Large Arabic Web Corpora of High Quality: The Dimensions Time and Origin

Thomas Eckart, Faisal Alshargi, Uwe Quasthoff, Dirk Goldhahn
Natural Language Processing Group, University of Leipzig, Germany
Email: {teckart, alshargi, quasthoff, dgoldhahn}@informatik.uni-leipzig.de

Abstract

Large textual resources are the basis for a variety of applications in the field of corpus linguistic. For most languages spoken by large user groups a comprehensive set of these corpora are constantly generated and exploited. Unfortunately for modern Arabic there are still shortcomings that interfere with systematic text analysis. The use of the Arabic language in many countries with different cultural backgrounds and the political changes in many of these countries over the last years require a broad and steady text acquisition strategy to form a basis for extended analysis. This paper describes the Arabic part of the Leipzig Corpora Collection (LCC) which is a provider of freely available resources for more than 200 languages. The LCC focuses on providing modern text corpora and wordlists via web-based interfaces for the academic community. As an example for the exploitation of these resources it will be shown how wordlists reflect political and cultural concepts that can be automatically exploited for diachronic or spatial comparisons.

Keywords: Arabic corpus generation, text acquisition, comparative corpus analysis

1. Availability of Arabic Text Resources

For a language with such a large group of native speakers Arabic has (compared with similar languages) still a strong demand for large corpora and tools. Existing corpora like the Al-Hayat Corpus (Roeck, 2002), the Corpus of Contemporary Arabic (CCA) or the An-Nahar News Paper Text Corpus are valuable resources and widely used. Unfortunately many of the existing corpora or resources lack properties that are strongly desirable for their use in the scientific context. These shortcomings contain problems with their availability (in some cases only by using very specific interfaces), lack of currentness (a problem especially when dealing with ongoing political developments), high costs or strict licences that permit reuse and data aggregation. As some of these problems can't be eliminated in general (like in the context of copyright and personality rights) it would be desirable to have more resources that can be used with as less restrictions as possible and that can be useful for further progress in the exploitation of Arbic corpora and other text-based resources.

2. Arabic Resources at the LCC

The *Leipzig Corpora Collection* (LCC) collects digital text material for more than 20 years. Starting with a focus on European languages it became apparent that a lot of the developed strategies and tools could be reused for other languages as well. Over the last years the used tool chain for text acquisition and text processing was adopted to deal with non-Latin scripts and especially Arabic resources were created and constantly improved.

2.1. Text Acquisition Strategies

The *Leipzig Corpora Collection* (Goldhahn et al., 2012) combines different strategies for collecting textual data from the WWW. The main goal is to ensure that corpora of large extent and high diversity concerning topics or genres can be created for specific languages. Especially a language like Arabic that is spoken in many countries requires a variety of approaches to achieve this objective.

2.1.1. Generic Web Crawling

A framework for massively parallel Web crawling is applied that utilizes the standard Web crawler and archiver *Heritrix*[1] of the Internet Archive. Among other enhancements, it was enriched with means for the automatic generation of crawling jobs.

Heritrix is used in several ways. On the one hand whole Top Level Domains are crawled. In this case a small list of domains of a country of interest is used as an input. Heritrix is then configured to follow links within this top-level domain (TLD). This has been conducted for several countries where Arabic is an official language.
On the other hand News sources are downloaded using the Heritrix based Web crawler. Basis is a list of more than 32,000 news sources in about 120 languages provided by *ABYZ News Links*[2]. This service offers URLs and information regarding country and language. This way news texts for several Arabic countries were collected. This includes text data excluded in the TLD crawling because of non-country TLDs used such as ".com".

[1] http://webarchive.jira.com/wiki/display/Heritrix/Heritrix

[2] http://www.abyznewslinks.com

2.1.2. Distributed Web Crawling

FindLinks (Heyer and Quasthoff, 2004) is a distributed Web crawler using a client-server architecture. The Java-based client runs on standard PCs and processes a list of URLs, which it receives from the *FindLinks*-server.

FindLinks has been used with community support for several years and allowed us to crawl the WWW to a large extent.

2.1.3. Bootstrapping Corpora

In addition an approach similar to Baroni (2004) and Sharoff (2006) was applied. Frequent terms of Arabic are combined to form Google search queries and retrieve the resulting URLs as basis for the default crawling system.

A small set of frequent terms is needed for languages in question. Therefore existing corpora of the LCC or other sources such as the *Universal Declaration of Human Rights* (*UDHR*)[3] were utilized as a resource.

Based on these lists tuples of three to five high frequent words are generated. These tuples are then used to query Google and to collect the retrieved URLs, which are then downloaded.

2.1.4. Crawling of special Domains

Certain domains are beneficial sources for Web corpora since they contain a large amount of text in predefined languages.

One example is the free Internet encyclopedia Wikipedia, which is available in more than 200 languages and of course also contains a version in Arabic.

Wikipedia dumps for these languages, among them Arabic, were downloaded. *Wikipedia Preprocessor*[4] was used for further processing and text extraction.

2.2. Corpus Creation Toolchain

Necessary steps for the creation of dictionaries are text extraction (mostly based on HTML as input material), language identification (Pollmächer, 2011), sentence segmentation, cleaning, sentence scrambling, conversion into a text database and statistical evaluation.

An automatic and mainly language independent tool chain has been implemented. It is easily configurable and only few language-dependent adjustments, concerning e.g. abbreviations or sentence boundaries, have to be made.

In a final step statistics-based quality assurance is applied to achieve a satisfying quality of the resulting dictionaries (Quasthoff, 2006b) (Eckart, 2012). Using features such as character statistics, typical length distributions, typical character or n-gram distributions, or tests for conformity to well-known empirical language laws problems during corpora creation can be detected and corrected.

The processing of Arabic text required several changes to the existing toolchain. Most of the developed tools could be reused but specific configurations had to be changed. This includes changes to components like sentence segmentation or quality assurance procedures. Besides some minor problems the general system again proved to be stable enough as for other languages or scripts before.

2.3. Sentence Scrambling

For all corpora the sentences had to be "scrambled" to destroy the original structure of the documents due to copyright restrictions. This inhibits the reconstruction of the original documents. With respect to German copyright legislation this approach is considered safe.

2.4. Available Resources

Corpora of this collection are typically grouped regarding the dimensions language, country of origin, text type (newspaper text, governmental text, generic Web material, religious texts etc.) and time of acquisition. The following table gives an introduction of currently available resources. As the crawling is an ongoing process new corpora are added at least every year.

Currently there are country specific corpora for Algeria, Bahrain, Egypt, Iraq, Jordan, Kuwait, Libanon, Mauritania, Morocco, Oman, Palestine, Qatar, Sudan, Syria, Tunisia, United Arab Emirates and Yemen which mostly consist of newspaper articles. As an example table 1 shows the most frequent sources of input material for a Morocco Web corpus.

Domain	*Number of documents*
www.riyada.ma	40,488
bayanealyaoume.press.ma	40,005
www.aktab.ma	39,309
www.goud.ma/	37,270
www.almassae.ma	37,203
www.almassae.press.ma	33,371
www.attarikh-alarabi.ma	33,255

Table 1: Number of documents for the most frequent sources used for a Moroccan Web corpus from 2013

2.5. Available Interfaces

The corpora are available via different Web-based interfaces. There is a freely available web portal where a variety of information can be accessed based on a word level (like sample sentences, word co-coocurrences, co-co-occurrence graphs etc.)[5]. Furthermore many corpora can be downloaded for free in different formats. These include plain text versions of the textual material and also MySQL databases[6]. For the later the platform-independent

[3] http://www.ohchr.org

[4] http://sourceforge.net/projects/wikiprep/

[5] Arabic portal: http://wortschatz.uni-leipzig.de/ws_ara/

[6] http://corpora.informatik.uni-leipzig.de/download.html

browsing tool is provided which allows examining the corpus locally.

3.

3.1. Linguistic Variants and Spell Checking

There are large sets of linguistic variants for many Arabic terms in the corpora. This is due to different reasons: there are many Arabic dialects spoken in different countries (like the term خمسا *(five)* which is used in Saudi Arabia, or خمستلاف*(five thousand)* which is used in Egypt). Besides these regional specifics there are of course also a lot of spelling errors like خمسئة (*Five hundred*).

Table 2 gives a short impression of different variants of the same word including their word rank in a Arabic mixed corpus with more than 4 million sentences.

3.2. Diachronic Comparisons

The availability of diachronic corpora can be used to detect political, economic and even cultural changes. These changes directly reflect in journalistic texts and user generated content.

Table 3 shows an example of such a diachronic comparison. The word rank of several terms being used in political contexts are calculated for six newspaper corpora based on input material from several Arabic speaking countries for the years 2007 to 2012. As expected words being part of current controversial topics are subject of strong changes in their relative frequency which is reflected in their word class. As an example *Obama* does hardly occur before 2008, but has a dramatically increase in frequency over the next years, with its peak in 2009 with the election of Barack Obama as US president in January.

3.3. Comparisons between Countries and Regions

By using texts from different top level domains it is furthermore possible to compare the contextual use of words in different countries. Based on sentence co-occurrences the generated co-occurrences graphs directly reflect typical usage of a word in a country and hence political situation and opinions. By comparing these graphs it is possible to extract similarities and differences in the public perception of different kind of topics.

Figure 1 shows the typical contexts of the word الانتخابات (Election) for text corpora from Bahrain and Egypt from 2013. Apparently some of the co-occurring terms are the same for both corpora (like *parliament*, *politics, voting* and similar election-related terms). However there are also differences: in Bahrain we also see the term *women.* This is because of the novelty of women allowed to vote in elections in Bahrain. On the other side both graphs contain different words for *vote:* الاقتراع in Bahrain and its Egypt correspondent التصويت.

4. Outlook

This corpora collection will continue in aggregating Web-based text material to extend the amount and quality of available resources. The result of these efforts will be furthermore provided to all interested users. Until mid of 2014 a new Web portal will be deployed that provides extended functionality and a more user-friendly interface. The underlying RESTful web services are also openly available and can be used for external applications as well. As a next step in exploiting word lists as a valuable resource in information extraction and language comparison it is planned to publish a book in the series of frequency dictionaries focusing on word frequency information in the Arabic language.

5. References

Al-Sulaiti L., Atwell E. (2004), Designing and developing a corpus of Contemporary Arabic. In *Proceedings of the sixth TALC conference*. Granada, Spain.

Baroni, M., Bernardini, S. (2004). BootCaT: Bootstrapping corpora and terms from the web. In *Proceedings of LREC 2004*.

De Roeck, A. (2002). ELRA's Al-Hayat Dataset: Text Resources in Arabic, Language Engineering. In *ELRA Newsletter 2002*, Vol.7 No.1.

Eckart, T.; Quasthoff, U.; Goldhahn, D. (2012). Language Statistics-Based Quality Assurance for Large Corpora. In *Proceedings of Asia Pacific Corpus Linguistics Conference 2012*, Auckland, New Zealand.

Goldhahn, D., Eckart, T., Quasthoff, U. (2012). Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. In *Proceedings of LREC 2012* (pp. 759-765)

Heyer, G., Quasthoff, U. (2004). Calculating Communities by Link Analysis of URLs. In *Proceedings of IICS-04*, Guadalajara, Mexico and Springer LNCS 3473.

Pollmächer, J. (2011). Separierung mit FindLinks gecrawlter Texte nach Sprachen. Bachelor Thesis, University of Leipzig.

Quasthoff, U., Biemann, C. (2006). Measuring Monolinguality. In *Proceedings of LREC 2006 Workshop on Quality assurance and quality measurement for language and speech resources.*

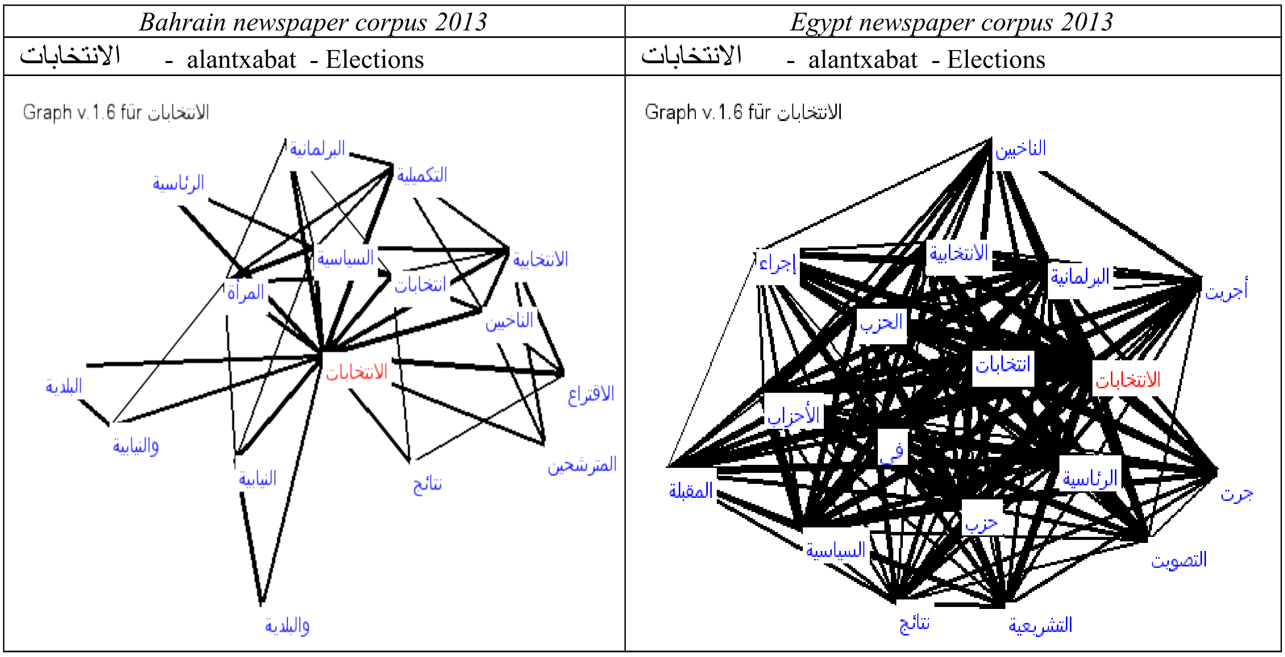
Sharoff, S. (2006). Creating general-purpose corpora using automated search engine queries. In M. Baroni and S. Bernardini, editors, WaCky! Working papers on the Web as Corpus. Gedit, Bologna.

Words	*Correct form*	*Translation in English*	*Rank in Wordlist*	*Comment*
خمة	خمسة	Five	751,951	Spelling error
خمسا	خمسة	Five	85,625	Used in Saudi dialect
خمس	خمسة	Five	1,122	All Arabic MSA and Dialects
خمسائة	خمسمائة	Five hundred	359,873	Spelling error
خمستعشر	خمسة عشر	Fifteen	1,438,010	Used in Yemen dialect
خمستلاف	خمسة آلاف	Five thousand	1,438,011	Used in Egypt dialect
خمسئة	خمسمائة	Five hundred	1,438,019	Spelling error
خمسيه	خمسمائة	Five hundred	751,973	Used in Jordan dialect

Table 2: Examples for language variants and spelling errors in MSA

English	*Term*	*2007*	*2008*	*2009*	*2010*	*2011*	*2012*
Democracy	الديمقراطية	631	721	453	655	347	500
Israel	إسرائيل	168	118	88	99	114	195
Obama	أوباما	10485	173	93	195	187	630
Elections	الانتخابات	170	141	97	153	138	158
Rights	الحقوق	683	2063	1180	1590	2892	1507
Iran	إيران	141	190	104	147	215	291
Freedom	الحريه	1635	1372	1175	656	699	636
Gaddafi	القذافي	1959	1894	2134	3804	79	589
Brotherhood	الاخوان	6556	5763	23147	15725	5122	2895

Table 3: Word rank of different terms in Arabic newspaper corpora from 2007 to 2012


Bahrain newspaper corpus 2013

الانتخابات - alantxabat - Elections

Egypt newspaper corpus 2013

الانتخابات - alantxabat - Elections

Figure 1: Word co-occurrences graphs of two newspaper corpora based on material from Bahrain and Egypt in 2013