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

James 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 in recent years more interest in the Arabic language. 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 is key for any objective research.

In this paper we present the results of experiments in building 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 representatives of this corpus using standard metrics and show its suitability for Language Engineering experiments.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing – *Indexing methods, Linguistic processing.*

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *Information filtering.*

General Terms

Measurement, Experimentation, Languages.

Keywords

Modern Standard Arabic, Corpus, Zipf's Law.

1. INTRODUCTION

The amount of Arabic data available on the World Wide Web is dramatically increasing daily. According to DITnet (a Middle East Internet Technology site), there were 12 million Arab Internet users by the end of 2002. 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 [11]. Providing users with quality web portals and efficient search engines is essential to keep up with the growth. The issued become even more serious when it comes to finding specific information on the net.

One other 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 [8,9]. 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 a very laborious nature. But this was done for 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 [4,7].The importance of corpora to linguistic study can be 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 this type of tools were initiated by researchers in the NLP field. Hmeidi el al in 1997 constructed a corpus of 242 abstracts collected from the proceedings of the Saudi Arabian national conference. (Goweder A and De Roeck A.) [3] produced an Arabic corpus using 42591 articles from Al-Hayat newspaper archive of the year 1998. The last experiment was mainly to reproduce and confirm results made

Conference '00, Month 1-2, 2000, City, State.

on small-scale corpus about the sparseness of the Arabic comparing to English.

In our experiments, we aimed to develop an Arabic corpus or several Arabic corpora that would help in the study of Modern Standard Arabic. We gathered samples of newspapers from different countries for the purpose of comparing the language used in different parts of the Arabic world. 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 couldn't get the text in an appropriate format to analyze. In these cases, we had to replace the most common newspapers in an area with other less common or widely read ones, which were at least available in reasonable quantity.

The table 1 shows the countries from which we collected newspapers.

Table 1. List of newspapers and countries of origin

Newspaper	URL	Country
Ahram	www.ahram.org.eg	Eygpt
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
Aorocco	www.morocco-today.info	Morocco
Petra	www.petra.gov.jo	Jordan
Raya	www.raya.com	Qatar
Teshreen	www.teshreen.com	Syria
Uruklink	www.uruklink.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 towards improving this analysis with a larger and more representative corpus.

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 go. 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 any problem that could be caused to the server by successive hits from the spider. We kept the spider running for a period of more than 3 months and collected 107 days of daily issues. Details about the size/number of files per newspaper are shown in the table 2.

Table 2. Files collected by newspaper

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

4. Preliminary processing

Following the collection and spidering of the data, it was prepare for processing.

The steps included filtering and tagging the data collected, mainly HTML pages coded in different Arabic encodings -- see Table 3. 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 weighting 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 [2,10].

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 (Arabic taweel), which may be added for cosmetic purpose and has no effect on the text, for example, “صـاحـب السـمو” “الأحـداث” “مـدة” which are same as “صاحب السمو” “الأحداث” “مدة” respectively.

Table 3. Text encoding for the collected newspapers

Newspaper	Web URL	Encoding
Ahram	www.ahram.org.eg	cp1256
Alraialaam	www.alraialaam.com	cp1256
Alwatan	www.alwatan.com	cp1256
Aps	www.aps.dz	cp1256
Assafir	www.assafir.com	cp1256/iso8859-6
Jazirah	www.al-jazirah.com	cp1256
Morocco	www.morocco-today.info	cp1256
Petra	www.petra.gov.jo	cp1256
Raya	www.raya.com	cp1256
Teshreen	www.teshreen.com	cp1256
Uruklink	www.uruklink.net	cp1256

We considered as well 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 [3,5], we kept the text close of its original format other than the previous mentioned changes.

5. Corpus Assessments

A corpus by itself can do nothing at all; being nothing other than a store of used language [6]. Corpus access software can re-arrange 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:

$$f.r = k \qquad -(1)$$

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”. 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:

$$f = P(r+r)-B$$

Where P, B, and r 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 begin 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 we 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. The table 4 and the Figure 1 show details about this. 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.

For the analysis purpose, we used the function got from the interpolation of the table 4

$$y = k .x –a \qquad (eq .1)$$

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

$$y’ = -k .a.x –a-1 \qquad (eq .2)$$

than $$y’’ = - k .a.(-a-1).x –a-2$$

which is $$y’’ = k .a.(a+1).x –a-2 \ (eq .3)$$

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

The next step, we confirm the balance and the completeness of every dataset. We are considering every newspaper as its own dataset. We had to analyze every set by its self, which helped us to acquire information and details about every single newspaper.

For the 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.

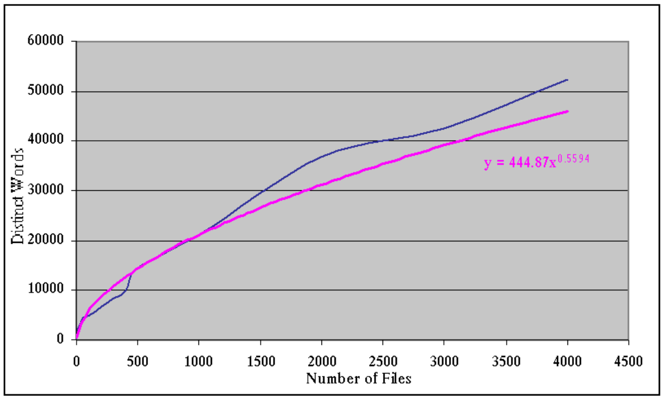


Figure 1. Presentation of the contribution of the document to the corpus case of Aps

Table 4. Contribution of the document to the corpus case of Aps

Number of files	Number of words	Number of distinct words
1	569	338
2	1215	596
3	1936	895
4	2545	1126
5	3131	1236
10	6287	2151
25	15754	3113
50	30606	4405
100	58423	5034
200	118045	6724
300	177438	8444
400	236612	9736
500	286950	14553
1000	571665	21050
2000	1140573	36810

These results reflect well for the representativeness of the dataset, especially for the case of collections such as Aljazirah, Raialaam, Alwatan, Assafir. Only in one case are the results unsatisfactory: the case of the Moroccan newspaper morocco-today, the ratio of the sparseness is very low compared to the other newspapers. After doing some investigation, we found that the reasons behind 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 that the frequency of the words to be skewed by adding the same document more than once. But for the limitation of the program we are using, we couldn't index more than 4000 files, although the collection contains 17196 (See table 2). Before making a final decision we will try to check the rest of the files and index the while collection.

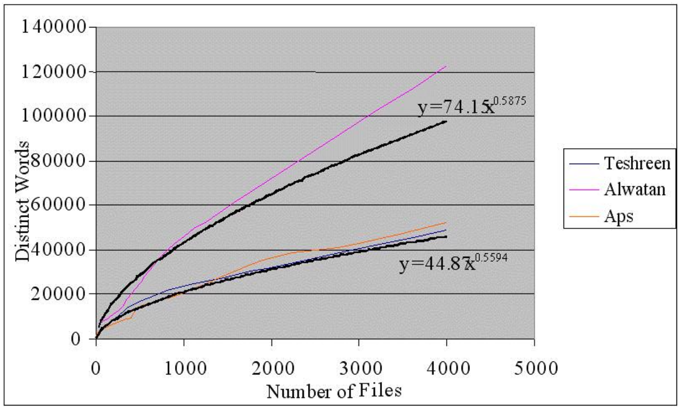


Figure 2. Presentation of the contribution of the document to the corpus case of Aps, Alwatan, Teshreen

Table 5. Number of words per collection

Newspaper	Number of Files	Total Words	Distinct Words	Ratio
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

For the case of the Uruklink as we can read from the table 5, the ratio of the distinct word to the total number is very low. Understanding the problem still was vague. The only point that could enlighten the issue is when we index the whole collection, which contains more than twice of the files we have indexed so far.

As result, from the Table 5, which presents a summary of the collection, for number of this datasets, there is no reason to believe that the datasets are imbalanced; the figures in the appendix support this as well.

Except the Moroccan dataset and the Iraqi which, are still under investigation and analysis, the rest of the datasets we believe that they are a real complete representative corpus for the area and that a serious study on these corpuses would bring and reveals very important information about this corpus and the Arabic language in general.

Table 6,7. Ratio of sparseness per newspaper

Newspaper	Sparseness
Ahram	55.10
Alraialaam	71.29
Alwatan	53.95
Aps	52.77
Assafir	49.87
Jazirah	73.06
Morocco	13.00
Petra	63.20
Raya	63.51
Teshreen	53.02
Uruklink	72.71

Newspaper	Sparseness
Jazirah	73.06
Uruklink	72.71
Alraialaam	71.29
Raya	63.51
Petra	63.20
Ahram	55.10
Alwatan	53.95
Tteshreen	53.02
Aps	52.77
Assafir	49.87
Morocco	13.00

7. Conclusions

Arabic data available on the net is a suitable candidate for building a significant corpus for purpose of studying the language. The approach of considering the online Arabic newspapers as a resource for data will be a boost for improving different researchers in Information Retrieval, Machine Translation and Arabic Language processing in general.

8. References

[1] Abdelali, A. Localization in Modern Standard Arabic. Accepted to be published on Journal of the American Society for Information Science and technology (JASIST).

[2] Abdelali, A. Cowie, J. Farwell, D. Ogden, W., UCLIR: a Multilingual Information Retrieval tool VIII Iberoamerican Conference on Artificial Intelligence, Sevilla (Spain), November 2002.

[3] Goweder, A. and De Roeck, A. Assessment of a significant Arabic corpus. Presented at the Arabic NLP Workshop at ACL/EACL 2001, Toulouse, France, 2001.

[4] Hunston, S. Corpora in applied linguistics Cambridge University Press May 2002.

[5] Larkey, L. S., Ballesteros, L., and Connell, M. (2002) Improving Stemming for Arabic Information Retrieval, Proceedings of SIGIR 2002, pp. 275-282

[6] Manning, C., Schütze, H. Foundations of Statistical Natural Language Processing, MIT Press. Cambridge, MA: May 1999.

[7] Meyer, Charles F. English corpus linguistics: an introduction Cambridge University Press July 2002.

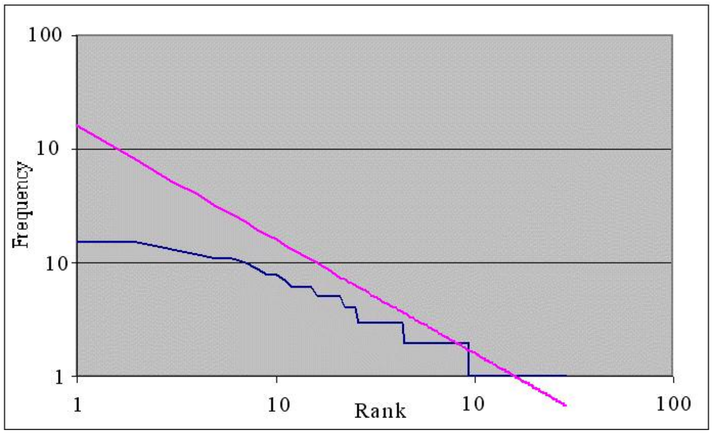
[8] Moreh, S. Studies in Modern Arabic Prose and Poetry, Leiden, E.J. Brill, 1988.

[9] Stetkevych, Jaroslav The Modern Arabic Literary Language Lexical and Stylistic Developments University of Chicago 1970.

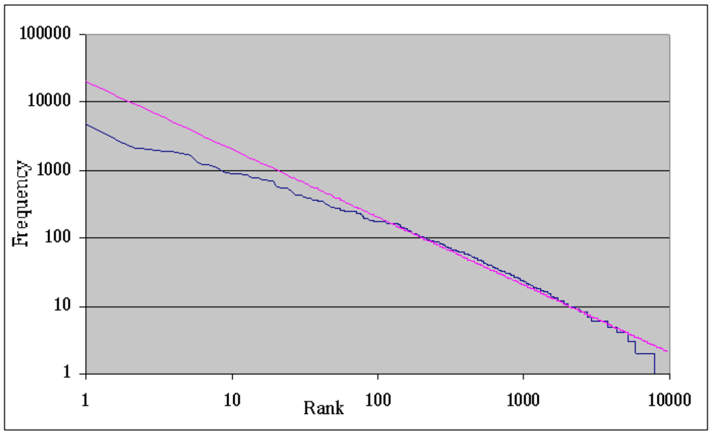
[10] Ogden, W. Cowie, J. Davis, M. Ludovik, E. Nirenburg, S. Molina-Salgado, H. and Sharples, N. (1999) Keizai: An Interactive Cross-Language Text Retrieval System. Paper presented at the Workshop on Machine Translation for Cross-language Information Retrieval, Machine Translation Summit VII, September 13-17, 1999, Singapore.

[11] Worldwide Internet Population www.commerce.net/other/research/stats/wwstats.html Retrieved Sept 11, 2002.

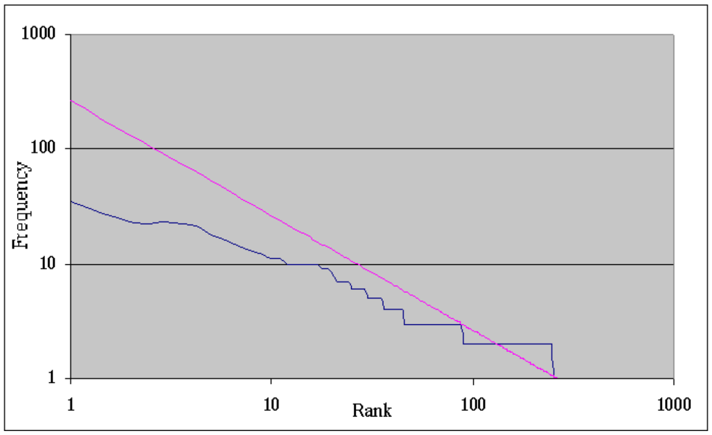
Appendix A.



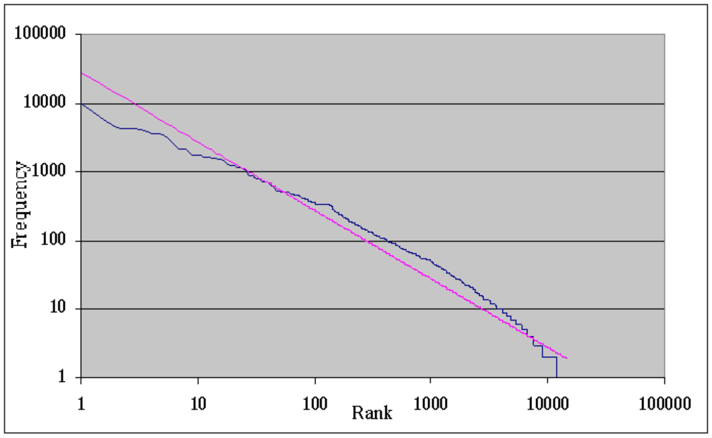
Graph 1. Word frequency versus rank in one document from Aps.



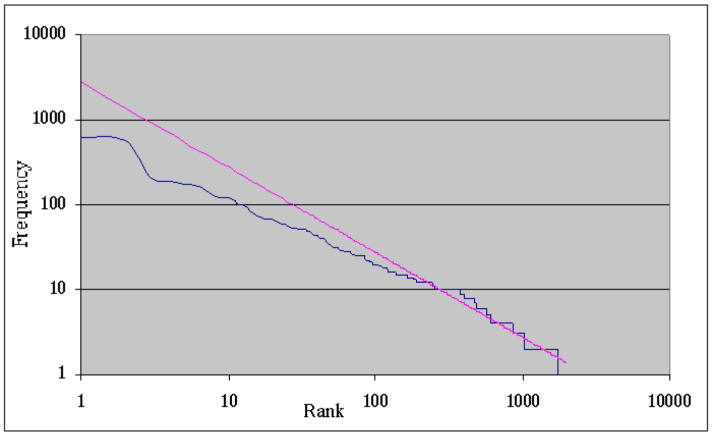
Graph 4. Word frequency versus rank in 250 documents from Aps



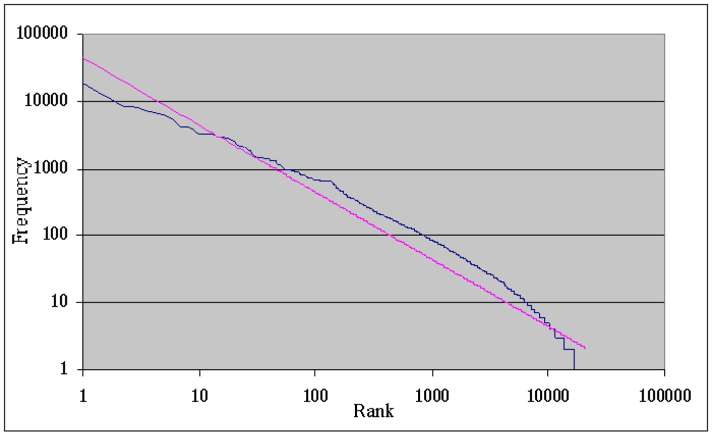
Graph 2. Word frequency versus rank in two document from Aps



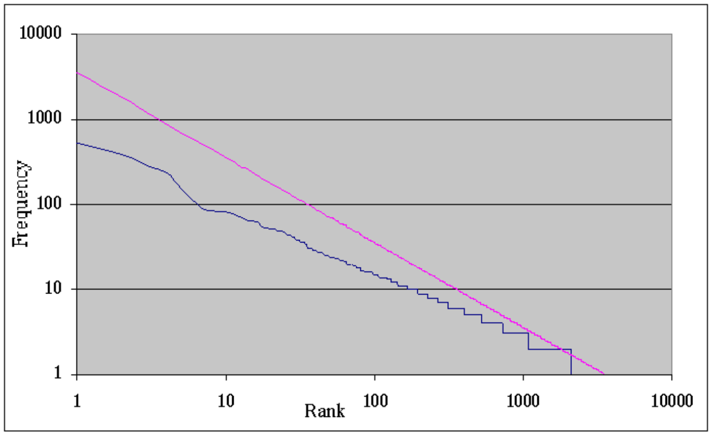
Graph 5. Word frequency versus rank in 500 documents from Aps



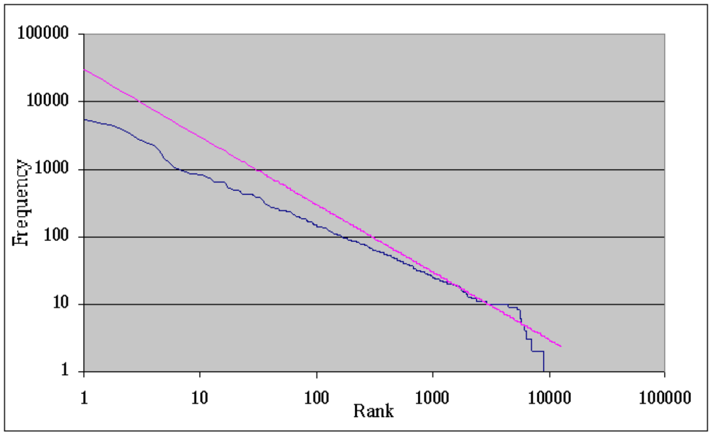
Graph 3. Word frequency versus rank in 25 documents from Aps



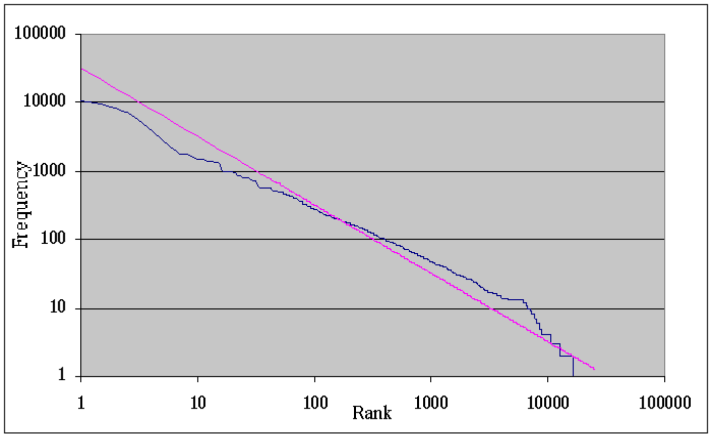
Graph 6. Word frequency versus rank in 1000 documents from Aps



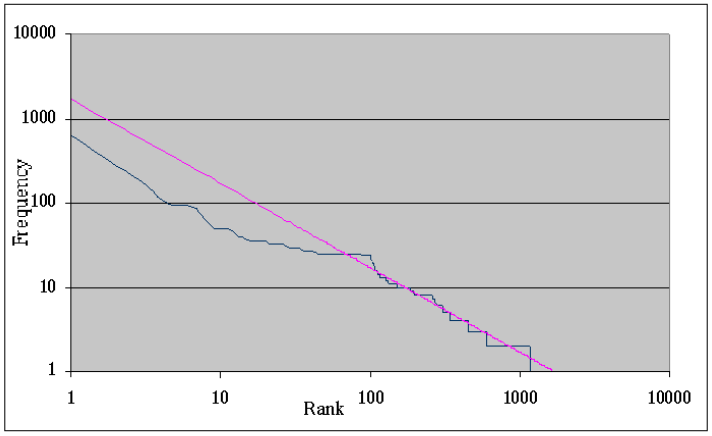
Graph 7. Word frequency versus rank in 25 documents from Alwatan



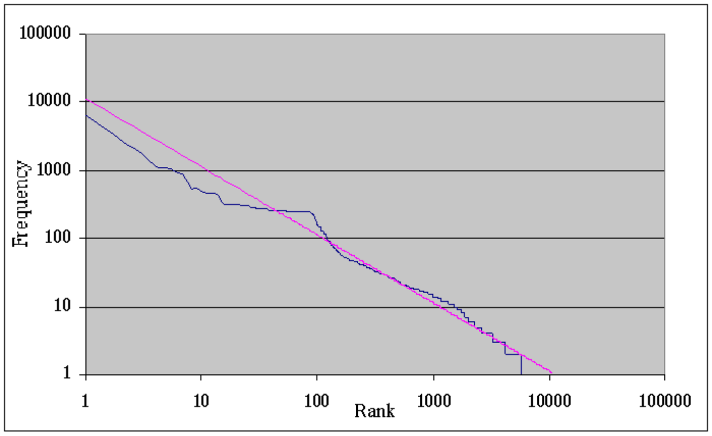
Graph 8. Word frequency versus rank in 250 documents from Alwatan



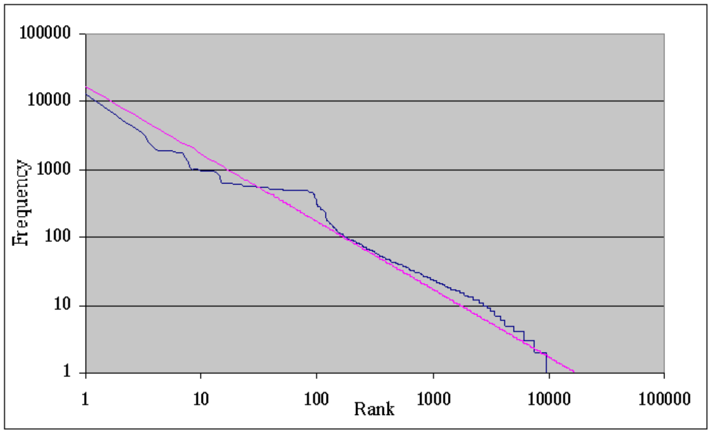
Graph 9. Word frequency versus rank in 500 documents from Alwatan



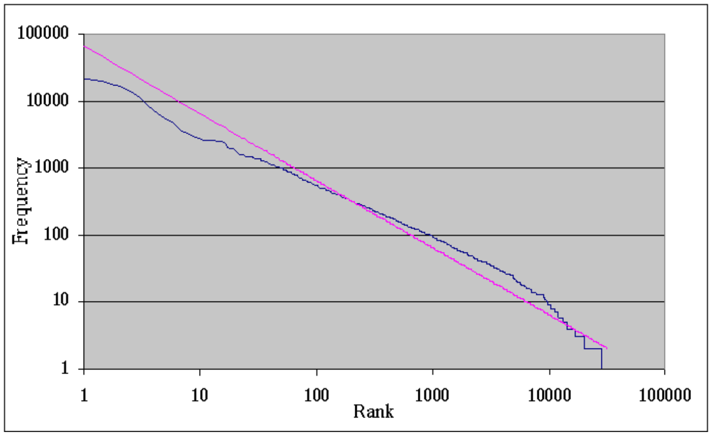
Graph 10. Word frequency versus rank in 25 documents from Teshreen



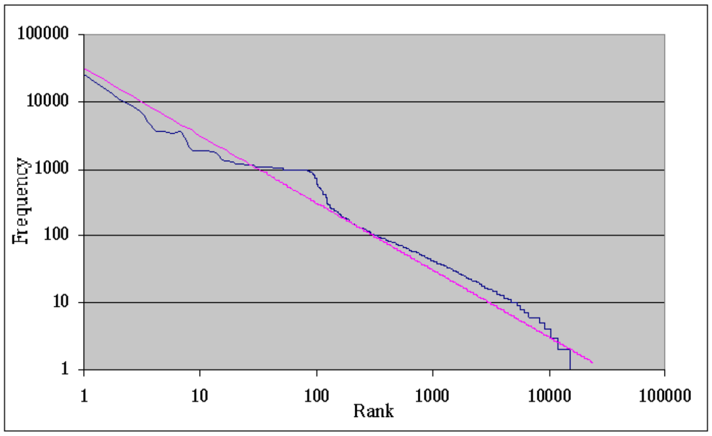
Graph 11. Word frequency versus rank in 250 documents from Teshreen



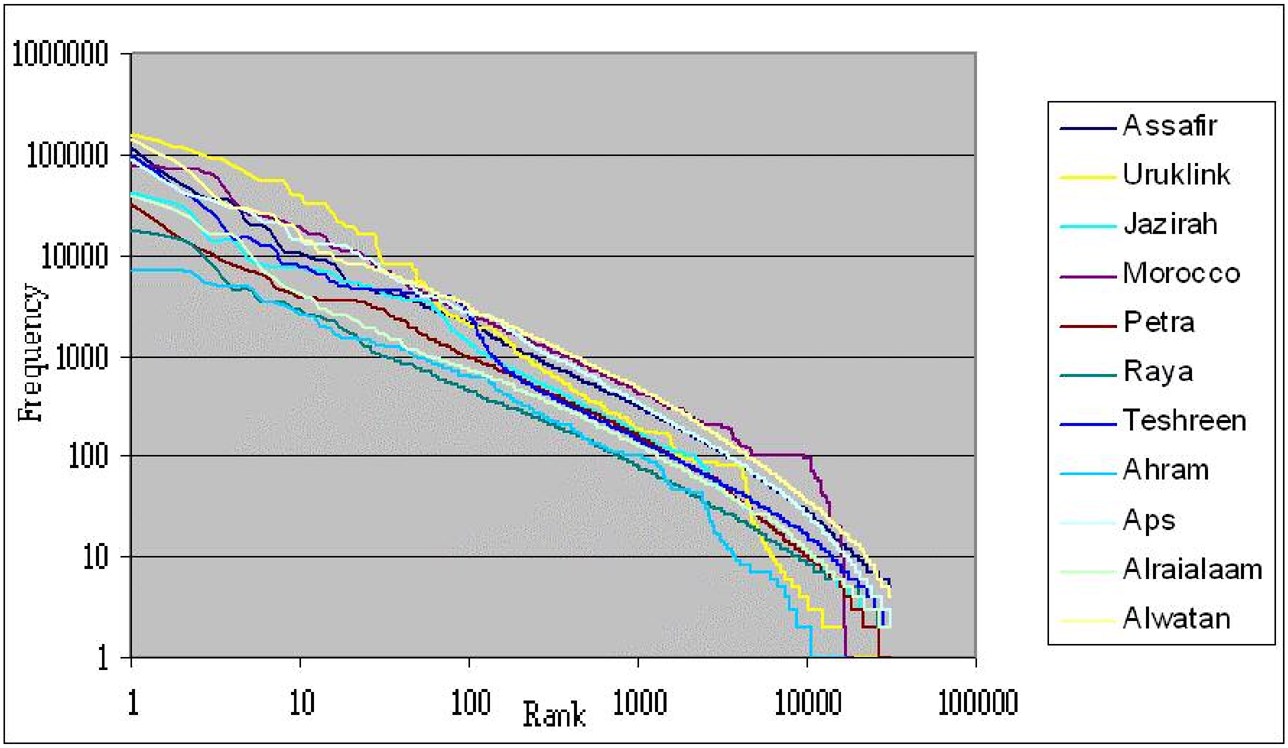
Graph 12. Word frequency versus rank in 500 documents from Teshreen



Graph 13. Word frequency versus rank in 1000 documents from Alwatan



Graph 14. Word frequency versus rank in 1000 documents from Teshreen



Graph 15. Word frequency versus rank in 4000 documents from the list of the newspapers