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 entstand ein türkisches Spracherkennungssystem. Die türkischen Daten enstammen dem GlobalPhone Projekt, welches die Erstellung einer multilingualen Datenbasis und eines multilingualen Erkenners zum Ziel hat. Beim Sprachspenden wurden Zeitungstexte vorgelesen, also diktiert, welches sich von spontaner Sprache durch sorgfältigeres Sprechen und grammatische Korrektheit unterscheidet und daher einfacher zu erkennen ist. Dieses hier entstandene Diktiersystem wurde mit Hilfe des Janus Speech Recognition Toolkits entwickelt. Da die Artikel mehrere Gebiete umfassen, wie Innen-, Außenpolitik und Wirtschaft, ist das Vokabular größer als bei eingeschränktem kleinen Gebiet. Eine spezielles Hauptmerkmal der türkischen Sprache ist, daß sie aufgrund der Eigenschaft der Agglutination (lat.'anleimen'), viele verschiedene Flexionen eines Wortes einbringt. Dieses hat ein starkes Vokabularwachstum zur Folge und damit auch einen hohen Anteil an unbekannten Wörtern im Testset. In weiterführenden Ansätzen wurde das Problem der hohen 'Out-of-vocabulary'-Rate näher betrachtet und versucht durch Reduktion der OOV-Rate eine Verbesserung der Word-Accuracy zu erreichen.

Inhaltsverzeichnis

Tabellenverzeichnis

Abbildungsverzeichnis

Kapitel 1

Einleitung

In der vorliegenden Arbeit werden wir die Entwicklung eines türkischen Spracherkennungssystems für gelesene Zeitungstexte beschreiben. 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 wurde es Ziel des GlobalPhone Projektes einen multilingualen Spracherkenner für weit verbreitete Sprachen zu erstellen. Hierzu wurde eine Datensammlungsaktion z.B. für Arabisch, Portugiesisch, Russisch, Chinesisch und weitere Sprachen gestartet [16]. Durch die Tatsache, daß ich der türkischen Sprache mächtig bin, wurde durch mich und Dipl. Inform. Uwe Meier, wissenschaftlicher Mitarbeiter am Institut, 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 ist. Das zu entwickelnde 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 z.B. nur über die Person oder Kasus gibt. Man nennt diese Eigenschaft der Suffixe Monosemie. Dadurch wächst das Vokabular stark, weil von vielen Worten verschiedenste Flexionen einfliessen. Dies hat zur Folge, daß es eine hohe Out-of-vocabulary Rate [1] gibt, welche ein Problem für die Spracherkennung ist. Jeder OOV-Prozentpunkt vermindert die

[1] Prozentsatz der unbekannten Worte im Testset

Erkennungsleistung erfahrungsgemäß um 1-2 % . Da die Daten nicht nur auf ein Themengebiet begrenzt sind, welches ein kleines Vokabular hat, sondern die Bereiche Wirtschaft, Innen- und Außenpolitik umfassen, wird hier ein Erkennungssystem mit großen Vokabularen entwickelt.

Eine Übersicht über die folgenden Kapitel.

Türkische Sprache

Dieses Kapitel soll die Grundlagen der türkischen Sprache und spezielle Eigenschaften darlegen. Es wird dabei auf die Phonetik, Morphologie und das Schriftsystem eingegangen, um das Verständnis für die Systementwicklung zu erleichtern.

Daten

Es wird in diesem Abschnitt auf die Datensammlung, die anschließende Aufbereitung und die Romanisierung eingegangen. Zudem wird das erste Sprachmodell vorgestellt und Anmerkungen zu den verschiedenen Ausbesserungen an der Datenbasis gemacht.

Erkennungssysteme

Wir werden von 'scratch' aus die Entwicklung der Diktiersysteme beschreiben, ihre erste Initialsierung und den Schritt vom kontextunabhängigen zum kontextabhängigen System. Es wird dabei zudem auf die stufenweise Erweiterung des Sprachmodells eingegangen.

Morphembasierte Ansätze

Die Problematik der hohen Out-of-Vocabulary(OOV)-Rate im Türkischen machte es notwendig, auch weiterführende Ansätze zu betrachten. Es bot sich zur Reduktion der unbekannten Worte (OOV-Worte) eine morphembasierte Erkennung an. Wir haben im konkreten Fall Silben betrachtet. Hier werden verschiedene Wortzerlegungen durch Zusammenfassen von Silben zu größeren Einheiten betrachtet und für einige von ihnen kontextunabhängige Systeme entwickelt. Für die 'Top 2' dieser Systeme werden, dann zudem die Ergebnisse von kontextabhängigen Systemen verglichen.

Danksagung

Dank an:
Mutlu Yalçın, die mit stets viel Elan und großem Einsatz bei der Sammlung und dem späteren Transkribieren der Daten vorging. Auch für die enorme Leistung innerhalb kurzer Zeit, solche eine große Datensammlung zu erstellen.
Klaus Ries für die kollegialen und ergebnisreichen Gespräche bei der Entwicklung der morphembasierten Systeme.
Tanja Schultz, meiner Betreuerin, für die gute Zusammenarbeit und vielen hilfreichen Anregungen, welche sie mir während der ganzen Zeit gab.
Meinen Eltern und Geschwistern, die mich während des gesamten Studiums, stets unterstützten.
Mutlu Yalçın, diesmal als Freundin, für ihre Unterstützung und daß sie stets an mich glaubte.

Kapitel 2

Die türkische Sprache

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 nach dem Stand der letzten Volkszählung im Dezember 1997 ca. 65 Millionen Einwohner hat.

2.1 Die Sprache

Historisches und Verbreitungsgebiet

Die Türkische Sprache gehört zu der Familie der altaischen Sprachen, welche sich in 3 Zweige aufteilt.

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

Durch die Gemeinsamkeiten der 3 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 z.B. häufig mit den uralischen Sprachen in eine uralisch-altaische Familie unterteilt.

Das Verbreitungsgebiet der Turksprachen (auch früher turk-tatarisch genannt) erstreckt sich vom seinem westlichen Rand in Thrakien bis tief nach Persien und Sibirien hinein.

Die für uns interessante Sprache das sogenannte "Türkeitürkisch" ist nur eine von vielen Einzelsprachen dieses Zweiges. So wird z.B. in 5 von 6 ehemaligen Sowjetrepubliken mit mehrheitlich islamischer Bevölkerung, eine Sprache gesprochen, die zur Familie der Turksprachen zählt. Das sind Aserbaidschanisch, Kasachisch, Kirgisisch, Usbekisch und Turkmenisch.([19])

2.2 Schriftsystem

Das ursprüngliche Schriftsystem war zu Zeiten des Osmanischen Reiches arabisch. Nach der Gründung der Türkischen Republik, wurde 1928 ein neues lateinisches Alphabet eingeführt, welches das alte arabische ersetzte.[5]
In Tab. 2.1 sind alle 29 Buchstaben des Alphabets eingetragen. Als Vorgriff auf Kap.2.3 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 Phonetik-Laute der Sprache

Im Türkischen gibt es keine Diphthonge (wie z.B. au,eu im Deutschen), gelegentlich gibt es Doppelvokale (z.B. m*aa*lesef). Sie werden so ausgesprochen, daß zwischen ihnen kein Stimmabsatz hörbar wird, und wie eine Verlängerung klingen. Bei Doppelkonsonanten hingegen wird jeder Bestandteil ausgesprochen (z.B. a*nn*e, man spreche an-ne).[15]

Phonemmenge

Das IPA-Raster der 21 türkischen Konsonanten ist in Tab.2.2 zu sehen [11]: Bei weiterem Interesse siehe [7].

Die 8 türkischen Vokale sind in Tab.2.3 in das IPA Raster eingetragen:

2.4 Morphologie und Syntax

Eigenschaften der türkischen Sprache

Agglutination Wörtlich bedeutet agglutinierend *anleimend*, d.h in dieser Sprache wird statt der Flexion des Wortes, der Mechanismus des Aneinanderhängens angewendet.

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	
H h	halk	*H*ahn
I ı	imza	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	öğrençi	
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 Raster der türkischen Konsonanten

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 Raster der türkischen Vokale

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

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, die eindeutig Auskunft über Person und Kasus gibt ([19]). Jede Funktion wird durch ein eigenes Suffix ausgedrückt. Ein besonders langes Adverb ist in Tab. 2.4 [11]:

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 im Testset relativ viele Wörter nicht im Vokabular auftauchen. Dies sind die Out-of-Vocabulary Worte. Die Rate beträgt bei uns bis zu 16, 8%.

Weitere Eigenschaften Türkisch hält sich an das Prinzip der Vokalharmonie. Dies bedeutet, daß sich Vokale an den Stamm anpassen. Adverbien und Partikel sind nicht erweiterbar. Alle bestimmenden Aussagen stehen vor dem zu bestimmendem Wort.

Kasus

Es gibt 7 Kasus im türkischen, welchen zur Übersicht in Tab.2.5 dargestellt sind.[3]

Bei weiterem Interesse am Türkischen empfehlen sich [11] [8].

Tabelle 2.5: Die türkischen Kasus

Kasus	Beispiele
Nominativ	ev, kapı
Genetiv	evin, kapının
Dativ	eve, kapıya,çocuğa
Akkusativ	evi, kapıyı
Ablativ	evden, kapıdan
Lokativ	evde, kapıda
Komitativ/Instrumentalis	evle, kapıyla, çocukla

Kapitel 3

Die Daten

3.1 Die Datensammlung

Woher sollte man die Trainingsdaten, d.h. 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 die Wissenschaftliche Hilfskraft und türkisch Muttersprachlerin Mutlu Yalçın, eine Studentin der Sozialwissenschaften, 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 größten türkischen Zeitungshäuser zu diesem Zeitpunkt Mai 1996 noch in der Vorbereitungsphase der Internetauftritte. Andere hatten noch keine umfangreichen Archive erstellt, die wir nutzen konnten. Letztlich fand sich nur eine Zeitung, die uns genügend Daten bot: ZAMAN [2]. Der erste Gedanke, die Texte hier in Karlsruhe auszudrucken und dann mitzunehmen nach Istanbul, gestaltete sich als problematisch, da es nicht möglich

[1] World Wide Web
[2] Website: www.zaman.com

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 wri-Dokumente [3] mit türkischen Fonts zu erhalten.

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.
Die Verwandten unterstützten sie in İstanbul tatkräftig. Sie spendeten selbst Sprache, und 'überredeten' Nachbarn, Mitarbeiter, Kollegen, Schulkameraden, Kommilitonen, praktisch alle Leute dazu, ihren Teil durch eine Sprachspende zum Gelingen der Datensammlung beizutragen. So wurden auch z.B. in der Kantine einer Firma, während der Mittagspause Aufnahmen gemacht und die Arbeiter opferten einen Teil Ihrer Pause dafür.

Der Zeitaufwand

Der Aufenthalt in İstanbul[4] dauerte vom 01.06.-12.06.1996. In diesen 11 Tagen hatte man mit vielen technischen Problemen 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.
Die gesamte Datensammlung dauerte mit Vorbereitung, Aufnahmen und Reise 130 Stunden, wobei hier aber nicht die Zeit eingerechnet ist, welche dritte Personen z.B. 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 48kHz stereo gesampled wurden die Daten in 16 kHz 16 Bit mono umgewandelt.

[3] Microsoft Write-Format
[4] Dies ist die korrekte türkische Schreibweise

Fakten zur Datenbasis

Eine Übersicht mit den wichtigsten Fakten ist in Tab. 3.1 zusammengefaßt.

Tabelle 3.1: Datenbasis: Fakten

Sprecheranzahl	100
Trainingssprecher	78
Entwicklungstestsprecher	11
Evaluierungstestsprecher	11
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

Es empfiehlt sich die Aufnahmen in einem ruhigen Raum vorzunehmen, um eine saubere Aufnahme zu haben. Die Sprecher sollte die räumliche Gegebenheit nicht ablenken und ihnen die Möglichkeit zur Konzentration gegeben werden. 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 lasen, 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.A.1). Für zukünftige Datensammlungen ist es mittlerweile problemlos möglich eine entsprechend große Anzahl an Texten hier auszudrucken und diese dann in die Türkei mitzunehmen. Es würde sich

auch anbieten schon vor der eigentlichen Reise, über entsprechende Kontakte im Vorfeld freiwillige Sprachspender zu suchen. Die Belohnung der Sprachspender mit 'Give-Aways' im Form von z.B. GlobalPhone Kugelschreibern mit ansprechendem Design hat sich auch als gut bewährt, leider bekam nicht jeder einen.

3.2 Datenaufbereitung

Die auf DAT-Tapes gesammelten Daten wurden auf den 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 werden mußte.

Transliteration

Das Testhören hatte einen hohen Zeitaufwand. Für jeden Sprecher, der jeweils ca. 15 Minuten gesprochenen Text lieferte, sind 4 Stunden Arbeitseinsatz zu rechnen. Es wurde dabei jeder Sprecher nur einmal angehört, also kein 2. Durchgang gemacht. Im Mai 1997 war die Transliteration abgeschlossen.

Bereinigung

Nach dem einmaligen Hören der Sprecherdaten, waren jedoch immer noch eine Vielzahl von Fehlern in den Datensätzen. So mußten mehrere Checks noch folgen, weil das Bearbeitungstool, einige Fehler produzierte, wie z.B. falsches Abschneiden der Transkription, zu kurze Audiodateien etc. Andere Probleme gab es mit fehlerhafter Rechtschreibung der Zeitungsartikel, die auch mühevoll beseitigt werden mußten.

3.3 Romanisierung

Die türkischen Daten liegen in ISO 8859-9 Schrift vor. Um die Transliterationen für Janus zu nutzen, erfolgte eine Umwandlung in 7-bit ASCII Zeichen. Es müßten die Sonderzeichen umgesetzt werden. Die Abbildung ist Tab. 3.2 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.

Tabelle 3.2: Romanisierungsabbildung

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

Im Rahmen der Romanisierung wurden dabei auch andere störende Sonderzeichen, welche nicht für die Spracherkennung erforderlich sind, entfernt, wie Kommas, Apostrophe, und einiges mehr etc. 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 Apostrophen, nicht mehr gängig sind. In unseren Zeitungsquellen, denoch auftauchende Diakritiken wurden in ihre aktuelle Form übergeführt.

Tabelle 3.3: Umsetzung alter Buchstaben

Alte Schreibweise	Umsetzung
â	a
î	i
í	i
û	u

3.4 Language Model

Verschiedene Sprachmodelle (engl.language model, Abkürzung LM) wurden mit dem Tool *ngrammodel*[5] erstellt. Es wurden die vorhandenen Texte der Datensammlung genommen. welche jedoch ca. 112.000 Worte waren. Auf eine Vergrösserung des Korpus wird dann in Kap. 4.5 eingegangen.

Perplexität und OOV

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

3.5 Verbesserungen

Datenbasis

Bei der Arbeit mit der Datenbasis stellte es sich heraus, daß diese noch mit etlichen Fehlern behaftet war. Zu einem großen Teil waren diese durch einen 'Bug' des Transkribiertools produziert worden. So z.B. wenn Text und Audiodatei nicht übereinstimmten, da bei der Speicherung der Fehler reingebracht wurde. Dadurch bedingt, wurde es nötig mehrmals in größeren Aktionen, die Daten nochmals Testzuhören und zu bereinigen.

Tabelle 3.4: Verbesserungen an Datenbasis

Manualchecktest 1.Stufe (check-Skript)	
# Testgehörte Dateien	206
Aschenputtel 2.Stufe (Aschenputtel-Skript)	
# Testgehörte Äußerungen	> 700
# Änderungen an Dict	> 50

Wie in Tabelle 3.4 zu sehen wurden in 2 größeren Stufen fast 1000 Äußerungen testgehört, wobei oftmals ein mehrmaliges Hineinhören nötig war, 4-5 mal war keine Seltenheit. Wir kommen somit auf 4000-5000 Audiodateien, welche angehört wurden.

[5] von Klaus Ries geschrieben

Skripten

Im Laufe der Arbeit wurde Skripten für verschiedenste Aufgaben benötigt, um diese Aufwand zu schätzen wurden diese ingesamt gezählt. Die Summe bewegt sich in der Größenordnung von 15 K Code in diversen Programmiersprachen, wie Tcl, Perl und Shell-Programmierung.

Bereinigung der Groß-/Kleinschreibung

Bei den ersten Tests zeigte sich, daß beim Alignment einige Fehler aufgrund von verschiedenen Schreibweisen resultierten. Dazu wurde ein Algorithmus zur Vereinheitlichung der Groß- und Kleinschreibung entwickelt.

Algorithmus Idee: Es werden dabei nur die Wörter betrachtet, welche 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 3 Varianten geben kann.
Diese sind komplett klein geschrieben, erster Buchstabe groß und komplett groß geschrieben, z.B. in Überschriften. Es wird dann von diesen Varianten, wenn vorhanden, die ausgewählt, welche am häufigsten auftritt. Es muß noch darauf geachtet werden, daß die verschiedenen Varianten auf die letzlich meist vorkommende Version umgesetzt werden.

Überblick Datenbasis

Im Laufe dieser Arbeit wurde der Stand der Datenbasis immer weiter verbessert und damit auch bei entscheidenden Verbesserungen eine neue Datenbasis erzeugt. In der Übersicht sind diese in Tab. 3.5 abgebildet. Die Größe hat sich dabei nicht signifikant verändert, daher wird hier auch nur ein kurzer Kommentar angegeben.

Tabelle 3.5: 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

Kapitel 4

Erkennungssysteme

4.1 Initialisierung

Dieses erste System wurde mit den Gewichten des multilingualen Erkenners initialisiert. Dies erforderte eine Abbildung des türkischen Phonemsatzes auf die Phoneme des multilingualen Erkenners, welche akustisch am nächsten lagen (siehe Tab. 4.1).
Mapping der türkischen Phoneme auf die Phoneme des multilingualen Erkenners (für deutsch, spanisch, englisch, japanisch).

Phonemsystem

Es wurde in Kap. 2.3 anhand von [7] die türkische Phonemmenge vorgestellt, indem es in das IPA-Raster nach Konsonanten und Vokalen getrennt eingetragen wurde.
Nun muß ein Phonemsystem für das Erkennungssystem festgelegt werden. Dieses Phonemsystem soll in einem rechnerkompatiblem Format sein. Es wurde mit unseren muttersprachlichen Kenntnissen in Betracht gezogen, daß im Türkischen, alles so gelesen wird, wie es geschrieben wird. Nach einer Einführung zur Aussprache der diakritischen Zeichen im Türkischen, sollte man in der Lage sein, korrekt vorlesen zu können. Aus diesem Grund hält sich das von uns gewählte Phonemset weitgehend an das türkische Alphabet.
Es wurde die Konvention eingeführt jeweils TU_ vor das entsprechende Phonem zu setzen. Die letzten 3 wurden für Stille, Fragmente und 'human noise' eingeführt. So ergab sich eine Menge von 32 Phonemen mit denen wir weitergearbeitet haben.
Das Türkische Phonemset ist in Tab. 4.1 zusammengefasst.

Tabelle 4.1: Mapping der türkischen Phoneme auf die Phoneme des multilingualen Erkenners

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

Aussprachewörterbuch

Zu einer der arbeitsintensivsten, aber auch wichtigsten Aufgabe gehört die Erstellung eines Aussprachewörterbuches oder auch englisch Dictionary genannt. Dieses sollte zu jedem Worteintrag 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 jedoch nur in der Endung/Suffix unterscheiden, zum Ausdruck der Konjugation und Deklination.
Zur Erzeugung des Wörterbuches wurde die 'Graphem2Phonem' Methode gewählt, da sich im Türkischen die Aussprache an das geschriebene Wort hält. Es ist dazu nur eine Abbildung eines Graphems auf das entsprechende Phonem zu erzeugen. Die Datei *TUgraph2phon*, welche die Abbildung der Grapheme in Phoneme enthält ist in Tab. 4.2 eingetragen. Dabei ist die Umsetzung der Zahlen jedoch nicht abgebildet. Es ist dazu aber anzumerken, daß jede Zahl einzeln umgesetzt werden muss. Einige Sonderzeichen sind ebenfalls zu beruecksichtigen, wie z.B. Prozentzeichen, Dollarsymbol etc. Bei Akrononymen wurden zudem noch Aussprachevarianten automatisch eingefügt. Auftretende Fremdwörter mußten jedoch manuell mit Aussprachevarianten versehen werden.

Im Rahmen dieser Arbeit wurden verschiedene Wörterbücher erzeugt, die für die Tests relevanten werden in Tab. 4.3 vorgestellt. Durch eine Erweiterung des TUdict-I um die 25000 meisten Wörter des größten Korpus (siehe Kap. 4.5) gelangt man zu Tudict-II. In beiden Wörterbüchern sind die Testwörter enthalten, was aber Tests mit unbekannten Wörtern weiterhin ermöglicht, in dem ein Vokabular für die Tests angegeben wird, welches die Testwörter nicht enthält.

4.2 Bootstrapping

Vorverarbeitung

Alle 10ms werden aus einem kurzen Abschnitt des digitalisierten Sprachsignals mehrere Merkmale extrahiert. Diese werden dann zu einem Vektor zusammengefaßt. Um diesen Vektor noch in den Dimensionen zu reduzieren, wird eine LDA (Lineare Diskriminanz Analyse) gemacht, um letztendlich einen von 43 auf 32 Merkmalen reduzierten Vektor zu erhalten.

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

Erste Labels

Da wir einen Erkenner für eine neue Sprache entwickeln werden, haben wir keine Parametereinstellungen, d.h. Codebooks und Distributions. Aus diesem Grunde haben wir den Erkenner mit den Gewichten des multilingualen Erkenners initialisiert.
Mit diesen Initialgewichten haben wir erste Labels berechnet, d.h. eine Zuordnung von Audioaufnahme zu der entsprechenden Transkription gemacht. Mit diesen Labels können wir dann eine erstes System angehen, und die nächsten Schritte in Richtung Training des Systems angehen.

4.3 Trainings- und Testdaten

Trainingsdaten

Die Trainingsdaten umfassten 78 Sprecher. Genauere Angaben sind in Tab. 4.5.

Tabelle 4.5: Fakten zu Trainingsdaten

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

Testdaten

Es wurden verschiedene Testmengen erstellt, welche in der Übersicht Tab. 4.6 angegeben sind. Die OOV-Rate OOV-1 wurde mit dem Vokabular Vokab-T25M berechnet, die OOV-Rate OOV-2 mit dem Trainingsvokabular Vokab-T.

Tabelle 4.6: Fakten zu Testdaten

Testset	#Sätze	#Worte	#Vokabular	Gesamtdauer	mittl. Dauer	OOV-1	OOV-2
D11	11	224	205	2min 16sec	12,33sec	11,1%	20,1 %
D14	14	216	204	2min 10sec	9,28sec	13,2%	28,2 %
D100	100	1788	1161	17min 43sec	10,63sec	15,3%	27%
D240	240	4125	1997	40min 20sec	10,09sec	16,8%	28,1%
E124	124	2439	1350	21min 56sec	10,61sec	11,5%	23,2%

Die relativ hohe OOV-Rate sieht man auch in der Abb.4.1. Hier ist die Self- und Crosscoverage des gesamten Sprachmodell Korpus (15,67 Mio. Worte) abgebildet. Bei einer Vokabulargrösse von selbst ca. 60000 Worten hat man nur eine Crosscoverage von 85%.

4.4 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.

Experimente

Für 14 Äußerungen hat sich folgendes Resultat nach 4 Trainingsiterationen ergeben (Tab. 4.7. Da hier alle Wörter bekannt waren wurde also mit OOV-O getestet. Mit WE ist die 'Word Error Rate'[1] gemeint.

Das Ergebnis mit dem Language Model (LM-OOV-0), welches die Testwörter als Unigramme enthält ist um einige Prozente besser.

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

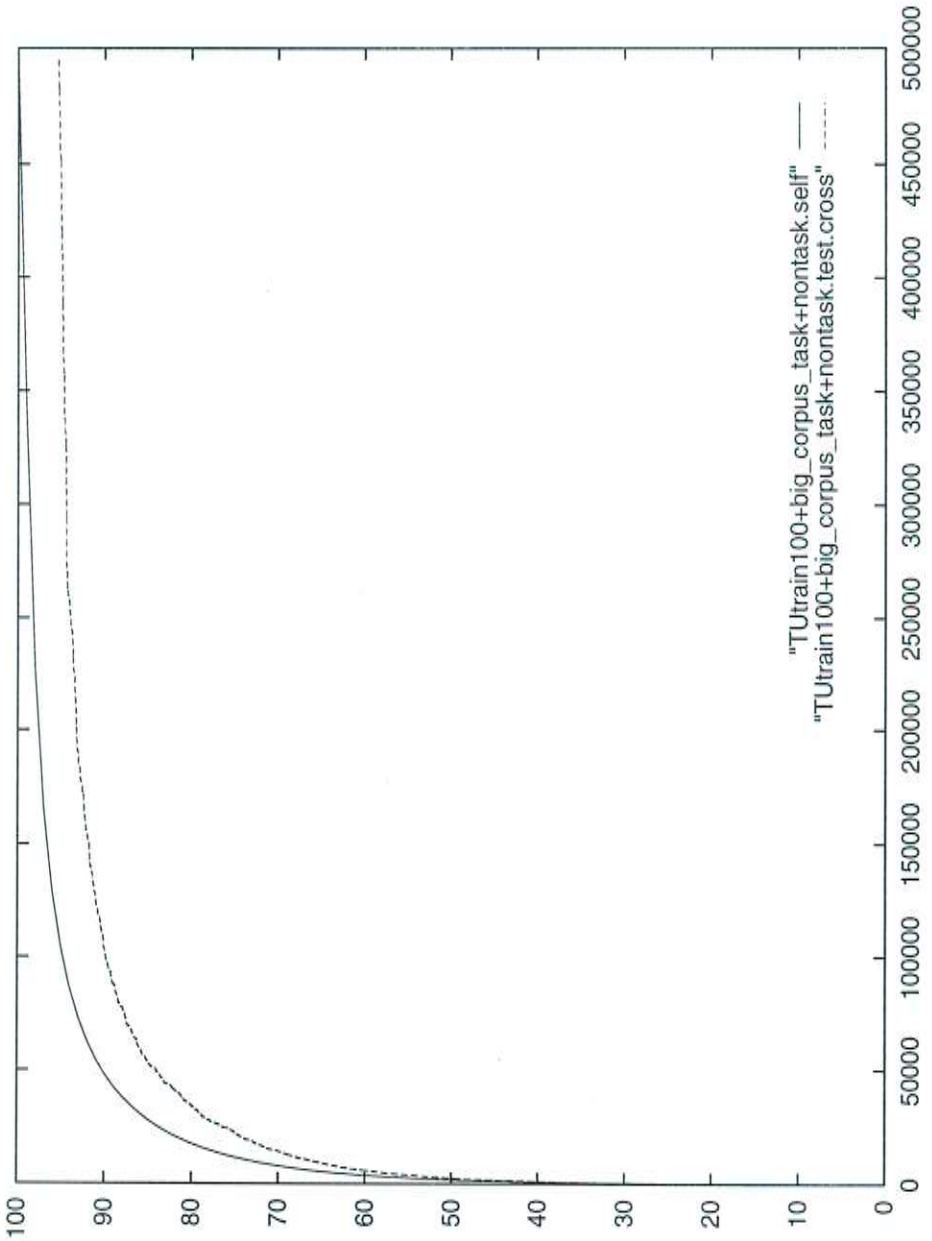

Abbildung 4.1: Self- und Cross-Coverage des 15,7 Mio Korpus

Tabelle 4.7: Experimente kontextunabhängige 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 %

4.5 Verbesserung des Sprachmodells

Das ursprüngliche Sprachmodell auf dem GlobalPhone Korpus erstellt, wurde schrittweise erweitert, da um die Aussagekraft des Sprachmodells zu steigern, der vorhandene Korpus noch zu klein war. Daher wurden Bestrebungen gemacht den Korpus für die Sprachmodellierung auszubauen.
Es gab 2 Stufen, in der ersten wurde als Korpusziel 1 Mio. gesetzt, bei der 2. Ausbaustufe nahmen wir das anspruchsvolle Ziel von 10 Mio. ins Visier.

Korpus

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 verschiendenen Verlagen.
Tasknah bedeutet, daß die Inhalte sich im Bereich der Trainingsdaten bewegen. Dies sind wie schon zu Beginn erwähnt: Innen-, und Außenpolitik, sowie Wirtschaftsthemen.

Tasknahe Daten sind 6,79 Mio. Worte und taskfremd, bzw. nontask Daten sind 8,84 Mio. Wörter. Insgesamt also 15,63 Mio. Wörter.

Für die Sprachmodelle wurde noch der Trainingscorpus hinzugenommen und zwar einfach gewichtet eingebracht.

Fakten

Mit diesen neuen Korpusdaten wurden 2 Language Models erstellt, eines welches die tasknahen Daten enthielt LM-Task und ein zweites zusätzlich noch mit nontask Daten LM-Task+Nontask. Die Tab. 4.10 zeigt alle für die Tests relevanten Language Models.

Tabelle 4.8: Korpus: 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/suberhaber	task	1,39 Mio.
		nontask	3,09 Mio.
Xn Online	http://www.xn.com.tr	task	0,11 Mio.
		nontask	0,16 Mio.

Tabelle 4.9: Übersicht: Task and Nontask Daten des Sprachmodells

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

Tabelle 4.10: Übersicht Language Models

Bezeichnung	Korpus	Vokab
LM	train	train
LM-OOV-0	train	dev+trainMost
LM-T-Z	train+ 1Mio Zaman	
LM-Task	train+Task	
LM-Task+Nontask	train+Task+Nontask	

4.6 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

Wenn, wie zuvor in Kap.4.4, Phoneme ohne ihren Kontext betrachtet werden, erfolgt keine Differenzierung der Phoneme in Abhängigkeit ihrer Umgebung. 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 ([7]) erstellt.

Phonetische Fragen

In Tab. 4.11 ist die von uns gewählte Unterteilung der verschiedenen phonetischen Fragen dargestellt. Auf der linken Seite sind die Fragen, gefolgt von den Phonemen.

Polyphone

Polyphone sind markierte Phoneme mit der Information, welches dieser dieser Phoneme das Bezugsphonem oder Zentralphonem ist. Es wurden 2 Systemzweige gebildet, die sich in der Breite des zu betrachtenden Kontextes unterschieden. Zum einen wurden Triphones mit Kontext +/- 1 und Quintphones mit Kontext +/- 2 betrachtet. Eine Einschränkung unter Janus ist, daß nur 1 Phonem aus dem nächsten Wort betrachtet werden darf. Zur Ermittlung der Polyphone wird die Trainingsmenge betrachtet und alle gefundenen Polyphone werden als Baumstruktur gespeichert.Durch phonetische Fragen werden diese nach dem Kontext geclustert.

Experimente mit kontextabhängigen Systemen

Der Vergleich in Tab. 4.13 zeigt, daß zwischen dem Triphone und Quintphone System das Quintphone System in der Performanz um einige Prozentpunkte besser ist, als das Triphone System. Trotz der Tatsache, daß die Quintphone weniger Trainingsmaterial zur Verfügung haben, als Folge der größeren Kontextbetrachtung. Die Systeme TS3a/b sind

Tabelle 4.11: Phonetische Fragen

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 4.12: Anzahl der Polyphone im Trainingskorpus

# Triphones	42925
# Quintphones	140125

Tabelle 4.13: Experimente kontextabhängige Systeme

System	#Modelle	dbase	LM	Testset	WE
TS3a	1500	dbase-LIII	LM-T-Z	d14	38 %
TS3a	1500	dbase-LIII	LM-T-Z	d11	40,4 %
TS3b	1500	dbase-LIII	LM-T-Z	d14	39,8 %
TS3b	1500	dbase-LIII	LM-T-Z	d11	41,7 %
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 %
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 %

die ersten kontextabhängigen Systeme mit Gewichten, die schon 8 Iterationen weitertrainiert wurden. Nochmals 4 Iterationen wurde mit TS4a/b trainiert, und dazu noch mit größeren LMs getestet. Es wurde stets mit der aktuellsten Version von TUdict-I getestet (OOV-O).

Die folgenden Test in Tab.4.14 wurden mit dem Vokabular Vokab-T25M und dem Wörterbuch: TUdict-II erzielt. Als Datenbasis wurde auch dbase-LIII benutzt.

Tabelle 4.14: Experimente-2 kontextabhängige Systeme

System	#Modelle	dbase	LM	Testset	OOV-Rate	WE
TS5a	1500	dbase-LIII	LMtask	d11	11,1 %	27,4 %
TS5b	1500	dbase-LIII	LMtask	d11	11,1 %	28,3 %

Tabelle 4.15: Experimente-3 kontextabhängige Systeme

System	#Modelle	dbase	LM	Testset	OOV-Rate	WE
S12	3000	dbase-LIII	LMtask	d11	11,1 %	31,8 %

In Tab. 4.15 ist ein anderes Systemzweig. Ein kontextabhängiges System welches 3000 Modelle hat und zudem mit Vokaltraktlängennormierung (VTLN) arbeitet.

Kapitel 5

Morphembasierte Ansätze

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

5.1 Vorüberlegungen

Um die morphologischen Bestandteile im Türkischen zu erhalten, muß eine Möglichkeit der Zerlegung gefunden werden. Auf der Suche kamen wir auch mit anderen Forschungsarbeiten in Berührung. In der Türkei wurde Anfang 1994 eine 'Turkish Natural Language Processing Initative' gestartet, welche es sich zur Aufgabe gemacht, Vorarbeit für Forschung im Bereich Sprachverarbeitung und -erkennung zu leisten.[12] Im Rahmen dieser Initative enstand auch eine Tool zur Zerlegung türkischer Texte in ihre grammatikalischen, morphologischen Partikel.[13] Dieses Tool 'Xcorpus' ist jedoch interaktiv ausgelegt. In 1 von 10 Fällen ist statistisch gesehen, keine eindeutige Zerlegung möglich.[14] Das Tool baut auf PC-Kimmo, dem von Kimmo Koskenniemi entwickelten 'Two Level Processor for Morphologial Analysis' auf.
Für unsere Zwecke konnte 'Xcorpus' aus den oben genannte Problemen nicht eingesetzt werden. Es mußte eine andere Zerlegungsmöglichkeit gefunden werden. Hier kam uns eine andere Idee zugute, welche uns ermöglichte eine Trennung der Wörter in Silben zu erhalten.[1]. Durch das Textsatzprogramm Latex wollten wir den Textkorpus zerlegen lassen, indem wir Latex dazu zwingen, nach entsprechender Vorlagenformatierung, jedes Wort zu trennen. Wir bauten daher ein türkisches Latex, welches mit türkischen Trennungsregeln unseren Korpus zerlegen sollte.

[1] Dong Hoon Van Uytsel,Gastwissenschaftler an den Interactive System Labs, gab den Tip

Tabelle 5.1: Umsetzung romanisierte 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:

5.2 Silbentrennung

Zur Silbentrennung läßt sich sagen, daß sie im Türkischen nach Sprechsilben erfolgt z.B. as-tar, ıs-pa-nak, pro-gram.[6] [4] Die Zerlegungsvorgang beinhaltete mehrere Schritte, die in Diagramm 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 betrachten.
Im ersten Schritt mußte eine Umsetzung der romanisierten Texte in türkische Latexfonts erfolgen. Die Abbildung ist in Tab. 5.1 erläutert.

Bei der Ersetzung der Kleinbuchstaben ist darauf zu achten, daß zuerst alle i's ohne eine folgende 2 zu i: konvertiert werden. Eine Mißachtung dieser Abhängigkeit führt zu einem nicht korrekten Korpus. Weitere Sonderzeichen, die in der Vokabularliste auftraten, mußten für Latex gequoted[2] werden. Im nächsten Schritt wurden 2 Jobs 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

[2]d.h. mit Slash Symbol versehen

Abbildung 5.1: Silbenzerlegung

alle Wörter zu erzeugen. Die silbengetrennte Version jedes Wortes wurde aus dem Latex-logfile extrahiert. Aus nicht bekannten Gründen war jedoch eine Nachbearbeitung der Silben nötig. Es gab fehlerhafte Trennungen, in denen z.T. nur einzelne Buchstaben auftragen, welches aber den türkischen Trennungsregeln zufolge nicht auftreten dürfte [6]. So wurden entsprechend der uns bekannten Regeln, einzeln auftretende Buchstaben, wenn sie am Wortanfang standen, der 2. Silbe angehängt, bzw. am Wortende an die vorletzte Silbe gehängt. Einen Ausschnitt der Vokabularliste, nach der Zerlegung gibt Tab. 5.2.

Tabelle 5.2: Ausschnitt der getrennten Vokabularliste

Einträge der Vokabularliste
..
Ada og2 lu
Ada pa za ri2
Ada pa za ri2 na
Ada pa za ri2n da
Ada pa za ri2n daki
Ada pa za ri2n dan
Ada pa za ri2n da yi2m
Ada pa za ri2 ni2n
Ada pa zar li2
Ada pa zar li2 lar
Ada pa zar li2 la ri2
Ada pa zar si2z
Adap la ri2
Adar
Ada ra
Ada ri2n
Ada sar han li2
Ada si2
Ada si2 na
Ada si2n da
Ada si2n daki
Ada si2n dan
Ada si2 ni2
..

5.3 Systemspezifische Betrachtungen

Für das neue System sind einige Vorarbeiten zu erledigen. Von der Erstellung eines Silbenwörterbuches, dem Schreiben neuer Labels bis natürlich auch dem Bauen eines neuen Sprachmodells.
Die Problematik des Kontextverlustes, welches bisher nur eine Betrachtung von maximal Trigrammen ermöglichte, hat ein Erweiterung der LM-Implementierung in Janus erfordert. Diese neuen 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 Kap.5.2 vollzogene Silbentrennung liefert uns diese Einheiten.
Die ursprüngliche Idee [3] ist es, daß bei der Berechnung der Wahrscheinlichkeit für ein Wort w_t nicht nur die beiden vorherigen Worte betrachtet werden, sondern auch die Klasse des letzten Wortes. Die Formulierung lautet in 5.1

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

In diesem Fall wird ein Sprachmodell für jede Klasse gebildet. Die Aussagekraft des 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 [4] aufgegriffen und ausgebaut. 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. Das Bigramm $P(c_t|c_{t-1})$ ist viel stärker als das in der ursprünglichen Idee erwähnte Bigramm. Um die Klassenzuordung zu erhalten gibt es die Funktion:

[3] Implementierung durch Dong Hoon Van Uytsel
[4] Erweiterung durch Klaus Ries

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

Mit Wort ist im folgenden immer eine Silbe bzw. eine Kombination mehrerer Silben gemeint.
Folgender Backoff wird angewandt, bzw. der betrachtende Kontext eines Wortes w_t ist c_t, w_{t-1}, w_{t-2} und für c_t wird $c_{t-1}, w_{t-1}, w_{t-2}$ so wird zum einen versucht 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 auch schon in [9] vorgestellt, wobei die Idee auch hier war, den Backoff auf Unigramme möglichst zu vermeiden.

5.4 Vergleich verschiedener Zerlegungen

Die gefundene Zerlegung in Silben stellt ein Problem für unser System dar, da wir lediglich maximal ein 64k Dictionary verarbeiten können und auch somit eine Zusammenfassung von einzelnen morphemischen Einheiten erfolgen muß. Um verschiedene Zerlegungen vergleichen zu können ohne jedesmal ein komplettes System hochzuziehen wurden verschiedene Faktoren beachtet. Es gibt drei Faktoren, die hierzu beachtet werden sollten. Zum ersten die OOV-Rate, die auf Wortebene und Silbenebene betrachtet wird. Um so geringer diese ist, desto besser ist auch die Zerlegung, da es weniger unbekannte Einheiten gibt. Der zweite Faktor ist die Grösse des Dictionaries, welche < 64k ist, umso kleiner dieses Wörterbuch ist, desto besser ist die Zerlegung. Als letzter Faktor die Anzahl der Tokens im Testkorpus, hier gilt auch je weniger es sind, desto besser ist die Zerlegung.
Für die verschiedenen Zerlegungen führen wir eine Notation ein (siehe Tab 5.3)

Tabelle 5.3: Notation für Zerlegungsbeispiel

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

Zu lesen ist die Notation wie folgt: In unserem Beispiel fällt die erste Silbe in Klasse POS0, 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.

Wir haben nun 12 verschiedene Zerlegungen erstellt, welche in Tab. 5.4 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.
In Tab. 5.5 werden alle Zerlegungen mit Ihren wichtigsten Vergleichsgrößen angegeben.

5.5 Experimente morphembasierte Systeme

Für das Prototypsystem ergab ein erster kleiner Test folgendes Ergebnis Tab. 5.6:

Diese "Word Error" Rate bezieht sich natürlich auf Silben. Zum Vergleich muß diese noch umgerechnet werden. Eine Betrachtung des selben Systems auf Wortebene, in dem nur die "FirstBest" Strategie angewandt wurde, um von der Silbenebene zu Wörtern zu gelangen, kam mit folgendem Ergebnis Tab.5.7.

Tabelle 5.4: Beschreibung der 12 Zerlegungen im Überblick

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%
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%

Tabelle 5.5: Vergleich verschiedener Zerlegungen

Split	#vocabsize	#corpus(ohne oov-words)	#words test	OOV-Rate
SP1	8808	29671520	9363	7,4%
SP2	14881	20711909	6744	10,3%
SP3	21868	15664351	5131	13,5%
SP4	21982	15664351	5131	13,5%
SP5	22043	15664351	5131	13,5%
SP6	17033	18556235	5988	11,6%
SP7	9266	29671520	9363	7,4%
SP8	9470	29671520	9363	7,4%
SP9	14627	24623962	7750	8,9%
SP10	14831	24623962	7750	8,9%
SP11	15529	24021733	7550	9,2%
SP12	19630	22468288	6995	9,9%

Tabelle 5.6: Experimente Morphem-Prototypsystem Silbenebene

System	dbase	LM	Testset	OOV-Rate	WE
PT1	dbasePT1	LMPT1	mini11	0 %	51,5 %

Tabelle 5.7: Experimente Morphem-Prototypsystem Wortebene

System	Testset	OOV-Rate	WE
PT1	mini11	0 %	62,9 %

Kapitel 6

Zusammenfassung und Ausblick

Im Rahmen dieser Arbeit wurde ein türkisches Diktiersystem erstellt. Die anfänglichen Ergebnisse wurden schrittweise verbessert. Durch Verbesserung der Datenbasis, Korrektur des Wörterbuches, Verbesserung des Sprachmodells und letzlich damit ein türkisches Spracherkennungssystem für große Vokabulare entwickelt. Dieses ist nun in der Lage diktierte türkische Zeitungstexte sprecherunabhängig zu erkennen.

Als Ausblick läßt ist noch zu erwähnen, daß das System in einen Multilingualen Erkenner [16, 17] integriert wurde und sich dort gut bewährte.

Literaturverzeichnis

[1] M. Finke, I. Rogina, M. Woosczyna, 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] Elvan Göçmen, Onur Şehitoğlu, Cem Bozşahin: *An Outline of Turkish Syntax*, März 1995.

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

[5] *The Turkish Language*, Website

[6] 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.

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

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

[9] John Miller, Fil Alleva: *Evaluation of a Language Model using a Clustered Model Backoff*

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

[11] Kemal Oflazer, Elvan Göçmen, Cem Bozşahin: *An Outline of Turkish Morphology*, Oktober 1994.

[12] Kemal Oflazer, H.Cem Bozşahin: *Turkish Natural Language Processing Initiative: An Overview* 1994.

[13] Kemal Oflazer,İlker Kuruöz: *Tagging and Morphological Disambiguation of Turkish Text* 1994.

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

[15] Pons, *Taschenwörterbuch: türkisch-deutsch, deutsch-türkisch*, Klett Verlag, ISBN 3-12-518820-2, 1993.

[16] 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.

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

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

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

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

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

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

Anhang

Anhang A

Anhang

A.1 Datenbasis

Aufteilung in Training- und Testset

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 wie z.B. l_p (LM Penalty) und l_z (LM Weight) ein. Das 2. Testset ist das Evaluationset, um die Word Accuracy zu ermitteln..

Die 3 Sets müssen 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. Die übrigen Sprecher stellen das Trainingsset dar.

Im konkreten Fall waren es 22 "Unikat"-Texte und somit 22 Sprecher, die für die Testmengen in Betracht kamen. Jeweils die Hälfte kommt in eines der Testsets. Die restlichen 78 Sprecher bilden die Trainingsmenge.

A.2 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.[19]
Wie kann man die bedeutungsunterscheidenden Einheiten erhalten? Durch Minimalpaare, d.h. 2 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

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)

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 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 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. Hier müßte man für die englische Sprache PHONASCII von George D. Allen, für die deutsche Sprache SAMPA von K. Kohler und WORLDBET von James L. Hieronymous erwähnen.

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.

A.3 Systeme

Parameterbetrachtungen

Es gab einige Parameter, welche es einzustellen galt. Diese sind der 'LM-penalty' l_p, die 'LM-Gewichtung' l_z, der 'Beam', welcher den Zweig einer Hypothese abbricht, wenn er überschritten wird und 'topN', der Verzweigungsgrad der verschiedenen Wortfolgen bestimmt.

Zur Einstellung von l_z und l_p wurden verschiedene Kombinationen eingestellt, um die beste Einstellung zu finden.

Language Model - Klassenbasiert

Bei dem Sprachmodellansatz gibt es nun eine Änderung, daß 3 Sprachmodelle benötigt werden. Zum einen aus einem Silbenbasierten Sprachmodell und zum anderen aus einem Klassenbasierten Sprachmodell, wobei Klassen von Silben gemeint sind. Das dritte Language Model verzahnt diese beiden, es handelt sich um die interne Modellierung, nach außen bestehen jedoch nur 2 Sprachmodelldateien.