Abu El-Khair Corpus: A Modern Standard Arabic Corpus

Ibrahim Abu El-Khair
Information Science Dept., Faculty of Social Sciences, Umm Al-Qura University

Abstract: Language Corpus is a collection of texts written in a single language, or more than one. It is collected, stored, and processed, automatically in an orderly fashion on computers according to its usage and purpose. Language corpus is one of the most important sources in the field of information retrieval, natural language processing, as well as computational linguistics. Building corpora is a common and familiar task in these areas of research for years. Corpora size have increased dramatically in recent years; because of the rapid developments in the technology used in their construction. This study is an attempt to build a contemporary linguistic corpus for Arabic language. The corpus produced, is a text corpus includes more than five million newspaper articles. It contains over a billion and a half words in total, out of which, there is about three million unique words. The data were collected from newspaper articles in ten major news sources from eight Arabic countries, over a period of fourteen years. The corpus was encoded with two types of encoding, namely: UTF-8, and Windows CP-1256. Also it was marked with two mark-up languages, namely: SGML, and XML.

Keywords: Language Corpora, Arabic Corpus, Natural Language Processing, Information Retrieval

I. INTRODUCTION

The efficiency of any information retrieval systems mainly depends on the experiments conducted by the researchers in the field, and commercial companies producing these systems. These experiments are done to emulate real world queries submitted to any system and the response of it to these queries. It is usually conducted in a closed laboratory environment. Elements of the retrieval process in this type of experiments are controlled by the researchers, in order to determine causes of success or failure and fixing it.

Language corpora are one of the most important elements for information retrieval experiments in particular and for natural language processing in general. This is because the corpus represents the actual everyday use of the language. Corpus use in retrieval has improved significantly in most languages especially Latin based languages. As for Arabic language it is still relatively new.

Arabic Language is the language of the holy Quran. It is used by more than a billion and a half Muslims around the world in the daily rituals. It is the mother tongue of about two hundred and fifty million people around the world. It is also, the official language of twenty-two countries and an official language for non-Arabic countries like Chad, Eritrea, Mali, and Turkey (Encyclopaedia Britannica Almanac, 2009). Moreover, it is one of the six official languages of the United Nation (UN, 2015), since 1973 (UN, 1973).

In spite of all of the above, Arabic language Corpora still in need for more research and studies. There is an ongoing need for more Arabic Corpora. The majority of available corpora now are relatively small in size, or rather expensive. The main purpose of this paper is producing a new free corpus. A corpus with a large size, representative of the language, from different countries, different writing styles, from more than one source, and distributed over many years. It will be available for researchers in the field of information retrieval, computational linguistics, and natural language processing.

II. AVAILABLE ARABIC CORPORA

This part of the paper reviews some previous attempts to create Arabic corpora. It should be noted that the review will be limited to textual monolingual corpora, not word lists, lexicons, speech, and
opinion corpora, all types were reviewed by Zaghouani, (2014). It is divided into two sections, freely available corpora with its related studies, and commercially available corpora with its related studies.

2.1. Free corpora

2.1.1. SACS Corpus (Saudi Arabian National Computer Science Conference):
Created by Hany Abu Salem (Abu Salem, 1992), modified and supplemented by Hmeidi, Kanaan, and Evens (1997). It consists of 242 Arabic abstracts collected from the proceedings of the Saudi Arabia National Computer Science Conference, containing 46,968 words in all. Each abstract includes 36 fields including title, authors, sources, and abstract. Each field starts with three characters that represent the name of the field followed by a space and the text.

2.1.2. Al-Raya Corpus:
Created by Ahmed Hasnah (Hasnah, 1996). It contains of 187 full newspaper articles selected from the Al-Raya newspaper, which is published in Qatar. This corpus contains 219,978 words in all and 30,096 unique words.
Although very small, in terms of their size, these two corpora are worth mentioning. They were built by individuals working with systems that had limited Arabic computational abilities. They were created at the Illinois Institute of Technology at the early stages of work on Arabic information retrieval experiments (Abu El-Khair, 2003; 2007). Both of them were available by personal contact with their producers.

2.1.3. Arabic Corpus Project:
A project (Saleh, 1999; Arabic Corpus Project, 2008) adapted by ALECSO (Arab League Educational, Cultural, and Science Organization). It aims at building a corpus that utilizes the actual use of Arabic language in all situations and all ages. It was supposed to incorporate: valuable literature and scientific articles in arts, science, and technology. It should include valuable lectures in arts and science. All Arabic dictionaries and lexicons, monolingual and multilingual. The main objective of the project was to build a digital bank for Arabic language, and its actual use. Unfortunately, there is no outcome of the project so far, except for some individual attempts to put some books online, four hundred books to be exact (Saleh, 2014).

2.1.4. Contemporary Arabic Corpus:
This corpus was created by Latifa Al-Sulaiti and Eric Atwell (Al-Sulaiti & Atwell, 2005; Al-Sulaiti & Atwell, 2006). It was originally created as part of Al-Sulaiti Master thesis (Al-Sulaiti, 2004). The main sources of the corpus are newspapers, emails, and websites from the beginning of the 1990s till 2004. It consists of 842,684 words and available for download for free. The main purpose of creating this corpus was to be used in information retrieval research, and teaching Arabic for non-native speakers. The corpus is one of the good ones available because of its balance in topics. It is tagged in XML language.

2.1.5. Akhbar El-Khaleeg:
This corpus was built by Abbas, (Abbas and Samaili, 2005) for research in the field of natural language processing, and it is available for free to researchers. It is a collection of "Akhbar El-Khaleeg" newspaper articles in Bahrain. It consists of three million words (4MB).

2.1.6. Al-Watan:
This is a collection of "Al-Watan" newspaper articles in Oman. It is created by Abbas and his colleagues (Abbas, Samaili, &Berkani, 20110), as an addition to the previous corpus. It consists of about ten million words (10MB). Both Abbas corpora are available for download for free.
2.1.7. KALIMAT Corpus:
This corpus (El-Haj, Koulali, 2013) used a part of both "Akhbar El-Khaleeg", "Al-Watan" newspapers as the starting point. KALIMAT is a multipurpose annotated corpus, with 20,291 articles, with 1,816,783 annotated words in raw text. The resulting corpus was in 2057 collocated files. It includes name entity annotation, and POS tagging, with full morphological analysis of all words.

2.1.8. Arabic Modern Standard Corpus:
This corpus was built by Ahmed Abdalali and his colleagues (Abdalali et al., 2005) especially for research in information retrieval and natural language processing. It is a collection of newspaper articles from ten Arab countries, namely: Egypt, Kuwait, Oman, Algeria, Lebanon, Kingdom of Saudi Arabia, Morocco, Jordan, Qatar, Syria, and Iraq. It has 102,134 articles, with 113 million words (800MB).

2.1.9. Open Source Arabic Corpus (OSAC):
This corpus was created by Moataz Saad (Saad, Ashour, 2010) from three main sources, namely: Arabic BBC, Arabic CNN, and various other sources. He collected 4763 articles in seven categories, with a total of 1.8 million words, with 106,733 unique words from the BBC. He also, collected 5070 articles, with 2.2 million words, and 144,460 unique words from CNN. Finally, he gathered 22,429 articles from several sources in ten subject categories, with 18 million words, and 449,600 unique words. The total number of words in the corpus was 22 million words. All articles were encoded with UTF-8.

2.1.10. The International Corpus of Arabic (ICA):
A team of researchers from Alexandria Library started working on this corpus in 2006 to help researchers in the field of Arabic language retrieval. The original plan was to have 100 million words, but currently it has about 80 million words (Alansary, & Nagi, 2014; Alansary, Nagi, & Adly, 2007). The corpus was collected from a large number of sources to be representative of the language use. It depends on Dewey decimal classification, with some minor modifications, in classifying the texts. It includes texts in eleven subject categories, namely: strategic sciences, social sciences, sports, religion, literature, humanities, natural sciences, arts, bibliography, and some other various topics. The total number of articles in the corpus so far are 70,022 with about 80 million words, and 1,272,766 unique words. It was collected over a period of twenty-one years, from 1993 till 2014. The majority of the articles were collected from newspaper sites like, Al-Ahram (Egypt), Al-destor (Jordan), and Al-Hayah (Lebanon). The second source was articles taken from blogs and forums. The Third source was electronic books. The fourth source was Academic research papers and dissertations. The corpus is about 100MB and is free by demand from the researchers.

2.1.11. King Saud University Corpus of Classical Arabic (KSUCCA):
This corpus was created by MahaAlrabiah as part of her Ph.D. dissertation (Alrabiah, 2012; Alrabiah, Al-Salman, & Atwell, 2013; Alrabiah, Al-Salman, & Atwell, 2014). The corpus was designed and compiled with the goal of supporting research in linguistics, computational linguistics, literature and history. The corpus consists of 50 million words, in raw text and is freely available for researchers. It consists of books on Islamic heritage. It is divided into six subject categories, namely: religion literature, linguistics, sciences, biographies, and sociology. It was collected from classical texts from the seventh century till the eleventh century.

2.1.12. University of Jordan Arabic Corpus (UJAC):
This corpus is a relatively small new one created by four researchers from Jordan University (Hammo et al., 2013). It was part of project to build tools and resources for Arabic language processing. It consists of articles from fifteen newspapers and other resources from nineteen countries. It
includes 61,037 articles, with 7,522,941 words, with 707385 unique words. It is encoded in UTF-8, with XML formatting.

2.1.13. King Abdulaziz City for Science and Technology (KACST) Corpus:
Also, called The Arabic Blog for King Abdulaziz City for Science and Technology (Arabic Blog, 2013; Al-Thubaity, 2014). It is one of the strategic projects of King Abdullah initiative for Arabic content. The corpus contains seven hundred million words starting from the pre-Islamic era to the modern age. It is collected from different regions and countries. Taking into account the nature and volume of intellectual activity for each period. Also, and diversity of publishing media, such as manuscripts, newspapers, books, magazines, scientific periodicals, etc. It is also collected from different scientific and non-scientific fields. The corpus has a running website on the Internet as well as tools to search it with linguistic and statistical analysis. The corpus contains about 869,800 text sources, and essays, with exactly 732,780,509 words, out of which, 7,464,396 are unique.

2.2. Commercial corpora
2.2.1. LDC Corpus (Arabic Newswire: part 1):
It was created by David Graff and Kevin Walker (Graff, & Walker, 2001) at the University of Pennsylvania’s Linguistic Data Consortium (LDC). The corpus is composed of articles from the Agence France Presse (AFP) Arabic Newswire published between May 13, 1994 and December 20, 2000. The source material was tagged using TIPSTER-style Standard Generalized Mark-up Language (SGML) and was transcoded to Unicode (UTF-8). The data are stored in 2,337 compressed Arabic text data files. There are 209 megabytes of compressed data (869 Mb when uncompressed), with 383,872 documents containing 76 million tokens over approximately 666,094 unique words. This corpus is perhaps the most used one in the field of Arabic information retrieval for a long time (Abu El-Khair, 2007). This is due to the fact that it is the largest corpus available with queries and relevance judgments.

2.2.2. An-Nahar Newspaper Text Corpus:
This corpus (An-Nahar Newspaper Text Corpus, 2001) contains articles published in the Lebanese newspaper An-Nahar (Al-Nahar, or The Daytime) from 1995 to 2000, which are stored as Hypertext Mark-up Language (HTML) files. Each year contains 45,000 articles and 24 million words. Each article includes information such as title, newspaper’s name, date, country, type, and page.

2.2.3. Al-Hayat Arabic Corpus:
The Al-Hayat corpus (Al-Hayat Arabic Corpus, 2001) was developed in the course of a research project at the University of Essex, in collaboration with the Open University. It contains Al-Hayat newspaper articles with value added for Language Engineering and Information Retrieval applications development purposes. The data have been distributed into seven subject-specific databases in accordance with Al-Hayat’s subject tags: General, Car, Computer, News, Economics, Science, and Sport. Mark-up, numbers, special characters, and punctuation have been removed. The total size of the file is 268 megabytes. The dataset contains 18,639,264 distinct tokens in 42,591 articles.

2.2.4. Arabic Gigaword Corpus:
This is a comprehensive corpus for Arabic news from some of the major newspapers and news agencies in the Arab world. It has five editions so far. The first edition (Graff, 2003), included all three previous corpora, Arabic Newswire: part 1 (France), An-Nahar Newspaper Text Corpus (Lebanon), and Al-Hayat Arabic Corpus (Lebanon). In addition to these sources LDC added a fourth source; the Xinhua Chinese news Agency (China). The corpus was encoded with UTF8, and written in SGML. It includes 1,256,719 articles, with a total of 391,619 unique words.
The second edition (Graff et al., 2006) included one more source, namely: The Ummah Press. The total number of articles increased to 1,591,987 articles, with 481,906 unique words. Graff (2007)
states that the third edition was supplemented with a sixth source. He added articles from Assabah News Agency (Tunisia). The total number of articles increased to 1,994,735 articles, with 576,799 unique words.

The fourth edition (Parker et al., 2009), included three more news sources, namely: Al-Ahram (Egypt), Asharq Al-Awsat (England), and Al-Quds Al-Arabi (England). The total number of articles increased to 2,716,995 articles, with 848,469 unique words. The fifth edition (Parker et al., 2011) did not include new sources, rather adding new articles to the sources already available in the fourth edition. Covering nine news sources from six different countries, the final corpus included 3,346,167 articles, with 1,077,382,000 words, and 848,469 unique words.

2.2.5. NEMLAR Corpus:
This corpus is part of the Network for Euro-Mediterranean LAnguage Resources (NEMLAR) project for developing Arabic resources in 2003-2005. It is a relatively small size corpus. It consists of half a million words in thirteen different categories. It is available at the European Language Resources Association website (NEMLAR Written Corpus, 2003; NEMLAR Project, 2010).

There were some minor changes to the corpus during work at a supplementary project to NEMLAR called MEDAR (Mediterranean Arabic Language and Speech Technology). Morphological analysis, diacritics, and POS tagging were added to the corpus (MEDAR, 2010; MEDAR Evaluation Package, 2010).

III. DATA COLLECTION
Web scrapping or web copying programs were used to extract text from news sources in order to create the corpus. The researchers used wget(1), which is used by LDC, and htttrack(2) site copier, but both were very slow, so they were not used. Two other programs, Internet Download Manager (3), cyotekwebcopy(4), were used and eliminated as well because they stop working for no apparent reason, in addition to being slow. After several attempts the researcher used MetaProducts Offline Explorer Pro(5), Visual Web Ripper(6). Both programs were very good in extracting text and eliminating all unnecessary objects like images, videos, JavaScript files, and CSS files.

3.1. Corpus Sources:
There are a lot of news sources that could be used for creating a language corpus. At this paper, the researcher has chosen ten sources to be used in the corpus. Several news websites were tested before selecting the source that will be used. The fame of the website, and the news source, or the number of readers were not the criterion for selection. There were other criteria and technical reasons for selecting the news resources used in building the corpus.

- The first criterion is having no overlap with previous Arabic corpora. For example, Al-Ahram newspaper from Egypt has the largest digital news archive on the internet, but were not selected because it is a part of the Arabic Gigaword Corpus.
- The source should be online for a long time. This is simply to have a large volume of articles available. This was perhaps one the major obstacles in conducting this study. Knowing when the newspaper appeared online, was a problem. There was no way of knowing that without checking each one individually since there is no website that could have this information.
- All selected sources should represent different countries in the Arab world.
- The scrapped text should be in an editable form.

2. https://www.httrack.com
3. https://www.internetdownloadmanager.com
The selected news source website should allow the crawling programs to work on it and import the articles. Some websites have very tight security procedures, and do not allow spidering. It should be noted that the news websites crawling was done between December 2013 and June 2014. Two of the sites, almustaqbal, sabanews, were re-crawled because of errors discovered in the quality control phase. There was a problem importing the publication date in them.

Table one, indicates the selected sources for the corpus, its name in English and in Arabic, its abbreviation, the time period for each one of them, country of origin, and its website. Nine newspapers, and one news agency from eight countries were selected as shown in the table. Egypt and Saudi Arabia are represented with two newspapers each, since they are the pioneers in online journalism, and have some of the oldest online newspapers in the Arab world.

<table>
<thead>
<tr>
<th>Source (English)</th>
<th>Source (Arabic)</th>
<th>Abbrev.</th>
<th>Country</th>
<th>From</th>
<th>To</th>
<th>Website</th>
</tr>
</thead>
</table>

Table 1. Corpus resources

The coverage period varies from one source to the other. The starting time in each news source is basically the time it first appeared online. The ending date depended on the time of the data collection. Some websites allowed harvesting the news archive but not the current news like Alyaum from Saudi Arabia, and Almasryalyoum from Egypt.

3.2. Metadata:

Two tagging schemes were used with the corpus in hand. All articles in the current corpus were tagged with SGML (Standard Generalized Markup Language), which is used in TREC corpora. The other scheme was using XML (Extensible Markup Language) tagging, which is used in the LDC corpora.

Each article will have an ID using the source abbreviation, table one, Arabic language abbreviation, and a serial number, e.g.

<ID> RYD_ARB_0000001 </ID>, or
<br/>&lt;DOCNO&gt;RYD_ARB_0000001 &lt;/DOCNO&gt;.
3.3. Encoding:
The corpus will be encoded with windows cp-1256\(^7\) for Arabic language. It will also be encoded with UTF-8\(^8\). Having two versions of the corpus with two different encoding schemes will be of great use for researchers in the field of Arabic information retrieval, and Natural language processing.

IV. RESULTS
As mentioned earlier, the corpus by itself is useless unless it is used to serve some a research area. The main purpose for creating this corpus, is to have a free tool for Arabic language available for researcher. It is made specifically for work in the field of information retrieval, or natural language processing.

The corpus is not limited to one subject. It is multitopic news corpus covering Politics, literature, arts, technology, sports, economy, culture, and many other subject matters. It is also, a good representation of Arabic language. It covers a period of fourteen years and eight countries. These countries have a very large portion of Arabic native speakers. Finally, all ten sources used in creating the corpus are well represented.

<table>
<thead>
<tr>
<th>Source</th>
<th>Articles</th>
<th>Percentage</th>
<th>Words</th>
<th>Percentage</th>
<th>Unique Words</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alriyadh</td>
<td>858,188</td>
<td>16.43%</td>
<td>271,353,697</td>
<td>17.79%</td>
<td>1,451,320</td>
<td>15.39%</td>
</tr>
<tr>
<td>youm7</td>
<td>1,025,027</td>
<td>19.63%</td>
<td>261,700,304</td>
<td>17.15%</td>
<td>1,020,444</td>
<td>10.82%</td>
</tr>
<tr>
<td>Alyaum</td>
<td>888,068</td>
<td>17.00%</td>
<td>237,914,494</td>
<td>15.59%</td>
<td>1,319,996</td>
<td>13.99%</td>
</tr>
<tr>
<td>Alqabas</td>
<td>817,274</td>
<td>15.65%</td>
<td>233,741,575</td>
<td>15.32%</td>
<td>1,260,511</td>
<td>13.36%</td>
</tr>
<tr>
<td>Alitihad</td>
<td>349,342</td>
<td>6.69%</td>
<td>139,962,699</td>
<td>9.17%</td>
<td>932,628</td>
<td>9.89%</td>
</tr>
<tr>
<td>Almustaqbal</td>
<td>446,873</td>
<td>8.56%</td>
<td>135,446,906</td>
<td>8.88%</td>
<td>982,765</td>
<td>10.42%</td>
</tr>
<tr>
<td>Tishreen</td>
<td>314,597</td>
<td>6.02%</td>
<td>94,695,378</td>
<td>6.21%</td>
<td>905,169</td>
<td>9.60%</td>
</tr>
<tr>
<td>Almasryalyoum</td>
<td>291,723</td>
<td>5.59%</td>
<td>93,398,135</td>
<td>6.12%</td>
<td>760,511</td>
<td>8.06%</td>
</tr>
<tr>
<td>Echorouk Online</td>
<td>139,732</td>
<td>2.68%</td>
<td>40,978,911</td>
<td>2.69%</td>
<td>543,799</td>
<td>5.77%</td>
</tr>
<tr>
<td>Saba News Agency</td>
<td>92,149</td>
<td>1.76%</td>
<td>16,530,153</td>
<td>1.08%</td>
<td>255,098</td>
<td>2.70%</td>
</tr>
<tr>
<td>Totals</td>
<td>5222973</td>
<td>100.00%</td>
<td>1,525,722,252</td>
<td>100.00%</td>
<td>3,303,723</td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Corpus Statistics according to the source.

Table two shows the statistics of the corpus in details, and what has been assembled from each source of ten sources. It includes the number and percentage of articles that have been imported from each source, and the total number and percentage of words and unique words for each source. It has been arranged based on the number of words; because they determine the value of each source for corpus. It should be noted that the total number of "unique words" is not equal to the addition of the values in the column; because all repeated words between sources are excluded.

V. CONCLUSION
Language corpus is a representation of the language use. It should be, according to Mansour’s principles (2013), large, have a specific purpose, diverse, representative, and well balanced. In order to have a general idea about the corpus in hand, in terms of size. Table one, shows the general statistics of the corpus. It indicates that the corpus has over five million articles from ten news sources. The total number of words exceeds 1.5 billion words, and the total number of unique words exceeds 3.3 million words.

---

8. [http://unicode.org/resources/utf8.html](http://unicode.org/resources/utf8.html)
### Table 3. General Statistics of the corpus

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of resources</td>
<td>Nine Newspapers, One news Agency</td>
</tr>
<tr>
<td>Number of countries covered</td>
<td>Eight Countries</td>
</tr>
<tr>
<td>Years covered</td>
<td>14 Years</td>
</tr>
<tr>
<td>Corpus Size</td>
<td>10GB (CP-1256) / 16GB (UTF-8)</td>
</tr>
<tr>
<td>Number of articles</td>
<td>5,222,973 Articles</td>
</tr>
<tr>
<td>Number of Words</td>
<td>1,525,722,252 Words</td>
</tr>
<tr>
<td>Number of Unique Words</td>
<td>3,303,723 Words</td>
</tr>
</tbody>
</table>

The KACST Corpus (Al-Thubaity, 2014), the largest free corpus available, created by a team from King Abdulaziz City for Science and Technology. They also outsourced 25% of the corpus to external specialists. It has 700 million words with about 1.5 million articles. The Arabic GigaWord corpus, which is the largest paid corpus available, was created by an institution like the LDC over a period of over ten years. It has 3.3 million articles, and 1.077 billion words.

### REFERENCES


39. Saleh, Abdul Rahman Al-Haj. (2014). This is the Arabic Linguistic Corpus Project; and this is the Algerian perception of it. [In Arabic]. Algerian news today, 06.25.2014. Retrieved on: 10/25/2015, from: http://www.akhbarelyoum.dz/ar/200243/200256/109357


