Latex, welches mit türkischen Trennungsregeln arbeitete, installiert und damit der Korpus zerlegt.

5.2 Silbentrennung

Zur Silbentrennung läßt sich sagen, daß sie im Türkischen nach Sprechsilben erfolgt z.B. as-tar, ıs-pa-nak [6, 8]. Der Zerlegungsvorgang beinhaltete mehrere Schritte, die in Ablaufplan 5.1 zu sehen sind.
Es wurde die Vokabularliste des gesamten Textkorpus gebildet, um den Vorgang der Zerlegung zu beschleunigen. In diesem Falle war es bedeutend schneller eine Vokabularliste aufzuteilen und später im Korpus alle Wörter durch ihr zerlegtes Pendant zu ersetzen, als den laufenden Text von ca. 16 Mio. Wörtern direkt zu bearbeiten.
Im ersten Schritt mußte eine Umsetzung der romanisierten Texte in türkische Latexfonts erfolgen. Die Abbildung ist in Tab. 5.1 erläutert.
Sonderzeichen, die in der Vokabularliste auftraten, mußten für Latex gequoted[3] werden. Im nächsten Schritt wurden zwei Jobs[4] gestartet, welche jeweils die Hälfte der ca. 500.000 Wörter zu zerlegen hatten. Es wurden dabei jeweils Blöcke von 1000 Wörtern genommen, da Latex sonst aufgrund der speziellen Konfigurierung (alle Wörter in einer langen Zeile) nicht in der Lage war, einen Output für alle Wörter zu erzeugen. Die silbengetrennte Version jedes Wortes wurde aus dem Latex-logfile extrahiert. Es war jedoch eine Nachbearbeitung der Silben nötig, weil es fehlerhafte Trennungen gab, in denen zum Teil nur einzelne Buchstaben auftraten, welches aber den türkischen Trennungsregeln zufolge nicht auftreten dürfte [8]. So wurden entsprechend der bekannten Regeln, einzeln auftretende Buchstaben, wenn sie am Wortanfang standen, mit der 2. Silbe verschmolzen bzw. am Wortende an die vorletzte Silbe gehängt. Einen Ausschnitt der Vokabularliste nach der Zerlegung gibt Tab. 5.2. Man sieht dabei zur linken das vollständige Wort und zur rechten die zerlegte Form.

[1] Idee von Dong Hoon Van Uytsel, Gastwissenschaftler an den Interactive System Labs
[2] Auf www.dante.de gibt es die Quellen.
[3] d.h. mit Slash Symbol versehen
[4] technisch bedingt aufgrund der Größe des Korpus

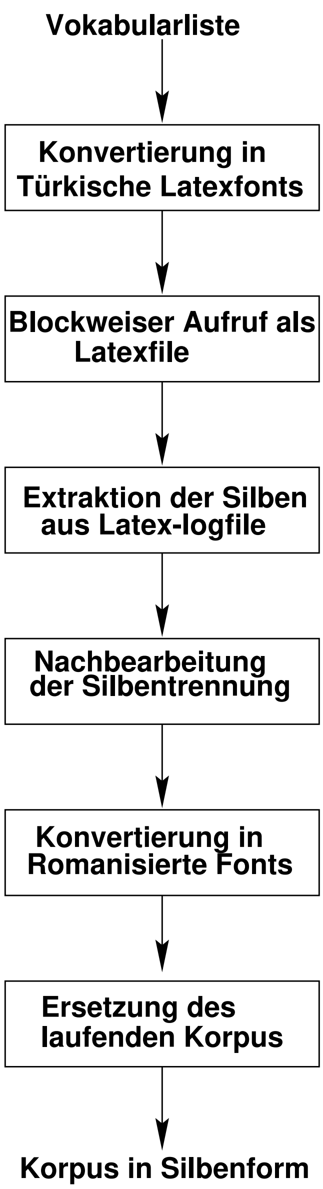

Abbildung 5.1: Ablaufplan der entwickelten Silbenzerlegung

Tabelle 5.1: Umsetzung romanisierter Texte in türkische Latexfonts

romanisierte Darstellung	türkische Latexfonts
Großbuchstaben	
TscH	C:
O^	O:
U^	U:
I2	I:
Ssch	S:
Kleinbuchstaben	
tscH	c:
g2	g:
o^	o:
u^	u:
i	i:
i2	i
sscH	s:

5.3 Systemspezifische Betrachtungen

Für das neue System sind einige Vorarbeiten zu erledigen. Von der Erstellung eines Silbenwörterbuches, dem Schreiben neuer Labels bis hin zu der Erstellung eines neuen Sprachmodells.
Die Schwierigkeit des Kontextverlustes im Sprachmodell bei Betrachtung von Worteinheiten durch Benutzung von Trigrammen, hat eine Erweiterung der Sprachmodellierungsmöglichkeiten in Janus erfordert. Diese Änderungen wurden in eine neue Janusversion eingebunden.

Erweiterung der Sprachmodellierung

Da das Türkische aufgrund seiner agglutinierenden Eigenschaften an einen Wortstamm mehrere Suffixe in einer grammatikalisch vorgegebenen Reihenfolge hängt, enstand die Idee kleinere Einheiten als Worte zu betrachten. Die in Kapitel 5.2 vollzogene Silbentrennung liefert uns diese Einheiten. In diesem Abschnitt ist mit Wort eine Silbe bzw. eine Kombination mehrerer Silben (Zerlegungseinheit) gemeint.
Die ursprüngliche Idee[5] ist es, daß bei der Berechnung der Wahrscheinlichkeit für ein Wort

[5] Implementierung durch Dong Hoon Van Uytsel

Tabelle 5.2: Ausschnitt der getrennten Vokabularliste

Einträge der Vokabularliste	Silbenzerlegung
..	..
Adaog2lu	Ada og2 lu
Adapazari2	Ada pa za ri2
Adapazari2na	Ada pa za ri2 na
Adapazari2nda	Ada pa za ri2n da
Adapazari2ndaki	Ada pa za ri2n daki
Adapazari2ndan	Ada pa za ri2n dan
Adapazari2ndayi2m	Ada pa za ri2n da yi2m
Adapazari2ni2n	Ada pa za ri2 ni2n
Adapazarli2	Ada pa zar li2
Adapazarli2lar	Ada pa zar li2 lar
Adapazarli2lari2	Ada pa zar li2 la ri2
Adapazarsi2z	Ada pa zar si2z
Adaplari2	Adap la ri2
Adar	Adar
Adara	Ada ra
Adari2n	Ada ri2n
Adasarhanli2	Ada sar han li2
Adasi2	Ada si2
Adasi2na	Ada si2 na
Adasi2nda	Ada si2n da
Adasi2ndaki	Ada si2n daki
Adasi2ndan	Ada si2n dan
Adasi2ni2	Ada si2 ni2
..	..

w_t nicht nur die beiden vorherigen Worte betrachtet werden, sondern auch die Klasse des letzten Wortes. Hierzu müssen Kriterien zur Klassenbildung gesetzt werden, um die Worte dementsprechend einzuteilen. Die Formulierung lautet in Formel 5.1

$$\boxed{P(w_t) = P(w_t|c_{t-1}, w_{t-1}, w_{t-2}) = P_{c_{t-1}}(w_t|w_{t-1}, w_{t-2})} \tag{5.1}$$

In diesem Fall wird ein Sprachmodell für jede Klasse gebildet. Wenn dieses 4-gramm nicht vorhanden ist, wird ein 'Backoff' angewandt. Dies bedeutet eine Betrachtung der nächst kleineren Einheit, des Trigramms bis hin zum Unigramm, daß heißt dem Wort selbst. Die Aussagekraft des dabei zu untersuchenden Bigramms $P(w_t|c_{t-1})$ ist jedoch nicht sehr groß, da von der Klasse des letzten Wortes schwer auf das aktuelle Wort geschlossen werden kann.
Dieser Ansatz wurde nun aufgegriffen und ausgebaut [6]. Die neue Berechnungsformel sieht dann wie folgt aus:

$$\boxed{P(w_t) = P(w_t|c_t, w_{t-1}, w_{t-2})} \tag{5.2}$$

wobei die Klassenwahrscheinlichkeit mit

$$\boxed{P(c_t) = P(c_t|c_{t-1}, w_{t-1}, w_{t-2})} \tag{5.3}$$

ermittelt wird. Der Unterschied ist hierbeit, daß in diesem 4-gramm nicht c_{t-1}, sondern die *aktuelle* Klasse c_t benutzt wird. Das dann beim Backoff zu betrachtende Bigramm $P(c_t|c_{t-1})$ ist viel stärker als das in der ursprünglichen Idee erwähnte Bigramm $P(w_t|c_{t-1})$. Um die Klassenzuordung für ein Wort zu erhalten gibt es die Funktion:

$$\boxed{c_t = c(w_t).} \tag{5.4}$$

Es wird versucht zum einen das Wort zu prädizieren und zum anderen die Klasse vorauszusagen.
Eine Trigrammanfrage wird auf die 4-gram Anfrage abgebildet. Die Idee beim Sprachmodell ein klassenbasiertes Backoff einzusetzen, wurde in ähnlicher Weise in [11] vorgestellt, wobei die Idee war, den Backoff auf Unigramme möglichst zu vermeiden.

Kopplung von wortbasiertem und klassenbasiertem Sprachmodell

Man hat durch diese erweiterte Sprachmodellierung eine Verzahnung von zwei Sprachmodellen erreicht, und zwar die des wortbasierten Sprachmodells mit dem klassenbasierten

[6] Erweiterung durch Klaus Ries

Sprachmodell. Man hat nach außen jedoch weiterhin nur ein Sprachmodell, welches angesprochen wird.

5.4 Vergleich verschiedener silbenbasierter Wortzerlegungen

Nach der Erweiterung der Sprachmodellierung folgt nun die Anwendung. Für die Bildung der Klassen wurde die Idee von "silbenbasierten Wortzerlegungen" entwickelt. Dies besagt, daß mit der erreichten Silbenzerlegung durch Zusammenfassung aufeinanderfolgender Silben, verschiedene Wortzerlegungen entstehen. Um verschiedene Zerlegungen vergleichen zu können, ohne jedesmal ein komplettes System zu erstellen, wurden diese analysiert.
Für die verschiedenen Zerlegungen wurde folgende Notation eingeführt (siehe Tab. 5.3).
Zu lesen ist die Notation wie folgt: In diesem Beispiel fällt die erste Silbe in Klasse POS0,

Tabelle 5.3: Notation für Zerlegungsbeispiel

Zerlegungsklasse	Elemente eines Wortes
POS0	0
POS1	1 2 +
POS2	3 FF

in POS1 fallen die zweite Silbe und die dritte Silbe miteinander verknüpft, was durch das + ausgedrückt wird. In die 3. Klasse POS2 fällt die 4. Silbe und alle folgenden, das FF bezeichnet dies.
Es wurden zwölf verschiedene Zerlegungen untersucht, welche in Tab. 5.4 und 5.5 dargestellt sind. Hier wird zudem noch die Anzahl der Elemente jeder Klasse angegeben und der Abdeckungsgrad für den 15,67 Mio. Korpus angegeben. Die Prozentzahl beinhaltet nur dem Vokabular bekannte Wörter, weil der Korpus nicht aus Platzgründen umgesetzt wurde, sondern eine Zähltabelle benutzt wurde, welche aber nur bekannte Wortuntereinheiten enthielt, und nicht die OOV-Worte.
In Tab. 5.6 werden alle Zerlegungen mit ihren wichtigsten Vergleichsgrößen angegeben.

Tabelle 5.4: Beschreibung der zwölf Zerlegungen im Überblick - Teil I

Zerlegung	Klasse	Klassenelemente	#Klassenvokab.	#Klassenelte.	Abdeckungsgrad
SP1	POS0	0	5049	12772173	43,04%
	POS1	1	2357	8959611	30,20%
	POS2	2 FF	1402	7939736	26,76%
SP2	POS0	0 1 +	13021	12772173	61,67%
	POS1	2	1293	5047558	24,37 %
	POS2	3 FF	567	2892178	13,96%
SP3	POS0	0 1 2 +	21097	12772173	81,54%
	POS1	3	521	2155674	13,76%
	POS2	4 FF	250	736504	4,7%
SP4	POS0	0 1 2 +	21097	12772173	81,54%
	POS1	3	521	2155674	13,76%
	POS2	4	235	602229	3,84%
	POS3	5 FF	129	134275	0,86%
SP5	POS0	0 1 2 +	21097	12772173	81,54%
	POS1	3	521	2155674	13,76%
	POS2	4	235	602229	3,84%
	POS3	5	119	116434	0,74%
	POS4	6 FF	71	17841	0,12%
SP6	POS0	0 1 +	13021	12772173	68,83%
	POS1	2 3 +	3762	5047558	27,2%
	POS2	4 FF	250	736504	3,97%

Tabelle 5.5: Beschreibung der zwölf Zerlegungen im Überblick - Teil II

Zerlegung	Klasse	Klassenelemente	#Klassenvokab.	#Klassenelte.	Abdeckungsgrad
SP7	POS0	0	5049	12772173	43,05%
	POS1	1	2357	8959611	30,20%
	POS2	2	1293	5047558	17,01%
	POS3	3 FF	567	2892178	9,74%
SP8	POS0	0	5049	12772173	43,05%
	POS1	1	2357	8959611	30,20%
	POS2	2	1293	5047558	17,01%
	POS3	3	521	2155674	7,27%
	POS4	4 FF	250	736504	2,47%
SP9	POS0	0	5049	12772173	51,87%
	POS1	1 2 +	9011	8959611	36,39%
	POS2	3 FF	567	2892178	11,75%
SP10	POS0	0	5049	12772173	51,87%
	POS1	1 2 +	9011	8959611	36,39%
	POS2	3	521	2155674	8,75%
	POS3	4 FF	250	736504	2,99%
SP11	POS0	0	5049	12772173	53,17%
	POS1	1 2 +	9011	8959611	37,30%
	POS2	3 4 +	1340	2155674	8,97%
	POS3	5 FF	129	134275	5,59%
SP12	POS0	0	5049	12772173	56,84%
	POS1	1 2 3 +	14331	8959611	39,88%
	POS2	4 FF	250	736504	3,28%

Tabelle 5.6: Vergleich verschiedener silbenbasierter Wortzerlegungen

Split	#Vokabgr.	#Worte Korpus(o.OOV-W.)	#Worte D240
SP1	8808	29671520	9363
SP2	14881	20711909	6744
SP3	21868	15664351	5131
SP4	21982	15664351	5131
SP5	22043	15664351	5131
SP6	17033	18556235	5988
SP7	9266	29671520	9363
SP8	9470	29671520	9363
SP9	14627	24623962	7750
SP10	14831	24623962	7750
SP11	15529	24021733	7550
SP12	19630	22468288	6995

Tabelle 5.7: Beispiel für die Zerlegung eines OOV-Wortes

Wortebene	yarati2lmasi2nda
Silben	ya- ra- ti2l- ma- si2n- da
SP2	yara_0 ti2l_1 ma_2 si2n_2 da_2
SP3	yarati2l_0 ma_1 si2n_2 da_2
SP6	yara_0 ti2lma_1 si2n_2 da_2

Erkennen von OOV-Wörtern

Mit diesen gebildeten Zerlegungen ist ein Zusammensetzen eines OOV-Wortes auf bekannten Einheiten möglich. Für die auch später weitergeführten Systeme SP2, SP3 und SP6 wird zur Verdeutlichung ein Beispiel in Tab. 5.7 angegeben.

5.5 Experimente zerlegungsbasierte Systeme

Prototypsystem

In Tab. 5.8 ist das Ergebnis des ersten Tests für das Prototypsystem PT1 zu sehen. Diese

Tabelle 5.8: Experimente Zerlegungs-Prototypsystem auf Zerlegungsebene

System	dbase	LM	Testset	OOV-Rate	WE
PT1	dbasePT1	LMPT1	D11	3,7%	51,5 %

"Word Error" Rate bezieht sich auf Silben. Zum Vergleich muß diese noch umgerechnet werden, in eine Wortfehlerrate auf Wortebene. Eine Betrachtung desselben Systems auf Wortebene, in dem nur die "FirstBest" Strategie[7] angewandt wurde, um von der Silbenebene zur Wortebene zu gelangen, ergab folgendes Ergebnis in Tab. 5.9.
Dieses Experiment bereitete den Weg für weitere Versuche, da dies auch als Test für

Tabelle 5.9: Experimente Zerlegungs-Prototypsystem auf Wortebene

System	Testset	WE
PT1	D11	62,9 %

diesen neuen Ansatz galt, in welchem eine neue Sprachmodellierung benutzt wurde. Mit dem Ziel der Reduzierung der Wortfehlerrate wurden Folgesysteme entwickelt.

Folgesysteme

Es wurden von den zwölf erzeugten Zerlegungen sieben ausgewählt und mit ihnen erste CI-Systeme erstellt. Bei der Auswahl wurde darauf geachtet, daß möglichst verschiedene Zerlegungstypen betrachtet wurden. In Tab. 5.10 ist auf der linken Seite der Typ der Zerlegung angegeben und auf der rechten die Zerlegungen, welche in diese Kategorie fallen. Zu Typ I gehören alle Zerlegungen, welche keine Silben zusammenfassen. Typ II faßt die Anfangssilben zusammen. In Typ III fallen die Zusammenfassung der ersten drei Silben und eine Erweiterung der Klassen hinein. Typ IV lässt die zweite und dritte Silbe zusammenfassen mit Variationen der Folgeklassen und Typ V fasst von der zweiten bis zur vierten Silbe zusammen. Alle ausgewählten Zerlegungen sind **fett** hervorgehoben. Für diese sieben mußte dann der gesamte Systementwicklungsprozeß einzeln durchgeführt werden. Dies beinhaltete, von der Erstellung des Wörterbuches, der Zerlegung der Datenbasis, der Berechnung der beiden Sprachmodelle (wort- und klassenbasiert) und der

[7] Eine Betrachtung der 'lattice' wurde hier nicht angewandt, nur das beste Wort wurde genommen, und nicht noch die rangnächsten Worte

Tabelle 5.10: Auswahl der sieben Zerlegungen

Typ	Zerlegung
I	SP1, **SP7**, SP8
IIa	**SP2**
IIb	**SP6**
III	**SP3**, SP4, SP5
IVa	SP9, **SP10**
IVb	**SP11**
V	**SP12**

anschließenden Initalisierung und des Trainings alle Schritte einer Neuentwicklung. Zur Intialisierung wurden die Gewichte des besten kontextunabhängigen Systems TS1 genommen.

Die Angaben für Datenbasis und Sprachmodell sind hier eindeutig, d.h. jedes System hatte hierzu eine eigene Datenbasis, in der die entsprechende Zerlegung angewandt wurde und ein Sprachmodell, welches über den Taskdaten des großen Sprachmodellkorpus (6,8 Mio.) berechnet wurde. Es wurde aus Zeit- und Platzgründen nicht der gesamte Korpus von knapp 16 Mio. Worten genommen. Die Größe des Vokabulars ist schon in Tab. 5.6 angegeben. Getestet wurde jeweils mit Testmenge D11.

Die Ergebnisse in Tab. 5.11 zeigen eine große Spanne in der Wortfehlerrate. Mit den

Tabelle 5.11: Experimente der sieben Zerlegungs-CI-Systeme

Split	WE (Zerlegungsebene)	WE (Wortebene)
SP2	40,0 %	44,8 %
SP3	37,1 %	38,6 %
SP6	36,3 %	39,5 %
SP7	47,5 %	57,0 %
SP10	43,7 %	53,4 %
SP11	49,1 %	53,8 %
SP12	48,9 %	52,5 %

Systemen SP2, SP3 und SP6, welche die ersten beiden Silben, bzw. sogar die ersten drei

Silben zusammenfassen, haben wir die besten Resultate. Die übrigen vier Systeme, welche die erste Silbe alleine in der ersten Zerlegungsklasse belassen, sind um bis zu 12 % auf Zerlegungsebene und bis zu 14 % auf Wortebene schlechter als das beste System. Durch Zusammenfassen der ersten mit einer oder mehreren Folgesilben ergibt sich eine höhere Erkennungsleistung, Eine mögliche Ursache ist, daß dadurch der Wortstamm besser modelliert wird.

Kontextabhängige Zerlegungssysteme

Mit den drei besten Systemen wird dann der Schritt zum kontextabhängigen System gemacht. Dies sind CD-SP2, CD-SP3 und CD-SP6.

Mittels eines selbst entwickelten Skriptes wurden nun auch die OOV-Rate auf Wortebene für diese besten Zerlegungen berechnet, d.h. alle OOV-Wörter wurden in die bekannten Bestandteile zerlegt, daß sich im Idealfall das komplette Wort aus den bekannten Fragmenten zusammensetzen ließ. Die OOV-Raten sind in Tab. 5.12 eingetragen. Man sieht, daß es eine Verminderung der OOV-Rate um bis zu 61 % gibt, jedoch schlägt sich das nicht direkt in einer besseren Erkennung nieder.

Tabelle 5.12: OOV-Raten der drei besten Zerlegungen für D100

System	OOV-Rate	Verbesserung (baseline=15,3 %)
SP2	6,0 %	-9,3 %
SP3	7,6 %	-7,7 %
SP6	6,8 %	-8,5 %

In Tab. 5.13 sieht man, daß wiederum CD-SP3 am besten abschneidet. Es fällt zudem auf, daß die Wortfehlerrate auf Zerlegungsebene um ein vielfaches höher ist, als die auf Wortebene. Zwischen den einzelnen Systemen, die im kontextunabhängigen Fall mit ihren Ergebnisse relativ nahe zusammenlagen, gibt es diesmal gravierende Unterschiede von bis zu 42 % auf Zerlegungsebene bei Test von D100. Das gute Abschneiden der Systeme auf Wortebene zeigt auch, daß auf Zerlegungsebene zwar Fehler dadurch entstehen, daß oftmals größere Einheiten erkannt werden, welche in der Referenz durch mehrere Einheiten modelliert sind, und dies damit zwangsläufig zu Fehlern auf Silbenebene führt, welche jedoch auf Wortebene durch das Zusammenfügen zu ganzen Worten behoben werden.
Die im Vergleich zu dem besten System aus Kap. 4.3 um 1,6 % schlechtere Erkennungsleistung ist damit zu begründen, daß es zum einen eine etwas schwächere Modellierung

des Aussprachewörterbuches gibt, welche keine Sonderbehandlung von Akrononymen, sowie von Aussprachevarianten hat. Zudem kommt hier die Eigenschaft der Datenbasis zum Tragen, daß aufgrund der hohen Sprechgeschwindigkeit der türkischen Sprachspender teilweise Endungen verschluckt wurden oder nicht immer richtig ausgesprochen wurden. Durch die nun detailliertere Betrachtung von Wortuntereinheiten können Endungen genauer erkannt werden, was uns in diesem Falle jedoch zum Nachteil gereicht, da dies Fehler produziert.
Das in Tab. 5.7 gezeigte OOV-Beispielwort wird auch von CD-SP2 erkannt und zusammengesetzt. Dies zeigt Tab. 5.14

Tabelle 5.13: Experimente der drei kontextabhängigen Zerlegungssysteme

System	Testmenge	WE (Zerlegungsebene)	WE (Wortebene)
CD-SP2	D11	32,3 %	35,0 %
CD-SP2	D100	37,9 %	39,0 %
CD-SP3	D11	72,3 %	30,0 %
CD-SP3	D100	79,9 %	35,7 %
CD-SP6	D11	48,0 %	33,6 %
CD-SP6	D100	55,4 %	37,0 %

5.6 Vergleich mit anderen Verfahren

Das Verfahren namens 'Hypothesis Driven Lexical Adaptation' wurde von Geutner in [2] vorgestellt. Es basiert auf der Idee, daß OOV-Wörter in agglutinierenden Sprachen oftmals aus bekannten Wörtern bestehen, welche lediglich eine anderes Suffix besitzen. Durch

Tabelle 5.14: Beispiel für ein erkanntes OOV-Wort in CD-SP2

Referenz	¨ yer_0 alan_0 sag2li2k_0 li2_1 tschevre_0 nin_1 YARATI2LMASI2NDA ¨
Hypothese	¨ yer_0 alan_0 sag2li2k_0 li2_1 tschevre_0 nin_1 YARA_0 TI2L_1 MA_2 SI2N_2 DA_2 ¨

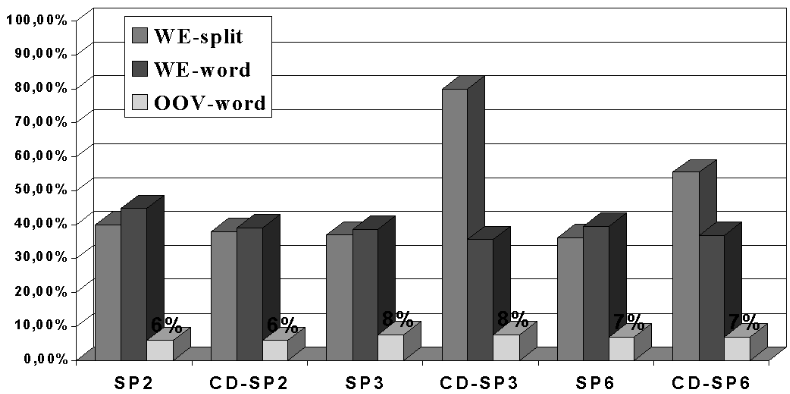

Abbildung 5.2: Vergleich der drei besten Zerlegungssysteme

die Hinzunahme von ähnlichen Wörtern in das bestehende Vokabular und Entfernen von nicht so guten Worten , die über den phonetischen Abstand oder über morphologische Ähnlichkeit ermittelt werden, erreicht man eine Reduktion der OOV-Rate. Mit dieser Neukonstruktion des Vokabulars wird ein nochmaliger Erkennungslauf gestartet. Dieser Ansatz zeigte beispielsweise für einen kroatischen Spracherkenner mit einer OOV-Rate von 8,7 %, eine Reduktion auf 4,8 %.
Für ein türkisches System wurden Experimente gemacht [3]. Dabei wurde jedoch nur der morphologische Ansatz benutzt. Das heißt jeweils die letzten 1-2 Silben wurden abgespaltet und der dadurch entstehende Wortstamm zur Ähnlichkeitsanalyse benutzt. In Tab. 5.15 ist die Reduktion der OOV-Rate zu erkennen, welche sich aber nicht ganz so verringert, wie bei den silbenbasierten Wortzerlegungen. Weitere Ergebnisse konnten aus Zeitgründen leider nicht berechnet werden.

Tabelle 5.15: Experiment zur Reduktion der OOV-Rate mittels HDLA

OOV-Rate (baseline)	OOV-Rate(morphology-based approach)
14,9 %	10,9 %

Kapitel 6

Zusammenfassung

Im Rahmen dieser Arbeit wurde ein türkisches Diktiersystem für große Vokabulare entwickelt. Die einzelnen Ergebnisse lassen sich wie folgt zusammenfassen.

Die für das GlobalPhone Projekt gesammelten Daten wurden in mehreren Stufen von Fehlern bereinigt und verbessert. Für die Arbeit mit den türkischen Daten wurde eine umkehrbare Romanisierung entwickelt. In einer weiteren Stufe wurde mit einem eigenen Algorithmus eine Vereinheitlichung der Groß-/Kleinschreibung erreicht, welche eine Fehlerreduzierung im Erkenner, sowie eine bessere Modellierung der Wahrscheinlichkeiten des Sprachmodells ermöglichte. Um eine bessere Schätzung der Wahrscheinlichkeiten im Sprachmodell zu erhalten, wurde der Textkorpus auf ca. 16 Mio. Worte ausgebaut.

Es wurden mit einer eigens eingeführten Phonemmenge mehrere Aussprachwörterbücher erstellt, mit einer halbautomatischen Behandlung von Aussprachvarianten für Akronoyme. Fremdwörter wurden zudem manuell mit Aussprachevarianten versehen.

Die erstellten Erkenner zeigten ansprechende Ergebnisse in den Testläufen, aber wiesen auch auf die Schwierigkeit der hohen OOV-Rate im Türkischen hin. Das beste System hatte eine Wortfehlerrate von 34,1 % mit einer OOV-Rate von 15,3 %. Dies veranlaßte Untersuchungen von Wortuntereinheiten, um durch diese die OOV-Rate zu reduzieren.

Es wurde eine Möglichkeit zur automatischen Generierung einer silbenzerlegten Worteliste mittels Latex entwickelt. Dann wurden silbenbasierte Wortzerlegungen eingeführt und untersucht. Dabei ergab sich, daß die Zerlegungsart wichtigen Einfluß auf die Erkennungsleistung besitzt. Die Ergebnisse zeigen, daß durch ausgewählte Zerlegungen unbekannte Worte aus dem bekannten Vokabular zusammengesetzt werden können. Damit ist eine Reduktion der OOV-Rate um bis zu 61 % erreicht worden, jedoch keine signifikante Verbesserung der Erkennungsleistung.

Für diese Betrachtung von Wortuntereinheiten wurde eine Verzahnung von klassenbasiertem und wortbasiertem Sprachmodell entwickelt, welche erstmals mit Hilfe des Janus Systems möglich war.

Letztlich wurde im Rahmen dieser Arbeit ein Grundstein für weitere Forschungen im Bereich der türkischen Spracherkennung gelegt.
In der Praxis wurde das System auch weiterverwendet, als es in einen multilingualen Erkenner [19, 20] integriert wurde und sich dort gut bewährte.

Literaturverzeichnis

[1] M. Finke, I. Rogina, M. Woszczyna , M. Westphal und T. Sloboda: *The JanusRTk Tutorial*, Interactive Systems Laboratories, Pittsburgh, PA, USA and Karlsruhe 1993-1997.

[2] Petra Geutner, Michael Finke, Peter Scheytt, Alex Waibel und Howard Wactlar: *Transcribing Multilingual Broadcast News Using Hypothesis Driven Lexical Adaptation*, Proceedings of the 1998 Broadcast News Transcription and Understanding Workshop, Lansdowne, Februar 1998.

[3] Petra Geutner, Michael Finke, Alex Waibel: *Selection Criteria for Hypothesis Driven Lexical Adaptation*, submitted for publication.

[4] Elvan Göçmen, Onur Şehitoğlu, Cem Bozşahin: *An Outline of Turkish Syntax*, Technical Report 95-2, Computer Engineering Department, METU, März 1995.

[5] *The IPA 1989 Kiel Convention*, Journal of the International Phonetic Association 1989(19) S. 67-82.

[6] *İmla Kilavuzu*, Atatürk Kültür, Dil ve Tarih Yüksek Kurumu, Türk Dil Kurumu Yayınları: 525, ISBN 975-16-0034-0, Ankara, 1993.

[7] *Turkish Language and Literature*, http://www.turkey.org/c_litera.htm.

[8] Prof. Dr. Zeynep Korkmaz: *Türçede Eklerin Kullaniliş Şekilleri Ve Ek Kaliplaşması Oyları*, Atatürk Kültür, Dil ve Tarih Yüksek Kurumu, Türk Dil Kurumu Yayınları: 598, ISBN 975-16-0602-0, Ankara, 1994.

[9] Jack Kornfilt: *Turkish And The Turkic Languages*, Bernhard Comrie(Editor): *The Word's Major Languages*, chapter 30, Oxford University Press, 1990.

[10] G. L. Lewis: *Turkish Grammar*, Oxford University Press, 1991.

[11] John Miller, Fil Alleva: *Evaluation of a Language Model using a Clustered Model Backoff*, http://www.asel.udel.edu/icslp/cdrom/vol1.htm, International Conference on Spoken Language Processing, Proceedings of ICSLP 1994.

[12] H.-H.Nagel: *Skriptum zur Vorlesung Kognitive Systeme (SS 1990 und SS 1991)*, Karlsruhe 1992 (überarbeitete und ergänzte Fassung vom 12.Juni 1992).

[13] Kemal Oflazer, Elvan Göçmen, Cem Bozşahin: *An Outline of Turkish Morphology*, Report on Turkish Natural Language Processing Initiative Project, Oktober 1994.

[14] Kemal Oflazer, H.Cem Bozşahin: *Turkish Natural Language Processing Initiative: An Overview*, Proc. of 3rd Turkish AI and Neural Network Conf, TAINN'94.

[15] Kemal Oflazer, İlker Kuruöz: *Tagging and Morphological Disambiguation of Turkish Text*, Proceedings of the 4th ACL Conference on Applied Natural Language Processing, Stuttgart, Germany Oct. 1994.

[16] Kemal Oflazer and Gökhan Tür: *Combining Hand-crafted Rules and Unsupervised Learning in Constraint-based Morphological Disambiguation*, Proceedings of ACL Conference on Empirical Methods in Natural Language Processing, Philadelphia, PA, Mai 1996.

[17] Pons, Taschenwörterbuch: Türkisch-Deutsch, Deutsch-Türkisch, Klett Verlag, ISBN 3-12-518820-2, 1993.

[18] Klaus Ries, Bernhard Suhm and Petra Geutner: *Language Modeling in JANUS*, Interne Dokumentation, Interactive Systems Laboratories, Pittsburgh, PA, USA and Karlsruhe 1995.

[19] Tanja Schultz, Martin Westphal and Alex Waibel: *The GlobalPhone Project: Multilingual LVCSR with JANUS-3*, Multilingual Information Retrieval Dialogs: 2nd SQEL Workshop, pp 20-27, Plzen, Czech Republic, April 1997.

[20] Tanja Schultz and Alex Waibel: *Multilingual and Crosslingual Speech Recognition*, Proceedings of the DARPA Broadcast News Workshop 1998, Lansdowne, Februar 1998.

[21] E. G. Schukat-Talamazzini: *Automatische Spracherkennung*, Braunschweig/Wiesbaden, Friedr. Vieweg & Sohn Verlagsgesellschaft mbH, 1995.

[22] *Sota Turkish World*, http://www.turkiye.net/sota/language.html

[23] Hans-Joachim Störig: *Abenteuer Sprache: Ein Streifzug durch die Sprachen der Erde* Humboldt-Verlag (in Zusammenarbeit mit Langenscheidt), 1997.

[24] A. Waibel & K. Lee: *Readings in Speech Recognition*, Pittsburgh 1990.

[25] A. Waibel: *Skript zur Vorlesung Kognitive Systeme*, Karlsruhe 1994.

[26] A. Waibel: *Foliensammlung zur Vorlesung Automatische Spracherkennnung*, Karlsruhe 1995.

Anhang A

Hinweise und Erklärungen

A.1 Sprachliche Aspekte

Phonetik und Phonologie

Zunächst eine kleine Erklärung der Begriffe und Grundlagen aus dem Bereich der Phonetik. Mit Phonetik bzw. Phonologie bezeichnet man die sprachwissenschaftlichen Forschungsgebiete. Der Ursprung liegt im griechischen Wort *phon*: Laut, Ton, Stimme. Phonetik ist die Naturwissenschaft, welche sich mit der Untersuchung der Sprachlaute auf physiologische Art beschäftigt. Es wird die Art und Weise des Zustandekommens eines Lautes betrachtet. Die Phonologie hingegen versucht festzustellen, welche Laute und Lautvarianten einer Sprache als bedeutungsunterscheidende Einheit dienen. Phoneme sind die kleinsten bedeutungsunterscheidenden sprachlichen Einheiten [23].
Wie kann man die bedeutungsunterscheidenden Einheiten erhalten? Durch Minimalpaare, d.h. zwei Wörter verschiedener Bedeutung, die sich nur durch ein Phon an der gleichen Stelle unterscheiden, ist dies möglich. Indem durch den Austausch zweier Phone eine andere Bedeutung erhalten wird, ist dies der Hinweis, daß es sich hierbei um Phoneme handelt. Zur Verdeutlichung kann man z.B. das Minimalpaar 'können gönnen' nehmen. Da diese verschiedenen Wörter nur in dem 'k' und dem 'g' unterscheiden, sind 'k' und 'g' Phoneme.

Wenn Phoneme phonetische Aussprachevarianten besitzen, nennt man diese *Allophone*, welche wiederum nach stellungsbedingten und freien Varianten unterschieden werden. Stellungsbedingte Varianten sind verschiedene Phone, die nicht in derselben lautlichen Umgebung auftreten können. Z.B. hat das Phonem /x/ den Laut [x], welcher nach hinteren Vokalen, wie in "Dach" vorkommt, und als stellungsbedingte Variante [c], das nach vorderen Vokalen wie in “Dich” auftritt. Die Freien Varianten treten in derselben lautlichen Umgebung auf und können miteinander vertauscht werden, ohne einen Bedeutungsunterschied hervorzurufen. Als Beispiel läßt sich da das gerollte oder das geschlagene r anführen.

Mit Hilfe der Phonetik soll es möglich sein, die Zahl der Phoneme einer Sprache zu ermitteln. Für jede Sprache ist diese verschieden und variiert zwischen 15 und 80 Phonemen, d.h. im Mittel kann eine Sprache mit ca. 40 Phonemen abgedeckt werden.

Es wird durch die Phoneme ein System gebildet, das nach Lautklassen getrennt ist. Die beiden großen Lautklassen sind die uns bekannten *Vokale* (Selbstlaute) und *Konsonanten* (Mitlaute).

Bei Vokalen schwingen die Stimmlippen im Kehlkopf und die Atemluft kann ungehindert ausströmen. Eine Unterscheidung wird insofern gemacht, daß es Monophthonge (Einzelvokale) und Diphthonge (Doppelvokale) gibt. Die Unterscheidung von Vokalen erfolgt durch die Merkmale Geschlossenheit, Helligkeit und Rundung. Geschlossenheit und Helligkeit sind bedingt durch die Stellung des Zungenrückens im Moment der Artikulation. Je höher der Punkt des Zungenrückens liegt, desto geschlossener ist der Vokal. Je weiter vorne im Mund der höchste Punkt des Zungenrückens liegt, desto heller ist der Vokal. Von einem runden Vokal spricht man, wenn die Lippen während der Artikulation gerundet sind.

Die Konsonanten sind die Laute, bei denen ausströmende Atemluft während einer gewissen Zeit gehemmt oder eingeengt wird. Bei Konsonanten unterscheidet man Artikulationsart und Artikulationsstelle.

Die Artikulationsart ist bestimmt durch die Art des Durchganges und des Widerstands des Luftstromes bei der Lautbildung. Man unterscheidet hier zwischen Verschlußlauten, Nasallauten, Seitenlauten, Schwinglauten, geschlagenen Lauten, Reibelauten und Affrikaten.

Mit der Artikulationsstelle bezeichnet man den Ort, an dem die beteiligten Organe bei der Artikulation zusammentreffen. Die entsprechenden Laute sind Lippenlaute, Lippenzahnlaute, Zahnlaute, Palatoalveolar, Vordergaumenlaute, Hintergaumenlaute, Zäpfchenlaute und Stimmritzenlaute.

Die eben erwähnten Fachbegriffe nochmals aufgelistet und erläutert.
1989 wurde auf einem internationalen Workshop von 120 Mitglieder der International

Tabelle A.1: Erklärung der wichtigsten linguistischen Begriffe

labial	an den Lippen
dental	mit Zähnen
dentalalveolar	zwischen Zähnen und Zahndamm
alveolar	am Zahndamm (obererer Zahndamm = Alveolardamm)
alveopalatal	zwischen Zahndamm und hartem Gaumen
palatal	am harten Gaumen (Palatum)
velar	am weichen Gaumen (Velum)
uvular	am Zäpfchen
pharyngal	an der Rachenwand (Pharyngis)
epiglottal	am Kehldeckel (Epiglottis)

Phonetic Association versucht einen gemeinsamen Symbolstandard zu verabschieden, mit welchem zukünftig eine offiziell anerkannte einheitliche Darstellung aller formbaren Laute aller Sprachen der gesamten Welt möglich sein sollte. Dieser Standard wurde kurz IPA [5] getauft, identisch mit dem Akronym des Verbandnamens.

Mit dem IPA Symbolalphabet wird es ermöglicht eine einheitliche Darstellung der Phonetik/Phonologie der Sprachen zu haben, einheitliche Transliteration von Gesprochenem zu erstellen und die (orthographische) Romanisierung von Sprachen, die mit fremden Schriftsystemen dargestellt werden, zu bewerkstelligen. Die ausreichend große Anzahl an IPA-Symbolen, soll gewährleisten, daß eine feine Unterscheidung der Akustik möglich ist.

Der Nachteil des IPA-Rasters ist, daß es nicht in einem computertauglichen Format ist. Die Mitglieder des Verbandes, welcher überwiegend aus Linguisten besteht, war nicht in der Lage, sich auf eine gemeinsame Codierung zu einigen. Sie wollten keine Ascii-Codierung, um sich nicht für die Zukunft durch eine Begrenzung auf 7 bzw.8 Bit, den Weg zu verbauen. Diese Entscheidungsnot hat dadurch den Erfolg des IPA-Symbolalphabets verhindert. Für diese Arbeit hat es sich aber als geeignet erwiesen.

Im IPA-System sind die Konsonanten in einem Raster angeordnet. Dieses Raster wird durch die oben erwähnten Merkmale des Artikulationsortes und der Artikulationsstelle unterteilt.

Die Anordnung der Vokale erfolgt nach 3 Gesichtspunkten. Erstens die Geschlossenheit, welche es in den Ausprägungen geschlossen, halb-geschlossen, halb-offen und offen gibt. Zweitens die Helligkeit, mit den Klassen vorne/hell - mitte - hinten/dunkel) und zuletzt die Rundung mit den Arten rund und unrund. In Tab.2.3 sind die türkischen Vokale eingeteilt, nicht vorhandene Klassen wurden hierbei nicht aufgeführt.

Ausgehend von dem IPA-System gab es einige Bestrebungen eine rechnertaugliche Kodierung zu entwickeln. Geeignet sind für die englische Sprache PHONASCII von George D. Allen, für die deutsche Sprache SAMPA von K. Kohler und WORLDBET von James L. Hieronymous.

WORLDBET ist nicht auf eine spezielle Sprache eingeschränkt, doch enthält dieser Ansatz einige Sonderzeichen, die bei der Verarbeitung mit JANUS Schwierigkeiten bereiten können. Es ist daher sinnvoll bei Namen mit Sonderzeichen andere Bezeichnungen zu wählen, sonst aber sich an den Phonemnamen zu halten, da die Verfasser von WORLDBET bemüht waren, intuitiv lesbare Namen zu finden.