universität wien

# DIPLOMARBEIT

Titel der Diplomarbeit

## Ein digitales Wörterbuch der 200 häufigsten Wörter der Wikipedia in ägyptischer Umgangssprache - Corpusbasierte Methoden zur lexikalischen Analyse nicht-standardisierter Sprache


Verfasser

Omar Siam


angestrebter akademischer Grad

Magister der Philosophie (Mag. phil.)

Wien, 2013

| | |
|---|---|
| Studienkennzahl lt. Studienblatt | A 385 |
| Studienrichtung lt. Studienblatt: | Diplomstudium Arabistik |
| Betreuer: | Dr. Karlheinz Mörth |

# Abstract

Natural Language Processing (NLP) for Arabic is still a challenging task. Only recently there have been some efforts to create tool sets for accomplishing more advanced tasks in the field of Arabic NLP. Even more challenging it seems to create such a tool set for Arabic colloquials, which are the mother tongue of every citizen of the Arab World. The present work tries to sketch the process of creating a corpus from a freely available source of Egyptian colloquial on the Internet, namely the Wikipedia Masry, to extract frequency information of the words therein. To achieve this end, the theoretical background of corpus creation is presented as well as tools, which were used to create texts, which can be processed using the computer to obtain first results in the field. Furthermore some tools are discussed which could be used to enhance the data and refine the analysis. After presenting current standards in the encoding of dictionaries the work is concluded by a dictionary part, which will be available in the encoding currently best suited for scholarly exchange of lexicographic data, namely the Guidelines of the Text Encoding Initiative.

## Zusammenfassung

Natural Language Processung (NLP) für Arabisch ist immer noch eine Herausforderung. Erst kürzlich gab es mehrere Versuche Softwarewerzeugsammlungen zusammenzustellen, um komplexere Aufgaben im Bereich NLP für Arabisch zu lösen. Eine noch größere Herausforderung scheint es darzustellen eine solche Werkzeugsammlung für arabische Umgangssprachen zu erstellen, die die Erstsprache aller Menschen in der arabischen Welt sind. Diese Arbeit versucht einen Prozess zu skizzieren, mit dem ein Corpus aus einer frei verfügbaren Ressource im Internet, der Wikipedia Masry, erstellt werden kann, aus dem Frequenzinformationen der darin enthaltenen Wörter extrahiert werden sollen. Um dieses Zeil zu erreichen, wird der theoretische Hintergrund von Corpuslinguistik dargestellt. Außerdem werden die Werkzeuge beschrieben, die benutzt werden, um die Texte zu extrahieren, die mittels Computer verarbeitet werden können, um erste Ergebnisse zu erhalten. Weiters werden einige Werkzeuge diskutiert, die verwendet werden können, um die Daten mit Informationen anzureichern und die Analysen zu verfeinern. Nachdem die aktuellen Standards für die Codierung vorgestellt werden, wird die Arbeit mit einem Wörterbuchteil abgeschlossen, die auch in dem heute in den Geisteswissenschaften am weitesten verbreiteten Standard zur Codierung lexikographischer Daten, den Richtlinien der Text Encoding Initiativem, vorliegen soll.

# Danksagung

An dieser Stelle möchte ich mich für die Unterstützung bei der Erstellung und der Korrektur dieser Arbeit durch meine Familie bedanken. Meine Geschwister Naschaat Siam und Dr$^{in}$. Monira Siam haben mir mit ihren Kommentaren und Korrekturen sehr geholfen. Auch meiner Mutter Mag$^{a}$. pharm. Maria Siam möchte ich für die vorgeschlagenen Korrekturen danken.

Auch meinem Betreuer Dr. Karlheinz Mörth möchte ich dafür danken, dass er die Entwürfe dieser Arbeit gelesen und mich mit wertvollen Anregungen unterstützt hat.

# Inhaltsverzeichnis

# Einleitung

Ein typisches Merkmal des arabischen Sprachraums ist die Diglossie. Dieses Beschreibungsparadigma definierte Ferguson 1959 in folgender Weise:

> DIGLOSSIA is a relatively stable language situation in which, in addition to the primary dialects of the language (which may include a standard or regional standards), there is a very divergent, highly codified (often grammatically more complex) superposed variety, the vehicle of a large and respected body of written literature, either of an earlier period or in another speech community, which is learned largely by formal education and is used for most written and formal spoken purposes but is not used by any sector of the community for ordinary conversation.[1]

Ferguson erarbeitete seine Theorie und seine Definitionen an Beispielen. Er benutzte das ägyptische Arabisch, Schweizerdeutsch und zwei weitere Sprachen um sein Konzept zu belegen. Ganz ähnlich den deutschsprachigen Schweizern spielt sich die Kommunikation im Alltag auch im arabischen Raum in einer anderen Sprache ab als diejenige, in der die schriftliche Kommunikation in herkömmlichen Medien abläuft, also etwa Bücher und Zeitungen. Daher bietet sich die Situation in der deutschsprachigen Schweiz zu einem gewissen Grad als Vergleichsmöglichkeit an. Ähnlich verhält es sich in Ägypten: Aufgrund dessen, dass das Hocharabische hauptsächlich in offiziellen Dokumenten und in schriftlichen Publikationen vorkommt, die gelesen, aber selten vorgelesen werden, tut man gut daran Hocharabisch zu verstehen. Jemand, der sich im arabischen Raum mit den Menschen unterhalten will, muss nicht nur die Schriftsprache erlernen, sondern auch zumindest einen Dialekt des gesprochenen Arabisch beherrschen. So sah es jedenfalls Ferguson.

Anfang der 1990er Jahre ging er in einem zweiten Text auf die Kritiken ein, die es zu seinem Konzept „Diglossia“ im Laufe der Zeit gegeben hat. Er stellte vor allem klar, dass ihm schon damals bewusst war, dass diese Definition nur die beiden Extreme festmacht, zwischen denen sich Sprache im Alltag in Diglossie Situationen wie der deutschsprachigen Schweiz oder Ägypten normalerweise abspielt. Er merkt auch an, dass in dieser Definition die Haltung, die Menschen zu den verschiedenen Sprachvarietäten, die sie gebrauchen, nicht besonders zur Geltung kommt. Eben die Thematik der Haltung gegenüber den Einsatzgebieten von Sprache ist auch für diese Arbeit relevant.[2]

Es ist schwierig eine passende Bezeichnung für die Sprache zu finden, um die es hier gehen soll. Es sind beinahe alle Termini in der Literatur kontroversiell diskutiert worden. In der Encyclopedia of Arabic Language and Linguistics, EALL, findet sich ein Artikel von Miller zur „Dialect Koine“, in dem die Autorin auch die zeitgenössischen dominanten Varietäten in den einzelnen arabischen Ländern, wie jene von Casablanca in Marokko oder eben die Kairoer Varietät in Ägypten, unter dem Begriff „Koine“ einordnet.[3] Man liest von ägyptischem Slang, von ägyptischem Arabisch als Dialekt oder als Akzent. Andere gebrauchten den Terminus Koine. Aus pragmatischen Gründen soll im Weiteren von der ägyptischen Umgangssprache gesprochen werden. Umgangssprache, engl. „vernacular“, ist hierbei als jene Sprache definiert, die die Menschen hauptsächlich aber nicht ausschließlich im direkten Umgang miteinander in einer bestimmten Region benutzen und die sich von der dort erwarteten Schriftsprache unterscheidet.

---

[1] Charles A. Ferguson, Anwar S. Dil: Language structure and language use. Essays by Charles A. Ferguson. Stanford, Calif., USA, 1971, S. 16.

[2] Charles A. Ferguson: Epilogue: Diglossia revisited. in: Alaa Elgibali, El-Said Festschrift Badawi (Hg.): *Understanding Arabic. Essays in contemporary Arabic linguistics in honor of El-Said Badawi,* Cairo, 1996, S. 49–67.

[3] Catherine Miller: Dialect Koine. in: Cornelis H. M. Versteegh (Hg.): *Encyclopedia of Arabic language and linguistics,* Leiden ( 1. A-Ed), S. 593–597, hier S. 595f.

Die schon aufgrund der Demografie meistgesprochene Varietät des Arabischen ist die ägyptische Umgangssprache, die man aufgrund der großen Deckungsgleichheit mit einem angesehenen der Hauptstadtdialekte auch als Kairenisch [4] bezeichnet.

Das Erlernen dieses Dialekts hat aufgrund der großen Bedeutung der ägyptischen Filmindustrie einen speziellen Wert[5], da dieser Dialekt auch heute noch im größten Teil der arabischen Welt verstanden und mehr oder weniger kompetent auch produziert wird. Spätestens seit den 1990er Jahren wird das Geld, das mit dem Erdölverkauf verdient wird, auch für die Finanzierung von Fernsehsendern verwendet. Die Dominanz der ägyptischen Medienindustrie ist seither rückläufig und es können über Satellit Musik, Talkshows und Serien etwa auf syrisch/libanesisch oder in Varietäten der Golf-Anrainer-Staaten empfangen werden[6], unter anderem syrisch/libanesisch synchronisierte erfolgreiche türkische Serien. Trotzdem wird das Kairenische noch oft eingesetzt und so kann man sich also in dieser Varietät mit besonders vielen Leuten in einem geografisch sehr ausgedehnten Raum unterhalten. Auch gibt es zahlreiche bekannte und gern gelesene ägyptische Autoren. In literarischen Texten kommt für die Erzählung in den meisten Fällen die Schriftsprache zum Einsatz. Bei Dialogen der Protagonisten in Romanen oder in Theaterstücken würde seit gut 100 Jahren reine Hochsprache nur dann noch zum Einsatz kommen, wenn deren Handlung in längst vergangener Zeit angesiedelt ist, etwa beim Roman al-Zaynī Barakāt von al-Ġitānī.[7] In einem Roman, der in der Zeit ab etwa 1800 spielt, wird mehr oder weniger offen der Dialekt in Dialogpassagen verwendet. Dies ist manchmal nicht offensichtlich, weil manche ältere Autoren bewusst eine Sprache gewählt haben, die man so oder so, also in Hochsprache oder im Dialekt, lesen kann, wobei der Sinn des Dialogs derselbe bleibt.[8] In jüngerer Vergangenheit scheint sich aber der Trend zur offenen Verwendung eines verschriftlichten Dialekts gefestigt zu haben.[9] So könnte man durchaus ein stattliches Corpus von schriftlichen Äußerungen im ägyptischen Dialekt zusammenstellen, das von Autoren für Dialogpassagen in Prosatexten aber vor allem in dramaturgischen Texten produziert wurde.

Neben gedruckt vorliegenden literarischen Werken gibt es auch ein frei verfügbares Corpus, dessen Inhalt nach dem Willen der Autoren in ägyptischem Arabisch sein soll: die Wikipedia in ägyptischem Arabisch (http://arz.wikipedia.org/). Ein großer Vorteil dieser Quelle ist, dass sie digital vorliegt, also nicht erst mittels optical character recognition (OCR) für die digitale Verarbeitung zugänglich gemacht werden muss, zumal diese für arabische Schrift längst nicht jene Erkennungsraten erreicht wie für Lateinschrift[10]. Ein zweiter Vorteil dieser digital vorliegenden Quelle liegt darin, dass sie strukturiert ist. Auch könnte man, was aber im Rahmen dieser Arbeit aus Zeitgründen nicht untersucht werden soll, diachrone Untersuchungen durchführen, da im Prinzip eine genaue Historie der Veränderung der Texte

[4] bewusst mit e nach Manfred Woidich. Siehe etwa seine kair**e**nische Grammatik. Manfred Woidich: Das Kairenisch-Arabische. Eine Grammatik. Wiesbaden, 2006.
[5] Josef Gugler (Hrsg.): *Film in the Middle East and North Africa. Creative dissidence.* Cairo, 2011, S. 4.
[6] Patricia Kubala: The Controversy over Satellite Music Television in Contemporary Egypt. in: Michael Aaron Frishkopf (Hg.): *Music and media in the Arab world,* Cairo, 2010, S. 173–224, hier S. 199.
[7] Sasson Somekh: Genre and language in modern Arabic literature. Wiesbaden, 1991, S. 34f.
[8] Somekh (1991), S. 24–29.
[9] Humphrey Davies: Dialect Literature. in: Cornelis H. M. Versteegh (Hg.): *Encyclopedia of Arabic language and linguistics,* Leiden ( 1. A-Ed), S. 597–604, hier S. 598–601.
[10] Volker Märgner, Haikal El Abed: Arabic Word and Text Recognition. Current Developments. http://www.elda.org/medar-conference/pdf/46.pdf (Zugriff am: 26. 01. 2012).. Sie merken an, dass es keine wissenschaftliche Forschung auf dem Gebiet der OCR für gedruckte Medien gibt und dass man auf die Angaben der kommerziellen Hersteller von OCR Software vertrauen muss. Die Erkennungsrate der Software Abbyy Finereader 11 ist dem Autor vertraut und sie ist nicht annähernd so hoch wie etwa bei Deutsch. Allerdings fehlt ein umfangreiches Wörterbuch für arabisch, was einen direkten Vergleich nicht erlaubt.

verfügbar ist. Beispielsweise wurde schon sehr bald nach dem Start der englischen Wikipedia eine Untersuchung zu den Veränderungen der Texte im Laufe der Zeit gemacht.[11]

Eine weitere Quelle für verschriftlichten ägyptischen Dialekt könnten Webforen und Social Media Plattformen im Internet sein. Es gibt dabei aber das Problem, ägyptisches Arabisch zu erkennen. Es fehlt ein computertaugliches Modell um ägyptisch Arabisch zu erkennen. Es gibt zwei Möglichkeiten die ägyptisch arabische Varietät zu notieren: in arabischer Schrift und in Lateinschrift. Letztere Möglichkeit wurde vor allem durch SMS gefördert und unterliegt den Einschränkungen dieses Dienstes. Auch wenn es heute allgemein möglich ist, arabische Schrift in SMS zu benutzen ist Lateinschrift hier sehr verbreitet. Es kann also weder aufgrund des verwendeten Schriftsystems noch aufgrund der verwendeten Wörter einfach auf die intendierte Varietät geschlossen werden, schließlich teilen sich Hocharabisch, Kairenisch und andere arabische Varietäten die Mehrzahl der Grapheme in arabischer Schrift. Man kann auch von der Herkunft der Benutzer oder der Websites nicht zuverlässig auf die dort gepflegte Varietät schließen. Theoretisch kann jeder Ägypter sowohl Hocharabisch als auch Kairenisch schreiben.

Der ägyptische Dialekt ist mittlerweile gut dokumentiert. Es existieren zahlreiche wissenschaftliche Untersuchungen und auch pädagogische Lehrwerke. Unter den pädagogischen Werken sind hier zu erwähnen Manfred Woidich (2002) und Manfred Woidich, Rabha Heinen-Nasr (2004), unter den wissenschaftlichen Untersuchungen vor allem Manfred Woidich (2006). Ein gewisser Grundwortschatz wird beim Erlernen der Grammatik vermittelt.

## Stand der Lexikographie im Hocharabischen und seinen Varietäten

Für das Hocharabisch gibt es kaum einsprachig arabische Wörterbücher, die in den letzten 200 Jahren erstellt wurden, die mehr bieten als ein Schulwörterbuch. Wörterbücher für die landesspezifische Umgangssprache werden überhaupt nicht für notwendig erachtet. Was es aber jetzt schon seit rund 100 Jahren in größerem Umfang gibt, sind bilinguale Wörterbücher englisch/französisch/deutsch – arabisch und seit den 1980er Jahren ein gutes englisch – kairenisch Wörterbuch[12].

An Wörterbüchern des ägyptischen Arabisch gibt es etwa Socrates Spiro (1895), zuletzt 1980 neu aufgelegt, Jacques Jomier (1976) und als bislang aktuellstes umfangreiches Werk Martin Hinds, as-Said Muhammad Badawi (1986). Weniger zu empfehlen wäre unter anderem Mohamed Abdel Aziz (2007a).

Heute gewinnen computerlesbare Wörterbücher immer mehr an Bedeutung. Das Nachschlagen von Wörtern ist effizient und der Zugang im Internet wird zum Standard. Neben Menschen können aber auch Programme diese nutzen, um darauf aufbauend Funktionen anzubieten. Dem Autor ist weder ein arabisches noch gar ein kairenisches elektronisches Wörterbuch bekannt, dass eine sinnvolle Durchsuchbarkeit hat und hilfreiche Ergebnisse liefert. Hierbei stellt einerseits der Wildwuchs an Transkriptionssystemen für Lateinschrift der verschiedenen westlichen arabischen Dialektologen ein Problem dar, und andererseits die eingebürgerte Eigenart der arabischen Produzenten von Dialekten die Vokalisierung und bei Dialekten noch zusätzlich die Deutung gewisser ambiguer Konsonanten dem Leser zu überlassen. Beides ist für Lernende eher keine Unterstützung. Auf einer soliden Datenbasis könnte man ein solches Wörterbuch mit den häufigsten Vokabeln erstellen und wenn entsprechend frei verfügbare Quellen genutzt werden, wäre es auch im Internet publizierbar.

[11] Fernanda B. Viégas, Martin Wattenberg, Kushal Dave: Studying Cooperation and Conflict between Authors with history flow Visualizations. http://alumni.media.mit.edu/~fviegas/papers/history_flow.pdf (Zugriff am: 11. 01. 2013).

[12] Martin Hinds, as-Said Muhammad Badawi: A dictionary of Egyptian Arabic. Arabic-English. Beirut, 1986.

## Ziel der Untersuchung

Auf der Grundlage der Wikipedia in ägyptischer Umgangssprache, der Wikipedia Masry, soll ein kurzes Frequenzwörterbuch erstellt werden. Dieses soll in einem digitalen Standardformat präsentiert werden, das auch eine weitere computerbasierte Nutzung erlaubt.

## Forschungsfragen

In der vorliegenden Arbeit geht es primär um zwei Dinge:

a) Den Workflow: Wie lassen sich die 200 häufigsten Wörter, die der Benutzer der Wikipedia in ägyptischem Arabisch vorfindet, ermitteln?
b) Eine Analyse der 200 häufigsten Wörter.

## Methode

Eine Version der ägyptischen Wikipedia wird computergestützt bearbeitet, die Wortformen werden gezählt, lemmatisiert und danach die zweihundert häufigsten Lemmata ermittelt. Diese Lemmata werden dann mit deutschen und englischen Übersetzungen versehen und in einer standardisierten digitalen Form aufbereitet.

## Über Corpuslinguistik

Linguistische Forschung findet zwischen zwei Extrempositionen statt: den Denker, der seinen Arbeitsplatz nie verlässt und durch Erforschung vor allem seines eigenen Sprachverständnisses Theorien über linguistische Phänomene entwickelt. Dies kann über Sprachgrenzen hinweg geschehen bis hin zu Universalgrammatiken, die versuchen, jede von Menschen gesprochene Sprache zu beschreiben. Am anderen Extrem ist der reine Beobachter angesiedelt. Ohne irgendeine Theorie geht er zu den Menschen und beobachtet deren Sprache, um dann anhand eben dieser Beobachtungen die Sprache zu beschreiben. Der vollkommene Empirismus steht also dem vollkommenen Rationalismus gegenüber.[13]

Nachdem in vielen Werken der Linguistik bis zur Mitte des 20. Jahrhunderts hauptsächlich Beobachtung und Beschreibung im Vordergrund standen, trat eine Figur der Linguistik auf, die Beobachtung und damit Corpora für entbehrlich hielt, Noam Chomsky. Er argumentierte, dass egal wie groß das Corpus, auf das sich Beobachtungen stützen, auch immer sei, es wäre unmöglich eine Sprache vollständig abzubilden. Er verfocht daher den Standpunkt, dass man Forschung über Sprache nur rational über Sprache nachdenkend betreiben könnte und dass Menschen einzig über deren Erstsprache wirklich zuverlässige Aussagen machen können.[14] Corpora, die als Grundlage für sprachwissenschaftliche Studien wie für Lexikographie dienten, hatten etwas gemeinsam, das die Fundamentalkritik Chomskys besonders leicht nachvollziehbar machte: Es waren undokumentierte Zusammenstellungen von Texten, deren Auswahlkriterien lediglich dem Zusammensteller bekannt waren oder im ungünstigsten Fall nicht einmal diesem bewusst waren. Der Beobachter wir immer nur die realisierte Sprache beschreiben können, die „performance“. Sprache besteht aber aus einer weit größeren Menge an Ausdrucksmöglichkeiten, der „competence“. Chomsky verwendet dafür später die Begriffe E-Language und I-Language, die er ein wenig anders definiert. „Performance“ entspricht also in etwa der „externalised language“ die man nachweisen und Aufzeichnen kann, „competence“ entspricht dem „internalised knowledge“ darüber, welche Möglichkeiten man hat, um gesprochene Sprache korrekt zu produzieren. An dieses Wissen kommt man selbst mit den heute verfügbaren Methoden nicht heran. Man kann es nur durch Introspektion befragen[15]. Mit statistischen Fachtermini ausgedrückt lautet die Herausforderung also: Die Grundgesamtheit, also die „[…] Objekte […], die einer statistischen Untersuchung zugrunde liegen“[16], ist nicht in „sachlicher, örtlicher und zeitlicher Form“[17] abgrenzbar.

Obwohl diese Erkenntnis die Linguistik für drei Jahrzehnte nachhaltig änderte und die früheren Ansätze als sehr einfach gestrickt und naiv enttarnte, war es Chomsky selbst, der unbeabsichtigt den Beweis lieferte, dass Rationalismus ohne Empirismus auch keine guten Ergebnisse liefert. Wenn es sich der Rationalist bequem macht und ein wenig nachdenkt, dann kann er mit einer Flut von Daten rechnen, mit denen er arbeiten kann. So kam es im Zuge der „Third Texas Conference on Problems of Linguistic Analysis in English” zu folgendem Dialog:

> Chomsky: The verb perform cannot be used with mass word objects: one can *perform a task* but one cannot *perform labour*.

[13] Lothar Lemnitzer, Heike Zinsmeister: Korpuslinguistik. Eine Einführung. 2. Auflage. Tübingen, 2010, S. 6f.
[14] Wie sich langer Aufenthalt außerhalb des Erstsprachgebiets auswirkt bleibt meist unerwähnt bei den Betrachtungen. Siehe etwa diverse computertechnische Übersetzungen, die Menschen deren Erstsprache durchaus die Zielsprache der Übersetzung ist unter dem Einfluss der Sprache ihres Arbeitsortes, meist der USA, liefern.
[15] Tony McEnery, Andrew Wilson: Corpus linguistics. An introduction. 2. Auflage. Edinburgh, 2001, S. 6f.
[16] Peter Pflaumer: Statistik für Wirtschafts- und Sozialwissenschaften. 3. Auflage. München, 2005, S. 13.
[17] Pflaumer (2005), S. 13.

> Hatcher: How do you know, if you don't use a corpus and have not studied the verb *perform*? Chomsky: How do I know? Because I am a native speaker of the English language.[18]

Das Problem mit dieser Aussage ist: Man kann im Britisch National Corpus nachweisen, dass es sehr wohl „mass word objects", also Kollektivnomen „mass nouns" wie „labour" oder „magic", gibt, die man mit „perform" verwenden kann.[19] Es findet sich folgender Satz:

> My latest favourites are sun-dried tomatoes in olive oil, pungent and sweet. Admittedly, they are expensive, but you only need one or two to transform a salad or perform magic when mixed with lemon, garlic and salami and stirred into pasta (see recipe for Fettucine with sun-dried tomatoes).[20]

Sieht man sich viel größere Corpora an, wie sie heute etwa durch das „Google Books" Projekt verfügbar sind, dann findet man mehr Nachweise, auch wenn diese anteilsmäßig immer noch sehr gering sind.

| DECADE | 1900 | 1910 | 1920 | 1930 | 1940 | 1950 | 1960 | 1970 | 1980 | 1990 | 2000 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| SIZE (MIL) | 2,119 | 904 | 1,305 | 1,125 | 813 | 1,463 | 1,818 | 1,799 | 2,139 | 2,914 | 5,407 |
| TOKENS | 8 | 5 | 8 | 22 | 9 | 14 | 9 | 20 | 33 | 56 | 140 |
| PER MIL | 0.00 | 0.01 | 0.01 | 0.02 | 0.01 | 0.01 | 0.00 | 0.01 | 0.02 | 0.02 | 0.03 |

Google Books American

| DECADE | 1900 | 1910 | 1920 | 1930 | 1940 | 1950 | 1960 | 1970 | 1980 | 1990 | 2000 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| SIZE (MIL) | 7,52 | 10,087 | 7,089 | 5,795 | 6,167 | 8,104 | 13,192 | 14,011 | 15,511 | 19,816 | 26,882 |
| TOKENS | 16 | 28 | 41 | 42 | 68 | 77 | 157 | 216 | 246 | 428 | 749 |
| PER MIL | 0.00 | 0.00 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.02 | 0.02 | 0.02 | 0.03 |

**Abbildung 1 "perform magic" im Google Books Corpus. Aus http://googlebooks.byu.edu/ (Zugriff 17. 12. 2012)**

Tatsache ist also, dass beide Ansätze in ihren Extremen versagen müssen. Weder kann ein Mensch alle gebräuchlichen Varianten seiner Erstsprache beschreiben, noch kann ein Corpus - so groß es auch seien mag - alle potenziell möglichen Wörter, Wortkombinationen oder Satzkonstruktionen enthalten. Wenn ein reiner Rationalist seine Theorien nicht an einem Corpus testet und Gegenbeweise nicht einarbeitet oder erklären kann, warum diejenigen falsch liegen, die für die in Corpora konservierten Aussagen verantwortlich sind, dann ist das wenig hilfreich. Wenn man etwas in Theorien ausschließt, die mit Hilfe von Corpora aufgestellt wurden, nur weil etwas in diesen Corpora nicht erfasst war, dann ist die Aussage ebenso nicht hilfreich[21]. Corpora sind generell nicht gut für negative Beweise zu gebrauchen. Viel mehr erlauben Corpora Aussagen gradueller Natur. Unter Berücksichtigung der Methoden, die angewandt wurden, um einen gewissen Grad der Ausgewogenheit des Corpus zu erreichen, können dann mehr oder weniger genaue Aussagen darüber extrapoliert werden, wie richtig oder falsch ein Wort, eine Wortkombination oder ein Satzkonstrukt angesehen wird. Untersuchungen, die unter Zuhilfenahme von

[18] Archibald A. Hill (Hrsg.): *Third Texas Conference on Problems of Linguistic Analysis in English (May 9-12, 1958).* Austin, Texas, USA, 1962.. Zitiert nach: McEnery, Wilson (2001), S. 11.
[19] McEnery, Wilson (2001).
[20] Aus: She magazine. London: The National Magazine Company Ltd, 1989. Enthalten im BNC, abgerufen unter http://corpus.byu.edu/bnc/ (Zugriff am 17. 12. 2012). Zugegebener maßen lässt es den Einwand gegen Chomsky in einem anderen Licht erscheinen wenn man bedenkt, dass dieser Satz Jahrzehnte nach dessen Aussage entstand und es ist auch das einzige Mal, dass diese Wortkombination im BNC auffindbar ist.
[21] Dylan Glynn: Corpus-Driven Cognitive Semantics: An introduction to the field. in: Dylan Glynn, Kerstin Fischer (Hg.): *Quantitative methods in cognitive semantics. Corpus-driven approaches,* Berlin2010 (= Cognitive linguistics research), S. 1–42

Corpora durchgeführt oder überprüft wurden, sind belastbarer als solche, die keine Corpora berücksichtigt haben, ebenso sind Theorien fundierter, wenn sie sich auch in der Praxis an Corpora bewährt haben.

Ob rationalistisch oder empirisch mittels Corpusabfragen, man wird nie zu einer 100% sicheren Bestätigung oder Widerlegung kommen. Schlüsse aus Corpusabfragen lassen sich auf das Corpus extrapolieren, solange dieser statisch ist. Nimmt man z. B. die Wikipedia Masry oder die deutsche „Wortwarte“[22] als nicht statisches Corpus, so werden Prognosen für sprachliche Entwicklung in der Zukunft zum Beispiel nicht sinnvoll sein. Andererseits sagen die Verhältnisse in der Wikipedia Masry nicht zwingend etwas über das Kairenische im Allgemeinen aus. Hohe Wahrscheinlichkeiten kann man aber erreichen, indem man beide Verfahren kombiniert. Während rationalistische Ansätze eine gewisse Anfälligkeit für grammatikalisch korrekte, aber konstruierte Beispiele, die man mit hoher Wahrscheinlichkeit nicht empirisch nachweisen kann, weil sie keinen Bezug zur Lebensrealität mehr haben, zeigen, kann man in Corpora manchmal Nachweise für Konstrukte finden, die der eigenen Ratio folgend nicht grammatikalisch sind. So können sich Prognose und Corpusevidenz ergänzen.[23]

In dieser Arbeit werden beide Ansätze zusammen genutzt. Die gewonnenen Rohdaten werden mit einer erlernten Grammatik interpretiert, um etwa Wortformen Lemmata zuzuordnen.

Um zu erreichen, dass Aussagen aus Corpusabfragen auch darüber hinaus mit großer Wahrscheinlichkeit anwendbar sind, muss ein Corpus „ausgewogen“ sein, sonst kann man nur von einer Sammlung elektronisch lesbarer Texte sprechen. Absolute Ausgewogenheit muss als unerreichbar gelten. Es ist aber machbar ein möglichst ausgewogenes Corpus zusammenzustellen bezogen auf einen bestimmten räumlichen, zeitlichen, lokalen oder Genre spezifischen Rahmen. Einen weitergehenden Nutzen kann man dann daraus ziehen, wenn genau dokumentiert ist, aus welchen Texten das Corpus zusammengestellt wurde. Diese Daten, die Metadaten eines Corpus, sind daher einer der wichtigsten Bestandteile eines Corpus.[24]

Für diese Metadaten haben sich festgelegte Kodierformen etabliert. Es handelt sich zwar nicht um eine Standardisierung im eigentlichen Sinn, aber es gibt Formate, die sich in einem mehr oder weniger breiten Gebiet von Sprachwissenschaften als de facto Standard etabliert haben. Aus dem Umfeld der Internettechnologien wurden etwa die Metadaten nach „Dublin Core“ übernommen. Ein andere aus der computerbasierten Sprachwissenschaft stammender de facto Standard ist der „Corpus Encoding Standard“ (CES), der in einem von der EU unterstützten Projekt entwickelt wurde und auf den in den Geisteswissenschaften recht beliebten TEI Richtlinien[25] aufbaut. Eine heute wohl oft genutzte XML[26] Variante ist als XCES bekannt.[27]

Eine mögliche Definition für ein Corpus lautet daher:

> Ein Korpus ist eine Sammlung schriftlicher oder gesprochener Äußerungen in einer oder mehreren Sprachen. Die Daten sind digitalisiert […]. Die Bestandteile des Korpus, die Texte oder Äußerungsfolgen, bestehen aus

[22] Siehe Seite 9.
[23] Lemnitzer, Zinsmeister (2010), S. 54–56.
[24] McEnery, Wilson (2001), S. 29–32.
[25] TEI Consortium (eds): P5: Guidelines for Electronic Text Encoding and Interchange. http://www.tei-c.org/Vault/P5/2.2.0/doc/tei-p5-doc/en/html/ (Zugriff am: 13. 01. 2013).
[26] Siehe Seite 57XML XML
[27] Lemnitzer, Zinsmeister (2010), S. 48f.

den Daten selber sowie möglicherweise aus Metadaten, die diese Daten beschreiben, und aus linguistischen Annotationen, die diesen Daten zugeordnet sind.[28]

Für die Lösung der beschriebenen methodologischen Probleme gibt es etwa folgende Ansätze, was das letztlich unlösbare Problem der Repräsentativität angeht:

- Man kann entweder seine Erkenntnisse ausschließlich auf das Corpus selbst beziehen. Meistens wird dies zwar in der Corpuslinguistik unerwünscht sein, manchmal aber kann man durchaus Aussagen finden, die, obwohl stark auf das konkrete Corpus konzentriert, weitergehende Hypothesen zulassen und weiter Untersuchungen anregen können, etwa wenn man sich die Sprache in Handbüchern international tätiger Konzerne wie IBM ansieht.[29]
- Oder man kann sich wie erwähnt um Ausgewogenheit bemühen. Dazu definiert man Kriterien, die dann auch dokumentiert werden sollen. Für bestimmte Untersuchungen werden Texte ganz bewusst nicht repräsentativ zusammengestellt. Untersucht man im Deutschen etwa die Möglichkeiten für befehlssatzähnliche subjektlose Konstruktionen, so bieten sich Corpora an, die nur aus Rezeptsammlungen bestehen. Sind solche Informationen in ausgewogeneren Corpora mittels Metadaten vermerkt, kann man diese daraus extrahieren. Auch kann es bei einer Untersuchung von seltenen Phänomenen nützlich sein, die Ausgewogenheit von Corpora zugunsten dieser Phänomene zu verschieben, sie werden überrepräsentiert. Corpora sind immer Stichproben und so kann eine Hypothese immer auch an anderen Stichproben überprüft werden. Es ist gut möglich, dass die Hypothese dadurch angezweifelt oder falsifiziert wird. So kann man linguistische Erkenntnisse aufgrund einer größeren und/oder anderen Materialbasis verfeinern.[30]

Moderne digitale Corpora bestehen oft aus Zeitungstexten, Chatprotokollen, SMS-Nachrichten, Tweets auf Twitter oder Texten aus Romanen. Älter Corpora bestehen meist aus jenen Drucksachen, zu denen die Ersteller Zugang hatten, also Zeitungen, Zeitschriften, Bücher und in eher seltenen Fällen auch transkribierte Gesprächssituationen. Bei Ausschnitten muss der Text sinnvollerweise mindestens so lang sein, dass die linguistischen Zusammenhänge erkennbar sind. Die ältesten digitalen Corpora sind für Englisch verfügbar. Dort wurde schon in den 1960er Jahren mit der Zusammenstellung von Corpora begonnen und die Möglichkeiten genutzt, die die damalige Computertechnologie bot. Im Vergleich zu heute waren das bescheidene Anfänge, aber für alle anderen Sprachen ergibt das einen Rückstand hinsichtlich computerlinguistischer Forschung von 20 Jahren oder mehr, der trotz des computertechnischen Fortschritts für Sprachen, bei denen es noch wenig computerlinguistische Tradition und andere Problemstellungen gibt, wie Arabisch, nur schwer aufholbar ist.[31]

Beim Aufbau eines Corpus empfiehlt es sich, folgende Punkte zu beachten:

- Die Daten liegen heute zum überwiegenden Teil schon digital vor, für ältere Texte oder gesprochene Sprache ist aber oft OCR oder manuelle Eingabe erforderlich oder die Entwicklung von Spezialwerkzeugen, was sehr zeit- und ressourcenintensiv ist.
- Gesammelte Daten unterliegen umfangreichen rechtlichen Einschränkungen. Sei es das Urheberrecht, der Datenschutz oder Persönlichkeitsrechte. Das größte Problem dabei ist oft die Möglichkeit der Weitergabe der Daten zu vereinbaren. Die Verfügbarkeit der Daten spielt für die

[28] Lemnitzer, Zinsmeister (2010), S. 40.
[29] McEnery, Wilson (2001), S. 165–185.
[30] McEnery, Wilson (2001), S. 50–54.
[31] Lemnitzer, Zinsmeister (2010), S. 42f.

wissenschaftliche Nachvollziehbarkeit von wissenschaftlichen Arbeiten eine wichtige Rolle und ist für aufbauende Studien sehr wünschenswert. Abgesehen vom Urheberrecht, gibt es manche Vorgehensweisen, die Wissenschaftler vor einigen Jahrzehnten wählten und denen wir wertvolles Material verdanken, die heute rechtlich nicht mehr nachvollziehbar sind.

- Für die Zeichencodierung im computertechnischen Sinn bietet sich heute eigentlich nur noch Unicode in seiner Variante UTF-8 an. Ältere Daten, die anders codiert sind, müssen erst umgewandelt werden, was aber meist kein schwieriges Unterfangen ist. Schwierig ist die Situation bei Zeichen, die im Unicodestandard keine Entsprechung haben. Mit diesem Problem sind spezielle Forschungsgebiete konfrontiert, wie Dialektologen, wenn es für Phänomene einer Varietät keine Ausdrucksmöglichkeiten in der Schrift der anerkannten Hochsprache gibt.
- Die Metadaten müssen zusammengestellt und beigefügt werden.
- Die Daten müssen je nach Forschungszweck annotiert, also um spezifische digitale Zusatzinformationen, angereichert werden. Dabei muss über die passende Codierung ebenso entschieden werden, wie über Techniken mit denen die Annotationen hinzugefügt aber auch wieder entfernt werden können, sodass der ursprüngliche Text wieder sichtbar wird. Eine Möglichkeit dies zu erreichen, bietet etwa eine Codierung, die auf XML[32] basiert.[33]

## Corpora in der Lexikographie

In der seriösen Lexikographie wird heute größtenteils auf der Basis digitaler Corpora gearbeitet. Alle großen Wörterbuchverlage (Duden, Longman, Larousse etc.) haben mittlerweile digitale Textsammlungen.[34] Schöne Beispiele für öffentlich zugängliche Corpora sind für das Deutsche etwa die Corpora des IDS (Institut für deutsche Sprache), allen voran das DeReKo, das „deutsche Referenzkorpus“[35] und jene des DWDS (Digitales Wörterbuch der Deutschen Sprache)[36] oder auch die „Wortwarte“, die automatisch deutsche Tages- und Wochenzeitungen auswertet und damit manchmal interessante Beiträge zur Lexikographie liefern kann.[37] Diese werden dann in die Wörterbücher eingepflegt wie etwa auch die Belegsammlungen aus Zeitungen für das arabische Wörterbuch für die Schriftsprache der Gegenwart von Hans Wehr und Lorenz Kropfitsch.[38] Oft werden bei Untersuchungen auch Belege aus großen Corpora extrahiert. Für das Englische ist etwa das Collins COBUILD English Language Dictionary als corpusbasiert zu nennen. Hier wurde auch nachgewiesen, dass es möglich ist, bestimmte Angaben in Wörterbüchern zu hinterfragen, wenn man diese gegen ein anderes und/oder größeres Corpus testet.[39]

Corpora werden schon länger etwa bei der Planung von Wörterbüchern eingesetzt.[40] Es ist oft nötig, angepasste Werkzeuge zur Extraktion, der für die Lexikographie relevanten Informationen zu verwenden. So können Corpora dann wichtige Hinweise für die Lemmaauswahl liefern. Für den korrekten Gebrauch von Worten und auch für leicht unterschiedliche Sinnvarianten sind Beispiele für das Umfeld, in dem sie vorkommen, sehr hilfreich. Hier können Corpora bessere und realistischere

[32] Siehe Seite 57
[33] Lemnitzer, Zinsmeister (2010), S. 57f.
[34] Lemnitzer, Zinsmeister (2010), S. 42.
[35] http://www.ids-mannheim.de/kl/projekte/korpora/ (Zugriff am 26. 01. 2013)
[36] http://www.dwds.de/ (Zugriff am 26. 01. 2013)
[37] Lemnitzer, Zinsmeister (2010), S. 148.
[38] Hans Wehr: Arabisches Wörterbuch für die Schriftsprache der Gegenwart. Arabisch - deutsch. 5. Auflage. Wiesbaden, 1985, S. vii und xv.
[39] McEnery, Wilson (2001), S. 107.
[40] Stefan Engelberg, Lothar Lemnitzer: Lexikographie und Wörterbuchbenutzung. 4. Auflage. Tübingen, 2009, S. 238–243.

Beispiele liefern als die Lexikographen selbst. Verschiedene Lesarten und Bedeutungsunterschiede können anhand von Corpusdaten so angeordnet werden, wie es ihrer Häufigkeit im Corpus und damit wahrscheinlich auch ihrer tatsächlichen Verwendungshäufigkeit entspricht. Weiters können in Corpora mit statistischen Verfahren Kollokationen aufgespürt werden, die auf bestimmte eingeschränkte Verwendungsmöglichkeiten oder idiomatischen Gebrauch hinweisen, wie im Deutschen etwa bei dem Substantiv *Hehl* und dem Verb *fackeln* die Verwendungsmöglichkeiten sehr beschränkt sind. Es könne bei Korrekturentscheidungen für Wörterbücher auch Corpora als Beweisquelle herangezogen werden. Eines der wichtigsten Werkzeuge in diesem Zusammenhang sind Programme, die aus Corpora Konkordanzen erstellen. Konkordanz bedeutet hier, dass ein Wort mit Kontext, mit allen Wörtern, die in einem bestimmten Abstand, meist 10 oder weniger, vor und nach diesem stehen, ausgegeben wird. Die Werkzeuge heißen oft Concordancer und sie erzeugen sogenannte KWIC (Key-word-in-context) Listen.[41]

Besonders gut geeignet für die Untersuchung an Corpora sind häufig vorkommende Wortkombinationen innerhalb von Sätzen. Man spricht hier von Kookkurrenzen (coocurences) und Kollokationen (collocations). Lothar Lemnitzer, Heike Zinsmeister (2010) definieren:

- Als Kookkurrenz soll das gemeinsame Vorkommen zweier Wörter in einem gemeinsamen Kotext betrachtet werden. Die Länge des betrachteten Kotextes kann als Textfenster bestimmter Länge festgelegt werden. Im Allgemeinen wird ein einzelner Beleg abstrahiert und das gemeinsame Vorkommen zweier Wörter in vielen Kotexten betrachtet werden. Es kann zudem die Reihenfolge des Auftretens beider Wörter in den Belegen als Unterscheidungskriterium zweier Kookkurrenzen festgelegt werden. Ferner kann festgelegt werden, dass die Wörter einer Kookkurrenz häufiger […] miteinander vorkommen, als dies der Fall wäre, wenn diese zufällig verteilt wären. Man spricht in diesem Fall von einem signifikanten Kovorkommen […].
- Eine Kollokation muss […] darüber hinaus aber auch eine innere Struktur, in Form einer Hierarchie zwischen Kollokationsbasis und Kollokator aufweisen. […] Die Glieder einer Kollokation [müssen] in einer syntaktischen Beziehung zueinander stehen […].[42]

Große Corpora machen das Auffinden solcher Phänomene über statistische Methoden möglich, da sich computerbasiert jene Kookkurrenzen aufspüren lassen, die eben signifikant über einem zufälligen Auftreten liegen. Die Ergebnisse bekommen durch die Größe des Corpus eine höhere Relevanz. Schließlich bieten sich Corpora noch für Untersuchungen zum Gebrauch und zum Aufbau von festen Redewendungen, Phrasemen, an. Weiters lässt sich das Einfließen von Neologismen anhand von Corpora oft gut untersuchen. Einerseits hinsichtlich des Auftretens, wenn das Corpus entsprechend diachron strukturiert ist und die Metadaten benutzt werden, andererseits kann der Grad der Integration in die Sprache abgeschätzt und entsprechender grammatikalischer Gebrauch nachgewiesen werden. Ein Spezialfall von Neologismen sind heute Anglizismen, die sich in vielen Sprachen auf technischem Gebiet - aber nicht nur dort - in die Sprachen integrieren.[43]

[41] Lemnitzer, Zinsmeister (2010), S. 139–141.
[42] Lemnitzer, Zinsmeister (2010), S. 143f.
[43] Lemnitzer, Zinsmeister (2010), S. 142–152.

## Corpora und statistische Erkenntnisse[44]

Einer der ersten Philologen, der sich mit den systematischen Zusammenhängen zwischen der Häufigkeit eines Wortes und seinem Bedeutungsinhalt beschäftigten war George Kingsley Zipf. Er lebte Anfang des 20. Jahrhunderts und absolvierte ursprünglich ein philologisches Studium. Dies erklärt teilweise zwei Kritikpunkte an Zipfs Werk: Er wusste nicht viel über andere Wissenschaftler, Mathematiker und Ökonomen etwa, die dieselben Zusammenhänge beschrieben hatten, die auch ihm auffielen. Zweitens beruhten seine Schlussfolgerungen auf seinem noch sehr mechanistisch geprägten Weltbild und stellten sich später als kaum haltbar heraus.[45] Bei einem Studienaufenthalt in Deutschland begann er auf Anregungen seiner Professoren, die Philologie von einer naturwissenschaftlichen Seite aus zu betrachten. 1930 bekam er den Ph. D. in vergleichender Philologie verliehen. Er war danach unter anderem als Deutschlehrer in Harvard tätig. Er hatte mit seinen Vokabellisten der häufigsten Wörter dabei unter den gegebenen Bedingungen des Fremdsprachenerwerbs außerhalb des deutschen Sprachgebiets mehr Erfolg dabei seinen Studenten Deutsch beizubringen als andere Sprachlehrer. Er veröffentlichte nur eine literaturwissenschaftliche Arbeit, aber seine Arbeiten zu den „Gesetzen" hinter der Sprache waren interessant und inspirierend. Seine Dissertation aus dem Jahr 1932 trug den Titel „The Psycho-Biology of Language" und erschien 1968 als Buch, 18 Jahre nach Zipfs Tod. Besonders die quantitative Linguistik in der ehemaligen Sowjetunion beschäftigte sich in den folgenden Jahrzehnten mit seinem Werk.[46]

In seiner frühen akademischen Laufbahn gelingt es ihm mathematische Beschreibungen für sprachliche Beobachtungen zu finden und er versucht wissenschaftlich zu begründen, warum diese mathematischen Formeln die empirischen Daten begründen. So führt er etwa 1929 den Nachweis, dass die Häufigkeit eines Wortes mit „den Ausprägungen der verschiedenen Eigenschaften von Lauten, Silben und Wörtern" in Zusammenhang steht. Mit anderen Worten: Je häufiger ein Wort, eine Silbe oder ein Laut, desto weniger wird er als herausragend und wichtig im Gesamtgefüge der Sprache wahrgenommen. Damit ist eine Verschleifung, ein undeutlich Werden zu beobachten je häufiger ein Wort, eine Silbe oder ein Laut eingesetzt wird. Die Aussprache wird vereinfacht. Er erkennt zwei gegensätzliche Kräfte, welche die Sprache beeinflussen: Unterscheidbarkeit von sprachlich bedeutenden Einheiten für den Hörer und die Faulheit oder der Ökonomie des Sprechers beim Artikulieren. Extrem häufige Wörter müssen nicht genau verstanden werden, weshalb sie dann mit der Zeit nicht mehr deutlich artikuliert werden. Dies wurde gegen Ende des 20. Jahrhunderts als „Minimierung des Dekodierungsaufwands" und „Minimierung des Produktionsaufwandes" definiert.[47]

Kritik an Zipfs Ausführungen gab es damals vor allem in Bezug darauf, was deutlich und ausgeprägt oder schwach heißen sollte. Er versuchte schon, den Zusammenhang mit einer mathematischen Funktion zu beschreiben. Man kann diese Beobachtungen diachron im Zusammenhang mit Lautwandelprozessen sehen. Laute, die schwierig artikulierbar sind, aber keine spezielle Bedeutungsunterscheidung mehr haben werden verschliffen. Dadurch entstehen neue oder andere Laute, deren Häufigkeit dann wieder in den weiteren Wandlungsprozess Einzug hält. Zipf denkt etwa über Schwellwerte bei der Häufigkeit sprachlicher Phänomene nach, die einen Lautwandel hervorrufen. Zipf sieht die Betonung als von der Häufigkeit abhängig an, da sie energieintensiv ist. Er versucht das mit der Betonung von trennbaren

[44] Der folgende Abschnitt bezieht sich in wesentlichen Teilen auf Claudia Prün: Das Werk von G. K. Zipf. in: Reinhard Köhler (Hg.): *Quantitative Linguistik. Ein internationales Handbuch,* Berlin, 2005 (= Handbücher zur Sprach- und Kommunikationswissenschaft), S. 142–152.
[45] Anatol Rapoport: Zipf's Law Re-visited. in: Henry Guiter (Hg.): *Studies on Zipf's law,* Bochum, 1982 (= Quantitative linguistics).
[46] Prün (2005).
[47] Reinhard Köhler: Zur linguistischen Synergetik. Struktur und Dynamik der Lexik. Bochum, 1986, S. 50f.

Vorsilben im Deutschen zu belegen. Auch die Wortlänge steht in einem indirekten Zusammenhang mit der Häufigkeit. Ist ein Wort häufig aber lang, kommt es zu abgekürzter Aussprache. „Häufige, stärker erwartete Elemente“ tragen weniger Information und sind daher oft zu erraten. Eine genaue Artikulation wird unnötig. Er sieht einen Zusammenhang zwischen der Häufigkeit von Worten und der Anzahl der Worte die eine bestimmte Häufigkeit haben. Je häufiger ein Wort, desto weniger Wörter mit einer ähnlichen Häufigkeit wird man finden. In seiner Dissertation 1935 erweitert er diese Erkenntnis zu dem Gesetz, für das er berühmt wurde: das Rang-Frequenz-Gesetz, Zipf's Law. In einer nach Häufigkeit geordneten Liste wird ein Wort jene Position, jenen Rang, einnehmen, der in einem inversen Verhältnis zu seiner Frequenz steht. Das häufigste Wort hat den Rang 1, das zweithäufigste den Rang 2. Trägt man in einem Diagramm die Frequenz über dem Rang auf entsteht annähernd eine Hyperbel, nimmt man einen doppelt logarithmischen Maßstab ergibt sich eine Linie. Aber auch das Alter eines Wortes hat einen Zusammenhang mit seiner Häufigkeit, vermutete Zipf, denn obwohl häufige Wörter mit der Zeit immer mehr verschliffen werden als seltene, sind deren ursprüngliche Formen häufig für eine viel frühere Zeit nachweisbar.[48] Diese häufigen Wörter lassen sich nicht so einfach aus dem Wörterbuch verdrängen. Das wurde Ende des 20. Jahrhunderts auch bestätigt. Die unterschiedlichen Ebenen der menschlichen Sprache unterliegen auch einer Selbstregulation, deren Zusammenhang aus Zipfs Erkenntnissen erklärt werden kann. So wird es in einem Text etwa das Inventar an Wörtern größer sein als das an Silben und jenes größer als das an Morphemen. Das wirkt sich bei den Kurven aus, die man nach Zipfs Theorien erstellen kann: Sie werden flacher. Es gibt ihm Zufolge auch eine Beziehung zwischen den Bausteinen von Spracheinheiten und den Eigenschaften der übergeordneten Spracheinheiten, an denen sie beteiligt sind. Später findet er sein Rang-Häufigkeitsgesetz auch auf Gebieten der Soziologie und der Ökonomie bestätigt.[49] Seine Schlussfolgerungen darüber, warum diese Kurven erstellt werden können, sind allerdings unhaltbar, was schon 1965 klar war, als das Vorwort zu einer Neuauflage von Zipfs Werk erschien[50]. Es ist daher nicht abschließend geklärt, worauf alles seine (wieder-)entdeckten Gesetze anwendbar sind und worauf nicht und welche Prognosen man treffen könnte.[51]

## Frequenzwörterbücher[52]

Frequenzwörterbücher sind insofern etwas Besonderes, als dass sie Auskunft darüber geben, wie oft ein darin enthaltenes Wort in der betrachteten Textmenge vorkommt. Die Einträge sind entweder absteigend nach Häufigkeit sortiert oder traditionell alphabetisch. Jedenfalls ist bei jedem Eintrag ersichtlich, welche relative oder absolute Häufigkeit einem Wort zugeordnet werden kann. Oft sind auch beide Listen in einem Buch zu finden. Beinahe alle Frequenzwörterbücher sind nicht vollständig. Eine Auflistung aller Wörter bis hin zu den nur einzeln vorkommenden, den Hapax Legomena, wäre früher für den Druck einfach nicht durchführbar gewesen und ist auch heute relativ aufwendig. Trotzdem sollten in einem solchen Wörterbuch neben den abgedruckten Wörtern auch noch die restlichen Frequenzen und eine Angabe wie viele Wörter es mit diesen Frequenzen gibt abgedruckt sein. Da alle Frequenzwörterbücher auf Texten oder einem Corpus basieren und da sich, wie oben erwähnt, bei der Auswahl dieser Texte, auch wenn man um Ausgewogenheit bemüht ist, niemals das perfekt

[48] Vgl. Köhler (1986), S. 66–70.
[49] Prün (2005).
[50] George Kingsley Zipf: The psycho-biology of language. An introduction to dynamic philology. Cambridge, Mass., USA, 1965, S. v–x.
[51] Rapoport (1982).
[52] Der folgende Abschnitt bezieht sich in wesentlichen Teilen auf Pavel M. Alekseev: Frequency dictionaries. in: Reinhard Köhler (Hg.): *Quantitative Linguistik. Ein internationales Handbuch,* Berlin, 2005 (= Handbücher zur Sprach- und Kommunikationswissenschaft), S. 312–324.

ausgewogene Corpus für jeden Zweck ergeben wird, sind Frequenzwörterbücher, was Vorhersagen, die mit ihnen über einen zufällig ausgewählten Text gemacht werden können, mit Vorsicht zu genießen. Es ist daher bei einem Frequenzwörterbuch wichtig herauszufinden, welche Charakteristiken es hat. Dazu zählen:

> language; sublanguage, style, idiolect; input unit; the volume of the sampling or of the text used for compiling the dictionary; the number of different units found in the sampling; their number published in the dictionary; the form of indicating frequencies and sufficiency of information about them; the structure of the dictionary and of the entry; compiling techniques; the main aim and the addressee of the dictionary.[53]

Was die Namen angeht, so findet man einerseits wie zu erwarten „Häufigkeitswörterbuch, Rangwörterbuch, frequency dictionary, frequency word book, word count oder dictionnaire de fréquence. Manchmal sind auch Wörterbücher, die elementary, basic oder fundamental im Namen tragen Frequenzwörterbücher, bei denen man herausfinden kann, nach welcher Methode sie zusammengestellt wurden.

Die in heute vorhandenen Frequenzwörterbüchern enthaltenen Sprachen decken ein sehr breites Feld ab, von Altgriechisch und Akkadisch über Arabisch[54], Kirgisisch und Slowakisch bis Jiddisch. Die häufigsten Sprachen, zu denen man Frequenzwörterbücher finden kann, sind Englisch, Russisch, Deutsch, Französisch, Latein, Polnisch und Japanisch. Auch Konkordanzen, Indizes und Wörterbüchern oder -listen zu Texten von diversen Autoren kann man dazu zählen, wenn sie Frequenzangaben enthalten. Hier sind dann sehr viele Werke dabei, die zum besseren Verständnis lateinischer Texte angefertigt wurden.

Was Varietät, Stil und Idiolekt angeht, so sind sie gute Möglichkeiten, um die Menge an Text genau zu definieren, die man für ein Frequenzwörterbuch verwenden will. Die Aussagekraft in Bezug auf allgemeine Sprachproduktion ist dann natürlich sehr beschränkt. Frequenzwörterbücher werden für fiktionale, wissenschaftliche oder technische Teilgebiete der Sprache erstellt, für Prosa und Poesie, für ein bestimmtes literarisches Genre, für einen Autor, für einen Überblick über alle wissenschaftlichen oder technischen Texte oder für bestimmte technische oder wissenschaftliche Teilgebiete. Zeitungsberichte und andere Nachrichtenquellen werden gerne in als allgemein verstandene Corpora eingebaut oder einzeln behandelt (für arabische Mediensprache vergleiche etwa Orhan Elmaz (2010)). Bei historischen Texten sind Frequenzwörterbücher teilweise auch den heiligen Schriften gewidmet, etwa der Bibel und auch dem Koran.

In einem Wörterbuch mit alphabetischer Zugriffsstruktur[55], jene Wörterbücher, die man häufig in der Schule kennenlernt, ist traditionell relativ klar welche Formen (Alekseev bezeichnet sie als input units) dort als Einträge vorkommen und welche nicht. Was als Lemma definiert ist und wie genau die Anordnung nach dem Alphabet erfolgt, ist von Wörterbuch zu Wörterbuch unterschiedlich und folgt in verschiedenen Sprachräumen unterschiedlichen Traditionen. Besonders bei phraseologischen Wörterbüchern sind die Kriterien der Auswahl der Phraseme und deren Anordnung nicht einheitlich. Es lassen sich ganz unterschiedliche linguistische Phänomene im Hinblick auf ihre Frequenz analysieren und in wörterbuchartigen Repräsentationen darstellen. Von möglichen Wortformen über Morpheme,

[53] Alekseev (2005).
[54] Tim Buckwalter, Dilworth B. Parkinson: A frequency dictionary of Arabic. Core vocabulary for learners. London, 2011.
[55] Herbert Ernst Wiegand: Zugriffsstrukturen in Printwörterbüchern: Ein zusammenfassender Beitrag zu einem zentralen Ausschnitt einer Theorie der Wörterbuchform, in: *Lexicographica* 24 (2008), hier S. 214.

Grapheme bis hin zu Lauten ist alles möglich. Auch Graphemkombinationen, Silben, Flexionen und syntaktische Konstruktionen oder Sätze sind mögliche Kandidaten für einen Eintrag. Auch zu Anthroponymen, also Eigennamen, wurden Frequenzwörterbücher herausgegeben. Relativ viele Frequenzwörterbücher beschäftigen sich mit den Termini technicii auf verschiedensten Gebieten.

Ein wichtiges Merkmal um die Aussagekraft einer frequenzbasierten lexikographischen Darstellung auf ihrem Bestimmungsgebiet abschätzen zu können, ist die Menge an Text, auf dessen Basis die Frequenzzählung erfolgte. Bevor Computer und entsprechende Software allgemein verfügbar waren, konnten sich nur wenige Frequenzwörterbücher auf wesentlich mehr als eine Million Wörter oder andere sprachliche Einheiten stützen, eher waren die Texte, die als Grundlage dienten in der Größenordnung von einigen 1000 bis 10000 Einheiten. Die Rolle des Zusammenstellers der Wortfrequenzen kann heute der Computer übernehmen, wenn die Randbedingungen zum Auffinden der sprachlichen Einheiten einmal klar und für Maschinen umsetzbar sind, was bei Hochsprachen und den gebräuchlicheren sprachlichen Einheiten oft schon der Fall ist. So können heute Frequenzwörterbücher über einige Millionen Einheiten in einigen Minuten erstellt werden und auch eine Abfrage von Corpora, die im Internet abrufbar oder anders elektronisch zugänglich sind, ist verglichen mit den Bedingungen, die bis in die 1980er Jahre vorherrschten, viel einfacher.

Elektronisch veröffentlicht ist es zwar heute möglich alle Frequenzen bis hinunter zu den einzeln vorhandenen Wörtern darzustellen und für Englisch ist zum Beispiel eine Liste mit 100.000 Häufigkeitsrängen verfügbar, die auf renommierten englischen Corpora basiert. [56] Google stellt die Datengrundlage seines Google Books Dienstes online zur Verfügung. Dort sind Wortformen und deren Kombinationen (n-grams) bis zu einer Mindestfrequenz von 40 aufgelistet und das in den Sprachen Englisch, Spanisch, Französisch, Deutsch, Russisch, Italienisch, Chinesisch und Hebräisch. Kombinationen sind bis zu fünf Worte lang. Es ist also ein minimaler Kontext verfügbar. Die Daten sind in viele Dateien aufgeteilt, können aber wieder zusammengeführt werden.[57] Auch in Frequenzwörterbüchern sind allerdings meist Lemmata angegeben und nicht flektierte Wortformen. Wenn es für eine Sprache Werkzeuge gibt, die mit hinreichender Zuverlässigkeit Wortformen innerhalb eines derart beschränkten Kontexts disambiguieren, Homonyme auflösen und entsprechende Wortformen ihren Lemmata zuordnen können, dann kann man damit ganz ähnliche Untersuchungen durchführen wie an den aufwendiger erstellten Standardcorpora. Solche Werkzeuge sind vor allem für morphologisch relativ einfache Sprachen wie Englisch gut vorstellbar und dementsprechend sind Abfragemasken, die eine Suche nach allen Formen eines Lemmas erlauben, vorhanden.[58] Eine Zuverlässigkeit von nahezu 100 % ist bei einem derart beschränkten Kontext allerdings kaum zu erwarten. Für morphologisch komplexe Sprachen wie Arabisch und seine Varietäten, bei denen es zu mehr Homonymen kommen kann, wird der sehr eingeschränkte Kontext wahrscheinlich zu großen Fehlerraten führen. Ein Vergleich mit Untersuchungen und Lösungen für Hebräisch bietet sich an. Abgesehen davon bleibt das Problem eine Gruppe von fünf Worten der richtigen Varietät zuzuordnen. Dies wird wohl in der Mehrzahl der Fälle undurchführbar sein, da ja der größte Teil der Grapheme unvokalisiert bleibt und damit in den meisten arabischen Varietäten ident aussieht, aber unterschiedlich ausgesprochen wird. Nur an vergleichsweise wenigen eindeutig einer Varietät zuzuordnenden Worten lässt sich auch ohne Vokalisierung deren Zugehörigkeit ablesen.

[56] http://www.wordfrequency.info/100k_purchase.asp (Zugriff am: 11. 1. 2013) als Tabelle für Tabellenkalkulationen für $ 125 bzw. $ 250 für kommerzielle Anwendungen,
[57] http://storage.googleapis.com/books/ngrams/books/datasetsv2.html (Zugriff am 11. 1. 2013). Zusammengeführt wurden die Daten etwa hier: http://googlebooks.byu.edu/
[58] Etwa http://googlebooks.byu.edu/ (Zugriff am 27. 01. 2013).

Soll das Wörterbuch gedruckt vorliegen, dann wird man einen wesentlich geringeren Umfang festlegen und genaue Einträge für alle weiteren Frequenzen weglassen müssen. Es gibt nur ganz wenige Frequenzwörterbücher, die vollständig bis zu den einzelnen Wörtern gedruckt wurden. Wie sich nach Zipf auch nicht anders erwarten lässt, gibt es bei der Erstellung von jedem Frequenzwörterbuch eine sehr große Anzahl von Wörtern mit niedriger Frequenz. Die meisten Nutzer eines solchen Wörterbuches werden aber eher an den häufigsten Wörtern interessiert sein. So liegt es dann am Ersteller, den Umfang festzulegen und den Rest nur zusammenzufassen. Es mag auch schwierig sein ein Frequenzwörterbuch am Massenmarkt zu positionieren, da der Nutzen eines solchen Wörterbuchs nicht allgemein bekannt ist. Eine Ausnahme sind wohl auf diese Weise erstellte Basis- oder Elementarwörterbücher und ähnliche, wobei sich diese einem weiteren Nutzen verschließen, wenn sie keine oder zu vage Angaben über ihre Grundlagen enthalten.[59] Eine andere grundsätzliche Problematik, die sich auch aus Zipfs Beobachtungen ableiten lässt, ist, dass der Anstieg der Anzahl der gefundenen sprachlichen Einheiten mit steigender Textgröße des Ausgangscorpus immer flacher wird, man also um eine signifikante Anzahl eher seltener Einheiten zu finden um Größenordnungen mehr Ausgangstext benötigt.

Welche quantitativen Angaben je Eintrag ein Frequenzwörterbuch enthält, ist höchst unterschiedlich. Als Standard kann gelten, dass angegeben wird, welche *absolute Häufigkeit* im Ausgangstext eine sprachliche Einheit hat. Zusammen mit der Angabe zur *Größe und Art des Ausgangstexts* ergeben sich die notwendigen und ausreichenden Angaben zur möglichen Zuverlässigkeit und Repräsentativität des Werks innerhalb einer Varietät. Aus diesen Angaben kann man dann etwa die *relative Häufigkeit* der Einheiten, bezogen auf alle Einheiten des Ausgangstexts, errechnen. Ein Problem bei der Zusammenstellung von Frequenzlisten stellt oft die ungleichmäßige Verteilung von Wörtern in bestimmten Teilen des Ausgangstexts dar. Dies tritt etwa durch Texte in Fachsprachen auf, die im Corpus enthalten sind oder Sprachen, die im Zusammenhang mit bestimmten Medien benutzt werden, wie SMS, Tweets oder Foren. Man kann dem damit begegnen, dass man mit einem bestimmten Koeffizienten gewichtet, um so das angenommene Ungleichgewicht auszugleichen. Man kann etwa den Teil mit der ungewöhnlichen Häufigkeit im Verhältnis zum Gesamttext betrachten, um auf einen Ausgleichsfaktor zu kommen. Angegeben wird dann die korrigierte absolute Häufigkeit. Andere mögliche statistische Angaben sind der Variationskoeffizient als Streumaß über Teile des Ausgangstexts oder der Standardhäufigkeitsindex, der angibt, nach welcher Anzahl anderer Wörter das betrachtete Wort statistisch wieder auftritt. All das versucht eine Brücke zwischen der Häufigkeit eines Wortes und der Wahrscheinlichkeit seines Auftretens zu schlagen, was aber nur in dem Rahmen allgemeingültige Ergebnisse liefern wird, der durch den Ausgangstext abgesteckt wurde. Eine weitere Größe, die durch die absolute Frequenz angegeben ist, ist der *Rang*. Er ist die Positionsnummer eines Wortes in der Liste aller Wörter, die absteigend nach der Häufigkeit geordnet ist. Ab einem gewissen Rang werden immer mehr Wörter mit derselben Häufigkeit auftreten. Für alle Wörter derselben Häufigkeit kann man dann das gleiche Rang Intervall angeben.

Häufigkeit und Rang bilden die Hauptcharakteristika eines Eintrags in einem Frequenzwörterbuch. Es ist durchaus möglich, dass das Wörterbuch genau wie ein herkömmliches aufgebaut, es also im Falle des Arabischen alphabetisch nach den üblichen Wurzelwörtern angeordnet, und mit den Zusatzangaben zur Frequenz versehen ist.

[59] Vgl. etwa Gerhard Fink: Langenscheidts Grundwortschatz Latein. Ein nach Sachgebieten geordnetes Lernwörterbuch mit Satzbeispielen. Völlige Neubearb. Berlin, 2001, S. 7–9.

Frequenzwörterbücher werden heute praktisch immer computergestützt erstellt, aber die Zahl der rein computergenerierten Wörterbücher ist noch gering.

Wer sind nun die Adressaten solcher Wörterbücher? Eines der ersten großen Frequenzwörterbücher wurde erstellt um, Druckmaschinen mit chinesischen Schriftzeichen effizient bauen zu können. Ein anderes Anwendungsgebiet war die Verbesserung von Kurzschrift. Oft wurden und werden sie auch zur Verbesserung von Sprachlehrkursen und Sprachlehrwerken verwendet. Die Erstellung eines Grundwortschatzes anhand eines sinnvoll zusammengestellten Corpus von Texten kann didaktisch sehr hilfreich sein. Oft sind Frequenzwörterbücher der Einstieg in eine tief gehende strukturelle Erforschung einer Sprache. In der Informationstechnologie spielen Frequenzwörterbücher, deren Ausgangstext nach anerkannten statistischen Methoden entstand, eine wichtige Rolle auf dem Gebiet der Informationsextraktion, bei Suchfunktionen oder bei der maschinellen Übersetzung.

Bei der Zusammenstellung muss nun beachtet werden, dass man sich der Grenzen und Möglichkeiten bewusst ist und sich kritisch damit auseinandersetzt. Statistik in der Linguistik ist oft schwer verständlich und verleitet zu der Annahme, dass mit den errechneten Zahlen die Arbeit erledigt ist. Statistik ist aber nur ein Werkzeug, das gültige und gute Ergebnisse liefert, wenn man weiß, wie man es einsetzt. Andernfalls können die Ergebnisse fragwürdig oder schlichtweg falsch sein. Den Erstellungsprozess eines Frequenzwörterbuchs kann man nun so skizzieren:

- Sprachwissenschaftler sollten sich als Erstes versichern, dass ihre Annahmen über die Verhältnisse der Textarten im untersuchten Gebiet zutreffend sind. An einer zufällig ausgewählten Menge Texte aus dem Untersuchungsgebiet, die daher eine Menge an sprachlichen Einheiten enthalten, deren Verteilung nicht repräsentativ ist, werden die textlichen Eigenheiten untersucht und es werden Experten zu den Metainformationen befragt usw. um ein Schema für die geordnete Zusammenstellung von Texten zu einem Corpus zu erhalten.
- Aus einer größeren Menge von Texten, in denen die sprachlichen Einheiten vorkommen, wird - wenn die Gesamtmenge nicht verarbeitbar ist - ein dem Schema entsprechendes Sample gezogen, in denen die sprachlichen Einheiten vorkommen werden.
- Dieses Sample - oder wenn möglich der gesamte Text -, in dem die sprachlichen Einheiten das Forschungsobjekt sind, wird analysiert und zusammen mit den Eigenschaften der Teiltexte festgehalten. Es wird auf Wort-, Satz- und/oder Textebene annotiert.
- Die Zusammenfassung der Wortformen unter den üblichen Lemmata oder eine sinnvolle Zusammenfassung größerer sprachlicher Einheiten führt dann zu den rohen Frequenzdaten einer Liste von sprachlichen Einheiten. Damit ist ein grundlegendes Frequenzwörterbuch erstellt.

In einem weiteren Schritt können Auswertungen der gesammelten Daten durchgeführt werden.

- Mit dem Frequenzwörterbuch, in dem die sprachlichen Einheiten mit den dazugehörigen Frequenzen verzeichnet sind, kann man nun Auswertungen durchführen, um etwa Tabellen mit der Verteilung der Einheiten zu erhalten.
- Aus Tabellen mit der Verteilung der sprachlichen Einheiten, in denen Häufigkeit und andere quantitative Daten enthalten sind, kann man durch Skalierung und der Verfolgung der Verteilung entsprechend aussagekräftige grafische Darstellungen der Verteilung erhalten.
- Mit den Tabellen der Verteilung und den Darstellungen der Verteilung, in denen die Häufigkeit und andere quantitative Daten enthalten sind und der Art und Form der grafischen Darstellungen, kann man durch Analyse der tabellarisch oder als Visualisierung aufbereiteten

Rohdaten Formeln auswählen, um die Daten analytisch zu repräsentieren, etwa mittels einer bestimmten Regressionsanalyse.

- Mit dieser analytischen mathematischen Darstellung der Verteilung und deren Parametern kann man durch Vergleich der empirischen Daten mit den theoretischen die Parameter weiter anpassen, um zu einer möglichst geringen Abweichung/einer möglichst guten Annäherung der empirischen und der theoretischen Verteilung zu kommen.

Ist man bei der Menge an Text, die ausgewertet werden kann, eingeschränkt, etwa weil man manuell auszählt, muss man sich Folgendes klar machen und auch den zukünftigen Anwendern des Wörterbuchs entsprechend mitteilen: Was für sprachliche Einheiten, welcher Sprache, welches Genre, Poesie, Prosa, Wissenschaftssprache, Umgangssprache sollen hier beschrieben werden? Soll eine Hochsprache beschrieben werden und dabei eventuell nicht normgerechte Sprache entsprechend angepasst werden oder soll die tatsächliche Verwendung beschrieben werden, ohne eine Abbildung von sprachlichen Einheiten in Varietäten auf entsprechende hochsprachliche Einheiten durchzuführen? Geht es um die mündliche oder die schriftliche Form?

Frequenzwörterbucher sind Erzeugnisse aus einem Corpus. Corpora sind wie erwähnt nicht in der Lage, eine lebende Sprache in ihrer Gesamtheit zu repräsentieren, die Grundgesamtheit ist nicht fassbar. Man kann versuchen die verschiedenen Gebiete einer Sprache proportional zu repräsentieren oder eine Gewichtung vorzunehmen. Beides ist sehr subjektiv, Ersteres, weil sich Sprache nicht streng in Gebiete einteilen lässt, und für Letzteres gibt es keine allgemein anerkannte Vorgehensweise. Die meisten Corpora für Frequenzwörterbücher, welche die „ganze“ Sprache erfassen wollen, bestehen aus 25 % Zeitungsartikeln, 25 % Radio- bzw. Fernsehnachrichten und 50 % Literatur ohne weitere Begründung.

Wenn man die Samples, die man aus der Gesamtmenge an Texten zieht, entsprechend der Anteile der verschiedenen Textkategorien an dieser gewichtet, dann hat man natürlich keine Repräsentation der ganzen Sprache, sondern kann seine Ergebnisse im besten Fall auf diese Gesamttextmenge übertragen. Die Zuverlässigkeit statistischer Aussagen ist abhängig von der Größe des erfassten Texts und dem Inhalt der Samples. Generell werden größere Mengen Text besser verallgemeinerbare Ergebnisse liefern und eine stärkere Beschränkung der Arten von Ausgangstexten wird bessere Ergebnisse innerhalb einer Auswahl ähnlicher Texte liefern.

Die Einheit, die gezählt wird, sollte auch verwendet werden, um das Corpus und die Samples zu beschreiben. Es ist nicht nützlich Mengen einmal in Seiten, einmal in Wörtern und einmal in Wortkombinationen anzugeben.

Im Zweifelsfall muss aber definiert und kommuniziert werden, was konkret mit „Wort“ in Frequenzwörterbuch gemeint ist. Neben zum Teil komplexen morphologischen Veränderungen neigen viele Sprachen zu Konstrukten, die ein Wort grafisch verändern, wobei aber die Teile des grafisch neuen Wortes als relativ eigenständig wahrgenommen werden. Wo es im Deutschen Komposita gibt, ist im Arabischen etwa die Frage zu klären, wie Partikel, die an das nächste Wort angefügt werden oder auch wie der arabische Artikel gezählt wird. Auch Endungen wie etwa in Turksprachen oder im Osmanischen würden ein Frequenzwörterbuch schnell unbrauchbar machen, wenn man sie einfach übergeht. Was wird also im Frequenzwörterbuch verzeichnet, Wörter oder Wortformen? Und was wird gezählt? Um den Problemen rund um die Definitionen von Wort, Wortform und Lemma auszuweichen, werden bei elektronischen Wörterbüchern oft „Token“ gezählt. Diese können dann wie bei Google Books zu „n-grams“ statt Wortkombinationen zusammengefasst werden. Token und N-Grams definieren Einheiten, die mit Software leichter fassbar sind, aber im sprachwissenschaftlichen Kontext teils falsche Ergebnisse

liefern. Computergestützte Lemmatisierung ist nicht perfekt und nur für einige Sprachen überhaupt weiter fortgeschritten. Bei manueller Zählung wird man eine repräsentative aber bearbeitbare Anzahl Samples ziehen. Dafür können Einträge mit beliebig komplexen syntaktischen und semantischen Informationen angereichert werden. Computergestützte Zählung basiert meist auf einer sehr einfachen Ebene von Zeichenketten mit bestimmten festgelegten Zeichen als Trenner.

Aus der alphabetischen Liste mit den Frequenzen und der Liste der Wörter nach Frequenz geordnet können dann die entsprechenden Daten errechnet werden, um die statistischen Auswertungen durchzuführen. Mit statistischen mathematischen Mitteln kann dann die Effizienz oder die Abdeckung an verschiedenen Texten erfasst werden. Wenn man mit Samples einer Elternpopulation gearbeitet hat, dann sollte vor allem in dieser Textmenge eine hohe Effizienz feststellbar sein. Einige zufällig ausgewählte Samples der Gesamttextmenge sollten Annahmen belegen oder widerlegen. Ideal wäre es mehrere Frequenzwörterbücher aus dem Ausgangstext zu erstellen, die qualitativ und quantitativ ähnlich sind und diese dann zu vergleichen. Praktikabler ist wohl das Frequenzwörterbuch an einem wesentlich kleineren Sample zu testen. Wenn es nicht allzu klein ist, kann man die Ergebnisse in das bestehende Wörterbuch integrieren. Je kleiner das Kontrollsample desto geringer ist die zu erwartende Effizienz. Eine Auswahlmöglichkeit für Kontrollen sind auch die verschiedenen Frequenzzonen des Wörterbuchs, etwa die 1000 häufigsten Wörter oder jene mit einer Häufigkeit von 1001 bis 2000.

Generell können Frequenzwörterbücher auf drei Gebieten verwendet werden:

- Man kann die häufigsten Vorkommen einer sprachlichen Einheit und diejenigen mit dem größten Bedeutungsgehalt ermitteln (nach Zipf sind die Ersteren am oberen Ende der Frequenzliste, die anderen am unteren Ende zu finden).
- Man kann einen Maßstab entwickeln, gewissermaßen ein Verhaltensmuster, einer Sprache, einer Varietät oder eines Idiolekts erstellen, anhand dessen man Texte einordnen kann.
- Mit diesen Mustern kann man dann Texte mit anderen Texten vergleichen, um daraus Schlüsse zu ziehen, etwa im Hinblick auf Autorenschaft oder Entstehungszeit und –ort.

Man erhält also ein probabilistisch-statistisches Modell der Lexik einer Sprache, mit dem man arbeiten kann.

In der theoretischen und angewandten Linguistik, in der Lexikographie und der Lexikologie, bei der Typologisierung von Texten, in der Stilistik oder etwa in der Sprachgeographie, der diachronen Untersuchung von Sprache und der Varietätenlinguistik sind Frequenzwörterbücher die ersten Produkte der linguistischen Statistik und werden dort etwa zum Erstellen statistischer Modelle benutzt. Des Weiteren sind Anwendungen in der Sprachdidaktik hervorzuheben, besonders in der L2-Didaktik, also beim Lernen von Zweit- und Fremdsprachen. Sorgfältig erstellte bilinguale, multilinguale und nach Thema geordnete Frequenzwörterbücher (beispielsweise Tim Buckwalter, Dilworth B. Parkinson (2011) oder Elisabeth Kendall (2012) für Hocharabisch) können eine effizientere Nutzung der Unterrichtszeit ermöglichen. Auch können sie als Grundlage für unterschiedliche Anwendungen im Bereich des computerunterstützten Lernens dienen. Weitere Anwendungsmöglichkeiten gibt es in der Rehabilitation und der Ausbildung von Blinden und Taubstummen.

In der Computerlinguistik kann der Computer nicht nur Hilfsmittel zur Erzeugung der Wörterbücher sein, sondern diese selbst auch verwenden, um tiefer gehende Probleme zu lösen, etwa in der maschinellen Übersetzung oder dem „information retrieval“.[60]

## Zur Problematik von statistischen Aussagen in der Sprachwissenschaft[61]

Er weist auf Gründe hin, durch die die Verwendung von Frequenzwörterbüchern manchmal unerwartete Resultate liefert: wie schon erwähnt erlauben Studien an Samples streng genommen nur Rückschluss auf die Menge an Text aus dem sie stammen. Eine Verallgemeinerung auf alle Texte oder gar eine Sprache ist meistens nicht sinnvoll. Das mögliche Lexikon einer jeden Sprache geht gegen unendlich und Änderungen selbst beim hochfrequenten Anteil der Wörter sind jederzeit möglich, wenn vielleicht auch wenig wahrscheinlich. Was niederfrequente Phänomene angeht, so sind Nachweise aus einem Textcorpus sehr stark an diesen gebunden. Man kann Ergebnisse mit einfachen Theorien bestenfalls von einer Untersuchung eines Corpus des Schaffens einer historischen Person auf deren gesamtes Schaffen übertragen. Es gilt aber durchaus, dass

> [T]he absolute difference between the relative frequency of a word A in a sample and its probability in the population tends to zero the larger the sample becomes.[62]

Die große Menge an Text ist heute verfügbar vornehmlich in den frei zugänglichen, wissenschaftlich gepflegten Corpora wie COCA (corpus of contemporary American English), COHA (corpus of historical American English) für amerikanisches Englisch, den IDS und DWDS Corpora für Deutsch, oder, wenn auch mit den erwähnten Einschränkungen in noch größerem Ausmaß die n-grams von Google Books. Das Problem der Repräsentativität von Corpora bleibt aber ungelöst. Somit gibt es auch keine wirklich repräsentativen Samples. Bemühungen um die maximale erreichbare Repräsentativität machen aber mit einer gewissen Unsicherheit behaftete Aussagen möglich.

Die Listen mit rohen Frequenzdaten vermitteln oft ein gewisses Gefühl der Stabilität, die so allerdings in lebendigen Sprachen nicht existiert. So gibt es etwa das Beispiel, dass einmal „bacon“ doppelt so oft gezählt wird wie „cheese“ und in einer anderen Untersuchung genau andersherum oder „October“ seltener vorkommt als „November“. Aus den Aussagen, die aus einem bestimmten Corpus gewonnen werden, lassen sich kaum Rückschlüsse ziehen, die jene mathematische Präzision haben, die mit Zahlenangaben oft assoziiert wird. Unter anderem ist dies deswegen der Fall, weil die absolute Anzahl von in Corpora erfassten Worten im Vergleich zu dem, was an Sprache im Laufe der Zeit produziert wird, eher gering ist.

Um dem Problem zu begegnen, dass in manchen Corpora bestimmte Ergebnisse von Sprachproduktion nicht in dem Maße repräsentiert sind, wie es ihrem Anteil an der realen Sprachproduktion entspricht, wurden Koeffizienten eingeführt. Etwa ist gesprochene Sprache gegenüber jener Sprache, die in rein schriftlicher Textproduktion verwendet wird oft unterrepräsentiert. Die Sprachproduktion im Internet, die oft zwischen gesprochener Sprache und rein schriftlicher Textproduktion angesiedelt ist, muss ebenso eingeordnet werden. Die Ermittlung der richtigen Berechnungsmethode für diese Koeffizienten

[60] Alekseev (2005).

[61] Der folgende Abschnitt bezieht sich im Wesentlichen auf Willy Martin: 143. The Frequency Dictionary. in: Franz Josef Hausmann, Oskar Reichmann, Herbert Ernst Wiegand, Ladislav Zgusta (Hg.): *Wörterbücher. Ein internationales Handbuch zur Lexikographie,* Berlin, 1990 (= Handbücher zur Sprach- und Kommunikationswissenschaft Bd. 5.2), S. 1314–1322.

[62] Martin (1990), S. 1317–1322.

hängt mehr oder weniger stark vom Sprachgefühl derjenigen ab, die die Corpora erstellen oder aus vorhandenen Corpora zusammenstellen, die die Grundlage eines Frequenzwörterbuchs bilden.

Andere subjektive Ideen zur besseren Abbildung von tatsächlich existierendem Sprachgebrauch sind etwa Leute, deren Erstsprache die untersuchte Sprache ist, zu fragen, welche 20 Wörter ihnen spontan einfallen und diese mit aufzunehmen und stark zu gewichten. Dies kompensiert wohl zu einem Teil den Umstand, dass die Texte in Corpora zum überwiegenden Teil keine freie mündliche Textproduktion betreffen oder übliche Situationen wie Essen, arbeiten im Haushalt etc. So wurde schon gezielt nach Ausdrücken gesucht, die jeder Sprecher einer Sprache zur Verfügung hat oder haben sollte, die aber in keinem Corpus häufig genug erfasst sind, als dass sie in kürzeren Frequenzlisten aufscheinen, etwa Geschirr, Besteck und Küchengeräte.

Eine andere Möglichkeit deutlich zu machen, dass die reinen Zahlen oft wenig Aussagekraft haben, ist, statt der Zahlen, nach einer eher frei gewählten Einteilung anzugeben, ob ein Ausdruck sehr häufig, häufig, weder häufig noch selten oder selten ist und für Ausdrücke, für die man sehr seltene Belege gefunden hat, gar keine Frequenz mehr anzugeben. Man kann die Einteilung auf Konzepte für Sprachkompetenz stützen und so versuchen Objektivität und Subjektivität zu verbinden.[63]

## Beispiele für Frequenzwörterbücher

Beispiele für Frequenzwörterbücher, die eine Dokumentation des Erstellungsprozesses enthalten, was eine Voraussetzung dafür ist, dass diese für weitergehende wissenschaftliche Untersuchungen verwendet werden können, sind zum Beispiel für Englisch Winthrop Nelson Francis, Henry Kučera (1982) und aus einer Reihe des Routledge Verlags, in der Frequenzwörterbücher für mehrere Sprachen erschienen sind, gibt es solche für Deutsch[64] oder für Arabisch[65]. Offensichtlich sind etwa auch andere Wörterbücher, die einen Grundwortschatz vermitteln sollen, aus Frequenzzählungen entstanden, etwa die auf Seite 18 genannten von Gerhard Fink (2001) und Magdi Fouad (2011). Es fehlen aber viele Angaben zu Methodik oder Häufigkeit, was eine genauere Beurteilung unmöglich macht. Das Frequenzwörterbuch von Francis ist zwar ein gutes Beispiel für ein vorbildlich aufgebautes Frequenzwörterbuch. Man sollte aber nicht übersehen, dass es auf dem Brown Corpus für amerikanisches Englisch basiert. Dieser trägt zwar „Present-Day“ in seinem Namen, gemeint ist aber aktuell für die 1960er Jahre. Er wurde 1963/64 aus Texten von 1961 zusammengestellt. Beschrieben ist die Zusammenstellung des Corpus, es sind 500 Samples aus allen möglichen Arten von Textproduktionen aus dem Jahr 1961 nach der oben beschriebenen Sample Methode (siehe Seite 16) zusammengestellt. Der Umfang beträgt 1.014.000 Worte. Die Autoren beschreiben die Annotationen, mit denen die Texte angereichert wurden mit einer Erklärung und Beispielen. [66] Weiters ist bei der Rangliste angegeben wie nach der Verteilung über die Kategorien, in die die Texte eingeteilt wurden, korrigierte Frequenzwerte errechnet wurden.[67] Es wird eine alphabetisch geordnete Liste von Worten geboten[68], bei der je Wort angegeben ist, wie oft es gezählt wurde, in wie vielen der Textsorten es gefunden wurde und auch in wie vielen der Samples es gefunden wurde. Diese Auflistung enthält alle

[63] Martin (1990).
[64] Randall L. Jones, Erwin Tschirner: A frequency dictionary of German. Core vocabulary for learners. 1. Auflage. London, 2006.. Dieses basiert auf einem viel kleineren Corpus als jenes für Arabisch von nur 4 Millionen Worten. Allerdings ist die Zusammenstellung des Corpus sorgfältig ausbalanciert um eine möglichst gute Repräsentativität zu erreichen S. 1f.
[65] Buckwalter, Parkinson (2011).
[66] Winthrop Nelson Francis, Henry Kučera: Frequency analysis of English usage. Lexicon and grammar. Boston, 1982, S. 4.
[67] Francis, Kučera (1982), S. 461–464.
[68] Francis, Kučera (1982), S. 18–460.

gefundenen Wörter, auch Hapaxes[69]. Danach ist eine Rangliste abgedruckt. Diese ist zwei Mal aufgeführt. Einmal geordnet nach den rohen Frequenzdaten und einmal geordnet nach den korrigierten Frequenzen. In dieser zweiten Auflistung der Wörter werden alle Frequenzen absteigend geordnet bis zu einem Minimum von fünf angegeben.[70] Weiters sind Frequenzangaben nach Wortklassen[71] enthalten sowie eine Diskussion der vorgefundenen Sätze nach Länge und Struktur[72]. Man sollte bedenken, dass eine Verarbeitung von einer Million Wörtern Ende der 1970er bzw. Anfang der 1980er Jahre technisch aufwendig war und daher für die Zeit des Erscheinens als wegweisend gelten muss. Heute ist das Brown-Corpus leicht im Internet zugänglich und dient als Beispieldatensatz im weitverbreiteten Programmpaket NLTK[73]. Der Bezug auf ein Corpus, das 20 Jahre zuvor erstellt wurde, ist der Aussagekraft für aktuellen Sprachgebrauch für die 1980er Jahre allerdings nicht zuträglich. Für historische Untersuchungen ist es andererseits eine wichtige Quelle.

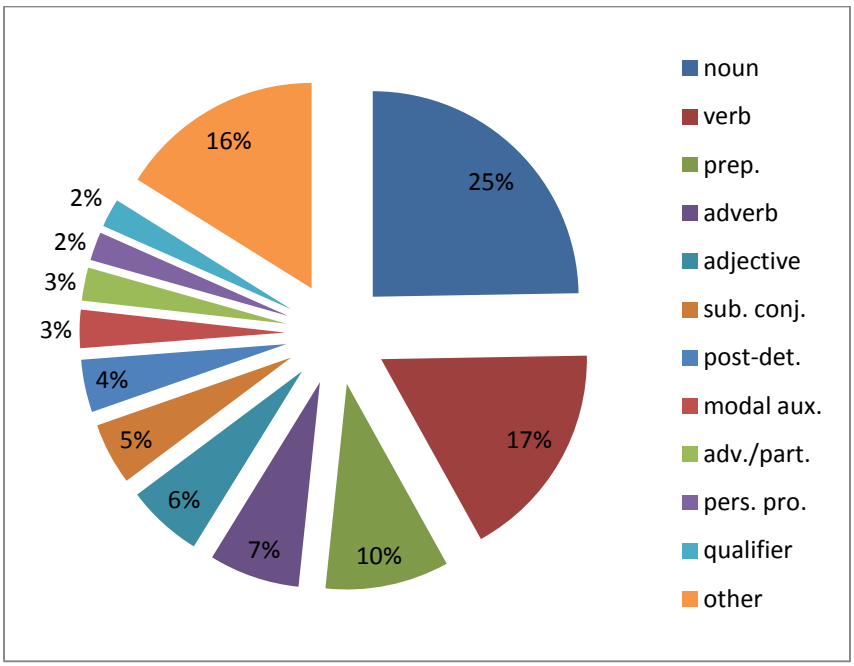
Als Beispiel soll hier die Liste der 15 häufigsten Wörter nach roher (linke Spalte) bzw. korrigierter (rechte Spalte) Frequenz angegeben werden:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 1 | the | article | 69975 | 69792.94 | 1 | the | article | 69975 | 69792.94 |
| 2 | be | verb | 39175 | 39109.95 | 2 | be | verb | 39175 | 39109.95 |
| 3 | of | prep. | 36432 | 35786.01 | 3 | of | prep. | 36432 | 35786.01 |
| 4 | and | co. conj. | 28872 | 28821.11 | 4 | and | co. conj. | 28872 | 28821.11 |
| 5 | a | article | 23073 | 22984.95 | 5 | a | article | 23073 | 22984.95 |
| 6 | in | prep. | 20870 | 20685.17 | 6 | in | prep. | 20870 | 20685.17 |
| 7 | he | pers. pro. | 19427 | 17280.77 | 7 | he | pers. pro. | 19427 | 17280.77 |
| 8 | to | inf. mark. | 15025 | 14990.82 | 8 | to | inf. mark. | 15025 | 14990.82 |
| 9 | have | verb | 12458 | 12192.06 | 9 | have | verb | 12458 | 12192.06 |
| 10 | to | prep. | 11165 | 11129.57 | 10 | to | prep. | 11165 | 11129.57 |
| 11 | it | pronoun | 10942 | 10836.51 | 11 | it | pronoun | 10942 | 10836.51 |
| 12 | for | prep. | 8996 | 8899.55 | 12 | for | prep. | 8996 | 8899.55 |
| 13 | I | pers. pro. | 8387 | 6885.48 | 13 | they | pers. pro. | 8284 | 8162.08 |
| 14 | they | pers. pro. | 8284 | 8162.08 | 14 | with | prep. | 7286 | 7267.37 |
| 15 | with | prep. | 7286 | 7267.37 | 15 | I | pers. pro. | 8387 | 6885.48 |

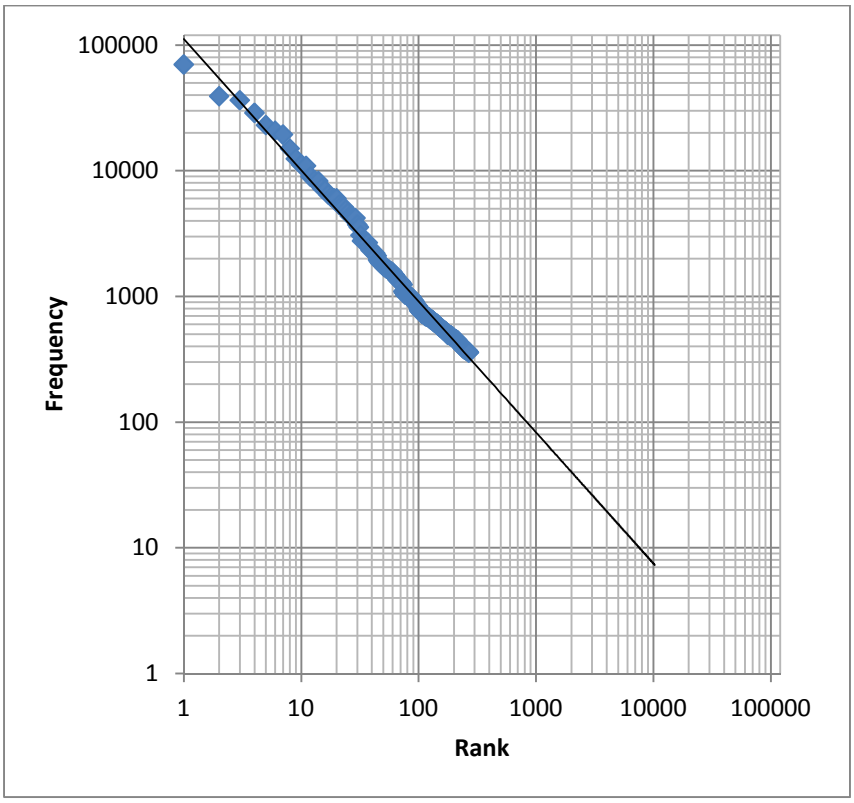
**Abbildung 2 Die 15 häufigsten Wörter im Brown Corpus von 1961**[74]

Folgende Anteile von Wortkategorien finden sich hier unter den 200 häufigsten Wörtern (unter „other“ sind solche Kategorien von Wörtern zusammengefasst, die weniger als fünf Mal vorkommen:

[69] Francis, Kučera (1982), S. 16f.
[70] Francis, Kučera (1982), S. 465–532.
[71] Francis, Kučera (1982), S. 533–548.
[72] Francis, Kučera (1982), S. 549–556.
[73] Steven Bird, Ewan Klein, Edward Loper: Natural language processing with Python. [analyzing text with the natural language toolkit]. 1. Auflage. Sebastopol, Calif., USA, 2009, S. 42–44.
[74] Francis, Kučera (1982), S. 18.




Das Rang-Frequenzdiagramm der Wörter folgt in etwa dem, wonach man nach Zipf ausgehen muss:

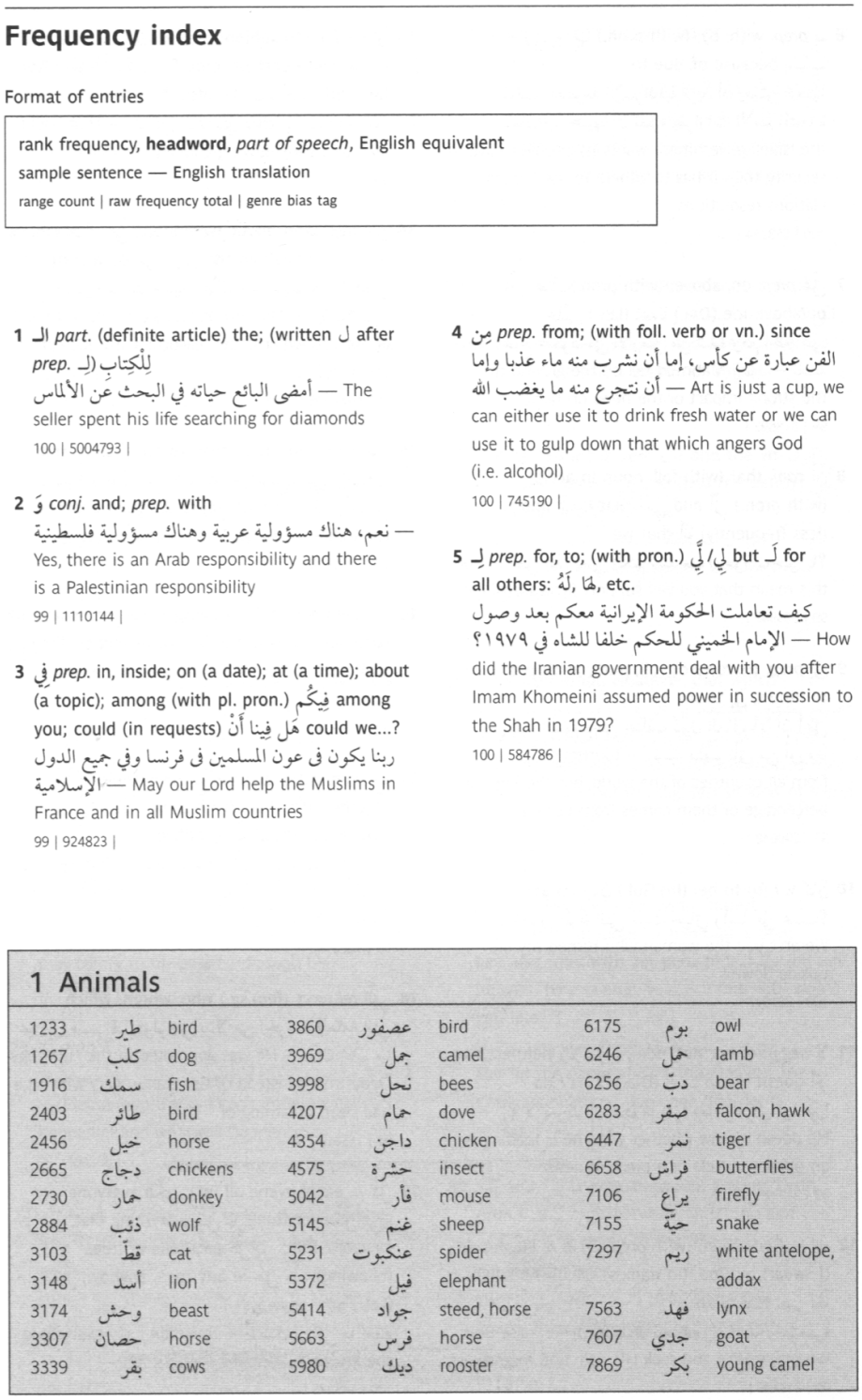



Das zweite Werk, das als Beispiel für ein Frequenzwörterbuch dienen soll, ist 2011 erschienen. Auch hier ist sowohl die Datenbasis beschrieben als auch die Berechnungen, die durchgeführt wurden, um ein häufigeres Vorkommen eines Wortes in einer bestimmten Kategorie von Texten auszugleichen. Dieses Frequenzwörterbuch des Arabischen basiert auf einer respektablen Menge von 30 Millionen Worten, wobei mit 3 Millionen transkribierten Worten aus spontanen Sprechsituationen der Problematik der Diglossie in gewisser Hinsicht Rechnung getragen wurde.

Der Hauptteil des Buches ist die nach korrigierten Frequenzen absteigend angeordnete Liste lemmatisierter Formen bis zu einer Mindestfrequenz von 5000.[75] Dabei werden einige unerwartete Zuordnungen vorgenommen. Es werden beispielsweise manche feste Wortverbindungen aufgelöst. So ist آن „Zeit“ angegeben, dessen Hauptvorkommen wohl الآن „jetzt“ ist oder أيض „das Wiederzurückkehren“ für أيضاً „auch“. Die Wortklasse wird als Substantiv angegeben, was formal richtig ist. Auch Länderadjektive werden nicht extra ausgewiesen, sondern unter Adjektive bzw. Substantive subsumiert. Weiters findet man in dieser Liste Kästen mit thematisch gruppierten Worten, deren Frequenz auch weniger als 5000 betragen kann, etwa für den menschlichen Körper, Essen, Berufe, Nationalitäten u. a. m.[76] Danach folgen Indizes der Wörter in alphabetischer Anordnung[77] und nach Wortarten und absteigender Frequenz aufgeschlüsselt.[78]

[75] Buckwalter, Parkinson (2011), S. 9–432.
[76] Buckwalter, Parkinson (2011), S. vi.
[77] Buckwalter, Parkinson (2011), S. 433–514.
[78] Buckwalter, Parkinson (2011), S. 515–578.

Zur Illustration sei hier die erste Seite des Frequenzwörterbuchs wiedergegeben:



## Frequency index

Format of entries

rank frequency, **headword**, *part of speech*, English equivalent
sample sentence — English translation
range count | raw frequency total | genre bias tag

**1 الـ** *part.* (definite article) the; (written ل after *prep.* لِـ) لِلْكِتابِ
أمضى البائع حياته في البحث عن الألماس — The seller spent his life searching for diamonds
100 | 5004793 |

**2 وَ** *conj.* and; *prep.* with
نعم، هناك مسؤولية عربية وهناك مسؤولية فلسطينية — Yes, there is an Arab responsibility and there is a Palestinian responsibility
99 | 1110144 |

**3 فِي** *prep.* in, inside; on (a date); at (a time); about (a topic); among (with pl. pron.) فِيكُم among you; could (in requests) هَل فِينا أَنْ could we...?
ربنا يكون فى عون المسلمين فى فرنسا وفي جميع الدول الإسلامية — May our Lord help the Muslims in France and in all Muslim countries
99 | 924823 |

**4 مِن** *prep.* from; (with foll. verb or vn.) since
الفن عبارة عن كأس، إما أن نشرب منه ماء عذبا وإما أن نتجرع منه ما يغضب الله — Art is just a cup, we can either use it to drink fresh water or we can use it to gulp down that which angers God (i.e. alcohol)
100 | 745190 |

**5 لِـ** *prep.* for, to; (with pron.) لِي/لَّي but لَـ for all others: لَهُ, لَهَا, etc.
كيف تعاملت الحكومة الإيرانية معكم بعد وصول الإمام الخميني للحكم خلفا للشاه في ١٩٧٩؟ — How did the Iranian government deal with you after Imam Khomeini assumed power in succession to the Shah in 1979?
100 | 584786 |

### 1 Animals

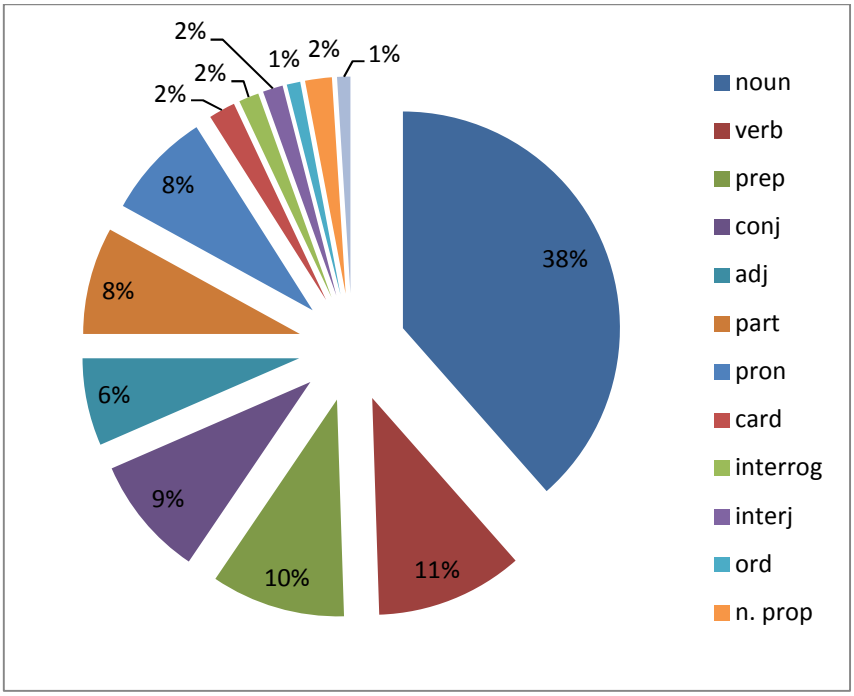
| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 1233 | طير | bird | 3860 | عصفور | bird | 6175 | بوم | owl |
| 1267 | كلب | dog | 3969 | جمل | camel | 6246 | حمل | lamb |
| 1916 | سمك | fish | 3998 | نحل | bees | 6256 | دبّ | bear |
| 2403 | طائر | bird | 4207 | حمام | dove | 6283 | صقر | falcon, hawk |
| 2456 | خيل | horse | 4354 | داجن | chicken | 6447 | نمر | tiger |
| 2665 | دجاج | chickens | 4575 | حشرة | insect | 6658 | فراش | butterflies |
| 2730 | حمار | donkey | 5042 | فأر | mouse | 7106 | يراع | firefly |
| 2884 | ذئب | wolf | 5145 | غنم | sheep | 7155 | حيّة | snake |
| 3103 | قطّ | cat | 5231 | عنكبوت | spider | 7297 | ريم | white antelope, addax |
| 3148 | أسد | lion | 5372 | فيل | elephant | | | |
| 3174 | وحش | beast | 5414 | جواد | steed, horse | 7563 | فهد | lynx |
| 3307 | حصان | horse | 5663 | فرس | horse | 7607 | جدي | goat |
| 3339 | بقر | cows | 5980 | ديك | rooster | 7869 | بكر | young camel |

**Abbildung 3 Erste Seite des Frequenzwörterbuchs für Arabisch**[79]

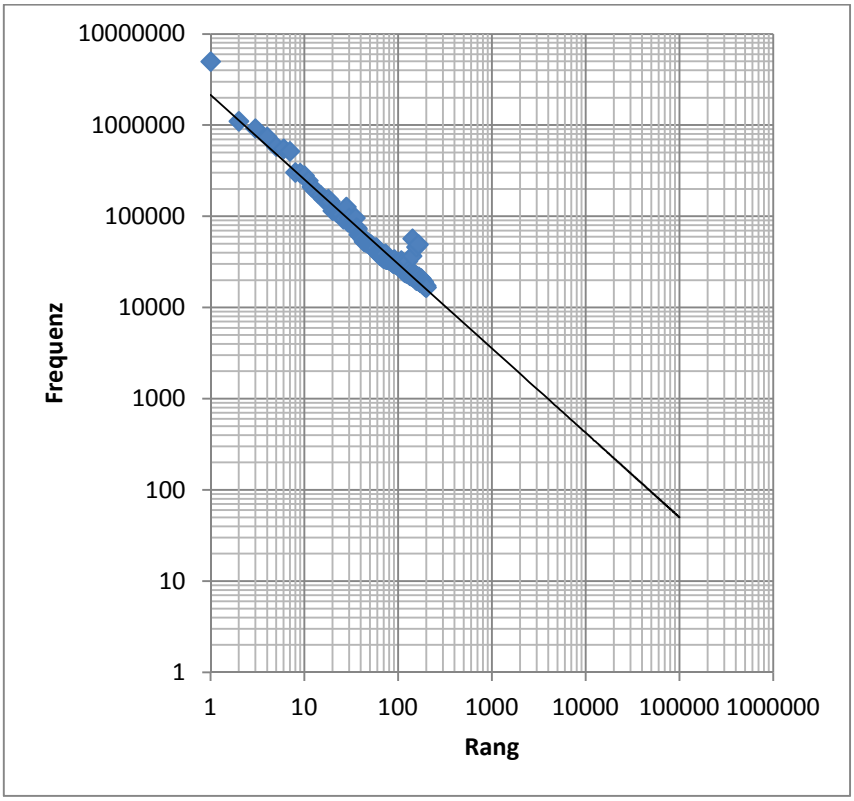
Die Wortart muss nicht eindeutig sein, etwa bei Ländernamen und Länderadjektiven. Die korrigierte Frequenz ist nicht angegeben, lässt sich aber näherungsweise mit den angegebenen Zahlen ermitteln, wobei das Verteilungsmaß gerundet ist.[80]

[79] Buckwalter, Parkinson (2011), S. 9.
[80] Buckwalter, Parkinson (2011), S. 6.

Unter den 200 häufigsten Wörtern finden sich hier folgende Anteile von Wortkategorien:




Auch das Rang-Frequenz-Diagramm folgt im Großen und Ganzen Zipfs Gesetz, die Ausreißer ergeben sich aus der nicht durchgeführten Korrektur:




Weiter frequenzbasierte Werke für Hocharabisch sind etwa: Jacob M. Landau (1959), von dem es 2011 eine Neuauflage gab. Einen praktisch ausgerichteten Grundwortschatz, dessen Zusammenstellung allerdings nicht genau dokumentiert ist, gibt es mit Magdi Fouad (2011). Für Medienarabisch gibt es Elisabeth Kendall (2005) bzw. Elisabeth Kendall (2012). Für ägyptisches Arabisch gibt es mit Mohamed

Abdel Aziz (2007b) einen Grundwortschatz ohne eine genauere Angabe darüber, wie es zusammengestellt wurde.

# Über Wikipedia[81]

Im folgenden Kapitel soll versucht werden, das untersuchte Corpus, sowie die Regeln nach denen es erstellt wurde und das, was man über die Urheber weiß, zu beschreiben. Es soll auch um die Rahmenbedingungen gehen, mit denen Personen, die in einer Wikipedia schreiben, konfrontiert sind. Diese Bedingungen, technischer aber auch sozialer Natur, prägen die Art, wie Texte in eine Wikipedia einfließen und welchen Veränderungen sie unterworfen sind oder welche Veränderungen bewusst unterlassen werden, da sie der „Kultur von Wikipedia" nicht angemessen erscheinen.

## Die Entstehungsgeschichte des Wikipedia Projekts

Wikipedia ist verglichen mit anderen Internetphänomenen, ein junges Produkt. Eine der ersten Ideen, die schließlich zu dem führten, was wir heute als Wikipedia kennen, ist laut der Wikipedia Seite History of Wikipedia[82], auf das Jahr 1993 zurückzuführen. In diesem Jahr stellte Rick Walters, ein Student an der University of Arizona, ein Konzept zur Diskussion, mit dem er das Internet nutzen wollte, um Experten aller nur denkbaren Disziplinen ein Forum zu bieten, ihre Fachmeinungen in eine neue Art von Enzyklopädie einzubringen. Er dachte an die CD-ROM, die er zu Hause liegen hatte und auf der die Enzyklopädie eines renommierten Verlags gespeichert war. Er wollte diese Idee einer elektronischen Enzyklopädie direkt im Internet umsetzen. Von allem Anfang an hoffte er, dass man in und mit einer solchen Wissenssammlung aufschlussreiche neue Forschungsarbeiten durchführen könnte. Er machte sich auch Gedanken darüber, wie das Unterfangen finanziert werden könnte.[83] Bedenkt man die Zeit, scheint es kein Wunder zu sein, dass diese Idee nicht sehr bekannt wurde.

Sieben Jahre später hatte Richard Stallman, der Mann hinter der GPL[84], die sicherstellen soll, dass Software für jeden Interessierten zugänglich und veränderbar ist, das Gefühl, dass es etwas Ähnliches auch für enzyklopädisches Wissen geben müsste.

## Die erste Idee zu einer freien Enzyklopädie[85]

Er war von dem Gefühl getrieben, dass Firmen den Menschen etwas wegnehmen, um es ihnen dann zu verkaufen. Er meinte, Firmen würden irgendwann so und so das Internet abgrasen und ihre Funde verkaufen. Da wäre es besser, Daten gleich frei zugänglich anzubieten. Ihm schwebte eine Enzyklopädie vor (und ebenso konstruierte Kurse mit denen sich jeder in ein Gebiet einarbeiten kann), die überall im Internet verfügbar wäre und überall von der gesamten Internetgemeinde vorangetrieben wird. Er wollte die Enzyklopädie komplett dezentral organisieren, ohne eine zentrale Infrastruktur oder eine Organisation, die alles kontrolliert und entscheidet. Ein Punkt trifft auf die Wikipedia von heute eindeutig nicht zu. Es gibt eine zentrale und damit rechtlich oder physisch angreifbare Infrastruktur.

[81] Wikipedia steht für zwei Dinge: Einmal ist es der gängigen Ausdruck für die enzyklopädieartige Wissenssammlung im Internet etwa auf Deutsch, die heute jeder kennt. Diese liegen in verschiedensten Sprachen vor. Zum anderen ist es jener Name den die WikiMedia Foundation für jene Klasse von enzyklopädieartigen Wissenssammlungen verwendet im Gegensatz etwa zu Wiktionary das für eine Klasse von wörterbuchartigen Sammlungen steht. Deshalb soll hier von Wikipedia ohne Artikel die Rede sein wenn es um alle Wikipedias in den verschiedenen Sprachen geht und der Artikel soll dazu dienen auf eine konkrete Wikipedia in einer bestimmten Sprache zu verweisen.

[82] Wikipedia contributors: History of Wikipedia. http://en.wikipedia.org/wiki/History_of_Wikipedia (Zugriff am: 16. 11. 2012).

[83] Rick Gates: The Internet Encyclopedia. http://listserv.uh.edu/cgi-bin/wa?A2=ind9310d&L=pacs-l&T=0&P=1418 (Zugriff am: 16. 11. 2012).

[84] Richard Stallman: GNU General Public License, version 2 (GPL-2.0). http://opensource.org/licenses/GPL-2.0 (Zugriff am: 16. 11. 2012).

[85] Der folgende Abschnitt bezieht sich im Wesentlichen auf Richard Stallman: The Free Universal Encyclopedia and Learning Resource. http://www.gnu.org/encyclopedia/anencyc.txt (Zugriff am: 16. 11. 2012).

Stallman sah Lehrerende und Schüler bzw. Studenten, engl. „students“, die außergewöhnlich gute Arbeiten zu einem bestimmten Thema verfassen, als diejenigen, die die Einträge verfassen und sie so im Internet zugänglich machen. Auch sehr kurze Beiträge hielt er für wichtig, um die Motivation der Schreibenden zu erhalten. Er prophezeite, dass das Projekt erst langsam Geschwindigkeit aufnehmen würde und dass die ersten Beitragenden viel Geduld aufbringen müssten. Er hoffte aber darauf, dass es irgendwann einen Lawineneffekt geben und das Projekt wahrscheinlich nach 20 Jahren abgeschlossen sein werde. Der erste Teil der Voraussage hat sich wohl bewahrheitet, der Zweite hat mit der heutigen Realität eines zeitlich unbegrenzten Projekts nichts mehr zu tun. Er schlug einerseits vor, die Enzyklopädie nach dem Vorbild gedruckter Vorlagen thematisch zu limitieren, war sich andererseits aber offensichtlich auch nicht sicher, ob das dem Universalitätsanspruch des Projekts nicht zuwiderliefe. Er schlug selbstverständlich vor, die freie Nutzung und Verbreitung rechtlich abzusichern. Er warnte davor Beiträge anzunehmen, die an für das Ziel kontraproduktive Bedingungen geknüpft sind. Da das Ziel in einer allgemein verfügbaren, frei verwendbaren Enzyklopädie besteht, ist es etwa nicht sinnvoll möglich Bilder, Texte, Audio- oder Videodateien zu integrieren, die nicht kommerziell genutzt werden dürfen. Dies schließt viele Verwendungsmöglichkeiten aus und bringt rechtliche Unsicherheit im Bezug darauf, ab wann eine Verwendung nicht mehr nicht-kommerziell ist. Stallman machte sich in seiner Schrift auch Gedanken darüber, dass zentrale Organisationsformen anfällig für Naturkatastrophen, politische Entscheidungen oder schlicht Geldmangel sind, womit derart organisierte Projekte auch schnell zusammenbrechen können. Dezentralität und das Vermeiden einer Kontrollinstanz waren ihm so wichtig, dass er dies sogar noch einmal herausstrich. Seine Lösung ist technisch einfach: Die Informationen werden bei allen interessierten Institutionen rund um die Welt in Kopie vorgehalten, gespiegelt. Dies geschieht bei Wikipedia ziemlich genau so, wie er es postulierte, allerdings werden die Daten meist nicht in der Form vorgehalten, wie man sie im Internet sehen kann (sondern als „Dump“, also als Komplettauszug der Datenbank die hinter den einzelnen Wikipedias steht). Er setzte sich dafür ein, Übersetzungen willkommen zu heißen und rechtlich abzusichern, da er annahm, dass die englische Wikipedia die größte werden würde. Dies hat sich bestätigt, wie ein Blick in die Statistiken der WikiMedia Foundation zeigt[86]. Das Thema Übersetzung wurde und wird allerdings kontrovers diskutiert.[87] Nicht weiter verwunderlich ist auch, dass er sich für eine freie Lizenzierung aller weiteren Teile einer typischen Enzyklopädie, wie Bilder oder Videos, einsetzte. Eine weitere seiner Forderungen wurde inzwischen in der Wikipedia implementiert: Er meinte damals, es müsste einen Mechanismus zur Kennzeichnung anerkanntermaßen guter Artikel geben. Auch schlug er vor, eine sehr restriktive Politik für die Verlinkung von weiterführender Literatur oder Quellen zu verfolgen. Hat ein potenzielles Ziel eines Links zu restriktive Nutzungsbedingungen, dann sollte es gar nicht verlinkt werden.[88]

[86] Sowohl die Anzahl der Artikel (http://stats.wikimedia.org/EN/TablesArticlesTotal.htm) als auch die Anzahl der Beitragenden (http://stats.wikimedia.org/EN/TablesWikipediansContributors.htm) weist die englische Wikipedia als die größte aus. Andere ermittelte Kenngrößen bestätigen die führende Rolle der englischen Wikipedia.

[87] Vgl. Wikipedia contributors: Wikipedia:Translation. http://en.wikipedia.org/wiki/Wikipedia:Translation (Zugriff am: 16. 11. 2012)., Wikipedia contributors: Wikipedia:Übersetzungen. http://de.wikipedia.org/wiki/Wikipedia:%C3%9Cbersetzungen (Zugriff am: 16. 11. 2012)., Wikipedia contributors: ويكيبيديا:ترجمة مقالات إلى العربية. http://ar.wikipedia.org/wiki/%D9%88%D9%8A%D9%83%D9%8A%D8%A8%D9%8A%D8%AF%D9%8A%D8%A7:%D8%AA%D8%B1%D8%AC%D9%85%D8%A9_%D9%85%D9%82%D8%A7%D9%84%D8%A7%D8%AA_%D8%A5%D9%84%D9%89_%D8%A7%D9%84%D8%B9%D8%B1%D8%A8%D9%8A%D8%A9 (Zugriff am: 16. 11. 2012). und Wikipedia contributors: Rootschläg för Übersetzige. http://als.wikipedia.org/wiki/Hilfe:Ratschl%C3%A4ge_f%C3%BCr_%C3%9Cbersetzungen (Zugriff am: 16. 11. 2012).

[88] Richard Stallman: The Free Universal Encyclopedia and Learning Resource. http://www.gnu.org/encyclopedia/anencyc.txt (Zugriff am: 16. 11. 2012).

## Vom Wiki Konzept zur frei verfügbaren Software

Der folgende Abschnitt bezieht sich in wesentlichen Teilen auf Bo Leuf, Ward Cunningham (2001). Die Idee zu Wiki entstand schon 1994. Ward Cunningham hatte sie als er über Möglichkeiten nachdachte, wie er möglichst effektiv mit anderen Programmierern gemeinsam an Entwurfsmustern für Software arbeiten konnte. Was er benötigte, war ein Werkzeug zur Zusammenarbeit an und zur Diskussion von Texten. Vor ein paar Jahren nannte man solche Software Groupware oder Collaboration Software[89], heute wird es als Social Software bezeichnet, weil es etwas Ähnliches ermöglicht, wie die Social Media des Web 2.0. Eine der ersten Softwarelösungen für solche Aufgaben war etwa Lotus Notes (heute von IBM) oder Novell Groupwise aber auch Exchange von Microsoft wird firmenintern häufig so verwendet. Allen Letztgenannten ist gemeinsam, dass sie relativ teuer, relativ kompliziert zu betreiben und zueinander völlig inkompatibel sind und eingegebene Informationen in ihren internen Strukturen „gefangen" halten. Unter anderem deswegen suchte Ward Cunningham einen neuen Zugang zum Thema Zusammenarbeit. Die Lösung sollte kostengünstig, leicht einzurichten und von möglichst jedem verfügbaren Computer aus nutzbar sein. Die Dokumente sollten immer in der aktuellsten Fassung verfügbar sein und es erlauben, Informationen sinnvoll miteinander zu verknüpfen sowie durch die Auszeichnung der Struktur der Information diese über eine Suchfunktion gut zugänglich zu machen. Daneben sollten die Informationen auf verschiedenste Arten ausgebbar sein, also nicht nur als HTML für verschiedene Webbrowser, was in den 1990er Jahren ein beträchtliches Problem darstellte, sondern auch in anderen Formaten wie etwa PDF für den Druck u. a. m. Die Software sollte es so einfach wie möglich machen, Texte miteinander zu verknüpfen. Die Informationen sollten in einem unabhängig von der eingesetzten Software lesbaren Format vorgehalten werden. Bei den schon 1994 bekannten Ansätzen wie E-Mail Mailinglisten missfiel ihm, dass man häufig zu sehr damit beschäftigt war, neue Informationen mitzubekommen und miteinander im Kopf zu verknüpfen. Außerdem kann man einmal versandte E-Mails nicht mehr ändern. Sie sind dann vielleicht in einem öffentlichen Archiv für jedermann lesbar, aber für niemanden änderbar. Etwas Ähnliches gilt für Foren. Bei den oben genannten Produkten namhafter Softwarehersteller war das gemeinsame Arbeiten an Dokumenten auch schon länger möglich, allerdings mussten und müssen sich die Benutzer öfter darum kümmern, dass jene Dokumente, die am gemeinsam genutzten Speicherplatz im Netzwerk liegen auch zu dem passen, was in den Nachrichten besprochen wurde. Des Weiteren gab es auch damals schon teilweise interaktive Websites, auf denen man etwa Kommentare im Gästebuch hinterlassen konnte, aber der eigentliche Inhalt der Seiten wurde ausschließlich von den Betreibern gestaltet. Das Wichtigste aber ist: Ein Server für Diskussionen muss einfach zu nutzen sein. Weiters muss es einfach möglich sein andere Informationen oder andere elektronische Dinge zu referenzieren (also sie zu verlinken).[90]

Ward Cunningham schrieb dann eine Software für seine Website namens WikiWikiWeb. Diese bestand insgesamt nur aus einigen Hundert Zeilen Quellcode. Die 2001 publizierte Version für „The Wiki Way" ist jedenfalls nicht länger.[91] Sie stellte eine der frühesten Möglichkeiten für Benutzer des Internets dar, eine

[89] Die deutsche Übersetzung Kollaborationssoftware ist eine IT-typische Neubelegung, die im deutschen eher anrüchig klingen mag. Es soll nur den Aspekt der Zusammenarbeit ausgedrückt werden.

[90] Bo Leuf, Ward Cunningham: The Wiki way. Quick collaboration on the web. 1. Auflage. Boston, Mass., USA, 2001, S. 3–12.

[91] Bo Leuf, Ward Cunningham: Quellecode für The Wiki Way. http://web.archive.org/web/20070822133615/http://www.leuf.net/ww/tww?WikiWaySources (Zugriff am: 24. 11. 2012).

Seite, die sie gerade sehen, direkt zu ändern und sie ist heute noch in Betrieb.[92] Man findet dort heute rund 35000 Seiten.[93]

Ward Cunningham stellte die Software hinter WikiWikiWeb auf Anfrage anderen zur Verfügung und fand viele Nachahmer seiner Idee. Jeder, der in der Lage war einen Webserver mit CGI einzurichten (oder zumindest die Skriptsprache Perl auf seinem Server zu installieren), konnte sein eigenes Wiki aufmachen. Viele fügten Funktionen hinzu, die sie vermissten, oder betrieben eine Version in ihrer Lieblingsprogrammiersprache.[94]

Was charakterisiert nun ein Wiki? Es stellt eine ganz simple Methode zur Verfügung, um sich in den Informationen zu orientieren. Durch schnelles und einfaches kreuzweises Referenzieren werden die Benutzer ermutigt Konzepte miteinander zu verbinden, die nur „einen Klick weit" weg sind. Auch das Erstellen oder Ändern einer Seite ist nur „einen Klick" weit weg. Das Gestalten von Text ist mit einer wirklich simplen Auszeichnung möglich. Jeder auf der Welt kann alles ändern und alle Änderungen von anderen wieder rückgängig machen. Der Zugriff ist schnell, die Suche auch. Das kann einerseits benutzt werden um mehrere Diskussionsabläufe nebeneinander zu haben und eine strikte Reihenfolge der Diskussionen aufzubrechen. Andererseits ermöglicht es ein leichtes Verfassen von Beiträgen und damit eine leichte Zusammenarbeit bei der Zusammenstellung von Information.[95]

Schon 2001 war klar, dass das Wikikonzept funktioniert. Es ist allerdings bis heute nicht ganz klar, warum es funktioniert.[96] Eine mögliche Erklärung ist wohl, dass es mit so wenig Aufwand möglich ist, Informationen über Dinge, die einen interessieren, auf eine wohl strukturierte Art und Weise zu veröffentlichen. Manche soziale Interaktionsphänomene wurden allerdings auch erst durch die Bekanntheit und die Themenbreite von Wikipedia offensichtlich, etwa „Edit Wars", bei denen Anhänger verschiedener Ansichten über längere Zeit hinweg ständig die Änderungen ihrer Gegner wieder ausbessern.[97] Die meisten Beitragenden zu einem Wiki sind höflich zueinander. Wenn es jemand in einem Wiki gar nicht mehr aushält, dann arbeitet man an einem anderen Wiki mit. Wenn jemand einmal herausgefunden hat, dass die Inhalte des Wiki regelmäßig bei jeder Änderung gesichert werden, verliert mutwillige Beschädigung stark ihren Reiz. Die absolute Offenheit ermöglicht es jedem, der das will, sich um derartige Beschädigungen entsprechend zu kümmern. Eine Gemeinschaft- die Community - trägt sicher auch zum Verantwortungsbewusstsein für das gemeinsame Werk bei. Das kann man dadurch beschleunigen, dass man zu Beginn eines neuen Wikiprojekts ein gutes Grundgerüst zur Verfügung stellt. Die Art, wie geschrieben wird, wird in der Gemeinschaft der Autoren ausdiskutiert. Eine einmal getroffene Entscheidung bezüglich des Stils übt einen gewissen Druck auf neue Autoren aus, den vorgefundenen Stil nachzuahmen. Grundsätzlich kann man zwei Arten unterscheiden, wie in einem Wiki geschrieben wird: Eine Seite kann als ein Diskussionsverlauf gestaltet werden oder es kann sich um ein Dokument handeln. Ersteres zeichnet sich vor allem durch einen sprachnahen Stil aus, in dem viel in der ersten und zweiten Person geschrieben ist und in dem alle ernst zu nehmenden Diskussionsteilnehmer ihre Beiträge unterschreiben, etwa mit einem Link auf eine Seite, die sie persönlich beschreibt. Letzteres zeichnet sich üblicherweise dadurch aus, dass der Text in einem typischen Dokumentstil abgefasst ist,

[92] Ward Cunningham: WikiWikiWeb. http://c2.com/cgi/wiki?WikiWikiWeb (Zugriff am: 24. 11. 2012).
[93] o. A.: WikiWikiWeb - Wiki Page Counts. http://c2.com/cgi/wikiPages (Zugriff am: 24. 11. 2012).
[94] Leuf, Cunningham (2001), S. 24–29.
[95] Leuf, Cunningham (2001), S. 21.
[96] David Ludwig, Tobias Schumann: Wikipedistik. in: Catrin Schoneville (Hg.): *Alles über Wikipedia und die Menschen hinter der größten Enzyklopädie der Welt,* Hamburg, 2011, S. 225–241.
[97] Wikipedia contributors: Wikipedia:Edit-War. http://de.wikipedia.org/wiki/Wikipedia:Edit-War (Zugriff am: 24. 11. 2012).

häufig in der dritten Person unpersönlich und ohne Unterschrift. Solche Seiten werden dann auch als das Werk der gesamten Gemeinschaft verstanden.[98]

## Freie Enzyklopädie und WikiWikiWeb ergibt Wikipedia

Ein Jahr nachdem Richard Stallman seine Ideen veröffentlicht hatte, im Jänner 2001, setzten dann Jimmy Wales und Larry Sanger Wikipedia um. Ben Kovitz hatte Sanger die WikiWikiWeb Technologie nähergebracht.[99] Wales und Sanger hatten schon zuvor ein zentral überprüftes Modell einer Internetenzyklopädie gestartet, deren Name Nupedia war. Hier schrieben Experten, was durch einen sieben-stufigen Überprüfungsprozess abgesichert war. Dieses Lexikon wurde allerdings im September 2003 eingestellt, nachdem es innerhalb von drei Jahren lediglich 24 Artikel durch den gesamten Überprüfungsprozess geschafft hatten.

Wikipedia andererseits soll von jeder und jedem für jeden und jede sein. Dies kommt unter anderem in den fünf Säulen, den fünf grundlegenden Prinzipien für Wikipedia zum Ausdruck:

1. Wikipedia is an encyclopedia.
   It incorporates elements of general and specialized encyclopedias, almanacs, and gazetteers. Wikipedia is not a soapbox, an advertising platform, a vanity press, an experiment in anarchy or democracy, an indiscriminate collection of information, or a web directory. It is not a dictionary, a newspaper, or a collection of source documents; that kind of content should be contributed instead to the Wikimedia sister projects.
2. Wikipedia is written from a neutral point of view.
   We strive for articles that document and explain the major points of view in a balanced and impartial manner. We avoid advocacy and we characterize information and issues rather than debate them. [...] [We do] not presenting any point of view as "the truth" or "the best view". All articles must strive for verifiable accuracy: unreferenced material may be removed [...]. Editors' personal experiences, interpretations, or opinions do not belong here. [...]
3. Wikipedia is free content that anyone can edit, use, modify, and distribute.
   Respect copyright laws, and do not plagiarize sources. Non-free content is allowed under fair use, but strive to find free alternatives to any media or content that you wish to add to Wikipedia. Since all your contributions are freely licensed to the public [100], no editor owns any article; all of your contributions can and will be mercilessly edited and redistributed.
4. Editors should interact with each other in a respectful and civil manner.
   Respect and be polite to your fellow Wikipedians, even when you disagree. Apply Wikipedia etiquette, and avoid personal attacks. Find consensus, avoid edit wars [...]. Act in good faith [...]. Be open and welcoming, and assume good faith on the part of others. When conflict arises, discuss details on the talk page, and follow dispute resolution.
5. Wikipedia does not have firm rules.
   Rules in Wikipedia are not carved in stone, as their wording and interpretation are likely to change over time. The principles and spirit of Wikipedia's rules matter more than their literal wording [...].

[98] Leuf, Cunningham (2001), S. 322–327.
[99] Ben Kovitz: The conversation at the taco stand.
http://en.wikipedia.org/wiki/User:BenKovitz#The_conversation_at_the_taco_stand (Zugriff am: 17. 11. 2012).
[100] o. A.: Terms of Use. Licensing of Content.
http://wikimediafoundation.org/wiki/Terms_of_Use#7._Licensing_of_Content (Zugriff am: 08. 01. 2013).

> Be bold (but not reckless) in updating articles and do not worry about making mistakes. Prior versions of pages are saved, so any mistakes can be corrected. [101]

Die englische Wikipedia bestätigte mit ihrem Wachstum eine der Voraussagen von Richard Stallman:

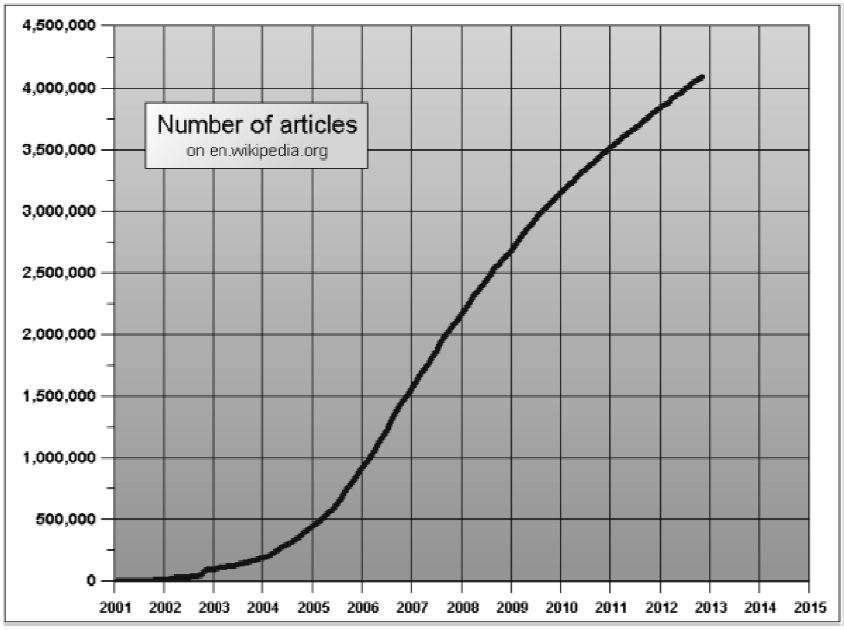



**Abbildung 4 Anzahl der Artikel in der englischen Wikipedia**

Larry Sanger startete 2007 einen zweiten Versuch einer Internetenzyklopädie mit einem formalisierten Überprüfungsprozess, bei dem Artikel an dessen Ende eine Art Gütesiegel erhalten und damit vor jenen Versionen angezeigt werden, die noch ungeprüft sind[102]. Sie hat es bis heute immerhin auf 16319 Artikel gebracht, aber lediglich 165 davon sind von Experten überprüft und bestätigt. Diese Zahlen, genauso wie einige weitere im Internet auffindbare Beispiele für Enzyklopädien, die viel Wert auf Überprüfung und Bestätigung legen, lassen vermuten, dass eine solche Vorgehensweise mit einem quälend langsamen Wachstum jedenfalls der geprüften Beiträge einhergeht.

Zuerst wurde Wikipedia von Wales Firma wikia.com, sie stellt gegen Entgelt Wikis für beliebige Zwecke zur Verfügung, betrieben. Er übergab aber Mitte 2003 alle Namens- und andere Rechte an eine Stiftung, die WikiMedia Foundation mit Sitz in Florida, USA, die heute alle unter „wikipedia.org“ erreichbaren Wikipedias betreibt.[103] Hinter der heutigen Wikipedia steht also mit der WikiMedia Foundation eine zentrale Organisation, die ihre Aufgabe aber hauptsächlich in der unentgeltlichen Bereitstellung von Infrastruktur hat und sich in die Erstellung von Artikeln nicht einmischt. [104]

## MediaWiki: Die Software, die Wikipedia ermöglicht

2001 startete Wikipedia noch mit einem damals frei unter der GPL verfügbaren Wiki-Klon namens UseModWiki, der in Perl geschrieben war. Dieser legte die Daten als einfache Textdateien ab, was dazu

[101] Wikipedia contributors: Wikipedia:Five pillars. http://en.wikipedia.org/wiki/Wikipedia:Five_pillars (Zugriff am: 17. 11. 2012).
[102] Zu finden unter http://en.citizendium.org/wiki/Welcome_to_Citizendium (Zugriff am 17.11.2012)
[103] Wikimedia Foundation: Board member. https://wikimediafoundation.org/wiki/Board_member (Zugriff am: 17. 11. 2012).
[104] Wikimedia Foundation: Bylaws - ARTICLE II - STATEMENT OF PURPOSE. https://wikimediafoundation.org/wiki/Wikimedia_Foundation_bylaws#ARTICLE_II_-_STATEMENT_OF_PURPOSE (Zugriff am: 17. 11. 2012).

beitrug, dass dieses System der wachsenden Menge an Anfragen schon Ende 2001 nicht mehr gerecht wurde. So musste eine Software erstellt werden, die besser mit der Anzahl der Anfragen an den Server zurechtkam. Markus Manske, damals an der Universität Köln, schrieb eine Wiki-Software, die nicht mehr Dateien nutzte, sondern die frei verfügbare Datenbank MySQL (eine der beliebtesten Datenbanken im Umfeld von dynamischen Websites, aber nicht unbedingt so sicher und zuverlässig wie Datenbanken im Unternehmensumfeld)[105]. Außerdem wechselte die Programmiersprache von Perl zu PHP. Die Wikipedia unter der ersten Software UseModWiki wurde mit Phase I bezeichnet, jene unter der neuen Software mit Phase II. Die Verbesserungen reichten aber, wie sich schon im Jahr 2002 herausstellte, nicht aus und so wurde mit den gewonnenen Erkenntnissen nochmals eine neue Software geschrieben, die mit Phase III bezeichnet wurde. Diese Software wurde ab Juli 2002 verwendet und brachte die Möglichkeit, Teile zu identifizieren, die zu viele Rechenzeit benötigten oder die Datenbank zu stark belasteten, was auch dringend notwendig war. Es war in der Folgezeit immer wieder notwendig, kurzfristig Änderungen an der Software durchzuführen, da Probleme durch zu viele Zugriffe Serviceeinschränkungen zur Folge hatten. Nur so konnte die Wikipedia überhaupt weiter angeboten werden. Trotzdem wurde die Software ab diesem Zeitpunkt nicht mehr radikal neu entwickelt, sondern Schritt für Schritt angepasst, wo Probleme auftraten und neue Funktionen gewünscht waren.[106]

Heute benutzen die Wikipedias die 20. größere Revision der Software, deren Entwicklung im Jahr 2002 begonnen hatte.[107] MediaWiki setzt nun sehr viele Ideen um, die schon Ward Cunninghams Ur-Wiki auszeichneten, macht aber auch einiges bewusst anders. Es ist sogar so, dass man, gibt man bei einer der bekanntesten Vergleichsseiten für Wikis die typische Merkmale von Wikipedia an, nur genau MediaWiki als Software genannt bekommt.[108] Diese Spezialitäten von MediaWiki sollen im Folgenden ein wenig genauer betrachtet werden:

- Unterstützung für Unicode: Schon sehr früh war klar, dass Wikipedia tatsächlich in vielen Sprachen verfügbar sein sollte und so ist es nur konsequent, dass die verwendete Software mit Unicode umgehen kann, also jedenfalls alle dort verzeichneten „Buchstaben“[109] enthalten kann. Heute umfasst der Unicode Standard 6.2 110.182 (http://en.wikipedia.org/wiki/Unicode) „Buchstaben“ was bis auf 17 x $2^{16}$ (mehr als 1.000.000) Zeichen ausgebaut werden kann.[110]

[105] Magnus Manske: MedieWiki - oder: Das Web 0.0. in: Catrin Schoneville (Hg.): *Alles über Wikipedia und die Menschen hinter der größten Enzyklopädie der Welt,* Hamburg, 2011, S. 283–295.. Er bezeichnete sie erst als phpwiki, allerdings gab es damals schon einen Wiki Klon mit diesem Namen, der allerdings nichts mit seinen Bemühungen zu tun hatte. Der Autor hätte gerne einen entsprechenden Link angegeben, aber der Benutzerfreundliche Zugriff auf das SVN Repository von PHPWiki bei sourceforge.net funktionierte zum Zeitpunkt der Erstellung dieses Absatzes nicht. Es konnte aber anhand einer lokalen Kopie des SVN Repository überprüft werden, dass sich der Code auch 2001 nicht im Geringsten ähnelte. So wurde die Software dann immer häufiger als das „PHP script“ bezeichnet bis jemand mit der „cleveren“ Idee vom Wikimedia <-> MediaWiki Wortspiel ankam. Im Nachhinein betrachtet eine Quelle endloser Verwirrungen.

[106] Mediawiki contributors: MediaWiki history. http://www.mediawiki.org/wiki/MediaWiki_history (Zugriff am: 25. 11. 2012).

[107] Man kann die aktuelle Version in der Quelltextansicht jeder Wikipedia Seite nachlesen. Dort ist etwa am 25. 11. 2012 als eines der ersten das Tag <meta name="generator" content="MediaWiki 1.21wmf4" /> zu finden. Die erste Dokumentierte Version von MediaWiki war 1.1.

[108] o. A.: WikiMatrix - Search for Wikis. Kein CamelCase. http://www.wikimatrix.org/search.php?sid=137647 (Zugriff am: 25. 11. 2012).

[109] Eigentlich: Code Points, Zeichen die irgendwo auf der Welt in Schriften vorkommen oder -kamen.

[110] Julie D. Allen, Deborah Anderson, Joe Becker, Richard Cook, Mark Davis, Peter Edberg, Michael Everson, Asmus Freytag, Jenkins. John H., Rick McGowan, Lisa Moore, Eric Muller, Addison Phillips, Michel Suignard, Ken Whistler: The Unicode Standard. Ch 2.: General Structure. http://www.unicode.org/versions/Unicode6.2.0/ch02.pdf (Zugriff am: 24. 11. 2012).

- Unterstützung für von rechts nach links geschriebene Sprachen: Eine tief greifende Unterstützung solcher Sprachen steht nicht bei vielen Wiki-Plattformen zur Verfügung. Auch in MediaWiki wurde diese Unterstützung noch 2011 weiter verbessert.[111]
- Die Art, wie man kommentieren kann: Man kann wie oben beschrieben in jedem Wiki kommentieren und diskutieren. Es hängt mehr davon ab, wie andere das Geschriebene auffassen. Die Art allerdings, wie Diskussionen in MediaWiki gehandhabt werden, ist wenig verbreitet. Jede Seite hat eine „Zwillingsseite" für Diskussionen über diese Seite.
- Bearbeiten einzelner Abschnitte eines Texts: Auch nicht weit verbreitet ist die Möglichkeit nur einen Abschnitt eines Texts zu bearbeiten. Dies ist allerdings bei längeren Artikeln, wie sie ja in Enzyklopädien durchaus öfter vorkommen ein Vorteil für den Autor. Man wird, wenn man will, dabei unterstützt den Überblick zu behalten.
- Die Möglichkeit Seiten, auf die keine Links verweisen, schnell zu finden: Diese Funktion erleichtert das Auffinden von „herrenlosen" oder „verwaisten" Seiten enorm. Eine Seite sollte zwar normalerweise immer von irgendeiner anderen referenziert sein, aber es kann auch vorkommen, dass irgendwann alle Referenzen gelöscht sind. Dann sind alle Mitwirkenden aufgefordert einzugreifen und die Situation zu verbessern. Eine Trennung in jene die Inhalte produzieren und jene die Wartungsarbeiten an Wikipedia durchführen existiert meist nicht.
- Am häufigsten bzw. am seltensten benutzte Seiten: Abgesehen davon, dass man mit dieser Funktion sehen kann, was gerade die Gemeinschaft bewegt, kann man diese Funktion auch dazu verwenden die eigenen Bemühungen um eine sichtbare Verbesserung der freien Enzyklopädie in die richtigen Bahnen zu lenken.
- Mathematische Formeln: Wikipedia enthält in vielen Sprachen eine exzellente Sammlung an mathematischen Formeln mit Erklärungen zu deren Funktion und Bedeutung. Dadurch lassen sich diese auch einfach aus Artikeln heraus referenzieren, die sich um konkrete Anwendungen drehen. Dies war wohl nur möglich, weil MediaWiki seit den frühesten Versionen eine relativ eingängige Möglichkeit zur Verfügung stellte Formeln aufzuschreiben, sodass sie dann in ihrer vollen mathematischen Gestalt dargestellt werden.
- Namensräume: Sie bilden einen Rahmen, in dem Seiten mit einem gewissen Titel mehrfach existieren können. Bei MediaWiki sind etwa die Diskussionsseiten unter demselben Namen wie die Seite aber eben im Namensraum „Diskussion" zu finden. Normalerweise wird eine Seite auf die man referenziert nur angelegt, wenn sie noch nicht existiert. Andere Namensräume sind etwa „Benutzer" für Angaben der Benutzer über sich oder „Spezial" für Seiten, die nicht von Menschen erstellt, sondern vom Server generiert werden. Ein MediaWiki-typischer Namensraum sind auch die sogenannten „Vorlagen".
- Vorlagen: In MediaWiki sind Vorlagen Teile einer Seite, die mit Platzhaltern einmal erstellt werden und dann vonseiten zu einem Thema bei Bedarf eingebunden werden, wobei der Autor die Platzhalter einsetzt. So kann etwa das Gerüst einer Infobox mit den wichtigsten Daten eines Landes mit eben diesen gefüllt und dann an der entsprechenden Position im Artikel über dieses Land angezeigt werden.
- XML Export[112]: Diese Funktion steht schon seit der Version 1.1 aus dem Jahr 2003 zur Verfügung, wenn sie auch immer wieder verändert und erweitert wurde. [113] Es ist in diesen Dateien auch die

[111] Wikipedia contributors: MediaWiki version history. http://en.wikipedia.org/wiki/MediaWiki_version_history (Zugriff am: 25. 11. 2012).
[112] Für XML siehe Seite 57
[113] Wikipedia contributors: MediaWiki version history. http://en.wikipedia.org/wiki/MediaWiki_version_history (Zugriff am: 25. 11. 2012).

gesamte Geschichte der Änderungen nachvollziehbar, wenn man bei einer Wikipedia in der Größe der Englischen über die entsprechenden Computer und Programme verfügt. Ein Nachteil Implementierung dieser Funktion, wie sie bis jetzt existiert ist, dass der Inhalt der Artikel direkt als Wikitext in der XML-Datei enthalten ist. Das macht es notwendig für die Arbeit mit dem strukturierten Inhalt diesen mittels einer anderen Software wieder als Struktur zugänglich zu machen. Ohne spezielle Software zu bemühen, kann man auch einfach alle Zeichen löschen, die für MediaWiki eine Bedeutung haben. Dadurch verliert man aber einiges an Information, die die Autoren der Artikel zur Verfügung stellen.

Nimmt man diese Kriterien, dann bleiben drei Wiki-Systeme über, die die oben genannten Funktionen unterstützen. MediaWiki, Wikia und DrupalWiki. Wikia ist, wie schon erwähnt die ehemalige Betreiberfirma von Wikipedia und DrupalWiki ist das Produkt einer deutschen Firma, die die Funktionalität von MediaWiki offensichtlich mit dem beliebten CMS (Content Management System, ein Verwaltungssystem für Websites) Drupal kombinieren will.[114]

- CamelCase für WikiWords: Dies ist eine der zentralen Ideen von Ward Cunningham, die es nochmals erleichtern sollte, die Inhalte eines Wikis miteinander zu verknüpfen. Die Software erkennt, was verknüpft werden soll, indem sie einfach nach einem Artikel im Wiki sucht, der dem in CamelCase geschriebenen Wort entspricht, wenn aus dem Titel alle Leerzeichen entfernt und alle Wörter mit großem Anfangsbuchstaben geschrieben werden. Allerdings würde eine solche Schreibweise – EineSolcheSchreibweise - in einer Enzyklopädie wohl fehl am Platze sein. Deshalb erlaubt MediaWiki solche Verknüpfungen nicht.

Bei all den Gemeinsamkeiten die Wiki-Systeme ihrem Konzept nach haben, hat sich eine Sache bis heute nicht zu einem Standard verdichtet. Die konkrete Schreibweise, mit der man in einem Wikiartikel Textbausteine wie Überschriften, Aufzählungen oder eben auch Links auf andere Seiten im Internet oder innerhalb des Wiki auf andere Artikel setzt, ist sehr verschieden. Allerdings hat MediaWiki dank seiner Größe so de facto etwas Ähnliches wie einen Standard etabliert.[115]

Zwei Eigenschaften von MediaWiki sind für die Verarbeitung der Texte der Wikipedia besonders wichtig:

1. Vorlagen: Sie sind selbst Wiki Seiten, die nur das Gerüst eines bestimmten Elements einer Seite enthalten. In dieses Gerüst füllt der Ersteller der eigentlichen Seite dann die konkreten Daten für seinen Artikel ein und bekommt ein einheitlich gestaltetes Element an der von ihm gewünschten Stelle in seinen Artikel eingebettet. Ein Beispiel wären die Infoboxen für Länder, Personen, etc. Die feststehenden Worte in diesen Vorlagen tauchen also effektiv so oft auf, wie die Vorlage verwendet wird.
2. Funktionen: Es gibt in der MediaWiki Software, die von der WikiMedia Foundation angepasst und eingesetzt wird, eine Unzahl von Funktionen, die der Ersteller eines Texts nutzen kann, wenn er möchte. Er kann etwa das aktuelle Datum einfügen, aber auch abhängig vom aktuellen Datum einen bestimmten Text. Da dies nur ein sehr einfaches Beispiel für die Möglichkeiten von Kombinationen ist und nur die MediaWiki Software der WikiMedia Foundation alle Funktionen

[114] o. A.: WikiMatrix - Search for Wikis. http://www.wikimatrix.org/search.php?sid=137646 (Zugriff am: 25. 11. 2012).
[115] o. A.: WikiMatrix. Feature Comparison. http://www.wikimatrix.org/compare/MediaWiki+UseMod+DokuWiki+MojoMojo+PmWiki+TWiki+Zwiki (Zugriff am: 25. 11. 2012).

aktuell auch ausführen kann, kann es dadurch zu Fehlern bei der Erstellung des Grundtexts kommen.

## Die Probleme von Wikitext

Seit 2008 bietet die MediaWiki Software im Prinzip die Möglichkeit alle Artikel, in XML und einigen anderen Darstellungen programmgesteuert abzurufen. Dies ist allerdings vor allem dazu gedacht, dass man mit auf dem lokalen Rechner installierten Programmen Artikel bearbeitet. Auch ist der Inhalt des Artikels keineswegs mittels XML strukturiert. Der MediaWiki parser ist zurzeit nicht fähig etwas anderes als HTML auszugeben. Es ist ferner nicht vorgesehen, dass man eine komplette Wikipedia so herunterlädt und WikiMedia behält sich auch vor, solche Versuche zu unterbinden.[116]

Der folgende Abschnitt bezieht sich in weiten Teilen auf Hannes Dohrn, Dirk Riehle (2011). MediaWiki ist dazu gedacht aus Wikitext HTML zu generieren und etwas andere Formate sind nicht vorgesehen. Um Wikitext unabhängig von MediaWiki zu interpretieren, entsprechend in reinen Text oder andere Formate umzusetzen und dabei die von den Autoren gelieferte Struktur zu erhalten und eventuell daraus einen Nutzen zu ziehen, ist also spezialisierte Software erforderlich. Für die gegenständliche Arbeit wurde dazu auf die frei verfügbare in Java geschriebene Software Sweble Wikitext Parser zurückgegriffen.[117] Dies vor allem deshalb, weil andere Werkzeuge, die zur Verarbeitung von Wikipedia Dumps verwendet wurden, auch in der Java-Umgebung ablaufen. So wurde die Einbindung der Parserfunktion für Wikitext wesentlich erleichtert. Es gibt noch eine Reihe anderer Werkzeuge, die Wikipedia aufführt.[118] Sweble verwandelt Wikitext in einem fünfstufigen Prozess in eine Baumstruktur, in der alle strukturellen Elemente einer Wikipedia Seite, wie man sie in einem Browser sehen kann, enthalten sind. In dieser Baumstruktur ist der semantische Gehalt des Wikitexts klar ersichtlich. Der Parser von MediaWiki, der historisch gewachsen ist, hat inzwischen eine Größe und Komplexität erreicht, die ihn nur noch schwer wartbar und erweiterbar machen. Dieser hatte niemals eine „Grammatik“ als Grundlage. Der in Java neu erstellte Parser Sweble versucht MediaWiki Wikitext aufgrund einer wohl definierten „Grammatik“ (einer „parsing expression grammar“) in eine Baumstruktur überzuführen, in der jedes Element eine Eltern-Kind- oder Geschwisterbeziehung zu seinen umliegenden Elementen steht. So ist dann ein Abschnitt ein Kind einer Seite und seine Kinder sind die Überschrift im Absatz und die Absätze des Abschnitts. Die Software Sweble enthält ein Modul, das eine Zugriffsmöglichkeit mittels XPath[119] bietet. So können auf standardisierte Art und Weise bestimmte Teile des so entstandenen Baumes ausgewählt werden und im besten Fall Informationen anhand eines definierten Zugriffspfads aus vielen gleichartigen Seiten extrahiert werden. Ein anderes Modul ermöglicht es, den entstandenen Baum als XML[120] auszugeben.

Das Programm hat Grenzen, was die genaue Nachbildung des Verhaltens der MediaWiki Software angeht: MediaWiki stellt eine nicht leicht erfassbare Menge an „functions, parser variables, tag extensions, templates, etc.“ bereit, „[i]n short, the whole context in which a page is rendered“[121] wird im

[116] Siehe http://en.wikipedia.org/w/api.php. Je nachdem für welche konkrete Wikipedia man Funktionen aufrufen will muss man die Subdomain anpassen. Also etwa http://arz.wikipedia.org/w/api.php. Interessant sind die beiden Funktion action=parse bzw. auch action=expandtemplates. Diese sind seit 2008 in MediaWiki eingebaut.
[117] Die Originalversion dieser Software kann unter http://www.sweble.org/ (Zugriff am 17. 01. 2013) bezogen werden.
[118] Für eine Übersicht siehe http://www.mediawiki.org/wiki/Alternative_parsers (Zugriff am 17. 01. 2013)
[119] Die Spezifikation befindet sich unter http://www.w3.org/TR/xpath/ (Zugriff am 17. 01. 2013).
[120] Siehe Seite 57.
[121] Hannes Dohrn, Dirk Riehle: Design and Implementation of the Sweble Wikitext Parser. Unlocking the Structured Data of Wikipedia. in: Andrea Forte (Hg.): *Proceedings of the 7th International Symposium on Wikis and Open Collaboration,* New York, NY, USA, 2011, S. 72–81, hier S. 81.

ungünstigsten Fall benötigt, um auf genau dieselbe Darstellung zu kommen, wie sie eine Wikipedia anbietet. Auch ist es aufgrund der Komplexität des Parsers der MediaWiki Software nur möglich das Ergebnis eines Parsevorgangs zu ermitteln, wenn man eben den Wikitext mit exakt jener MediaWiki Instanz bearbeitet für die er geschrieben wurde, also etwa Texte der Wikipedia Masry mit der Software auf arz.wikipedia.org. Tatsächlich kann man durch Aufruf von speziellen Seiten, die nicht für menschliche Leser, sondern für andere Programme bestimmt sind (Application Programming Interface, API), einige der benötigten Informationen erhalten. Zum Teil ist es auch nicht sinnvoll genau dasselbe zu produzieren wie MediaWiki, da dies ungültiges HTML wäre, das ein Browser trotzdem darzustellen versucht. Dies gelingt auch meistens, da Browser aus historischen Gründen nie sehr empfindlich auf falsches HTML reagieren durften.

Will man also den Inhalt der Wikipedia weiterverarbeiten, muss man den Wikitext in ein besser spezifiziertes Zwischenformat oder ein Exchange-Format überführen. Eine Möglichkeit für ein solches Format ist ein Abstract Syntax Tree (AST) wie er auch für Programmiersprachen erstellt werden kann, um sie dann in Maschinencode zu übertragen, ein weiterer Vorschlag ist XML, konkret etwa XML Wikitext Markup Language (XWML)[122]. Der Sweble Wikitext Parser verwendet als interne Repräsentation einen AST. Aufgrund der oben angegebenen Möglichkeit eine Fülle von Erweiterungsfunktionen in MediaWiki einzubauen, ist es allerdings nicht zu erwarten, dass jemals garantiert werden kann, dass ein Ersatz für den eingebauten MediaWiki-Parser wie Sweble, sich exakt so verhält wie das Original. Trotzdem wird bei Sweble versucht jene Fälle, in denen sich Sweble anders verhält als sein Vorbild, auf ein Minimum zu begrenzen. Eine mögliche Anwendung des Sweble Parsers ist zu versuchen Texte in MediaWiki Wikitext in einem dokumentierten Baumformat auszugeben, das in XWML repräsentiert ist. Dies ergibt dann eine ähnliche Repräsentation wie sie moderne Browser aus HTML oder XHTML generieren und die als Dokument Object Model, DOM, bekannt ist. Über diese Repräsentation kann man in Browsern den aktuellen Zustand und Aufbau einer geladenen Webpage abfragen und manipulieren. Abfragen an ein als Wiki Object Model, WOM, geladenes XWML oder Wikitext Dokument wären ebenfalls auf eine standardisierte Art und Weise möglich.[123]

Die Idee eines WOM, das einen einfacheren Zugriff und eine einfachere Möglichkeit für Änderungen an Wikipedia Artikeln ermöglicht, ist unabhängig davon als eine Erweiterung für die MediaWiki Software verfügbar.[124] Diese Erweiterung hat allerdings noch einige weitere Einschränkungen verglichen mit Sweble und ist bei der Wikipedia Masry nicht verfügbar.[125]

Das Softwaremodul, mit dem die von Sweble gelesenen Wikitext Artikel als WOM repräsentiert und als XWML ausgegeben werden können, wurde erst während dieser Arbeit verfügbar gemacht.

## Wikipedia in anderen Sprachen als Englisch

Schon bald nach Projektbeginn äußerte Jimmy Wales selbst den Wunsch, auch andere Sprachen zu unterstützen. Unter den ersten nicht-englischsprachigen Versionen war, da Wales dies selbst im März 2001 vorschlug, die deutschsprachige Wikipedia.[126] Nach dem Übergang zur WikiMedia Foundation wurde die Aufnahme neuer Sprachen formalisiert. Im Jänner 2006 wurde ein Special Projects Commitee

[122] Hannes Dohrn, Dirk Riehle: WOM: An object model for Wikitext. http://sweble.org/downloads/wom-tr.pdf (Zugriff am: 27. 01. 2013).
[123] Dohrn, Riehle (2011).
[124] Siehe http://www.mediawiki.org/wiki/Extension:Wiki_Object_Model (Zugriff am: 27. 01. 2013).
[125] Alle verfügbaren Erweiterungen und die aktuell eingesetzte Version der MediaWiki Software sind unter http://arz.wikipedia.org/wiki/Special:Version (Zugriff am: 27. 01. 2013) verfügbar.
[126] Jimmy Wales: Alternative language wikipedias. http://lists.wikimedia.org/pipermail/wikipedia-l/2001-March/000048.html (Zugriff am: 17. 11. 2012).

gegründete, welches im August desselben Jahres das New Languages Subcommittee gründete. Während das Special Projects Commitee 2007 das letzte Mal aktiv wurde und 2009 aufgelöst wurde, ist das New Language Subcommittee als Language Commitee heute für die Aufnahme neuer Sprachen zuständig.[127]

Um eine Wikipedia in einer noch nicht vorhandenen Sprache aufzunehmen, muss der WikiMedia Foundation ein entsprechender Vorschlag gemacht werden. Dieser Vorschlag wird nach der „Language proposal policy" beurteilt. Darin sind folgenden Kriterien für die Zulässigkeit eines Vorschlags aufgestellt:

1) Für die Sprache darf noch kein Projekt existieren, was anhand entsprechender Übersichtsseiten über vorhandene Wikimediaprojekte festgestellt werden kann.
2) „Der Sprache muss eine Abkürzung nach ISO 639 1 – 3 zugewiesen sein. Wenn das noch nicht der Fall ist, [was bedauerlicherweise immer noch für viele linguistische Varietäten gilt] dann muss man sich bemühen, eine solche Abkürzung zu bekommen. […]"[128]]
3) „Die Sprache sollte einigermaßen einzigartig sein, jedenfalls so einzigartig, dass Artikel, die darin verfasst wurden, nicht in einer anderen Sprachversion koexistieren können. […]" [Hier wird explizit darauf verwiesen, dass dies wohl für Dialekte oft gelte. Andererseits wird von Fall zu Fall geprüft, ob eine Koexistenz wirklich sinnvoll möglich ist. Die heute verfügbare Anzahl von Wikipedias in Varietäten spricht allerdings dafür, dass dieses Ausschlusskriterium, in der liberalen Art wie es für Wikipedia typisch ist, in den meisten Fällen nicht angewandt wird.]
4) Weiters soll versucht werden zu belegen, dass der Vorschlag auf eine genügend große Anzahl von potenziellen Beitragenden zurückgreifen kann, deren Erstsprache diese Sprache ist, um eine existenz-, überlebens- und entwicklungsfähige Gemeinschaft von Verfassern und Lesern zu gewährleisten. Es sind auch Projekte von WikiMedia denkbar, die Sprachen abdecken, die niemandes Erstsprache sind. [Konkret wird das Beispiel Esperanto genannt. In einem solchen Fall wird über eine Aufnahme nach einem Diskussionsprozess entschieden.][129]

Um schließlich die Annahme eines neuen Wikimediaprojekts beantragen zu können, müssen folgende Kriterien erfüllt sein:

1) Es muss einen erfolgreichen Testlauf gegeben haben. WikiMedia stellt dafür eine eigene Infrastruktur zur Verfügung, den Incubator. Hier können alle Typen von Wikimediaprojekten schnell erstellt werden und in kleinem Rahmen auf die Veröffentlichung als offizielles Projekt vorbereitet werden.
2) Es müssen schon dort die wichtigsten Teile der Benutzeroberfläche in die jeweilige Sprache übertragen werden. Zu diesem Zweck gibt es eine Liste jener 500 Texte, die Benutzer eines

[127] Wikimedia Foundation: Special projects committee. http://meta.wikimedia.org/wiki/Special_Projects_Committee (Zugriff am: 17. 11. 2012).

[128] Siehe dazu unten „Zur Frage der Identifikation linguistischer Varietäten". Dies mag erst einmal nach einer hohen Hürde klingen. Da aber auch auf ISO 639 – 3 abgestellt wird stellt sich das dann ganz anders dar. Die ISO stellt hier nur den grundlegenden Prozess sicher und delegiert die eigentliche Durchführung an mindestens drei „authorities". SIL, das ISO 639 – 3 verwaltet, versucht aufgrund ihrer Glaubensansichten möglichst vielen Varietäten einen Code zuzuweisen. Alleine 2011 wurden 64 neue Sprachen aufgenommen (siehe Melinda Lyons, SIL International: ISO 639-3 Change Requests Series 2011 Summary of Outcomes. http://www.sil.org/iso639-3/cr_files/639-3_ChangeRequests_2011_Summary.pdf (Zugriff am: 11. 12. 2012).) Dementsprechend ist allerdings auch das Gewicht dieses Standards. Allerdings geht es Wikimedia wohl vor allem um eine halbwegs sinnvolle und standardisierte Namensgebung der Subdomains für die einzelnen Projekte.

[129] Wikimedia Meta Contributors: Language proposal policy. http://meta.wikimedia.org/wiki/Language_proposal_policy (Zugriff am: 18. 11. 2012).: Requisites for eligibility, übersetzt vom Autor

> Projekts, das die Software hinter allen Wikimediaprojekten, MediaWiki, nutzt, darunter auch eine Wikipedia, am häufigsten zu sehen bekommen. Die Liste beruht auf einer Statistik, die nicht auf die Häufigkeit der Verwendung durch Autoren abstellt, sondern auf dem, was Benutzer tatsächlich zu einem bestimmten Testzeitraum in der englischen Wikipedia zu sehen bekamen. Es geht hier um jene Texte, auf die Autoren in der Wikipedia keinen Einfluss haben, da sie im umgebenden „Gerüst" der Seiten zu finden sind. Dazu wurden in einem definierten Zeitraum die Ergebnisse konkreter Anfragen an die WikiMedia Server ausgewertet. Die Zählung wurde seit der Einführung dieser Technik im Jahr 2007 zweimal, 2009 und 2011, wiederholt. 2007 wurde nur die englische Wikipedia berücksichtigt, die anderen beiden Male wurden zusätzlich die deutschsprachige Wikipedia und alle anderen Wikis der Wikimediaprojekte berücksichtigt.[130] Die Übersetzung erfolgt mithilfe von Translatewiki, einem eigenständigen europäischen Übersetzungsprojekt und Übersetzungswerkzeugprojekt, das von der WikiMedia Foundation genutzt und ein wenig unterstützt[131] wird. [132]

2003 gab es eine Diskussion über die Zulässigkeit von „Dialekten" als eigene Sprachvarianten. Grund dafür war die Anfrage des Elsässers Alexis Dufrenoy, ob man nicht eine elsässische Wikipedia aufmachen könnte.[133] Jimmy Wales sagt keinen halben Tag später: Ja, so schnell als möglich.[134] Die Wikipedia wurde noch im November gegründet, aber nachdem sie ein Jahr später immer noch sehr wenig enthieltS wurde sie als alemannische Wikipedia für westoberdeutsche Dialekte (Elsässisch, Schwyzerdütsch und andere[135]) und Sprachen weitergeführt[136].

Am 30. März 2008 reichte Benutzer „Ghaly" einen Antrag auf Überprüfung der Zulässigkeit einer Wikipedia in ägyptischem Umgangsarabisch ein. Schon einen Tag später am 1. April war klar, dass dieser Antrag zulässig war.[137] Die Wikipedia war nur relativ kurz im Incubator. Das Projekt wurde im Juni 2008 dort angelegt und im Juli 2008 wurde es offiziell genehmigt. [138] Währenddessen wurde eine Diskussion darüber geführt, ob dieses Projekt offiziell angelegt werden sollte. Schon im November 2008 wurde die

---

[130] o. A.: Most often used messages in MediaWiki.
http://translatewiki.net/wiki/Most_often_used_messages_in_MediaWiki (Zugriff am: 18. 11. 2012).
[131] o. A.: Project:About. http://translatewiki.net/wiki/Project:About (Zugriff am: 18. 11. 2012).
[132] Wikimedia Meta Contributors: Language proposal policy.
http://meta.wikimedia.org/wiki/Language_proposal_policy (Zugriff am: 18. 11. 2012).: Requisites for final approval, übersetzt vom Autor
[133] Alexis Dufrenoy: Alsatian Wikipedia.
http://comments.gmane.org/gmane.science.linguistics.wikipedia.international/2338 (Zugriff am: 25. 11. 2012).
[134] Jimmy Wales: Re: Re: Intlwiki-l Alsatian.
http://permalink.gmane.org/gmane.science.linguistics.wikipedia.international/2375 (Zugriff am: 25. 11. 2012).
[135] Badisch mit Sprachgebieten in Bayern, Vorarlbergisch mit Sprachgebieten in Tirol, Liechtensteinisch, Sprachgebiete im Piemont
[136] Wikipedia contributors: D Alemannisch Wikipedia.
http://als.wikipedia.org/w/index.php?title=Wikipedia&stable=0#D_Alemannisch_Wikipedia (Zugriff am: 25. 11. 2012).
[137] Vgl. Wikimedia Meta Contributors: Requests for new languages/Wikipedia Egyptian Arabic. 30. März 2008.
http://meta.wikimedia.org/w/index.php?title=Requests_for_new_languages/Wikipedia_Egyptian_Arabic&oldid=937275 (Zugriff am: 18. 11. 2012). und Wikimedia Meta Contributors: Requests for new languages/Wikipedia Egyptian Arabic. http://meta.wikimedia.org/wiki/Requests_for_new_languages/Wikipedia_Egyptian_Arabic (Zugriff am: 13. 12. 2012).
[138] Wikimedia Meta Contributors: Requests for new languages.
http://meta.wikimedia.org/w/index.php?title=Requests_for_new_languages&oldid=1096985 (Zugriff am: 18. 11. 2012).. Sieht man sich mithilfe der Versionsgeschichte dieser Seite die Seite in ihren jeweiligen früheren Versionen an so findet man dort vor der Genehmigung mit 17. Juli 2008, eingetragen am 23. Juli 2008, einen Hinweis auf eine Prüfung der Zulässigkeit im März 2007. Dies ist aber mit hoher Wahrscheinlichkeit ein Fehler.

ägyptisch-arabische Wikipedia als neue offizielle Wikipedia erstellt und ist seit dem unter der Subdomain arz.wikipedia.org erreichbar.[139]

## Die Diskussion über die Legitimität und Nützlichkeit der Wikipedia in ägyptischem Umgangsarabisch

In allen Literaturgattungen wird das Kairenische schon längere Zeit als Stilmittel eingesetzt[140], aber was die Vermittlung von Wissen angeht, wie man es von einer Enzyklopädie erwartet, so ist die Wikipedia Masry im arabischen Umfeld bisher einzigartig. So stand am Beginn des Wikipedia Masry Projekts eine Diskussion, die hier in Auszügen kurz besprochen werden soll.

„Ghaly" eröffnete die englischsprachige Diskussion im Ende März 2008[141] um die Zulässigkeit einer eigenen Wikipedia mit der Feststellung, dass es einen ISO 639 – 2 und einen ISO 639 – 3 Code für ägyptisches Arabisch gäbe[142] und dass man schon zu diesem Zeitpunkt einfach wegen der Bevölkerungszahl Ägyptens von 76 Millionen Sprechern ausgehen muss. Er bringt einen Vergleich mit der Wikipedia für „Simple English". Später ergänzt er aber, dass das Bild, das durch das letzte Argument entstanden ist, wohl nicht von ihm beabsichtigt wurde. Er möchte also keine simplifizierte Version der arabischen Wikipedia, sondern eine eigenständige Enzyklopädie, die sehr wohl auch kompliziert sein darf.

Er vertrat in der Diskussion auch immer die etablierte Meinung, dass ägyptisches Arabisch eine semitische Sprache aus dem afro-asiatischen Zweig dieser Gruppe ist. Er verortet den Ursprung der Sprache im Nildelta in Unterägypten rund um die Hauptstadt Kairo. Es stamme ab vom gesprochenen Arabisch, das während der arabischen Eroberungen im 7. Jahrhundert nach Ägypten kam. Er sieht Einflüsse vom Koptischen, das damals dort die gesprochene Sprache war und spätere Einflüsse aus dem Türkischen. Er erwähnt auch, dass aufgrund der Medienlandschaft, im arabischsprachigen Raum, in der Ägypten zumindest aufgrund seiner schieren Größe eine herausragende Position einnimmt, das ägyptisch Arabische in sehr weiten Teilen der arabischen Welt verstanden wird. Abgesehen davon erwähnte er, dass es eine der am besten untersuchten Varietäten des Arabischen ist.[143]

Die Darstellung der Diskussion im Internet ist nicht ausgewogen, da die Argumente dafür genauso wie die Widerlegung einiger Gegenargumente als Erstes kommen, wodurch die Gegenargumente immer weiter nach unten wandern. Bedenkt man allerdings die liberalen Grundprinzipien, dann mag das nicht zwangsläufig ein Problem sein. Es ging in der Diskussion auch nicht um mehr als die Erstellung eines MediaWiki unter einer Subdomain von Wikipedia und damit um eine Möglichkeit die Idee einer ägyptischen Wikipedia einmal breiter bekannt zu machen.

Die Diskussion verlief, wie sie schon öfters verlaufen ist. In einer Domäne, die bisher dem Hocharabischen vorbehalten war, sei das nun Drama, Prosaliteratur oder Zeitungsartikel, fühlte ein

[139] Incubator contributors: Revision history of "Wp/arz". http://incubator.wikimedia.org/w/index.php?title=Wp/arz&action=history (Zugriff am: 18. 11. 2012).
[140] Siehe dazu Somekh (1991). Für jüngere Entwicklungen etwa Gisela Kitzler: Von Taxifahrern und Heiratskandidaten - zwei moderne ägyptische Bestseller im Kontext des Schreibens im Dialekt ("ʿāmmiyya"). Eine interdisziplinäre Analyse der Texte "Ich will heiraten" von Ġāda ʿAbd al-ʿĀl und "Taxi" von Ḫālid al-Ḫamīsī, 2012.
[141] Wikimedia Meta Contributors: Requests for new languages/Wikipedia Egyptian Arabic. 1. April 2008. http://meta.wikimedia.org/w/index.php?title=Requests_for_new_languages/Wikipedia_Egyptian_Arabic&oldid=939972 (Zugriff am: 18. 11. 2012).
[142] Es gibt keinen ISO 639 – 2 Code für ägyptisches Umgangsarabisch.
[143] Wikimedia Meta Contributors: Requests for new languages/Wikipedia Egyptian Arabic. http://meta.wikimedia.org/wiki/Requests_for_new_languages/Wikipedia_Egyptian_Arabic (Zugriff am: 13. 12. 2012).

Autor eine Notwendigkeit in Teilen oder in einer gesamten Publikation die Umgangssprache zu benutzen. Die Gründe mögen unterschiedlich sein, meist haben sie aber mit dem Ansprechen von Schichten zu tun, die das Hocharabischen nicht mehr oder nicht mehr gut genug beherrschen, da sie es nach der Schullaufbahn nicht mehr verwende. Damit sehen Autoren die Gefahr, dass die Botschaft, die sie vermitteln wollen, nicht mehr von denjenigen verstanden wird, die sie ansprechen wollen. Oder es hat damit zu tun, dass es offensichtlich ist, dass im alltäglichen Leben niemand Hocharabisch in normalen Gesprächssituationen benutzt. Beispiele sind die frühen satirischen Zeitungen, deren Botschaft für alle von Interesse hätte sein sollen, nur nicht für jene gebildeten herrschenden Eliten, die Hocharabisch sicher verstanden hätten, oder die sozialkritischen Dramen aus der Zeit des Nasserismus, die das „wahre Leben" darstellen sollten.[144]

Daher sollen hier auszugsweise nur einige wenige Beispiele aus der Diskussion um die Zulässigkeit einer ägyptisch arabischen Wikipedia wiedergegeben werden. Dies soll lediglich verdeutlichen, wie tief diese Debatte in der ägyptischen Kultur verankert ist und dass auch jene interessierten Laien, die die Wikipedia schreiben, durch den entsprechenden Diskurs geprägt sind und so mühelos aber unreflektiert jene Stereotypen reproduzieren, die seit 200 Jahren verwendet werden.

Eine starke persönliche Motivation des Erfinders der ägyptischen Wikipedia „Ghaly" war, dass er manche Artikel in der hocharabischen Wikipedia, in der er auch mitgearbeitet hatte, nicht mehr verstand. Er führt das zum Teil darauf zurück, dass in Ägypten auch ein anderes Hocharabisch verwendet wird als in anderen Teilen der arabischen Welt.

[144] siehe auch Yasir Suleiman: A War of words. language and conflict in the Middle East. New York, NY, USA, 2004. und Manfred Woidich: Von der wörtlichen Rede zur Sachprosa: Zur Entwicklung der Ägyptisch-Arabischen Dialektliteratur. http://www.opus.ub.uni-erlangen.de/opus/volltexte/2010/2199/pdf/04_Woidich_Aegyptisch_Arabische_Dialektliteratur.pdf (Zugriff am: 13.12.2012).

| pro: | contra: |
| --- | --- |
| Manches in modernem Standard-Arabisch (MSA) ist für Ägypter nur mehr schwer verständlich. Zeitungen und Literatur in Ägypten benutzen ein anderes Vokabular. Auch medizinische Artikel oder medizinisches Vokabular im Allgemeinen sind auf Hocharabisch nur schwer verständlich, weil sich andere Worte eingebürgert haben. Wenn ein Fachmann in Ägypten, also z. B. ein Arzt, ein Zahnarzt oder ein Apotheker, mit einem Patienten spricht, dann versucht er sich so auszudrücken, dass der Patient das versteht, also benutzt er oder sie ägyptisches Arabisch. Das ist Teil der Ausbildung. Niemand kann behaupten, ägyptisches Arabisch sei zu beschränkt, um auch komplizierte Sachverhalte auszudrücken. Dass Hocharabisch die offizielle Sprache in Ägypten ist, ist kein Argument gegen ägyptisches Arabisch. In Indien gibt es ja auch einen Unterschied zwischen der offiziellen Sprache und der Umgangssprache. | Fachartikel sind immer schwer zu verstehen und die „ägyptischen“ Bezeichnungen sind einfach falsche Verwendungen unterschiedlicher Fachbegriffe. Die meisten Fachbegriffe im ägyptischen Arabisch sind Hocharabisch. Ägyptisches Arabisch ist zu beschränkt um komplizierte Sachverhalte auszudrücken, deshalb sind etwa gute Romane auf hocharabisch abgefasst. Wenn das offizielle Arabisch in Ägypten anders ist als MSA, was hilft es dann, Informationen in einer nochmals anderen Sprache zur Verfügung zu stellen? Nur die Hocharabische Sprache ist offizielle Sprache in Ägypten, alle Regierungsmitteilungen erfolgen auf hocharabisch ebenso wie Pressekonferenzen. Die Nationalhymne ist Hocharabisch. Es gibt keinen signifikanten Unterschied zwischen ägyptischem Fachjargon und jenem in klassischem Arabisch, so etwas müsste erst nachgewiesen werden. [145] |

Mit diesen Argumenten im Zusammenhang steht eine Beobachtung, die ein User namens „Mamduh“ in die Diskussion eingebracht hatte. Diese ist in der ägyptischen Wikipedia bis heute aber nicht berücksichtig. Er schrieb:

> The Arabic alphabet tends to mislead many to think that Egyptian and Arabic are one language, because Arabic doesn't have vowels, but instead uses "tashkeel", so because many of the differences between Arabic and Egyptian are "vowel-based", it is very hard to "see" those differences when we write down Egyptian in the Arabic alphabet. Let me elaborate on the examples used by C2: **سمحله** in Egyptian is pronounced "samahlo", while **سمح له** in Arabic is pronounced "samaha lahu", the same case with **اتوجت** (ittawwigit) in Egyptian versus **توجت** (tuwwijat) in Arabic, and many other words that are written the same way but pronounced very differently.

Deshalb wollte User „AnonMoos” das Projekt verschieben:

> Mamduh -- this difference between the pronunciations ittawwigit and tuwwijat etc. is relevant to the issue of a standardized Egyptian colloquial Arabic orthography which I raised previously. If the way you write Egyptian colloquial Arabic doesn't adequately indicate the special Egyptian colloquial pronunciation features of ittawwigit etc., then it will just look like a failed illiterate attempt to write the standard written Arabic form. I think it might be best to first focus on developing a form of orthography with some clear independence from Classical Arabic spelling conventions (similar to Nizar Habash's "Palestinian Arabic Spelling Standardization Project"), because trying to write Egyptian colloquial Arabic in an ad-hoc seat-of-

[145] Wikimedia Meta Contributors: Requests for new languages/Wikipedia Egyptian Arabic. http://meta.wikimedia.org/wiki/Requests_for_new_languages/Wikipedia_Egyptian_Arabic (Zugriff am: 13. 12. 2012).

> the-pants way with Classical Arabic spelling conventions seems to involve endless difficult issues, which can never really be satisfactorily resolved in terms of Classical Arabic spelling conventions.

Diesen Umstand machen sich seit Langem Autoren mit einer Abneigung gegen die geschriebene Umgangssprache zunutze, indem sie Sätze derart konstruieren, dass es von der nicht gedruckten Vokalisierung abhängt, ob dieser Satz hocharabisch oder umgangssprachlich verstanden wird.[146]

Ein weiteres Diskussionsgebiet betrifft die Lese- und Schreibfähigkeiten, die jeder Ägypter nach diversen offiziellen Statistiken hat oder haben sollte und welche Sprache in der Literatur zum Einsatz kommt:

| pro: | contra: |
|---|---|
| Es gibt schon ägyptisch arabische Literatur in der Mamlukenzeit, dann mehr im 19. Jahrhundert. Der erste in Ägypten erschienene Roman ist in ägyptischem Arabisch verfasst. Man findet in der Zeitung ganze Artikel in ägyptischem Arabisch. Ägyptisch ist die am meisten unterrichtete arabische Varietät weltweit. Modernes Standardarabisch und ägyptisches Umgangsarabisch wird wegen des Fehlens von Vokalzeichen oft für ähnlich gehalten. Ägyptisches Arabisch schreiben zu wollen ist kein Anzeichen für mangelnde Bildung. Als Ägypter kann man Hocharabisch schon richtig hassen. Es ist mühsam und kompliziert zu schreiben. | Selbst dieser erste in Ägypten erschienene Roman ist auf Hocharabisch abgefasst. Man muss sich nur die ersten paar Zeilen ansehen. Wenn jemand in der Zeitung ägyptisches Arabisch verwendet, dann nur irgendwelche Zuarbeiter. Das sind alles nur unbedeutende Fragmente. Es gibt keine ernst zu nehmende Literatur in ägyptischem Arabisch, nur eine Handvoll Gedichte und Romane. Das Schreiben von ägyptischem Arabisch propagieren nur Außenstehende mit dubiosen Absichten. Ägyptisches Arabisch ist einfach nur komisch. Man verwendet es höchstens in kabarettistischen Blogs. Es passt gut zu Scherzprojekten wie der Uncyclopedia[147]. Es gibt niemanden in Ägypten, der hocharabische Literatur nicht versteht. Jeder 6-jährige Erstklässler am Ende des Schuljahres versteht Hocharabisch.[148] |

In einer derart öffentlichen Diskussion auf einem so exponierten Feld sollte man wohl jedes Wort auf die Goldwaage legen. Eine Aussage wie:

> Indeed the first piece of modern Egyptian literature [Zaynab] was written in Masri.[149]

ist nicht dazu angetan Menschen mit der festen Vorstellung, ägyptisches Arabisch sei schlicht zu beschränkt für einen Roman, zu überzeugen. In der Entstehungszeit dieses Romans war es selbstverständlich, dass die Handlung in Hocharabisch verfasst ist, ebenso die Dialoge der gebildeten Leute oder Studenten. Andererseits schreibt er für die Sprache auf dem Dorf eben Umgangssprache nieder. Damit findet man die erste geschriebene Umgangssprache in einem anerkannten Werk der

---

[146] Somekh (1991), S. 27.

[147] o. A.: الصفحة الرئيسية.
http://beidipedia.wikia.com/wiki/%D8%A7%D9%84%D8%B5%D9%81%D8%AD%D8%A9_%D8%A7%D9%84%D8%B1%D8%A6%D9%8A%D8%B3%D9%8A%D8%A9 (Zugriff am: 28. 11. 2012).

[148] Wikimedia Meta Contributors: Requests for new languages/Wikipedia Egyptian Arabic. http://meta.wikimedia.org/wiki/Requests_for_new_languages/Wikipedia_Egyptian_Arabic (Zugriff am: 13. 12. 2012).

[149] Wikimedia Meta Contributors: Requests for new languages/Wikipedia Egyptian Arabic. http://meta.wikimedia.org/wiki/Requests_for_new_languages/Wikipedia_Egyptian_Arabic (Zugriff am: 13. 12. 2012).

ägyptischen Literatur schon 1913[150]. Tatsache ist auch, dass die Umgangssprache lange Zeit vor allem in komischen oder tragikomischen Stücken gepflegt wurde. Dies und auch die regelmäßige Verwendung in Karikaturen in Zeitungen fördert natürlich die gedankliche Verbindung zwischen Dialekt und komischen Situationen.[151]

Dass es Ägypter gibt, die Hocharabisch regelrecht hassen, hängt nicht zuletzt mit dem Schulsystem zusammen. Dieses ist derart unterdotiert, dass ein sinnvoller Unterricht für alle schon lange nicht mehr möglich ist. Eine Untersuchung von Niloofar Haeri (1997) Anfang der 1990er Jahre zeigte schon damals, dass diejenigen, die es sich leisten können, auf Schulen gehen, in denen Hocharabisch nicht einmal formal die Unterrichtssprache ist, und selbst wenn der Unterricht formal in Hocharabisch gehalten wird, sind häufig alle Erklärungen in der Umgangssprache. So ist es nur logisch, dass mit fortschreitender Ausbildung in allen Bereichen außer den religiösen Wissenschaften die Fähigkeiten, was Hocharabisch angeht, zunehmend verkümmern. Die Mehrheit der Menschen in Ägypten wird im Laufe ihres Lebens langsam zu „Illiterates", nicht in dem Sinne, dass sie nicht mehr lesen könnten, aber insofern, als Sinn erfassendes Lesen von beliebigen hocharabischen Texten zunehmend schwierig wird und manche es zu hassen beginnen. [152]

Ein weites Gebiet für Diskussionen ist auch die Frage wie, wenn überhaupt, man ägyptisches Arabisch aufschreiben kann:

| pro: | contra: |
|---|---|
| Es gibt einige Leute, die ägyptisches Arabisch schreiben, etwa in persönlichen Briefen (und persönlichen Beleidigungen) oder in ihren eigenen Notizen. Ägyptisches Arabisch ist eine lebendige Sprache, die viel verwendet wird. Die für die ägyptische Wikipedia vorgeschlagene Schreibweise ist offensichtlich von vielen Leuten entzifferbar. Es gibt eine deutsche Studie über den Schreibstil von Saʿad ad-Dīn Wahba. Auf dem Stil, den auch er pflegte, baut die Verschriftung des ägyptisch Arabischen auf. Argumente gegen ägyptisches Arabisch basieren auf irrationalen Annahmen und ideologischen Motiven. Bei der Beurteilung von ägyptischem Arabisch fehlt oft das Wissen über die einfachsten linguistischen Grundlagen. | Viele Leute schreiben auch ihre privaten Briefe in ordentlichem Hocharabisch. Es gibt keine standardisierte Rechtschreibung für ägyptisches Arabisch im Gegensatz zu Hocharabisch. Jeder der Dialekte aufschreibt, erfindet seine eigenen Regeln. Ägyptisch ist eine Untersprache, ein Dialekt, Slang von Arabisch. Oder Ägyptisch ist einfach ein Akzent des Arabischen. Die Ägypter selbst bezeichnen ihren Slang auch nicht als Sprache, bestenfalls als Dialekt. Es gäbe ja Varietäten von Arabisch, die schon eigenständige Sprachen wären, etwa Darija in Marokko, aber sicher nicht ägyptisches Arabisch.[153] |

Bei den Gegenargumenten wurde schon in der Diskussion unter anderem angemerkt, dass gewisse Zuordnungen arabischer zu englischen Begriffen mittels Interwiki Links einen starken Einfluss auf die Wahrnehmung des Standpunkts „der Ägypter" haben: عامية(ʿāmmīya) wird mit Slang, لغة (luġa) mit

[150] Somekh (1991), S. 25.
[151] Somekh (1991), S. 39.
[152] Niloofar Haeri: The sociolinguistic market of Cairo. Gender, class, and education. 1. Auflage. London, 1997, S. 162–166.
[153] Wikimedia Meta Contributors: Requests for new languages/Wikipedia Egyptian Arabic. http://meta.wikimedia.org/wiki/Requests_for_new_languages/Wikipedia_Egyptian_Arabic (Zugriff am: 13. 12. 2012).

Sprache und لهجة (lahğa) mit Dialekt gleich gesetzt. Wo die Befürworter zum Teil die Diskussion um die Begrifflichkeiten kennen, verlassen sich die Gegner auf die Links um diese Begriffe abwertend gegenüber der Umgangssprache in Stellung zu bringen. In den entsprechenden Sprachgemeinschaften etwa für Darija wird man allerdings auf dieselben Argumente treffen, warum das ebenso ein Akzent, ein Dialekt ja ein Slang ist. Das Kairenische, dass zweifellos als ʿāmmīya bezeichnet wird, kann aufgrund seiner Funktion als zeitgenössische Koine in Ägypten[154], als Prestigedialekt, sicher nicht mit dem Konzept hinter Slang gleichgesetzt werden. Bemerkenswert auch das Detail, dass für Ägypter unverständlich klingenden Varietäten des Arabischen durchaus zugestanden wird, eigenständige Sprachen zu sein.

Dass es keine Regeln für Kairenisch gibt, ist eine jener erwähnten irrationalen und ideologischen Argumente. Manfred Woidich hat sie beschrieben[155]. Obwohl natürlich bei einer derart beschreibenden Grammatik nichts verhindert, dass sich die Sprecher vollkommen neue Regeln einfallen lassen, ist dies innerhalb kurzer Zeiträume doch unwahrscheinlich. Zudem haben einige Einflüsse zu einer gewissen Stabilität beigetragen[156]. Auch die Schreibung des Kairenischen mit arabischen Buchstaben ist nicht halb so unvorhersehbar und vom individuellen Schreiber abhängig, wie die Gegner der Wikipedia Masry glauben machen wollen[157].

Man sollte bedenken, dass die Gegenargumente, so traditionsreich sie auch seien mögen, für die Entscheidungsfindung des „New Language Commitee“ beinahe alle irrelevant sind. Dieses Komitee beurteilt, ob es vollkommen pragmatisch gesehen eine Chance gibt, dass das neue Projekt die notwendige Zahl an Unterstützern und Lesern hat. Dabei handelt es sich regelmäßig um kleine Zahlen von direkt Beitragenden. Wenn also wie in dieser Diskussion vier bis fünf Leute ihren Enthusiasmus öffentlich kundtun, ist das Beweis genug, dass man das Projekt starten kann.

Die ersten Beitragenden mit ägyptischem Nationalismus in Verbindung zu bringen oder ihr Engagement mit ihrem Bekenntnis zu ihrer Religion als Kopten[158], ist durch deren eigene Aussagen nicht bestätigt. „One last Pharao“, der die Idee unterstützt, schrieb in der Diskussion um die Zulässigkeit des Projekts:

> u have pointed to a very good example of an educated egyptian who is not able to contribute in the arabic wikipedia – Ghaly [dessen Beiträge in der arabischen Wikipedia nicht geschätzt wurden wegen seiner Ansichten zur Orthographie gewisser Städte, etwa Venedig] -; he, and any other egyptian, have the right to contribute using their own mother tongue, instead of having to learn a language to contribute. it happens that for example me, and Zerida can speak english well, other egyptians can speak arabic well too, but what about egyptians who cannot speak either well - and they are the majority -, or just egyptian that can, but want their right, and freedom to use their own language? do not they have such right? (sic!)[159]

Auch die frühe ausführliche Beschreibung von Persönlichkeiten, die dem ägyptischen Nationalismus zuzuordnen sind[160], mag damit zusammenhängen, dass es historisch einfach diese Leute waren, die sich

[154] Miller
[155] Woidich (2006).
[156] Miller, S. 595–597.
[157] Gabriel M. Rosenbaum: Egyptian as a written language, in: *Jerusalem Studies in Arabic and Islam* 29 (2004), S. 281–326.
[158] Ivan Panovic: The Beginnings of Wikipedia masry, in: *al-Logha, Series of Papers in Linguistic,* hier S. 98–99.
[159] Wikimedia Meta Contributors: Requests for new languages/Wikipedia Egyptian Arabic. http://meta.wikimedia.org/wiki/Requests_for_new_languages/Wikipedia_Egyptian_Arabic (Zugriff am: 13. 12. 2012).
[160] Ivan Panovic, S. 98–99.

dem Dialekt überhaupt widmeten und so könnte die Beschäftigung mit dem Dialekt dazu geführt haben, dass es einfach und naheliegend war, Artikel zu diesen Persönlichkeiten zu verfassen. Daraus einen Schluss auf die Gesinnung der Gründer der Wikipedia Masry zu ziehen ist gewagt.

Einen positiven Einfluss auf die Beschäftigung mit der eigenen Muttersprache hat wohl das Fehlen der Vision des islamisch dominierten Panarabismus bei koptischen Christen. Es ist für sie wohl nicht so attraktiv, sich die sprachliche Realität zurechtzubiegen.

Von einem anderen Mitglied der ersten Stunde ist im Artikel von Ivan Panovic zu lesen:

> Nabil is a Muslim. What unites Nabil and Ghaly is not necessarily hatred towards Islam or Arabs. It is their similarly articulated nationalism, and their love and appreciation for their native tongue that brings them together.[161]

Ein anderer bedeutender Punkt, den sowohl Panovic[162] als auch die Gegner des Projekts[163] ansprechen, ist die Idee, Ägyptisch sei eine eigene Sprache aus einer anderen Sprachfamilie als Hocharabisch. Bei manchen im ägyptischen Arabisch gebrauchten Strukturen ist es für den Laien wahrscheinlich schwer nachvollziehbar, wie diese aus dem Arabischen stammen könnten, etwa die zirkumfigierte Verneinung ma-…-š. Dies wurde auch im August 2008 in den Artikel über modernes Ägyptisch eingebaut.[164] Dort ist seither zu lesen, dass modernes Ägyptisch eine „hamitische" Sprache sei.[165] Für diese Idee ist ein längerer Artikel als Quelle angeführt, aus dem wohl ursprünglich ein Beleg für die Funktion als Muttersprache kommen sollte.

Die Bewährungsprobe in Form der Frage, ob es genügend Autoren geben und ob die Qualität der Artikel ausreichend sein würde, stand zu diesem Zeitpunkt noch bevor. Es gibt bei WikiMedia nicht nur einen Prozess zum Erstellen von Projekten in noch nicht abgedeckten Sprachen, sondern auch einen Prozess zum Schließen von Projekten, die nicht von den Nutzern angenommen wurden.[166] Abgesehen davon, dass heute nicht mehr davon auszugehen ist, dass die Unterstützung der ägyptischen Wikipedia zu gering wäre, gab es auch nie einen Vorschlag sie wieder zu schließen, auch nicht vor der offiziellen Formalisierung dieses Prozesses.[167]

## Vorschläge zur Schreibung des Ägyptischen in der Wikipedia Masry

Um klar darzustellen, dass es für das Schreiben von Kairenisch ein paar Regeln gibt, führte „Ghaly" schon zu einem frühen Zeitpunkt als die Wikipedia Masry noch im „Incubator" war eine Seite ein, die

[161] Ivan Panovic, S. 102.
[162] Ivan Panovic, S. 98–99.
[163] Wikimedia Meta Contributors: Requests for new languages/Wikipedia Egyptian Arabic. http://meta.wikimedia.org/wiki/Requests_for_new_languages/Wikipedia_Egyptian_Arabic (Zugriff am: 13. 12. 2012).
[164] Wikipedia contributors: « اللغه المصريه الحديثه»: تاريخ التعديل. http://arz.wikipedia.org/w/index.php?title=%D8%A7%D9%84%D9%84%D8%BA%D9%87_%D8%A7%D9%84%D9%85%D8%B5%D8%B1%D9%8A%D9%87_%D8%A7%D9%84%D8%AD%D8%AF%D9%8A%D8%AB%D9%87&offset=20080812042737&action=history (Zugriff am: 13. 12. 2012).
[165] Wikipedia contributors: اللغه المصريه الحديثه. http://arz.wikipedia.org/wiki/%D8%A7%D9%84%D9%84%D8%BA%D9%87_%D8%A7%D9%84%D9%85%D8%B5%D8%B1%D9%8A%D9%87_%D8%A7%D9%84%D8%AD%D8%AF%D9%8A%D8%AB%D9%87 (Zugriff am: 13. 12. 2012).
[166] Wikimedia Meta Contributors: Proposals for closing projects. http://meta.wikimedia.org/wiki/Proposals_for_closing_projects (Zugriff am: 19. 11. 2012).
[167] Wikimedia Meta Contributors: Proposals for closing projects/Archive. http://meta.wikimedia.org/wiki/Proposals_for_closing_projects/Archive (Zugriff am: 19. 11. 2012).

beschreibt, wie für die ägyptische Wikipedia geschrieben werden soll. Der „Incubator“ ist ein Platz, den die WikiMedia Foundation zur Verfügung stellt um schnell eine Wikipedia, ein Wiktionary u. a. m. auszuprobieren. Man findet auf dieser Seite also eine Mischung aus zwei Themenbereichen: Zum einen die in der Wikipedia üblichen Aufforderungen einen einer Enzyklopädie angemessenen Schreibstil zu pflegen. Diese findet man in jeder Wikipedia und die Anweisungen sind auch weitgehend dieselben in den verschiedenen Sprachen. Zum anderen findet man aber auch einiges an Gedanken und Vorschlägen zum Thema der Verschriftung des ägyptischen Dialekts. Die Überlegungen lassen immer noch eine Menge Mehrdeutigkeiten zu und die Betonung, dass es jeder so machen soll, wie er mag, solange es die Ägypter verstehen, lässt nicht auf einen einheitlichen Stil hoffen. Es wundert also kaum, wenn eine kurze Stichprobe schon Abweichungen ergibt[168]. Einige Probleme, die der oder die Autoren gesehen haben, werden allerdings versucht einer Lösung zuzuführen. Beachtenswert ist etwa die Bitte um Satzzeichen, die so im Arabischen keine Tradition haben.

> **Punkte und Beistriche sind sehr wichtig.** Es wäre schön, wenn du sorgfältig bist [tāḫud bālak] bei Punkten und Beistrichen, damit man nicht verloren geht mitten im Text. Sie werden direkt mit dem Wort verbunden [lāzi'īn fi] geschrieben, das vor ihnen geschrieben wurde, und nach ihnen folgt ein Leerzeichen [misāfa].[169]

Aber auch die Disambiguierung von Aussprachen mit Hilfe von Zeichen, die ins arabische Alphabet Einzug hielten, als Perser versuchten ihre Sprache damit zu schreiben, ist eine Neuerung und Abweichung von dem, was in der hocharabischen Wikipedia praktiziert wird. Heute sind ja eine Reihe von Erweiterungen des arabischen Zeichensatzes[170] einschließlich Ligaturen durch Unicode problemlosauf den meisten Computern verfügbar.[171]

- Verwende möglichst Buchstaben, die klar sind, im Inneren der Seite, wie پ [p], چ [ʒ] ڤ[engl. v]

  Schreib فيفا oder ڤيڤا [viva]. Schreib ايفا oder ايڤا [engl. Eva], [beides] kein Problem[172]

Auch scheint es dem Autor oder den Autoren darauf anzukommen Genitivverbindungen zumindest bei Leitwörtern mit Femininendungen klar von sonstigen gemeinsamen Vorkommen von femininen und anderen Wörtern zu trennen.

> **Das t der Weiblichkeit [tah ᵉt-ta'nīṯ, ة]:** Das t der Weiblichkeit schreibt man immer der Aussprache nach entweder /ة/ oder /ه/, das heißt, man sollte schreiben <انا رايح المكتبه[der letzte Buchstabe ist ه, weil es

[168] Ivan Panovic

[169] Wikipedia contributors: ويكيبيديا:سياسات/طريقة الكتابه. http://arz.wikipedia.org/wiki/%D9%88%D9%8A%D9%83%D9%8A%D8%A8%D9%8A%D8%AF%D9%8A%D8%A7:%D8%B3%D9%8A%D8%A7%D8%B3%D8%A7%D8%AA/%D8%B7%D8%B1%D9%8A%D9%82%D8%A9_%D8%A7%D9%84%D9%83%D8%AA%D8%A7%D8%A8%D9%87 (Zugriff am: 28. 11. 2012). Alle Übersetzungen ins Deutsche vom Autor. Eine möglichst wortgetreue Übersetzung wurde angestrebt.

[170] Alan S. Kaye: Arabic Alphabet for Other Languages. in: Cornelis H. M. Versteegh (Hg.): *Encyclopedia of Arabic language and linguistics,* Leiden ( 1. A-Ed), S. 133–147.

[171] Seit dem Erscheinen von Microsoft Windows Vista im Jahr 2007 sind auf jedem handelsüblichen Computer alle beschriebenen Zeichen in zumindest ein paar Schriftarten verfügbar. Auf anderen Betriebssystemen war das teils schon einige Jahre vorher der Fall.

[172] Wikipedia contributors: ويكيبيديا:سياسات/طريقة الكتابه. http://arz.wikipedia.org/wiki/%D9%88%D9%8A%D9%83%D9%8A%D8%A8%D9%8A%D8%AF%D9%8A%D8%A7:%D8%B3%D9%8A%D8%A7%D8%B3%D8%A7%D8%AA/%D8%B7%D8%B1%D9%8A%D9%82%D8%A9_%D8%A7%D9%84%D9%83%D8%AA%D8%A7%D8%A8%D9%87 (Zugriff am: 28. 11. 2012).

als /a/ gesprochen wird]> und man sollte schreiben <انا رايح مكتبة الكليه [ة steht hier am Ende des dritten Wortes, weil es in einer Genitivverbindung mit dem 4. Wort steht und /it/ gesprochen wird. Das vierte Wort endet wieder auf /a/]>.[173]

Nicht zuletzt werden auch Probleme beschrieben und Lösungen vorgeschlagen für Situationen, die sich erst durch die Schreibung, die mehr an der Aussprache orientiert ist, ergeben.

**ق**: Wenn es ein Wort gibt, von einer arabischen Wurzel in dem es den Buchstaben <ق> gibt, geht man darauf zurück, wie es ist.

**ث**: Wenn es ein Wort gibt von einer arabischen Wurzel oder aus irgendeiner Sprache, in der es diesen Buchstaben/diese Aussprache gibt, (ث[θ]) und es wird /س[s]/ ausgesprochen, dann geht man darauf zurück, wie es ist. Wenn es /ت[t]/ ausgesprochen wird, dann schreibt man <ت>

**ذ**: Wenn es ein Wort gibt von einer arabischen Wurzel oder aus irgendeiner Sprache, in der es den Buchstaben/die Aussprache (ذ[ð]) gibt, und es wird /ز[z]/ ausgesprochen, dann geht man darauf zurück, wie es ist. Wenn es /د[d]/ ausgesprochen wird, schreibt man <د>.

Die Hamza [glottal stop, Glottisschlag, ء mit seinen diversen Trägern]:Es gibt keine trennenden Hamza und keine stummen Anfangshamza [hamzāt waṣl], alle sind <ا> (außer wenn es im Text ein anderes Wort gibt), aber wenn du willst, dann schreib trennende Hamza innerhalb der Artikel, kein Problem.

**ى**: Es gibt kein <ي>, es gibt <ى>; weil <ى> ist die Schreibweise, wie es die Ägypter schreiben. <ي> ist kein Problem, weil <ى> und <ي> sind ein Buchstabe.[174]

und

**Partikel, die den Genitiv regieren, und Konjunktionspartikel** [Partikel hier hauptsächlich in dem Sinn, dass es sich um einzelne Buchstaben handelt, die grafisch mit dem nächsten Wort verbunden werden.]: Das Schreiben von Partikel, die den Genitiv regieren, und Konjunktionspartikel ist klarer und einfacher, wenn sie getrennt geschrieben werden von dem Wort, das nach ihnen kommt, das heißt, schreib <فى اروبا [in Europa]> oder <ف اروبا [das gleiche, aber es berücksichtigt, dass /fi/ eben mit einem kurzen i gesprochen wird und es ist nicht mehrdeutig mit /fi/ mit langem ī im Sinne von „es gibt“]>, [das ist] klarer und

[173] Wikipedia contributors: ويكيبيديا:سياسات/طريقة الكتابه. http://arz.wikipedia.org/wiki/%D9%88%D9%8A%D9%83%D9%8A%D8%A8%D9%8A%D8%AF%D9%8A%D8%A7:%D8%B3%D9%8A%D8%A7%D8%B3%D8%A7%D8%AA/%D8%B7%D8%B1%D9%8A%D9%82%D8%A9_%D8%A7%D9%84%D9%83%D8%AA%D8%A7%D8%A8%D9%87 (Zugriff am: 28. 11. 2012).

[174] Wikipedia contributors: ويكيبيديا:سياسات/طريقة الكتابه. http://arz.wikipedia.org/wiki/%D9%88%D9%8A%D9%83%D9%8A%D8%A8%D9%8A%D8%AF%D9%8A%D8%A7:%D8%B3%D9%8A%D8%A7%D8%B3%D8%A7%D8%AA/%D8%B7%D8%B1%D9%8A%D9%82%D8%A9_%D8%A7%D9%84%D9%83%D8%AA%D8%A7%D8%A8%D9%87 (Zugriff am: 28. 11. 2012).

einfacher als das Lesen von <فاروبا [denn das kann auch /fa-…/ gelesen werden was dann „und, und auch usw.“ bedeuten kann]>.[175]

Zum überwiegenden Teil folgen diese Rechtschreibanweisungen dem, was seit Jahrzehnten im verschrifteten ägyptischen Arabisch üblich ist, da sich ohne offizielle Anerkennung kein Satz von Regeln wirklich weitgehend durchgesetzt hat. Die gebräuchliche Schreibung der kairenischen Varietät und bekannte Abweichungen davon wurden etwa schon von Rif'at al Farnawānī (1981) anhand der damals schon zahlreich verfügbaren Theaterstücke beschrieben. Dort heißt es etwa:

> Im Allgemeinen können wir drei Hauptgrundsätze, die die Verschriftung des Dialekts beherrschen wie folgt vorstellen:
>
> 1) Die Neigung zur historischen bzw. etymologischen Schreibweise, die sich bei Wörtern deren Beziehung zum Hocharabischen noch klar ist, findet. […]
> 2) Die Neigung zur phonemischen Schreibweise, die sich bei den Wörtern deren Beziehung zum Hocharabischen unklar ist, findet […]
> 3) Die Neigung zu persönlichen Schreibgewohnheiten der Schriftsteller. Diese findet sich vor allem bei Fremdwörtern, die von gebildeten Personen benutzt werden.
>
> […] Die Meinungsverschiedenheiten unter Schriftstellern über die richtige Schreibweise dauern trotz der großen Zahl der vorhandenen Werke, die im Dialekt verfaßt wurden, an. Niemand hat eindeutige Regel[n] für die Verschriftung des Dialekts festgelegt, vielleicht auch deshalb, weil alle Versuche, die hocharabische Schreibung zu reformieren, bis heute keinen Erfolg gehabt haben.[176]

Zur Getrennt- oder Zusammenschreibung heißt es:

> Während die Schriftsprache eine einfache Regel für Zusammen- und Getrenntschreibung kennt […], läßt sich bei der Schreibung von Dialekttexten keine bestimmte Regel erkenn[en].[177]

Weiters gibt es etwa eine Untersuchung von drei Theaterstücken von Saʿad ad-Dīn Wahba, die auf Kairenisch verfasst sind, von Renate Malina (1987) unter anderem im Hinblick auf die Orthographie. Darin bemerkt die Autorin anfangs, dass es sich bei dem ihr vorliegenden Text um einen Druck handelt. Sie mutmaßt, dass der Autor vielleicht etwas anderes notiert haben könnte. Sie findet alle Beobachtungen von El-Farnawany bestätigt. Der Autor scheint des Öfteren spontan zwischen möglichen Schreibvarianten auszuwählen.[178] Was die Femininendung angeht, so bemerkt sie:

> Die Femininendung von Substantiven und Adjektiven wird im vorliegenden Text sowohl mit ۃ als auch mit ہ notiert, wobei die Notierung mit ۃ den größeren Anteil aufweist. […]

[175] Wikipedia contributors: ويكيبيديا:سياسات/طريقة الكتابه. http://arz.wikipedia.org/wiki/%D9%88%D9%8A%D9%83%D9%8A%D8%A8%D9%8A%D8%AF%D9%8A%D8%A7:%D8%B3%D9%8A%D8%A7%D8%B3%D8%A7%D8%AA/%D8%B7%D8%B1%D9%8A%D9%82%D8%A9_%D8%A7%D9%84%D9%83%D8%AA%D8%A7%D8%A8%D9%87 (Zugriff am: 28. 11. 2012).

[176] Rif'at al Farnawānī: Ägyptisch-Arabisch als geschriebene Sprache. Probleme der Verschriftung einer Umgangssprache, 1981, S. 52f.

[177] Farnawānī (1981), S. 162.

[178] Renate Malina: Zum schriftlichen Gebrauch des kairinischen Dialekts anhand ausgewählter Texte von Saʿdaddīn Wahba. Berlin, 1987, S. 120–136.

> Wenn an das fem. Partizip ein Suffix angefügt wird, wird die Längung von –a immer durch ا wiedergegeben.[179]

aber:

> In Genitivverbindungen wird die Endung –(i)t des fem. Leitwortes immer durch ة ausgedrückt.[180]

Die jüngsten Untersuchungen zu Kairenisch als geschriebene Sprache stammen von Rosenbaum[181].

> This situation resulted in the view, prevalent in scholarship and literary criticism to this day, that *ʿāmmiyya* has no script or spelling of its own, and thus its writers are forced to use an ill-adapted script and orthography, which do not reflect the full range of its phonetic phenomena. [...] This approach, even if it did reflect the situation of *ʿāmmiyya* writing in the past, is inadequate for describing the situation prevailing in Egyptian literature toward the end of the twentieth and the beginning of the twenty-first century.[182]

Rosenbaum beschreibt die heute gängige Verschriftung. Sie enthält immer noch jede Menge Varianten, die die Autoren weidlich nutzen. Es gäbe eine Diskussion über Rechtschreibung, die auf Hocharabisch zurückgreife. Aber der Punkt einer festgelegten Orthographie sei noch lange nicht erreicht.

## Wer trägt zur Wikipedia Masry bei und wie viel?

Die WikiMedia Foundation stellt unter http://stats.wikimedia.org/EN/TablesWikipediansContributors.htm eine vergleichende Statistik bzw. unter http://stats.wikimedia.org/EN/TablesWikipediaARZ.htm eine recht ausführliche Statistik speziell für die Wikipedia Masry bereit, die im Folgenden nach dem Stand Dezember 2012 interpretiert werden sollen.

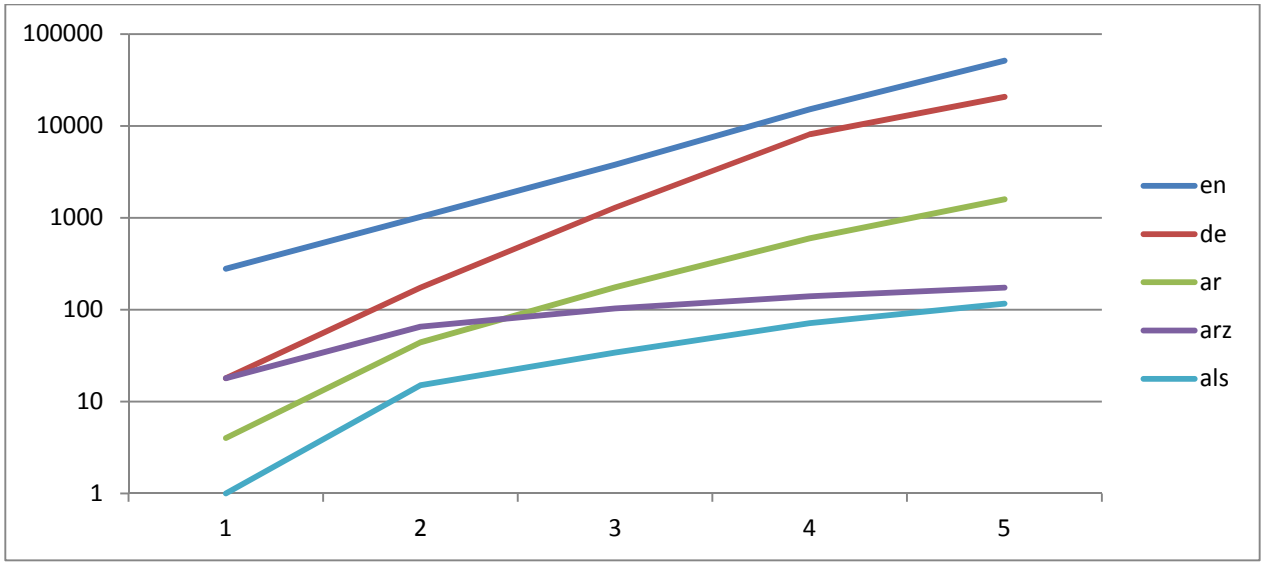
Die Wikipedia Masry ist - wie ein Vergleich der Zahl der Mitwirkenden mit einer anderen Wikipedia einer typischen Umgangssprache, Schweizerdeutsch, zeigt - eine typische Wikipedia einer Umgangssprache in einem Sprachgebiet, in dem eine andere Sprache die dominierende Schriftsprache ist. Was sich sicher nicht vergleichen lässt, ist das ökonomische und soziale Umfeld. Es ist jedem Schweizer, der das will, sicher möglich zur alemannischen Wikipedia beizutragen. Unter den arabischsprachigen Menschen stellen die Ägypter sicher einen großen Anteil, wenn nicht die Mehrheit. Aber es wird längst nicht jeder Ägypter genug Zeit oder die technischen Voraussetzungen haben, um in der Wikipedia Masry zu schreiben. In diesem Sinne ist es bezeichnend, dass sich die beiden Wikipedias trotz ihres unterschiedlichen Alters in ihren Kennzahlen heute noch ähneln.

Das Wachstum von Wikipedias folgt offensichtlich einer annähernd zur Potenz verlaufenden Entwicklung. Das jedenfalls kann man bei den großen Wikipedias für Englisch und Deutsch beobachten. Je mehr Artikel vorhanden sind, desto mehr Leute lesen sie, desto mehr Leute ergänzen Informationen oder gestalten eigene Artikel. Ein Multiplikatoreneffekt bleibt nicht aus, weil jemand, der gute Informationen gefunden hat, diese Entdeckung wahrscheinlich auch anderen mitteilt usw. Wenn es nicht gelingt, Wikis, wie eine Wikipedia, mit genügend attraktiven Inhalten zu füllen, dann werden

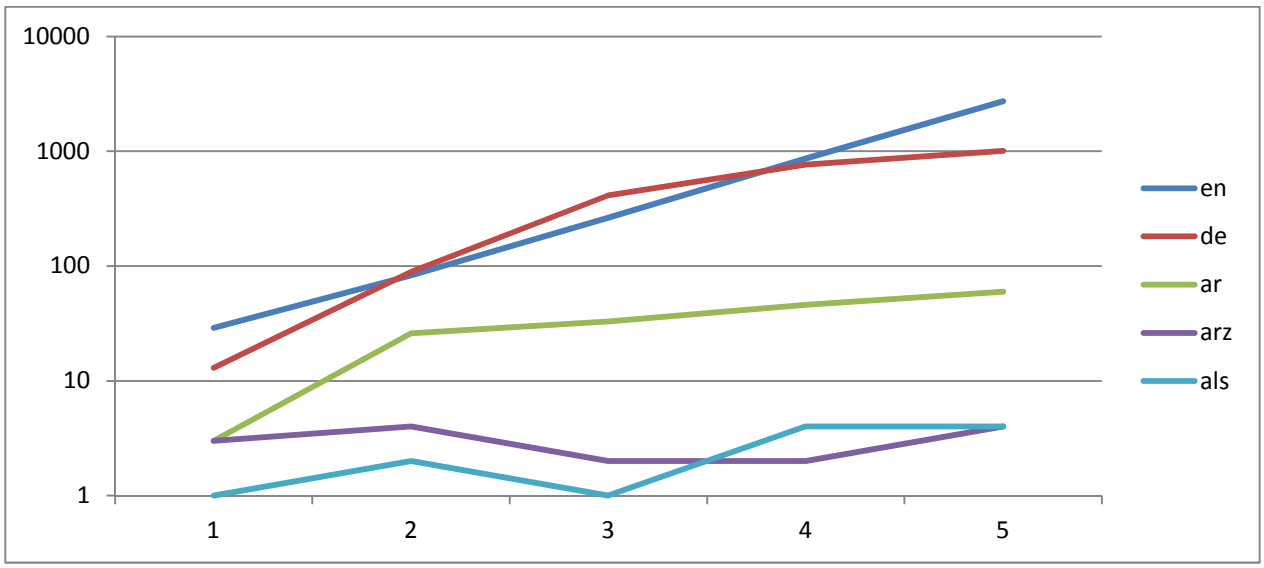
[179] Malina (1987), S. 125f.
[180] Malina (1987), S. 126.
[181] Rosenbaum (2004).
[182] Rosenbaum (2004), S. 282f.

irgendwann die an Information Interessierten wegbrechen und sie wird verwaisen. Eine große Anzahl von Lesern und Autoren erhöht auch die Wahrscheinlichkeit, dass Fehler gefunden und ausgebessert werden. Von einem Zusammenbruch sind die Wikipedias in den beiden Umgangssprachen Schweizerdeutsch und Ägyptisch weit entfernt, aber ihre Entwicklung ist stark gebremst, vielleicht von einer Erwartungshaltung in Bezug auf die Sprache, die mit verlässlichen Informationen assoziiert wird. Solche werden eher mit Hochsprache assoziiert, auch wenn Informationen in jeder Wikipedia mit Vorsicht zu verwenden sind.

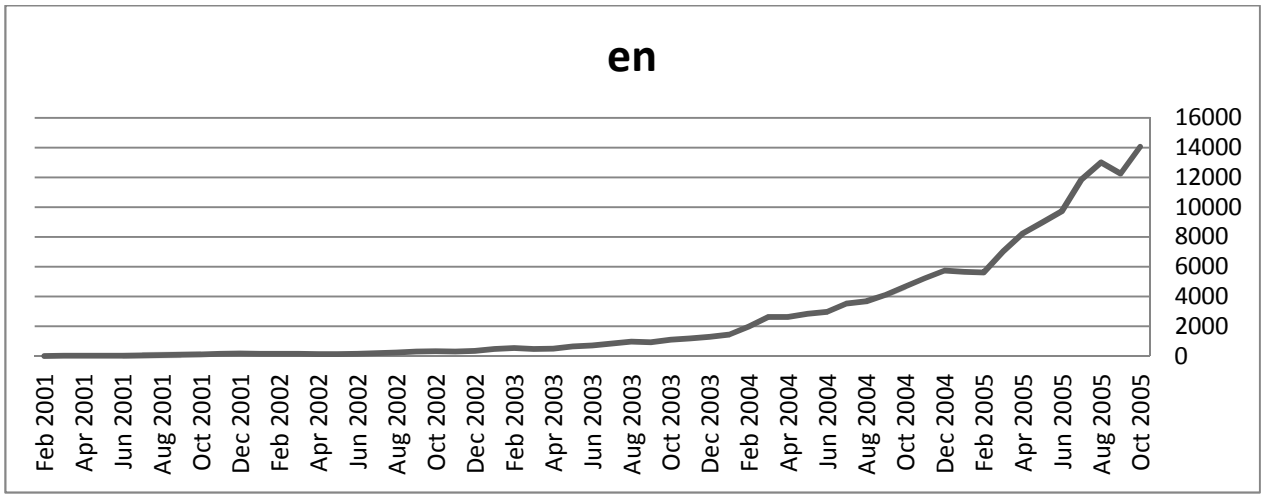
Die folgenden Diagramme stellen die Werte der jeweiligen Kennzahlen im Verlauf der Jahre seit dem Start der Wikipedia dar. Die Diagramme stellen jeweils die Daten der ersten 5 Jahre nach dem Start der Wikipedia in der angegebenen Sprache dar. Dies ist, wegen des oben erwähnten Multiplikatoreneffekts, eine notwendige Vereinheitlichung. Eine interessante Kennzahl ist die Anzahl jener Leute, die eine Benutzerkonto in einer Wikipedia angelegt haben und seither mindestens 10 Beiträge verfasst haben. Da die Zahlen erwartungsgemäß bei den größten Wikipedias exponentiell ansteigen, wurde zur besseren Darstellbarkeit eine logarithmische Skalierung gewählt.

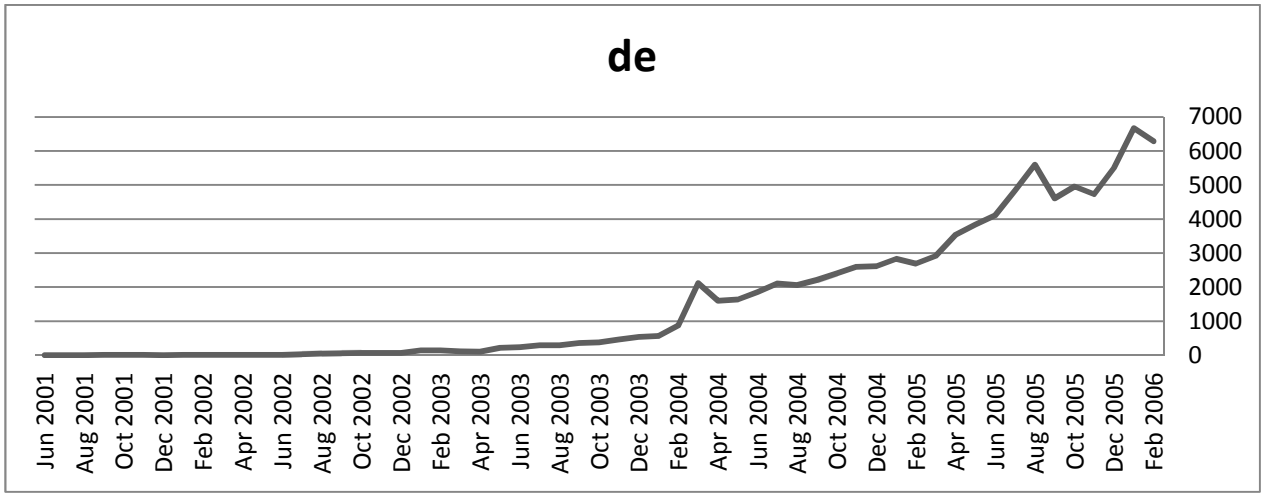


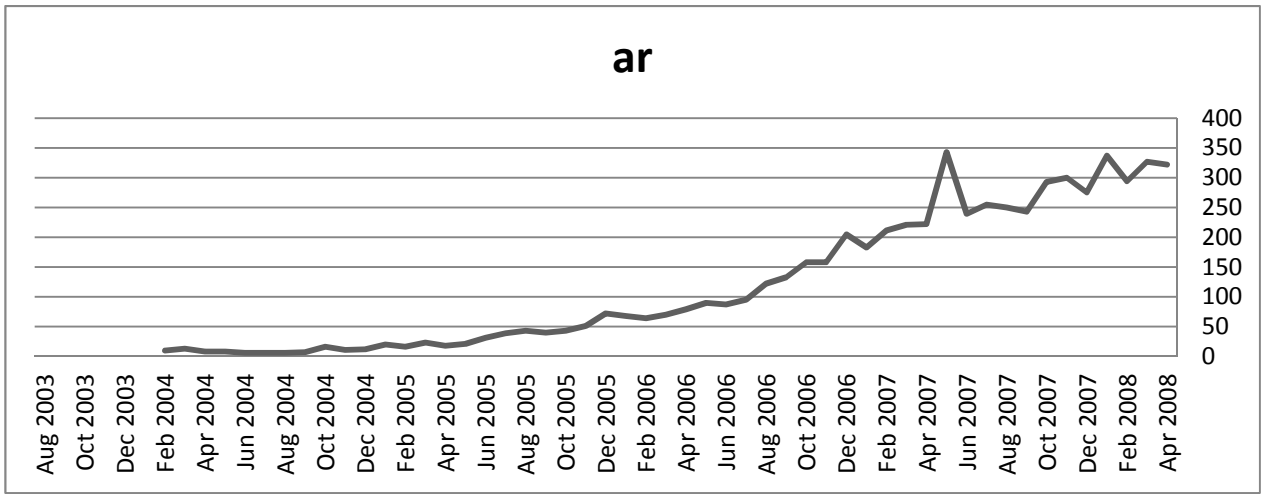

Eine weitere Kenngröße ist die Zahl jener Beitragenden, die sehr aktiv sind, was gleichgesetzt wird damit, dass sie zu 100 Artikeln in einem Zeitraum von einem Monat beigetragen haben (hier jeweils der Dezember eines jeden Jahres des Bestehens das Wikipedia).



Auch die Anzahl jener, die eine kleine Anzahl von Artikeln - mindestens fünf - in einem Monat beitragen, zeigt ein ähnliches Bild für die großen hochsprachlichen Wikipedias. Auch hier soll ein Vergleich des gleichen Zeitraums nach dem Start der Wikipedia durchgeführt werden.

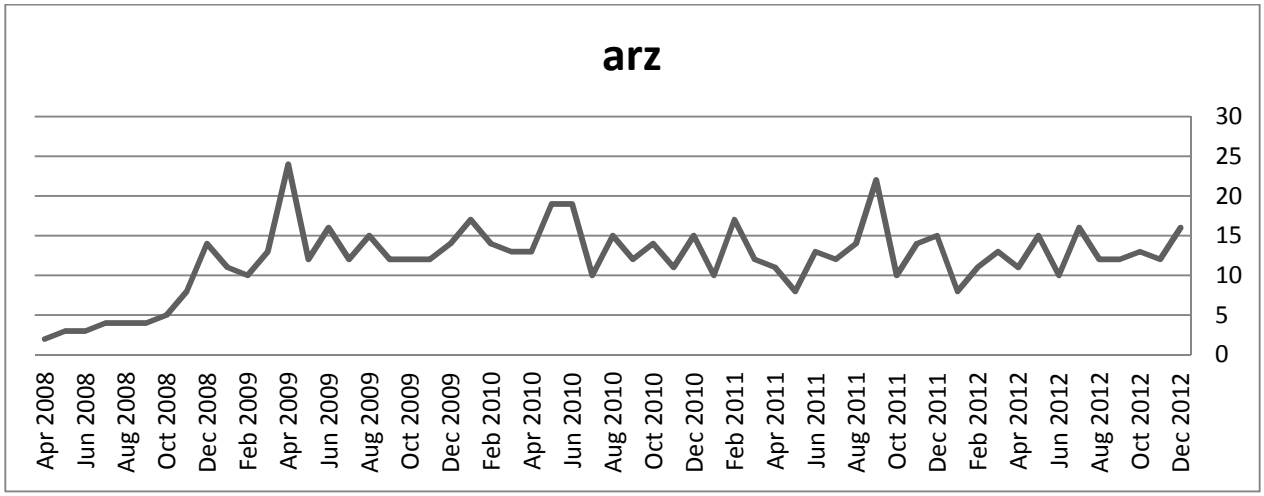


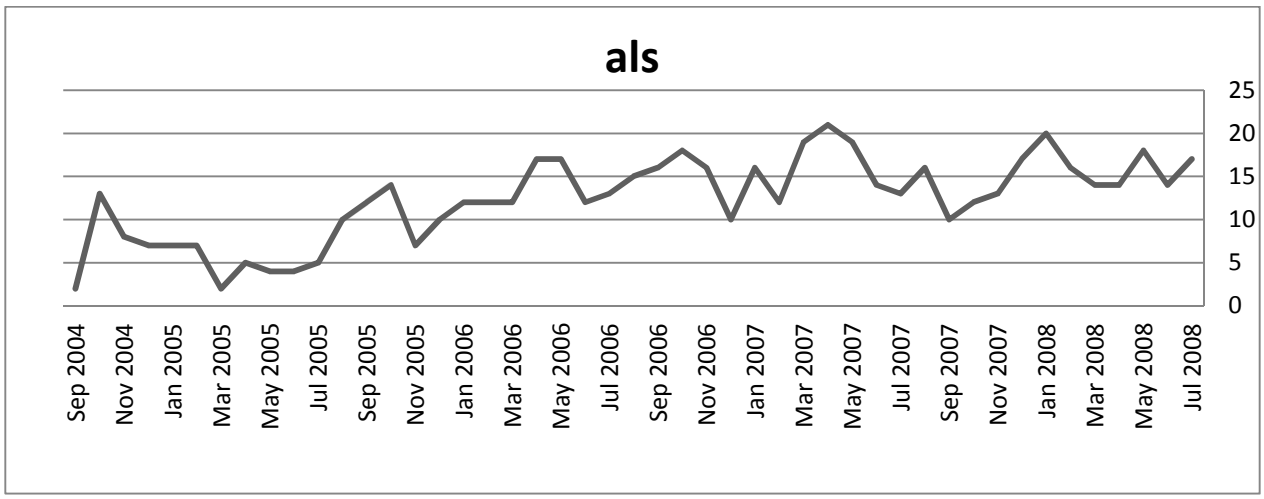




Deutlich ist bei allen hochsprachlichen Wikipedias, dass der Anstieg steiler wird, wenn sich auch die absoluten Zahlen um Größenordnungen unterscheiden. Auffällig ist, dass Arabisch, das nominell um eine Größenordnung mehr Sprecher und Autoren hat als Deutsch, bei der Zahl der Beiträge eine um eine Größenordnung kleinere Anzahl vorzuweisen hat. Dies mag vielleicht mit dem oben erwähnten Umstand

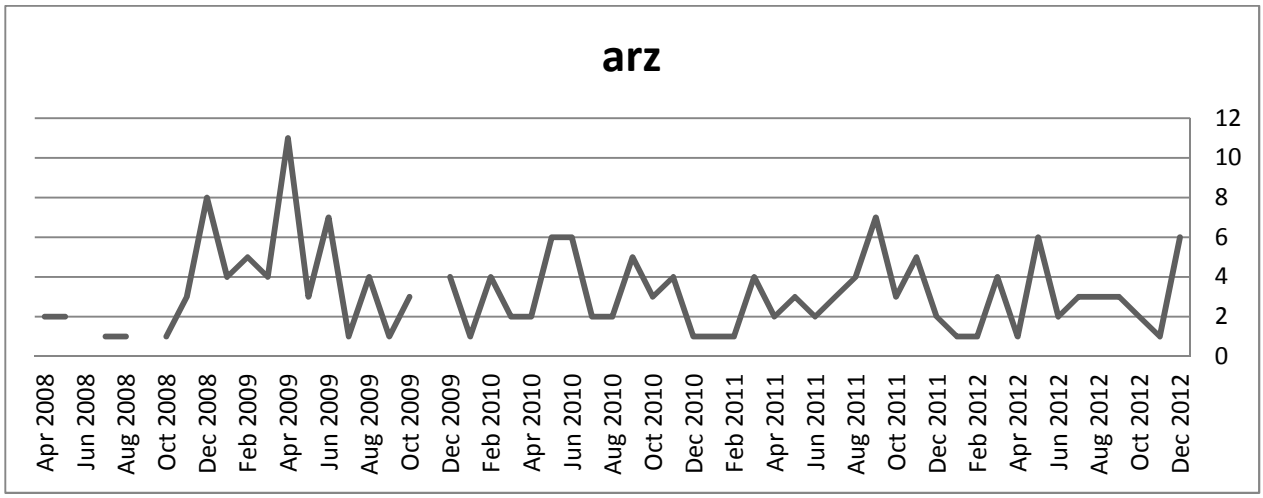
der Zugangsmöglichkeiten zusammenhängen. Anderes verhält es sich bei den Wikipedias in Umgangssprachen: Sie zeigen ebenfalls ähnliche Entwicklungen.

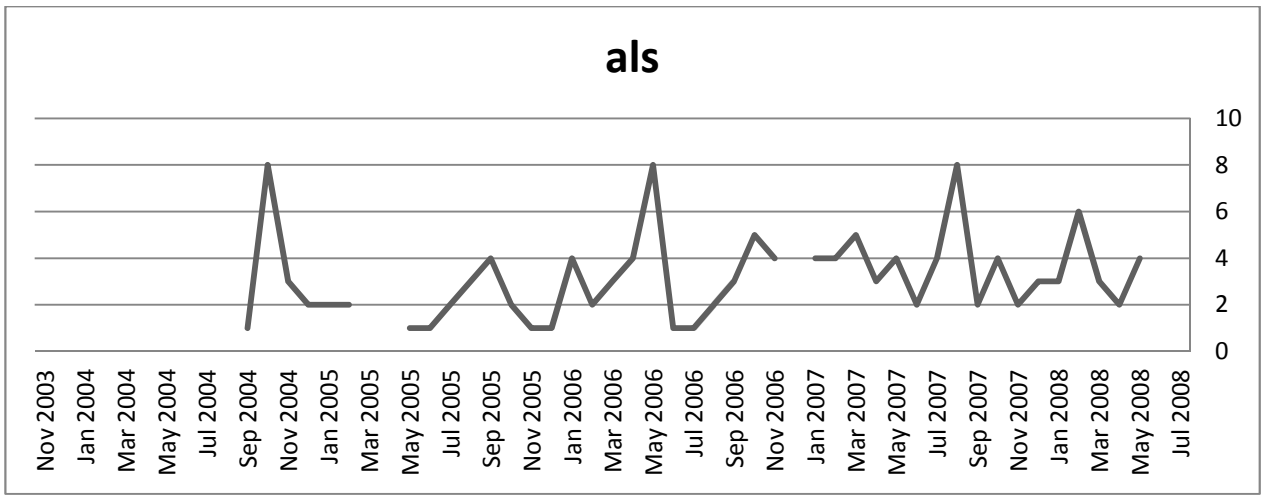


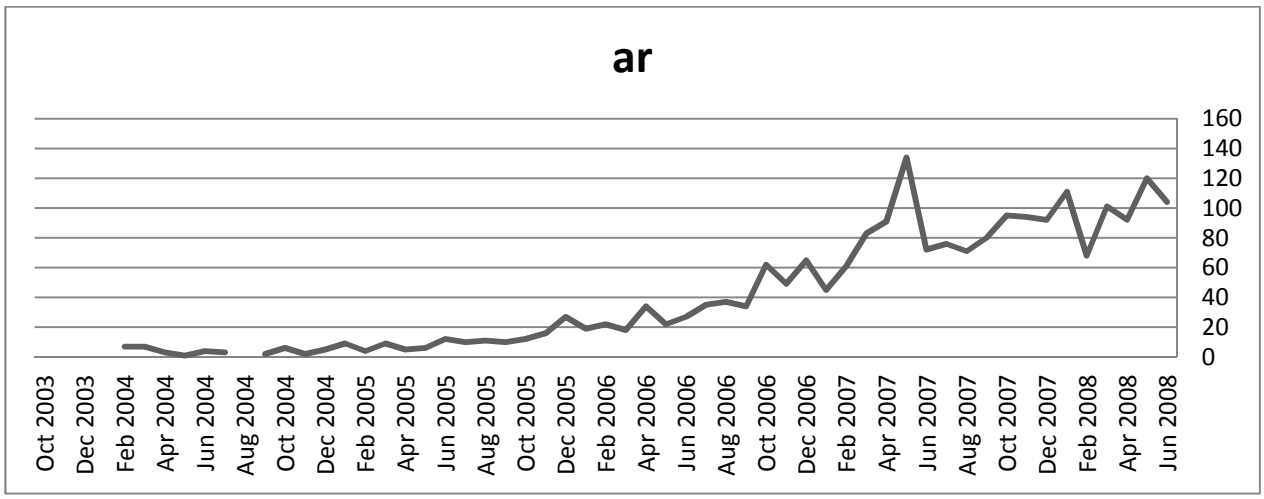

Die Zahlen bleiben aber auf einem niedrigen Niveau relativ konstant. Das Wachstum ist also relativ stark limitiert.

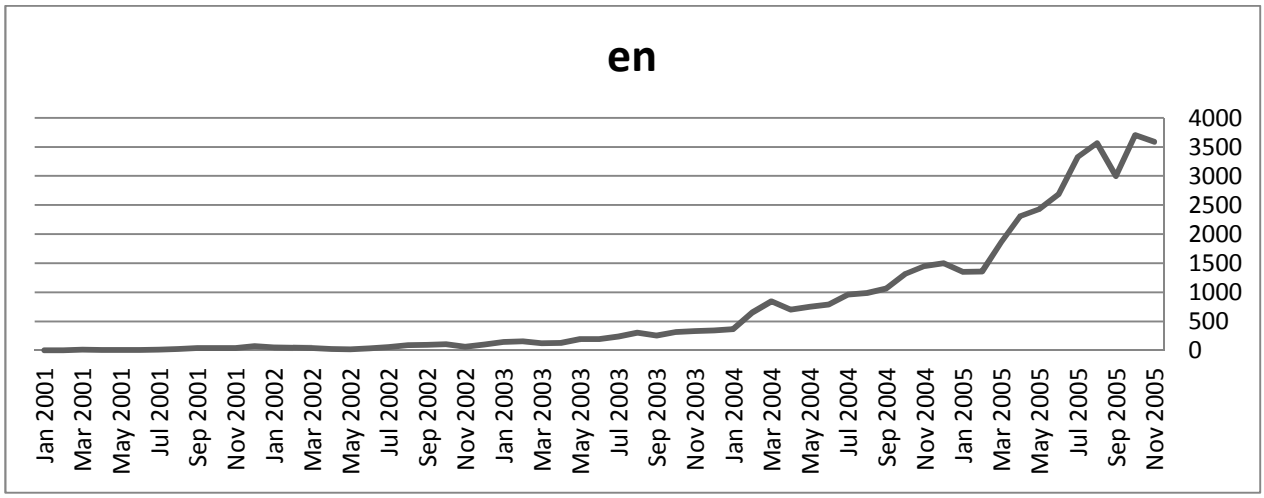
Die ernsthaften Neuanmeldungen von Autoren, wobei von einem ernsthaften Engagement ausgegangen wird, nachdem sie zehn Artikel verfasst oder bearbeitet haben, zeigen ein ähnliches Bild.






Im Vergleich dazu steigen in den hochsprachlichen Wikipedias auch diese Anmeldungen mit der Zeit steiler an:




Für die Frage, was man aus den an der Wikipedia gewonnenen Ergebnissen in Bezug auf ägyptisches Arabisch schließen kann, darf man eines nicht aus den Augen verlieren: An der Wikipedia haben bis heute nicht mehr als ca. 170 Personen mit einem nennenswerten Anteil an Artikeln - mehr als zehn - mitgewirkt. Betrachtet man die Rangliste der Beitragenden, ergibt sich ein noch engeres Bild:

| User | Edits | | | Creates | | | |
|---|---|---|---|---|---|---|---|
| | Rank | Articles | | Articles | | % | |
| User Contributions | now | Total | Last 30 days/last Edit | Total | Last 30 days | | |
| Samsam22 | 1 | 49.898 | 18.10.2012 | - | - | 43% | |
| Raafat | 2 | 16.190 | 362 | 1.796 | 49 | 14% | 58% |
| Ghaly | 3 | 13.205 | 147 | 1.751 | 10 | 12% | 69% |
| Mahmudmasri | 4 | 10.585 | 27.11.2012 | - | - | 9% | 78% |
| Ramsis II | 5 | 9.702 | 27.11.2012 | - | - | 8% | 87% |
| Eskandarany | 6 | 5.068 | 213 | 496 | 13 | 4% | 91% |
| Abu Amal Bahraini | 7 | 1.362 | 21.10.2010 | - | - | 1% | 92% |
| El Gaafary | 8 | 1.306 | 27.11.2012 | - | - | 1% | 94% |
| D'ohBot | 9 | 887 | 26.04.2012 | - | - | 1% | 94% |
| Dudi | 10 | 722 | 24.11.2008 | - | - | 1% | 95% |

95 % aller von bekannten Benutzern erstellten Artikeln (114.720) wurden von zehn Leuten verfasst. Das bedeutet einerseits, dass sich im aktuellen Text nicht alle bekannten Schreibvarianten wiederfinden werden. Andererseits ist es auch nicht ungewöhnlich, dass ein und derselbe Autor Worte in zwei Varianten schreibt. Es ist andererseits auch ein Zeugnis von viel Durchhaltevermögen, wenn diese zehn Autoren eine Textmenge von 2,3 Millionen Worten verfassen, allerdings unter Zuhilfenahme der Vorlagenfunktion von Wikipedia, die Texte im Prinzip vervielfältigt.

Ohne genauere Untersuchungen des Hintergrunds dieser Leute in Bezug auf Schicht, Herkunft, Ausbildung etc. aber auch im Hinblick auf die im Vergleich zur ägyptischen oder auch nur kairinischen Bevölkerung kann man nur sagen: Wikipedia ist grundsätzlich nicht repräsentativ. Was man mit der Wikipedia hat, ist ein Corpus, das gewollt ägyptischen Dialekt enthält. Damit kann in diesem Rahmen angenommen werden, dass das Kairenische die gewünschte Sprache und das Hocharabische die Ausnahme darstellt. Bei den meisten anderen Textquellen wird es genau umgekehrt sein.

## Über die maschinelle Verarbeitung von Texten in arabischer Sprache und in arabischen Dialekten

Das Arabische stellt durch manche Eigenheiten und einige Gemeinsamkeiten mit anderen semitischen Sprachen höhere Anforderungen an die Verarbeitung mit dem Computer. Die Themenbereiche, in denen Natural Language Processing (NLP) eingesetzt wird, um automatisch große Mengen von Text mit linguistischer Information anzureichern, decken für jene Sprachen, bei denen sich die Sprachwissenschaft schon längere Zeit der Computerunterstützung bedient, weite Teile der sprachwissenschaftlichen Disziplinen ab. Computer liefern dabei zwar selten 100 % richtige Ergebnisse, aber für die am besten derart untersuchten Sprachen, allen voran Englisch, aber auch Deutsch, 80 % bis über 90 % richtige Ergebnisse, was ein darauf aufbauendes Arbeiten recht komfortabel werden lässt. Die Tatsache, dass die computergestützte Beschäftigung in Amerika und England ihren Ausgang nahm, hat bezogen auf viele andere Sprachen ein Problem zur Folge: Englisch ist keine morphologisch komplexe Sprache im Vergleich zu den meisten indoeuropäischen Sprachen und Sprachen anderer Weltregionen. Die Funktion eines Wortes im Satz, die genaue Person eines Verbs und anderes mehr, werden durch die Satzstruktur - durch die Syntax - ausgedrückt und ist nicht aus dem einzelnen Wort alleine heraus ersichtlich. In anderen Sprachen, wie etwa auch Deutsch ist die Syntax oft variabler und die morphologische Vielfalt höher. So kann man durchaus hilfreiche Publikationen zum Thema NLP finden, die morphologische Analyse kaum behandeln. Im Englischen wird in diesem Zusammenhang oft nur von stemming gesprochen. Ein einfaches Entfernen aller Prä- und Suffixe eines Wortes, die relativ einfach zu bestimmen sind, reicht, um den Stamm zu finden, der auch in einem Wörterbuch zu finden ist.[183] Alle Wortformen, bei denen man mit diesen sehr einfachen Regeln nicht weiterkommt, sind aufzählbar und bekannt. Im Deutschen ist das schon komplexer. Es gibt Veränderungen am Stamm, etwa bei Pluralen wie Haus - Häuser oder bei unregelmäßigen Verbformen wie nehmen – nahm - genommen, die im Deutschen viel zahlreicher sind als im Englischen. Es gibt auch insgesamt mehr Suffixvarianten, die bei einer Analyse, die nur den Stamm herausfinden möchte, noch recht einfach entfernt werden können.[184] Für die besser untersuchten Sprachen in Bezug auf computergestützte Verarbeitung unter den indoeuropäischen Sprachen können solche Aufgaben mit vorhandenen Werkzeugen gelöst werden.

Arabisch zählt nicht zu jenen Sprachen, bei denen es einfach ist, computergestützte Werkzeuge zu deren Verarbeitung zu entwickeln. Hocharabisch ist eine morphologisch sehr komplexe Sprache, deren Verarbeitung noch zusätzlich dadurch komplexer wird, dass ihre Verschriftung kaum je vollständig erfolgt.[185] Das Fehlen der Vokalzeichen und den als diakritische Zeichen realisierte Hamzae und Lautverdoppelungen (šadda)[186] sowie der Antritt von kurzen Worten[187], die nur aus einem Zeichen bestehen, an das nachfolgende Wort erhöht die Menge an möglichen Lesarten einiger Zeichenketten enorm.[188]

Für beinahe jede weiterführende Beschäftigung mit einem Text ist es also erst einmal nötig, eine möglichst genaue morphologische Analyse der einzelnen Wörter vorliegen zu haben. Diese wird in

[183] Bird, Klein, Loper (2009).
[184] Jörg Caumanns: A Fast and Simple Stemming Algorithm for German Words. http://edocs.fu-berlin.de/docs/servlets/MCRFileNodeServlet/FUDOCS_derivate_000000000350/tr-b-99-16.pdf?hosts= (Zugriff am: 04.01.2013).
[185] Nizar Y. Habash: Introduction to Arabic natural language processing. San Rafael, Calif., USA, 2010, S. 11.
[186] Habash (2010), S. 31–34.
[187] Habash (2010), S. 47f.
[188] Nizar Y. Habsh, Owen Rambow: Arabic Tokenization, Part-of-Speech Tagging and Morphological Disambiguation in One Fell Swoop. http://dl.acm.org/citation.cfm?id=1219911 (Zugriff am: 04. 01. 2013).

vielen Fällen eine stattliche Menge unterschiedlicher Lösungen zu ein und derselben Zeichenkette ergeben. Will man die Lösungen danach wieder eingrenzen, muss man den Satzzusammenhang analysieren. Damit sind dann nur noch wenige der gefundenen nicht kontextbezogenen Analysen tatsächlich möglich. Die meisten Ansätze zur computerunterstützten Analyse von Texten in natürlichen Sprachen versuchen heute mittels statistischer Methoden ihre Zuverlässigkeit zu erhöhen. Dazu müssen bestimmte Wahrscheinlichkeiten aus einem bekanntermaßen richtig analysierten Text (Goldstandard) errechnet werden.[189] Dieser wird meist von Linguisten erstellt, umfasst einige 100.000 bis wenige Millionen Worte und erfordert einen erheblichen Aufwand an Arbeitszeit und/oder Geld. Für das Arabische ist speziell die Interaktion zwischen der Orthographie und der Syntax eine Herausforderung, etwa wenn ة bei Antritt eines normalerweise suffigierten Objekts zu ت wird. Diese könnte natürlich auch zum Wortstamm gehören. Auch der Buchstabe ه kann entweder Teil des Wortstammes sein oder ein suffigiertes Objekt der 3. Person männlich Singular. Auch Verdoppelungen und schwache Radikale führen schnell zu einer mehrdeutigen Lösung für isoliert betrachtete arabische Worte.[190]

Für Dialekte müssen die Werkzeuge zur morphologischen Analyse wesentlich angepasst werden. Wenn sie verschriftet vorliegen, dann kann man das in den meisten Fällen an jenen Schreibweisen erkennen, die im Hocharabischen so nicht verwendet werden. Das betrifft vor allem die Verben, deren Konjugation (andere und meist weniger Endungen), Modi (mehr und andere Verbprefixe, dafür keine Bedeutung der Vokalisierung hierfür mehr) und Negation (im Kairenischen und anderen Dialekten als eine Art Zirkumfigierung mit einem vorangestellten ما einem weiteren Buchstaben ش am Ende des Wortes, die Ausdrücke für ein indirektes Objekt können ebenfalls antreten). Zusätzlich erschwert wird die Arbeit mit Varianten des Arabischen, wie eben dem hier untersuchten kairenischen Dialekt, der hier untersucht werden soll, dadurch, dass einige der in Publikationen öfter zitierten Programme nicht im Quellcode zugänglich sind. Das schließt eine Veränderung der Programme zur Unterstützung eines Dialekts von vornherein aus. Einige Programme sind zwar im Quellcode verfügbar, jedoch nur unter einer Lizenz, unter der eine veränderte Version eines eventuell sogar gekauften Programmes nicht weiter verbreitet werden darf. Daher scheiden einige der vorhandenen Werkzeuge für die hier diskutierte Aufgabe aus. Hinzu kommt, dass es im arabischen Raum mit sehr wenig Prestige behaftet ist, einen Dialekt zu untersuchen. Daher sind hier nochmals weniger Ressourcen zu erwarten.

Einige Ansätze werden in Abdelhadi Soudi, Antal den van Bosch, Günter Neumann (2007) vorgestellt.

## Übersicht über einige Programme

Für das Hocharabische wurden in der Vergangenheit bereits einige Werkzeuge zur computergestützten Verarbeitung entwickelt[191]:

Die einfachste Form der Verarbeitung bieten Stemmer. Sie versuchen für Zeichenfolgen den Stamm zurückzugeben, vereinfacht gesagt also, genau das, was man in indoeuropäischen Sprachen meistens im Wörterbuch findet. Stemmer versuchen einem Wort genau einen Stamm zuzuordnen. Dies ist, bedenkt man, dass eine relativ genaue morphologische Analyse schon das eine oder andere Mal mehrdeutig ausfallen wird, eine grobe Vereinfachung der Realität. Als Grundlage für eine weitergehende

[189] Lemnitzer, Zinsmeister (2010), S. 138–139.
[190] Habash (2010), S. 39–63.
[191] Imad A. Al-Sughaiyer, Ibrahim A. Al-Kharashi: Arabic morphological analysis techniques: A comprehensive survey, in: *Journal of the American Society for Information Science and Technology* 2004 (2004), S. 189–213.

computergestützte Verarbeitung von natürlichen Sprachen ist zu aller erst eine Klassifizierung der Worte nach ihrer Funktion im Satz notwendig (Part-Of-Speech tagging). Dies erfordert im Arabischen, wie in allen Sprachen, die morphologisch komplexer als Englisch sind, eine morphologische Analyse. Diese an sich ist noch kein direkt verwendbares Ergebnis, da oft vieldeutig, aber eine unverzichtbare Grundlage.[192] Es sind allerdings auch einige Stemmer verfügbar, von denen hier zwei genannt werden sollen, da sie im Quelltext vorliegen und eine weiterer auch weil er schnell beschrieben ist. Dieser Letztgenannte ist ein sogenannter light stemmer. Sein Algorithmus ist besonders simpel und oberflächlich. Er entfernt lediglich alle Präfixe فال, كال, بال, وال, الund و dann entfernt er die Suffixe ة, ه, ية, يه, ين, ون, ات, ان, هاund ي . Dies wird lediglich mit wenigen Normalisierungen und Einschränkungen, was die Wortlänge betrifft, kombiniert.[193] Bei den anderen beiden handelt sich zum einen um den Arabic Stemmer von Shereen Khoja (http://zeus.cs.pacificu.edu/shereen/research.htm), *Khoja Stemmer* genannt, und zum anderen um den *ISRI Stemmer.* Dessen Algorithmus ist in [194] beschrieben und eine Umsetzung findet sich im Natural Language Toolkit für Python http://www.nltk.org/ . Während der Khoja Stemmer noch mit einem überschaubaren Lexikon an speziell zu behandelnden Wurzeln arbeitet, versucht der ISRI Stemmer ganz ohne Eingabedaten, nur mit einem Satz an Schemata direkt im Programmcode eine korrekte Lösung zu liefern.

Für die morphologische Analyse des Hocharabischen sollen hier beispielhaft folgende Programme genannt werden:

*BAMA* (Buckwalter Arabic Morphological Analyzer, 1.0 und 2.0) und sein Nachfolger *SAMA* (LDC Standard Arabic Morphological Analyzer, 3.0). Diese wurden vom und für das Linguistic Data Consortium, einer Plattform betreut von der University of Pennsylvania für die Bereitstellung linguistischer Softwarewerkzeuge und Daten. Beide verwenden die Programmiersprache Perl.

Von der Version 1.0 des BAMA wurde eine unter GPL v2 verfügbare Variante abgeleitet. Diese trägt den Namen *AraMorph* und ist im Internet frei verfügbar (http://sourceforge.net/projects/aramorph/). Ein Problem der dieser frei verfügbaren Version ist, dass sie stark auf die Codierung cp1256, wie sie für arabische Windows-Systeme üblich ist, setzt und damit unter anderem etwa auf einem deutschen Windows-System erst einmal keine Ergebnisse liefert. Außerdem gibt es eine frei verfügbare (GPL v3) Java Implementierung, die den Namen denselben Namen, AraMorph, trägt (http://www.nongnu.org/aramorph/english/index.html).

BAMA 1.0 und somit auch AraMorph haben bekannte Schwächen[195], gehören heute aber immer noch zu den weiter verbreiteten Werkzeugen für die Verarbeitung des Hocharabischen. Vor allem die von Buckwalter benutzte Codierung der Ergebnisse und teilweise derEingabedaten, sowie seine orthographische Transliteration, mit den 26 Buchstaben des lateinischen Alphabets und einigen Sonderzeichen wie $ und {} war bis zur breiteren Einführung von Unicode eine Voraussetzung für sinnvolles computerunterstütztes Arbeiten. Diese Buchstaben sind im ASCII (American Standard Code for Information Interchange) definiert, das seit den 1980er Jahren bis weit in die 2000er Jahre die

[192] Habash (2010), S. 65.
[193] Leah S. Larkey, Lisa Ballesteros, Margaret E. Connell: Improving Stemming for Arabic Information Retrieval: Light Stemming and Co-occurrence Analysis. http://ciir.cs.umass.edu/pubfiles/ir-249.pdf (Zugriff am: 10. 01. 2013).
[194] Kazem Taghva, Rania Elkhoury, Jeȉrey Coombs: Arabic Stemming Without A Root Dictionary. http://jeffcoombs.com/isri/Taghva2005b.pdf (Zugriff am: 10. 01. 2013).
[195] Mohammad Altabbaa, Ammar Al-Zaraee, Mohammad Arif Shukairy: An Arabic Morphological Analyzer and Part-Of-Speech Tagger. قُطُوف.http://qutuf.com/Files/Report_1.06.pdf (Zugriff am: 08. 01. 2013).

Darstellung von Text auf Computern und im Internet dominierte.[196] Heute wird diese Darstellung noch gerne für die Darstellung von arabischem Text verwendet, wenn arabische Schriftzeichen nicht als nutzbringend angesehen werden, etwa wenn eine deutliche Darstellung von unterschiedlicher Vokalisierung gewünscht ist. Es sind mehrere Varianten aufgekommen[197], da einige Sonderzeichen schon Bedeutung in Programmiersprachen haben und somit mehr Aufwand beim Programmieren nötig ist, um diese zu verarbeiten.[198]

Ein Programm jüngeren Datums, dessen Funktion als ausgezeichnet bewertet wurde, ist *Alkhalil Morpho Sys* (http://sourceforge.net/projects/alkhalil/). Dieses ist in Java geschrieben und ebenfalls für arabische Windows-Systeme eingerichtet. Was bei diesem Programm hinzukommt, ist, dass es das arabische Beschreibungssystem der arabischen Sprache verwendet. Dieses ist mit dem auf das Lateinische zurückgehende nicht direkt vergleichbar (Beispiel: Bezeichnung رفع sowohl für den Indikativ des Verbs als auch den Nominativ des Nomens). Obwohl das Programm intern Kategorien feiner differenziert als in der Ausgabe ersichtlich, ist in der vorliegenden Version ein Weiterverarbeiten der Ergebnisse durch Programme, die auch für andere Sprachen und deren lateinische Beschreibungssystematik ausgelegt sind, erst einmal nicht möglich.

Ein weiterer Ansatz wurde in Prag entwickelt: ElixirFM (http://sourceforge.net/apps/trac/elixir-fm/). Bei diesem Programm wird einerseits das Lexikon, das Buckwalter erstellt hat, benutzt, andererseits wird dieses mit einem Ansatz zur besseren Lesbarkeit im Hinterkopf umgeformt und erweitert. Die Ausgabe der Ergebnisse erfolgt strukturiert als Baum.[199] Der Autor des Programms glaubt, dass es sich relativ einfach für weitere Varietäten anpassen lässt.[200] Eine der großen Herausforderungen bei einer Anpassung ist, dass der Kern des Programms, der die Regeln zur Erkennung der Worte enthält, in der heute eher selten verwendeten Programmiersprache Haskell geschrieben ist.[201] Da aber ElixirFM mit allen seinen Komponenten im Quellcode unter GPL und anderen Lizenzen frei verfügbar ist und seine Varianten weiterverbreitet werden dürfen, wenn diese ebenfalls mit Quelltexten zugänglich gemacht werden, könnte dieses Programm in Zukunft eine größere Rolle spielen. Das Projekt wird jedenfalls noch von Otokar Smrž betreut, der es gestartet hat.[202]

Die erwähnten Programme kann man als lexikonbasiert bezeichnen. Ohne ein ausführliches Lexikon mit Angaben zu Wurzeln und den damit möglichen Wortbildungen, also der Verbindung zu den Schemata, liefern diese Programme keine sinnvollen Ergebnisse.

Ausdrücklich für arabische Dialekte ist bisher nur ein Werkzeug entwickelt worden:

---

[196] W3C: utf-8 Growth On The Web. http://www.w3.org/QA/2008/05/utf8-web-growth.html (Zugriff am: 06. 01. 2013).
[197] Habash (2010), S. 25f.
[198] Habash (2010), S. 20f.
[199] Habash (2010), S. 75f.
[200] Otakar Smrž: Functional Arabic Morphology. Formal System and Implementation. Prag, 2007, S. 97.
[201] Diese Programmiersprache ist keine der heute allgemein benutzten Sprachen und folgt auch mit dem funktionalen Modell nicht dem verbreiteten prozeduralen und objektorientierten Model, wie etwa C, C++, Java, Objective-C, Perl, Python usw. Andererseits sind einige heute relevante Programmiersprachen funktional, allen voran XSLT und XQuery. Bei XSLT kann man vermuten, dass diese nicht als Programmiersprache im eigentlichen Sinn wahrgenommen wird.
[202] Otakar Smrž: ElixirFM / Code / Commit [e0f0bd]. http://sourceforge.net/p/elixir-fm/code/ci/e0f0bdef24fac74815c795263b2ee73755bb44de/ (Zugriff am: 08. 01. 2013).

*MAGEAD (*Morphological Analyzer and Generator for Arabic and its Dialects). Es benutzt Finite State Transducers, Zustandsautomaten, die „vorwärts“ und „rückwärts“ ablaufen können. Damit kann man von einer Analyse zu einem Wort und umgekehrt kommen, wenn einmal die entsprechenden Regeln erstellt wurden. Dieser Ansatz wurde für verschiedene Sprachen benutzt, da die Software, die diese Zustandsautomaten erzeugt und ablaufen lässt (ATT FSM Toolkit zusammen mit den Lextools), lange Zeit gut verfügbar war. Als Beispiel werden Regelsätze für modernes Standard-Arabisch (MSA) und für levantinisches Arabisch mitgeliefert.

*MAGEAD* hat aus heutiger Sicht allerdings einen schweren Nachteil: Es stützt sich in seiner Funktion auf Software, die nicht im Quellcode verfügbar ist und die der Hersteller nicht mehr öffentlich zugänglich vorhält (ATT Lextools) oder wartet (das gilt auch für das ATT FSM Toolkit) . Daneben ist das Programm nur in einer Version 0.5 im Internet auffindbar. Der Autor selbst hält es offenbar nur für halb fertig: Es kann nur mit Verben umgehen, nicht jedoch mit andern Wortformen.

Wären die nun beschriebenen Werkzeuge für Dialekte anpassbar, dann könnte man über eine weitergehende Verarbeitung nachdenken, wie etwa die automatische Bestimmung der Funktion eines Wortes im Satz (Part Of Speech, POS) oder eine automatisierte oder computerunterstützte hierarchische Aufarbeitung von Satzgliedern. Vieles davon ist für Englisch heute tatsächlich möglich. Bei anderen Sprachen ist es oft noch nicht möglich. Bei Hocharabischen ist noch eine POS-Bestimmung möglich. Dazu wurden Werkzeuge entwickelt, die auf BAMA bzw. SAMA oder besser dessen Wörterbüchern aufbauen wie etwa MADA+TOKAN (http://www1.ccls.columbia.edu/MADA/index.html). Dieses Toolkit verwendet seinen eigenen morphological Analyzer ALMOR oder Almorgeana. [203] Ein anderes Werkzeug für diesen Zweck von derselben Universität, der Columbia University, ist AMIRA 2.1 (http://www.flintbox.com/public/project/8335/). Hier werden keine morphologischen Analysen und Regelwerke verwendet, sondern AMIRA „lernt“ aus einem Corpus.[204] Dieser Ansatz würde sich auch für Dialekte anbieten, nachdem man ein entsprechendes Referenzcorpus erstellt hat.

Der nächste Schritt in der Aufbereitung von Textdaten ist dann auf Ebene der Syntax die Erstellung von Baumbanken, in denen die Satzstruktur dargestellt ist. Solche gibt es für Hocharabisch von der Pennsylvania State University, der Columbia University und der Karls Universität Prag.[205]

Auf Ebene der Semantik ist nach dem Vorbild des englischen WordNet ein Arabic WordNet entstanden (http://www.globalwordnet.org/AWN/).[206]

Schließlich gibt es heute auch mehr oder weniger gelungene Ansätze, um mit dem Computer Arabisch in andere Sprachen zu übersetzen. Diese statistischen oder regelbasierten Methoden kennt man im Internet von Google Translate oder Bing Translator. Die Qualität der Übersetzungen ist durchwachsen. Das kann man durchaus direkt auf die Zuverlässigkeit der Technologien zurückführen, auf denen die Übersetzung aufbaut, und diese befinden sich wie oben beschrieben in Entwicklung.[207]

[203] Habash (2010), S. 86–89.
[204] Habash (2010), S. 89f.
[205] Habash (2010), S. 93–112.
[206] Habash (2010), S. 113–118.
[207] Habash (2010), S. 119–124.

# Über Standards und Normen zur Repräsentation digitaler Wörterbücher

Herbert Ernst Wiegand (2010) beschreibt Invarianten, die er bei der Struktur von Einträgen in Wörterbüchern identifizieren konnte. Er kommt zu dem Schluss, dass sich zur Beschreibung der Struktur von Wörterbucheinträgen Baumstrukturen gut eignen könnten.[208] Diese Arbeit liefert eine mögliche theoretische Erklärung dafür, warum es möglich und sinnvoll ist, Wörterbucheinträge in einem standardisierten Format, das eine Baumstruktur darstellen kann, zu verwalten und zugänglich zu machen. Durch eine derart standardisierte Austauschmöglichkeit von Wörterbuchdaten können aus den Verknüpfunge der Daten neue Erkenntnisse gewonnen werden können.

Die ISO (International Organization for Standardization) ist die größte internationale Organisation, die international einheitliche Standards und Normen entwickelt. Die Einhaltung der Standards und Normen ist freiwillig. Die ISO ist im Prinzip nicht für die Bereiche Elektronik und Kommunikation zuständig, wo es eigene Normungsinstitutionen gibt, IEC bzw. ITU.[209] Die ISO gibt aber durchaus Normen für digital genutzte Austauschformate heraus, wie etwa Bürodokumente, heraus. Dort stehen mehrere ISO-Normen zur Auswahl.

Bei der Mehrzahl dieser Wörterbücher werden die Daten wohl in keinem standardisierten Format vorliegen bevor sie als gedrucktes Werk erscheinen, auch wenn das Ergebnis letztendlich für sich spricht. Trotzdem gibt es seit Langem eine Norm für die Gestaltung von Wörterbucherdaten herausgegeben von der ISO, die gemeinhin als internationale Autorität für Standards und Normen aller Art gilt.

## ISO 1951

Für die Repräsentation von Wörterbuchdaten gibt es seit 1973 die Norm ISO 1951 „Lexicographical symbols particularly for use in classified defining vocabularies“[210] Diese Norm wurde gepflegt und überarbeitet, das letzte Mal 2007. Die in der letzten Version vorgeschlagene Repräsentation nutzt XML. Allein, es ist nur von wenigen Firmen bekannt, dass sie diese Norm auch einsetzen. Einer der wenigen ist der Langenscheidt Verlag.[211] ISO 1951 findet in der lexikographischen Community heute keine Beachtung mehr.

## LMF oder ISO 24613[212]

Ein Grund dafür ist, dass die ISO eine zweite Norm hat, die sich etwas weiter gefasst mit Daten im NLP beschäftigt und von Anfang an für lexikograpghische Zwecke ausgelegt war. LMF (Lexical Markup Framework) oder ISO 24613 in der Version von 2008. Ein nicht zu unterschätzender Vorteil dieser Norm ist, dass eine der letzten Überarbeitungsrevisionen für den offiziellen Standard online zugänglich ist[213], wohingegen die vorher erwähnte, wie die meisten ISO-Normen, ausschließlich gegen eine nicht geringe Gebühr erhältlich ist.

---

[208] Herbert Ernst Wiegand: Semantik, Pragmatik und Wörterbuchform in einsprachigen Wörterbüchern, in: *Zeitschrift für germanistische Linguistik* 38 (2010), S. 405–441, hier S. 290–441.

[209] International Organization for Standardization: About ISO. http://www.iso.org/iso/home/about.htm (Zugriff am: 18. 11. 2012).

[210] http://www.iso.org/iso/home/store/catalogue_ics/catalogue_detail_ics.htm?csnumber=6677 (Zugriff am: 11. 01. 2013)

[211] Marie-Jeanne Derouin, André Le Meur: Presentation/Representation of Entries in Dictionaries. http://www.lrec-conf.org/proceedings/lrec2002/pdf/344.pdf (Zugriff am: 12. 01. 2013).

[212] Der folgende Abschnitt basiert in wesentlichen Teilen auf ISO/TC 37/SC4: Language resource management — Lexical markup framework (LMF). Rev 16.
http://www.tagmatica.fr/lmf/iso_tc37_sc4_n453_rev16_FDIS_24613_LMF.pdf (Zugriff am: 13. 01 2013).

[213] http://www.lexicalmarkupframework.org/ (Zugriff am: 12. 01. 2013)

Die Norm ISO 24613 versucht eine optimale Produktion, Wartung und Erweiterung von lexikalischen Ressourcen bei Sprachtechnologien (Human Language Technologies, HLT) und Technologien zur Sprachverarbeitung (NLP) sicherzustellen. Dazu wird ein Metamodell für den lexikalischen Teil dieser Technologiegebiete definiert. Es soll ein gemeinsames, standardisiertes Gerüst für die Erstellung von Computerlexika darstellen. Es stellt sicher, dass einmal eingegebene Daten in verschiedenen Anwendungen und Einsatzgebieten wiederverwendbar sind. Es stellt außerdem eine gemeinsame Repräsentation lexikalischer Objekte bereit, die morphologische, syntaktische und semantische Aspekte umfassen. Dies soll sowohl für kleine, als auch für große Projekte möglich sein.Durch die standardisierte Struktur der Daten soll ein Austausch erleichtert werden. So sollen die auf der ganzen Welt elektronisch vorhandenen Sprachressourcen miteinander verknüpfbar gemacht werden. Das große Ziel wäre, dass alle diese lexikalischen Sprachressourcen von jedem, der sich an diesen Standard hält, genutzt werden können, die Daten wären also leicht austauschbar.

Im Kern definiert das LMF vier Dinge:

- Kategorisierungen, die von verschiedenen Teilen des Frameworks benutzt werden;
- die Rahmenbedingungen oder Einschränkungen für die Beziehungen dieser Kategorisierungen zum Metamodell und dessen Erweiterungen. Begriffe für Kategorisierungen, sowohl die jene, die in der Norm direkt verwendeten werden, als auch neue Begriffe, die sich im Laufe der Anwendung der Norm ergeben, werden in einem zentralen Verzeichnis unter http://www.isocat.org verwaltet. Dort können alle derzeit verfügbaren Kategorisierungen eingesehen werden;
- Normen um diese Kategorisierungen auszudrücken;
- ein Vokabular, um auszudrücken, dass bestimmte Informationen miteinander verbunden sind und wie solche Verbindungen für Erweiterungen des Modells angepasst werden können. Außerdem umfasst es Methoden zur Analyse und zum Entwurf solcher untereinander verbundenen Systeme.

In Anhängen sind fertige Erweiterungen enthalten für:

- maschinenlesbare Wörterbücher;
- lexikalische Ressourcen für das NLP (etwa zum Ausdruck von Syntax und Semantik, Morphologie und speziell morphologische Schemata oder multilinguale Verknüpfungen über den Sinn oder das Verhalten eines Lexems).

Konkrete Realisierungen des LMF können etwa einfache elektronische Wörterbücher umfassen oder komplexere Formen, die sich besser für NLP und maschinelle Übersetzung eignen, also etwa mono-, bi-, oder multilinguale elektronische Wörterbücher. LMF bietet eine Norm für den Entwurf und die Analyse solcher Ressourcen, aber keine konkrete Implementierung, und die Möglichkeit bestehende lexikalische Ressourcen zu beschreiben. LMF bietet also auch die Rahmenbedingungen, die es ermöglichen, Daten, die schon in einem konkreten Format vorliegen mittels des LMF darzustellen und so den Datenaustausch zu ermöglichen.[214] Eine konkrete Umsetzung von LMF für Arabisch wurde beschrieben von Khemakhem et al.[215]

[214] ISO/TC 37/SC4: Language resource management — Lexical markup framework (LMF). Rev 16. http://www.tagmatica.fr/lmf/iso_tc37_sc4_n453_rev16_FDIS_24613_LMF.pdf (Zugriff am: 13. 01 2013).
[215] Aida Khemakhem, Bilel Gargouri, Abdelmajid Ben Hamadou: LMF standardized dictionary for Arabic Language. http://www.taibahu.edu.sa/iccit/allICCITpapers/pdf/p522-khemakhem.pdf (Zugriff am: 22. 01. 2013).

## TEI[216]

Die TEI-Richtlinien sind kein Standard im eigentlichen Sinn einer eindeutigen und reproduzierbaren vorgegebenen Handlungsanweisung oder festgelegten Form. Die Gemeinschaft, die diesen „Standard" pflegt und weiterentwickelt, will ihn nicht als „single fully specified encoding scheme for use in a well-defined application domain" [217] sehen, etwas was die meisten „Industriestandards" auszeichnet. Das TEI Consortium (http://www.tei-c.org/index.xml), das als Gremium über die im Internet veröffentlichte Dokumentation wacht, gibt nur einen sehr detaillierten Satz von Richtlinien heraus, zurzeit. Zurzeit ist es das TEI P5 in der Version 2.3.0[218]. Es bleibt den Benutzern dieser Richtlinien überlassen welche von mehreren möglichen Varianten sie benutzen um ihre Texte so zu codieren, wie es ihren Problemstellungen am ehesten entspricht. Mit diesen Richtlinien ist gewährleistet, dass zwar alle Interessierten einen gemeinsamen Referenzrahmen haben und sich über genau definierte Dinge unterhalten können. Es ist aber nicht das Ziel der Richtlinien, die individuellen Entscheidungen eines Forschers, was er wie in seinen Dokumenten darstellt, einzuschränken. Daraus ergibt sich nur eine negative Folge: Es existieren nur ganz wenige fertig verfügbare Werkzeuge, die TEI umfassend auf einer höheren Abstraktionsebene bearbeiten können, z. B. ähnlich einem Textverarbeitungsprogramm, und das Ergebnis ist bei Verwendung bestimmter Teile der TEI-Richtlinien weit von einem einfach publizierbaren Endresultat entfernt[219]. So bieten sich Dokumente, die auf den TEI-Richtlinien basieren, vor allem für die Archivierung und für den Austausch von Daten an.

Ein TEI-Dokument besteht im Wesentlichen aus zwei Bereichen: Im Kopf (teiheader) sind alle Metainformationen ersichtlich, inklusive dem Autor, der Revisionsgeschichte und der Angabe der rechtlichen Rahmenbedingungen, unter denen diese Dokument verwendet werden kann. Darunter folgt dann der eigentlichen Text, oft eine Digitalisierung eines Drucks, eine Edition von Handschriften und anderes mehr. Dieser kann in einen Vorspann, einen Hauptteil und einen Anhang unterteilt sein, aber nur der Hauptteil muss vorhanden sein. Die TEI-Richtlinien werden in der von der TEI vorgeschlagenen Sprache als XML Dokument gepflegt.

XML ist ein im Wesentlichen ein sehr einfach gehaltener Standard für das Internet (W3C Standard), der die Darstellung von Daten in Baumstrukturen im Sinne eines gewurzelten Baumes der Graphentheorie ermöglicht. Für das was im XML Standard spezifiziert ist existieren nur einige wenige, sehr abstrakte Anwendungsmöglichkeiten. XML selbst definiert noch keine konkrete Auszeichnungssprache, es ist vielmehr ein Metasstandard, welcher die Definition von Auszeichnungssprachen ermöglicht. Standards oder Richtlinien, die für ein bestimmtes Anwendungsgebiet entwickelt wurden, wie etwa TEI oder auch XHTML, bauen auf XML auf. Für die Verarbeitung von in XML aufbereiteten Daten existieren mittlerweile unzählige Werkzeuge, spezialisierte wie der Webbrowser für XHTML oder allgemeinere wie XML Editoren oder XML-Datenbanken, die dann TEI oder beliebige andere Arten von Daten, die in XML dargestellt sind, verarbeiten können. Der Metastandard XML ermöglicht es Programmen ab einem bestimmten Grad von Abstraktion vom eigentlichen Aufgabengebiet der Software auf ein umfangreiches und gut getestetes Arsenal von Standardsoftwarekomponenten zurückzugreifen. Im Grunde genommen

[216] Der folgende Abschnitt bezieht sich in wesentlichen Teilen auf TEI Consortium (eds): P5: Guidelines for Electronic Text Encoding and Interchange. http://www.tei-c.org/Vault/P5/2.2.0/doc/tei-p5-doc/en/html/ (Zugriff am: 13. 01. 2013).

[217] TEI-Guidelines, „Design Principles". http://www.tei-c.org/release/doc/tei-p5-doc/en/html/AB.html#ABTEI2 (Zugriff am 07. 02. 2012).

[218] Diese wie auch alle vorherigen Versionen findet man unter http://sourceforge.net/projects/tei/files/TEI-P5-all/.

[219] Etwa ist es dem Autor nicht gelungen schnell eine Druckreife Darstellung für Wörterbücher zu bekommen indem ein durchaus diesen Richtlinien entsprechende XML Dokument mit den angebotenen Werkzeugen in ein Word-Dokument umgewandelt wurde.

ist heute jedes Dokument einer Textverarbeitung eine kleine Ansammlung von XML Dokumenten. XML hatte und hat ein Ziel: Das Format soll einerseits von Maschinen verarbeitbar sein und andererseits immer noch von Menschen gelesen werden können, was etwa bei Formaten, in denen etwa die mit früheren Textverarbeitungsprogrammen erstellten Daten vorliegen, nicht gewährleistet ist.

Insofern kann auch TEI als ein weiterer Schritt hin zu einer Form gesehen werden, die Menschen gut lesen können sollen und bei denen sich der Sinn einer Bezeichnung für ein Stück Text direkt erschließen soll. Benutzen Textverarbeitungen abstrakte bzw. dem Druckereiwesen entnommene Beschreibungen für den Inhalt eines Dokuments, haben die von den TEI-Richtlinien vorgeschlagenen Bezeichner für Komponenten eines Texts (Tags, „Fähnchen“) einen hohen, genau definierten semantischen Gehalt. Nimmt man die TEI-Richtlinien mit der dort recht ausführlichen Beschreibung und den Beispielen hinzu, ist jedenfalls bei den schon länger verwendeten Teilbereichen - wie dem für Textstruktur – klar, wofür eine bestimmte Auszeichnung steht und wofür nicht.

TEI enthält eine Unterkategorie speziell zur Darstellung von Wörterbüchern. Es ist auch schon mit sehr positiven Aussichten untersucht worden, inwiefern sich TEI in LMF überführen lässt.[220] Der Bereich Wörterbücher in den TEI P5 Richtlinien ist, was die vorhandenen, allgemein verfügbaren Werkzeuge zur Bearbeitung und Weiterverarbeitung von TEI Dokumenten angeht, immer noch in der Entwicklung, das heißt, es sind noch keine Werkzeuge verfügbar, die eine einfache Eingabe von Wörterbucheinträgen oder eine einfache Publikation ermöglichen.

Ein einfacher Eintrag eines einsprachigen englischen Wörterbuchs sieht also etwa so aus:

```
<entry>
  <form>
    <orth>competitor</orth>
    <hyph>com|peti|tor</hyph>
    <pron>k@m"petit@(r)</pron>
  </form>
  <gramGrp>
    <pos>n</pos>
  </gramGrp>
  <def>person who competes.</def>
</entry>
```
[221]

In mehrsprachigen Wörterbüchern soll <sense n="x"/> verwendet werden, um mehrere mögliche dem Sinn nach unterschiedliche Übersetzungen anzugeben, wobei mit n, einer Zahl, die verschiedenen möglichen Bedeutungen unterschieden werden.

Wie das TEI-Dokument genau aussieht, ist zurzeit oft abhängig davon, mit welchen konkreten Werkzeugen darauf zugegriffen werden soll oder mit welchen es in eine von gedruckten Wörterbüchern gewohnte Darstellung umgewandelt werden soll.

[220] Laurent Romary, Inria & HUB-IDSL: TEI and LMF crosswalks. http://hal.inria.fr/docs/00/77/28/01/PDF/TEI_and_LMF_crosswalks.pdf (Zugriff am: 13. 01. 2013).
[221] TEI Consortium (eds): P5: Guidelines for Electronic Text Encoding and Interchange. http://www.tei-c.org/Vault/P5/2.2.0/doc/tei-p5-doc/en/html/ (Zugriff am: 13. 01. 2013).

Da bis jetzt nur ganz wenige Forscher den Bedarf sahen Frequenzangaben zu Wörtern zu codieren, gibt es in den Richtlinien noch keinen allgemeinen Vorschlag, wie man eine solche Angabe aufnehmen könnte[222].

## RDF und OWL

„Industrietandards" im Sinne einer festgelegten Darstellungsform, die zur Darstellung lexikographischer Daten dienen können, sind hingegen RDF (Resource Data Framework) und OWL (Web Ontology Language). Diese stehen im Zusammenhang mit Bemühungen um das „Semantic Web", also einer Darstellungsform von Daten im Internet, die es, ähnlich wie auch HTML, ermöglicht, Daten im World Wide Web, miteinander zu verknüpfen. Statt aber dem Benutzer das Erkennen des Sinns hinter den Verknüpfungen zu überantworten, sollen im „Semantic Web" Maschinen dazu in die Lage versetzt werden, Daten nach semantisch Kriterien zu verarbeiten und so den Maschinen eine „intelligente" Verknüpfung der Daten zu ermöglichen.[223] Es gibt auch einen Vorschlag LMF mittels RDF umzusetzen.[224]

Diese Standards sind seit einiger Zeit auch für die Macher der verschiedenen WordNet Varianten interessant. Die Idee hinter WordNet ist es Beziehungen zwischen Worten festzuhalten und diese Beziehungsdaten elektronisch und im Internet zugänglich zu machen. Hierfür bieten sich dieselben Konzepte an, die hinter der Idee des „Semantic Web" stehen und damit auch die Standards mit denen diese Konzepte ausgedrückt werden.[225]

Während aber die entsprechenden XML Sprachen standardisiert sind und so auch Werkzeuge zur Abfrage der entsprechend aufbereiteten Daten existieren[226], ist für das „Semantic Web" eine leicht auffindbare Liste der verfügbaren Ressourcen noch nicht vorhanden, ebenso wenig wie einfach handhabbare Werkzeuge um Daten entsprechend aufzubereiten.

## Zur Frage der Identifikation linguistischer Varietäten

Auch für die Identifikation von Sprachen gibt es eine ISO-Norm. ISO 639 ist eine Gruppe von Standards, welche die zwei und drei (Latein-)Buchstaben langen Abkürzungen für Sprachen bzw. den Prozess zur Registrierung solcher definiert. Diese Kürzel sind für Terminologie, Lexikografie und Linguistik gedacht. So werden sie etwa in Bibliotheken zur Kennzeichnung der Sprache von Werken verwendet. Die drei erwähnten Teile werden von drei verschiedenen Registrierungsautoriäten oder Registraren verwaltet.

- ISO 639 – 1 enthält die Abkürzungen für die am weitesten verbreiteten und bekanntesten Sprachen der Welt. Diese werden mit Kürzeln aus zwei Buchstaben dargestellt, also etwa en, fr, de, ar, tr oder cu (Tschuwaschisch). Diese Bezeichner werden von Infoterm vergeben, dem internationalen Informationszentrum für Terminologie, einem Verein mit Sitz in Wien.[227]
- ISO 639 – 2 enthält Abkürzungen für Terminologie und Bibliografie für die alle Sprachen aus ISO 639 – 1 sowie zusätzliche Abkürzungen für alle Sprachen, in denen es eine bedeutende Menge

[222] Vgl. http://listserv.brown.edu/archives/cgi-bin/wa?A2=TEI-L;hn%2BksQ;20071227081957%2B0100 und http://listserv.brown.edu/archives/cgi-bin/wa?A2=TEI-L;Q7vTuA;20071217190037%2B0000

[223] Dean Allemang, James A. Hendler: Semantic web for the working ontologist. Effective modeling in RDFS and OWL. 2. Auflage. Amsterdam, 2011, S. 1–12.

[224] Gil Francopoulo: Strategy for an OWL specification of LMF. http://www.tagmatica.fr/lmf/StrategyForLMFInOWL29october2007.pdf (Zugriff am: 22. 01. 2013).

[225] Lothar Lemnitzer, Claudia Kunze: Integrating Wordnets into the Resource Description Framework. http://www.cs.vassar.edu/~ide/events/NLPXML03_review/papers/Lemnitzer.pdf (Zugriff am: 22. 01. 2013).

[226] siehe etwa die WordNet Abfrage mit der Abfragesprache SPARQL unter http://factforge.net/sparql

[227] Infoterm - Internationales Informationszentrum für Terminologie: STATUTEN des internationalen Vereines Infoterm - Internationales Informationszentrum für Terminologie. http://www.infoterm.info/pdf/about_us/InfotermStatuten_2011_FV.pdf (Zugriff am: 18. 11. 2012).

an Literatur gibt. Auch sind Bezeichner für Sprachgruppen enthalten, die zusammengenommen alle Sprachen der Welt abdecken sollen. Die Abkürzungen sind drei Buchstaben lang, etwa eng, fra, deu, ara, tur oder chv aber auch enm (Middle English 1100 - 1500), ang (Old English ca. 450 – 1100) und ota (Osmanisch 1500 - 1928) oder trk (Turkic Languages). Verwaltet werden diese Abkürzungen von der „Library of Congress" in Washington D. C., USA.

- ISO 639 – 3 will schließlich alle der Menschheit bekannten Sprachen abdecken. Es wird also ISO 639 – 2 nochmals erweitert und genauso drei Buchstaben Kürzel vergeben. Daher findet man neben eng, fra, deu, ara, tur, chv, enm, ang und ota auch arz (ägyptisches Umgangsarabisch), ary (marokkanisches Umgangsarabisch) und weitere Bezeichner für arabische Dialekte (oder Dialektgruppen) sowie etwa auch als (für Alemannisch, also hauptsächlich Schwyzerdütsch) aber keine Gruppenbezeichner wie etwa trk.[228]

Verwaltet werden die Bezeichner der ISO-Norm 639 – 3 von SIL International mit Sitz in Dallas, Texas, USA. SIL, Summer Institute of Linguistics, Inc wurde 1934 als kleiner Studienkreis gegründet. Sie ist eine religiös motivierte nicht-kommerzielle Organisation, die ihre Aufgabe darin sieht, es Gesellschaften in aller Welt zu ermöglichen, die Entwicklung ihrer Sprache voranzutreiben. Sie stellen ihre Dienste jedem zur Verfügung ohne Ansehen des Glaubens, der politischen Einstellung, des Geschlechts, der Rasse oder des ethnolinguistischen Hintergrunds. Obwohl christlich religiös motiviert, beschränkt sich SIL auf Sprachentwicklung und beteiligt sich weder direkt an Missionierung oder Bekehrungsversuchen noch an der Verbreitung religiöser Schriften.[229]

SIL gibt seit 1951 immer wieder aktualisierte Ausgaben des Nachschlagewerks Ethnologue[230] heraus. Dort ist mindestens seit 1996 ägyptisches Umgangsarabisch als eigene Sprache verzeichnet. Die spätere 15. Auflage des Werks diente als Grundlage für ISO 639 - 3[231], was erklärt, warum es im Rahmen der ISO 639 - 3 als eigenständige Varietät des Arabischen wahrgenommen wird, ohne jeden ägyptischen Einfluss, ob nationalistisch, koptisch.

Eine weitere umfassende Kategorisierung aller bekannten Sprachen und Varietäten, der „Linguasphrere", liegt mit David Dalby, David Barrett, Michael Mann (1999) vor. Dieses Werk ist seit 2012 auch im Internet in einer minimal überarbeiteten Variante abrufbar und darf im Rahmen nicht-kommerzieller Aktivitäten genutzt werden[232]. Anders als etwa der ISO 639 – 3 ist die Einteilung der Sprachen stark hierarchisiert. Es gibt eine äußere Haupteinteilung in Sprachgruppen mit einem Nummernsystem und eine äußere sowei eine innere Unterteilung mittels Buchstabencodes, die auf eine gewisse Weise die Distanz zwischen den einzelnen Sprachen und Sprachvarietäten zum Ausdruck bringen. Es sind über 70000 Referenznamen und Alternativnamen für Sprachen verzeichnet und klassifiziert.

[228] o. A.: ISO 639.2 Registration Authority - Frequently Asked Questions (FAQ). http://www.loc.gov/standards/iso639-2/faq.html#1 (Zugriff am: 18. 11. 2012).
[229] SIL International: What is SIL International? http://www.sil.org/sil/ (Zugriff am: 08. 01. 2013).
[230] Barbara F. Grimes, Richard S. Pittman, Joseph Evans Grimes: Ethnologue. Languages of the world. 13. Auflage. Dallas, Texas, USA, 1996.
[231] o. A.: ISO 639.2 Registration Authority - Frequently Asked Questions (FAQ). http://www.loc.gov/standards/iso639-2/faq.html#1 (Zugriff am: 18. 11. 2012).
[232] David Dalby: The Linguasphere Register of the world's languages and speech communities. http://www.linguasphere.info/lcontao/fichiers-pdf.html (Zugriff am: 04. 02. 2013).

## Graphemische Besonderheiten der ägyptischen Wikipedia

Wie oben erwähnt[233] erwähnt gibt es einen Artikel in der Wikipedia Masry, der Ratschläge für die Orthographie gibt. Für die hier bearbeitete Problemstellung ist diese Hilfestellung für Autoren wenig förderlich. Für Wikipedia ist typisch, dass es keine festen Regeln gibt. Solange alles verständlich bleibt, darf man schreiben was und wie man will. Aber auch wenn sich die Autoren an die Vorschläge halten, werden dadurch manchmal noch zusätzliche Ambiguitäten eingeführt, anstatt die Verschriftung eindeutiger werden zu lassen. Ein Beispiel dafür ware die Schreibung von ة als ه oder ة was zu einer Ambiguität mit dem Possesifsuffix der dritten Person männlich Singular führt das auch ه lauten kann. Auch die vorgeschalgene phonemische Schreibweise für Wörter mit ث oder ذ unter ihren Wurzelkonsonanten, die aber von Ägyptern normalerweise als ت und د ausgesprochen werden, führt zu neuen Kombinationsmöglichkeiten und damit möglicherweise zu neuen Mehrdeutigkeiten. Was eine Disambiguierung und Lemmatisierung von Wörtern auch nicht erleichtert ist, dass es keinen Ratschlag gibt zu vokalisieren. Eine Verarbeitung der Daten wird also komplizierter, weil nicht zugunsten einer klaren Orthographie entschieden wurde, sondern ein sehr liberaler Satz von Richtlinien verwendet wird, um möglichst wenige potenzielle Beitragende vor den Kopf zu stoßen. Der allgemeinen Verständlichkeit wird das in den meisten Fällen keinen Abbruch tun. Für Leser, deren Erstsprache - oder jedenfalls die am häufigsten verwendete Umgangssprache - eine Sprache ist, ist klar, welche Vokale in den Text eingesetzt werden müssen, den sie lesen. Andererseits sind die Texte auch für einen größeren Personenkreis lesbar, denn wirklich markante Unterschiede zwischen Varietäten einzelner Städte oder Regionen bestehen meist in den Vokalen und nicht im Konsonantengerüst. Damit bleiben noch die vielen Schreibvarianten für dasselbe Wort oder dieselbe gramatikalische Konstruktion, die etwa in Rif'at al Farnawānī (1981) aufgezählt sind oder die sich in den Arbeiten von Renate Malina (1987) und jüngst Gabriel M. Rosenbaum (2004) finden. In diesem Sinne folgt die Schreibung der Wikipedia dem üblichen Standard für die Schreibung der ägyptischen arabischen Varietät. Mehr noch, sie folgt jedem „Standard“, den sich die diversen Autoren gegeben haben, deren Werke wiederum die Beitragenden der Wikipedia Masry gelesen haben. Im Wesentlichen machen sie das nach, was ihnen aus ihrer persönlich präferierten Dialektliteratur bekannt ist. Die Wikipedia scheint daher ein gutes Testfeld zu sein, um Werkzeuge zu entwickeln, die flexibel und stabil genug sind, um ein größeres Spektrum an möglichen Schreibweisen des Ägyptischen zu verarbeiten.

Im Grunde genommen sind die Autoren der ägyptischen Wikipedia im Hinblick auf die Orthographie wenig innovativ. Eine Besonderheit was die Textsegmentierung im extrahierten Text angeht hängt mit dem Verfahren zusammen, das benutzt wurde, um allen für Benutzer sichtbaren Text as der Wikipedia Masry zu extrahieren. Dabei werden vor allem bei Links zwischen Artikeln meist alle Zeichenkombinationen abgespalten, die regelmäßig an einen Wortstamm antreten können, sodass dieser dann übrig bleibt. Zum Beispiel wird im Kurzartikel رسم القلب (rasm 'alb, Herzgeräusch) die Zeichenkombination لِلـ , bestehend aus einem Artikel الـ , der präfigiert wird und der Präposition لـ , die ebenfalls mit dem nachfolgenden Wort verbunden wird, sichtbar. So findet sich dort folgender Satz:

[233] Siehe Seite 45

**رسم القلب** هو تسجيل النشاط الكهربى للقلب على مر الزمن ,على ورق الرسم البيانى.

Der Wikitext dazu sieht wie folgt aus:

'''رسم القلب''' هو تسجيل النشاط الكهربى لل[[قلب]] على مر الزمن ,على ورق الرسم البيانى.

Der Autor des Wikipediaeintrags hat hier der Einfachheit halber, die Verknüpfung zum Eintrag über قلب ('alb, Herz) dadurch hergestellt, dass er das Wort, das der Titel des Eintrags ist, auf den verwiesen werden soll, doppelt eingeklammert hat. MediaWiki fügt diese Trennung beim Erstellen des HTML-Codes, den der Browser anzeigt, wieder zusammen. Titel von Wikipediaeinträgen sind, wie auch in der arabischen Wikipedia üblich, immer ohne Artikel und fast zwangsläufig ohne Suffixe. Da für die Autoren die Verlinkung auf diese Weise einfach ist, wird diese Methode von ihnen offenbar gerne angewandt, wie die Einzelzählung einiger Prä- und Suffixe unter den 200 häufigsten Wörtern zeigt. Diese mussten nicht mittels Software erkannt und abgespalten werden, sondern waren bereits durch die Autoren als solche gekennzeichnet.

Diese Abtrennung kann durch die hier präsentierte Methode auch im Text, der der Zählung der Wörter zugrunde liegt, als Leerzeichen erhalten bleiben.

رسم القلب هو تسجيل النشاط الكهربى **لل قلب** على مر الزمن ,على ورق الرسم البيانى.

Natürlich sind dadurch nicht alle Präfixe abgedeckt, die aufgrund der technischen Gegebenhieten von den Autoren abgetrennt wurden, ohne dass diese dies als orthographisch korrekt empfinden würden. Das Präfix لل ist zum Beispiel noch 15100 Mal mit einem Wort verbunden im Corpus aufzufinden.

Von der hocharabischen Orthographie wird verlangt, dass Wörter, die nur aus einem Buchstaben bestehen, mit dem nächsten Wort verbunden geschrieben werden. Interessant ist daher die sehr häufige Getrenntschreibung von و, also „und“ in der Wikipedia Masry. Auch hier kommen sicher wieder eine Menge Fälle hinzu, in denen die Zusammenschreibung benutzt wurde.

## Die Lexik der ägyptischen Wikipedia

Im Rahmen dieser Arbeit konnte keine Software gefunden werden, die mit hinreichender Genauigkeit Worte, die sich lediglich in ihren morphologisch bedingten Formen unterscheiden, sinnvoll gruppiert, also etwa lemmatisiert. Es wurden daher primär Token gezählt und diese dann halb-automatisch nachbearbeitet. Token seien hier definiert als Zeichenketten, die durch Leerzeichen oder Satzzeichen voneinander getrennt sind.

Danach wurde eine umfangreiche Liste von Prä- und Suffixen mit den gefundenen Token kombiniert, um alle möglichen Kombinationen im Corpus zu suchen. Bei diesem Verfahren werden durch die große Zahl an Homographen zwangsläufig auch falsche Kombinationsmöglichkeiten gefunden. Wo dies die Zählung markant verfälschte, wurden für diese Fehler manuell Ausnahmen definiert, die eine Gruppierung verhindern.

Mehrworteinheiten wurden vorerst nicht untersucht.

Zuerst sollen hier dieselben Diagramme, wie für die oben beschriebenen Frequenzwörterbücher angeführt, dargestellt werden. Die Anteile der Wortkategorien wurden in den Diagrammen nicht nach der Häufigkeit in der Wikipedia angeordnet, sondern nach der Häufigkeit in Tim Buckwalter, Dilworth B. Parkinson (2011). Übernommen wurde auch eine Eigenheit: In diesem Werk sind Positiv und Elativ getrennt als Adjektive aufgenommen.

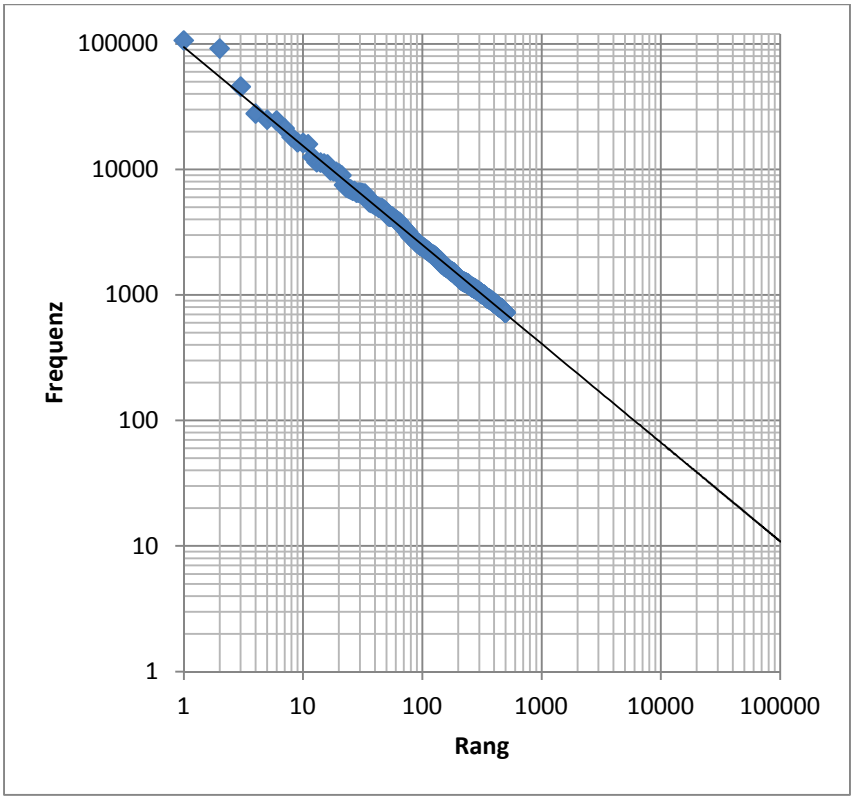
Für drei Wortkategorien konnten unter den 200 häufigsten keine eindeutigen Belege gefunden werden: Interrogativa, Interjektionen und Kardinalzahlen. Die ersten beiden Kategorien mögen vereinzelt in den Artikeln zu finden sein, aber sicher nicht häufig genug um unter den 200 häufigsten Wörtern zu finden zu sein. Dies entspricht dem Stil, den man wohl allgemein von einer Enzyklopädie erwarten würde. Eine Untersuchung der Diskussionen zu den Artikeln würde wohl andere Ergebnisse liefern. Des Weiteren scheint es zum bevorzugten Stil der Wikipedia Masry zu gehören, Zahlen nicht auszuschreiben.

Eine im Vergleich zur Einteilung auf Seite 25 neue Kategorie sind Abkürzungen. Dies deshalb, weil eine gängige Abkürzung eindeutig häufig vorkommt und es drei Abkürzungen mit einer Länge von einem Buchstaben gibt, die im Kontext der Wikipedia eine spezielle Bedeutung haben.

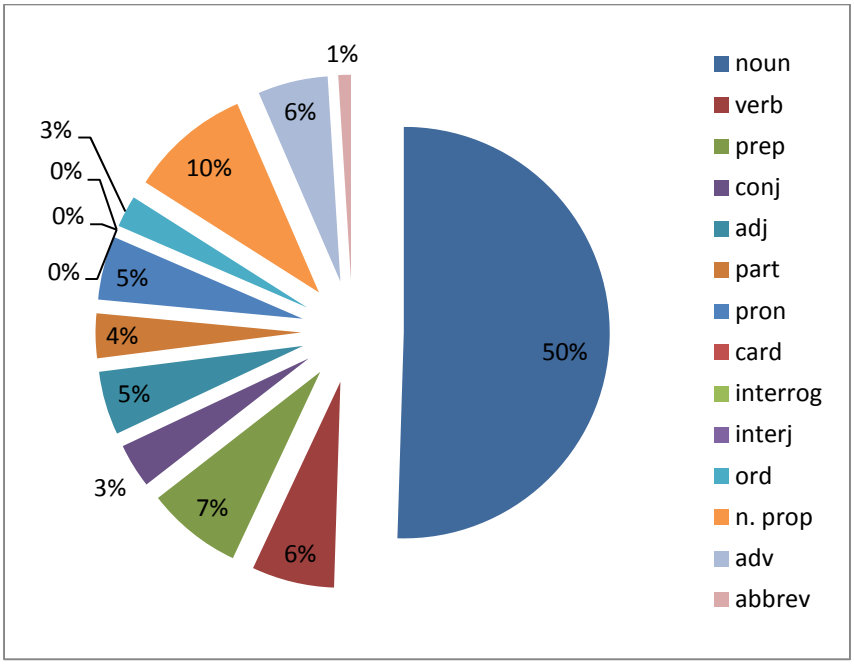
Eine Zuordnung der gefundenen Ergebnisse zu Wissensgebieten ist auf zwei Arten prinzipiell möglich. Es hat sich eingebürgert, dass Seiten nach Möglichkeit einer oder mehreren Kategorien zugeordnet werden. Diese werden am unteren Ende des Artikels mit dem Stichwort تضنيف angegeben. Es gibt in den Daten, die dieser Untersuchung zugrunde liegen 3596 unterschiedliche Einzelkategorien mit teils nur einigen wenigen zugeordneten Artikeln. Aus diesen Einzelkategorien wird eine Hierarchie gebildet sodass Unterkategorien unter einer gemeinsamen Kategorie zusammengefasst und über eine entsprechende Kategorieseite zugänglich sind. Dies ist aber eine manuelle Aufgabe bei der Pflege einer Wikipedia, der aktuelle Zustand dieser Hierarchie wurde nicht untersucht. Diese Hierarchie befindet sich in einem eigenen Namespace, der nicht bearbeitet wurde. Ohne Berücksichtigung der Hierarchie ist eine Zuordnung also nur zu sehr eng abgegrenzten Themenbereichen möglich.

Es gibt aber noch eine zweite Möglichkeit sich einen Überblick über die wichtigen Themen in der Wikipedia Masry zu verschaffen: Bestimmte Worte in der Liste der 200 häufigsten Wörter lassen Rückschlüsse darauf zu, welche Themen in der Wikipedia Masry besonders häufig behandelt werden.

Für die Wikipedia Masry ergibt sich folgendes Bild eines Rang-Frequenz-Diagramms:




Die Verteilung nach Wortkategorien sieht wie folgt aus:




Was kann man nun aus den gefundenen Worten schließen? Zipf und seine Nachfolger lassen uns vermuten, dass aus den meisten Worten nicht sehr viel zu schließen ist. Die Häufigkeit ist sehr hoch, aber die Aussagekraft des Wortes an sich ohne seinen Kontext ist niedrig. Es sind dies jene Worte, die man meistens in Stoppwortlisten findet, wenn es darum geht, möglichst effizient zu suchen.

## Abweichungen von „A Frequency Dictionary of Arabic“

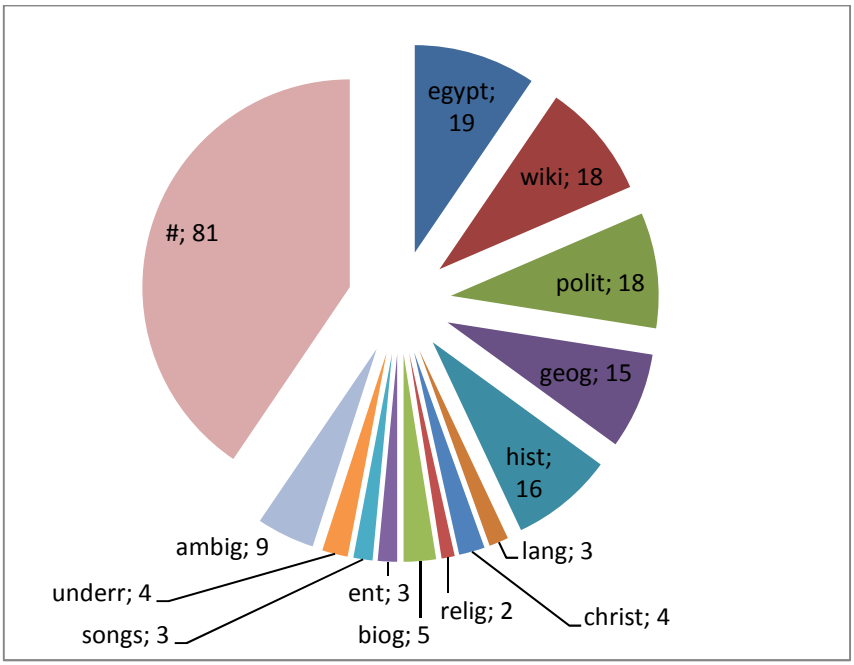
Auffällig ist, dass Nomina propria häufig vorkommen. Dies hat zwei Gründe. Vergleicht man zum Einen etwa die Methodik der Zählung mit der von Tim Buckwalter, Dilworth B. Parkinson (2011) gehören sie zu jenen Worten, die dort in eine eigene Liste ausgelagert und aus der Hauptliste ausgeschieden wurden[234].

Zum Zweiten kommt es in der Wikipedia Masry zu einer Vervielfältigung von Namen. Zum Beispiel in der Kategorie ممثلين, „Schauspieler“, ist bei jedem Schauspieler eine Liste von Schauspielern angegeben, das heißt, diese Liste ist mit der Anzahl von Schauspielern vervielfältigt. Diese Vorgehensweise der Autoren der Wikipedia rechtfertigt durchaus eine Reihung von Namen unter den 200 häufigsten Wörtern, da ein Benutzer mit hoher Wahrscheinlichkeit auf diese Namen stoßen wird. Abgesehen davon deckt sich die Reihung der Namen in etwa mit dem, was in Tim Buckwalter, Dilworth B. Parkinson (2011) zu finden ist. Die vorkommenden Namen entsprechen dem, was üblicherweise als die häufigsten Namen im arabischen Sprachraum angenommen wird (محمد, احمد, حسن, حسين „Muhammad“, „Ahmad“, „Hasan“, „Hussein“).

Die Abkürzung م für ميلادي, also für Zeitangaben nach Christi Geburt, kommt sehr häufig vor. Die andere gebräuchliche Zeitangabe هـ für islamische Zeitangaben findet sich nicht unter den 200 häufigsten Wörtern. Diese Angabe scheint den Autoren dennoch in vielen Artikeln wichtig zu sein, obwohl die in der Wikipedia Masry angegebenen Jahre grundsätzlich nach dem westlichen Kalender zu verstehen sind, so nicht anders angegeben.

## Die häufigsten Themenbereiche

Die folgende Auflistung teilt die Wörter danach ein, welche Themen man an den isoliert betrachteten Wortformen erkennen kann:




[234] Buckwalter, Parkinson (2011), S. 155 und 162.

Unter den 200 Worten sind 81, die keine spezielle Bedeutung haben. Sie kommen so oder jedenfalls in der angegebenen hauptsächlich gebrauchten Bedeutung ebenso im Hocharabischen vor.

19 Worte stellen eindeutige Dialektmarker dar. Darunter finden sich die Pronomen دي ,ده und إللى.

## Für Wikipedia Masry spezifische Ausdrücke

Weitere 18 Worte sind Wikipedia-spezifisch beziehungsweise dem Computerjargon entnommen. Zu beachten ist vor allem, dass sich darunter einige befinden, die den Übersetzern der Oberfläche passend erschienen, heute aber bei den ins Hocharabische übersetzten Computerprogrammen kaum Verwendung finden: darunter لينك und فيلات.

Es finden sich die Transliterationen von Wikipedia und WikiMedia Commons.

Andere Worte sind nur unter den 200 häufigsten Wörtern gelistet, weil sie dank ihrer Verwendung in häufig vorkommenden und Wikipedia typischen Phrasemen in Meldungsboxen häufig vorkommen. Die folgende Textbox in einer Vielzahl von unzugänglichen Artikeln erklärt einige vorkommende Worte. Sie lautet:

„المقاله دى يتيمه علشان مافيش مقالات بتوصل ليها.

"ساعد من فضلك بإضافة وصلات ليها فى المقالات اللى ليها علاقه بيها."

auf Deutsch: „Dieser Artikel ist verwaist, weil es keinen Artikel gibt, der hierher verlinkt. Hilf bitte mittels des Hinzufügens von Links in Artikel, die mit diesem in Verbindung stehen.". Dieser Hinweis wird automatisch von Programmen eingefügt. Ebenso erklären sich die beiden Worte مصادر und تصنيف.

Quellenangaben sollte jeder Artikel in einer Wikipedia bieten und jeder Artikel sollte mindestens einer Kategorie zugeordnet sein. Das ist aber offensichtlich nicht immer der Fall.

Einige der Worte gehen auf Arbeitsanweisungen, für diejenigen, die eine Änderung einpflegen wollen, zurück, etwa: „تحويل قالب:لينك مقاله مختاره" auf Deutsch: „Umstellen auf Vorlage: Link des gewählten Artikels". Vorlagen sorgen wie erwähnt dafür, dass bestimmte Teile einer Seite immer gleich aussehen, etwa die Infoboxen mit den wichtigsten Daten eines Landes.

Die restlichen Worte, die nur im Kontext der Wikipedia so häufig vorzufinden sind, gehen auf folgende Standardmeldung zurück:

„الصفحه دى فيها تقاوى مقاله. و انت ممكن تساعد ويكيپيديا مصرى علشان تكبرها"

auf Deutsch: „Auf dieser Seite gibt es einen Stub (wörtliche Übersetzung des Arabischen: „Samen") eines Artikels. Und du kannst der Wikipedia Masry helfen, sie zu vergrößern."

Zuletzt gibt es noch eine Abkürzung, die so nur in Wikipedia zu finden ist, die Abkürzung ن. Diese kommt in der folgenden Struktur vor: „ ع • ن • ش ". Diese Abkürzungen stehen für: „شوف القالب ده", „ناقش ".

القالب ده“ und „عدّل القالب ده“ auf Deutsch: „Schau dir diese Vorlage an“, „Diskutiere diese Vorlage“ und „Modifiziere diese Vorlage“. Diese Abkürzungen kommen etwa in den Vorlagen der Listen der ägyptische Schauspieler und Sänger vor und ermöglichen so den schnellen Zugriff auf die Vorlage aus dem Artikel über einen Schauspieler oder Sänger heraus.

## Hinweise auf behandelte Themengebiete

Einige Worte deuten auf die Beschreibung von politischen Ereignissen hin, etwa الملك, حكم, رئيس oder قوات, العسكريه und نظام, „König“, „Regierung“, „Präsident“, „(Streit-) Kräfte“, „Militär“ und „System“.

Ein anderes gut erkennbares Themengebiet ist die Geographie. Einerseits sieht man, dass es in dieser Wikipedia, wie zu vermuten war, hauptsächlich um Ägypten geht. المصريه und مصر, „ägyptisch/kairinisch“ und „Ägypten/Kairo“, sind die häufigsten Wörter mit geographischer Bedeutung. Begriffe wie مدينة, منطقة und البحر, „Stadt“, „Region“ und „Meer“, oder alle vier Himmelsrichtungen deuten ebenfalls auf häufige geographische Beschreibungen hin.

Andere wiederum lassen vermuten, dass sich viele Artikel mit historischen Ereignissen und Entwicklungen beschäftigen, wie Monatsnamen, سنة, تاريخ und دلوقتى, „Jahr“, „Geschichte“ und „jetzt bzw. heute“.

Die am häufigsten erwähnten Sprachen sind العربية und انجليزى, „Arabisch“ und „Englisch“.

Sechs Wörter in der Liste weisen eindeutig auf die Beschäftigung mit Religion hin, davon sind fünf der Thematik christliche Religion zuordenbar, etwa der Name يوأنس, „Johannes“, und المسيحيين und المسلمين, also „Christen“ und „Muslime“, wobei Christen häufiger vorkommt. Ebenso kommt der Titel بابا, „Papst, das religiöse Oberhaupt der Kopten“ sehr häufig vor.

Als letztes Themengebiet lässt sich noch Unterhaltung identifizieren. Vor allem فيلم, ممثلين und مسلسل, „Film“, „Schauspieler“ und „Serie“, sind sehr häufig. Daneben finden sich noch einige unerwartete Zeichenketten im Kontext einer Enzyklopädie, etwa يا oder قلبي, die Vokativpartikel („O …“) und „mein Herz“. Dies erklärt sich daraus, dass einige Texte bekannter Lieder auf Kairenisch Einzug in die Wikipedia Masry gehalten haben.

## Probleme bei der Zählung

14 Wörter sind in ihrer Interpretation problematisch. Für einige Wörter, die im Hocharabischen mit dem nächsten Wort verbunden werden, wird eine zu geringe Häufigkeit ausgewiesen. Diese sind zwar in der Wikipedia Masry häufig getrennt zu finden, ein Algorithmus, der diese Wörter abtrennt, wenn sie verbunden geschrieben sind, wurde aber nicht eingesetzt. Trotzdem werden diese kurzen Wörter genau wie im Hocharabischen, wo dies die einzig allgemein anerkannte orthographisch Schreibweise darstellt,

auch oft verbunden geschrieben. Damit sind die Wörter و, بـ und لـ in ihrer Häufigkeit unterrepräsentiert.

Einige der 200 Wörter sind durch das regelmäßige Fehlen von Vokalzeichen mehrdeutig. Der Buchstabe ع kann im Kairenischen entweder als Kurzform von عَلَى, „auf“, verwendet werden oder er steht wie erwähnt für عَدَّل, „modifiziere“. Die Wurzel كتب kann entweder als كَتَبَ, „er schrieb“, oder كُتُب, „Bücher“, gelesen werden. Es kann كِتَابَه, „das Schreiben, die Schreibweise“, nicht von كِتَابُه, „sein Buch“, unterschieden werden. Das bedeutet, dass diese Gruppe mit den hier verwendeten Werkzeugen nicht sinnvoll in zwei Gruppen rund um die Lexeme „schreiben“ und „Buch“ aufgespalten werden kann. Ähnliches gilt für عَمَلَ und عَمَل, „er machte“ und „Arbeit“, und مَارَسَ und مَارِس, „er übte aus“ und „März“. Zwei Namen entsprechen ohne Vokalzeichen anderen Worten: عُمَر, „Omar“, ist ohne Einbeziehung des Kontexts nicht von عُمْر, „Lebensalter“,zu unterscheiden, genauso wenig wie نَجِيب, „Nagib“, von نِجِيب „wir bringen“. Auch ununterscheidbar ist سِتّ, „Frau“, von سِتّ im Sinne von sechs genannten Dingen. Schließlich kann ش entweder eine Abkürzung für شُوف, „sieh“, in der oben erwähnten Wikipedia typischen Wendung sein oder es ist ein abgespaltenes ش des Zirkumfix der Verneinung ما ... ش

# Schluss

Die vorliegende Arbeit hat gezeigt, dass eine Verarbeitung von Text in einem arabischen Dialekt am Computer mit heute verfügbaren, wenn auch nicht auf unbedingt auf arabischen Text spezialisierten, Tools in relativ kurzer Zeit möglich ist. Die verwendeten Programme waren allesamt kostenlos und im Quellcode verfügbar und konnten so an die Aufgabenstellung angepasst werden. Dennoch muss man feststellen, dass zurzeit fertige NLP-Werkzeuge für Arabisch noch fehlen, die mit für Nichtinformatiker überschaubarem Aufwand eingesetzt werden können. Die vorgestellten ersatzweise entwickelten Ansätze könnten in Zukunft zu einem solchen Werkzeug ausgebaut werden. Dies wird aber noch beträchtliche Forschungs- und Entwicklungsarbeit erfordern.

Die Zahl der Ausdrücke, die nur in einer für Wikipedia typischen Bedeutung so oft auftreten, dass sie zu den 200 häufigsten Wörtern zählen, ist relativ niedrig. Auch eine vergleichende Untersuchung wie die die typische Verteilung der Wörter in einem Lexikon aussieht bzw. in einer Wikipedia wäre sicher eine interessante Fragestellung, steht aber noch aus.

Die Anzahl der maßgeblich Beitragenden zur Wikipedia Masry hat sich als noch kleiner herausgestellt als zu Anfang angenommen. Obwohl als gesichert gelten kann, dass die Autoren ägyptische Umgangssprache schreiben wollen und Hocharabisch als Ausnahme gelten muss, ist die Repräsentativität der Daten daher für die gesprochene Varietät des ägyptischen Arabischen sehr gering.

Für die Darstellung von Wörterbüchern in elektronischer, standardisierter Form müssen noch Werkzeuge geschaffen werden, die es Linguisten einfach machen sie zu Nutzen und die Ergebnisse in einem gewohnten Format zu präsentieren.

Insgesamt stellt die hier durchgeführte Untersuchung sicher nicht mehr als einen ersten Baustein zum Aufbau eines Corpus des ägyptischen Arabisch dar, aus dem dann vielleicht in einer zukünftigen Untersuchung eine repräsentative Liste von Wörtern nach Häufigkeit erstellt werden kann.

## Anhang I: Das Wörterbuch

Eine Möglichkeit einen Eintrag für ein ägyptisch-arabisches Fremdwörterbuch mit Frequenzangabe darzustellen könnte beispielsweise wie folgt aussehen:

```
<?xml version="1.0" encoding="UTF-8"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0">
 <teiHeader n="1">
  <fileDesc>
   <titleStmt>
    <title>A machine-readable glossary of Egyptian Arabic </title>
   </titleStmt>
   <editionStmt>
    <edition>1.0 (Alpha)</edition>
    <author>x</author>
   </editionStmt>
   <publicationStmt>
    <pubPlace>x</pubPlace>
    <date>2011</date>
    <availability status="restricted">
     <p>
      <ref                    type="license"                    target="
http://creativecommons.org/licenses/by-sa/3.0/at/"/>
     </p>
    </availability>
   </publicationStmt>
   <sourceDesc>
    <p>This is an original digital text</p>
   </sourceDesc>
  </fileDesc>
  <revisionDesc>
  </revisionDesc>
 </teiHeader>
 <text>
  <body>
   <entry xml:id="wi_0" n="1377">
    <form type="lemma">
     <orth xml:lang="ar-arz-x-cairo-latin">wi-</orth>
     <orth xml:lang="ar-arz-x-cairo-arabic">و</orth>
    </form>
    <gramGrp>
     <gram type="pos">conjunction</gram>
     <gram type="root" xml:lang="ar-arz-x-cairo-latin">w</gram>
    </gramGrp>
    <usg>
     <cit>
      <ref type="url">http://arz.wikipedia.org/</ref>
      <fs>
       <f name="frequency"><numeric value="106446"/></f>
      </fs>
     </cit>
     <cit>
      <ref type="url">http://some-arabic-forum.eg/</ref>
      <fs>
       <f name="frequency"><numeric value="2203"/></f>
      </fs>
     </cit>
    </usg>
    <sense>
     <cit type="translation" xml:lang="en">
      <quote>and</quote>
     </cit>
     <cit type="translation" xml:lang="de">
      <quote>und</quote>
     </cit>
    </sense>
    <sense>
```

```
      <def xml:lang="en">particle used to introduce oaths</def>
      <def xml:lang="de">Schwurpartikel</def>
     </sense>
    </entry>
   </body>
  </text>
</TEI>
```

Im Text, der auf dem Datenbankauszug der Wikipedia Masry vom 21. 12. 2012 beruht, wurden 2937751 Token gefunden. Davon enthielten 667982 kein einziges arabisches Zeichung und wurden daher nicht weiter bearbeitet.

| | | | | | | |
|---|---|---|---|---|---|---|
| و | 106446 | و(100%), وـ(0%) | | conj | underr | und |
| فى | 91783 | فى(64%), في(17%), فيه(9%), فيها(6%), مافيش(3%) | und 27 weitere Formen. | prep | # | in; am |
| من | 45538 | من(91%), منها(3%), منهم(2%), منه(2%), ومن(1%) | und 14 weitere Formen. | prep | # | aus |
| على | 27832 | على(77%), ع(7%), عليه(7%), عليها(5%), عليهم(2%) | und 24 weitere Formen. | prep | # | auf |
| كان | 24848 | كان(50%), كانت(23%), كانوا(5%), كانو(4%), يكون(3%) | und 127 weitere Formen. | verb | ambig | ist, soll sein |
| تصنيف | 24441 | تصنيف(84%), تصانيف(16%), التصنيف(0%), تصنيفها(0%), بتصنيفها(0%) | und 4 weitere Formen. | noun | wiki | Kategorie |
| إللى | 21217 | اللى(84%), اللي(13%), واللي(2%), واللى(1%), إللى(0%) | und 2 weitere Formen. | pron | egypt | welcher/welche/ welches |
| مصرى | 18095 | المصريه(19%), مصرى(18%), مصريين(16%), المصرى(14%), المصريين(11%) | und 29 weitere Formen. | adj | geog | ägyptisch/Ägyptisch/ Ägypter; Kairo |
| إن | 16391 | ان(36%), انه(22%), إن(10%), انهم(7%), انها(7%) | und 62 weitere Formen. | conj | # | dass |
| سنه | 16222 | سنة(67%), سنه(13%), السنة(8%), سنين(5%), السنه(4%) | und 22 weitere Formen. | noun | hist | Jahr |

| | | | | | | |
|---|---|---|---|---|---|---|
| مصر | 15864 | مصر(94%), لمصر(4%), بمصر(1%), مصرنا(0%), المصرين(0%) | und 19 weitere Formen. | noun | geog | Ägypten; Kairo |
| مقال | 12515 | مقالات(46%), مقاله(23%), المقالات(15%), المقاله(15%), مقالة(0%) | und 12 weitere Formen. | noun | wiki | Artikel |
| ل | 11360 | ليها(28%), ليه(15%), له(13%), ل(13%), لـ(6%) | und 51 weitere Formen. | prep | underr | für |
| عن | 11332 | عن(91%), عنه(4%), عنها(2%), عنهم(1%), عني(0%) | und 24 weitere Formen. | prep | # | über |
| تحويل | 11149 | تحويل(99%), بتحويل(0%), تحويلها(0%), التحويل(0%), لتحويل(0%) | und 11 weitere Formen. | noun | wiki | Umwandeln |
| محمد | 10939 | محمد(99%), لمحمد(0%), ومحمد(0%), بمحمد(0%), محمدكم(0%) | und 2 weitere Formen. | n. prop | # | Muhammad |
| تانى | 9752 | التانى(49%), تانيه(14%), التانيه(13%), تانى(12%), تاني(3%) | und 27 weitere Formen. | ord | # | zweiter, II. |
| دى | 9645 | دى(89%), دي(9%), ودي(1%), ودى(1%) | | pron | egypt | diese |
| بعد | 9541 | بعد(78%), بعده(9%), بعدها(5%), وبعد(4%), وبعدها(1%) | und 23 weitere Formen. | prep | # | nach |
| ما | 9290 | ما(99%), وما(1%) | | part | # | nicht; was |
| عشان | 8961 | علشان(56%), عشان(43%), وعشان(1%), وعلشان(0%), عشانها(0%) | und 7 weitere Formen. | conj | egypt | weil; um zu |
| عبد | 7513 | عبد(83%), عبده(15%), | und 38 weitere | n. prop | # | Abd …, Abdu |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | العبد(0%), وعبد(0%), لعبد(0%) | Formen. | | | |
| يوم | 7169 | يوم(48%), ايام(17%), اليوم(13%), أيام(5%), اياميها(3%) | und 66 weitere Formen. | noun | # | Tag |
| أحمد | 7047 | احمد(64%), أحمد(35%), واحمد(0%), وأحمد(0%), لأحمد(0%) | und 4 weitere Formen. | n. prop | # | Ahmad |
| دول | 6911 | الدوله(23%), دول(19%), الدول(9%), دوله(8%), الدولة(8%) | und 40 weitere Formen. | noun | # | Staat |
| مع | 6741 | مع(79%), معاه(8%), معاهم(5%), معاها(3%), معاك(2%) | und 23 weitere Formen. | prep | # | zusammen mit |
| هو | 6735 | هو(88%), وهو(11%), ماهو(0%), وهوه(0%), ماهوش(0%) | | pron | # | er |
| كتاب | 6520 | كتاب(19%), كتب(15%), كتابة(10%), الكتاب(10%), الكتب(9%) | und 125 weitere Formen. | noun | ambig | Buch; schreiben; Koranschule |
| سبتمبر | 6518 | سبتمبر(100%), وسبتمبر(0%) | | noun | hist | September |
| قالب | 6518 | قالب(100%), قالبه(0%), قالبوا(0%) | | noun | wiki | Vorlage |
| بن | 6487 | بن(49%), ابن(36%), ابنه(7%), ابنها(1%), إبن(1%) | und 47 weitere Formen. | noun | # | Sohn |
| أول | 6432 | الأول(27%), اول(19%), الاول(17%), أول(17%), الأولى(7%) | und 32 weitere Formen. | ord | # | erster |
| أو | 6410 | أو(57%), او(43%), لاو(0%), لأو(0%) | | conj | # | oder |
| عمل | 6156 | عمل(20%), يعمل(5%), عملت(4%), العمل(4%), اعمال(4%) | und 173 weitere Formen. | verb | ambig | machen; Arbeit, Werk |

| | | | | | | |
|---|---|---|---|---|---|---|
| أولانى | 5745 | الاولانى(93%), الاولانيه(3%), الأولانى(1%), الأولانيه(1%), الاولاني(0%) | und 11 weitere Formen. | ord | egypt | I. |
| ملك | 5467 | الملك(39%), ملك(34%), ملكة(3%), الملكى(3%), للملك(2%) | und 57 weitere Formen. | noun | polit | König |
| دين | 5407 | الدين(89%), دين(5%), لدين(1%), للدين(1%), دينهم(1%) | und 30 weitere Formen. | noun | relig | Religion |
| يتيم | 5370 | يتيمه(99%), اليتيم(0%), يتيم(0%), اليتيمة(0%), يتيمة(0%) | | noun | wiki | verwaister Artikel, d. h. unzugänglich |
| ده | 5280 | ده(71%), دا(19%), وده(6%), ودا(2%), دة(2%) | | pron | egypt | dieser |
| ب | 5260 | بيها(17%), بـ(15%), ب(15%), بيه(14%), بنا(5%) | und 71 weitere Formen. | prep | underr | mittels |
| لكن | 5104 | لكن(93%), ولكن(3%), لكنه(2%), لكنها(1%), ولكنه(1%) | und 9 weitere Formen. | conj | # | aber |
| مدينه | 5089 | مدينة(32%), مدن(17%), مدينه(16%), المدينه(9%), المدن(7%) | und 47 weitere Formen. | noun | geog | Stadt |
| قال | 4961 | قال(18%), بيقول(12%), بيتقال(8%), بتقول(7%), قالت(4%) | und 187 weitere Formen. | verb | # | sagen |
| يكيبيديا | 4946 | ويكيبيديا(60%), ويكيبيديا(39%), الويكيبيديا(0%), لويكيبيديا(0%), الويكيبيديات(0%) | und 2 weitere Formen. | n. prop | wiki | Wikipedia |
| كل | 4941 | كل(73%), كله(7%), كلها(6%), بكل(3%), كلهم(3%) | und 24 weitere Formen. | adv | # | ganz |
| قاهره | 4933 | القاهرة(49%), القاهره(46%), بالقاهرة(2%), للقاهره(1%), | und 8 weitere | noun | geog | Kairo |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | والقاهرة(1%) | Formen. | | | |
| أب | 4799 | أبو(35%), ابو(32%), ابوه(11%), أبوه(3%), أبى(3%) | und 73 weitere Formen. | noun | # | Vater |
| إسم | 4759 | اسم(28%), اسمها(19%), باسم(13%), الاسم(12%), بإسم(10%) | und 39 weitere Formen. | noun | # | Name |
| ال | 4660 | ال(95%), الـ(2%), وال(2%), الكم(0%), الهم(0%) | und 10 weitere Formen. | part | underr | der/die/das |
| بين | 4599 | بين(74%), مابين(7%), بينهم(5%), وبين(2%), بينه(2%) | und 51 weitere Formen. | prep | # | zwischen |
| فيلم | 4479 | فيلم(91%), الفيلم(6%), فيلمه(1%), وفيلم(0%), لفيلم(0%) | und 13 weitere Formen. | noun | ent | Film |
| ناس | 4216 | الناس(45%), ناس(15%), الانسان(12%), انسان(5%), الإنسان(5%) | und 45 weitere Formen. | noun | # | Leute |
| كمان | 4210 | كمان(92%), وكمان(8%), الكمان(0%), والكمان(0%), كماني(0%) | | adv | egypt | auch |
| تأريخ | 4202 | تاريخ(64%), التاريخ(28%), بتاريخ(2%), تأريخ(2%), للتاريخ(1%) | und 8 weitere Formen. | noun | hist | Geschichte |
| زى | 4200 | زى(78%), زي(18%), زيه(1%), زيها(1%), زيت(1%) | und 14 weitere Formen. | adv | egypt | wie |
| ساعد | 4200 | ساعد(49%), تساعد(42%), يساعد(1%), ساعده(1%), يساعده(1%) | und 56 weitere Formen. | verb | wiki | helfen (hilf…) |
| حكم | 4139 | حكم(34%), الحكم(30%), حكمه(5%), حكمت(3%), | und 71 weitere | noun | polit | Regierung |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | للحكم(2%) | Formen. | | | |
| لغه | 4070 | اللغه(25%), لغه(16%), لغة(11%), باللغه(10%), لغات(10%) | und 33 weitere Formen. | noun | lang | Sprache |
| ممكن | 4009 | ممكن(98%), وممكن(1%), ممكنة(0%), ممكنه(0%), الممكن(0%) | und 2 weitere Formen. | adv | # | vielleicht |
| كبير | 3987 | كبير(40%), كبيره(32%), الكبير(15%), الكبيره(5%), كبيرة(4%) | und 9 weitere Formen. | adj | # | groß |
| هى | 3947 | هى(70%), هي(17%), وهي(7%), وهى(6%), وهيه(0%) | | pron | # | sie |
| كومونز | 3876 | كومونز(100%) | | n. prop | wiki | Commons |
| فضل | 3838 | فضلك(60%), فضل(16%), فضلت(9%), الفضل(2%), فضلو(2%) | und 40 weitere Formen. | adv | wiki | bitte … |
| كده | 3834 | كده(72%), بكده(11%), كدا(8%), كدة(3%), وبكده(2%) | und 10 weitere Formen. | adv | egypt | so |
| ويكيميديا | 3704 | ويكيميديا(100%), الويكيميديا(0%) | | noun | wiki | WikiMedia |
| فايلات | 3685 | فايلات(100%), الفايلات(0%) | | noun | wiki | Dateien (-> File) |
| تالت | 3597 | التالت(85%), التالته(8%), تالت(4%), التالتة(1%), تالته(1%) | und 7 weitere Formen. | ord | # | dritter |
| تقاوى | 3503 | تقاوى(100%) | | noun | wiki | Stub (wtl.: Samen) |
| يا | 3500 | يا(100%) | | part | songs | Vokativpartikel (O …) |
| ممثل | 3392 | ممثلين(31%), ممثل(21%), ممثلات(13%), ممثله(13%), ممثلة(11%) | und 24 weitere Formen. | noun | ent | Schauspieler |

| | | | | | | |
|---|---|---|---|---|---|---|
| مش | 3380 | مش(98%), ومش(2%) | | part | egypt | nicht … |
| غير | 3330 | غير(58%), غيرها(9%), غيرهم(9%), غيره(3%), وغيرها(2%) | und 66 weitere Formen. | adv | # | anders |
| صفح | 3285 | الصفحه(88%), صفحة(6%), صفحات(4%), صفحه(1%), الصفحة(0%) | und 11 weitere Formen. | noun | wiki | Seite (-> Webpage) |
| إسكندريه | 3190 | اسكندريه(84%), اسكندرية(10%), إسكندرية(3%), إسكندريه(2%), لاسكندريه(0%) | und 5 weitere Formen. | noun | geog | Alexandria |
| عربى | 3181 | العربية(21%), العربيه(20%), العربى(19%), عربى(11%), العربي(9%) | und 25 weitere Formen. | noun | lang | Arabisch/arabisch |
| نفس | 3181 | نفس(26%), نفسه(22%), نفسها(10%), نفسهم(8%), النفس(7%) | und 46 weitere Formen. | pron | # | selbst |
| حسن | 3050 | حسن(94%), الحسن(4%), لحسن(1%), وحسن(1%), بحسن(0%) | und 6 weitere Formen. | n. prop | # | Hasan |
| طريق | 3010 | طريق(42%), طريقة(23%), بطريقه(10%), الطريق(7%), طريقه(4%) | und 28 weitere Formen. | noun | # | Methode |
| رئيس | 2960 | رئيس(62%), الرئيس(23%), رئيسه(5%), ورئيس(3%), لرئيس(2%) | und 15 weitere Formen. | noun | polit | Präsident |
| محمود | 2891 | محمود(99%), ومحمود(1%), لمحمود(0%), المحمود(0%), بمحمود(0%) | | n. prop | # | Muhammad |
| شاف | 2877 | شوف(65%), شاف(8%), | und 106 weitere | verb | egypt | sehen |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | شافت(3%), يشوف(2%), شافوا(2%) | Formen. | | | |
| لما | 2870 | لما(94%), ولما(6%) | | conj | # | wenn; als |
| مولود | 2833 | مواليد(97%), مولود(1%), المولود(1%), مولودة(0%), المولودين(0%) | und 9 weitere Formen. | prep | biog | Geburt/Geborene |
| حد | 2795 | لحد(70%), حد(21%), الحد(3%), ولحد(1%), وحد(1%) | und 19 weitere Formen. | noun | # | bis |
| عصر | 2784 | العصر(36%), عصر(27%), العصور(20%), عصور(7%), عصره(3%) | und 36 weitere Formen. | noun | # | Ursprung |
| بلد | 2755 | بلاد(29%), البلد(18%), بلد(14%), البلاد(13%), بلادى(4%) | und 50 weitere Formen. | prep | polit | Land |
| قبل | 2718 | قبل(85%), قبله(8%), قبلها(3%), وقبل(1%), قبلي(0%) | und 19 weitere Formen. | noun | # | vor (zeitl.) |
| منطق | 2652 | منطقة(31%), المنطقه(18%), مناطق(17%), منطقه(10%), المناطق(9%) | und 37 weitere Formen. | noun | geog | Region |
| حديث | 2647 | احداث(36%), الحديث(19%), الحديثه(15%), الحديثة(7%), حديث(6%) | und 27 weitere Formen. | noun | hist | Ereignis |
| بابا | 2580 | بابا(56%), البابا(42%), للبابا(1%), وبابا(0%), والبابا(0%) | und 4 weitere Formen. | noun | christ | Pabst(, Papa) |
| بحر | 2574 | البحر(50%), بحر(13%), البحريه(9%), بحريه(7%), البحرين(6%) | und 30 weitere Formen. | noun | geog | Meer |
| عالم | 2573 | العالم(76%), عالم(18%), | und 16 weitere | noun | geog | Welt |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | للعالم(2%), وعالم(1%), عالمة(1%) | Formen. | | | |
| مكان | 2571 | المكان(38%), مكان(31%), مكانه(9%), اماكن(4%), مكانها(3%) | und 36 weitere Formen. | noun | geog | Ort |
| كتير | 2563 | كتيره(45%), كتير(42%), كتيرة(5%), الكتير(3%), بكتير(2%) | und 7 weitere Formen. | adj | # | groß |
| يوسف | 2503 | يوسف(98%), اليوسف(1%), ويوسف(0%), ليوسف(0%) | | n. prop | # | Josef |
| سيد | 2460 | السيد(45%), سيد(41%), سيدي(3%), السيدة(3%), السيده(2%) | und 18 weitere Formen. | noun | biog | Herr … |
| إنت | 2456 | انت(80%), وانت(10%), أنت(4%), إنت(4%), وأنت(1%) | und 7 weitere Formen. | pron | # | du/ihr |
| مصدر | 2443 | مصادر(80%), مصدر(8%), المصادر(8%), المصدر(2%), مصادرة(0%) | und 23 weitere Formen. | noun | wiki | Quellen |
| سلطان | 2438 | السلطان(64%), سلطان(27%), للسلطان(4%), سلطانه(1%), بالسلطان(1%) | und 14 weitere Formen. | noun | polit | Sultan |
| الله | 2415 | الله(83%), بالله(11%), والله(4%), ياالله(1%), اللات(0%) | und 9 weitere Formen. | noun | # | Gott |
| جيش | 2405 | الجيش(55%), جيش(27%), جيشه(5%), بالجيش(3%), بجيش(3%) | und 18 weitere Formen. | noun | polit | Heer |
| جامعه | 2387 | جامعة(62%), الجامعه(12%), الجامعة(7%), جامعه(5%), الجامعات(4%) | und 18 weitere Formen. | noun | geog | Universität |

| | | | | | | |
|---|---|---|---|---|---|---|
| يوأنس | 2380 | يوأنس(100%) | | n. prop | christ | Johannes |
| شرق | 2345 | الشرق(39%), شرق(23%), الشرقيه(17%), الشرقى(6%), الشرقية(6%) | und 16 weitere Formen. | noun | geog | Osten |
| حرب | 2327 | الحرب(49%), حرب(44%), للحرب(2%), حربه(1%), لحرب(1%) | und 11 weitere Formen. | noun | polit | Krieg |
| وصلات | 2325 | وصلات(100%) | | noun | wiki | Links |
| بإضافة | 2291 | بإضافة(100%), باضافة(0%) | | noun | wiki | Hinzufügen |
| راح | 2291 | راح(36%), يروح(8%), الروح(7%), راحت(7%), روح(6%) | und 86 weitere Formen. | verb | egypt | gehen |
| أحد | 2287 | واحد(87%), أحد(4%), ماحدش(3%), حدش(2%), الأحد(1%) | und 10 weitere Formen. | pron | egypt | jemand/niemand |
| ليل | 2253 | ليلى(36%), ليلة(16%), الليل(15%), ليله(8%), ليل(5%) | und 38 weitere Formen. | noun | hist | Nacht |
| قديم | 2241 | القديمه(37%), القديم(21%), قديمه(13%), القديمة(11%), قديم(10%) | und 12 weitere Formen. | adj | hist | alt |
| مملوك | 2236 | المماليك(32%), المملوكيه(27%), مماليك(17%), المملوكى(9%), مملوكيه(4%) | und 27 weitere Formen. | noun | hist | Mamluk |
| سلام | 2212 | السلام(31%), سلامه(22%), سلامة(17%), سلام(16%), للسلام(3%) | und 32 weitere Formen. | noun | polit | Islam |
| حزب | 2203 | حزب(51%), الحزب(40%), للحزب(4%), لحزب(2%), | und 14 weitere Formen. | noun | polit | Partei |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | حزبه(1%) | | | | |
| كنيس | 2182 | الكنيسه(43%), كنيسة(31%), الكنيسة(17%), كنيسه(5%), كنيس(1%) | und 12 weitere Formen. | noun | christ | Kirche(; Synagoge) |
| مسيحى | 2172 | المسيحيين(27%), المسيحيه(19%), مسيحيه(15%), مسيحيين(10%), المسيحية(10%) | und 14 weitere Formen. | noun | christ | Christen |
| قدر | 2158 | قدر(32%), يقدر(11%), قدرت(6%), تقدر(5%), قدروا(4%) | und 90 weitere Formen. | verb | # | wollen |
| م | 2143 | م(100%), مـ(0%) | | abbrev | hist | n. Chr. |
| عدد | 2119 | عدد(78%), عددهم(5%), العدد(5%), عددها(2%), لعدد(2%) | und 15 weitere Formen. | noun | # | Anzahl; eine Anzahl von |
| حب | 2111 | الحب(41%), حب(21%), حبك(7%), حبي(6%), حبه(4%) | und 42 weitere Formen. | noun | songs | Liebe |
| وفاه | 2100 | وفيات(81%), وفاة(11%), الوفاه(5%), وفاه(1%), الوفاة(1%) | und 4 weitere Formen. | noun | biog | Sterben/Verstorbene |
| بقى | 2098 | بقى(62%), بقية(8%), بقوا(7%), بقو(4%), وبقى(4%) | und 30 weitere Formen. | verb | # | bleiben |
| رابع | 2082 | الرابع(84%), الرابعه(9%), رابع(4%), الرابعة(2%), رابعه(1%) | und 4 weitere Formen. | ord | # | vierter |
| إنجليزى | 2079 | انجليزى(60%), الانجليزى(11%), بالانجليزى(4%), إنجليزى(4%), الانجليزيه(2%) | und 24 weitere Formen. | adj | lang | englisch/Englisch |
| عمر | 2074 | عمر(56%), العمر(16%), عمره(8%), عمرها(5%), عمري(5%) | und 31 weitere Formen. | noun | ambig | Alter, Omar (n. prop) |
| بس | 2051 | بس(98%), وبس(2%) | | adv | egypt | nur; aber |

| | | | | | | |
|---|---|---|---|---|---|---|
| عند | 2045 | عند(38%), عنده(29%), عندهم(13%), عندها(11%), عندك(2%) | und 17 weitere Formen. | prep | # | bei |
| حصل | 2044 | حصل(37%), حصلت(31%), بيحصل(10%), بتحصل(7%), يحصل(4%) | und 43 weitere Formen. | verb | # | geschehen |
| عام | 2035 | عام(28%), العام(24%), العامة(19%), العامه(12%), عامة(7%) | und 21 weitere Formen. | noun | hist | Jahr |
| سبب | 1992 | بسبب(58%), سبب(15%), السبب(8%), بيسبب(2%), بتسبب(2%) | und 42 weitere Formen. | noun | # | Grund; aufgrund |
| حياه | 1990 | حياته(25%), حياة(21%), الحياه(14%), الحياة(14%), حياتها(6%) | und 18 weitere Formen. | noun | biog | Leben |
| بتاع | 1985 | بتاع(41%), بتاعة(27%), بتاعه(8%), بتاعتها(7%), بتاعته(5%) | und 12 weitere Formen. | part | egypt | Gen. Exponent |
| شهر | 1985 | شهر(46%), الشهر(14%), شهرية(11%), تشهر(6%), شهره(4%) | und 37 weitere Formen. | noun | hist | Monat |
| فتره | 1951 | فترة(34%), الفتره(30%), فتره(13%), الفترة(11%), لفتره(4%) | und 15 weitere Formen. | noun | hist | Zeit(raum) |
| شمال | 1930 | شمال(40%), الشمال(24%), الشماليه(19%), الشمالى(6%), الشمالية(5%) | und 15 weitere Formen. | noun | geog | Nord |
| فؤاد | 1924 | فؤاد(99%), الفؤاد(1%), فؤاده(0%), والفؤاد(0%), وفؤاد(0%) | | n. prop | # | Fuad |
| جنوب | 1908 | جنوب(48%), الجنوبيه(18%), | und 18 weitere | noun | geog | Süd |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | الجنوب(17%), الجنوبية(7%), الجنوبى(3%) | Formen. | | | |
| أمريكا | 1894 | امريكا(86%), أمريكا(10%), لامريكا(2%), لأمريكا(1%), وأمريكا(1%) | und 4 weitere Formen. | noun | polit | Amerika |
| علم | 1892 | علم(74%), العلم(17%), لعلم(2%), بالعلم(1%), وعلم(1%) | und 16 weitere Formen. | noun | ambig | Flagge; Wissenschaft |
| لينك | 1892 | لينك(55%), لينكات(44%), اللينكات(0%) | | noun | wiki | Link |
| كامل | 1852 | كامل(63%), الكامل(17%), بالكامل(11%), كامله(4%), الكامله(1%) | und 10 weitere Formen. | adv | # | vollständig, ganz |
| حق | 1821 | حقوق(25%), حق(13%), الحق(13%), الحقوق(11%), لحقه(11%) | und 64 weitere Formen. | noun | # | Wahrheit |
| حسين | 1813 | حسين(93%), الحسين(6%), وحسين(0%), بالحسين(0%) | | n. prop | # | Hussein |
| وقت | 1766 | وقت(67%), وقتها(30%), وقته(1%), للوقت(0%), بوقت(0%) | und 6 weitere Formen. | noun | hist | Zeit |
| لغايه | 1761 | لغاية(97%), لغايه(1%), ولغاية(1%), الغايه(0%), الغاياتي(0%) | und 2 weitere Formen. | prep | egypt | bis |
| قرن | 1760 | القرن(89%), قرن(4%), القرنين(3%), للقرن(2%), قرنين(1%) | und 17 weitere Formen. | noun | hist | Jahrhundert |
| محافظ | 1757 | محافظة(52%), محافظه(13%), المحافظه(11%), محافظات(9%), المحافظين(5%) | und 13 weitere Formen. | noun | polit | Gouvernement |
| مره | 1755 | مره(31%), مرة(25%), مرات(20%), مراته(16%), | und 16 weitere | noun | # | Mal |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | للمرة(2%) | Formen. | | | |
| إتولد | 1748 | اتولد(68%), اتولدت(23%), إتولد(7%), إتولدت(1%), اتولدو(1%) | und 6 weitere Formen. | verb | egypt | geboren |
| غرب | 1745 | غرب(25%), الغربيه(23%), الغرب(20%), الغربية(11%), الغربى(9%) | und 30 weitere Formen. | noun | geog | Osten |
| فرنسا | 1723 | فرنسا(97%), لفرنسا(2%), بفرنسا(1%), وفرنسا(0%), ولفرنسا(0%) | | noun | polit | Frankreich |
| أكبر | 1699 | اكبر(48%), أكبر(32%), الاكبر(9%), الأكبر(7%), واكبر(2%) | und 8 weitere Formen. | adj | # | größer/am größten |
| حال | 1685 | الحال(24%), حالة(23%), حالات(7%), حاله(6%), الحاله(6%) | und 56 weitere Formen. | noun | # | Zustand |
| دخل | 1685 | دخل(38%), دخلت(16%), تدخل(6%), يدخل(5%), دخلوا(4%) | und 80 weitere Formen. | verb | # | hineingehen |
| يناير | 1684 | يناير(100%), لينایر(0%), بيناير(0%) | | noun | hist | Jänner |
| أى | 1676 | اى(30%), أى(21%), باى(10%), اي(10%), أي(7%) | und 23 weitere Formen. | pron | # | irgendein |
| باشا | 1672 | باشا(97%), الباشا(3%), باشاه(0%), للباشا(0%), والباشا(0%) | | noun | egypt | Herr … (osman. Ehrentitel) |
| دور | 1659 | دور(36%), دورة(7%), الدور(6%), بدور(5%), دوره(5%) | und 65 weitere Formen. | noun | # | Rolle |
| صلاح | 1655 | صلاح(97%), لصلاح(2%), وصلاح(0%), الصلاح(0%), | und 2 weitere Formen. | n. prop | # | Salah |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | بالصلاح(0%) | | | | |
| أكتر | 1645 | اكتر(61%), أكتر(28%), لأكتر(2%), لاكتر(2%), واكتر(2%) | und 15 weitere Formen. | adj | # | mehr |
| سياسه | 1644 | السياسى(24%), السياسيه(23%), السياسية(10%), سياسه(9%), السياسه(8%) | und 33 weitere Formen. | noun | polit | Politik |
| إلى | 1641 | إلى(47%), الى(15%), الي(14%), والى(12%), إليك(2%) | und 17 weitere Formen. | prep | # | zu |
| إبراهيم | 1635 | ابراهيم(76%), إبراهيم(23%), وابراهيم(1%), لإبراهيم(0%), وإبراهيم(0%) | | n. prop | # | Ibrahim |
| جديد | 1630 | الجديد(22%), جديده(22%), جديد(21%), الجديده(15%), الجديدة(12%) | und 10 weitere Formen. | adj | # | neu |
| أمير | 1616 | الأمير(36%), أمير(19%), امير(12%), الامير(9%), الأميره(4%) | und 40 weitere Formen. | noun | polit | Prinz |
| ست | 1615 | ست(36%), الستات(28%), الست(16%), للستات(4%), لست(2%) | und 29 weitere Formen. | noun | ambig | Frau; sechs |
| هم | 1596 | هما(64%), هم(9%), بتهمة(6%), وهما(6%), وهم(3%) | und 39 weitere Formen. | pron | egypt | sie |
| أكتوبر | 1589 | اكتوبر(75%), أكتوبر(25%), وأكتوبر(0%) | | noun | hist | Oktober |
| أم | 1579 | ام(35%), أم(24%), الأم(8%), الام(8%), امها(4%) | und 54 weitere Formen. | noun | biog | Mutter |
| حوالى | 1568 | حوالى(72%), حوالي(12%), حواليه(6%), لحوالى(4%), | und 10 weitere Formen. | adv | # | ungefähr |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | بحوالى(3%) | | | | |
| مارس | 1560 | مارس(100%), لمارس(0%), ومارس(0%) | | noun | ambig | März; ausüben, praktizieren |
| حاجه | 1554 | حاجه(63%), حاجات(16%), حاجة(15%), بحاجات(1%), حاجتهم(1%) | und 12 weitere Formen. | noun | egypt | Sache, etwas |
| وصل | 1554 | وصل(45%), بيوصل(8%), يوصل(8%), توصل(6%), بتوصل(6%) | und 74 weitere Formen. | verb | # | ankommen |
| متحد | 1541 | المتحده(61%), المتحدة(36%), متحده(1%), المتحد(0%), المتحدين(0%) | und 3 weitere Formen. | adj | polit | vereinigt |
| قلب | 1538 | قلبي(30%), القلب(18%), قلب(15%), قلبى(13%), قلبك(6%) | und 39 weitere Formen. | noun | songs | Herz |
| مركز | 1538 | مركز(59%), المركز(24%), مركزها(4%), ومركز(3%), مركزه(3%) | und 16 weitere Formen. | noun | geog | Zentrum |
| ش | 1525 | ش(100%) | | part | ambig | Teil der Verneinung |
| مسلم | 1525 | المسلمين(66%), مسلمين(13%), مسلم(9%), للمسلمين(3%), المسلم(3%) | und 20 weitere Formen. | noun | relig | Muslim |
| عين | 1524 | عين(39%), العين(9%), عينه(8%), عيني(6%), العينى(4%) | und 83 weitere Formen. | noun | # | Auge |
| زكى | 1523 | زكى(85%), زكيه(9%), زكي(6%), زكية(0%), الزكي(0%) | und 2 weitere Formen. | n. prop | # | Zaki |
| ناصر | 1516 | الناصر(71%), ناصر(17%), الناصرى(5%), الناصريه(1%), | und 13 weitere Formen. | n. prop | # | Nasser |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | للناصر(1%) | | | | |
| قوه | 1512 | قوات(29%), القوات(27%), قوة(8%), للقوات(7%), قوه(6% | und 15 weitere Formen. | noun | polit | (Streit-)Kräfte |
| لو | 1509 | لو(87%), ولو(6%), بلو(1%), حلوين(1%), لوكى(1%) | und 13 weitere Formen. | conj | # | wenn |
| مسلسل | 1504 | مسلسل(90%), المسلسل(6%), المسلسلات(3%), مسلسله(0%), للمسلسل(0%) | und 7 weitere Formen. | noun | ent | Serie |
| عسكرى | 1488 | العسكريه(25%), عسكريه(25%), عسكرى(18%), العسكرى(12%), العسكرية(6%) | und 12 weitere Formen. | adj | polit | militärisch |
| ضد | 1465 | ضد(92%), ضده(5%), ضدهم(2%), ضدها(1%), ضدكم(0%) | und 3 weitere Formen. | prep | # | gegen |
| دلوقت | 1457 | دلوقتى(84%), دلوقتي(12%), دلوقت(2%), ودلوقتي(1%), ودلوقتى(1%) | | adv | hist | jetzt |
| نجيب | 1446 | نجيب(100%), لنجيب(0%), ونجيب(0%) | | n. prop | ambig | Nagib; wir bringen |
| يد | 1443 | ايد(33%), إيد(5%), ايده(4%), يدى(3%), بيدى(3%) | und 113 weitere Formen. | noun | # | Hand |
| دار | 1438 | دار(80%), الدار(10%), دارك(2%), دارت(1%), لدار(1%) | und 27 weitere Formen. | noun | # | Haus |
| سعاد | 1438 | سعاد(99%), اسعاد(0%), وسعاد(0%), سعادت(0%), لسعاد(0%) | | n. prop | # | Suad (w.) |
| رجع | 1437 | رجع(38%), يرجع(13%), بيرجع(9%), رجعت(8%), | und 65 weitere | verb | # | zurückkehren |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | بترجع(4%) | Formen. | | | |
| إسماعيل | 1425 | اسماعيل(56%), إسماعيل(43%), واسماعيل(0%), وإسماعيل(0%), لاسماعيل(0%) | und 2 weitere Formen. | n. prop | # | Ismail |
| كلمه | 1425 | كلمات(39%), كلمة(28%), كلمه(9%), الكلمه(9%), الكلمات(5%) | und 22 weitere Formen. | noun | # | Wort |
| لأ | 1419 | لا(97%), لأ(3%) | | part | # | nicht |
| فبراير | 1410 | فبراير(100%) | | noun | # | Februar |
| نظام | 1408 | نظام(54%), النظام(28%), لنظام(7%), بنظام(3%), للنظام(2%) | und 13 weitere Formen. | noun | polit | System |
| سليمان | 1393 | سليمان(99%), لسليمان(0%), وسليمان(0%) | | n. prop | # | Suleiman |
| ن | 1389 | ن(100%), نـ(0%) | | abbrev | wiki | Diskussion |
| رجل | 1376 | رجال(27%), رجل(17%), الرجل(10%), الرجاله(8%), الرجال(6%) | und 50 weitere Formen. | noun | # | Mann |

Die angegebene Liste deckt 1024164 Token ab. Es sind also im Text noch weitere 1913587 Token zu finden, wenn man die nicht verarbeiteten mitzählt bzw. 1245605 wenn man sie nicht mitzählt.

# Anhang II: Für diese Arbeit verwendete und erstellte Programme

Für das Arbeiten mit XML Dateien hat sich <oXygen/> XML Editor[235] in den letzten Jahren immer mehr zum Standardtool der Geisteswissenschaftler entwickelt. Er wurde auch hier verwendet. Dieser ist zwar nicht gratis, aber für akademische und private Zwecke sehr günstig zu haben. Dieser Editor beinhaltet eine uneingeschränkte Lizenz für Saxon 9.4[236], ein Werkzeug, mit dem man XML mittels XQuery[237] abfragen und mittels XSLT 2.0[238] umformen kann. Man kann beliebige Java Objekte von Saxon aus nutzen als wären es Funktionen in XQuery oder XSLT. Auf diese Weise wurde der Wikitext Parser Sweble eingebunden. Dieser wurde leicht geändert, um die Weiterverarbeitung zu vereinfachen. Außerdem wurde das Problem des Auszeichnens von einzelnen Worten in einem Satz in eine Java-Funktion ausgelagert, da die prinzipiell mögliche Lösung in XQuery in der gegebenen Zeit nicht erstellt werden konnte.

Als erster Schritt wird ein XML-Dump der Datenbank von Wikipedia Masry so bearbeitet, dass alle Worte von XML-Tags umschlossen sind und sie in der Baumstruktur als Kinder von Sätzen aufscheinen, die Kinder von Absätzen sind, die Kinder vonseiten sind. Das wird in seltenen Fällen nicht erreicht, sodass Absätze wieder von Absätzen umgeben sind. Aufzählungsartige Listen werden als Absätze interpretiert. Links zu Wikipedias in anderen Sprachen werden unterdrückt, da sich darunter auch links zu Sprachen wie Arabisch und Farsi befinden, bei denen die Worte sonst nicht ausscheidbar wären.

Sweble nutzt das Buildsystem Maven[239], um das Erstellen des Programms und das Auflösen der Abhängigkeiten von Programmen von Dritten zu automatisieren. Daher wurde auch das Modul, das Sweble mit Saxon kombiniert mit diesem Werkzeug erstellt. Es kann im Quellcode oder als JAR-Datei unter https://github.com/simar0at/swc-saxon-interface (Stand 01. 2013) heruntergeladen werden. Weiters wird eine Erweiterung für Saxon benötigt, um die Worte mittels XML Tags zu markieren. Diese findet sich hier: https://github.com/simar0at/word-marker-for-saxon. Beide Male sind fertige JAR Dateien vorhanden, die in <oXygen/> als Extension für Saxon geladen werden müssen.

Das recht umfangreiche XQuery Skript für diesen Arbeitsschritt ist ebenfalls unter https://github.com/simar0at/word-marker-for-saxon zu finden.

---

[235] http://www.oxygenxml.com/

[236] Eine frei verfügbare Version, die Home Edition HE, ist unter http://sourceforge.net/projects/saxon/files/ verfügbar. Es ist mindestens die Version 9.3 notwendig um die hier selbst erstellte Zusatzfunktionen einzubinden. Eine kommerzielle Version ist relativ teuer. Diese kann unter http://www.saxonica.com/download/download_page.xml bezogen werden. Der Unterschied liegt in den angebotenen Funktionen, die aber auch bei der HE Version in den meisten Fällen, so auch hier, ausreichen. <oXygen/> enthält eine Version von Saxon, bei der alle Funktionen verwendet werden können. Die einzig wirklich interessante davon für größere Wikipedias wie die englische oder deutsche, ist das sogenannte Streaming. Dabei ist es nicht nötig das gesamte Eingabedokument im Speicher zu halten, was bei der englischen Wikipedia heute nicht möglich ist. Streaming konnte allerdings mit der Kombination <oXygen/> - Saxon nicht erfolgreich getestet werden.

[237] Priscilla Walmsley: XQuery. Farnham, Calif., USA, 2007.

[238] Jeni Tennison: Beginning XSLT 2.0. From novice to professional. Berkeley, Calif., USA, 2005.. Dies ist vor allem deshalb interessant, da die Umwandlung von TEI in andere Formate wie etwa Word (docx) mittels XSLT 2.0 Stylesheets erfolgt.

[239] http://maven.apache.org/

Es trennt die nach einer Bearbeitung des Wikitexts mit Sweble entstandenen Rohtexte an folgenden Zeichen auf und markiert die dazwischenliegenden Zeichenketten mit dem Tag <w/>:

Zeichenfolgen mit spezieller Funktion in XML-Dateien:

„&“ = &, „>“ = >, „<“ = <

Zeichen, die auch in Zahlen als Trenner vorkommen können:

„.“, „,“, „%“

Satzzeichen:

„-“, „–“, „—“, „.-“ (arabischer Punkt), „=“, „|“, „(“, „)“, „{“, „}“, „[“, „]“, „<“, „>“, „⟨“, „⟩“, „‘“, „’“, „"“, „«“, „»“, „‹“, „›“, „“, „„“, „#“, „&“, „/“, „*“, „•“, „;“, „:“, „?“, „؟“ (arabisches Fragezeichen), „!“, „،“ (arabisches Komma)

Sowie diverse Leerzeichen und nicht sichtbare Zeichen, die die Darstellung von Text beeinflussen:

Right-to-left mark, right-to-left override, pop directional formatting, zero width joiner.

Schreibweisen, die im ägyptischen bekanntermaßen Varianten desselben Wortes darstellen wurden zusammengefasst (Worte, die auf ـي und ـى beziehungsweise ـه und ـة enden).

Das so entstandene XML kann einerseits einfach in reinen Text umgewandelt werden:

```
xquery version "1.0";
declare default element namespace "http://www.mediawiki.org/xml/export-
0.8/";
declare boundary-space strip;
declare option saxon:output "method=text";
(: In principle
string-join(//text, " ")
would do but there are many useless newlines and whitespaces one can get
rid of. :)
(: Compact the whitespaces, separate every tagged entity by space and
every article
by a newline :)
string-join(
for $text in //text/*
return
string-join(normalize-space($text), " ")
, "&#10;")
```

Andererseits wurden die entstandenen Daten in eine eXist-db XML-Datenbank geladen, um so direkt nach Worten in ihrem Kontext suchen zu können. Dies kann man mit folgendem XQuery Skript durchführen.

```
xquery version "3.0";
declare default element namespace "http://www.mediawiki.org/xml/export-
0.8/";
let $timestamp := "20121221"
```

```
let $search-for := ( )
let $path := collection(concat("/db/arzwiki-tagged/", $timestamp))
let $all-words-count := count($path//w)
return
( <comment>{("Number of words to search for:",
count($search-for),
"out of", $all-words-count,
"tagged words (including of unknown type)")}
</comment>,
 for $word in subsequence($search-for, 1, 50)
 let $count := count((# exist:force-index-use #) {
  $path//w[. = $word]}
  )
 let $example-sentences :=
  for $wordIC in subsequence((# exist:force-index-use #) {
  $path//w[. = $word]},
  1, 100)
  let $title := $wordIC/ancestor::page/title
  return
   ($title,
   <s>{string-join($wordIC/../*|$wordIC/../text(), " ")}</s>,
   <pre>{
   string-join($wordIC/preceding-sibling::*[position() > last() - 5]|
    $wordIC/preceding-sibling::text()[position() > last() - 5],
   " ")
   }
   </pre>,
   <w>{$word}</w>,
   <post>{
   string-join($wordIC/following-sibling::*[position() < 5]|
   $wordIC/following-sibling::text()[position() < 5],
   " ")
    }
   </post>,
   "")
 order by $count descending
 return
 (<n>{$count}</n>,<w>{$word}</w>,<context>{$example-sentences}</context>)
 )
```

Die Möglichkeiten sind durch die technischen Gegebenheiten der exist-db Datenbank etwas eingeschränkt. Massenabfragen und Zählungen großer Mengen von Wörtern führen zu unbrauchbar langen Abfragezeiten.

Die eigentliche Zählung wurde mit einem für ägyptisches Arabisch entwickelten Java Programm am reinen Text durchgeführt. Das Programm ist unter https://github.com/simar0at/word-counter abrufbar.

Es werden zuerst alle Zeichenketten mit mindestens einem arabischen Buchstaben gezählt. Dann werden Gruppen gebildet, indem alle möglichen Varianten, etwa für den Antritt von Prä- und Suffixen, ه und Varianten, ي und ى, ausprobiert werden. Durch Prüfung dieser Ergebnisse wurde eine Liste von Kombinationen erstellt, bei der für die ersten 200 manuell geprüft wurde, welche nicht zusammengeführt werden dürfen, und für weitere 300, welche sinnvollerweise zusammengeführt müssen. Dadurch wurde das Ergebnis so lange verfeinert, bis die 200 häufigsten Wörter gefunden waren.

Es ist nicht sicher, ob mit diesem Programm andere Aufgaben ebenso erfüllt werden können, und es wurde nicht geprüft, wie hoch die Fehlerrate bei den auf 200 folgenden Rängen tatsächlich ausfällt.

## Anhang III Richtlinien zum Schreibstil in der Wikipedia Masry

Auf der Seite ويكيبيديا:سياسات/طريقة الكتابه „*siyāsāt/ṭarīqit il-kitāba*“ in der Wikipedia Masry steht nachzulesen, welchen Stil die Gemeinschaft der Autoren dort pflegen möchte.

Die Art zu Schreiben in der Wikipedia Masry, der ägyptischen Ausgabe der „Wikipedia“, die in moderner ägyptischer Sprache geschrieben ist, ist jene, in der die Ägypter miteinander reden. Das Geschriebene ist wie das, was sie in ihren Briefen schreiben und in ihrem täglichen Leben. Die Ägypter schreiben das Ägyptische in ihren Erzählungen und Theaterstücken, in ihren volkstümlichen Gedichten und auch auf dem Gebiet der Komödie, der Werbung und auch zum Teil in den Journalen. Man kann das Ägyptische schreiben auf jede Art, solange jeder Ägypter es verstehen kann. Mag es auch dauern, bis du weißt, wie du ägyptisch schreibst, schreib, wie du es in der Wikipedia Masry siehst (tiʿrafuh).

Inhalt:



## Allgemeine Grundlagen

1. Sei entspannt [ḫad rāḥtak] solange du niemanden ausgrenzt. [Beim Schreiben darf man annehmen]
2. Versuch denjenigen als Quellen zu folgen, die dir deine Sprache gegeben haben.
3. Schreib, ändere, übersetze, übertrag Artikel aber vereinfache nicht zu sehr [ma tihayyinš ḥadd]
4. Schreib in der Art, die du magst, aber das Kairenisch [al-qāhirīya], das ist die bekannteste Art um ägyptische Sprache aufzuschreiben, das Wichtigste ist, dass es die Ägypter verstehen.
5. Versuch neutral zu sein.
6. Stell niemanden bloß.
7. Bring keine Texte ein, deren Recht auf Verbreitung vorbehalten ist. [Also kein urheberrechtlich geschütztes Material.]
8. Versuch dir deinen Benutzernamen [Log in, Eintreten] zu registrieren, damit deine Beiträge mithilfe deines Namens registriert werden nicht mit dem Namen deiner IP-Adresse und vergiss nicht zu unterschreiben mithilfe von „~~~~“ [das ist das Kürzel mit dem eine Unterschrift/Signatur im Namen des gerade angemeldeten Benutzers in einem MediaWiki ausgelöst wird] auf den Diskussionsseiten.

9. Rechtschreibfehler werden passieren werden passieren [sic! Es dient gleich als Beispiel für einen Rechtschreibfehler], kein Problem, weil jeder Einzelne von den Usern [yūzirāt] sie ausbessern kann.
10. Versuch immer den Grund deiner Änderung zu erklären und besonders den Grund, warum du eine Änderung von einem der anderen User wieder rückgängig gemacht hast.
11. Versuch neue Beitragende zu ermutigen und zu unterstützen.
12. Wikipedia ist eine offene Enzyklopädie, in der du ändern und erweitern kannst und falls du etwas gesehen hast, das falsch ist, dann stell es richtig, solange du damit nicht beleidigst.

# Ratschläge

Wähl eine Art, die dir zusagt, und **halt dich** daran.

- Wenn du dich entscheidest, in einem Artikel zu schreiben, in dem eine bestimmte Art der Rechtschreibung verwendet wird, dann bitte halt dich an dieselbe [Art] im Rest des Artikels.
- Es nutzt nichts, wenn in einem Artikel auf so und so eine Art geschrieben ist und in derselben Zeile [nutzt du deine Art], und es nutzt nichts, wenn ein Wort auf so und so eine Art geschrieben ist und im selben Artikel [nutzt du deine Art] *(außer wenn deutlich ist, dass man es auf so und so eine Art schreibt, von Anfang an).*
- Bitte, **ändere nicht die Art der Rechtschreibung** in den Artikeln solange der ganze Artikel in eine und derselben Art Rechtschreibung geschrieben wurde. Und wenn du ihn erweitern möchtest, schreib in derselben Art Rechtschreibung, die er schon hat.
- Verwende möglichst diese Buchstaben in den Titeln: ا أ إ آ ب ت ث ج ح خ د ذ ر ز س ش ص ض ط ظ ع غ ف ق ك ل م ن ه ة و ي ء ى
- Verwende möglichst Buchstaben, die klar sind, im inneren der Seite, wie پ [p], چ [ʒ] ڤ [engl. v]

  Schreib فيفا oder ڤيڤا [viva]. Schreib ايفا oder ايڤا [engl. Eva], [beides] kein Problem

Das Wichtigste ist du hältst dich an eine Art Rechtschreibung beim Schreiben des Texts auf derselben Seite. Ein Text ist weiter unten über das Schreiben in lateinischem Alphabet.

## Die Buchstaben

**ق**: Wenn es ein Wort gibt, von einer arabischen Wurzel in dem es den Buchstaben <ق> gibt, geht man darauf zurück, wie es ist.

**ث**: Wenn es ein Wort gibt von einer arabischen Wurzel oder aus irgendeiner Sprache, in der es diesen Buchstaben/diese Aussprache gibt, (ث[θ]) und es wird /س[s]/ ausgesprochen, dann geht man darauf zurück, wie es ist. Wenn es /ت[t]/ ausgesprochen wird, dann schreibt man <ت>

**ذ**: Wenn es ein Wort gibt von einer arabischen Wurzel oder aus irgendeiner Sprache, in der es den Buchstaben/die Aussprache (ذ[ð]) gibt, und es wird /ز[z]/ ausgesprochen, dann geht man darauf zurück, wie es ist. Wenn es /د[d]/ ausgesprochen wird, schreibt man <د>.

Die Hamza [glottal stop, Glottisschlag, ء mit seinen diversen Trägern]:Es gibt keine trennenden Hamza und keine stummen Anfangshamza [hamzāt waṣl], alle sind <ا> (außer wenn es im Text ein anderes Wort gibt), aber wenn du willst, dann schreib trennende Hamza innerhalb der Artikel, kein Problem.

**ى**: Es gibt kein <ي>, es gibt <ى>; weil <ى> ist die Schreibweise, wie es die Ägypter schreiben. <ي> ist kein Problem, weil <ى> und <ي> sind ein Buchstabe.

**Das t der Weiblichkeit [tah ᵉt-ta'nīṯ, ة]**: Das t der Weiblichkeit schreibt man immer der Aussprache nach entweder /ة/ oder /ه/, das heißt man sollte schreiben <انا رايح المكتبه[der letzte Buchstabe ist ه, weil es als /a/ gesprochen wird]> und man sollte schreiben <انا رايح مكتبة الكليه[ة steht hier am Ende des dritten Wortes, weil es in einer Genitivverbindung mit dem 4. Wort steht und /it/ gesprochen wird. Das vierte Wort endet wieder auf /a/]>.

**Partikel, die den Genitiv regieren, und Konjunktionspartikel** [Partikel hier hauptsächlich in dem Sinn, dass es sich um einzelne Buchstaben handelt, die grafisch mit dem nächsten Wort verbunden werden.]: Das Schreiben von Partikel, die den Genitiv regieren, und Konjunktionspartikel ist klarer und einfacher, wenn sie getrennt geschrieben werden von dem Wort, das nach ihnen kommt, das heißt schreib <فى اروبا [in Europa]> oder <ف اروبا [das gleiche, aber es berücksichtigt, dass /fi/ eben mit einem kurzen i gesprochen wird und es ist nicht mehrdeutig mit /fī/ mit langem ī im Sinne von „es gibt"]>, [das ist]klarer und einfacher als das Lesen von <فاروبا [denn das kann auch /fa-.../ gelesen werden was dann „und, und auch usw." bedeuten kann]>.

**Vokalzeichen**: Vokalzeichen sind immer dann [saʿāt bitkūn] wichtig, wenn dasselbe Wort mit ihnen irgendeine Bedeutung hat, wenn es auf diese Art ausgesprochen wird.

## Punkte und Beistriche

**Punkte und Beistriche sind sehr wichtig.** Es wäre schön, wenn du sorgfältig bist [tāḫud bālak] bei Punkten und Beistrichen, damit man nicht verloren geht mitten im Text. Sie werden direkt mit dem Wort verbunden [lāziʿīn fi] geschrieben, das vor ihnen geschrieben wurde, und nach ihnen folgt ein Leerzeichen [misāfa].

## Eigennamen

Schreib die Eigennamen so, wie sie die Ägypter schreiben, das heißte schreib über <ڤينيسيا [vinisiya, Venedig]> nicht über <البندقية [al-Bunduqīya, Venedig, wie es hocharabisch richtig geschrieben wird. Das ist homonym mit بندقية als Flinte, Gewehr, was wohl in Ägypten geläufiger ist. Welchen arabischen Buchstaben man bei einer Transkription für /v/ nimmt, ist nicht zuletzt vom Dialekt des oder der Transkribierenden abhängig, auch (ب) ist möglich. Genauso kann (ق) in anderen arabischen

Dialekten als /g/ gesprochen werden.[240]] > und schreib <انجلترا [Ingiltīra, England] > nicht <إنكلترا [England, wie es hocharabisch richtig geschrieben wird. Ein Grund dafür, warum die ägyptische Schreibweise auf Hocharabisch keinen Sinn ergibt, ist sicherlich, dass in anderen Teilen der arabischen Welt (ج) als /d͡ʒ/ oder /ʒ/ ausgesprochen wird und nicht regelmäßig als/g/ wie in Ägypten.] >.Und schreib <البرت اينشتاين [Albert Einstein] > nicht <البرت ,اينشتاين [Einstein, Albert] >.

## Warum Wörter ägyptisieren?

Hauptartikel: ويكيبيديا:أسئله متكرره [FAQ]

Wenn der Ägypter über etwas modernes Sprechen möchte, dann äyptisiert er ein Fremdwort; beispielsweise eine Seite **saven** [سيڤ /seyiv/], ein Wort, das von „save“ kommt, wird hier [also in der Benutzeroberfläche, die man angezeigt bekommt, wenn man eine Seite ändern möchte, was genau bei dieser Seite nicht möglich ist] verwendet, weil es in der ägyptischen Sprache verwendet wird, und das Wort احفظ [speichern auf Hocharabisch, aber in anderem Kontext be- oder verwahren, behüten und auswendig lernen des Koran], dessen Bedeutung ist „erinnere dich“, wie etwa beim Lernen und Wiederholen. Immer wenn über eine Sache geredet wird am Computer, dann ägyptisiert er [der Ägypter] Fremdwörter, wie الديسك [die Disk], الدرايڤ [das Drive] und السكرين [der Screen].

## Schreibstil

Außer bei einer buchstäblichen Übertragung einer Aussage, die jemand macht, nimm Abstand vom Gebrauch der persönlichen Personalpronomen in Artikeln und vom Stil einer Ansprache, wie *(Ich, wir, …)*, denn diese Art ist nicht enzyklopädisch.

Beispiel: „Wir sind Ägypter – Wir sind Araber – So [zayy ma] führten ~~**wir**~~ 4 Kriege in 25 Jahren. ~~**Wir haben**~~ Verbindungen zu den Griechen, zu den Italienern und anderen. […] und wenn ~~**dir**~~ die Wahrheit über Alzheimer nicht ~~**selbst**~~ klar ist, dann **geh** und lass ~~**dich**~~ beraten.

## Artikel in lateinischem Alphabet

Hauptartikel: Artikel in lateinischer Schrift – lateinisches ägyptisches Alphabet – Beobachtungen über das Schreiben in lateinischer Schrift

Es gibt ganz wenige Artikel in der ägyptischen Wikipedia, bei denen der Text in lateinischem Alphabet ist, denn es gab seit dem Jahr 1948 Vorschläge zum Schreiben von ägyptisch in lateinischen Buchstaben, und obwohl die Mehrheit der Ägypter nicht an das lateinische Alphabet gewöhnt ist, gibt

[240] Eine Diskussion über dieses und andere Beispiele ist in den Argumenten gegen eine eigene ägyptische Wikipedia nachzulesen. Dem Autor ist der korrekte Name für Venedig auf hocharabisch bis heute auch noch nicht untergekommen.

es ganz wenige Materialien in der ägyptischen Wikipedia in Lateinisch (in „الفرانكو“), die in ihr schreiben sollen [Text unklar].

## Übernahmen [von Artikeln], die auf Arabisch sind

Übernahmen [von Artikeln] in klassischem Arabisch oder in modernem Standardarabisch, die bleiben mit den Buchstaben geschrieben, die die Sprache von jemandem hat oder Gedichte:

**Die Hamza:**.

**ى**: <ى> ohne Punkte, wie es in Ägypten verwendet wird.

**Das t der Weiblichkeit**:.

**Partikel, die den Genitiv regieren, und Konjunktionspartikel:**.

**Vokalzeichen:** Vokalzeichen sind immer dann [saʿāt bitkūn] wichtig, wenn dasselbe Wort mit ihnen irgendeine Bedeutung hat, wenn es auf diese Art ausgesprochen wird.

Siehe auch

Hilfe für neue User

Problem des Font

Wikipedia: Manual of Style

Kategorie: Richtlinien der ägyptischen Wikipedia[241]

[241] Wikipedia contributors: ويكيبيديا:سياسات/طريقة الكتابه. http://arz.wikipedia.org/wiki/%D9%88%D9%8A%D9%83%D9%8A%D8%A8%D9%8A%D8%AF%D9%8A%D8%A7:%D8%B3%D9%8A%D8%A7%D8%B3%D8%A7%D8%AA/%D8%B7%D8%B1%D9%8A%D9%82%D8%A9_%D8%A7%D9%84%D9%83%D8%AA%D8%A7%D8%A8%D9%87 (Zugriff am: 28. 11. 2012).

# Bibliographie

## Wörterbücher

### Arabisch

Wehr, Hans: Arabisches Wörterbuch für die Schriftsprache der Gegenwart. Arabisch - deutsch. 5. Auflage. Wiesbaden, 1985.

### Kairenisch

Abdel Aziz, Mohamed: Ägyptisch-Arabisches Wörterbuch. 7200 Wörter ; deutsch-äyptisch-arabisch, ägyptisch-arabisch-deutsch. 2. Auflage. Zürich, 2007.

Hinds, Martin; Badawi, as-Said Muhammad: A dictionary of Egyptian Arabic. Arabic-English. Beirut, 1986.

### Frequenzwörterbücher

#### *Arabisch*

Buckwalter, Tim; Parkinson, Dilworth B.: A frequency dictionary of Arabic. Core vocabulary for learners. London, 2011.

#### *Andere Sprachen*

Francis, Winthrop Nelson; Kučera, Henry: Frequency analysis of English usage. Lexicon and grammar. Boston, 1982.

Jones, Randall L.; Tschirner, Erwin: A frequency dictionary of German. Core vocabulary for learners. 1. Auflage. London, 2006.

### Grundwortschätze

#### *Arabisch*

Fouad, Magdi: Grundwortschatz Arabisch. 5000 Wörter zu 85 Themen ; [Niveau A1 - B2]. 1. Auflage. Ismaning, 2011.

Kendall, Elisabeth: The top 1,000 words for understanding media Arabic. Washington, DC, USA, 2005.

Kendall, Elisabeth: The top 1,300 words for understanding media Arabic. Washington, DC, USA, 2012.

#### *Kairenisch*

Abdel Aziz, Mohamed: Wörterbuch Grundwortschatz. Deutsch - Ägyptisch-Arabisch : Ägyptisch-Arabisch - Deutsch : mit Lautschrift. 4. Auflage. Zürich, 2007.

Jomier, Jacques: Lexique pratique français–arabe (parler du Caire). Kairo, 1976.

#### *Andere Sprachen*

Fink, Gerhard: Langenscheidts Grundwortschatz Latein. Ein nach Sachgebieten geordnetes Lernwörterbuch mit Satzbeispielen. Völlige Neubearb. Berlin, 2001.

## Monographien

Allemang, Dean; Hendler, James A.: Semantic web for the working ontologist. Effective modeling in RDFS and OWL. 2. Auflage. Amsterdam, 2011.

Bird, Steven; Klein, Ewan; Loper, Edward: Natural language processing with Python. [analyzing text with the natural language toolkit]. 1. Auflage. Sebastopol, Calif., USA, 2009.

Dalby, David; Barrett, David; Mann, Michael: The linguasphere register of the world's languages and speech communities. 1. Auflage. Hebron, Wales, UK, 1999.

Engelberg, Stefan; Lemnitzer, Lothar: Lexikographie und Wörterbuchbenutzung. 4. Auflage. Tübingen, 2009.

Ferguson, Charles A.; Dil, Anwar S.: Language structure and language use. Essays by Charles A. Ferguson. Stanford, Calif., USA, 1971.

Grimes, Barbara F.; Pittman, Richard S.; Grimes, Joseph Evans: Ethnologue. Languages of the world. 13. Auflage. Dallas, Texas, USA, 1996.

Habash, Nizar Y.: Introduction to Arabic natural language processing. San Rafael, Calif., USA, 2010.

Haeri, Niloofar: The sociolinguistic market of Cairo. Gender, class, and education. 1. Auflage. London, 1997.

Köhler, Reinhard: Zur linguistischen Synergetik. Struktur und Dynamik der Lexik. Bochum, 1986.

Landau, Jacob M.: A word count of modern Arabic prose. New York, NY, USA, 1959.

Lemnitzer, Lothar; Zinsmeister, Heike: Korpuslinguistik. Eine Einführung. 2. Auflage. Tübingen, 2010.

Leuf, Bo; Cunningham, Ward: The Wiki way. Quick collaboration on the web. 1. Auflage. Boston, Mass., USA, 2001.

Malina, Renate: Zum schriftlichen Gebrauch des kairinischen Dialekts anhand ausgewählter Texte von Saʿdaddīn Wahba. Berlin, 1987.

McEnery, Tony; Wilson, Andrew: Corpus linguistics. An introduction. 2. Auflage. Edinburgh, 2001.

Pflaumer, Peter: Statistik für Wirtschafts- und Sozialwissenschaften. 3. Auflage. München, 2005.

Somekh, Sasson: Genre and language in modern Arabic literature. Wiesbaden, 1991.

Spiro, Socrates: An Arabic-English vocabulary of the colloquial Arabic of Egypt. Containing the vernaculer idioms and expressions, slang, phrases, etc., etc., used by the native Egyptians. Cairo, 1895.

Suleiman, Yasir: A War of words. language and conflict in the Middle East. New York, NY, USA, 2004.

Tennison, Jeni: Beginning XSLT 2.0. From novice to professional. Berkeley, Calif., USA, 2005.

Walmsley, Priscilla: XQuery. Farnham, Calif., USA, 2007.

Woidich, Manfred: Ahlan wa sahlan. eine Einführung in die Kairoer Umgangssprache. 2. Auflage. Wiesbaden, 2002.

Woidich, Manfred: Das Kairenisch-Arabische. Eine Grammatik. Wiesbaden, 2006.

Woidich, Manfred; Heinen-Nasr, Rabha: Kullu tamam! An introduction to Egyptian colloquial Arabic. 2. Auflage. Berkeley, Calif., USA, 2004.

Zipf, George Kingsley: The psycho-biology of language. An introduction to dynamic philology. Cambridge, Mass., USA, 1965.

## Sammelwerke

Elgibali, Alaa; Badawi, El-Said Festschrift (Hrsg.) (1996): *Understanding Arabic. Essays in contemporary Arabic linguistics in honor of El-Said Badawi.* Cairo

Forte, Andrea (Hrsg.) (2011): *Proceedings of the 7th International Symposium on Wikis and Open Collaboration.* New York, NY, USA

Frishkopf, Michael Aaron (Hrsg.) (2010): *Music and media in the Arab world.* Cairo

Gugler, Josef (Hrsg.) (2011): *Film in the Middle East and North Africa. Creative dissidence.* Cairo

Guiter, Henry (Hrsg.) (1982): *Studies on Zipf's law.* Bochum

Hausmann, Franz Josef; Reichmann, Oskar; Wiegand, Herbert Ernst; Zgusta, Ladislav (Hrsg.) (1990): *Wörterbücher. Ein internationales Handbuch zur Lexikographie.* Berlin

Hill, Archibald A. (Hrsg.) (1962): *Third Texas Conference on Problems of Linguistic Analysis in English (May 9-12, 1958).* Austin, Texas, USA

Köhler, Reinhard (Hrsg.) (2005): *Quantitative Linguistik. Ein internationales Handbuch.* Berlin

Munske, Horst (Hrsg.) (2010): *Dialektliteratur heute – regional und international. Forschungskolloquium am Interdisziplinären Zentrum für Dialektforschung an der Friedrich-Alexander-Universität Erlangen-Nürnberg, 19.11.2009–20.11.2009.* Nürnberg

Schoneville, Catrin (Hrsg.) (2011): *Alles über Wikipedia und die Menschen hinter der größten Enzyklopädie der Welt.* Hamburg

Soudi, Abdelhadi; van Bosch, Antal den; Neumann, Günter (Hrsg.) (2007): *Arabic Computational Morphology. Knowledge-based and Empirical Methods.* Dordrecht

Versteegh, Cornelis H. M. (Hrsg.): *Encyclopedia of Arabic language and linguistics.* Leiden

## Internetdokumente

Allen, Julie D.; Anderson, Deborah; Becker, Joe; Cook, Richard; Davis, Mark; Edberg, Peter; Everson, Michael; Freytag, Asmus; Jenkins. John H.; McGowan, Rick; Moore, Lisa; Muller, Eric; Phillips, Addison; Suignard, Michel; Whistler, Ken: The Unicode Standard. Ch 2.: General Structure. http://www.unicode.org/versions/Unicode6.2.0/ch02.pdf (Zugriff am: 24. 11. 2012).

Caumanns, Jörg (1999): A Fast and Simple Stemming Algorithm for German Words. http://edocs.fu-berlin.de/docs/servlets/MCRFileNodeServlet/FUDOCS_derivate_000000000350/tr-b-99-16.pdf?hosts= (Zugriff am: 04.01.2013).

Cunningham, Ward: WikiWikiWeb. http://c2.com/cgi/wiki?WikiWikiWeb (Zugriff am: 24. 11. 2012).

Dalby, David (2012): The Linguasphere Register of the world's languages and speech communities. http://www.linguasphere.info/lcontao/fichiers-pdf.html (Zugriff am: 04. 02. 2013).

Derouin, Marie-Jeanne; Le Meur, André (2002): Presentation/Representation of Entries in Dictionaries. http://www.lrec-conf.org/proceedings/lrec2002/pdf/344.pdf (Zugriff am: 12. 01. 2013).

Dohrn, Hannes; Riehle, Dirk (2011): WOM: An object model for Wikitext. http://sweble.org/downloads/wom-tr.pdf (Zugriff am: 27. 01. 2013).

Dufrenoy, Alexis: Alsatian Wikipedia. http://comments.gmane.org/gmane.science.linguistics.wikipedia.international/2338 (Zugriff am: 25. 11. 2012).

Francopoulo, Gil (2007): Strategy for an OWL specification of LMF. http://www.tagmatica.fr/lmf/StrategyForLMFInOWL29october2007.pdf (Zugriff am: 22. 01. 2013).

Gates, Rick (1993): The Internet Encyclopedia. http://listserv.uh.edu/cgi-bin/wa?A2=ind9310d&L=pacs-l&T=0&P=1418 (Zugriff am: 16. 11. 2012).

Habsh, Nizar Y.; Rambow, Owen (2005): Arabic Tokenization, Part-of-Speech Tagging and Morphological Disambiguation in One Fell Swoop. http://dl.acm.org/citation.cfm?id=1219911 (Zugriff am: 04. 01. 2013).

Incubator contributors: Revision history of "Wp/arz". http://incubator.wikimedia.org/w/index.php?title=Wp/arz&action=history (Zugriff am: 18. 11. 2012).

Infoterm - Internationales Informationszentrum für Terminologie: STATUTEN des internationalen Vereines Infoterm - Internationales Informationszentrum für Terminologie. http://www.infoterm.info/pdf/about_us/InfotermStatuten_2011_FV.pdf (Zugriff am: 18. 11. 2012).

International Organization for Standardization: About ISO. http://www.iso.org/iso/home/about.htm (Zugriff am: 18. 11. 2012).

ISO/TC 37/SC4 (2008): Language resource management — Lexical markup framework (LMF). Rev 16. http://www.tagmatica.fr/lmf/iso_tc37_sc4_n453_rev16_FDIS_24613_LMF.pdf (Zugriff am: 13. 01 2013).

Khemakhem, Aida; Gargouri, Bilel; Ben Hamadou, Abdelmajid (2012): LMF standardized dictionary for Arabic Language. http://www.taibahu.edu.sa/iccit/allICCITpapers/pdf/p522-khemakhem.pdf (Zugriff am: 22. 01. 2013).

Kovitz, Ben: The conversation at the taco stand. http://en.wikipedia.org/wiki/User:BenKovitz#The_conversation_at_the_taco_stand (Zugriff am: 17. 11. 2012).

Larkey, Leah S.; Ballesteros, Lisa; Connell, Margaret E.: Improving Stemming for Arabic Information Retrieval: Light Stemming and Co-occurrence Analysis. http://ciir.cs.umass.edu/pubfiles/ir-249.pdf (Zugriff am: 10. 01. 2013).

Lemnitzer, Lothar; Kunze, Claudia (2003): Integrating Wordnets into the Resource Description Framework. http://www.cs.vassar.edu/~ide/events/NLPXML03_review/papers/Lemnitzer.pdf (Zugriff am: 22. 01. 2013).

Leuf, Bo; Cunningham, Ward: Quellecode für The Wiki Way. http://web.archive.org/web/20070822133615/http://www.leuf.net/ww/tww?WikiWaySources (Zugriff am: 24. 11. 2012).

Lyons, Melinda; SIL International: ISO 639-3 Change Requests Series 2011 Summary of Outcomes. http://www.sil.org/iso639-3/cr_files/639-3_ChangeRequests_2011_Summary.pdf (Zugriff am: 11. 12. 2012).

Märgner, Volker; El Abed, Haikal (2009): Arabic Word and Text Recognition. Current Developments. http://www.elda.org/medar-conference/pdf/46.pdf (Zugriff am: 26. 01. 2012).

Mediawiki contributors: MediaWiki history. http://www.mediawiki.org/wiki/MediaWiki_history (Zugriff am: 25. 11. 2012).

o. A.: ISO 639.2 Registration Authority - Frequently Asked Questions (FAQ). http://www.loc.gov/standards/iso639-2/faq.html#1 (Zugriff am: 18. 11. 2012).

o. A.: Most often used messages in MediaWiki. http://translatewiki.net/wiki/Most_often_used_messages_in_MediaWiki (Zugriff am: 18. 11. 2012).

o. A.: Project:About. http://translatewiki.net/wiki/Project:About (Zugriff am: 18. 11. 2012).

o. A.: Terms of Use. Licensing of Content. http://wikimediafoundation.org/wiki/Terms_of_Use#7._Licensing_of_Content (Zugriff am: 08. 01. 2013).

o. A.: WikiMatrix. Feature Comparison. http://www.wikimatrix.org/compare/MediaWiki+UseMod+DokuWiki+MojoMojo+PmWiki+TWiki+Zwiki (Zugriff am: 25. 11. 2012).

o. A.: WikiMatrix - Search for Wikis. http://www.wikimatrix.org/search.php?sid=137646 (Zugriff am: 25. 11. 2012).

o. A.: WikiMatrix - Search for Wikis. Kein CamelCase. http://www.wikimatrix.org/search.php?sid=137647 (Zugriff am: 25. 11. 2012).

o. A.: WikiWikiWeb - Wiki Page Counts. http://c2.com/cgi/wikiPages (Zugriff am: 24. 11. 2012).

o. A.: الصفحة الرئيسية.

http://beidipedia.wikia.com/wiki/%D8%A7%D9%84%D8%B5%D9%81%D8%AD%D8%A9_%D8%A7%D9%84%D8%B1%D8%A6%D9%8A%D8%B3%D9%8A%D8%A9 (Zugriff am: 28. 11. 2012).

Romary, Laurent; Inria & HUB-IDSL: TEI and LMF crosswalks. http://hal.inria.fr/docs/00/77/28/01/PDF/TEI_and_LMF_crosswalks.pdf (Zugriff am: 13. 01. 2013).

SIL International: What is SIL International? http://www.sil.org/sil/ (Zugriff am: 08. 01. 2013).

Smrž, Otakar: ElixirFM / Code / Commit [e0f0bd]. http://sourceforge.net/p/elixir-fm/code/ci/e0f0bdef24fac74815c795263b2ee73755bb44de/ (Zugriff am: 08. 01. 2013).

Stallman, Richard: GNU General Public License, version 2 (GPL-2.0). http://opensource.org/licenses/GPL-2.0 (Zugriff am: 16. 11. 2012).

Stallman, Richard (2000): The Free Universal Encyclopedia and Learning Resource. http://www.gnu.org/encyclopedia/anencyc.txt (Zugriff am: 16. 11. 2012).

Taghva, Kazem; Elkhoury, Rania; Coombs, Jeȉrey: Arabic Stemming Without A Root Dictionary. http://jeffcoombs.com/isri/Taghva2005b.pdf (Zugriff am: 10. 01. 2013).

TEI Consortium (eds) (2012): P5: Guidelines for Electronic Text Encoding and Interchange. http://www.tei-c.org/Vault/P5/2.2.0/doc/tei-p5-doc/en/html/ (Zugriff am: 13. 01. 2013).

Viégas, Fernanda B.; Wattenberg, Martin; Dave, Kushal (2004): Studying Cooperation and Conflict between Authors with history flow Visualizations. http://alumni.media.mit.edu/~fviegas/papers/history_flow.pdf (Zugriff am: 11. 01. 2013).

W3C: utf-8 Growth On The Web. http://www.w3.org/QA/2008/05/utf8-web-growth.html (Zugriff am: 06. 01. 2013).

Wales, Jimmy: Re: Re: Intlwiki-l Alsatian. http://permalink.gmane.org/gmane.science.linguistics.wikipedia.international/2375 (Zugriff am: 25. 11. 2012).

Wales, Jimmy (2001): Alternative language wikipedias. http://lists.wikimedia.org/pipermail/wikipedia-l/2001-March/000048.html (Zugriff am: 17. 11. 2012).

Wikimedia Foundation: Board member. https://wikimediafoundation.org/wiki/Board_member (Zugriff am: 17. 11. 2012).

Wikimedia Foundation: Bylaws - ARTICLE II - STATEMENT OF PURPOSE. https://wikimediafoundation.org/wiki/Wikimedia_Foundation_bylaws#ARTICLE_II_-_STATEMENT_OF_PURPOSE (Zugriff am: 17. 11. 2012).

Wikimedia Foundation: Special projects committee. http://meta.wikimedia.org/wiki/Special_Projects_Committee (Zugriff am: 17. 11. 2012).

Wikimedia Meta contributors: Language proposal policy. http://meta.wikimedia.org/wiki/Language_proposal_policy (Zugriff am: 18. 11. 2012).

Wikimedia Meta contributors: Proposals for closing projects. http://meta.wikimedia.org/wiki/Proposals_for_closing_projects (Zugriff am: 19. 11. 2012).

Wikimedia Meta contributors: Proposals for closing projects/Archive. http://meta.wikimedia.org/wiki/Proposals_for_closing_projects/Archive (Zugriff am: 19. 11. 2012).

Wikimedia Meta contributors: Requests for new languages. http://meta.wikimedia.org/w/index.php?title=Requests_for_new_languages&oldid=1096985 (Zugriff am: 18. 11. 2012).

Wikimedia Meta contributors: Requests for new languages/Wikipedia Egyptian Arabic. 1. April 2008. http://meta.wikimedia.org/w/index.php?title=Requests_for_new_languages/Wikipedia_Egyptian_Arabic&oldid=939972 (Zugriff am: 18. 11. 2012).

Wikimedia Meta contributors: Requests for new languages/Wikipedia Egyptian Arabic. http://meta.wikimedia.org/wiki/Requests_for_new_languages/Wikipedia_Egyptian_Arabic (Zugriff am: 13. 12. 2012).

Wikimedia Meta contributors: Requests for new languages/Wikipedia Egyptian Arabic. 30. März 2008. http://meta.wikimedia.org/w/index.php?title=Requests_for_new_languages/Wikipedia_Egyptian_Arabic&oldid=937275 (Zugriff am: 18. 11. 2012).

Wikipedia contributors: D Alemannisch Wikipedia. http://als.wikipedia.org/w/index.php?title=Wikipedia&stable=0#D_Alemannisch_Wikipedia (Zugriff am: 25. 11. 2012).

Wikipedia contributors: History of Wikipedia. http://en.wikipedia.org/wiki/History_of_Wikipedia (Zugriff am: 16. 11. 2012).

Wikipedia contributors: MediaWiki version history. http://en.wikipedia.org/wiki/MediaWiki_version_history (Zugriff am: 25. 11. 2012).

Wikipedia contributors: Rootschläg för Übersetzige. http://als.wikipedia.org/wiki/Hilfe:Ratschl%C3%A4ge_f%C3%BCr_%C3%9Cbersetzungen (Zugriff am: 16. 11. 2012).

Wikipedia contributors: Wikipedia:Edit-War. http://de.wikipedia.org/wiki/Wikipedia:Edit-War (Zugriff am: 24. 11. 2012).

Wikipedia contributors: Wikipedia:Five pillars. http://en.wikipedia.org/wiki/Wikipedia:Five_pillars (Zugriff am: 17. 11. 2012).

Wikipedia contributors: Wikipedia:Translation. http://en.wikipedia.org/wiki/Wikipedia:Translation (Zugriff am: 16. 11. 2012).

Wikipedia contributors: Wikipedia:Übersetzungen.
http://de.wikipedia.org/wiki/Wikipedia:%C3%9Cbersetzungen (Zugriff am: 16. 11. 2012).

Wikipedia contributors: اللغه المصريه الحديثه.

http://arz.wikipedia.org/wiki/%D8%A7%D9%84%D9%84%D8%BA%D9%87_%D8%A7%D9%84%D9%85%D8%B5%D8%B1%D9%8A%D9%87_%D8%A7%D9%84%D8%AD%D8%AF%D9%8A%D8%AB%D9%87 (Zugriff am: 13. 12. 2012).

Wikipedia contributors: «اللغه المصريه الحديثه»: تاريخ التعديل.

http://arz.wikipedia.org/w/index.php?title=%D8%A7%D9%84%D9%84%D8%BA%D9%87_%D8%A7%D9%84%D9%85%D8%B5%D8%B1%D9%8A%D9%87_%D8%A7%D9%84%D8%AD%D8%AF%D9%8A%D8%AB%D9%87&offset=20080812042737&action=history (Zugriff am: 13. 12. 2012).

Wikipedia contributors: ويكيبيديا:ترجمة مقالات إلى العربية.

http://ar.wikipedia.org/wiki/%D9%88%D9%8A%D9%83%D9%8A%D8%A8%D9%8A%D8%AF%D9%8A%D8%A7:%D8%AA%D8%B1%D8%AC%D9%85%D8%A9_%D9%85%D9%82%D8%A7%D9%84%D8%A7%D8%AA_%D8%A5%D9%84%D9%89_%D8%A7%D9%84%D8%B9%D8%B1%D8%A8%D9%8A%D8%A9 (Zugriff am: 16. 11. 2012).

Wikipedia contributors: ويكيبيديا:سياسات/طريقة الكتابه.

http://arz.wikipedia.org/wiki/%D9%88%D9%8A%D9%83%D9%8A%D8%A8%D9%8A%D8%AF%D9%8A%D8%A7:%D8%B3%D9%8A%D8%A7%D8%B3%D8%A7%D8%AA/%D8%B7%D8%B1%D9%8A%D9%82%D8%A9_%D8%A7%D9%84%D9%83%D8%AA%D8%A7%D8%A8%D9%87 (Zugriff am: 28. 11. 2012).

## Beiträge zu Sammelwerken

Alekseev, Pavel M. (2005): Frequency dictionaries. in: Reinhard Köhler (Hg.): *Quantitative Linguistik. Ein internationales Handbuch,* Berlin (= Handbücher zur Sprach- und Kommunikationswissenschaft), S. 312–324.

Davies, Humphrey: Dialect Literature. in: Cornelis H. M. Versteegh (Hg.): *Encyclopedia of Arabic language and linguistics,* Leiden ( 1. A-Ed), S. 597–604.

Dohrn, Hannes; Riehle, Dirk (2011): Design and Implementation of the Sweble Wikitext Parser. Unlocking the Structured Data of Wikipedia. in: Andrea Forte (Hg.): *Proceedings of the 7th International Symposium on Wikis and Open Collaboration,* New York, NY, USA, S. 72–81.

Ferguson, Charles A. (1996): Epilogue: Diglossia revisited. in: Alaa Elgibali, El-Said Festschrift Badawi (Hg.): *Understanding Arabic. Essays in contemporary Arabic linguistics in honor of El-Said Badawi,* Cairo, S. 49–67.

Kaye, Alan S.: Arabic Alphabet for Other Languages. in: Cornelis H. M. Versteegh (Hg.): *Encyclopedia of Arabic language and linguistics,* Leiden ( 1. A-Ed), S. 133–147.

Kubala, Patricia (2010): The Controversy over Satellite Music Television in Contemporary Egypt. in: Michael Aaron Frishkopf (Hg.): *Music and media in the Arab world,* Cairo, S. 173–224.

Ludwig, David; Schumann, Tobias (2011): Wikipedistik. in: Catrin Schoneville (Hg.): *Alles über Wikipedia und die Menschen hinter der größten Enzyklopädie der Welt,* Hamburg, S. 225–241.

Manske, Magnus (2011): MedieWiki - oder: Das Web 0.0. in: Catrin Schoneville (Hg.): *Alles über Wikipedia und die Menschen hinter der größten Enzyklopädie der Welt,* Hamburg, S. 283–295.

Martin, Willy (1990): 143. The Frequency Dictionary. in: Franz Josef Hausmann, Oskar Reichmann, Herbert Ernst Wiegand, Ladislav Zgusta (Hg.): *Wörterbücher. Ein internationales Handbuch zur Lexikographie,* Berlin (= Handbücher zur Sprach- und Kommunikationswissenschaft Bd. 5.2), S. 1314–1322.

Miller, Catherine: Dialect Koine. in: Cornelis H. M. Versteegh (Hg.): *Encyclopedia of Arabic language and linguistics,* Leiden ( 1. A-Ed), S. 593–597.

Prün, Claudia (2005): Das Werk von G. K. Zipf. in: Reinhard Köhler (Hg.): *Quantitative Linguistik. Ein internationales Handbuch,* Berlin (= Handbücher zur Sprach- und Kommunikationswissenschaft), S. 142–152.

Rapoport, Anatol (1982): Zipf's Law Re-visited. in: Henry Guiter (Hg.): *Studies on Zipf's law,* Bochum (= Quantitative linguistics).

Woidich, Manfred (2010): Von der wörtlichen Rede zur Sachprosa: Zur Entwicklung der Ägyptisch-Arabischen Dialektliteratur. in: Horst Munske (Hg.): *Dialektliteratur heute – regional und international. Forschungskolloquium am Interdisziplinären Zentrum für Dialektforschung an der Friedrich-Alexander-Universität Erlangen-Nürnberg, 19.11.2009–20.11.2009,* Nürnberg .

## Zeitschriftenaufsätze

Al-Sughaiyer, Imad A.; Al-Kharashi, Ibrahim A.: Arabic morphological analysis techniques: A comprehensive survey, in: *Journal of the American Society for Information Science and Technology* 2004 (2004), S. 189–213.

Ivan Panovic: The Beginnings of Wikipedia masry, in: *al-Logha, Series of Papers in Linguistic.*

Rosenbaum, Gabriel M.: Egyptian as a written language, in: *Jerusalem Studies in Arabic and Islam* 29 (2004), S. 281–326.

Wiegand, Herbert Ernst: Zugriffsstrukturen in Printwörterbüchern: Ein zusammenfassender Beitrag zu einem zentralen Ausschnitt einer Theorie der Wörterbuchform, in: *Lexicographica* 24 (2008).

Wiegand, Herbert Ernst: Semantik, Pragmatik und Wörterbuchform in einsprachigen Wörterbüchern, in: *Zeitschrift für germanistische Linguistik* 38 (2010), S. 405–441.

## Diplomarbeiten

Altabbaa, Mohammad; Al-Zaraee, Ammar; Shukairy, Mohammad Arif (2010): An Arabic Morphological Analyzer and Part-Of-Speech Tagger. قُطُوْف. http://qutuf.com/Files/Report_1.06.pdf (Zugriff am: 08. 01. 2013).

Elmaz, Orhan: Frequenzbasierte Lehrmaterialien für die arabische Mediensprache, 2010.

Kitzler, Gisela: Von Taxifahrern und Heiratskandidaten - zwei moderne ägyptische Bestseller im Kontext des Schreibens im Dialekt ("ʿāmmiyya"). Eine interdisziplinäre Analyse der Texte "Ich will heiraten" von Ġāda ʿAbd al-ʿĀl und "Taxi" von Ḫālid al-Ḫamīsī, 2012.

## Dissertationen

Farnawānī, Rif'at al: Ägyptisch-Arabisch als geschriebene Sprache. Probleme der Verschriftung einer Umgangssprache, 1981.

Smrž, Otakar: Functional Arabic Morphology. Formal System and Implementation. Prag, 2007.

## Lebenslauf

| | |
|---|---|
| Name | Omar |
| | SIAM |
| Geburtsdatum | 03.11.1981 |
| Geburtsort | Judenburg in der Steiermark |
| | |
| Nationalität | Österreich |
| | unverheiratet |
| | |
| seit 01.09.2007 | **Universität Wien** |
| | Studium der Arabistik und Islamwissenschaft |
| bis 31.10.2007 | **Logotronic GmbH** |
| | Wien |
| | Softwareentwickler im Bereich Fernabfrage von Messnetzen |
| ab 01.09.2002 | |
| bis 10.06.2002 | **Höhere Technische Bundeslehr- und Versuchsanstalt TGM** |
| | **Fachbereich Nachrichtentechnik/Microelektronik** |
| | Wien |
| 10.06.2002 | Matura |