Entwicklung einer digitalen Korpusanwendung zur türkeitürkischen Dialektologie

Masterarbeit

zur Erlangung des akademischen Grades

Master of Arts (M.A.)

im Fach „Empirische Sprachwissenschaft“

Goethe-Universität Frankfurt am Main

FB 09 Sprach- und Kulturwissenschaften

Institut für Empirische Sprachwissenschaft,

Phonetik und Slavische Philologie

eingereicht von: Manuel Raaf
geboren am: 11.05.1984 in Nördlingen
Betreuung: Prof. Dr. Uwe Bläsing

eingereicht am: 12.03.2012

Inhaltsverzeichnis

Abkürzungsverzeichnis

allgemeine Abkürzungen:

Abk.	Abkürzung	lat.	Latein
bzw.	beziehungsweise	Pl.	Plural
deutl.	deutlich	S.	Seite
dt.	Deutsch	trans.	Übersetzer
ebd.	ebenda	translit.	Transliteration
engl.	Englisch	trk.	Türkisch
entspr.	entspricht	u. a.	unter anderem, und andere
et al.	und andere (Autoren)	uvm.	und viele/s mehr
etc.	et cetera	v. a.	vor allem
ggbfs.	gegebenenfalls	vgl.	vergleiche
Hrsg.	Herausgeber	wörtl.	wörtlich
i. d. R.	in der Regel	z. B.	zum Beispiel

linguistische Coversymbole:

A	a, e (zweiförmige Vokalharmonie)
C	Konsonant
D	d, t
K	k, ğ
V	Vokal
X	ı, i, u, ü (vierförmige Vokalharmonie)

technische Abkürzungen:

CPU	Prozessor (central processing unit)
HDD	Festplatte (harddisk drive)
ID	Identifikationsnummer
MB	Megabyte
MB/s	Megabyte pro Sekunde
RAM	Arbeitsspeicher (random access memory)

Abstract

The present paper is the written and mainly theoretical part of the master thesis „developement of a digital corpus application to the dialectology of Turkish of Turkey“. The thesis' ambition is to create a useful linguistical tool and to spark linguists' interest in creating corpora and working with them.
First the work's aims and some important definitions are given in chapter one. The second chapter then deals with the linguistic relevance of corpus linguistics. Furthermore, the current state of the art with respect to Turkish corpora is given. The main part of this paper follows afterwards in chapter three. In this section the dictionary of dialects of Turkey, called „Derleme Sözlüğü“, as the basic of the corpus, is described in detail. Additionally, a complete user manual as well as two examples of the application's linguistic benefits are presented. Chapter four deals with the technical side of the thesis, however, it does not go into its depth, since for this purpose the commented source code is contained in the appendix. Finally, the conclusion and an outlook on future works of this kind are given in the fifth chapter.

Keywords: application, corpora, corpus, corpus application, corpus linguistics, Derleme Sözlüğü, developement, dialectology, dictionary, Turkey, Turkish

1 Einleitung

1.1 Überblick

Inhalt dieser Arbeit ist die Entwicklung einer digitalen Korpusanwendung zur türkeitürkischen Dialektologie, in welcher der Inhalt des in Kapitel 3.1 beschriebenen Wörterbuchs komplett enthalten und umfangreich durchsuchbar sein soll.
Da das Dialektwörterbuch insgesamt 126.468 Einträge enthält, ist eine manuelle Recherche darin äußerst zeitintensiv und unkomfortabel. Selbst die Suche innerhalb einer digitalen Version, die z. B. in einer Word-Datei gespeichert sein kann, ist nicht komfortabel möglich, sobald Optionen bzw. Einschränkungen getätigt werden sollen, wie z. B. die Beschränkung der Ergebnisse auf das Vorkommen in einer bestimmten Provinz.
Daher ist es das Ziel der Entwicklung, den Anwendern folgende Suchmöglichkeiten zu bieten: Das Wörterbuch soll entweder komplett im Volltext durchsucht werden können, oder nur innerhalb des Lemmas[1]. Dabei soll jeweils die Angabe einer oder mehrerer Provinzen möglich sein, sodass die Suchergebnisse nur jene Einträge beinhalten, in denen Suchbegriff und Provinz in der gleichen Zeile enthalten sind. Die Treffer der Suche sollen farbig hervorgehoben sein, damit insbesondere bei großen Wörterbucheinträgen nicht manuell nach den darin enthaltenen Suchbegriffen gesucht werden muss, sondern diese dem Anwender direkt auffallen. Außerdem ist es wünschenswert, mittels Platzhaltern und sogenannten Coversymbolen nach Wortformen und phonetisch bedingten Allomorphen suchen zu können. Letzteres ist für Linguisten, die sich mit dem Türkischen beschäftigen, bei der Eingabe der Suchbegriffe äußerst praktisch, da das Türkische eine vokalharmonische Sprache ist, in der phonetisch bedingte Varianten üblicherweise durch die Schreibung mit Coversymbolen dargestellt werden: z. B. *-lAr* für die Pluralallomorphe *-lar* und *-ler.*

All dies soll eine enorme Zeitersparnis bei Recherchen zur türkeitürkischen Dialektologie ermöglichen und darüber hinaus Linguisten, die bisher wenig mit Korpora arbeiteten, die Möglichkeit geben, einen unkomplizierten Einstieg in dieses Gebiet zu wagen. Ferner sollen sie erkennen, dass es für ein modernes Betreiben der Sprachwissenschaft unabdingbar ist, technische Hilfsmittel und deren stetig steigende Leistungsfähigkeit[2] zu nutzen. Außerdem

1 Ein Lemma ist ein „Eintrag in einem Lexikon“ (Metzler – S. 376), oder auch die „Grundform einer lexikalischen Einheit“ (Lemnitzer – S. 189)
2 Dem *Moor'schen Gesetzt* nach verdoppelt sich die Leistung eines Computers durchschnittlich alle zwei Jahre (Vgl. Wikimedia Foundation Inc. – http://en.wikipedia.org/wiki/Moore's_law)

soll durch die Mitgabe des kommentierten Quellcodes das Interesse am eigenständigen Entwickeln ähnlicher Anwendungen geweckt werden, da es auch für Linguisten grundsätzlich möglich ist, Programmiersprachen zu erlernen und effektiv einzusetzen.

Die Arbeit besteht aus zwei Teilen: einem technischen, der online unter http://www.manuelraaf.de/master zu finden ist und dem mit diesem Text vorliegenden schriftlichen. Ersterer kann auf Anfrage an master@manuelraaf.de eingesehen werden, da auf dem im Anhang vorhandenen Datenträger lediglich der Quellcode mitgeliefert wird, aus urheberrechtlichen Gründen jedoch nicht das Wörterbuch. Der Autor dieser Arbeit durchlief eine dreijährige betriebliche Berufsausbildung zum „*Fachinformatiker in Anwendungsentwicklung*“, besaß zum Beginn der Abschlussarbeit eine elfjährige Erfahrung als Softwareentwickler und wies daher die nötigen Kenntnisse auf, um auch den technischen Teil durchzuführen.

Im nun folgenden schriftlichen Teil wird zunächst auf einige wichtige themabezogenen Definitionen eingegangen, bevor im zweiten Kapitel die linguistische Relevanz der Korpuslinguistik erläutert und der aktuelle Stand der Technik bezüglich türkischer Korpora dargestellt wird.
Anschließend folgt mit dem dritten Kapitel der Kern der Arbeit. In ihm wird das digitale Dialektwörterbuch beschrieben, eine Bedienungsanleitung für die Korpusanwendung gegeben sowie durch zwei Beispiele deren linguistischer Nutzen demonstriert.
Kapitel 4 beschäftigt sich mit der technischen Seite der Masterarbeit, geht jedoch auf diese nur oberflächlich beschreibend ein, da ihr Kern der im Anhang enthaltene Quellcode ist. Dieser ist mit Kommentaren versehen und kann daher von Lesern mit Programmierkenntnissen leicht verstanden werden.
Schließlich folgt im fünften Kapitel das Fazit und ein Ausblick auf künftige Arbeiten dieser oder ähnlicher Art.

1.2 Definitionen

1.2.1 Korpus

Ein Korpus (lat. Körper; Pl. Korpora) besteht aus sprachlichen Daten, die „einer sprachwissenschaftlichen Analyse als Grundlage dienen“[3]. Diese spielen u. a. „eine zentrale Rolle … bei der Bearbeitung ausgewählter sprachlicher Phänomene“[4]. Ferner dienen sie „zur Ermittlung und Beschreibung sprachlicher Regularitäten bzw. zur Überprüfung von Hypothesen und Theorien und sind Grundlage der Korpusanalyse“[5].

Es[6] ist „eine Sammlung schriftlicher oder gesprochener Äußerungen in einer oder mehreren Sprachen. Die Daten des Korpus sind digitalisiert, d.h. auf Rechnern gespeichert und maschinenlesbar. Die Bestandteile des Korpus, die Texte oder Äußerungsfolgen, bestehen aus den Daten selber sowie möglicherweise aus Metadaten, die diese Daten beschreiben, und aus linguistischen Annotationen, die diesen Daten zugeordnet sind.“[7].

1.2.2 Korpuslinguistik

Korpuslinguistik ist ein Ansatz zur Untersuchung von Sprache, welcher durch den Einsatz von großen Textsammlungen und computergestützten Analyseverfahren charakterisiert ist[8].

Mit diesem Ansatz kann bezüglich der Art der Forschungsfrage eine große Vielfalt abgedeckt werden und es wird ermöglicht, spezifische, auf die jeweilige Frage angepasste Techniken anzuwenden[9].

„Als Korpuslinguistik bezeichnet man die Beschreibung von Äußerungen natürlicher Sprachen, ihrer Elemente und Strukturen, und die darauf aufbauende Theoriebildung auf der Grundlage von Analysen authentischer Texte, die in Korpora zusammengefasst sind. Korpuslinguistik ist eine wissenschaftliche Tätigkeit, d.h. sie muss wissenschaftlichen Prinzipien folgen und wissenschaftlichen Ansprüchen genügen. Korpusbasierte Sprachbeschreibung kann verschiedenen Zwecken dienen, zum Beispiel dem Sprachunterricht, der Sprachdokumentation, der Lexikographie oder der maschinellen

3 Metzler – S. 357
4 Ebd.
5 Ebd.
6 Das Genus des Lexems „Korpus“ innerhalb der Linguistik ist neutral.
7 Lemnitzer – S. 40
8 Vgl. Conrad – S. 385
9 Vgl. Conrad – S. 385

Sprachverarbeitung“[10].

1.2.3 Korpus- vs. Computerlinguistik

Unter Computerlinguistik ist die „Repräsentation und Verarbeitung natürlicher Sprachen mit den Methoden der Informatik“[11] zu verstehen.

Bei Carstensen et. al wird Computerlinguistik wie folgt beschrieben: „Gegenstand der Computerlinguistik ist die Verarbeitung natürlicher Sprache (als Abgrenzung zu z. B. Programmiersprachen) auf dem Computer, was sowohl geschriebene Sprache (Text) als auch gesprochene Sprache ... umfasst. Computerlinguistik ist im Kern und von ihrer Historie her ... eine Synthese informatischer und linguistischer Methoden und Kenntnisse“[12].

Ferner werden Ansichten dazu aufgelistet, worum es sich bei diesem Begriff innerhalb verschiedener Bereiche handelt[13]. Im Bezug auf die entwickelte Korpusanwendung ist nachfolgende Spezifizierung passend: „Computerlinguistik als Disziplin für die Entwicklung linguistik-relevanter Programme und die Verarbeitung linguistischer Daten ('Linguistische Datenverarbeitung'). Diese Auffassung hat ihre Wurzeln in den Anfängen der Informatik und hat insbesondere durch die zunehmende Wichtigkeit empirischer Untersuchungen anhand umfangreicher Sprachdatenkorpora ... eine Renaissance erfahren“[14].

Korpus- und Computerlinguistik sind also voneinander getrennt zu betrachten, da sie beide nicht Teil des jeweiligen anderen Bereichs sind. Zweitere ist ein Teilgebiet der künstlichen Intelligenz, die wiederum innerhalb der Informatik angesiedelt ist. Darüber hinaus stellt die Computerlinguistik eine Schnittstelle zwischen der Sprachwissenschaft und der Informatik dar. Die Korpuslinguistik hingegen ist ein Teilbereich der Sprachwissenschaft und bedient sich lediglich der Methoden der Informatik.

Ein Computerlinguist nutzt darüber hinaus Korpora i. d. R. nicht für wissenschaftliche linguistische Arbeiten, sondern um anhand der verfügbaren Sprachdaten Programme zu entwickeln, mit denen Linguisten arbeiten, oder bestehende Programme bzw. Korpora mit Daten zu füllen bzw. zu erweitern[15].

10 Lemnitzer – S. 10
11 Metzler – S. 119
12 Carstensen – S. 2
13 Vgl. Carstensen – S. 2
14 Carstensen – S. 2
15 Vgl. Lemnitzer – S. 137f

Korpus- und Computerlinguistik sind aufgrund der Überschneidungen der Einsatzgebiete und der in diesen angewandten Methoden als Nachbardisziplinen zu sehen[16].

1.2.4 Dialekt

Als Dialekt wird eine „besondere Sprech- (und z. T. auch Schreib-)weise innerhalb einer National- oder Standardsprache“[17] bezeichnet. Dabei erstreckt sich die Besonderheit „auf alle Sprachebenen (Lautebene, Phonologie, Morphologie, Lexik, Syntax, Idiomatik)“, hat allerdings „v. a. in Lautung und Wortschatz eine deutl. Ausprägung, die von anderen Sprachteilhabern der Standardsprache als abweichend bzw. von den Sprechern eines D. selbst so wahrgenommen wird“[18].
Chambers und Trudgill geben in ihrer Arbeit an, dass nicht immer klar ist, wo die Grenze zwischen einer „Sprache“ und einem „Dialekt“ zu ziehen ist[19]. Die „gegenseitige Verständlichkeit“ (engl.: mutual intelligibility) ist, wie sie anhand des Norwegischen, des Schwedischen und des Dänischen aufzeigen, nicht immer ein geeignetes Kriterium[20], da sich die Sprecher untereinander zwar verstehen, dies allerdings in unterschiedlichen Qualitäten[21]. Neben linguistischen Kriterien für die Klassifizierung als Sprache oder Dialekt gibt es Fälle, in denen Sprachen aus politischen, geografischen, historischen, soziologischen und/oder kulturellen Gründen als Dialekte deklariert werden und anders herum[22].
Die beiden Autoren sind darüber hinaus der Ansicht, dass auch die Standardsprache als Dialekt anzusehen sei[23] und somit die Bezeichnung „Varietät“[24] besser geeignet ist, um Unterschiede in der Sprech- bzw. Schreibweise innerhalb einer übergeordneten Einheit (~ Sprache) terminologisch festzulegen und die damit jeweilige sprachliche Entität zu benennen, die sich von allen anderen Entitäten durch phonetische, phonologische, lexikalische, syntaktische oder idiomatische Eigenschaften unterscheiden lässt.

16 Vgl. Carstensen – S. 3
17 Metzler – S. 139
18 Ebd.
19 Vgl. Chambers – S. 3f
20 Vgl. Chambers – S. 4
21 Ebd.: „It is often said … that Danes understand Norwegians better than Norwegians understand Danes“
22 Vgl. Chambers – S. 4
23 Vgl. Chambers – S. 3
24 Vgl. Chambers – S. 5

1.2.5 Dialektologie

Die Dialektologie ist eine Teildisziplin der Linguistik, in der primär durch Feldforschungsarbeiten mit Dialektsprechern in allen Bereichen der Linguistik Daten erhoben werden, um Dialekte synchron und diachron möglichst umfassend zu beschreiben und damit auch einen wichtigen Beitrag zur Sprachgeschichte der jeweiligen Hoch- bzw. Standardsprache zu leisten[25]. Darüber hinaus dienen in der Dialektologie gewonnene Kenntnisse oftmals Kulturwissenschaftlern im Bereich der Kulturraumforschung[26].

1.2.6 Türkeitürkisch

Das Türkische (Eigenbezeichnung *Türk dili*, *Türkçe* oder *Türkiye Türkçesi*; innerhalb der Sprachwissenschaft meist Türkeitürkisch (trk.: *Türkiye Türkçesi*) genannt, um es von anderen Turksprachen abzugrenzen; seltener Osmanisch-Türkisch)[27] ist „die Staatssprache der Republik Türkei sowie der Türkischen Republik Nordzyperns“[28] und neben Griechisch die der Republik Zypern[29]. Das Türkische hat mit derzeit rund 83 Millionen Muttersprachlern[30] und weiteren 5 Millionen Sprechern[31] die größte Sprecherzahl innerhalb der Sprachfamilie der Turksprachen. Es wird primär in den genannten Republiken sowie in Deutschland, Österreich, der Schweiz, Belgien, Aserbaidschan, Kasachstan, Kirgisistan, Usbekistan, Rumänien und in den Balkanstaaten gesprochen[32].

Um den Titel der Masterarbeit nicht zu verkomplizieren und zu verlängern, wurde in ihm bewusst auf eine nähere Spezifizierung dessen verzichtet, was unter „Türkeitürkisch“ im Rahmen dieser Arbeit zu verstehen ist, da diese nun erfolgt: Das Wörterbuch zur türkeitürkischen Dialektologie, welches die Grundlage der Korpusanwendung ist, behandelt mit wenigen Ausnahmen die Dialektwörter innerhalb der Grenzen der Republik Türkei. Türkeitürkische Sprecher aus anderen Ländern und deren Dialektwortschatz ist darin (mit wenigen Ausnahmen) nicht erfasst (siehe auch S. 12).

25 Vgl. Metzler – S. 140f
26 Vgl. ebd.
27 Vgl. Metzler – S. 700
Vgl. Wikimedia Foundation Inc. – http://de.wikipedia.org/wiki/Türkische_Sprache
28 Ersen-Rasch – S. 1
29 Vgl. Wikimedia Foundation Inc. – http://de.wikipedia.org/wiki/Republik_Zypern
30 Wikimedia Foundation Inc. – http://tr.wikipedia.org/wiki/Türkçe
31 Ebd.
32 Vgl. SIL International – http://www.ethnologue.com/show_language.asp?code=tur
Vgl. Wikimedia Foundation Inc. – http://de.wikipedia.org/wiki/Türkische_Sprache
Vgl. Wikimedia Foundation Inc. – http://tr.wikipedia.org/wiki/Türkçe

2 Korpuslinguistik

2.1 Linguistische Relevanz

Durch den Einsatz einer digitalen Korpusanwendung können tausende Seiten Text innerhalb weniger Sekunden, oder gar nur in Bruchteilen davon nach beliebigen linguistischen Mustern durchsucht werden.

Aufgrund der Leistung moderner Computer wird eine Textdatei, welche ca. 4000 gefüllte DIN-A4-Seiten enthält und eine Dateigröße von knapp 13 Megabyte (Abk.: MB) aufweist[33], in weniger als einer Sekunde eingelesen. Der Datendurchsatz einer Festplatte beträgt in privaten PCs oder Arbeitsplatzrechnern zwischen 25 und 100 Megabyte pro Sekunde (Abk.: MB/s), abhängig von der Qualität der Festplatte, der Anzahl der zu lesenden Dateien bzw. Dateifragmente[34] und der verwendeten Datenschnittstelle[35]. Auch wenn von der niedrigsten Rate von 25 MB/s ausgegangen wird, kann eine knapp 8000 Seiten lange Textdatei mit ca. 26 MB innerhalb einer Sekunde eingelesen werden.

Durch die Leistungsfähigkeit des Prozessors, mehrere Milliarden Rechenoperationen pro Sekunde ausführen zu können[36], ist es darüber hinaus möglich, diese schnell eingelesenen Daten ebenfalls sehr schnell zu verarbeiten, d. h. darin zu suchen, sie zu annotieren, Vorkommnisse bestimmter Muster zu zählen etc.

Diese beiden Punkte verdeutlichen am besten, dass durch digitale Korpora gestützte linguistische Forschungen um ein Vielfaches schneller sind, als die Bearbeitung der jeweiligen Fragestellung ohne technische Unterstützung.

Durch empirische Untersuchungen innerhalb eines Korpus können konkrete Aussagen zu einem bestimmten Thema getroffen sowie linguistische Theorien untermauert oder widerlegt werden[37]. Oftmals wird insbesondere von Generativisten davon ausgegangen, dass die Sprachkompetenz eines gesunden (d. h.: ohne mentale Beeinträchtigung) Muttersprachlers

33 Entspricht der digitalisierten Version des türkeitürkischen Dialektwörterbuchs (siehe S. 12)

34 Dateien werden auf dem Dateisystem oftmals an verschiedenen Stellen abgelegt und erscheinen dem Anwender nur als einzelne Datei. Beim Lesen dieser fragmentierten Dateien muss der Lesekopf der Festplatte für jedes Fragment neu positioniert werden. Dadurch wird unweigerlich Geschwindigkeit eingebüßt.

35 Selbst Festplatten, die dem IDE-Standard „Ultra DMA 4" des Jahres 1997 entsprechen und nur noch selten in Computern anzutreffen sind, erreichen durchschnittlich mehr als 25 MB/s beim Lesevorgang. Alle mit moderneren Schnittstellen konstruierten Festplatten (S-ATA ab I, USB ab 2.0, eSATA, Firewire ab 400) erreichen diese Rate und übertreffen sie oftmals um ein Vielfaches. (vgl. Wikimedia Foundation Inc. – http://de.wikipedia.org/wiki/ATA/ATAPI, http://de.wikipedia.org/wiki/Festplatte)

36 Vgl. Wikimedia Foundation Inc. – http://en.wikipedia.org/wiki/Instructions_per_second

37 Vgl. Carstensen – S. 490

ausreicht, eine Sprache gänzlich zu beschreiben und anhand dessen eine Grammatik zu erstellen[38]. Die Nutzung von Korpora wird von ihnen oftmals als „irrelevant und nutzlos“[39] beschrieben. In dieser Arbeit soll dieser Sichtweise nicht der Boden unter den Füßen entzogen werden, jedoch muss bedacht werden, dass in Abhängigkeit des Sprachgefühls eines jeden Muttersprachlers die innerhalb des Generativismus' mit solch hohem Stellenwert versehene Sprachkompetenz nicht immer ausreichend ist, eine Sprache in Gänze zu beschreiben, sodass alle anderen Sprecher derselben Sprache dieser Beschreibung widerspruchslos zustimmen[40]. Ferner ist es selbst einer Gruppe kompetenter Muttersprachler nicht möglich, alle morphologischen und syntaktischen Formen einer Sprache zu bilden[41]. Darüber hinaus sind diese Formen unter Umständen „künstlich“, d. h., dass sie bildbar und womöglich auch semantisch einwandfrei interpretierbar sind, jedoch nie benutzt werden (vgl. Chomskys grammatisch korrekten, jedoch semantisch nicht interpretierbaren Satz „*Colorless green ideas sleep furiously“*[42]). Schließlich gilt es außerdem zu bedenken, dass ausschließlich ein Korpus einen Anhaltspunkt darüber liefern kann, wie häufig ein Wort bzw. eine Phrase verwendet wird[43].

Empiriker stehen obigen generativistischen Aussagen gegenüber, da sie die tatsächliche Verwendung einer Sprache erforschen wollen und nicht nur wenig oder womöglich nie gebrauchte, aber bildbare Konstruktionen. Sie benötigen daher idealerweise eine große Menge echter Sprachdaten, welche üblicherweise durch Korpora zugänglich ist. Hierbei muss bedacht werden, dass aufgrund der unendlich möglichen Bildungen kein Korpus alle Formen einer Sprache erfassen kann[44]. Dies ist Linguisten, die Korpora für ihre Recherche nutzen, bewusst.

Durch die Nutzung einer Korpusanwendung als linguistisches Werkzeug ist es Sprachwissenschaftlern möglich, nach Morphemen, syntaktischen Konstruktionen oder Wortpaaren zu suchen. Dadurch können Unregelmäßigkeiten, Lehnwörter, seltene

38 Vgl. Lemnitzer – S. 1f
39 Lemnitzer – S. 7
40 Vgl. Lemnitzer – S. 21
41 Vgl. ebd.
Vgl. Mulzer – S. 64f, S. 91f
42 Vgl. Wikimedia Foundation Inc. – http://en.wikipedia.org/wiki/Colorless_green_ideas_sleep_furiously
43 Ein Muttersprachler kann diesen Anhaltspunkt bzw. eine klare Aussage allein aufgrund seiner sprachlichen Kompetenz nicht nennen.
44 Vgl. Lemnitzer – S. 21
Vgl. Mulzer – S. 64f, S. 91f

morphologische oder syntaktische Phänomene, Partikelreichweite etc. gefunden und untersucht, oder Aussagen zu deren Frequenz getroffen werden. Des Weiteren können Übersetzungs- und Spracherwerbssysteme erstellt und genutzt werden[45]. Durch eine Annotation der Daten ist es außerdem möglich, den Computer eine automatische morphologische und/oder syntaktische Segmentierung der Daten vornehmen zu lassen und diese somit nicht mehr zeitaufwändig manuell erstellen zu müssen.

Es ist theoretisch möglich, jegliche linguistische Fragestellung zu beantworten bzw. jegliche linguistische Untersuchung anzustellen[46]. Gerade im Bereich der Semantik oder Pragmatik, aber auch in der Morphologie und Syntax, gibt es trotz geeigneter (d. h., dass sowohl Korpus als auch Abfrageanwendung ausreichend Daten und Funktionen aufweisen, um den Forschungsansprüchen zu genügen) Korpusanwendung jedoch oftmals Schwierigkeiten, die Bedeutung bzw. syntaktische Relation eindeutig anzugeben. Die Semantik vieler türkischer Lexeme, z. B. die des Verbs „çekmek“ (dt.: herunterladen, reißen, rücken, ziehen, strecken, locken, mahlen), ist kontextbedingt und kann daher nicht nur anhand des Lexems selbst angegeben werden. Auch im Deutschen gibt es Probleme dieser Art, z. B. im Satz „Die alte Frau jagte den Dieb mit dem Regenschirm über den Marktplatz“: Ob die Phrase „mit dem Regenschirm“ sich adverbiell auf das Verb „jagte“ bezieht oder attributiv zum Objekt „Dieb“ steht, ist entweder kontextabhängig oder nicht klar zu sagen, sofern kein Kontext gegeben sein sollte.
Deshalb ist es lediglich theoretisch möglich, jegliche linguistische Fragestellung mittels einer geeigneten Korpusanwendung hundertprozentig klar zu beantworten. Sie klar zu bearbeiten hingegen ist (bei entsprechend vollständigen Daten und gut durchdachter Abfrageanwendung) deutlich schneller möglich, als alle Recherchen und Analysen manuell vorzunehmen. Ambiguitäten und andere Unklarheiten können anschließend manuell gelöst werden.

Im Bezug auf das hier verarbeitete Dialektwörterbuch des Türkeitürkischen hätte eine Suche ohne Verwendung der Korpusanwendung zur Folge, dass alle Seiten aufmerksam durchgelesen werden müssten. Bei Auffinden eines Treffers müsste dieser abgeschrieben

45 Für weitere Anwendungsmöglichkeiten vgl. Lemnitzer – S. 124ff und Carstensen – S. 553ff
46 Diese Aussage bezieht sich auf alle Teilbereiche der Linguistik: Mit jeweils geeignetem Korpussystem kann innerhalb der Phonetik, Phonologie, Morphologie, Syntax, Semantik, Pragmatik und Typologie geforscht werden (vgl. hierzu auch Carstensen – S. 169ff).
Siehe außerdem 1.2.2 auf S. 3.

werden. Auch wenn nur nach Lemmata gesucht wird und hierfür keine Volltextsuche erforderlich ist, sind wohl mehrere Tage bis wenige Wochen für die Durchführung der Suche erforderlich, abhängig von der Lesegeschwindigkeit und der täglich aufgewendeten Stundenanzahl. Sobald jedoch eine Volltextsuche nötigt ist, müssten hierzu alle zwölf Bände der gedruckten Variante bzw. alle 3826 Seiten der digitalisierten Version vollständig durchgelesen werden. Dafür wären deutlich mehr als „wenige Wochen“ nötig.
Eine solche manuelle Recherche würde daher nicht nur viel Zeit, sondern auch Geld kosten. Durch entsprechende personelle Ressourcen könnte das Ergebnis schneller geliefert und der zeitliche Faktor somit reduziert werden. Auf den Kostenfaktor hingegen wirkt sich dies nicht positiv aus. Darüber hinaus ist eine manuelle Recherche durch das anstrengende und individuell womöglich als monoton empfundene Lesen des Materials fehleranfälliger, d. h., dass wichtige Einträge überlesen werden könnten.

Korpora und Korpusanwendungen sind zusammenfassend aus folgenden Gründen ein nützliches Werkzeug für Linguisten: Die Suchergebnisse werden schnell und zuverlässig geliefert, sofern die Korpusanwendung gut durchdacht wurde. Dies ist selbst dann der Fall, wenn es sich bei den zu durchsuchenden Daten um mehrere tausend Seiten Text handelt[47]. Eine Korpusanwendung ermöglicht eine zeitsparende Recherche, trotz der anfänglich in die Erstellung zu investierende Zeit. Die Ergebnispräsentation erfolgt dabei automatisch und auch Diagramme oder PDF-Dateien sind innerhalb weniger Sekunden generiert. Außerdem kann das Korpus immer wieder verwendet sowie angepasst bzw. erweitert werden.

47 Je größer das Korpus, desto mehr Zeit wird für die Ausführung der Suche benötigt. Allerdings ist sie verglichen mit einer manuellen weiterhin um ein Vielfaches schneller.

2.2 Türkische Korpora

Für das Türkeitürkische existieren momentan wenige Korpora. Zwar gibt es diverse Projekte in der Türkei, allerdings wurde aus diesen bisher nichts Konkretes publiziert[48]. Innerhalb der gesamten Turkologie gibt es bis dato sehr wenige veröffentlichte Projekte im Bereich der Korpus- oder Computerlinguistik.

Nebst des Online-Bestandes der *Türk Dil Kurumu*, welcher ausschließlich Wörterbücher enthält, unter http://www.tdk.gov.tr sowie der Folgeseiten[49] zu finden ist und auch das dieser Arbeit zugrundeliegende Dialektwörterbuch der türkeitürkischen Sprache beinhaltet, ist leider kein weiterer Korpus zum Türkeitürkischen zugänglich.

Alle bei der *Türk Dil Kurumu* verfügbaren Wörterbücher sind jedoch nur beschränkt durchsuchbar. In ihnen kann nicht mittels Platzhalter oder Angaben zur Konsonanten- bzw. Vokalharmonie gesucht werden kann, um Wortformen oder Allomorphe zu finden. Außerdem ist die Beschränkung auf Suchbegriffe innerhalb einer oder mehrerer Provinzen nicht möglich. Des Weiteren scheint die Suche nur eingeschränkt und nicht in allen Wörterbucheinträgen möglich zu sein, wie sehr leicht anhand der Suche nach „anne“ (dt.: Mutter) festgestellt werden kann: Nach Eingabe des Suchbegriffs auf http://tdkterim.gov.tr/ttas/?kategori=derlay liefert die Webseite lediglich sieben Treffer, wohingegen die in dieser Arbeit entstandene Korpusanwendung 48 Treffer ausgibt. Dies lässt den Schluss zu, dass es sich bei dieser Onlineversion des Dialektwörterbuchs nicht um ein im Volltext durchsuchbares Exemplar handeln kann.

48 Zitiert nach: Prof. Dr. Uwe Bläsing, Arya International University, Erevan, Armenien; am 06.02.2012.
49 Ersichtlich, wenn in der Navigationsleiste mit dem Mauszeiger über „*SÖZLÜKLER*“ gefahren wird.

3 Digitales Dialektwörterbuch

3.1 Datengrundlage

Das Wörterbuch, welches die Datengrundlage der Korpusanwendung darstellt, beinhaltet Ausdrücke, die bis 1932 noch nicht in die türkeitürkische Standardsprache übernommen waren und in zwei großen Datenerhebungen sowie einer anschließenden Ergänzung zwischen 1932 und 1960 gesammelt und zusammengestellt wurden[50].

In der ersten Datenerhebung von 1932 bis 1934 schickten vier bis fünftausend Freiwillige aus allen Regionen der Türkei Karteikarten ein[51], auf welchen jeweils ein Dialektwort mit seiner Bedeutung bzw. Erklärung stand. Aus diesen ersten Einsendungen entstand ein sechsbändiges Dialektwörterbuch[52].

In der zweiten Zusammentragung der Daten, welche von 1952 bis 1959 durch die dafür von Ömer Asım Aksoy gegründete „*gönüllü derleyiciler örgütü*“[53] (dt.: freiwillige Datenerheber-Organisation) durchgeführt wurde, halfen 917 Personen, die namentlich aufgelistet sind[54], den Sammelbestand auf insgesamt ca. 600.000 dieser Karteikarten zu bringen[55].

Aus der Zusammenfassung der beiden Datenerhebungen und der anschließenden Ergänzung um Wörter, die nicht gesammelt, allerdings in einem anderen Wörterbuch namens „*Ana Dilden Derleme*“[56] (dt.: Zusammenstellung der Muttersprache; muttersprachliche Sammelzeitschrift) aufgeführt sind, entstand das seit dem existierende mehrbändige „*Türkiye'de Halk Ağzından Derleme Sözlüğü*“ (dt., wörtl.: Sammelwörterbuch des Volksmundes der Türkei; entspr.: türkeitürkisches Dialektwörterbuch), kurz „*Derleme Sözlüğü*“ (dt., wörtl.: zusammengestelltes Wörterbuch).

Das primäre Ziel dieser Datenerhebung- und veröffentlichung war es, die nicht in der türkeitürkischen Hochsprache vorhandenen, aber in diversen Mundarten auftretenden Lexeme zu dokumentieren und allen daran Interessierten eine umfassende Sammlung dazu anbieten zu können. Wie die Initiatoren einräumen, ist diese Sammlung nicht vollständig, da es nicht möglich war, die hierzu nötig gewesenen qualifizierten Linguisten im ganzen Land zu

50 Vgl. Türk Dil Kurumu Yayınları – S. V
51 Vgl. Türk Dil Kurumu Yayınları – S. VI
Vgl. Türk Dil Kurumu Yayınları – S. XIVf
52 Vgl. Türk Dil Kurumu Yayınları – S. V
53 Vgl. Türk Dil Kurumu Yayınları – S. VI
54 Vgl. Türk Dil Kurumu Yayınları – S. XXff
55 Vgl. Türk Dil Kurumu Yayınları – S. VI
56 Vgl. Türk Dil Kurumu Yayınları – S. V

beschäftigen[57]. Des Weiteren wird angegeben, dass durch die stetige Entwicklung der Sprache bzw. der Mundarten eine vollständige und fehlerfreie Dokumentation niemals möglich sei[58].

Die Digitalisierung des gesamten *Derleme Sözlüğü* geschah unabhängig dieser Masterarbeit im Jahre 2004 an der *Adana Üniversitesi*[59]. Dort wurden alle Bände eingescannt, mittels OCR-Software[60] in Text umgewandelt und anschließend Korrektur gelesen. Das Ergebnis ist eine Worddatei (.doc-Format) gespeichert worden, welche bei Seitenrändern von jeweils 2,5 cm, einfachem Zeilenabstand, der Schriftart Times New Roman und der Schriftgröße 10 insgesamt 3826 Seiten umfasst. Diese Datei zirkuliert seitdem innerhalb der Turkologie.

3.2 Korpusanwendung

3.2.1 Planung

Wie bereits in der Einleitung (siehe S. 1) geschildert, ist das Ziel dieser Masterarbeit die Erstellung einer digitalen Korpusanwendung zur türkeitürkischen Dialektologie. Aufgrund der dort genannten Anforderungen an die zu erstellende Anwendung wurde auf eine Annotation der Korpusdaten verzichtet, da diese für die Suche und anschließende Ergebnisdarstellung nicht von Nutzen wäre.
Somit wurde in der Planung im Hinblick auf die in der Einleitung genannten Ziele festgelegt, den Text des digitalisierten Wörterbuchs in ein passendes Format zu konvertieren, damit dieser anschließend in eine geeignete Datenbankstruktur (siehe S. 35 sowie Datenträger im Anhang) gespeichert werden kann und aus dieser heraus mittels eigens entwickelter Webabfrageanwendung (siehe S. 17, S. 38 sowie Datenträger im Anhang) durchsuchbar ist.

3.2.2 Datenverarbeitung

Da die unter 3.1 im letzten Abschnitt beschriebene Datei nicht direkt mit *PHP* (siehe S. 39) verarbeitet werden konnte, wurde sie im ersten Schritt der Arbeit durch die „*Speichern unter*“-Funktion des Textverarbeitungsprogramms *OpenOffice 3.3.0*[61] in das *HTML*-Format (siehe S. 38) konvertiert. Dadurch blieben die türkeitürkischen Grapheme ğ, ı, ç, Ç, ş, Ş und İ erhalten.

57 Vgl. Türk Dil Kurumu Yayınları – S. V
58 Vgl. Türk Dil Kurumu Yayınları – S. VI
59 Zitiert nach: Prof. Dr. Uwe Bläsing, Arya International University, Erevan, Armenien; am 06.02.2011
60 Texterkennung (vgl. Wikimedia Foundation Inc. – http://de.wikipedia.org/wiki/Texterkennung)
61 Vgl. Apache Software Foundation – http://www.openoffice.org/de/

In der resultierenden Datei waren darüber hinaus unzählige Formatierungsangaben vorhanden (Schriftart, -größe, Abstände), die für die weitere Verarbeitung unnötig waren und daher nach dem Einlesen der gesamten Datei und der anschließenden Umwandlung der *HTML*-Codierung der genannten Grapheme in Unicode entfernt wurden. Unicode ist ein Kodierungssystem für Schriftzeichen mit derzeit 110.116 enthaltenen Schriftzeichen, mit welchem diese plattform- und programmunabhängig dargestellt werden können[62]. Konvertierung und Entfernung der Formatierungen waren die ersten Schritte im für die Importierung der Textdaten in die zuvor erstellte Datenbankstruktur (siehe S. 35) programmierten *PHP*-Skripts *upload.php*. Anschließend wurde die Ersetzung eines jeden Buchstaben q durch x vorgenommen, da im gedruckten *Derleme Sözlüğü* an dieser Stelle jeweils das durch die OCR-Software als q gelesene ḥ geschrieben steht, welches den velaren Frikativ (Transkription gemäß des internationalen phonetischen Alphabets: [x]) darstellt und deshalb üblicherweise durch das Graphem x wiedergegeben wird. Aufgrund der einheitlichen Trennung der Lemmata wurde sodann jeder Eintrag des Wörterbuchs in die Datenbank geschrieben. Diese einheitliche Trennung sieht wie folgt aus: Jedes Lemma steht direkt am Zeilenanfang und jeder dazu gehörende Untereintrag (mit jeweiliger standardtürkischer Bedeutung und Ortsangaben) ist mit einem Zeilenumbruch und dem Tabulatorzeichen markiert. Nach der oben beschriebenen Entfernung der Formatierungsangaben enthielt der übrig gebliebene Text zwar zusätzliche Zeilenumbrüche, jedoch konnten diese alle gelöscht werden, da auch sie in einheitlicher Art und Weise auftraten. Dadurch war sowohl die Segmentierung eines Wörterbucheintrages selbst eindeutig und fehlerfrei möglich als auch die der jeweiligen Untereinträge.
Weitere Informationen hierzu sind in der Beschreibung der Datenbankstruktur ab Seite 35 gegeben. Weitere Informationen und Erklärungen der einzelnen Funktionen zur Datenverarbeitung finden sich in den Kommentaren innerhalb des Skripts *upload.php*.

Der gesamte Vorgang der Speicherung des Wörterbuchs in die Datenbank fand auf dem lokalen Arbeitsrechner statt, da der gemietete Webserver aufgrund seiner Speicherbeschränkungen (siehe S. 42) die Eingabedatei nicht verarbeiten konnte.

62 Vgl. Unicode Inc. – http://www.unicode.org/versions/Unicode6.1.0/#Character_Additions
Vgl. Unicode Inc. – http://www.unicode.org/standard/WhatIsUnicode.html

3.3 Struktur der Einträge

Um mit der Korpusanwendung arbeiten zu können, muss der Aufbau des Wörterbuchs und damit die Struktur eines jeden Eintrags bekannt sein. Bevor die Bedienung der Korpusanwendung im nächsten Unterkapitel geschildert wird, folgt daher zunächst eine kurze Beschreibung der Struktur der Einträge des *Derleme Sözlüğü*:

Jeder Wörterbucheintrag führt zuerst das Lemma auf, welches aus einem einzelnen Lexem bzw. einer Phrase besteht. Das Lemma stellt jeweils das Dialektwort bzw. den dialektalen Ausdruck dar. Sofern dazu alternative Formen existieren, sind diese in eckigen Klammern angegeben. Nach einem Zeilenumbruch folgt anschließend die Bedeutung im Standardtürkischen, welche von Ortsangaben gefolgt ist. Diese stehen in runden Klammern und beinhalten als letztes Element die entsprechend abgekürzte Provinz mit vorangestelltem Bindestrich. Durch diese Angabe ist die Provinz immer eindeutig zu erkennen. Nach jeder Provinz folgt ein Semikolon, sofern weitere Ortsangaben existieren. Vor ihr ist der durch ein Komma getrennte Landkreis angegeben, welchem ein * vorangestellt ist. Vor dem Landkreis sind Gemeinden und Dörfer ohne weitere Kennzeichnung angegeben und durch Kommata voneinander getrennt. Angaben, die sich auf die gleichnamige Hauptstadt der Provinz beziehen, stehen unmittelbar vor dieser.

Im weiteren Verlauf der Arbeit werden manche Ortsnamen als Dorf oder Gemeinde angegeben. Da zwischen ihnen eine Unterscheidung nur anhand der Kennzeichnung im Dialektwörterbuch nicht möglich ist, wurde hierfür jeweils eine Recherche im Internet angestellt, die durch die jeweilige Fußnote nachvollzogen werden kann.

Angesichts pragmatischer Gründe wurden die Provinzangaben von der *Türk Dil Kurumu* nicht überarbeitet, nachdem die Zahl der Provinzen durch eine Umstrukturierungen der Verwaltungsgliederung auf bis heute 81 stieg. Somit bestehen innerhalb des *Derleme Sözlüğü* nur die zur Zeit der Datenerhebung- und verarbeitung existierenden 67 Provinzen[63].

<u>Beispiel</u>: ** Karaman, Kurthasanlı * Kadınhanı, Sille -Kn.;*

Bei der Volltextsuche nach *anne* enthält das erste Suchergebnis u. a. diese Ortsangaben im zweiten Untereintrag. Das gefundene Dialektwort *aba* mit der Bedeutung *anne* (dt.: Mutter)

63 Vgl. Türk Dil Kurumu Yayınları – S. XI
Nach der Gründung der Republik existierten 63 Provinzen.
(Vgl. Wikimedia Foundation Inc. – http://tr.wikipedia.org/wiki/Tükiye#nin_illeri)

ist im Landkreis *Karaman*, im Dorf *Kurthasanlı*[64] des Landkreises *Kadınhanı* sowie im zur Stadt *Konya* gehörenden Dorf *Sille*[65] attestiert. Sie alle gehören zur Provinz *Konya*.

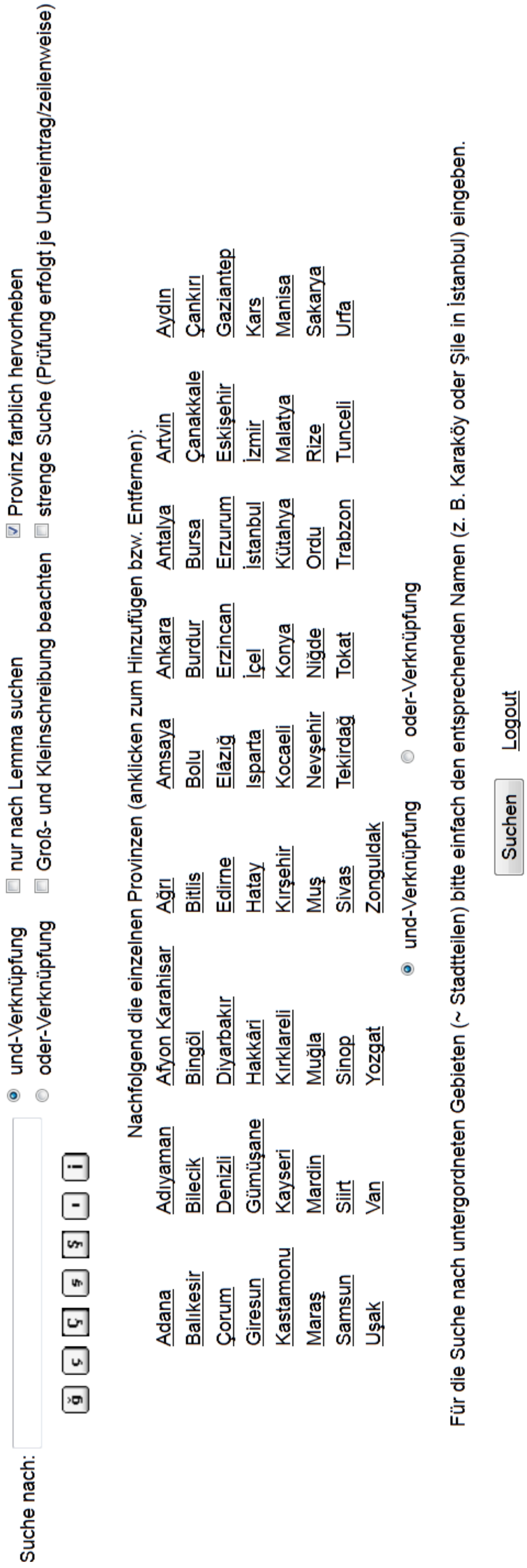
3.4 Bedienungsanleitung

Die Benutzeroberfläche (siehe S. 17) der Korpusanwendung beinhaltet neben einem Textfeld für die Eingabe des Suchbegriffs (bzw. der Suchbegriffe) diverse Möglichkeiten zur Spezifizierung der Suche. Darüber hinaus ist es durch die Verwendung sogenannter *Wildcards* (dt.: Platzhalter) möglich, nach beliebigen Zeichen zu suchen (siehe S. 21). Des Weiteren sind die innerhalb der Turkologie typischerweise verwendeten Coversymbole inkludiert, um damit Angaben zur Vokal- oder Konsonantenharmonie zu tätigen und somit nach Allomorphen zu suchen (siehe S. 22).

In der nachfolgenden Grafik ist der Hauptteil der Benutzeroberfläche zu sehen, in welchem Sucheingaben getätigt werden können. Der Rest wurde zur besseren Darstellung abgeschnitten und beinhaltet ohnehin nur die Informationen, die anschließend in den Punkten 3.4.2 bis 3.4.10 ab Seite 18 beschrieben sind.

64 Vgl. Wikimedia Foundation Inc. – http://tr.wikipedia.org/wiki/Kadınhanı
65 Vgl. Wikimedia Foundation Inc. – http://tr.wikipedia.org/wiki/Sille

3.4.1 Benutzeroberfläche der Korpusanwendung



Im Folgenden werden die einzelnen Bereiche der Benutzeroberfläche beschrieben. Sofern keine Suchoption angegeben ist, wurden die Standardeinstellungen verwendet: „*und-Verknüpfung*“, alles andere deaktiviert.

3.4.2 Das Suchfeld

Im Suchfeld stehen alle Suchbegriffe, nach denen gesucht werden soll. Sollte das verwendete Tastaturlayout die türkischen Buchstaben Ç, ç, ı, İ, Ş, ş, oder ğ nicht beinhalten, so können sie durch das Klicken auf die unter dem Suchfeld liegenden Grafiken eingefügt werden.

Die Trennung der Wörter erfolgt mittels Leerzeichen, welche durch die nachfolgende erste Suchoption für die Verknüpfung der einzelnen Suchbegriffe sorgen.

3.4.3 Verknüpfungsoptionen

Die durch Leerzeichen getrennten Wörter des Suchfeldes können eine „*und-Verknüpfung*“ oder eine „*oder-Verknüpfung*“ erwirken.

Bei der Wahl der ersten Option enthält das Suchergebnis nur Einträge, in denen alle Suchbegriffe enthalten sind.

<u>Beispiel</u>: *erkek kardeş büyük*
<u>Ergebnis</u>: Einträge die *erkek* <u>und</u> *kardeş* <u>und</u> *büyük* enthalten; Reihenfolge egal.

Durch die *oder-Verknüpfung* enthält das Resultat alle Einträge, die mindestens einen der Begriffe beinhalten.

<u>Beispiel</u>: *erkek kardeş büyük*
<u>Ergebnis</u>: Einträge die *erkek* <u>oder</u> *kardeş* <u>oder</u> *büyük* enthalten; Reihenfolge egal.

3.4.4 nur nach Lemma suchen

Falls nur innerhalb der Lemmata selbst gesucht und der Eintrag an sich nicht beachtet werden soll, so muss diese Option gewählt werden. Es wird in der gesamten Lemmazeile und daher auch innerhalb der dort stehenden Alternativen gesucht.

Beispiel: *baba* Suchoption: *nur nach Lemma suchen*

Ergebnis: Ergebnisse, die allesamt *baba* in der Lemmazeile enthalten und *baba* somit das Dialektwort bzw. ein Teil einer dialektalen Phrase ist.

3.4.5 Groß- und Kleinschreibung beachten

Durch den in der Datenbank verwendeten binären türkischen Zeichensatz „*latin5*“ (siehe S. 36) ist eine Unterscheidung zwischen Groß- und Kleinschreibung möglich.
Mit der Wahl dieser Suchoption wird die Groß- und Kleinschreibung beachtet. Die Option ist standardmäßig deaktiviert.

Beispiel: *İnsan* Suchoption: *Groß- und Kleinschreibung beachten*

Ergebnis: Einträge, in denen *İnsan* vorhanden ist und daher höchstwahrscheinlich am Satzanfang steht. Einträge, in denen lediglich *insan* existiert, werden nicht beachtet.

Bitte beachten: Bei versehentlicher Eingabe des Großbuchstaben *I* über das deutsche Tastaturlayout erfolgt keine Ausgabe der Ergebnisse zu *İnsan*, da *İ* und *I* in der türkischen Orthographie zwei verschiedene Grapheme sind, die jeweils eine eigene Lautung repräsentieren!

3.4.6 strenge Suche (Prüfung erfolgt je Untereintrag)

Durch Auswahl dieser Option enthält die Ergebnisliste nur Einträge, in denen die Suchbegriffe innerhalb einer Zeile vorhanden sind. Eine Zeile gibt entweder den Inhalt des Lemmas (samt Alternativen) und dessen ersten Untereintrag wieder, oder je einen zu einem Lemma gehörenden Untereintrag (siehe 3.3 auf S. 15).

Bei der Auswahl einer Provinz wird diese Option automatisch verwendet, da die Auflistung eines Eintrags, welcher das Suchwort im dritten Untereintrag enthält, die gewählte Provinz allerdings im siebten, nicht sinnvoll ist. Ungeachtet dessen werden im gesamten Wörterbucheintrag alle Vorkommnisse der gewählten Provinz farbig hervorgehoben, da Alternativen des Dialektwortes in der gleichen Region andernfalls womöglich übersehen werden.

Beispiel:	*abla*	Suchoption: Provinz *Kayseri* ausgewählt
Ergebnis:	Dialektwort *aba* mit *abla* als Bedeutung in der Provinz *Kayseri* innerhalb des ersten Untereintrages, sowie Hervorhebung der Provinz im Untereintrag der Alternative *abba* (dritter Untereintrag).	

Ohne Hervorhebung der Provinz innerhalb des gesamten Eintrages würde die Alternative (hier *abba*) womöglich übersehen werden.

Durch die Option „*nur nach Lemma suchen*“ (siehe 3.4.4 auf S. 19) wird die strenge Suche automatisch deaktiviert.

3.4.7 Provinzen

Wie im oberen Eintrag bereits kurz geschildert, enthält das Wörterbuch sehr viele lokale Angaben. Für die korrekte Lesart der Ortsangaben siehe Kapitel 3.3 (S. 15).
Aufgrund der Vielzahl von Gemeinden und Dörfern ist eine Auswahl aus Gründen der Übersichtlichkeit und Ambiguität[66] nicht sinnvoll. Um eine genauere örtliche Spezifizierung innerhalb einer Provinz vornehmen zu können, muss der Name der Gemeinde bzw. des Dorfs unter Verwendung der „*und-Verknüpfung*“ mit in das Suchfeld eingetragen werden.

Durch die Wahl mindestens einer Provinz wird automatisch die *„strenge Suche“* (siehe 3.4.6 auf S. 19) aktiviert. Die Verknüpfung mehrerer Provinzen funktioniert wie die Verknüpfung der Suchbegriffe, welche unter 3.4.3 auf S. 18 beschrieben ist.
Alle gefundenen Provinzen werden automatisch durch fette Schrift und blaue Schriftfarbe hervorgehoben.

66 Das Dorf bzw. die Gemeinde *Karaköy* existiert u. a. in Ankara, Istanbul und Bolu

3.4.8 Suche nach Phrasen

Soll nach mehreren Wörtern gesucht werden, die unmittelbar nacheinander auftreten (~ Phrase), so müssen sie in doppelte Anführungszeichen "" gesetzt werden.
Es können auch mehrere Phrasen als Suchbegriffe eingetragen und mit Einzelwörtern kombiniert werden.

Beispiel: *„erkek kardeş“ büyük*
Ergebnis: Einträge, in denen die exakte Wortfolge *erkek kardeş* vorkommt, sowie/oder[67] *büyük* an beliebiger Stelle; Reihenfolge egal.

3.4.9 Platzhalter und Coversymbole

3.4.9.1 Platzhalter

Es existieren zwei Platzhalter, die für Buchstaben im Allgemeinen verwendet werden:

* steht für beliebig viele Buchstaben (~ 0 bis unendlich)
Wenn * nachgestellt ist, endet die Suche an der Wortgrenze, welche durch die Satzzeichen, den Apostroph, den Bindestrich oder das Zeilenende eindeutig markiert ist. Wenn * vorangestellt ist, so beginnt die Suche an der Wortgrenze. Wird ein Bindestrich oder ein Apostroph von zwei Sternen eingeschlossen, gelten Bindestrich und Apostroph nicht als Wortgrenzen.

? exakt ein Buchstabe.

Beispiel: *insan**
Ergebnis: Einträge, die *insan*, *insanlar*, *insana*, *insan*, *insanların* etc. enthalten. Ohne Berücksichtigung der Groß- und Kleinschreibung enthält die Ergebnisliste auch alle Großschreibungen dieser Wortformen.

Beispiel: *kard??*
Ergebnis: Einträge, die die Buchstabenfolge *kard* sowie exakt zwei weitere beliebige Buchstaben enthalten (*kardeş, kardak, kardoş, kardeç* etc.). Auch hier enthält die Ergebnisliste ohne Berücksichtigung der Groß-

67 In Abhängigkeit der Verknüpfungsoption (siehe 3.4.3 auf S. 18)

und Kleinschreibung alle Großschreibungen dieser Wortformen.

Beispiel: *-*
Ergebnis: Alle mittels Bindestrich miteinander verbundenen Lexeme (z. B. *Yüz-ikiyüz*, *Otuz-kırk*, *Mehr-i*).

Beispiel: *yüz-*
Ergebnis: keine Ergebnisse, da kein Eintrag nach der Zeichenabfolge *yüz-* eine Wortgrenze beinhaltet.

3.4.9.2 Coversymbole

Die innerhalb der Turkologie gebräuchlichen Coversymbole sind in der Suchmaske wie folgt einzugeben (der Backslash \ darf dabei nicht vergessen werden!):

\X	steht für	i, ı, ü, u
\A	steht für	a, e
\I	steht für	i, ı
\D	steht für	d, t
\C	steht für	c, ç
\K	steht für	k, ğ

Beispiel: **\C\A*
Ergebnis: alle Einträge, in denen Lexeme mit dem Äquativsuffix (*-CA: -ca / -ce / -ça / -çe*) vorhanden sind.

Beispiel: *??*\D\X\K** Suchoption: *nur nach Lemma suchen*
Ergebnis: alle Einträge, in denen Lemmata mit dem Derivationssuffix *-DXk* (*-dık, -dik, -duk, -dük, -tık, -tik, -tuk, -tük*) stehen und anschließend beliebig vielen Zeichen folgen können. Dadurch werden Dialektwörter bzw. -ausdrücke gefunden, in denen *-DXk* alleine vorhanden ist, aber auch jene, in denen anschließend Possessiv- und Kasusmarker auftreten. Um Ergebnisse auszufiltern, in denen das Suchmuster den Wortanfang darstellt (z. B. *düğü*), wird mittels *??* dafür gesorgt, dass ein Basislexem, welches mindestens einsilbig ist, vor dem Suchmuster auftreten muss.

3.4.10 Sonstiges

Die Option „*Groß- und Kleinschreibung beachten*“ wird bezüglich der Coversymbole nicht beachtet. Bei allen Suchanfragen sollte bedacht werden, dass es unter Umständen mehrere Minuten benötigt, bis alle Ergebnisse geladen und dargestellt sind. Dies hängt von der Internetverbindung, der Auslastung des Servers, der Rechenleistung des Benutzercomputers und der Aktualität des auf diesem verwendeten Browsers ab. Bei großen Ergebnissen, wie sie bei z. B. der Suche nach dem Pluralallomorph *-lAr* durch **l\Ar* erwartet werden können, müssen einige Megabyte Daten übertragen und durch den Browser verarbeitet werden. Auf älteren Computern bzw. mit älteren Browsern kann die Darstellung mehrere Minuten beanspruchen.

3.5 Anwendungsbeispiele

Mit den nachfolgenden Beispielen soll die eingangs erwähnte linguistische Relevanz der Korpusanwendung gezeigt werden. Es werden nur diese beiden Suchen und deren Ergebnisse angegeben, da sie klar den Nutzen der Anwendung demonstrieren und aufgrund ihres Umfangs weder zu groß noch zu klein für eine Darstellung innerhalb dieser Arbeit sind.
Die Ortsangaben wurden jeweils eins zu eins aus der Korpusanwendung übernommen. Um diese Angaben besser verstehen zu können, bedarf es vorher noch folgender Erläuterungen:

• afşar köyleri, avşar köyleri:	dt.: Dörfer der Afşar/Avşar. Die Afşar sind ein türkmenischer Stamm, welcher insbesondere in Südanatolien anzutreffen ist[68].
• avşar aşireti:	dt.: Volksstamm der Afşar.
• ... ve köyleri:	... und seine Dörfer. Die Angaben zu den Dörfern sind bei diesen Einträgen leider nicht vollständig (siehe auch S. 12).
• … ve çevresi	… und Umgebung. Auch hier sind nähere Angaben nicht vorhanden.
• -Ky	Angabe der Provinz, hier Kayseri (siehe auch S. 15)
• Limasol, -Kıbrıs	Limasol, eine Hafenstadt[69] auf Zypern (trk.: Kıbrıs)

68 Steuerwald – S. 11
Vgl. Wikimedia Foundation Inc. – http://de.wikipedia.org/wiki/Afşar
69 Wikimedia Foundation Inc. – http://de.wikipedia.org/wiki/Limassol

3.5.1 denominales Nominalsuffix *-(X)msX*

Das denominale Nominalsuffix *-(X)msX* bildet Adjektive mit der Bedeutung „ähnlich wie“[70]. Um die dialektalen Wörter bzw. Ausdrücke zu finden, die dieses Suffix beinhalten, muss im Suchfeld **ms\X* eingetragen und die Suchoption „*nur nach Lemma suchen*“ aktiviert werden. Die Ergebnisliste enthält folgende (nach Lemmata gruppierte) Treffer:

1)	ağamsı	Beyazımsı (dt.: weißlich)	Sarıhamzalı * Sorgun -Yz.
2)	ağımsı	Beyazımsı (dt.: weißlich)	İncirgediği * Karaisalı -Ada.
3)	ağlamsı olmak, ağlayımsı olmak	Ağlıyacak gibi olmak (dt.: weinerlich sein)	Orhaniye * Marmaris -Mğ.; Türkmenaraplısı -Yz.; Dereçine * Sultandağı -Af.
4)	ağlamsı	Ağlıyacak hale gelmiş, ağlamaklı (dt.: in einem fast weinenden Zustand sein, weinerlich)	İğdir * Çivril -Dz.
5)	ağlamsı ağlamsı	Ağlamaklı ağlamaklı: Doğru söyle bir şey mi var? Niçin gözlerin ağlamsı ağlamsı? (dt.: die letzten Sekunden vor dem Weinen: Sag mir die Wahrheit, ist etwas los? Warum sind deine Augen so glasig (wörtl.: kurz vor dem Weinen)?)	* İnebolu -Ks.; * Elmalı -Ant.
6)	çamsı	çam gibi (dt.: glasig)	-Ank.
7)	gırımsı	Verimsiz, kolay sulanmayan toprak (dt.: karger/unfruchtbarer schwach bewässerter Boden)	-Çr.
8)	gırımsı	Hafif kar ya da küçük dolu. (dt.: schwacher/leichter Schnee oder kleiner Hagel)	-Ba.
9)	gomsu	Sevimli (dt.: lieb / nett / sympathisch)	Eldirek, Seki, * Fethiye -Mğ.
		Küçük yapılı (dt.: klein gebaut, Kleinwüchsige/r)	Eldirek, Seki, * Fethiye -Mğ.

70 Ersen-Rasch – S. 264

10)	gomsu	Bilgiçlik taslayan (dt.: Besserwisser)	Eldirek * Fethiye -Mğ.
		Birinden ötekine söz taşıyan (kimse). (dt.: jemand, der Gehörtes weitererzählt / Tratschtante)	* Tirebolu -Gr.
11)	hamsi	Hepsi (dt.: alle, Ganzheit)	-Ama.; Karakoyunlu -Kr.
12)	kekirimsi	Buruksu, kekremsi (dt.: säuerlich, herb)	Genek * Yatağan -Mğ.
13)	komsu	İki yüzlü, dalkavuk, boşboğaz, söz getirip götüren (dt.: häuchlerisch, unaufrichtig, geschwätzig, Tratschtante)	* Perşembe -Or.; -Tr.; Andifli * Kaş, Çandır * Serik -Ant.; Kınık * Fethiye -Mğ.
14)	komsi	İki yüzlü, dalkavuk, boşboğaz, söz getirip götüren (dt.: heuchlerisch, unaufrichtig, geschwätzig, Tratschtante)	Dont * Fethiye -Mğ.
15)	komsu	Süslü, şık: Şu kız çok komsudur. (dt.: schick / elegant: Dieses Mädchen ist sehr schick.)	Bozburun * Marmaris -Mğ.
16)	omsu	Sevimli: Omsu çocuk kendini sevdirir. (dt.: lieb / nett / sympathisch: ein liebes Kind sorgt selbst dafür, dass man es liebt; wörtl.: ein liebes Kind lässt sich lieben.)	Hadım * Çal -Dz.; * Milas -Mğ.
17)	samsı	Parça, bölüm (dt.: Stück / Teil / Abschnitt)	* Biga -Çkl.; -Kü.
18)	samsı	İçine kıyma, patates, haşlanmış yumurta konularak dört köşe katlanan bir çeşit börek. (dt.: viereckiges Börek, mit Hackfleisch, Kartoffeln und gebratenem Ei gefüllt.)	Ağlı * Küre -Ks.
		Hamur tatlısı (dt.: Knödel / Kloß)	Limasol, -Kıbrıs
19)	sıxlamsu	Çok ıslanmış, sudan çıkmış (dt.: sehr nass)	* Ardanuç ve köyleri -Ar.

20)	silimsi	Yemek seçen, boğazsız (dt.: wählerisch im Bezug auf das Essen)	Çığrı * Dinar -Af.; Bereketli * Tavas, Oğuz * Acıpayam -Dz.; * Bozdoğan -Ay.; * Bodrum -Mğ.
21)	silimsi	Turşu suyu, ekmek ve sarmısaktan yapılan yemek. (dt.: eine aus einer Art Gemüsewasser, Brot und Knoblauch gefertigte Mahlzeit)	Süksün * Bünyan -Ky.
22)	silimsi	Temiz, titiz (kimse) (dt.: sauber / rein, exakt /strikt (auf eine Person bezogen))	Çığrı * Dinar -Af.
23)	söbemsi	Yumurta biçiminde, oval (dt.: eiförmig, oval)	* Düzce -Bo.; -Ks.
		Koni biçiminde (dt.: kegelförmig)	Oğuz * Acıpayam -Dz.; Peşman * Daday -Ks.
24)	sümsü	Tütün çubuğu ucundaki, tütünün konulduğu yer. (dt.: Aschenbecher; wörtl.: der Ort, an den man die Spitze des Tabakstengels platziert.)	-Ank.
25)	şemsi	İçki dağıtan (dt.: Spirituosenhändler)	Tahtacılar * Silifke, * Mersin -İç.
26)	şemsi	Şemsiye (dt.: Regenschirm)	Bereketli * Tavas -Dz.
27)	şıpırdımsu	Derinliği bir karışı geçmeyen gölcük. (dt.: Teich, der nicht tiefer als eine Handspanne ist.)	* Milas -Mğ.
28)	tamsı	Damla, sızıntı: Bu yıl gökten tamsı düşmedi. (dt.: Tropfen / Träne, Leck: Dieses Jahr fiel kein Tropfen vom Himmel.)	Erkinis * Yusufeli -Ar.
29)	tomsu	Küçük tepe (dt.: kleiner Hügel)	* Eğridir ve köyleri -Isp.
30)	tomsu	Yükseklik (dt.: Erhebung / (An)Höhe)	Tuzhisar * Bünyan -Ky.

31)	tomsu	Kısa ve kalın (daha çok salatalık için). (dt.: kurz und dick (eher im Bezug auf Gurken.))	-Nş.
32)	tömsü	Sevgi, yarenlik. (dt.: Liebe, Freundschaft)	-Ky.
33)	tümsü	Tepe, tümsek. (dt.: Hügel, Erdwall / (An)Höhung)	* Develi -Ky.; * Silifke -İç.
34)	ürümsü	İnce çalı çırpı. (dt.: dünner Reisig.)	Tepeköy * Torbalı -İz.
35)	ağlamsı	Ağlamaklı, ağlayacak gibi (dt.: weinerlich, dem Weinen nahe / zum Heulen)	Dereçine * Sultandağı -Af.
36)	apağlamsı	Ağlayacak gibi (dt.: weinerlich / dem Weinen nahe / zum Heulen)	* Fethiye ve çevresi -Mğ.
37)	samsı	Dilim, parça (baklava, börek için). (dt.: Scheibe / Stück / Abschnitt (im Bezug auf Baklava und Börek).)	-Ks. ve çevresi
38)	samsı, samsı	Yağda kızartıldıktan sonra şuruba batırılan, içine ceviz konulup muska biçiminde sarılmış yufkalardan yapılan bir çeşit tatlı. (dt.: eine Art Dessert, das mit Walnüssen gefüllt ist und amulettförmig in Blätterteig gehüllt angebraten wird, nachdem der Blätterteig in Öl getaucht und mit Sirup gebacken wurde.)	-Çr.
39)	sasımsı	Tatlımsı, acımsı bir tat. (dt.: süßlich, bitterer Geschmack)	-Çr.
40)	samsı	Eli sıkı. (dt.: geizig)	Boyalık * Ermenek -Kn.
41)	sasımsı	Tatlımsı, acımsı bir tat. (dt.: süßlicher, leicht bitterer Geschmack)	-Çr.
42)	tomsu	Düz alanda yüksek yer, tümsek. (dt.: erhöhte Stelle einer Ebene, Erdwall / (An)Höhung)	* Sütçüler -Isp.; Çepni * Gemerek -Sv.

Die Liste enthält alle Einträge, die standardtürkische Angaben enthalten (reine Verweise auf Wörterbucheinträge wurden ausgefiltert). Um die Vorkommnisse des Suffixes *-(X)msX* in Dialekten der Türkei aussagekräftig beschreiben zu können, bedarf es nun einer manuellen Auswertung der Suchergebnisse. Diese wurde im Rahmen dieser Arbeit nicht vorgenommen. Daher sind unpassende Einträge (d. h. solche, die nichts mit dem Suffix *-(X)msX* zu tun haben) nicht herausgefiltert, wie z. B. *şıpırdımsu*, das leicht mit Hilfe der standardtürkischen Beschreibung als Kompositum identifiziert werden kann, in welchem die letzte Silbe kein Teil des Morphems *-(X)msX* darstellt, sondern das Lexem *su* (dt.: Wasser).

3.5.2 Verwandtschaftstermini

Verwandtschaftstermini sind eine interessante Gruppe semantisch „relationaler Begriffe“[71], für die es nebst standardsprachlicher Lexeme oftmals eine Vielzahl dialektaler Formen gibt. Sie ist interessant, da sie Aufschluss über soziale Strukturen der Sprecherethnie liefern kann. Lexeme zur Bezeichnung der verwandtschaftlichen Verhältnisse geben an, welche Regeln und Normen im Verhalten zwischen Verwandten bestehen, aber auch zu Außenstehenden, sofern sie für diese als Anrede verwendet werden[72].

In der Standardsprache des Türkischen gibt es folgende Verwandtschaftsbezeichnungen:

abla	ältere Schwester
ağabey	älterer Bruder
amca	Onkel (väterlicherseits)
anababa, ana baba	Eltern
anne	Mutter
anne ve baba	Eltern
anneanne	Großmutter (mütterlicherseits)
baba	Vater
babaanne	Großvater (väterlicherseits)
bacanak	Schwager (Gatte der Schwester der Ehefrau)
baldız	Schwägerin (Schwester der Ehefrau)
büyük anne-baba	Großeltern
büyükanne	Oma

71 Vgl. Lehmann – http://www.christianlehmann.eu/ling/lg_system/sem/verwandtschaftsterminologie.php
72 Vgl. Lehmann – http://www.christianlehmann.eu/ling/lg_system/sem/verwandtschaftsterminologie.php
Vgl. Wikimedia Foundation Inc. – http://en.wikipedia.org/wiki/Kinship_terminology
Vgl. Wikimedia Foundation Inc. – http://de.wikipedia.org/wiki/Verwandtschaftsterminologie

büyükbaba	Opa
dayı	Onkel (mütterlicherseits)
dede	Opa
dede-nine	Großeltern
ebeveyn, ebeveynler	Eltern
elti	Schwägerin (Gattin des Bruders der Ehefrau)
enişte	Schwager (Gatte der Schwester) Onkel (angeheiratet)
erkek kardeş	jüngerer Bruder
görümce	Schwägerin (Schwester des Ehemanns)
hala	Tante (väterlicherseits)
kayınbirader	Schwager (Bruder des Ehemanns oder der Ehefrau)
kız kardeş, kızkardeş	jüngere Schwester
nine	Oma
teyze	Tante (mütterlicherseits)
üvey anne	Stiefmutter
üvey baba	Stiefvater
yenge	Schwägerin (Gattin des Bruders) Tante (angeheiratet)

Um den Nutzen der Wahlmöglichkeit der Provinzen zu demonstrieren, werden nachfolgend die Dialektwörter der Verwandtschaftstermini in der Provinz *Kayseri* aufgelistet. Eine Auflistung aller türkeitürkischen Dialektwörter zur Benennung der Verwandtschaftsbeziehungen würde durch ihren Umfang von mehreren dutzend Seiten primär die Seitenzahl unnötig nach oben treiben und dabei den Nutzen der Suchoption nicht besser darstellen, als die derer aus *Kayseri*. Womöglich würde gar dem Leser/der Leserin durch eine komplette Auflistung aufgrund deren Länge und scheinbaren Unübersichtlichkeit der Nutzen der Anwendung verschleiert werden.

Bei den folgenden Einzelsuchen wurde jeweils keine der Suchoptionen aktiviert, mit Ausnahme Auswahl der Provinz *Kayseri*.

Suchfeld: *abla*

1)	aba	-Krş.; Dadağı -Ky.
	abba	* Bünyan -Ky.
2)	ade	Karahisar, * İncesu, * Develi, Bürüngüz * Bünyan -Ky.
3)	edi	Şehşaban * İncesu -Ky.
4)	sitti	-Ky.

Darüber hinaus stellt *abla* selbst ein Dialektwort mit der Bedeutung „Frau, Fräulein, Dame“ (trk.: hanim, hanimefendi) im Dorf *Fakıekinciliği*[73] des Landkreises *Pınarbaşı* dar, in welchem es von der afscharischen Bevölkerung verwendet wird (Angabe im Wörterbuch: Avşar aşireti, Fakıekinciliği * Pınarbaşı -Ky.).

Suchfeld: *ağabey*

1)	â	Mahzemin -Ky.
2)	ede	* Pınarbaşı ve köyleri, * Bünyan -Ky.
3)	ede	Kızılcın * Sarıoğlan -Ky.
4)	guçca	Kızılcın * Sarıoğlan -Ky.
5)	küçükağa	-Ky.
6)	mınna	-Ky.

Suchfeld: *amca*

1)	emmi	Kızılcın * Sarıoğlan, Muncusun -Ky.
	emi	* Bünyan -Ky.

Suchfeld: *anne*

1)	aba	Gergeme, * Bünyan, * İncesu köyleri, Kelgin, * Develi, Kızılcın * Sarıoğlan -Ky.
2)	ade	* Develi -Ky.

73 Vgl. Wikimedia Foundation Inc. – http://tr.wikipedia.org/wiki/Fakıekinciliği,_Pınarbaşı

Suchfeld: *baba*

1)	çece	-Ky.
2)	ede	* Erkilet -Ky.
3)	munna	Kızılcın * Sarıoğlan -Ky.

Darüber hinaus ist *baba* selbst ein Dialektwort im Landkreis *İncesu* (Angabe im Wörterbuch: * İncesu -Ky.), das „eine Lungenkrankheit bei Menschen und Schafen" bezeichnet (trk.: Koyun ve insanların akciğerinde olan bir hastalık). Darüber hinaus ist es im Landkreis *Pınarbaşı* und in den die Provinz (oder die gleichnamige Stadt) *Kayseri* umgebenden Dörfern (Angabe im Wörterbuch: * Pınarbaşı, -Ky. ve köyleri) mit der Bedeutung „großes unheilbares Geschwür, Seuche, Elend, Krankheit" (trk.: Büyük ve onulmaz çıban, veba, dert, hastalık) attestiert.

Suchfeld: *büyükanne*

1)	bibi	Ortaköy * İncesu -Ky.
3)	ebe	* Pınarbaşı ve köyleri -Ky.
2)	eci	Akkışla -Ky.

Suchfeld: *dayı*

1)	dayni	Karahisar * Develi -Ky.

Darüber hinaus ist *dayı* selbst ein Dialektwort in afscharischen Dörfern und in der Gemeinde *Pazarören*[74] des Landkreises *Pınarbaşı* (Angabe im Wörterbuch: Avşar aşireti, Pazarören * Pınarbaşı -Ky.) für den „Vorarbeiter / Chef" (trk.: amele başı).

Suchfeld: *elti*

In ganz *Kayseri* (Provinz oder gleichnamige Hauptstadt; Angabe im Wörterbuch: -Ky.) ist *elti* für „eine abfällige Bezeichnung für zwei Frauen eines Mannes, welche sie sich gegenseitig geben; Nebenfrau, Nebenfrau" (trk.: Erkeğin iki karısının birbirine nisbetle aldıkları ad, ortak, kuma) dokumentiert.

74 Vgl. Wikimedia Foundation Inc. – http://tr.wikipedia.org/wiki/Pazarören,Pınarbaşı

Suchfeld: *"erkek kardeş"*

1)	â	Mahzemin -Ky.

Suchfeld: *görümce*

1)	boyugüzel	Küpeli, Sarıoğlan * Bünyan -Ky.
2)	görüm	Afşar, Pazarören * Pınarbaşı -Ky.

Suchfeld: *hala*

1)	ame	Gaziler * Sarıoğlan, * Bünyan, Mahzemin, -Ky.
2)	bibi	Türkmen * Bünyan, Bakırdağı, Taşçı * Develi, -Ky.
3)	böle	* İncesu -Ky.
4)	sitti	-Ky.

Suchfeld: *kayınbirader*

1)	nazlıbey	-Ky.

Suchfeld: *nine*

1)	bibi	Ortaköy * İncesu -Ky.
2)	ebe	* Pınarbaşı ve köyleri -Ky.

Suchfeld: *teyze*

1)	sitti	* Bünyan -Ky.

Suchfeld: *"üvey anne"*

1)	analık	Mahzemin -Ky.
	analıx	Avşar ve Türkmen aşiretleri * Pınarbaşı, * Bünyan -Ky.

Suchfeld: *"üvey baba"*

1)	babalık	Akkışla -Ky.

Suchfeld: *yenge*

Das standardtürkische *yenge* besitzt in Kayseri keinen dialektalen Ausdruck, stellt allerdings einen solchen mit der Bedeutung „eine Frau, die einer frisch verheirateten Frau zeigt, wie sie eine gute Ehefrau ist“ (trk.: Geline kılavuzluk eden kadın) in afscharischen Dörfern des Landkreises *Pınarbaşı* und im nahe der Provinzhauptstadt gelegenen Dorf *Akçakaya*[75] (Angabe im Wörterbuch: Afşar köyleri, * Pınarbaşı, Akçakaya -Ky.) dar.

Alle hier nicht gelisteten Verwandtschaftstermini sind in *Kayseri* nicht belegt. Im Gegensatz zur Ergebnisliste unter 3.5.1 (S. 24) ist die der Verwandtschaftsbezeichnungen bereits aus pragmatischen Gründen um manche Einträge gekürzt worden, da in diesen die Suchbegriffe lediglich in einer Redewendung oder einem Beispielsatz enthalten waren und keinen Bezug auf das Lemma hatten. Daher wurden sie herausgefiltert, um die oben stehenden Tabellen etwas kürzer halten zu können.

Die einzeln angegebenen Suchbegriffe können auch in einer einzigen Suchanfrage an die Datenbank übermittelt werden, indem *abla ağabey amca "ana baba" "anababa" anne "anne ve baba" anneanne baba babaanne bacanak baldız "büyük anne-baba" büyükanne büyükbaba dayı dede-nine ebeveyn ebeveynler elti enişte "erkek kardeş" görümce hala kayınbirader "kız kardeş" kızkardeş nine teyze "üvey anne" "üvey baba" yenge* in das Suchfeld eingegeben und die Provinz *Kayseri* ausgewählt wird.

75 Vgl. TODAİE – http://www.yerelnet.org.tr/koyler/koy.php?koyid=251499

4 Technik

Im Folgenden wird beschrieben, was auf Hard- und Softwareseite zum Einsatz kam, um die digitale Korpusanwendung zu realisieren. Die unter 4.1, 4.2 und 4.3.2 genannten Einheiten wurden in einem Gesamtpaket namens *XAMPP*[76] (x-beliebiges Betriebssystem, *Apache*, *MySQL*, *PHP*, *Perl*) heruntergeladen, installiert und konfiguriert. Zur Konfiguration siehe auch 4.4 auf S. 42.

4.1 Webserver

Ein Webserver stellt den Besuchern einer Internetadresse deren Inhalte zur Verfügung.
Der auf dem Entwicklungssystem und dem Webspace (siehe jeweils 4.4 auf S. 42) zum Einsatz kommende Webserver ist *Apache*[77], welcher zur freien Software (engl.: freeware) gehört und daher kostenlos ist. Er ist der meistbenutzte Webserver im Internet[78] und ermöglicht die direkte Einbindung diverser Skriptsprachen, darunter auch *PHP* (siehe 4.3.2 auf S. 39).
Die Version des Webservers auf dem lokalen Entwicklungsrechner ist 2.2.17. Wie im Kapitel 4.4 (S. 42) zu lesen ist, konnten einige Informationen zur Hard- und Software des Webspace nicht in Erfahrung gebracht werden. Zu diesen zählt auch die Versionsnummer des Webservers *Apache*.

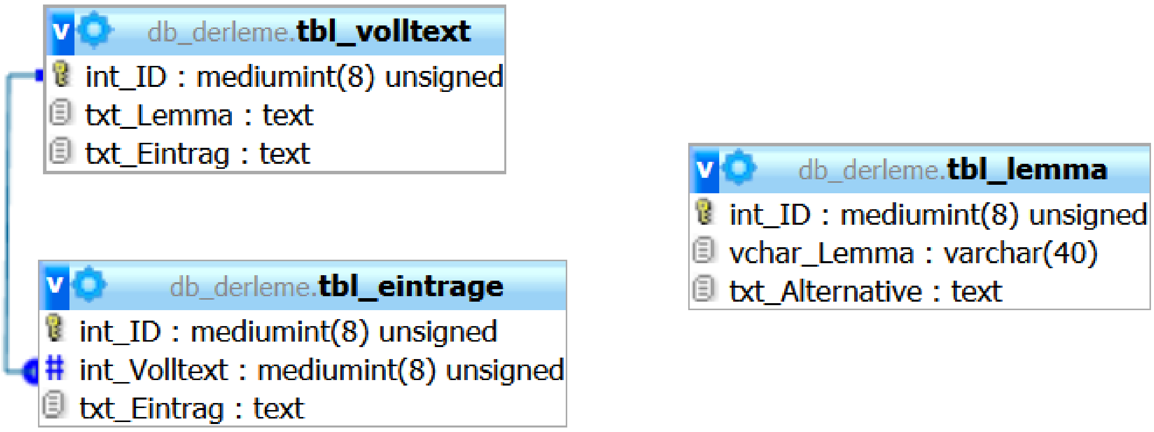
4.2 Datenbanksystem

Eine Datenbank ist ein „System zur Beschreibung, Speicherung und Wiedergewinnung von umfangreichen Datenmengen, die von mehreren Anwendungsprogrammen benutzt werden. Eine Datenbank ist meist Bestandteil eines umfassenden Informationssystems, das die Daten von der Datenbank anfordert, auswertet, nach Anwendungskriterien verarbeitet und Daten an die Datenbank zum Speichern abgibt. Datenbanken sind von zentraler Bedeutung für die Datenverarbeitung.“[79]

76 Vgl. Seidler – http://www.apachefriends.org
77 Vgl. Apache Software Foundation – http://httpd.apache.org/
78 Vgl. Apache Software Foundation – http://httpd.apache.org/
Vgl. Wikimedia Foundation Inc. – http://de.wikipedia.org/wiki/Apache_HTTP_Server
79 Schwill – S. 104f

Das in der Korpusanwendung zum Einsatz kommende *MySQL*[80] (*My* Structured Query Language, dt: *My*[81] strukturierte Abfragesprache) ist als Open-Source[82]-Datenbankmanagementsystem weit verbreitet, sehr leistungsfähig und durch entsprechende Schnittstellen in allen modernen Programmiersprachen verwendbar[83].
Aufgrund dieser Eigenschaften und bereits bestehender Kenntnisse im Umgang mit *MySQL* fiel die Wahl auf dieses System.

Das digitalisierte Wörterbuch wurde durch das *PHP*-Skript *upload.php* in die nachfolgend beschriebene Struktur der *MySQL*-Datenbank *db_derleme* eingelesen:



Es handelt sich jeweils um Tabellen des *MySQL*-eigenen Tabellentyps *MyISAM* (*My* Indexed Sequential Access Method, dt.: *My* Indizierte sequentielle Zugriffsmethode), welcher alle Inhalte kompakt speichert, wodurch sie sehr schnell gelesen werden können[84]. Die Schwachstellen bzw. Nachteile dieses Typs gegenüber anderen sind für die Korpusanwendung irrelevant, da die Tabellen ausschließlich gelesen werden und keine Fremdschlüsselüberprüfung (siehe nächster Absatz) erforderlich ist.

80 Vgl. Klaß & Ihlenfeld Verlag GmbH – http://www.golem.de/1009/78067.html
Vgl. Oracle Corporation – http://www.mysql.com
Vgl. Oracle Corporation – http://www.mysql.com/why-mysql/marketshare
81 *My* ist der Vorname der Tochter des *MySQL*-Erfinders Michael Widenius
(Vgl. Widenius – http://dev.mysql.com/doc/refman/5.1/en/history.html)
82 Open-Source bedeutet, dass der Quellcode offen zugänglich ist und dadurch Weiterentwicklungen gefördert werden sollen. Darüber hinaus ist Open-Source-Software frei/kostenlos (Vgl. Wikimedia Foundation Inc. – http://de.wikipedia.org/wiki/Open_Source). Es werden auch kommerzielle Varianten angeboten, welcher allerdings stets mit dem kompletten Titel bezeichnet werden (z. B. MySQL-Enterprise).
83 Vgl. Oracle Corporation – http://mysql.com/why-mysql/
84 Vgl. Klaß & Ihlenfeld Verlag GmbH – http://www.golem.de/1009/78067.html
Vgl. Oracle Corporation – http://dev.mysql.com/doc/refman/5.5/en/myisam-storage-engine.html
Vgl. Schwartz – S. 160

In jeder Tabelle wird jede Zeile durch eine eindeutige Identifikationsnummer (Abk.: ID) im Feld *int_ID* definiert. Dieses ist der Primärschlüssel (engl.: primary key), anhand dessen nebst der eindeutigen Identifizierung auch die Indizierung jedes Tabelleneintrags vorgenommen wird.

Der Datentyp aller ID-Felder ist „*mediumint*“[85], der durch die vorzeichenlose (engl.: unsigned) Eigenschaft Werte von 0 bis 1.677.215 annehmen kann. Der nächst-kleinere Datentyp „*smallint*“ reicht mit seiner vorzeichenlosen Obergrenze von 65.535 nicht für das komplette Wörterbuch aus, da es 126.468 Einträge enthält.

Die Tabelle *tbl_eintrage* enthält die ID der Tabelle *tbl_volltext*, um Datenredundanz zu verhindern. Andernfalls hätte in *tbl_eintrage* pro Zeile das zugehörige Lemma in einem Textfeld gespeichert werden müssen. Aufgrund der gewählten Speicherengine *MyISAM* erfolgt hierbei allerdings keine Relation der Werte und keine Zuweisung sogenannter Fremdschlüssel (engl.: foreign key). Jedoch ist die ID in einem tabelleninternen Index sortiert, damit eine Suche nach ihr schneller erfolgen kann. Die gesonderte Speicherung der ID der Volltexttabelle innerhalb der Tabelle *tbl_lemma* wäre der Vollständigkeit der Struktur wegen eigentlich vonnöten, kann jedoch aufgrund der ausnahmslosen Übereinstimmung beider Primärschlüssel (pro Wörterbucheintrag in *tbl_volltext* existiert nur eine Lemmaangabe in *tbl_lemma;* die ID wird im Skript *upload.php* hochgezählt und ist daher für beide Tabellen identisch) ausgespart werden.

Das Wörterbuch selbst ist in Textfeldern mit dem binären türkischen Zeichensatz „*latin5*“[86] gespeichert, damit Groß- und Kleinschreibung unterschieden werden kann. Eine Speicherung in einem der unterstützten Unicode-Zeichensätze ist aus folgendem Grund nicht möglich: Durch die innerhalb der Datenbank verwendete Suche mit *regulären Ausdrücken* und deren Schwachstelle (siehe S. 40) musste ein Zeichensatz gewählt werden, der pro Zeichen ein Byte Speicherplatz belegt. Damit alle türkischen Buchstaben dargestellt werden können, wurde daher „*latin5*“ gewählt. Die Verwendung des Unicode-Zeichensatzes in Zusammenhang mit der Schwachstelle bei *regulären Ausdrücken* hätte durch eine Konvertierung bei jeder einzelnen Suchanfrage umgangen werden können. Aufgrund der Tatsache, dass eine Speicherung in Unicode nicht zwingend erforderlich ist, wurde das Problem jedoch mittels Verwendung von „*latin5*“ gelöst.

85 Der Datentyp *Integer* (kurz *int*) wird zur Darstellung von Ganzzahlen verwendet.
86 Vgl. Wikimedia Foundation Inc. – http://de.wikipedia.org/wiki/ISO_8859-9

Die datenbankinterne Zeichencodierung, die der Datenbankverbindung sowie die aller *PHP*-Skripte hingegen ist Unicode „*utf8*“. Web- und Datenbankserver interagieren auf diese Art problemlos, da die „*latin5*“-codierten Daten automatisch problemlos in Unicode konvertiert werden.

Bei einer Suche mit der Option „*nur nach Lemma suchen*“ sowie der Suche ohne Angabe einer Provinz oder der Suchoption „*strenge Suche*“, wird auf die Tabelle *tbl_volltext* zugegriffen. In allen anderen Fällen muss in der Tabelle *tbl_eintrage* gesucht werden, in welcher alle Einträge des Wörterbuchs zeilenweise vorhanden sind. Aufgrund der Wörterbuchstruktur ist hierfür die Tabelle *tbl_lemma* nötig: Wenn zeilenweise gesucht wird, dürfen die Alternativen eines Lemmas in der Lemmazeile selbst nicht geprüft werden. Andernfalls gäbe es inkorrekte Suchergebnisse, wie z. B. bei der Suche nach *ede* in der Provinz *Çankırı*: die Provinz darf nicht innerhalb der ersten Eintragszeile markiert werden, da *ede* nicht in *Çankırı* auftritt, sondern *eci* (was das eigentliche Lemma des Eintrages ist). Damit also nur das Lemma und die erste Zeile des Eintrags miteinander in der strengen Suche verbunden werden können, wird dieses mit der ersten Zeile in *tbl_eintrage* gespeichert. Um in der Ausgabe der Ergebnisse jedoch die übliche Darstellung zu erhalten, ist es vonnöten, die entsprechenden Angaben aus *tbl_lemma* zu lesen und in fetter Schrift formatiert dem Suchergebnis voranzustellen.

Auf die Verwendung einer Wortliste wurde bewusst verzichtet, da aufgrund der geringen Datenmenge des Volltexts von lediglich 12,4 Megabyte und der datenbankinternen Verarbeitung dieser Daten in Tests (nach Erstellung der Wortliste und entsprechender Speicherung in separater Tabelle) keinerlei Geschwindigkeitsvorteile erreicht werden konnten. Bei der Aktivierung der Suchoption „*strenge Suche*“ verlangsamte sich die Ausgabe der Ergebnisse sogar durchschnittlich um den Faktor 10, da *MySQL* allem Anschein nach verschachtelte Abfragen mit Sortieranweisung nur sehr langsam ausführen kann.

Die Serverversion der Datenbank trägt die Nummer 5.1.57; die des lokalen Entwicklungssystems 5.5.10

4.3 Programmiersprachen, Seitenbeschreibungssprachen, reguläre Ausdrücke

Eine *Programmiersprache* ist eine „formale Sprache“[87], die zur „Formulierung von Rechenvorschriften … die von einem Computer ausgeführt werden können“[88] dient. „Programmiersprachen bilden die wichtigste Schnittstelle zwischen Benutzern und Computern“[89].

Seitenbeschreibungssprachen hingegen geben lediglich an, wie eine Seite (am Bildschirm oder für den Drucker) dargestellt werden soll. In ihr sind i. d. R. keine Funktionen oder Befehle kreierbar[90].

Reguläre Ausdrücke gehören ebenfalls zur Gruppe der formalen Sprachen, dienen jedoch im Unterschied zu Programmiersprachen lediglich der Beschreibung und Erstellung textueller Muster.

Nachfolgend werden alle in der Korpusanwendung genutzten formalen Sprachen aufgelistet und näher beschrieben.

4.3.1 HTML (Hypertext Markup Language, dt: Übertext-Auszeichnungssprache) ist die Standardsprache um Webseiten zu gestalten. Das Erstellen von Webseiten wird von Menschen, die nicht innerhalb der Informatik tätig sind, als Programmierung bezeichnet, obwohl dies falsch ist. Da in *HTML* keine eigenen Objekte oder Funktionen kreierbar sind, wird sie der Gruppe der Seitenbeschreibungs- bzw. Auszeichnungssprachen zugeordnet. *HTML*-Webseiten werden somit nicht programmiert, sondern geschrieben.

Die Benutzeroberfläche der Korpusanwendung sowie die Übermittlung der Sucheingaben an den Webserver ist mit HTML realisiert worden.

87 Wikimedia Foundation Inc. – http://de.wikipedia.org/wiki/Programmiersprache
88 Schwill – S. 379
89 Ebd.
90 Vgl. Schwill – S. 371
Vgl. Wikimedia Foundation Inc. – http://de.wikipedia.org/wiki/Auszeichnungssprache
Vgl. Wikimedia Foundation Inc. – http://de.wikipedia.org/wiki/Seitenbeschreibungssprache

4.3.2 PHP (PHP Hypertext Preprocessor, dt: PHP Hypertext Präprozessor[91]) ist eine weit verbreitete und serverseitig ausgeführte Programmiersprache (Skriptsprache), mit der primär dynamische Webseiten gestaltet werden. Mit ihr kann auf Benutzereingaben reagiert und somit der Inhalt der Webseite dynamisch dargestellt werden. Ferner sind der Zugriff auf das Dateisystem, die Erstellung von Grafiken oder PDF-Dateien, das Versenden von Emails, die Verbindung zu Datenbanken uvm. möglich.
Skripte, die in *PHP* geschrieben werden, werden auf dem Webserver ausgeführt. Lediglich die Ausgabe wird an den Benutzer übermittelt. Dies hat den Vorteil, dass der Quellcode nicht einsehbar ist, der Anwender Ausgaben erhält, die unabhängig seines Computers generiert wurden und diese dadurch, sowie aufgrund des Umstands, dass Serverhardware i. d. R. leistungsfähiger ist, deutlich schneller ausgeführt werden können. Ferner werden *PHP*-Skripte durch die direkte Einbindung in den Webserver äußerst schnell gestartet, da der zuständige Prozess bereits läuft und auf Aufgaben wartet[92].
Darüber hinaus unterstützt sie einige wichtige Datenbankformate, u. a. *MySQL*, und die Verwendung von sogenannten *regulären Ausdrücken* (siehe 4.3.4 auf S. 40).

Die beiden Hauptskripte *upload.php* und *index.php* wurden in *PHP* programmiert, ebenso die Funktionsskripte *config.php*, *db_connect.php*, *farbmarkierung.php* und *get_phrases.php* im Ordner *functions*. Erstere und die genannten Funktionsskripte bestehen ausschließlich aus *PHP*-Befehlen, zweitere enthält *HTML*- und *JavaScript*-Code, damit die Ausgaben und Benutzeraktionen entsprechend im Browser dargestellt werden. Die Datei *functions.js* im Ordner *functions* enthält ausschließlich *JavaScript*-Befehle.

Die Wahl der verwendeten Programmiersprache zur Verarbeitung des Wörterbuchs und zur Erstellung und Nutzung der Abfrageanwendung fiel auf *PHP*, da sie weit verbreitet und schnell ist, eine Vielzahl nützlicher Funktionen bereitstellt[93] und bereits vor Beginn dieser Arbeit sehr gute Kenntnisse in dieser Sprache vorhanden waren.
Zur möglichen Alternative *Perl* siehe Kapitel 4.4 (S. 42).

Die auf dem Server verwendete *PHP*-Version ist 5.3.10. Auf dem lokalen

91 „*PHP*“ ist ein rekursives Akronym
92 Vgl. Carstensen – S. 479
93 Vgl. Olson – http://php.net/manual/de/index.php

Entwicklungsrechner wird Version 5.3.6 gebraucht.

4.3.3 JavaScript ist ebenfalls eine Skriptsprache, die ausschließlich auf dem Benutzercomputer ausgeführt wird und primär dazu benutzt wird, dynamische *HTML*-Seiten zu erstellen. Mit ihr ist es möglich, Benutzereingaben zu überprüfen und die Eingaben nur bei erfolgreicher Prüfung an den Server zu übermitteln. Darüber hinaus kann mit ihr auf Benutzereingaben reagiert werden (z. B. farbliche Hervorhebung eines Elements, sobald die Maus sich darüber befindet). Auch die Inhalte aller *HTML*-Elemente können nach Abruf der Webseite verändert werden.
Damit *JavaScript*-Code ausgeführt werden kann, muss dies im Browser aktiviert sein. Standardmäßig ist dies der Fall, doch sobald der Benutzer die Verwendung deaktiviert, funktionieren *JavaScript*-Befehle nicht mehr. Im Falle der Korpusanwendung bedeutet dies, dass keine Provinzen mehr angegeben werden können.

4.3.4 reguläre Ausdrücke (engl.: regular expressions, Abk.: *RegExp* oder *Regex*)
Mithilfe *regulärer Ausdrücke* kann nach bestimmten Textmustern bzw. Mengen von Zeichenketten gesucht werden. Sie können darüber hinaus ersetzt oder generiert werden. Durch sie ist es möglich, sich „Dutzende, wenn nicht Hunderte von Zeilen prozeduralen Codes“[94] zu sparen. Durch *reguläre Ausdrücke* kann nach nicht vollständig bekannten Mustern gesucht werden, da diverse Platzhalter unterstützt werden. Damit ist die linguistisch oftmals relevante Suche nach Wortformen oder Derivaten möglich. Sie sind „ein mächtiges Mittel, um große Datenbestände nach komplexen Suchausdrücken zu durchforsten“[95].

Reguläre Ausdrücke zählen zu den formalen Ausdrücken innerhalb der theoretischen Informatik und daher zu formalen Sprachen, die in der Chomsky-Hierarchie den schwächsten Typ (Typ 3) darstellen[96]. Moderne *reguläre Ausdrücke* gehören aufgrund der Eigenschaft, ein Rücksetzverfahren (auch Rückverfolgung, engl.: backtracking[97]) zu ermöglichen, streng genommen nicht mehr zum dritten Typ, da sie durch das Backtracking nicht mehr kontextfrei sind[98].

94 Goyvaerts – S. 18
95 Selfhtml e. V. – http://de.selfhtml.org/perl/sprache/regexpr.htm
96 Vgl. Wikimedia Foundation Inc. – http://de.wikipedia.org/wiki/Regulärer_Ausdruck
Vgl. Wikimedia Foundation Inc. – http://de.wikipedia.org/wiki/Reguläre_Sprache
97 Vgl. Wikimedia Foundation Inc. – http://de.wikipedia.org/wiki/Backtracking
98 Vgl. Wikimedia Foundation Inc. –

Es gibt keinen allgemein gültigen Standard, der definiert, bei welchem Textmuster es sich definitiv um einen gültigen *regex* handelt[99]. Dadurch treten bei der Anwendung eines *regulären Ausdrucks* innerhalb verschiedener Programme bzw. Programmiersprachen oftmals große Unterschiede in der Ausführung und Ausgabe auf, da der eingegebene Ausdruck als falsch angesehen wird, obwohl er in einer anderen Programmiersprache bzw. in einem anderen Programm korrekt getätigt wird.
Die *Regexes* (Abk. zu engl. regular expressions) in *PHP* werden in allen verwendeten Funktionen durch die Bibliothek *PCRE* (engl.: perl compatible regular expression, dt.: mit Perl kompatibler regulärer Ausdruck) ausgeführt[100].
Die innerhalb *MySQL's* benötigen die Angabe von Zeichenklassen in eckigen Klammern, da sie die jeweilige Abkürzung mittels Backslash nicht unterstützen[101]. Ferner ist in ihnen kein Backtracking möglich. Darüber hinaus sind sie nicht *multibyte-sicher*[102], d. h., dass sie nur fehlerfrei funktionieren, wenn der verwendete Zeichensatz pro Zeichen ein Byte Speicherplatz belegt. Somit sind sie im Unicode nicht zu gebrauchen (siehe auch Beschreibung des verwendeten Zeichensatzes „*latin5*“ auf S. 36).

4.3.5 CSS (cascading stylesheets, dt.: stufenförmige Gestaltungsvorlage)

Durch *CSS*, welches eine *HTML* erweiternde Seitenbeschreibungssprache ist, werden Formatierungseigenschaften von *HTML*-Elementen angegeben. *CSS* können im Kopf der *HTML*-Seite (Element <*head*>) oder in der Eigenschaft namens *style* innerhalb nahezu jeden *HTML*-Elements stehen.
Sie kommen in der Korpusanwendung nur geringfügig vor, sind jedoch insbesondere für die farbliche Hervorhebung der Suchergebnisse unverzichtbar.

4.4 Hardware

Die gesamte Entwicklung fand auf einem lokalen Arbeitsplatzrechner statt, da auf diesem unabhängig der Internetverbindung und ohne die Notwendigkeit, jede einzelne Änderung neu auf den Server zu laden, entwickelt und getestet werden konnte.

http://de.wikipedia.org/wiki/Regulärer_Ausdruck#Reguläre_Ausdrücke_in_der_Praxis
99 Vgl. Goyvaerts – S. 18
100 Vgl. Goyvaerts – S. 23
Vgl. Olson – http://de2.php.net/manual/en/book.pcre.php
101 Vgl. Oracle Corporation – http://dev.mysql.com/doc/refman/5.5/en/regexp.html
102 Vgl. Oracle Corporation – http://bugs.mysql.com/bug.php?id=34473

Die für die Anwendung relevante Hardware ist wie folgt:

CPU	AMD Phenom X4 965 3,4 GHz
RAM	Mushkin Enhanced Silverline 8 GB (2 x 4 GB), DDR3 1333 MHz, Timing 9-9-9-24)
HDD	Western Digital Caviar Black, SATA III, 640 GB, 7200upm, 64 MB Cache (Modelnummer WD6402AAEX) Mittlerer Datendurchsatz: 85 Megabyte pro Sekunde[103]
Betriebssystem:	Windows 7 Ultimate Service Pack 1 64bit

Der für das Datenbanksystem und die Ausführung der *PHP*-Skripte zur Verfügung stehende Arbeitsspeicher wurde auf jeweils 512 Megabyte gesetzt.

Alle relevanten Dateien wurden nach Abschluss der Programmierung[104] auf den gemieteten Webspace geladen, welcher hinsichtlich der Hardware leider nicht näher beschrieben werden kann. Dies liegt am Umstand, dass der gemietete Webspace nicht auf einem eigenständigen, nur diese Webseite beherbergenden (ggbfs. virtuellen) Server liegt. Es kann daher nicht gesagt werden, welche Leistung tatsächlich für die Korpusanwendung zur Verfügung steht. Der Anbieter *alfahosting*[105] teilte auf eine diesbezügliche Anfrage mit, dass „derartige Auskünfte aus Sicherheitsgründen nicht zur Verfügung“[106] stehen. Lediglich die Menge des für ausgeführte *PHP*-Skripte zur Verfügung stehenden Arbeitsspeichers kann im Kundenbereich eingesehen werden: 92 Megabyte.
Durch eigene Nachforschungen bezüglich der Serverangaben konnte mittels Aufrufs der *PHP*-Funktion *phpinfo()* herausgefunden werden, dass im Server eine 64bit-CPU zum Einsatz kommt und als Betriebssystem eine Linuxvariante des Anbieters *Debian* mit dem Kernel 2.6.32 64bit genutzt wird. Weiteres konnte nicht in Erfahrung gebracht werden.

Zu Beginn der Arbeit wurde ein kostenloser Webspace genutzt, welcher sich nach einigen Wochen allerdings als unzuverlässig und zu langsam erwies. Die Anwendung war des Öfteren

103Test mittels „*DiskSpeed32*“ (Grinenko – http://software.vgrin.host56.com/diskspeed32)
104Streng genommen handelt es sich beim Schreiben des Programmcodes um den Vorgang der *Codierung.* Der Vorgang der *Programmierung* stellt innerhalb der Informatik die Entwicklung der Funktionen des Programms dar, meist in Form eines so genannten *Programmablaufplans* oder eines *Struktogramms*. Die Verwendung des Begriffs *Programmierung* hat sich jedoch allgemein etabliert, auch innerhalb der Informatik.
105Vgl. Alfahosting GmbH – http://www.alfahosting.de
106Supportanfrage auf https://alfahosting.de/kunden/ am 15.02.2011

nicht erreichbar und bearbeitete teilweise selbst einfache Suchanfragen im zweistelligen Sekundenbereich, was als eindeutig zu langsam zu bezeichnen ist.
Auf diesem anfänglichen Webspace war ausschließlich *PHP* verfügbar und somit wurde *Perl*[107] als mögliche Programmiersprache ausgeschlossen. *Perl* ist ebenfalls eine weit verbreitete Skriptsprache, die vom Linguisten Larry Wall[108] entwickelt wurde, insbesondere für die Verarbeitung von Textdaten sehr gut geeignet ist und darüber hinaus eine sehr gute Implementierung *regulärer Ausdrücke* aufweist[109].
Nach dem Wechsel des Anbieters bestand zwar die Möglichkeit *Perl* zu nutzen, allerdings wurde der bis dahin fast vollständige Programmcode auch im Hinblick auf den zeitlichen Rahmen der Arbeit nicht mehr auf *Perl* umgeschrieben und daher letztlich ausschließlich *PHP* genutzt.

4.5 Software

Der Quellcode aller Skripte wurde im Open-Source-Texteditor *Notepad++*[110] der Version 5.9.6.2 geschrieben. Die Erstellung und Verwaltung der Datenbank wurde mit *phpMyAdmin*[111] vorgenommen. Auf dem Server kam Version 3.4.7 zum Einsatz; auf dem lokalen Entwicklungsrechner 3.3.9. Darüber hinaus wurde auf beiden Systemen *MySQLDumper*[112] 1.24.4 verwendet, um die lokal gefüllten Tabellen auf den Webserver zu portieren und Sicherheitskopien zu erstellen[113].
Die Versionsinformationen zu *Apache*, *MySQL* und *PHP* sind in den entsprechenden Unterkapiteln ab Seite 35 gegeben.

4.6 Leistung

Die Leistung im Sinne der benötigten Zeit zur Ausführung der Suchanfrage und Verarbeitung sowie Darstellung der Ergebnisse wurde durch verschiedene, nachfolgend angegebene

107Vgl. Perl Foundation – http://www.perl.org
108Vgl. Wall – http://www.wall.org/~larry/
109Vgl. Carstensen – S. 475
110Vgl. Ho – http://notepad-plus-plus.org/
111Vgl. Delisle – http://www.phpmyadmin.net/home_page/index.php
112Vgl. Schlichtholz – http://www.mysqldumper.de/
113Sicherheitskopien über die entsprechende *phpMyAdmin*-Funktion sind aufgrund der vielen Einträge der Tabellen nicht möglich. Um komfortabel mit grafischer Oberfläche Backups anlegen zu können fiel daher die Entscheidung, *MySQLDumper* zu verwenden.

beispielhafte Suchanfragen ermittelt. Sie alle wurden mehrfach auf dem lokalen Entwicklungscomputer ausgeführt, wobei nach jeder Ausführung der Zwischenspeicher geleert wurde, um zu verhindern, dass das Suchergebnis aus diesem geladen wird[114]. Die Tests wurden lokal ausgeführt, da der Webserver hinsichtlich der Hardware leider nicht beschrieben werden kann und somit dessen Antwortzeiten letztlich nicht viel über die Leistungsfähigkeit aussagen. Allerdings muss angemerkt werden, dass sie sich nicht wesentlich von der des Entwicklungscomputers unterschieden.

Durch die genaue Kenntnis der Hard- und Softwareeigenschaften des Entwicklungs-PCs sind nachfolgende Daten bezüglich der Leistungsfähigkeit der Korpusanwendung aussagekräftig. Falls Erweiterungen oder umfangreichere Neuentwicklungen (siehe Kapitel 5.2 auf S. 47) durchgeführt werden und im Vorfeld abgeschätzt werden soll, welche Hardwarekomponenten welche Antwortzeiten zum Ergebnis haben könnten, stellen sie hierfür einen Anhaltspunkt dar. Der PC hatte während der Tests keine weiteren Prozesse geöffnet, die vordergründige Rechenleistung benötigten.

Bei jeder Suche ist daran zu denken, dass das komplette digitalisierte Wörterbuch in der jeweils genannten Zeit durchsucht wurde. Die Zeitangaben sind gerundet.

Suchbegriff(e)	**Optionen**	**Ergebnisse**	**benötigte Zeit in Sekunden**
insan* hayvan	und-Verknüpfung	179	0.17
??*\D\X\K\X*		3348	2.32 2.34
"erkek kardeş" büyük	und-Verknüpfung	4	0.14 0.14
"erkek kardeş" büyük	und-Verknüpfung strenge Suche	3	0.19 0.14
anne baba	oder-Verknüpfung Groß/Kleinschreibung	75	0.23
anne baba	oder-Verknüpfung	129	0.27
"erkek kardeş" büyük	oder-Verknüpfung	1355	0.39
"erkek kardeş" büyük	oder-Verknüpfung Groß/Kleinschreibung	704	0.25
*ms\X	nur nach Lemma	55	0.07

114Datenbankabfragen werden zwischengespeichert, um sie bei erneuter Abfrage direkt aus dem *Cache* (Zwischenspeicher) laden zu können. Dadurch muss die Abfrage nicht erneut ausgeführt werden und die Ergebnisse werden somit um ein Vielfaches schneller (< 0.0001 Sekunden) bereitgestellt.

*\Arg\A	nur nach Lemma	117	0.08
İnsan*	Groß/Kleinschreibung	158	0.13 0.14
abla ağabey amca "ana baba" "anababa" anne "anne ve baba" anneanne baba babaanne bacanak baldız "büyük anne-baba" büyükanne büyükbaba dayı dede-nine ebeveyn ebeveynler elti enişte "erkek kardeş" görümce hala kayınbirader "kız kardeş" kızkardeş nine teyze "üvey anne" "üvey baba" yenge	oder-Verknüpfung Provinz *Kayseri*	34	2.53 2.57

Anhand dieser beispielhaften Suchanfragen zur Ermittlung der Leistung ist klar zu erkennen, welch enorme zeitliche Ersparnis durch die Entwicklung der Anwendung möglich ist. Binnen Sekunden ist das Wörterbuch durchsucht und die Ergebnisse sind in geeigneter Weise ausgegeben. Eine manuelle Suche, welche aufgrund der Größe des *Derleme Sözlüğü* zumindest mehrere Tage benötigen würde, kann keinesfalls mit der Suche mithilfe der Korpusanwendung konkurrieren.

5 Fazit, Ausblick

5.1 Fazit

Nachdem die Entwicklungsarbeiten abgeschlossen sind und die Funktionalität der Korpusanwendung mehrfach getestet wurde, gilt das in der Einleitung genannte Ziel als erreicht. Aus sprachwissenschaftlicher Perspektive betrachtet gab es keine Probleme während oder nach der Erstellung der Anwendung. Aus der Sicht des Softwareentwicklers sind allerdings zwei nicht signifikante, aber insbesondere im Hinblick auf mögliche Weiterentwicklungen dennoch zu nennende Probleme aufgetreten:

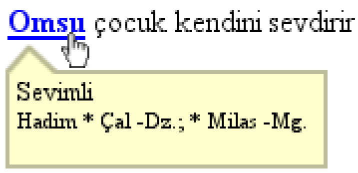
Reguläre Ausdrücke sind in *MySQL* derzeit nicht „*multibyte-sicher*“ (siehe S. 36). Dies führt zu Fehlern, sobald nach Begriffen gesucht werden soll, die ğ, ı, ç, Ç, ş, Ş, oder İ beinhalten. Das Problem konnte zwar durch die Nutzung des Zeichensatzes „*latin5*“ behoben werden, jedoch muss es für künftige Entwicklungen jeglicher Art unbedingt in der Planungsphase bedacht werden. Sofern Unicode unverzichtbar sein und diese *MySQL*-Schwachstelle noch bestehen sollte, müsste das Datenbanksystem gewechselt werden.

Ein weiteres Problem stellte die Zeichenkettenverarbeitung in *PHP* dar, die in einigen Funktionen (z. B.: *strtolower(), str_replace(), stripos()*) mit den oben gelisteten türkischen Buchstaben nicht fehlerfrei funktioniert. *PHP* besitzt allerdings in der Standardkonfiguration „*multibyte-sichere*“ Pendants dieser Funktionen, mit welchen alle Zeichenketten korrekt verarbeitet werden konnten.

Insgesamt verlief die Programmierung aus Sicht eines Entwicklers sehr gut. Die genannten Schwierigkeiten werden bei eventuellen Ergänzungen oder Neuentwicklungen ähnlicher Anwendungen berücksichtigt werden, indem ein für *reguläre Ausdrücke* besser geeignetes Datenbanksystem verwendet wird sowie eine Programmiersprache, die für die Suche nach Unicode-Textmustern zuverlässigere Funktionen bereitstellt.

5.2 Ausblick

Durch eine Erweiterung könnte die Anwendung künftig benutzt werden, um in Feldforschungen gesammeltes Material zu Dialekten der Türkei mit dem Inhalt des Dialektwörterbuchs zu verknüpfen und dadurch eine automatisierte Übersetzungshilfe bzw. Bedeutungsangabe des Standardtürkischen zu geben. Denkbar ist dies in z. B. folgender Art:



Eine weitere denkbare Erweiterung der Korpusanwendung wäre die Verknüpfung mit einem digitalisierten Wörterbuch des Standardtürkischen, um die dort vorhandenen Lemmata mit deren dialektalen Varianten zu verbinden.

Für das Deutsche, Englische und weitere Sprachen existieren derzeit einige teils sehr große Korpora, die von Linguisten weltweit täglich für deren Untersuchungen verwendet werden. Besonders hervorzuheben sind:

- der *British National Corpus* (BNC) mit 100 Millionen[115] Token (~ Gesamtzahl der Wörter eines Textes[116]) zum britischen Englisch,
- das *Digitale Wörterbuch der Deutschen Sprache* (DWDS) mit insgesamt zwei Milliarden[117] Token,
- das *Corpus Search, Management and Analysis System* (COSMAS II) mit insgesamt rund 5,4 Milliarden[118] Token,
- und das Projekt *Wortschatz* der Universität Leipzig, in welchem 136 monolinguale Wörterbücher enthalten sind[119], welche allesamt als Download verfügbar sind und dabei zusätzlich jeweils 100.000 bis 1.000.000 Sätze der jeweiligen Sprache (bzw. Varietät) beinhalten[120].

115 Vgl. University of Oxford – http://www.natcorp.ox.ac.uk
Vgl. Wikimedia Foundation Inc. – http://de.wikipedia.org/wiki/British_National_Corpus
116 Vgl. Metzler – S. 702
117 Vgl. Berlin-Brandenburgische Akademie der Wissenschaften – http://www.dwds.de/resource/zeitungskorpora
Vgl. Berlin-Brandenburgische Akademie der Wissenschaften – http://www.dwds.de/resource/spezialkorpora
Vgl. Berlin-Brandenburgische Akademie der Wissenschaften – http://www.dwds.de/resource/referenzkorpora
Vgl. Wikimedia Foundation Inc. – http://de.wikipedia.org/wiki/DWDS
118 Vgl. Institut für Deutsche Sprache – http://www.ids-mannheim.de/cosmas2/uebersicht.html
119 Vgl. Universität Leipzig – http://corpora.informatik.uni-leipzig.de/
120 Vgl. Universität Leipzig – http://corpora.informatik.uni-leipzig.de/download.html

Für das Türkische existiert bis dato leider keine auch nur annähernd vergleichbare Korpusanwendung (siehe S. 11). Das *Wortschatz*-Projekt der Universität Leipzig enthält zwar auch für das Türkische einen Datensatz mit 100.000 Sätzen, in denen rund 1,7 Millionen Token[121] enthalten sind, jedoch reicht diese Datenmenge nicht für aussagekräftige Studien aus. Viele Suchen enthalten aufgrund der geringen Textmenge keine oder nur wenige Treffer, obwohl es sich um eine übliche Konstruktion handelt (z. B. *-DXğX zamanda* mit lediglich zwei Treffern*).* Dies gilt es in naher Zukunft zu ändern.

Das Ziel ist die Realisierung eines türkischen Nationalkorpus, welches die Sprache der Türkei durch eine Vielzahl enthaltener Texte aus verschiedenen Textsorten und Jahrzehnten darstellt und mittels diverser Such- und Präsentationsoptionen ein mächtiges linguistisches Werkzeug zur Erforschung der Türkischen Sprache ist.

Aussagekräftige Recherchen sind nur mit einer entsprechend großen Textmenge zu erhalten. Aus diesem Grund beläuft sich die Wortanzahl verschiedener Korpora (siehe oben) auf mehrere hundert Millionen bis hin zu einigen Milliarden. Die Gesamtzahl der in dieser Neuentwicklung enthaltenen Wörter sollte sich daher ebenfalls auf mindestens 100 Millionen belaufen. Dadurch wäre dieses Korpus mit u. a. den nachfolgend genannten Eigenschaften und Funktionen hinsichtlich der Erforschung des Türkischen ebenso relevant und bedeutsam wie oben genannte (Teil)Korpora für das Englische oder Deutsche.

Alle im Korpus gespeicherten Texte und die darin enthaltenen Daten sollen annotiert werden, insbesondere morphologisch. Darüber hinaus soll neben der Möglichkeit der Suche mittels Platzhaltern und Coversymbolen nach Wortpaaren gesucht werden können. Die Suchergebnisse sollen farblich hervorgehoben werden und mit einer festen oder frei wählbaren Anzahl von Wörtern im Kotext ausgegeben werden (Konkordanz[122]). Ferner sollen statistische Angaben (z. B. Verteilung/Frequenz der Suchbegriffe, vergleichend auch in anderen Teilkorpora; Kollokationen[123]; Kookkurrenz-Analyse[124]) abrufbar sein und ggbfs. in Diagrammen dargestellt werden.

121 Analyse mittels *antconc* (Vgl. Anthony – http://www.antlab.sci.waseda.ac.jp/antconc_index.html)

122 Konkordanz ist „eine Sammlung von Kotexten eines bestimmten Schlüsselworts. Kotexte einer bestimmten Länge … um ein Schlüsselwort herum werden aus einem Korpus extrahiert ...“ (Lemnitzer – S. 188). Oftmals in Korpussystemen und als KWIC-Konkordanz angegeben (Keyword in context).

123 Als Kollokation wird „das wiederholte gemeinsame Vorkommen zweier Wörter in einer strukturell interessanten Einheit bezeichnet. In einer Kollokation beeinflusst ein Wort die Auswahl eines anderen Wortes zuungunsten von Wörtern mit gleicher oder ähnlicher Bedeutung.“ (Lemnitzer – S. 188).

124 Als Kookkurrenz wird „das gemeinsame Vorkommen zweier oder mehrerer Wörter in einem Kontext von fest definierter Größe bezeichnet. Das gemeinsame Vorkommen sollte höher sein, als bei einer Zufallsverteilung aller Wörter erwartbar wäre“ (Lemnitzer – S. 189).

Um den Gebrauch bestimmter Suffixe, Lexeme oder Phrasen diachron erforschen zu können, ist es außerdem vonnöten, nicht primär Texte aus den vergangenen fünf bis zehn Jahren in das Korpus aufzunehmen, sondern ebenfalls solche aus den vergangenen 50 oder mehr Jahren mit einzubeziehen.
Allgemein gilt, dass ein ausgewogenes Korpus aus Texten verschiedener Quelltypen besteht: journalistische Prosa, Belletristik, Fachliteratur, Märchen, Volkserzählungen, etc. All diese Textsorten sollen in einem eigenen Teilkorpus abgelegt werden, damit bei Bedarf speziell nur in diesem gesucht werden kann.

Mit einem solchen Nationalkorpus des Türkischen könnte innerhalb weniger Jahre eine digitale Korpusanwendung geschaffen werden, mit der Linguisten weltweit durch die einfache Zugänglichkeit via Internet Forschungen zu dieser Sprache anstellen können. Dazu gehören u. a., wie sich grammatische oder lexikalische Morpheme im Laufe der Zeit hinsichtlich ihrer Bedeutung (Semantik) oder Verwendung (Pragmatik, Semantik, Morphologie) verändert haben, welche neuen Wörter zu welcher Zeit in die Sprache gelangt sind (Lexikologie), welche Optionen bei der Anordnung syntaktischer Konstruktionen und Elemente (Syntax) bestehen, in welcher Umgebung Lexeme oder Phrasen auftreten oder wie sich das gesellschaftspolitische Geschehen der jeweiligen Zeit auf Texte bestimmter Gattungen auswirkte (Soziolinguistik, Literaturwissenschaft).

Problematisch bei diesem Vorhaben ist die legale Beschaffung der Texte, welche als Teil der Korpusanwendung allesamt in Auszügen veröffentlicht werden, sobald sie in der Ergebnisliste einer Suche auftreten. Aufgrund des Urheberrechts ist es nötig, die Einverständnisse der Verlage und Autoren zu erhalten und das könnte sich bei diesem Vorhaben durchaus als äußerst problematisch erweisen. Ferner könnte es schwierig sein, alle Texte in geeignetem Format (Word-Datei, HTML-Seite, reine Textdatei) zu erhalten und sie aus diesem direkt in die Datenbank einzupflegen. Das Extrahieren des Textes einer PDF-Datei ist oftmals nicht fehlerfrei möglich und wirkt sich daher aufgrund der nötigen Korrekturen negativ auf den Zeitaufwand aus.

Dennoch ist es an der Zeit, sich dieser Aufgabe zu widmen und für das Türkische innerhalb der nächsten Jahre eine große und funktional gut durchdachte Korpusanwendung zu liefern.

6 Bibliographie

ADLER, Olivia; HOLZGRAEFE, Hartmut: *PHP lernen*, Addison-Wesley, München, 2002

CALZOLARI, Nicoletta: „Computional lexicons and corpora“ in van Sterkenburg, Piet (Hrsg.): *Linguistics Today – Facing a Greater Challange*, John Benjamins Publishing Company, Amsterdam – Philadelphia, 2004, S. 89 - 107

CARSTENSEN, Kai-Uwe; EBERT, Christian, et. al (Hrsg.): *Computerlinguistik und Sprachtechnologie – Eine Einführung*, Spektrum, Heidelberg, 2010

CHAMBERS, J. K.; TRUDGILL, Peter: *Dialectology*, Cambridge University Press, Cambridge [u.a.], 2004

CORDTS, Sönke: *Datenbankkonzepte in der Praxis*, Addison-Wesley, München, 2002

DIPPER, Stefanie: „Theory-driven and corpus-driven computational linguistics, and the use of corpora“ in Ludeling, Anke; Kyto, Merja (Hrsg.): *Corpus Linguistics. An International Handbook. Volume 1*, de Gruyter, Berlin, 2008, S. 68 – 96.

DÜRR, Michael; SCHLOBINSKI, Peter: *Einführung in die deskriptive Linguistik*, Westdeutscher Verlag, Opladen, 1994

FERRARI, Giacomo: „State of the art in Computational Linguistics“ in van Sterkenburg, Piet (Hrsg.): *Linguistics Today – Facing a Greater Challange*, John Benjamins Publishing Company, Amsterdam – Philadelphia, 2004, S. 163 - 186

GLÜCK, Helmut (ed.): *Metzler Sprachlexikon*, Metzler, Stuttgart, 2005

GOYVAERTS, Jan; LEVITHAN, Steven / DEMMIG, Thomas (trans.): *Reguläre Ausdrücke Kochbuch*, O'Reilly, Köln, 2009

GÜVENIR, Altay; OFLAZER, Kemal: *Using a corpus for teaching Turkish morphology*, Bilkent University, Ankara, 1995

LEMNITZER, Lothar; ZINSMEISTER Heike: *Korpuslinguistik – Eine Einführung*, Narr, Tübingen, 2010

MEYER, Charles F.: *English Corpus Linguistics – An Introduction*, Cambridge University Press, Cambrigde [u.a.], 2002

MULZER, Klaus: *Sprachverständnis und implizites Wissen*, Herbert Utz Verlag, München, 2007

REIMERS, Stephan; THIES, Gunnar: *PHP 5.3 und MySQL 5.5 – Das umfassende Handbuch*, Galileo Press, Bonn, 2010

SCHWARTZ, Baron; ZAITSEV, Peter et. al.; LICHTENBERG, Kathrin (trans.): *High Performance MySQL – Optimierung, Backups, Replikation und Lastverteilung*, O'Reilly, Köln, 2009

SCHWARTZ, Randal L.; PHOENIX, Tom; FOY, Brian D.; LANG, Jorgen W. (trans.): *Einführung in Perl*, O'Reilly, Köln, 2009

SCHWILL, Andreas (Hrsg.): *Schülerduden Informatik*, Bibliographisches Institut – Dudenverlag, Mannheim, 2003

STEUERWALD, Karl: *Deutsch-Türkisches Wörterbuch*, Harrassowitz, Wiesbaden 1974

STEUERWALD, Karl: *Türkisch-Deutsches Wörterbuch*, Harrassowitz, Wiesbaden 1974

TÜRK DIL KURUMU YAYINLARI (Hrsg.): *Türkiye'de Halk Ağzından Derleme Sözlüğü*, Ankara Üniversitesi Basımevi, Ankara, 1993

Internetquellen:

ALFAHOSTING GmbH: *Webspace*, https://alfahosting.de, 15.02.2012

ANTHONY, Laurence: *antconc*, http://www.antlab.sci.waseda.ac.jp/antconc_index.html, 08.03.2012

APACHE SOFTWARE FOUNDATION: *Apache Webserver*, http://httpd.apache.org, 15.02.2012

APACHE SOFTWARE FOUNDATION: *OpenOffice*, http://www.openoffice.org, 14.12.2011

BERLIN-BRANDENBURGISCHE AKADEMIE DER WISSENSCHAFTEN: *Digitales Wörterbuch der Deutschen Sprache*, http://www.dwds.de, 08.03.2012

CONRAD, Susan et al. (Hrsg.): *TESOL Quartely Vol. 37 No. 3*, Pantagraph Printing, Bloomington (Illinois), 2003, http://www.twu.ca/library/tqd_2008/VOL_37_3.pdf, 15.02.2012

DELISLE, Marc et al.: *phpMyAdmin*, http://www.phpmyadmin.net/home_page, 02.02.2012

GRINENKO, Victor M.: *DiskSpeed32*, http://software.vgrin.host56.com/diskspeed32, 07.03.2012

HO, Don: *Notepad++*, http://notepad-plus-plus.org, 02.02.2012

INSTITUT FÜR DEUTSCHE SPRACHE: *COSMAS II*, http://www.ids-mannheim.de/cosmas2, 08.03.2012

KLAß & IHLENFELD VERLAG GMBH: *Golem.de*, http://www.golem.de, 07.03.2012

LEHMANN, Christian: *Verwandtschaftsterminologie*, http://www.christianlehmann.eu/ling/lg_system/sem/verwandtschaftsterminologie.php, 17.02.2012

OLSON, Philip (Hrsg.): *PHP User Manual*, http://www.php.net/manual/de, 15.02.2012

ORACLE CORPORATION : *MySQL*, http://www.mysql.de, 16.02.2012

PERL FOUNDATION: *Perl*, http://www.perl.org, 15.02.2012

SCHLICHTHOLZ, Daniel: *MySQLDumper*, http://www.mysqldumper.de, 02.02.2012

SEIDLER, Kai: *XAMPP*, http://www.apachefriends.org/de/xampp.html, 15.02.2012

SELFHTML e. V.: *SELFHTML*, http://de.selfhtml.org, 15.02.2012

SIL INTERNATIONAL: *Ethnologue*, http://www.ethnologue.com/show_language.asp?code=tur, 14.02.2012

TODAİE (TÜRKIYE VE ORTA DOĞU AMME İDARESI ENSTITÜSÜ): *Yerelnet.org*, http://www.yerelnet.org.tr, 07.03.2012

TÜRK DIL KURUMU: *Türkiye Türkçesi Ağızları Sözlüğü*, http://www.tdk.gov.tr, 07.03.2012

UNICODE INC.: *Unicode*, http://www.unicode.org, 07.03.2012

UNIVERSITÄT LEIPZIG: *Wortschatz*, http://corpora.uni-leipzig.de, 08.03.2012

UNIVERSITY OF OXFORD: *British National Corpus*, http://www.natcorp.ox.ac.uk, 08.03.2012

WALL, Larry: *private Homepage von Larry Wall*, http://www.wall.org/~larry, 28.02.2012

WIDENIUS, Michael; AXMARK, David et al.: *MySQL 5.5 Reference Manual*, http://dev.mysql.com/doc/refman/5.5/en, 16.02.2012

WIKIMEDIA FOUNDATION INC.:

http://de.wikipedia.org/wiki/Afşar, 02.03.2012
http://de.wikipedia.org/wiki/Apache_HTTP_Server, 15.02.2012
http://de.wikipedia.org/wiki/ATA/ATAPI, 07.03.2012
http://de.wikipedia.org/wiki/Auszeichnungssprache, 08.03.2012
http://de.wikipedia.org/wiki/Backtracking, 15.02.2012
http://de.wikipedia.org/wiki/British_National_Corpus, 08.03.2012
http://de.wikipedia.org/wiki/DWDS, 08.03.2012
http://de.wikipedia.org/wiki/Festplatte, 07.03.2012
http://de.wikipedia.org/wiki/Hypertext, 08.03.2012
http://de.wikipedia.org/wiki/ISO_8859-9, 15.02.2012
http://de.wikipedia.org/wiki/Konkordanz, 08.03.2012
http://de.wikipedia.org/wiki/Limassol, 09.03.2012
http://de.wikipedia.org/wiki/MyISAM, 08.03.2012
http://de.wikipedia.org/wiki/Open_Source, 08.03.2012
http://de.wikipedia.org/wiki/Programmiersprache, 08.03.2012
http://de.wikipedia.org/wiki/Reguläre_Sprache, 15.02.2012
http://de.wikipedia.org/wiki/Regulärer_Ausdruck, 15.02.2012
http://de.wikipedia.org/wiki/Republik_Zypern, 14.02.2012
http://de.wikipedia.org/wiki/Seitenbeschreibungssprache, 08.03.2012
http://de.wikipedia.org/wiki/Texterkennung, 07.03.2012
http://de.wikipedia.org/wiki/Textkorpus, 16.02.2012
http://de.wikipedia.org/wiki/Türkische_Sprache, 14.02.2012
http://de.wikipedia.org/wiki/Unicode, 07.03.2012
http://de.wikipedia.org/wiki/Verwandtschaftsterminologie, 17.02.2012
http://en.wikipedia.org/wiki/Colorless_green_ideas_sleep_furiously, 08.03.2012
http://en.wikipedia.org/wiki/Instructions_per_second, 16.02.2012
http://en.wikipedia.org/wiki/Kinship_terminology, 17.02.2012
http://en.wikipedia.org/wiki/Moore's_law, am 07.03.2012
http://tr.wikipedia.org/wiki/Fakıekinciliği,_Pınarbaşı, 02.03.2012
http://tr.wikipedia.org/wiki/Kadınhanı, 07.03.2012
http://tr.wikipedia.org/wiki/Pazarören,Pınarbaşı, 02.03.2012
http://tr.wikipedia.org/wiki/Sille, 07.03.2012
http://tr.wikipedia.org/wiki/Tükiye#nin_illeri, 02.03.2012
http://tr.wikipedia.org/wiki/Türkçe, 14.02.2012

Eigenständigkeitserklärung

Hiermit versichere ich, dass die vorliegende Arbeit titels

„Entwicklung einer digitalen Korpusanwendung zur türkeitürkischen Dialektologie“

nie anderweitig als Prüfungsleistung verwendet und als Masterarbeit noch nicht veröffentlicht worden ist.

Ich versichere außerdem, dass ich die Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe. Diesen wörtlich oder sinngemäß entnommene Stellen habe ich als solche gekennzeichnet.

Frankfurt, 12.03.2012

Ort, Datum

Unterschrift

Anhang

Auf der beiliegenden CD-ROM ist der Quellcode der Webanwendung enthalten, sowie der Code zur Erstellung der MySQL-Datenbank.

Der Code für die vom Webserver ausgeführten/eingelesenen Dateien ist in der jeweiligen Datei unverändert vorhanden. Lediglich die Zugangsdaten zum Webserver wurden entfernt.
Die MySQL-Befehle zum Anlegen der Datenbank sind in einer gesonderten, nicht auf dem eigentlichen Webserver hinterlegten Datei namens *sql_setup.sql* im Ordner *SQL* zu finden. Sie kann mit einem Texteditor geöffnet werden.

Aus urheberrechtlichen Gründen liegt das digitalisierte Wörterbuch nicht bei. Somit kann durch den mitgelieferten Programmcode keine Kopie der Anwendung erstellt werden.
Den Zugang zur Onlineversion der Korpusanwendung können Sie durch Anfrage an master@manuelraaf.de beantragen.