Die hocharabische Sprache und Romanisierung ihrer Schrift

Studienarbeit

von

Olfa Karboul

Institut für Logik, Komplexität und Deduktionssysteme
Fakultät für Informatik
Universität Karlsruhe (TH)

Betreuer:

Prof. Dr. Alexander Waibel
Dipl.-Inform. Tanja Schultz

Karlsruhe, den 14. Mai 1999

Zusammenfassung

Diese Arbeit beginnt im Kapitel 2 mit einer Einführung in die arabische Sprache. Nach einer Übersicht über die Geschichte und Entwicklung der Sprache, wird auf die Besonderheiten der Schrift eingegangen.
Im Kapitel 3 wird die Datensammlung, die in Tunesien durchgeführt wrude erläutert.
Im Kapitel 4 werden zuerst zwei Romanisierungsformen gezeigt. Danach wird die eigene Romanisierung.....

Inhaltsverzeichnis

Kapitel 1

Einleitung

Die vorliegende Arbeit *die hocharabische Sprache und Romanisierung ihrer Schrift* ist Teil des Globalphone Projekts, welches Forschung auf dem Gebiet der multilingualen Sprach-zu-Sprach Übersetzung, der multilingualen Spracherkennung und der automatischen Identifizierung zum Gegenstand hat.

Um all das zu ermöglichen, benötigt man eine multilinguale Datenbasis, die die Vielfältigkeit gesprochener Sprachen erfaßt, andererseits aber auch deren Verbreitungsgrad und wirtschafliche Bedeutung widerspiegelt.

Die Sprachen, die zur Zeit in Bearbeitung sind, sind Chinesisch, Deutsch, Englisch, Japanisch, Koreanisch, Kroatisch, Portugiesisch, Russisch, Schwedisch, Spanisch, Tamil, Türkisch und Arabisch.

Arabisch ist eine semitische Sprache, die von ca. 150 Millionen Sprechern in den arabischen Staaten, sowie von Minderheiten in anderen Ländern gesprochen wird.
In den arabischen Staaten ist die Sprachsituation durch Diglossie gekennzeichnet, d.h. der gemeinsamen arabischen Hochsprache stehen eine Vielzahl arabischer Dialekte gegenüber, die im Alltag verwendet werden, während die Hochsprache erst in der Schule gelernt und auf das öffentliche und religiöse Leben beschränkt ist.

Die arabische Schrift ist älter als die lateinische Schrift. Das arabische Alphabet hat 28 Buchstaben, wobei jedes Schriftzeichen in Abhängigkeit der

Stellung innerhalb eines Wortes vier verschiedene Formen ha und zwar am Anfang, in der Mitte, am Ende und alleinstehend.
Die einzelnen Schriftzeichen werden auf der vertikalen und horizontalen Ebene des Kreises angeordnet, wobei einige Buchstaben nur ein Viertel oder die Hälfte des Kreises einnehmen.
Die arabische Schrift wird wie das Hebräische von recht nach links geschrieben. Sie ist eine reine Konsonantenschrift - wie die phönizische. Die Vokalzeichen können für Leser arabischer Muttersprache fortbleiben; der Leser ergänzt sie aus dem Sinnzusammenhang, es gibt keine Groß- und Kleinschreibung und es gibt keine Druckschrift. Mit wenigen Ausnahmen sind alle Buchstaben miteinander verbunden.
In Tunesien wurden Sprachdaten für die arabische Hochsprache gesammelt. Es wurden ausschließlich gelesene Sprachdaten gesammelt. Hierzu wurden Texte aus überregionalen Tageszeitungen ausgesucht, die vom Sprecher vorgelesen wurden.

Unsere Texte waren auf dem Macintosh gespeichert. Die Macintosh Kodieurng ist nicht mit Unix-Kodierung kompatibel deshalb müssten die Texte romanisiert werden um sie nachher aufbereiten zu können.
Diese Arbeit beginnt im Kapitel 2 mit einer Einführung in die arabische Sprache. Nach einer Übersicht über die Geschichte und Entwicklung der Sprache, wird auf die Besonderheiten der Schrift eingegangen.
Im Kapitel 3 wird die Datensammlung, die in Tunesien durchgeführt wrude erläutert.
Im Kapitel 4 werden zuerst zwei Romanisierungsformen gezeigt. Danach wird die eigene Romanisierung dargestellt.

Kapitel 2

Die Sprache und Schrift des Hocharabisch

Um sich mit der arabischen Sprache zu beschäftigen, sollte man deren Geschichte und Entwicklung kennenlernen.
Obwohl in den arabischen Staaten eine Vielzahl arabischer Dialekte gesprochen wird und die Hochsprache erst in der Schule gelernt wird, ist Hocharabisch die Sprache des öffentlichen und religiösen Lebens.

2.1 Die hamitisch-semitische Sprachfamilie

Die hamitisch-semitische Familie wird neuerdings auch afro-asiatisch genannt. Das deutet auf ihr großes Verbreitungsgebiet hin, das vom Atlantik im Westen bis zum persischen Golf im Osten reicht, in Nordsüdrichtung vom Mittelmeer bis Somalia und Äthiopien.
In der zeitlichen Dimension überdeckt diese Familie einen gewaltigen Zeitraum, denn zu ihr gehören das Altägyptische, dessen älteste Denkmäler bis nahe an das Jahr 4000 v. Chr. zurückreichen; das Babylonisch-Assyrische, belegt seit dem 3. Jahrtausend v.Chr.; und das Althebräische, das als Sprache der meisten Bücher des alten Testaments lange Zeit als älteste Sprache der Menschheit gegolten hat, belegt ist es seit dem 9. Jahrhundert v.Chr.; schließlich das Aramäisch, die Sprache Jesus und seiner Jünger.
Diese Familie teilt sich in einen semitischen und einen hamitischen Zweig ein. [4]

2.1.1 Die hamitische Sprachfamilie

Die hamitische Sprache bildet ebenfalls zwei Zweige: Über den hamitischen Bestandteil gehen die Meinungen auseinander: Manchmal werden die beiden Gruppen, die zusammen diesen Zweig bilden, auch als selbständige Zweige betrachtet.
Diese beiden Gruppen sind:

- die Berbersprachen: Die Berber gelten als die Urbevölkerung Nordafrikas vor der Besiedlung durch die Araber.
  Trotz der heutigen Übermacht des Arabischen haben Berberstämme wie die Tuareg und die Rifkabylen in Marokko, die Kabylen Algeriens und die Bewohner der Insel Djerba (Südtunesien) an ihren angestammten Sprachen (oder Dialekten; die Berbersprachen sind einander sehr ähnlich) festhalten können. Ausgestorben ist die ebenfalls berberische Sprache der Guantschen, der Ureinwohner der Kanarischen Inseln;
- die kuschitischen Sprachen, zu denen u.a. das in Äthiopien verbreitete Galla (heute meist Oromo genannt) sowie Somali gehören.

2.1.2 Die semitische Sprachfamilie

Zu den semitischen Sprachen gehören als wichtigste Arabisch und Hebräisch, dazu erloschene Sprachen wie Phönizisch und Ugartisch.
Aramäisch ist die zweite Sprache des Alten Testaments und war in alttestamentlicher Zeit zugleich eine allgemeine Verkehrs- und Diplomatensprach des Alten Orients. Aramäisch wird heute noch von kleinen Gruppen in Iran, Irak und anderswo gesprochen.

Die Forschung ist sich hier auch nicht einig darüber, ob das Altägyptisch mit seiner *modernen* Fortsetzung, dem als Kirchensprache der koptischen Christen fortlebenden Koptisch, nicht besser als eigener Zweig innerhalb der Familie zu bezeichnen sei; Ähnliches gilt für das Amharische.

2.2 Die arabische Sprache

Die arabische Sprache ist die bedeutendste lebende semitische Sprache mit ca. 150 Mio Sprechern in den arabischen Staaten, sowie bei Minderheiten in

den Nachfolgestaaten der UdSSR, Afghanistan, Persien und der Südosttürkei. Mit der Verbreitung des Islams hat sich die arabische Sprache nach Mesopotamien, Syrien, Ägypten und weite Teile Nordafrikas ausgedehnt.
Aufgrund historisch-stilistischer, sozialer und typologischer Merkmale kann man drei Formen des Arabischen unterscheiden: klassisches Arabisch, modernes Hocharabisch und arabische Dialektsprachen.

- Das klassische Arabisch ist die Sprache des Korans. Es wird heute hauptsächlich als Sprache des Islams inner- und außerhalb der arabischen Welt verwendet.
  Vom 9.-16. Jh.n.Chr. war das klassische Arabisch internationale Verkehrssprache des Nahen und Mittleren Osten. Das klassische Arabisch hat viele Lehnwörter aus dem Aramäischen, dem Griechischen und weiteren Sprachen aufgenommen. Andere Sprachen wie z.B. Persisch, Türkisch, Urdu, afrikanische Sprachen (Swahili und Hausa) und Malayisch haben ihrerseits zahlreiche arabische Wörter übernommen.

- Das moderne Hocharabisch hat sich im vorigen Jahrhundert aus dem klassischen Arabisch entwickelt. Die Entwicklung des klassischen Arabisch zum modernen Hocharabisch mag als Ausdruck eines Wandels im politischen Bewußtsein und als Versuch angesehen werden, das klassische Arabisch der modernen Welt anzupassen.
  Das moderne Hocharabisch ist heute Symbol der kulturellen und politischen Zusammengehörigkeit aller arabischen Länder. Es wird heute in den arabischen Staaten als Standardsprache und als Schriftsprache in der Wissenschaft, Bildung, Literatur, im Theater und in den Massenmedien verwendet.

- In vielen arabischen Ländern wird Dialekt gesprochen. Diese Dialekte haben alle einen arabischen Ursprung. Es werden aber auch viele Fremdwörter benutzt, beispielsweise in Tunesine und Algerien: französische Wörter und in Marokko: französiche und spanische Wörter.
  Die sprachlichen Unterschiede zwischen den Dialektsprachen sind so groß, daß eine verbale Kommunikation zwischen vielen Dialektsprechern nur durch das moderne Hocharabisch möglich ist.

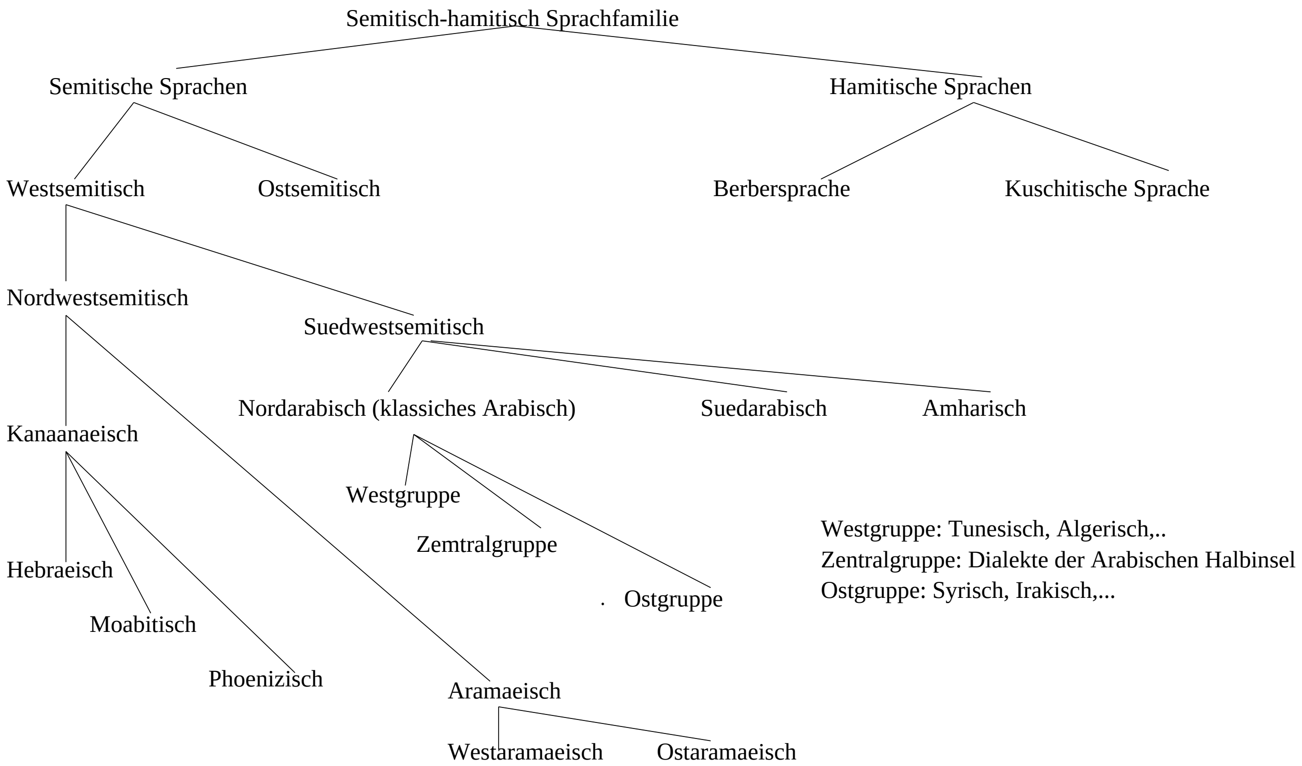

Abbildung 2.1: Die hamitisch-semitische Sprachfamilie

2.3 Arabisch heute

Die arabische Sprache verdankt ihre heutige Verbreitung den Eroberungszügen der Nachfolger des Propheten Mohammed, die sie bis nach Spanien trugen. Heute wird Arabisch ganz oder vorwiegend gesprochen in Ägypten, Algerien, Irak, Nord und Süd Jemen, Jordanien, Kuwait, dem Libanon, Libyen, Marokko, Maskat und Oman, Saudi-Arabien, Sansibar (Teil Tansanias), im Sudan, Syrien und Tunesien.

Das *Hocharabisch* ist in den genannten Gebieten Amtssprache und Sprache der Medien. Im täglichen Leben werden Dialekte gesprochen, bei denen man eine westliche (Marokko, Algerien, Tunesien), eine mittlere (Beduinen) und eine östliche Gruppe (Ägypten, Syrien, Irak) unterscheidet.

Die führende Rolle des Arabischen als Sprache der heiligen Schrift des Islams und als Sprache einer, den Nachbarn lange überlegenen Kultur, sowie der enge Kontakt mit Nachbarsprachen haben bewirkt, daß das Arabische Lehnwörter aufgenommen hat, vornehmlich aus dem Griechischen, auch aus dem Aramäischen; vor allem aber, daß arabische Wörter in zahlreiche Sprachen wie Türkisch, Urdu, Persisch, Malaiisch, Suaheli und Hausa eingedrungen sind.
Auch die europäischen Sprachen haben bereicherndes Sprachgut aus dem Arabischen aufgenommen, wie zum Beispiel im Spanischen.
Zwar stammt der überwiegende Teil des spanischen Grundwortschatzes aus dem Lateinischen, viele Wörter haben ihre lateinische Form sogar recht getreu bewahrt, aber arabische Lehnwörter - neben solchen keltischer und griechischer Herkunft - drangen in beträchtlicher Zahl ein, auch als Folge der zeitweise starken kulturellen Überlegenheit der Araber.
So sind nicht nur geographische Bezeichnungen wie Gibraltar oder Alcázar arabischer Herkunft (Dschebel-al-Tarik, *Felsen des Tarik*, des Anführers der arabischen Eroberer beim ersten Eindringen) , sondern auch wichtige Wörter der Umgangssprache, die oft mit der bezeichneten Sache selbst übernommen wurden.
So wurde aus dem amir-al-bahr, wörtlich *Emir des Meers*, im Spanischen über mehrere Zwischenstufen *almirante*, daraus über das Englische schließlich unser *Admiral.*
Arabische Wörter sind auf diesem Wege in die meisten übrigen europäischen

Sprachen, nicht nur die romanischen, eingebürgert worden, z.B. *Algebra*, *Alkohol*, *Magazin*, *Tarif*, *Ziffer* und *Zucker*.

2.4 Die arabische Schrift

Die arabische Schrift ist eine linksläufige Alphabetschrift, die sich zwischen dem 4. und 6. Jh. n.Chr. aus der nabatäischen Kursive des Aramäischen entwickelte. Die erste arabische Inschrift von Zebed (südöstlich von Aleppo) ist auf 512 n.Chr. datiert.

Da Arabisch mehr konsonantische Laute als das Aramäische besitzt und Laute z.T. zusammengefallen sind, waren diese in den ältesten Denkmälern der arabische Schrift nicht unterscheidbar. Unter Abdalmalik (608-705) wurde das heute noch gültige System der diaktrischen Punkte festgelegt, um diese Konsonanten in der arabischen Schrift zu unterscheiden. In der Folgezeit blieben die Buchstaben im wesentlichen unverändert.
Einige Zeichen besitzen am Wortanfang, -mitte oder -ende unterschiedliche Formen.
Die Kurzvokale a (Fatha), i (Kasra), u (Damma), die Vokallosigkeit (Sukun) und Verdopplung eines Konsonantes (Sadda) werden lediglich bei Bedarf und in wichtigen religiösen Texten durch Hilfszeichen über und unter der Schriftlinie dargestellt. Nur der Koran ist durchgängig vokalisiert.

Für die Interpunktion (Satzzeichensetzung) bestehen keine verbindlichen Regeln. Grundsätzlich werden Satzzeichen viel sparsamer verwendet als im Deutschen. Die Interpunktion überhaupt ist eine neuere Entlehnung aus Europa. Sie diente ursprünglich nur der Verseinteilung der Koransuren, erst im 20. Jh. wird im Buchdruck syntaktisch interpunktiert. Die hauptsächlich gebrauchten Satzzeichen sind: Punkt (.), Doppelpunkt (:), Beistrich (, oder ‘), Fragezeichen (?), Ausrufezeichen (!). Beistrich und Rufzeichen werden nur selten gesetzt, aber auch das Fragezeichen unterbleibt manchmal.

Mit der Ausbreitung des Islams wurde die arabische Schrift auch zur Schreibung anderer Sprache wie Persisch, Urdu, der Berbersprachen, Malaiisch, Hausa, Swahili und Türkisch (bis zur türkischen Schriftreform der 1920er Jahre) u.a. übernommen.

Die arabische Schrift hat aufgrund des Bilderverbots in der Islamischen Kunst große Bedeutung (Kalligraphie). Der Kufische Duktus geht auf die stark geometrisch gestaltete Monumentalformen der älteren Schrift zurück, die v.a. für Korancodices verwendet wurden. Wörter werden dabei ohne Rücksicht auf grammatische Formen getrennt, um den Buchstabenabstand gleichmäßig zu halten. Der Nashi-Duktus (runde Schriftzeichengestaltung) charakterisiert die normale Buchschrift. Zur Vereinfachung des Schriftsverkehrs im osmanischen Reich wurde als Kursive der Rup'a-Kuktus entwickelt.
Die Kalligraphen sahen in der kunstvollen Gestaltung und Ausschmückung der geschriebenen Worte Gottes eine Ehrerweisung und Würdigung ihres Schöpfers und unterstrichen damit den hohen Stellenwert der arabischen Schrift.

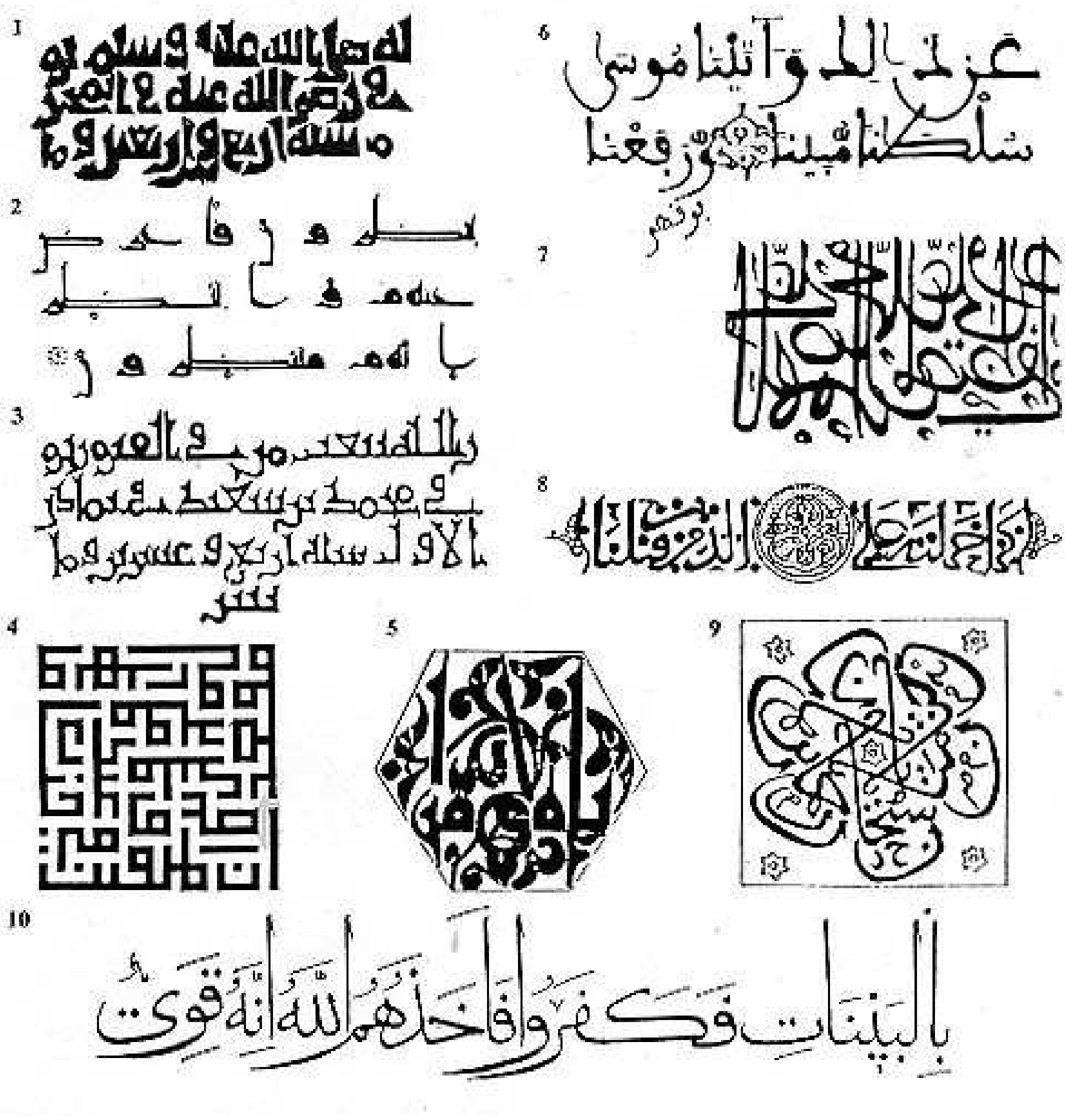

Abbildung 2.2: Arabische Kalligraphie

1 Drei Zeilen eines Grabsteins.Ägypten, datiert April 858 n. Chr. Kalkstein. Kufi-Duktus. [1].

2 Teil eines Koranfragments. Iraq oder Syrien, 8.-9 Jahrhundert. Pergament. Gestreckter Kufi-Duktus.

3 Drei Zeilen eines Grabmals. Ägypten, datiert Gumada März 838. Kalkstein. Blühender Kufi-Duktus.

4 Steintafel. Ägypten, 15. Jahrhundert. Marmor, Schiefer, Glasfritte. Quadrat-Kufi-Duktus.

5 Sechseckfliese. Iran, 12. Jahrhundert. Stuck. Blattwerk-Kufi-Duktus.

6 Drei Zeilen eines Koranfragments. Spanien, 12. Jahrhundert. Maghribi-Duktus.

7 Schriftband. Iran, 16 Jahrhundert. Seide. Tulut-Duktus.

8 Fragment eines Inschriftenfrieses. Ägypten, 15. Jahrhundert. Tannenholz. Naskhi-Duktus.

9 Fragment eines Nischenfrieses. Iran, 1. Hälfte 16. Jahrhundert. Fayencemosaik. Tulut-Duktus.

10 Koranfragment. Iran, Timuridische Zeit, 15. Jahrhundert. Tulut-Duktus.

[1] Die nach der Stadt Kufa in Mesopotamien bennante Kufi-Schrift hat eckig-geometrische Züge

2.5 Diakritische Hilfszeichen

- Die Verbindung der Buchstaben: Die arabische Buchstaben werden miteinander verbunden, soweit das möglich ist. Es sind darum bei jedem Buchstaben mehrere Schreibformen zu lernen, je nachdem ob der Buchstabe am Anfang, in der Mitte, am Ende des Wortes oder alleinstehend gebraucht wird. Es gibt sechs Buchstaben die können nur von rechts verbunden sein:
  A, D, Dd, R, Z, W:

  أ، د، ذ، ر، ز، و

- Vokalzeichen: Da die arabische Schrift eine Konsonantenschrift ist, werden die kurzen [a], [u], [i] und langen (auch Nunation gennant) [an], [un], [in] Vokale durch zusätzliche Zeichen angezeigt:

  - fatha über dem Konsonanten zeigt an, daß ihm ein [aa] folgt.
    zum Beispiel: B[aa]

    بَ

  - damma über dem Konsonanten zeigt an , daß ihm ein [uu] folgt.
    zum Beispiel: B[uu]

    بُ

  - kasra ünter dem Konsonanten zeigt an , daß ihm ein [ii] folgt.
    zum Beispiel: B[ii]

    بِ

  - doppel fatha über dem Konsonanten zeigt an, daß ihm ein [an] folgt.
    zum Beispiel: B[an]

    بًا

  - Doppel damma über dem Konsonanten zeigt an , daß ihm ein [un] folgt.
    zum Beispiel: B[un]

    بٌ

  - Doppel kasra unter dem Konsonanten zeigt an, daß ihm ein [in] folgt.

zum Beispiel: B[in]

بٍ

Im Wortauslaut dienen die kurzen und langen Vokale als Kasussuffix: Durch [u] oder [un] wird der Nominativ, durch [i] oder [in] der Genitiv und durch [a] oder [an] der Akkusativ angezeigt.
Häufig enden die indeterminierten Nomina mit einem Nun. Dieses Nun wird gesprochen, jedoch nicht geschrieben. Stattdessen wird der letzte Vokal verdoppelt; das wird [un], [in], [an] angesprochen. Dem [an] wird oft noch ein stummes Alif (arabisches A) hinzugefügt.

- Sukun: Wenn dem Konsonanten kein Vokal folgt, dann wird er durch Sukun gekennzeichnet. Sukun ist gleichbedeutend mit Null.

  لٰكِن

- Verdoppelung: Taschdid kennzeichnet die Verdoppelung des Konsonanten, über dem es steht. Betont ausgesprochene Konsonanten werden schriftlich nicht verdoppelt, sondern mit dem Verdopplungszeichen versehen.
  Beispiel

  عَدَّهَ

- Ta Marbuta: Weibliche Adjektive sowie viele weibliche Nomina werden mit Femininzeichen gekennzeichnet. Diese Zeichen ist ursprünglich eine Kombination der zwei Buchstaben T und h, das in der Verbindung als [tun] beziehungsweise [tu] ausgesprochen wird. Als Pausalform (vor Sprechpause) wird es h gesprochen, wie fatimah

  فَطِمَةٌ

  und in Kontextform [atun] ausgesprochen. Die Nationalgrammatik nennt dieses Zeichen ta marbuta.

- Hamza: Hamza ist ein stimmloser Glottalverschlußlaut. Das Hamzazeichen wird immer getrennt geschrieben, d.h. es wird weder mit dem vorausgehenden noch mit dem nachfolgenden Schriftzeichen noch mit seinem Träger verbunden.
  Es steht mit A, W oder Y, wenn diese den Lautwert des Kehlkopfverschlußlautes haben:

1 Am Anfang eines Wortes ist Alif (A) Träger von Hamza. Hamza wird über Alif geschrieben, wenn darauf ein [aa] oder [uu] folgt. Wenn auf Hamza ein [ii] folgt, wird es unter Alif geschrieben.

آ، أَ، أُ، إِ

2 In der Mitte des Wortes können alle drei Schriftzeichen A, W und Y unter folgenden Bedingungen Träger von Hamza sein:

. Alif wird als Träger von Hamza verwendet, wenn nach Hamza ein [aa] oder ein Konsonant folgt.

سأَلَ

Nach Alif wird Hamza auf die Schreiblinie gesetzt und hat keinen Träger.

سَاءَلَ

. W ist Träger von Hamza bei folgenden Lautverbindungen.

[u] + W

بُؤْسُن

[a] + W

رَؤُفَ

Nach W wird Hamza auf die Schreiblinie gesetzt und hat keinen Träger.

سُوءٌ

. Wenn Hamza in Umgebung von [ii] oder Y vorkommt, ist Y Träger von Hamza.

خَطِيئَةٌ

3 Am Ende des Wortes sind die Träger des Hamza abhängig von den vorausgehenden Vokalen:

. Ist der vorausgehenden Vokal ein [aa], so wird Hamza von Alif getragen.

قَرَأَ

. Wenn [ii] der vorausgehende Vokal ist, wird Hamza von Y getragen.

شَيٍءٍ

. Ist der vorausgehende Vokal ein [uu], so wird Hamza von W getragen.

لُؤلُؤ

. Nach Konsonanten und langen Vokalen wird Hamza allein geschrieben.

سُوءٌ

- Madda Anstelle von zwei AA (d.h langes A) schreibt man, um die Folge zweier Alif zu vermeiden:

  آ

  . Das Zeichen auf dem Alif heißt Madda (Dehnung). Beispiel:

  آكِلٌ

- Assimilation des Artikels Der Artikel Al wird mit dem Substantiv zusammengeschrieben. Das Al des Artikels assimiliert sich an die folgenden Konsonanten: T, D, Th, Dh, S, Z, N, L, R, Ss, Tt, Dt, Sd, Dd. Aufgrund der Assimilation des Artikels werden die arabische Schriftzeichen in zwei Gruppen mit zwei Kennwörtern eingeteilt: die Sonne ist Kennwort für die assimilierenden Schriftzeichen, der Mond ist Kennwort für die nicht assimilierenden Schriftzeichen.

2.6 Das Lautsystem des modernen Hocharabisch

Das Lautsystem des modernen Hocharabisch besitzt 28 Konsonanten, 3 Vokale und 2 Diphthonge:

2.6.1 Die arabischen Konsonanten

Buchstaben	Laut	Name	Lautwert
أ	A	alif	wie a, ä oder Stimmabsatz im Deutschen Beamter oder Theater
ب	B	be	wie b
ت	T	te	wie t
ث	Tt	th	stimmloses th, wie im Englischen there
ج	J	g	stimmhaftes g, wie im Englischen College
ح	H	h	kräftig gehauchtes h
خ	Kk	kh	immer Ach-Laut wie Nacht
د	D	dal	wie d
ذ	Dd	dhal	stimmhaftes the, im Englischen father
ر	R	re	wie r
ز	Z	zen	stimmhaftes s, wie Rose
س	S	sin	scharfer ß-Laut wie reißen
ش	Sc	schin	sch-Laut wie schön
ص	Ss	sad	emphatisches stimmloses s
ض	Td	ta	emphatisches stimmhaftes d

ط	Dt	dha	emphatisches stimmloses t
ع	Ar	re	gepreßter Kehllaut
غ	G	gen	entspricht fast dem Zäpfchen-r
ف	F	fe	wie f
ق	Q	qaf	emphatisches k
ك	K	kaf	wie k
ل	L	lam	wie l
م	M	mim	wie m
ن	N	nun	wie n
ه	h	he	wie h, wird auch im Silbenschluß gesprochen
و	W	waw	wie w im Englischen water
ي	Y	ye	wie j

Tabelle 2.1: Arabische Konsonanten

2.6.2 Die arabischen Vokale

	Name des Vokals	Buchstabe	Lautzeichen	mit dem Konsonanten
kurze Vokale	fatha	فَ	a	fa
	dhamma	فُ	u	fu
	kasra	فِ	i	fi
lange Vokale	fatha tawila	فًا	a	fa
	dhamma tawila	فٌ	u	fu
	kasra tawila	فٍ	i	fi

Tabelle 2.2: Die arabischen Vokale

2.7 Die Morphologie des Hocharabischen

2.7.1 Einführung

Das arabische Vokabular entsteht durch den Mechanismus 'al-ishtiqaq'. Al-ishtiqaq besteht aus der Ableitung eines oder mehrerer Wörter. Diese Ableitung soll die strukturelle und semantische Beziehung zwischen dem abgeleiteten Wort und dem Wort behalten.
Die meisten arabischen Wörter sind durch drei Konsonanten definiert, zu denen Affixe gebunden werden können.
Die trilaterale Wurzel ist die Basis der arabischen Sprache. Die Zahl der quadrilateralen Wurzeln im Vergleich zu den trilateralen ist sehr gering (zum Beispiel im Koran 15 gegen 1185).

2.7.2 Beispiele

h s b

حسب

ist ein trilaterale Stamm. Wenn er im Text nicht vokalisiert ist, kann er mehrere Bedeutungen besitzen:
hasiba = er dachte
hasaba = er zählt
hasba = entsprechend
hasab= adlige Abstammung

Die Wurzel s l m kann folgende Wortabstammung erzeugen:
salima = in Sicherheit
sallama= Handschütteln
aslama= zum Islam konvertieren

Die Konsonantenfolge k t b bedeutet schreiben, aber auch:
kataba = schreiben
aktab = schreib (Befehl)
kaatib = Schreiber
kitaab = Ein Buch
kutub = Bücher

kitaabah = Schriftsteller
kataba = er hat geschrieben
naktubu = wir schreiben

Alle Permutationen und einige mehr haben die gleiche Rechtschreibung im Arabischen.
Differenzierungsvermögen und gute Kenntnisse der Grammatik sind erforderlich, damit der Lesen den Sinn richtig erkennt.

Es gibt ungefähr 5000 arabische Wurzeln, die täglich benutzt werden, und ungefähr 400 phonologische unterschiedliche Muster, von denen die meisten mehrdeutig sind.
Jede Wurzel kann mit eine kleine Anzahl von verschiedene phonologische Pattern, diese Ableitung ist von eine alte lexicographische entschieden worden.
Der Stamm der meisten arabischen Wörter besteht aus drei Konsonanten (Trikonsonantismus), die, solange sie nicht mit Vokalen ausgefüllt werden, einen allgemeinen Begriff umschreiben.
Die Bildung solcher Formen erfolgt also durch eine Veränderung innerhalb der Wurzel, die Ablaut oder innere Flexion genannt wird. Neben diesen Bildungsmitteln werden häufig Suffixe verwandt.
Manchmal wird ein Vokal weggelassen, z.B. zur Steigerung des Adjektivs: kabir = groß; akbar = (der erste und der zweite Konsonant der Wurzel sind zusammengedrückt) größer; al akbaru = (mit dem bestimmten Artikel al davor) der Größte.
Beim arabischen Verbum werden viele Bedeutungsveränderungen, die im Deutschen hauptsächlich durch Vorsilben erzeugt werden (gehen - begehen - sich vergehen -entgehen), ebenfalls durch die eben geschilderte innere Flexion bewirkt.
Das Verbum kataba (in dieser Form also Vergangenheit der 3. Person: er schrieb) erhält z.B. durch Verlängerung des ersten Vokals eine reflexive Bedeutung: [kaataba] einander schreiben, miteinander korrespondieren; durch Verlängerung des zweiten Konsonanten entsteht [kat:aba] zum Schreiben veranlassen.
Die Formulierung beim einzelnen Wort erfolgt, wie hier angedeutet, durch Veränderung im und am Wort, also durch Flexion. Was dagegen das Verhältnis der Wörter im Satz anbehangt, so werden die Einzelwörter häufig - bei fest

geregelter Wortfolge - ohne weitere Kennzeichnung nebeneinander gestellt. In dieser Beziehung steht das Arabische den isolierenden Sprachen näher.
Die Funktion der Substantive ist beschrieben durch kurze vokalische Suffixe [u] Nominativ, [i] Genetiv und [a] Akkusativ.
Also [babun] die Tür (nom.), [babin] der Tür (gen.) und [baban] die Tür (akk.).

Kapitel 3

Datensammlung

Nachdem im vorausgegangenen Kapitel Aspekte der arabischen Sprache behandelt wurden, soll in diesem Kapitel beschrieben werden, wie benötigtes Datenmaterial für den Aufbau eines Spracherkenners, unter Berücksichtigung der Probleme, gesammelt und weiterverarbeitet wurde.

3.1 Das Globalphone Projekt

Auf der ganzen Welt gibt es etwa 6700 Sprachen, davon werden ca. 195 Sprachen von mehr als einer Million Menschen gesprochen. Angesichts einer zunehmenden Kooperation vieler Partner über Staatsgrenzen hinweg werden multilinguale sprachverarbeitende Systeme immer dringlicher.
Um Forschung auf dem Gebiet der multilingualen Sprach-zu-Sprach-Übersetzung, der multilingualen Spracherkennung und der automatischen Identifizierung zu ermöglichen, benötigt man eine multilinguale Datenbasis. Das Ziel des Projektes Globalphone ist die Erstellung einer solchen multilingualen Datenbasis, die die Vielfältigkeit gesprochener Sprachen erfaßt, andererseits aber auch deren Verbreitungsgrad und wirtschaftliche Bedeutung widerspiegelt[1].

[1]werner.ira.uka.de/ tanja

3.2 Erfahrungsbericht

3.2.1 Datensammlung

Es war vorgesehen, ausschließlich gelesene Sprachdaten von online verfügbaren Tageszeitungen des jeweiligen Landes zu sammeln.
Um die Domäne einzugrenzen aber gleichzeitig die Vergleichbarkeit über Staatsgrenzen hinweg zu gewährleisten, waren aus diesen Tageszeitungen jeweils tagespolitisch aktuelle Themen aus dem Ausland aufgegriffen.
Daneben werden Texte über innenpolitische und wirtschaftliche Themen ausgesucht. Die Texte waren von muttersprachlichen Sprechern vorgelesen. Jeder Sprecher sprach etwa 20 Minuten.
Ziel war es, pro Sprache 10.000 Äusserungen von insgesamt 100 Sprechern zur Verfügung stellen zu können. Die Datenbasis ist stetig am Wachsen, bisher wurden im Rahmen von GlobalPhone 9 Sprachen gesammelt: Arabisch, Chinesisch, Japanisch, Kroatisch, Koreanisch, Portugiesisch, Russisch, Spanisch und Türkisch.
Für die arabische Sprache bin ich zuständig.
Als Datenquelle für die Texte dachten wir an das Internet, aber ein Problem tauchte auf: Die arabischen Zeitungen haben ihre Texte immer als Bildformat gespeichert.
Deshalb wurde die Sammelaktion in Tunesien durchgeführt damit man die Texte in einer geeigneten Kodierung vorliegen, um sie danach bearbeiten zu können.
Ich habe Kontakt mit der Zeitung El Sabah aufgenommen und sie gebeten, 150 Texte für mich zu drucken, um gleich mit der Sprachdatensammlung anfangen zu können.

3.2.2 Umgebung

Nach dem Drucken der Texte, mußten 100 Personen gefunden werden, die bereit waren, ca. 20 Minuten lang Zeitungstexte vorzulesen.
Aufgrund der Erläuterung aus anderen Sammlungen nahm ich an daß Freunde und Verwandte ein guter Ausgangspunkt seien.
Und so ist auch die Sprachspende gelaufen.
Ungefähr 150 Texte wurden gedruckt, aber es wurde festgestellt, daß ein

Sprecher 8 bis 9 Texte braucht, um 20 Minuten zu lesen. Also hieß es, noch mehr Texte zu drucken.
Die Zeitung ist jetzt im Internet[2], aber das hilft uns noch nicht weiter, da die Texte als Bild gespeichert sind.
Die Sprachspende hat an verschiedenen Orten stattgefunden: Tunis, Sfax und auf der Insel Djerba.

3.2.3 Ausrüstung

Die Aufzeichnungen wuden mit einem portablen Sony Datenrecorder (Bandaufnahmegerät) TDC-8 und einem HD-440-6 Mikrophon gemacht. Die Daten wurden digital mit 48KHz in Stereoqualität aufgenommen.

3.2.4 Aufnahmeanweisungen

Die Aufnahmeanweisungen lauteten:

- Der Sprecher sollte bequem sitzen.

- Der Aufnahmeort sollte so gestaltet sein, daß sich der Sprecher gut konzentrieren kann und nicht gestört wird.

- Der Sprecher hält den Zeitungstext selbst in der Hand, dabei sollte er sie ruhig halten, um zuviel Papierrascheln zu vermeiden.

- Der Sprecher darf den Text vorher durchlesen (optional)

- Im Text sollte jeweils von einer Satzendmarkierung zur nächsten gelesen und dazwischen mindestens 2 Sekunden Pause gemacht werden.

- Bei Versprechen, Stottern, fehlerhaftem Vorlesen usw. den laufenden Satz abbrechen, 2 Sekunden Pause machen und dann den ganzen Satz nochmals vollständig vorlesen.

[2] www.tunisie.com/Assabah/

3.2.5 Probleme und Erlebnisse

Bei der Ein- und Ausreise habe ich keine Problem gehabt. Während der Sammlung ist es mir aufgefallen, wie hilfsbereit die Tunesier waren, sobald ich sagte, daß ich an der Uni bin, wollten sie mir unentgeltlich helfen.
Trotz der Hitze haben die Leute sich bereit erklärt, mir zu helfen und Sprache zu spenden. Sie haben gespendet und Schweiß stand auf ihrer Stirn .
Unsere Ausrüstung war bestens geeignet. Ich hatte kein Problem, damit zu arbeiten.
Die Sammelaktion hat insgesamt 3 Wochen gedauert, von denen 2 Wochen intensiv waren, was mehr als 8 Stunden Arbeit pro Tag bedeutete. Von dieser Sprachspende habe ich viel gelernt und dabei viele Erfahrungen gesammelt.

3.2.6 Beschreibung der Datenbasis

94 Sprachspender wurden gefunden, davon 37 Frauen und 53 Männer. Dies lag an der Tatsache, daß wir fast alle unsere Spender von der Straße erhielten, und es da sehr wenige Frauen gab.
In Tabelle 3.1 findet sich eine Zusammenfassung der wichtigsten Eigenschaften der Daten.

Sprecheranzahl	94
männiche Sprecher	53
weibliche Sprecher	37
Dialekt	Tunesisch
Origin: Djerba	50
Origin: Gabes	2
Origin: Sfax	9
Origin: Sousse	3
Origin: Tunis	29
Altersdurchschnitt Altersspanne	26 J(2532)
Raucher	22

Berufe	Anzahl
Beamte	12
Diplomat	1
Hausfrauen	3
Informatiker	1
Ingenieur	1
Journalisten	3
Koch	1
Lehrer	20
Schüler	27
Sekräterin	1
Studenten	21

Tabelle 3.1: Zahlen und Daten

Kapitel 4

Die Romanisierung

4.1 Einführung

Der Begriff *Romanisierung* bezeichnet den Vorgang, bei dem versucht wird, eine nicht in lateinischen Schriftzeichen dargestellte Sprache in diese Zeichen zu überführen.

4.2 Betrachtung der eigenen Daten

Wenn man in der Schule anfängt, arabisch zu lernen, werden die Texte in vokalisierter Form geschrieben, das heißt, man schreibt die Vokale über die Buchstaben. Sobald die Sprache beherrscht wird, schreibt man keine Vokale mehr. Deshalb sind die Zeitungen in nicht vokalisiertem Arabisch geschrieben.
Die arabischen Sprachdaten waren auf Macintosh unter dem Programm El Nasher El Sahafi gespeichert. Da die Bearbeitung der Daten unter Unix stattfinden wird, mußten wir die Daten konvertieren. Auf den Alphas haben wir dann keine arabischen Texte mehr, sondern Texte mit kodierten Zeichen.
Wir haben mit der Romanisierung angefangen, die Zeichen in romanisiertes Arabisch umzuwandeln. Die romanisierten Wörter werden ohne Vokale geschrieben. Man könnte die Vokale einfügen, aber eine trilaterale Wurzel besitzt ungefähr .... verschiedene Formen.

Man nimmt als Beispiel den trilateralen Stamm s l m:

s kann drei kurze Vokale besitzen.
l kann auch drei kurze Vokale haben, die Verdopplung und Sukun.
m kann lange und kurze Vokale haben und sukun.

4.3 Romanisierungsformen

4.3.1 Qalam

Qalam ist ein arabisch-lateinisch-arabisches Transliterationssystem zwischen dem arabischen Skript und dem lateinischen Skript, enthalten in dem ASCII Set.
Das Ziel des Qalam Systems ist die Romanisierung arabischer Skripte für die Rechnerkommunikation, um die Sprache richtig zu lesen und zu schreiben.
Qalam ist ein morphologisches System, weil die Romanisierung der arabischen Wörter auf der Spelling und Diacritis eher auf das Phonetic.......
Das macht es einfacher, das arabische Skriptwort von seiner Romanisierung abzuleiten. Die Aussprache der Wörter kann von der Romanisierung abgeleitet werden.
Jedes arabische Zeichen oder diakritische Maps wird in eine oder zwei ASCII-Character umgewandelt. Diese Wahl wurde gemacht zwecks Approximierung, um die arabische Aussprache so viel wie möglich die eine-to-eine morphologische Korrespandenz, die benötigt wird, um die Rücktranskription ins arabische Skript.
Die arabischen Zeichen, die nicht mit der lateinischen Aussprache korrespondieren, sind mit Großbuchstaben oder mit zwei Zeichensequenzen repräsentiert.
Qalam benutzt große ASCII-Zeichen, um arabische Zeichen anzudeuten, die von denen mit Kleincharakter abweichen.

4.3.2 Classical Arabic Transliteration (CAT)

Die Absicht von CAT ist die Repräsentation von arabischen Zeichen in ASCII Form, um computertextual Austausch zu ermöglichen.

In Tabelle 4.1 und 4.2 findet sich eine Zusammenfassung der wichtigsten

Unterscheidung der Romanisierung der Buchstaben.

Buchstaben	Qalam Romanisierung	CAT Romanisierung
hamza	’	’
alef	aa	aa
baa	b	b
taa	t	t
thaa	th	cˆ oder th
jym	j	j
Haa	H	h oder H
khaa	kh	k oder K oder kh
daal	d	d
dhaal	dh	zˆ oder zh [2]
raa	r	r
zayn	z	z
syn	s	s
shyn	sh	sˆ oder sh
Saad	S	s oder S
Daad	D	d oder D
Taa	T	t oder T
Zaa	Z	z oder Z
ayn	’	@
ghayn	gh	g ˆ oder gh
faa	f	f
qaaf	q	q
kaaf	k	k
laam	l	l
mym	m	m
nuwn	n	n
haa	h	h
waaw	w	w oder uu oder oo
yaa	y	y oder ii
taa marbuTah	t or h	nicht vorgegeben
haa marbuTah	h	nicht vorgegeben
alef maqSurah	ae	nicht vorgegeben
hamzat alwaSl	e	nicht vorgegeben

Abbildung 4.1: Romanisierung der Konsonanten

Buchstabe	Qalam Romanisierung	CAT Romanisierung
fatHah	a	a
kasrah	i	i
Dammah	u	u oder o
shaddah	Verdoppelung des Buchstaben	Verdoppelung des Buchstaben
maddah	aa	nicht vorgegeben
sukuwn	-	nicht vorgegebe
tanwyn	N	nicht vorgegebe

Abbildung 4.2: Romanisierung der Vokale

4.4 Romanisierung der Daten

Da die arabische Sprache mit normalem ASCII-7-bit-Code nicht dargestellt werden kann, muß diese auf geeignete Weise abgebildet werden.
Die Motivation bei der Auswahl einer geeigneten Romanisierung war, so nahe wie möglich an der Aussprache der Laute zu bleiben.
Um die Texte zu romanisieren, wurde ein Tcl-Skript erstellt.
Eine neue Romanisierungsform wurde erstellt, da die beiden anderen nicht geeignet waren.
Die Romanisierung muß eindeutig und umkehrbar sein. Die beiden Tabellen zeigen, wieso die Qalam- und CAT-Romanisierungsformen nicht geeignet sind:

Codierung	Buchstabe	Codierung	Buchstabe
307	A	352	Y
310	B	250	(
312	T	251	)
313	Tt	255	-
314	J	311	T
315	H	304	W
316	Kk	351	A
317	D	256	S
320	Dd	306	I
321	R	230	’
322	Z	214	’
323	S	240	
324	Sc	303	A
325	Ss	301	A
326	Sd	357	B
327	Td	302	A
330	Dt	272	:
331	Ar	340	-
332	G	245	%
341	F	@	Leerzeichen
342	Q	M	underline
343	K	254	,
344	L	260	.
345	M	300	*
346	N	257	/

Abbildung 4.3: Eigene Romanisierung

Buchstabe	Romanisierung Qalam	Romanisierung CAT	Eigene Romanisierung
aleef	aa	aa	A
baa	b	b	B
taa	t	t	T
thaa	th	th	Tt
jym	j	j	J
haa	H	h	H
khaa	kh	k	Kk
daal	d	d	D
dhaal	dh	zˆ oder zh	Dd
raa	r	r	R
zayn	z	z	Z
syn	s	s	S
shyn	sh	sˆ oder sh	Sc
saad	S	s oder S	Ss
daad	D	d oder D	Sd
taa	T	t oder T	Td
zaa	Z	z oder Z	Dt
ayn	'	@	Ar
ghayn	gh	g oder gh	G
faa	f	f	F
qaaf	q	q	Q
kaaf	k	k	K
laam	l	l	L
mym	m	m	M
nuwn	n	n	N
haa	h	h	h
waaw	w	w oder uu oder oo	W
yaa	y	y	Y

Abbildung 4.4: Vergleich

Kapitel 5

Ausblick

Für diese Arbeit wurde ein Teil, der für einen Spracherkenner nötig ist für die arabische Sprache erstellt.
Es ist aber nicht auszuschließen, daß die Romanisierung frei von Fehlern ist. Manche Sprecher lesen auch falsch von den Originaltexten ab oder fügen Wörter ein, die nicht geschrieben sind. In fast allen Aufnahmen kommen Häsitationen und Geräusche, wie Räuspern und Schmatzen vor.
Um diese Fehler auszuschließen, muß es von einem Muttersprachler noch überprüft werden.
Nach der Romanisierung und Erstellung der Datenbasis für JANUS sollte ein Aussprachewörterbuch und das Language-Model erstellt werden.

Literaturverzeichnis

[1] Tawfik Borg: *Modernes Hocharabisch Lehrbuch für Ausländer* in: S. 12-21, Verlag Borg GMBH, Hamburg 1986.

[2] Mohamed-Reza Majidi *Geschichte und Entwicklung der arabisch-persischen Schrift* in: Helmut Buske Verlag Hamburg 1986.

[3] *Language of the World*

[4] Hans Joachim Störig: *Abenteuer Sparache* in: Humboldt- Taschenbuchverlag Jacobi KG , München 1997.

Anhang A

Anhang

```
# ================================================================
#  Author  : Olfa Karboul Zouari
#  Module  : ara2rmn.tcl
#  Date    :  16.10.1997
#
#  Remarks : Romaniserung und Arabisierung der Daten
#
#=================================================================
if { ($argc != 1) || ([lindex $argv 1] == "-help")} {
    puts stderr "USAGE: $argv0 'Mode (roman,arab,diff)'"
    exit
}

#---------Parameter des Skriptes---------------------------------
set modus [lindex $argv 0]
puts $modus

set data_dir "/home/i13a2/gp_data/database/Arabisch/text-data"

#----------Romansierung---------------
puts $data_dir


proc ara2rmn {intext} {
```

```
set temp $intext

#----------------------
#nach arabisch geordnet
#----------------------
regsub -all \307 $temp "A"  temp
regsub -all \310 $temp "B"  temp
regsub -all \312 $temp "T"  temp
regsub -all \313 $temp "Tt" temp
regsub -all \314 $temp "J"  temp
regsub -all \315 $temp "H"  temp
regsub -all \316 $temp "Kk" temp
regsub -all \317 $temp "D"  temp
regsub -all \320 $temp "Dd" temp
regsub -all \321 $temp "R"  temp
regsub -all \322 $temp "Z"  temp
regsub -all \323 $temp "S"  temp
regsub -all \324 $temp "Sc" temp
regsub -all \325 $temp "Ss" temp
regsub -all \326 $temp "Sd" temp
regsub -all \327 $temp "Td" temp
regsub -all \330 $temp "Dt" temp
regsub -all \331 $temp "Ar" temp
regsub -all \332 $temp "G"  temp
regsub -all \341 $temp "F"  temp
regsub -all \342 $temp "Q"  temp
regsub -all \343 $temp "K"  temp
regsub -all \344 $temp "L"  temp
regsub -all \345 $temp "M"  temp
regsub -all \346 $temp "N"  temp
regsub -all \347 $temp "h"  temp
regsub -all \350 $temp "W"  temp
regsub -all \352 $temp "Y"  temp
regsub -all \250 $temp "\("  temp
regsub -all \251 $temp "\)" temp
regsub -all \255 $temp "\-"  temp
regsub -all \311 $temp "T"  temp
```

```
regsub -all \304 $temp "W"  temp
regsub -all \351 $temp "A"  temp
regsub -all \256 $temp "S"  temp
regsub -all \306 $temp "I"  temp
regsub -all \230 $temp "\'"  temp
regsub -all \214 $temp "\'"  temp
regsub -all \240 $temp " "  temp
regsub -all \303 $temp "A"  temp
regsub -all \301 $temp "A"  temp
regsub -all \357 $temp "B"  temp
regsub -all \302 $temp "A"  temp
regsub -all \272 $temp "\:"  temp
regsub -all \340 $temp "\-"  temp
regsub -all \245 $temp "\%"  temp
regsub -all \^@   $temp " "  temp
regsub -all {^M}   $temp "\n"  temp
regsub -all \254 $temp "\,"  temp
regsub -all \260 $temp "\."  temp
regsub -all \300 $temp "\*"  temp
regsub -all \257 $temp "\/"  temp
regsub -all \364 $temp "\n"  temp
#new underline ab t83
regsub -all \375 $temp "\n"  temp
#new underline ab t86
regsub -all \363 $temp "\n"  temp
regsub -all # $temp ".\n"  temp
#new underline ab t94
return $temp

}; #end ara2rmn

#-----------------------------
proc makeRoman {file} {
    global data_dir
    #-----
    set filein "$data_dir/$file"
    set fileout "$data_dir/$file.rmn"
```

```
    set FPin [open $filein r]
    set FPout [open $fileout w]

    while {![eof $FPin]} {
#Zeile einlesen
gets $FPin inline

#Romanisierung ausfuehren
set outline [ara2rmn "$inline"]
puts $FPout $outline
    }

    close $FPin
    close $FPout
}

#------------------Romanisierung=>Arabisierung------------------------
#Umkehrung der Romanisierung

proc ara2rmn {intext} {
set temp $intext

#----------------------
#nach arabisch geordnet
#----------------------
regsub -all "\A"   $temp "\307" temp
regsub -all "\B"   $temp "\310" temp
regsub -all "\T"   $temp "\312" temp
regsub -all "\Tt"  $temp "\313" temp
regsub -all "\J"   $temp "\314" temp
regsub -all "\H"   $temp "\315" temp
regsub -all "\Kk"  $temp "\316" temp
regsub -all "\D"   $temp "\317" temp
regsub -all "\Dd"  $temp "\320" temp
regsub -all "\R"   $temp "\321" temp
regsub -all "\Z"   $temp "\322" temp
```

```
regsub -all "\S"   $temp "\323" temp
regsub -all "\Sc"  $temp "\324" temp
regsub -all "\Ss"  $temp "\325" temp
regsub -all "\Sd"  $temp "\326" temp
regsub -all "\Td"  $temp "\327" temp
regsub -all "\Dt"  $temp "\330" temp
regsub -all "\Ar"  $temp "\331" temp
regsub -all "\G"   $temp "\332" temp
regsub -all "\F"   $temp "\341" temp
regsub -all "\Q"   $temp "\342" temp
regsub -all "\K"   $temp "\343" temp
regsub -all "\L"   $temp "\344" temp
regsub -all "\M"   $temp "\345" temp
regsub -all "\N"   $temp "\346" temp
regsub -all "\h"   $temp "\347" temp
regsub -all "\W"   $temp "\350" temp
regsub -all "\Y"   $temp "\352" temp
regsub -all "\(" $temp "\250" temp
regsub -all "\)" $temp "\251" temp
regsub -all "\-" $temp "\255" temp
regsub -all "T" $temp "\311" temp
regsub -all "W" $temp "\304" temp
regsub -all "A" $temp "\351" temp
regsub -all "S" $temp "\256" temp
regsub -all "I" $temp "\306" temp
regsub -all "\'" $temp "\230" temp
regsub -all "\'" $temp "\214" temp
regsub -all " " $temp "\240" temp
regsub -all "A" $temp "\303" temp
regsub -all "A" $temp "\301" temp
regsub -all "B" $temp "\357" temp
regsub -all "A" $temp "\302" temp
regsub -all "\:" $temp "\272" temp
regsub -all "\-" $temp "\340" temp
regsub -all "\%" $temp "\245" temp
regsub -all " " $temp "\"   temp
regsub -all "\n" $temp {^M}   temp
```

```
regsub -all "\," $temp "\254"  temp
regsub -all "\. $temp "\260"  temp
regsub -all "\*" $temp "\300"  temp
regsub -all "\/" $temp "\257"  temp
 regsub -all "\n" $temp "\364"  temp
#new underline ab t83
regsub -all "\n" $temp "\375"  temp
#new underline ab t86
regsub -all "\n" $temp "\363"  temp
regsub -all ".\n" $temp "\#"  temp
#new underline ab t94
return $temp

}; #end ara2rmn

#-----------------------------
proc makeArab {file} {
    global data_dir
    #-----
    set filein "$data_dir/$file.rmn"
    set fileout "$data_dir/$file.rmn"

    set FPin [open $filein r]
    set FPout [open $fileout w]

    while {![eof $FPin]} {
#Zeile einlesen
gets $FPin inline

if {[string index $inline 0] != ";"} {

#Arabisierung ausfuehren
set outline [rmn2ara "$inline"]
puts $FPout $outline
    }

    close $FPin
```

```
    close $FPout
}


#----------------------------------
#Romanisierung
proc roman {trl_list} {
    foreach elt $trl_list {
puts $elt
set elt_name [string trim $elt "t*"]
set elt_name $elt
puts $elt_name
makeRoman $elt_name
    }
}

#Arabisierung
proc arab {rmn_list} {
    foreach elt $rmn_list {
puts $elt
set elt_name [string trim $elt ".rmn*"]
puts $elt_name
makeArab $elt_name
    }
}


#---------------------------------------
#             MAIN
#--------------------------------------
#Liste der trl-Files, die heissen hier aber t{nr, daher auch lieber " "
    die Variable umbennen
set trl_list [glob t1]
puts $trl_list

puts "MODUS :$modus"
```

```
switch -glob -- $modus {
    {roman} {roman $trl_list}
    {arab} {set rmn_list [glob *.rmn];\
arab $rmn_list}

}
exit
```