

# Building A Modern Standard Arabic Corpus

Ahmed Abdelali

Computing Research Laboratory  
New Mexico State University  
Box 30001/MSC 3CRL  
Las Cruces, NM 88001  
+1 (505) 646 5711  
ahmed@crl.nmsu.edu

Jim Cowie

Computing Research Laboratory  
New Mexico State University  
Box 30001/MSC 3CRL  
Las Cruces, NM 88001  
+1 (505) 646 5711  
jcowie@crl.nmsu.edu

Hamdy S. Soliman

Computer Science Dept.  
New Mexico Institute of Mining  
and Technology  
Socorro, NM 87801-0389  
+1 (505) 835-5170  
hss@nmt.edu

## ABSTRACT

Language Engineering, including Information Retrieval, Machine Translation and other Natural Language-related disciplines, is showing more interest in the Arabic language in recent years. Suitable resources for Arabic are becoming a vital necessity for the progress of this research.

Until recently, only two Arabic corpora were commonly available for researchers: the AFP Arabic newswire from LDC and the Al-Hayat newspaper collection from the European Language Resources Distribution Agency. But the necessity of a suitable corpus with a wider coverage that samples the language used over the vast region is a key for any objective research.

In this paper we present preliminary results of experiments with a corpus for Modern Standard Arabic using data available on the World Wide Web. We selected samples of online published newspapers from different Arabic countries. The selection was driven mainly by the amount of data available. We will demonstrate the completeness and the representativeness of this corpus using standard metrics and show its suitability for Language Engineering experiments.

## Keywords

Modern Standard Arabic, language/vocabulary variations, Corpus.

## 1. INTRODUCTION

The amount of Arabic data available on the World Wide Web is dramatically increasing daily. The report by Madar Research Journal which includes statistics and forecasts on Internet users in 17 Arab countries estimates the size of the Internet community in the Arab world in excess of 25 million by end of 2005 from the current 7.4 million [11]. According to a new study from the Research Unit of Internet Arab World magazine, there are currently 1.9 million online websites in Arabic and number is expected to double every year [17]. Providing users with quality web portals and efficient search engines is essential to keep up with the growth. The issue becomes even more serious when it comes to finding specific information on the net.

Another important concern is the variation in the Arabic language across the wide area where Arabic is spoken, which includes a large number of Arab countries [14, 16]. Significantly

there are elements in the language that could lead to connecting a particular text to a specific country or region [1]. We investigate this concern in a scientific manner using the latest methodologies

in the field of Natural Language Processing (NLP). The first step is the collection of a significant amount of data, providing a representative sample of actually occurring language over a wide geographical area. This collection will be the source for a corpus that will contain useful information and be useful in experimentation.

## 2. Why an Arabic Corpus?

Collecting manuscripts, books and newspapers for analysis is very laborious in nature. But this was done for a long time, particularly by Academic researchers. Thankfully, as technological advances make the computerized storage of and access to large quantities of information easier, so the construction and use of text corpora will continue to increase. As a result the potential for research has widened considerably [8, 13]. The importance of corpora to linguistic study is appreciated. A corpus to a linguist is very valuable because it allows statements to be made about language in very convincing fashion. The actual use of the corpus includes studies in the grammar, lexicography, language variations, historical linguistics, language acquisition, and language pedagogy.

Attempts to study Arabic using these types of resources were initiated by researchers in the NLP field. To evaluate Information Retrieval systems and morphological analyzers Al-Kharashi & Evens in 1994 [4] used a collection of bibliographical records to test their Information Retrieval Micro-AIRS. Hmeidi et al in 1997 [7] constructed a corpus of 242 abstracts collected from the proceedings of the Saudi Arabian national conference. Goweder A and De Roeck A. [6] produced an Arabic corpus using 42591 articles from Al-Hayat newspaper archive of the year 1998. The experiment was mainly to reproduce and confirm results made on small-scale corpus about the sparseness of the Arabic comparing to English. In 2001 LDC released the Arabic Newswire, A corpus composed of articles from the Agence France Presse (AFP) Arabic Newswire. The corpus was tagged using SGML and was transcoded to Unicode (UTF-8). The corpus includes articles from 13 May 1994 to 20 December 2000 with approximately 76 Million tokens and 666,094 unique words. In 2003 LDC also released Arabic Gigaword a bigger and richer corpora compiled from different sources that includes Agence France Presse, Al Hayat News Agency, Al Nahar News Agency and Xinhua News Agency.

Other efforts for building corpora for IR evaluation purposes were carried by Darwish et al. [5]. They used a collection called Zad which was provided by Al-Areeb Electronic Publishers, LLC. The collection contains 4,000 documents. The documents were extracted from writings of the thirteenth century scholar Ibn Al-

Qayim and cover issues of history, jurisprudence, spirituality, and mannerisms. Also, there are 25 queries with their relevance judgments associated with the collection. The collection was used in ad-hoc experiments to test the effects of the choice of Arabic index terms and the use of morphology. Xu, Fraser, and Weischedel [18] collected a parallel corpus from the United Nations (UN) published documents. The United Nations web site (<http://www.un.org>) publishes all UN official documents under a document repository, which is accessible by paying a monthly fee. A special purpose crawler was used to extract documents that have versions in English and Arabic. After a series of clean-ups, they obtained 38,000 document pairs with over 50 million English words. The corpus was used to solve the issues of broken plural, which generally are not handled by stemmers.

**Table 1. Translation/Transliteration examples fro AFP**

| Words       | English       | Occurrences. |
|-------------|---------------|--------------|
| لوس انجليس  | Los Angeles   | 21           |
| لوس انجلوس  | Los Angeles   | 23           |
| لوس انجيلس  | Los Angeles   | 2            |
| لوس انجيليس | Los Angeles   | 34           |
| كارولاينا   | Carolina      | 26           |
| كارولينا    | Carolina      | 14           |
| ويسكونسين   | Wisconsin     | 8            |
| ويسكنسن     | Wisconsin     | 2            |
| ويسكونسن    | Wisconsin     | 16           |
| نيو هامبشير | New Hampshire | 15           |
| نيو هامبشر  | New Hampshire | 9            |
| جوهانس      | Johannes      | 4            |
| يوهانس      | Johannes      | 74           |
| يوهانيس     | Johannes      | 8            |
| جوهانز      | Johannes      | 1            |
| جوهانسبورغ  | Johannesburg  | 173          |
| جوهانسبرغ   | Johannesburg  | 15           |
| جوهانسبورغ  | Johannesburg  | 1            |
| جوهانسورغ   | Johannesburg  | 1            |
| فايمار      | Weimar        | 3            |
| فيمار       | Weimar        | 10           |

Besides the cost, most of the existing Arabic corpora are small, specialized or very regional. Abdelali et al. [3] discussed issues related to AFP corpus used in TREC conferences; although the size of the corpus is significant. There were many reservations about its representativeness for MSA. Issues about language style, structures and inconsistency in translation/transliteration were highlighted by examples from the corpus in Table 1 and 2, similarly, Xu, Fraser, and Weischedel [18] used additional articles from Al-Hayat and An-Nahar newspapers to get terms for automatic query expansion in addition to terms from the AFP corpus. A newer version of the corpus with additions from Al Hayat Newspaper (1994-2001), Nahar Newspaper (1995-2002) and Xinhua News Agency (2001-2003) was released from LDC

as a new corpus called “Arabic Gigaword”, very little is known about this collection.

**Table 2. Excerpts from AFP News collection**

|   |
|---|
| 1. لكن الانفجار احدث فجوة كبيرة في الحافلة وتشاهد برك من الدم في المكان.  |
| 2. المتمردون يشيرون الى معارك عنيفة في بويتو جنوب شرق جمهورية الكونغو   |
| 3. وقام المحققون البريطانيون مساء باستجواب الموقوفين الثلاثة يساعدهم عملاء من اجهزة الاستخبارات الاميركية   |
| 4. ان شخصا قتل واصيب اربعة آخرون بجروح امس الاحد في بوروندي في هجوم بالقتال اليدوية والسلاح الرشاشة استهدف قافلة تابعة لهذه المنظمة الفرنسية.   |
| 5. وتمكنا من العودة ليل الخميس الجمعة الى فريتاون، كما قال احدهما لوكالة فرانس برس.   |
| 6. ووضح الصحافي السيراليوني كريستو جونسون الذي يعمل لوكالة رويتر البريطانية انه افرج عنه مع عضو في بعثة الامم المتحدة في سيراليون (يونومسيل) المعنية بمشكلات مرتبطة باحترام حقوق الانسان. |

In our experiments, we aimed to develop an Arabic corpus or several Arabic corpora that would help in the study of Modern Standard Arabic and compare the language and styles used in different parts of the Arabic world. Evidences about the vocabulary and semantic variation could be noticed in words that are used in one region rather than others, or are used in different meanings, Tables 3 and 4 show examples of term usage to refer to the same subject. Tables 5 and 6 refer to names used in different regions for the same object or entity. Tables 7 and 8 show examples of words that carry different meanings in different regions. We intended for the corpus to be richer that captures all these features and representative than the existing corpora.

**Table 3. Example of “dormitory”**

| English Word | El-khabar Algeria | Addustur Jordan | Hayat London |
|--------------|-------------------|-----------------|--------------|
| Dormitory    | مرآقد             | عنبر            | إقامة        |

**Table 3. Example of “dormitory”**

| English Word | El-khabar Algeria | Addustur Jordan | Hayat London |
|--------------|-------------------|-----------------|--------------|
| Dormitory    | مرآقد             | عنبر            | إقامة        |

**Table 4. Examples of “arrest” and “tend to fall”**

| English Word | El-khabar Algeria | Al-anbaa Morocco |
|--------------|-------------------|------------------|
| Arrest       | توقيف             | حجز              |
| Tend to fall | معرضة للسقوط      | أيلة للسقوط      |

**Table 5. Example of naming differences “Ministry of Education”**

| Egypt, Saudi Arabia | Qatar, Kuwait, Bahrain, Jordan | Mauritania    |
|---------------------|--------------------------------|---------------|
| وزارة المعارف       | وزارة التربية والتعليم         | وزارة التهذيب |

**Table 6. Example of naming differences “Ministry of religious/Islamic affairs”**

|                      |                                 |
|----------------------|---------------------------------|
| Egypt, Algeria, UAE  | Kuwait, Qatar                   |
| وزارة الشؤون الدينية | وزارة الأوقاف والشؤون الإسلامية |

**Table 7. Example of usage differences for word “ملاحم”**

| Sense                 | Word  | Sentence  | Country |
|-----------------------|-------|---|---------|
| fierce battles; epics | ملاحم | ولكن افتقار الأدب العربي لهذا اللون من الشعر لا يعني عدم احتواءه على المعاني والمفردات كالبطولة والشجاعة والفخر والحماة التي تتصف بها الملاحم المعروفة            | Algeria |
| butcheries            | ملاحم | كما ينفذ قسم المراقبة الصحية حملات تفتيش واسعة النطاق على محلات الاستهلاك الأدمى المتمثلة في بقالات بيع المواد الغذائية والملاحم والمخابز والمطاعم ومحلات الحلاقة | Oman    |

**Table 8. Example of usage differences for word “دوار”**

| Sense                 | Word | Sentence  | Country |
|-----------------------|------|---|---------|
| rotating, turning     | دوار | دوار برج الصخرة من المعالم والمواقع الهامة والحيوية حيث يربط معظم ولايات السلطنة بمحافظة مسقط         | Oman    |
| vertigo               | دوار | كذلك الدوار، الناتج عن أمراض عصبية، كالإصابات المركزية في الدماغ والمخيخ                              | Syria   |
| Bedouin camp, village | دوار | وعلى بعد خطوات من مسكن أبويه اقترب منه شخصان، اعتقد في البداية أن الأمر يتعلق باثنين من أبناء الدوار، | Morocco |

For such purpose, we mined text from newspapers and news services from different Arab speaking countries. We encountered some difficulties in getting common newspapers in parts of the region: either they were not available in electronic format or if they were available, we could not obtain the text in an appropriate format to analyze. Quite a few websites publish their content in PDF files, from which Arabic text cannot be easily reconstituted. In these cases, we had to replace the most common newspapers or news sources in an area with other less common, which were at least available in reasonable quantity. Table 9 shows the countries from which we collected newspapers.

**Table 9. List of news sources and countries of origin**

| Source     | URL                    | Country      |
|------------|------------------------|--------------|
| Ahram      | www.ahram.org.eg       | Egypt        |
| Alraialaam | www.alraialaam.com     | Kuwait       |
| Alwatan    | www.alwatan.com        | Oman         |
| Aps        | www.aps.dz             | Algeria      |
| Assafir    | www.assafir.com        | Lebanon      |
| Jazirah    | www.al-jazirah.com     | Saudi Arabia |
| Morocco    | www.morocco-today.info | Morocco      |
| Petra      | www.petra.gov.jo       | Jordan       |
| Raya       | www.raya.com           | Qatar        |
| Teshreen   | www.teshreen.com       | Syria        |
| Uruklink   | www.ruklink.net        | Iraq         |

We did not consider the number of readers, or the popularity of the selected papers selected. Our choices were mainly governed by the considerations of availability already mentioned. This indeed must affect the analysis and conclusions, but we considered that for this preliminary study we could establish some initial results from this small survey, with an eye on improving this analysis with a larger and more representative corpus.

**Table 10. Files collected by Source**

| Newspaper  | Number of files | Size (Kb) |
|------------|-----------------|-----------|
| Ahram      | 1567            | 10348     |
| Alraialaam | 390             | 15784     |
| Alwatan    | 10932           | 141636    |
| Aps        | 7408            | 68508     |
| Assafir    | 13914           | 77290     |
| Jazirah    | 3723            | 28296     |
| Morocco    | 17196           | 165266    |
| Petra      | 3567            | 20960     |
| Raya       | 270             | 7740      |
| Teshreen   | 33703           | 403228    |
| Uruklink   | 9464            | 129688    |

### 3. Building the Corpus

We used a locally developed spider program to get the data from each site. The spider was initialized with one of the main

links in the top hierarchy of the site along with the level of depth to which it should collect document from. The spider will traverse the links and save the pages linked to the main page in a top-down fashion until it reaches the depth specified. The spider runs every morning, (basically evening in the Arab world), which avoids peak traffic time, when people will be reading the newspaper, and also avoid creating problems that could be caused to the server by successive hits from the spider robot. We kept the spider running for a period of more than 3 months in the year of 2002 and collected 107 days of daily issues. Details about the size/number of files per newspaper are shown in the Table 2.

#### 4. Collection processing

Following the mining of the data, it was prepared for processing.

The steps included filtering the data by stripping the HTML tags and extracting the raw text in the page; then tagging the data collected with appropriate tags for referencing the source and other information. Mostly Arabic web pages use Windows-1256 – cp1256- as the codeset for the pages, few other use other encodings such as ISO-9959-6 or UTF8. Therefore; the next step was to convert the data to a common encoding usable by the analysis tools.

We used URSA, a tool developed at CRL. URSA, Unicode Retrieval System Architecture, is a high-performance text retrieval system that can index and retrieve Unicode texts. URSA has the capacity to index and retrieve documents in UNICODE and provide word frequencies and other data. URSA also has a comprehensive set of query and document weighing functions commonly used for information retrieval. The complete suite of weighting and ranking functions implemented in URSA represents the bulk of the weighting schemes developed in the past 40 years of text retrieval research and includes many of the recent successful document weighting schemes from Cornell and City University of New York. Further, by using a posting compression scheme that is both simple enough to allow for the efficient merging of posting data as well as for its rapid decompression and yet is specifically tuned to the kinds of data in the postings, URSA indexes are only about 12%-25% larger than the original texts. Finally, the URSA tools are robust enough to be used in industrial grade applications and are based on a very simple object oriented API [2,15].

Before indexing the data, we reviewed all the data to check for specific formats that were added for general formatting of the text, such as the link character kasheeda (known also as taweel), which may be added for cosmetic purpose and has no effect on the text, for example, “صاحب السمو” “الأحداث” “مدة” which are same as “صاحب السمو” “الأحداث” “مدة” respectively.

We also considered removing all the diacritics because Modern Standard Arabic is generally written without diacritics, though in very rare cases people may use them in this type of media primarily for clarification purposes. Contrary to previous experiments [6,9,10], we kept the text close to its original format other than the previous mentioned changes and we did not apply any further processing of what is called Normalization which usually include:

- Replacing initial  $\text{إ}$  or  $\text{أ}$  with bare alif  $\text{ا}$ .
- Replacing  $\text{آ}$  with  $\text{ا}$
- Replacing the sequence  $\text{ءى}$  with  $\text{ئ}$
- Replacing final  $\text{ى}$  with  $\text{ي}$
- Replacing final  $\text{ة}$  with  $\text{ه}$

We believe that some of these normalizations will hide a lot of features and create more ambiguity knowing that replacing initial  $\text{إ}$  or  $\text{أ}$  with bare alif  $\text{ا}$  means  $\text{ان}$  could be  $\text{أَن}$ ,  $\text{إِن}$ ,  $\text{أَنَّ}$ ,  $\text{إِنَّ}$  or  $\text{أَنْ}$ . The same normalization could hide local variants of the same word as the case for the word “انترنت”. Usually in the Middle East they use “انترنت” in contrast to North Africa where they use “أنترنت” bearing in mind that there are reasons behind this; in the Middle east they use a transliteration of the word “Internet” from English versus in North Africa where the transliteration of the French word for Internet is used [1].

#### 5. Corpus Assessments

A corpus by itself can do nothing at all; being nothing other than a store of used language [12]. Corpus access software can rearrange that store so that observations of various kinds can be made. Using available tools we first experimented by applying some statistical and probability tests, such as Zipf’s law and the Mandelbrot formula. These tests are useful for describing the frequency distribution of the words in the corpus. Also they are well-known tests for gauging data sparseness and providing evidence of any imbalance of the dataset.

According to Zipf’s law, if we count up how often each word occurs in a corpus and then list these words in the order of their frequency of occurrence, then the relationship between the frequency of a given word  $f$  and its position in the list (its rank  $r$ ) will be a constant  $k$  such that:

$$f \cdot r = k \quad (1)$$

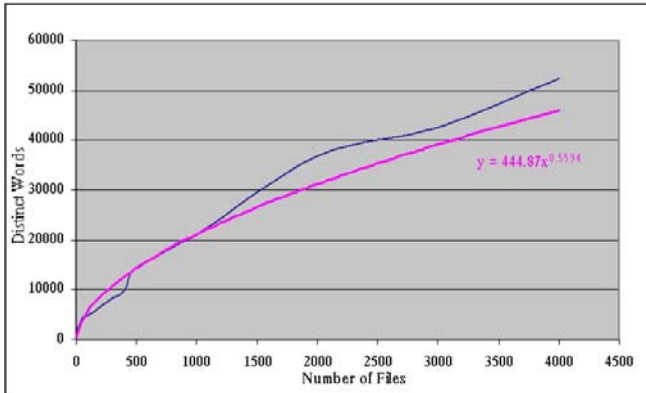
Ideally, a simple graph for the above equation will show a straight line with a slope  $-1$ . So we checked the situation in our corpus by starting with one file and increasingly adding more files to a corpus and checking the behavior of the relation between the rank and the frequency. An enhanced theory of the Zipf’s law is the Mandelbrot distribution; Mandelbrot notes that “although Zipf’s formula gives the general shape of the curves, it is very bad in reflecting the details” [12]. So to achieve a closer fit to the empirical distribution of words, Mandelbrot derived the following formula for relation between the frequency and the rank:

$$f = P(r+\rho)^{-B}$$

Where  $P$ ,  $B$ , and  $\rho$  are parameters of the text that collectively measure the richness of the text’s use of words. The common factor is that there is still a hyperbolic relation between the rank and the frequency as in the original equation of Zipf’s law. If this formula is graphed on doubly logarithmic axes, it closely approximates a straight line descending with a slope  $-B$  just as Zipf’s law, (See the appendix for Figures).

#### 6. Analysis

We began assessing the data collected after the pre-processing to have the data in a usable format that the tools could work on. First we started with one dataset and checked the contribution of every document to the corpus construction, by checking the number of distinct words added by every document and interpolating that in a function that simulate the behavior of this process. Table 4 and Figure 1 show details about this.



**Figure 1. Presentation of the contribution of the document to the corpus case of Aps.**

What we conclude from the function simulate the corpus construction that the speed and the amount of data added after a certain size of the corpus will not contribute significantly to the nature of the corpus itself.

**Table 11. Contribution of the document to the corpus case of Aps.**

| Number of files | Number of words | Number of distinct words |
|-----------------|-----------------|--------------------------|
| 1               | 569             | 338                      |
| 5               | 3131            | 1236                     |
| 10              | 6287            | 2151                     |
| 25              | 15754           | 3113                     |
| 50              | 30606           | 4405                     |
| 100             | 58423           | 5034                     |
| 500             | 286950          | 14553                    |
| 1000            | 571665          | 21050                    |
| 2000            | 1140573         | 36810                    |

For the analysis purpose, we used the function got from the interpolation of Table 11

$$y = k .x^{-a} \quad (\text{eq.1})$$

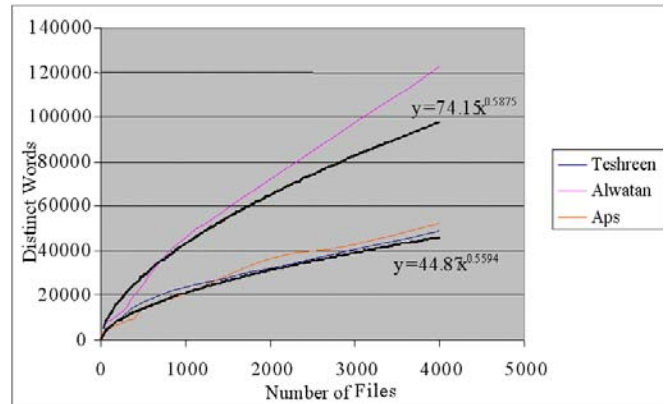
For this function the acceleration will be the second derivative of the function so,

$$y' = -k .a.x^{-a-1} \quad (\text{eq.2})$$

then  $y'' = -k .a.(-a-1).x^{-a-2}$   
 which is  $y'' = k .a.(a+1).x^{-a-2} \quad (\text{eq.3})$

Figure 1 shows a representation of the equation (eq.3) for values of  $a = 0.5594$  and  $k = 444.87$

In the next step, we confirmed the balance and the completeness of every dataset. We considered every newspaper as its own dataset. We had to analyze every set by itself, which helped us to acquire information and details about every single newspaper.



**Figure 2. Presentation of the contribution of the document to the corpus case of Aps, Alwatan, Teshreen.**

For completeness we applied the Zipf's law to check the behavior of every dataset. As we expected, graphs improved as the size of data increased, and the proportion of rare words declined.

These results reflect well the representativeness of the dataset, especially for the case of collections such as Jazirah, Raialaam, Alwatan, and Assafir. In only one case were the results unsatisfactory: in the case of the Moroccan newspaper Morocco-Today, the ratio of words to token (distinct words) is very low compared to the other sources. After doing some investigation, we found that the reasons was that the web site wasn't getting updated daily, only a few pages were updated, while the rest of the data was posted over and over for several days. This caused the frequency of the words to be skewed by adding the same document more than once.

For the case of the Uruklink as shown in Table 12, the ratio of distinct word to the total number is very low. Understanding the problem is still vague. The only thing that could clarify our understanding of this issue is when we index the whole collection, which contains more than twice of the files we have indexed so far.

As a result, from Table 12, which presents a summary of the collection, for number of this datasets, there is no reason to believe that the datasets are imbalanced; see figure 3.

Except for the Moroccan dataset and the Iraqi one which, we believe to be replaced either by collecting more data or looking for an alternative source from the same area, the rest of the datasets we believe are a real complete representative corpus for the area and that a serious study on these corpora would bring and reveals very important information about this corpus and the Arabic language in general.

**Table 12. Number of words per collection**

| Source     | Number of Files | (T) Total Words | (D) Distinct Words | Ratio D/T% |
|------------|-----------------|-----------------|--------------------|------------|
| Ahram      | 1567            | 455,366         | 16,569             | 3.639      |
| Alraialaam | 390             | 1,160,203       | 97,580             | 8.411      |
| Alwatan    | 4000            | 4,714,199       | 122,467            | 2.598      |
| Aps        | 4000            | 2,512,426       | 52,481             | 2.089      |
| Assafir    | 4000            | 3,448,639       | 121,911            | 3.535      |
| Jazirah    | 3723            | 1,405,083       | 84,638             | 6.024      |
| Morocco    | 4000            | 3,306,137       | 19,092             | 0.577      |
| Petra      | 3567            | 989,140         | 45,896             | 4.640      |
| Raya       | 270             | 612,409         | 55,868             | 9.123      |
| Teshreen   | 4000            | 1,467,368       | 49,067             | 3.344      |
| Uruklink   | 4000            | 2,378,499       | 32,899             | 1.383      |

## 7. Conclusions and Future Work

Arabic data available on the net is a suitable resource for building a significant corpus for the purpose of studying the language. The approach of mining data from online Arabic newspapers and news sources as resources for corpus use will be a boost for improving different researches in Information Retrieval, Machine Translation and Arabic Language processing in general, with an authentic quality and adequate quantity. The major issue with the approach is copyright which could prevent wide exploitation of the resources.

Although the corpus was tested for completeness and representativeness, expanding the collection for more newspapers and including other types of literature (i.e. official transcripts, novels, .etc ) will improve the quality and give more confidence for the results induced from it. We are investigating the collection to learn more about vocabulary differences in Modern Standard Arabic used in the different geographical regions.

## 8. References

- [1] Abdelali, A. (2004) Localization in Modern Standard Arabic. Journal of the American Society for Information Science and technology (JASIST), Volume 55, Number 1, 2004. pp. 23-28.
- [2] Abdelali, A. Cowie, J. Farwell, D. Ogden, W., (2002) UCLIR: a Multilingual Information Retrieval tool VIII Iberoamerican Conference on Artificial Intelligence, Sevilla (Spain), November 2002.
- [3] Abdelali, A. Cowie, J. Soliman S. H. (2004) Arabic Information Retrieval Perspectives. Proceedings of JEP-TALN 2004 Arabic Language Processing, Fez 19-22. April 2004.
- [4] Al-Kharashi, I. A. and Evans, M. W. (1994) Comparing words, stems, and roots as index terms in an Arabic information retrieval system. Journal of the American Society for Information Science (JASIS) 45(8), pp 548-560.
- [5] Darwish K, Doermann D, Jones R, Oard D & Rautiainen M (2001) TREC-10 experiments at University of Maryland CLIR and video. Text RE-trieval Conference TREC10 Proceedings, Gaithersburg, MD, pp 549-562.
- [6] Goweder, A. and De Roeck, A. (2001) Assessment of a significant Arabic corpus. Presented at the Arabic NLP Workshop at ACL/EACL 2001, Toulouse, France, 2001.
- [7] Hmeidi, I., Kanaan, G. and M. Evens (1997) Design and Implementation of Automatic Indexing for Information Retrieval with Arabic Documents. Journal of the American Society for Information Science, 48/10, pp. 867-881.
- [8] Hunston, S. Corpora in applied linguistics Cambridge University Press May 2002.
- [9] Larkey, L. S. and Connell, M. (2002) Arabic Information Retrieval at UMass in TREC-10 In Voorhees, E.M. & Harman, D.K. (Eds.) The Tenth Text Retrieval Conference, TREC 2001 NIST Special Publication 500-250, pp. 562-570.
- [10] Larkey, L. S., Ballesteros, L., and Connell, M. (2002) Improving Stemming for Arabic Information Retrieval, Proceedings of SIGIR 2002, pp. 275-282
- [11] Madar Research - In Focust Article <http://www.madarresearch.com/news/newsdetail.aspx?nwsId=6> Retrieved Sept 22, 2004
- [12] Manning, C., Schütze, H. Foundations of Statistical Natural Language Processing, MIT Press. Cambridge, MA. May 1999. ISBN 0-262-133600-1
- [13] Meyer, C. F. English corpus linguistics: an introduction Cambridge University Press July 2002.
- [14] Moreh, S. Studies in Modern Arabic Prose and Poetry, Leiden, E.J. Brill, 1988.
- [15] Ogden, W. Cowie, J. Davis, M. Ludovik, E. Nirenburg, S. Molina-Salgado, H. and Sharples, N. (1999) Keizai: An Interactive Cross-Language Text Retrieval System. Paper presented at the Workshop on Machine Translation for Cross-language Information Retrieval, Machine Translation Summit VII, September 13-17, 1999, Singapore.
- [16] Stetkevych, Jaroslav The Modern Arabic Literary Language Lexical and Stylistic Developments University of Chicago 1970.
- [17] Worldwide Internet Population [www.commerce.net/other/research/stats/wwstats.html](http://www.commerce.net/other/research/stats/wwstats.html) Retrieved Sept 14, 2002.
- [18] Xu, J. Fraser, A. Weischedel M. R. (2001) TREC 2001 Cross-lingual Retrieval at BBN NIST Text RE-trieval Conference TREC10 Proceedings, Gaithersburg, MD, pp. 68-77.

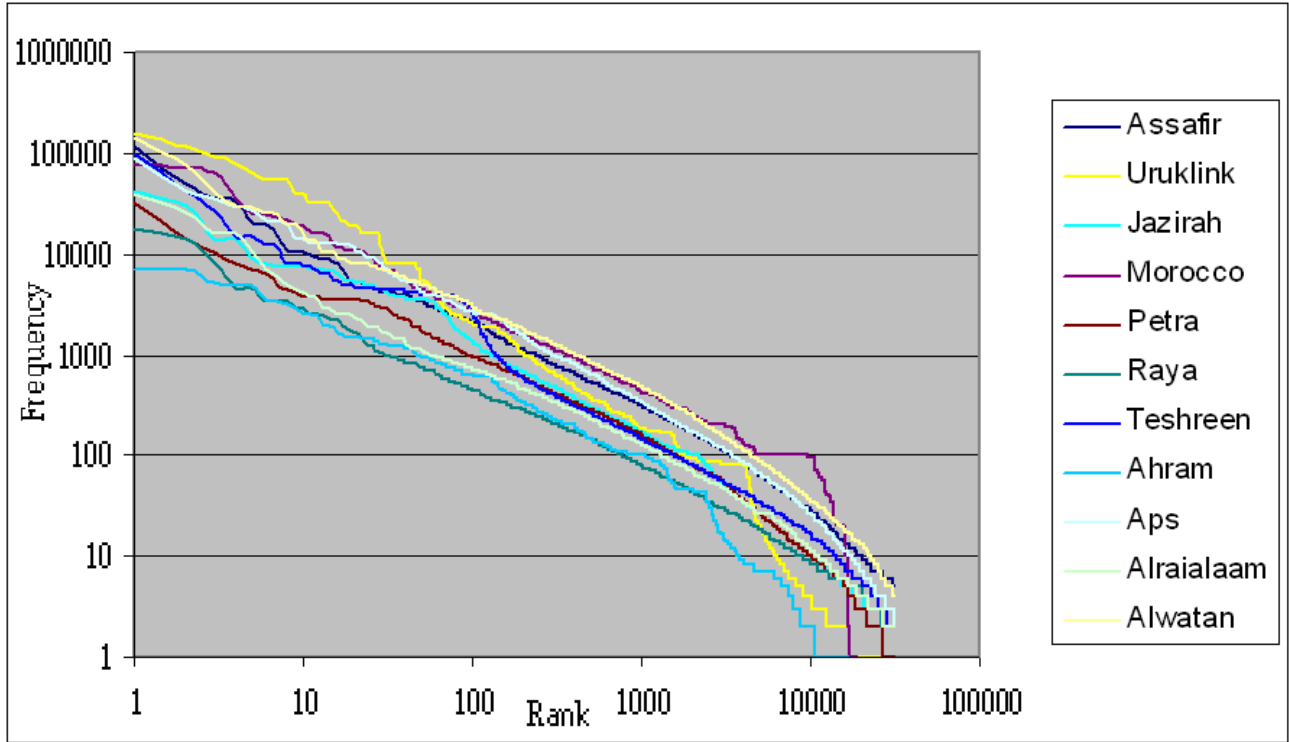


Figure 3. Word frequency versus rank in 4000 documents from the list of the newspapers