BUILDING A MULTILINGUAL AND MIXED ARABIC-ENGLISH CORPUS

Mohammed Mustafa, Hussein Suleman

Department of Computer Science, University of Cape Town , Rondebosch – 7700, Cape Town, Republic of South Africa
mohammed.mustafa@acm.org, hussein@cs.uct.ac.za

Keywords: Multilingual Query, Corpus, Zip’s Law, Heap’s Law

Abstract: Most currently available test collections and almost all CLIR collections have focused upon general-domain news stories. In addition, most of these corpora are built to help with retrieval of documents based on monolingual queries, even if they are translated. This paper presents the first-phase - building the corpus - of ongoing research to study the trends of multilinguality with special focus on Arabic/English multilingual texts in both queries and documents in scientific domains. The necessity of such a corpus would help a lot in providing good algorithms for Web searching of scholars in the Arabic World. The paper presented also the features of such corpus, how it is collected and how it has been validated in terms of terms frequencies, sparseness and vocabulary growth, using statistical tests. Results showed that the data is imbalanced at present

1. INTRODUCTION

Building text corpora is very common in Information Retrieval (IR), Natural Language Processing (NLP) and Computational Linguistics in order to support the ongoing research within these communities. However, most current available test collections and almost all Cross Language Information Retrieval (CLIR) collections have focused upon general-domain news stories, legal documents and encyclopedia articles. In addition, the majority of these collections is monolingual or, in the best case, consists of several monolingual documents in different languages, with each collection in a given language, rather than documents with different portions/terms that are tightly–integrated in multilingual forms.

In non-English languages, Arabic is an example where documents often are multilingual in scientific domains. A multilingual document is a mixed document that contains different languages. Multilinguality occurs in scientific non-English documents because most languages used in developing countries, including the Arabic-speaking world, suffer from a limited modern vocabulary and do not include up-to-date terminology. The typical Arabic speaker speaks a mixture of tightly-integrated words in both English and Arabic (and various slang variants) that will muddle most algorithms in IR. This mixed grammar is emerging in the Web.

However, the first phase of investigating trends of multilinguality in non-English languages is to gather a large corpus for experimentation. Thus, this paper presents the first-phase - building the corpus - of ongoing research to study the trends of multilinguality with special focus on Arabic/English multilingual texts in both queries and documents in scientific domains. The corpus contains mixed documents in both Arabic and English, monolingual English documents and monolingual Arabic documents. This corpus would serve as a representative sample of what actually occurs on the Web as well as being the test-bed for later experiments. This paper addresses the main features of the corpus and the statistical tests that have been applied during its construction phase.

2. RELATED WORK

Several corpora have been developed to serve as standard test collections. The first pioneering experiment to create such a test-bed was held in the late 1950s. The Cranfield corpus (Cleverdon, 1970) contains a few hundred abstracts collected from articles of aerodynamics journal.

However, current corpora can be classified in terms of: single language vs. multilingual; general vs. specialized (McEnery et. al, 2006); and synchronic vs. diachronic. For a given corpus, an overlap in this classification may occur, e.g., a given corpus may be monolingual and specialized.

In terms of their languages, current corpora can be categorized into two types: single language corpora and multilingual corpora (Lin and Chen, 2002). In the single language corpora, all documents are written in a single language. An example for a monolingual collection is the AFP (Agence France Presse)[1], which is an Arabic collection in the news genre, collected from articles from the AFP and created by the Linguistic Data Consortium (LDC).

In the second approach, which is multilingual, documents are usually written in several monolingual languages. Such types of multilingual corpora highlight language-specific, typological or cultural features. Parallel corpora, in which the same contents are translated into different languages, can be considered to be types of multilingual corpora. Multilingual corpora are the most dominant in the standards collections. The most widely known is the different editions of the Text Retrieval Conference (TREC) collections[2]. TREC is sponsored by NIST (the National Institute of Standards and Technology). It contains several monolingual corpora in different languages along with their queries and relevance judgments. Arabic has been included in TREC in 2001 in the crosslingual track.

NII Test Collection for IR Systems (NTCIR)[3] is a collection that contains languages in the Asian region (Chinese, Japanese and Korean) and their collections are of similar sizes to TREC. NTCIR focuses on CLIR.

The European Cross Language Evaluation Forum (CLEF)[4] is another valuable series of corpora, which is focused on European languages and CLIR.

However, most of these corpora are built to help with retrieval of documents based on monolingual queries, even if they are translated. Therefore, most documents are monolingual in several languages. However, some documents are multilingual, e.g. some Japanese documents. A multilingual document is written in two languages. Most such multilingual documents present the English translations for some words but not in a tightly-integrated manner.

In terms of vocabulary types, corpora can be classified as general corpora or specialized corpora. A general corpus, as the name indicates, usually contains different genres and domains such as regional and national newspapers, legal documents, encyclopedias and periodicals. In addition, general corpora may contain written or spoken data. CLEF, TREC and NTCIR can be considered to be general corpora because test documents in them are general domain news stories (Rogati and Yang, 2004).

In contrast, a specialized corpus contains terminology in a specific domain. Examples of specialized corpora include CACM (Dunlop and Rijsbergen, 1993), which was built from titles and abstracts of the Communications of the ACM from 1958-1979. Hmeidi et al. (1997) built an Arabic corpus with 242 documents gathered from the proceedings of the Saudi Arabian conference. NTCIR contains also some specialized documents, such as in NTCIR-1 and NTCIR-2, which contain abstracts of the Academic Conference papers. more than half are English-Japanese paired documents because abstracts are usually written in English and in Japanese but as parallel text, which is a text in a given language provided with its equivalent in another language.

Most of these specialized documents are either in a single language or constructed from abstracts, not mixed and complete documents with different lengths In addition, Arabic is rare among specialized corpora.

Corpora can be classified also, as synchronic or diachronic (McEnery et. al, 2006). Synchronic corpora are often used to compare regional varieties. Diachronic, or historical, corpora are usually used to compare vocabulary from the same language gathered from different time periods. To study regional variation in monolingual Arabic documents, Abdelali (2006) constructed a large corpus in Modern Standard Arabic (MSA) from different regional Arabic newspapers.

The corpus described in this paper will contain different regional variants as well as the general vocabulary of MSA. It is possible to say that the corpus under construction is specialized,

[1] http://www.ldc.upenn.edu/

[2] http://trec.nist.gov/

[3] http://research.nii.ac.jp/ntcir/index-en.htm

[4] http://www.clef-campaign.org/

synchronic and multilingual/mixed in both documents and queries.

3. WHY A MULTILINGUAL CORPUS OF COMMON COMPUTER SCIENCE

Most currently available test collections and almost all CLIR collections have focused upon general-domain news stories (Rogati and Yang, 2004). However, news collections have unique characteristics that are not provided in other genres, such as computer science (Gey, et al. 2005). Such characteristics include the regular use of proper nouns for places and names, the use of general purpose vocabulary and little use of dialects. In contrast, technical and scientific domains usually have rapidly developing terminology added to languages, especially non-English ones such as Arabic. Both NTCIR and CLEF have been working, to some extent, in domain-specific data, particularly in scientific abstracts and patents, but collections only cover a few languages. Arabic is not one of them.

Moreover, the majority of these collections are monolingual or consist of several monolingual documents in different languages, with each documents in a given language, rather than documents with different portions/terms that are tightly–integrated in multilingual forms. In non-English languages such as Arabic, documents are often multilingual – especially in the scientific domain. A multilingual document is a mixed document that contains different languages.

Scientific documents in Arabic have two distinguishing characteristics that are not found in English documents. Firstly, many multilingual documents contain different terms/portions/ snippets/phrases/paragraphs in two languages – usually English is one of them- but in a tightly-integrated manner. Secondly, a considerable number of multilingual documents contain similar description texts/snippets in multiple languages. In fact, a large number of bilingual terms/phrases/information in non-English scientific resources exists on the Web in the form of bilingual texts but not in a tightly integrated manner. For instance, in the multilingual phrase ‘(Hashing) ما هي البعثرة’ (meaning: what is Hashing) the English word ‘Hashing’ is presented as a translation for the Arabic word ‘البعثرة’ and hence removal of the English term will not make the sentence meaningless. This characteristic is prevalent in non-English documents. Zhang and Vines (2004) stated that, on Chinese Web pages, English terms are very likely to be the translations of their immediately preceding Chinese terms. Gey, et al. (2005) stated that an interesting characteristic of the document collections in non-English speaking countries is the number of technical terms and the existence of a partially paired corpus.

Moreover, sometimes the same term/word in the same multilingual document is written in different positions but in two different languages, each of which is tightly integrated with its neighbours. For example, the scientific term “deadlock” may occur in Arabic and English.

The phenomenon of multilinguality in scientific Arabic documents occurs for different reasons. First, at English was and still is the dominant language for scientific articles, lexicons, dissemination of information and different types of knowledge (Miniwatts Marketing Group, 2011). Second, many non-English-speaking users, such as Arabic speakers, do not know the exact translations/meanings for most terminology in scientific fields in their native languages. English scientific terms in the Arabic world are usually used to simplify ambiguous Arabic scientific terms. Third, translation/ transliteration of newly added terms to a non-English language, such as Arabic, is not usually performed on a regular basis. Fourth, is the problem of regional variation across the Arabic world, especially in scientific domains.

The majority of CLIR techniques focus on investigating the effectiveness of translation approaches but neither mixed-language queries nor searches for mixed-language documents have yet been adequately studied. This is because queries are usually translated to a monolingual target language. Most weighting algorithms, indexing methods and ranking functions were not designed for multilingual documents or queries.

In addition, CLIR techniques had proven their ability to retrieve and rank news stories but this does not mean that they are ready to be applied to scientific domains, especially in multilingual non-English collections. In such cases of documents, there is a possibility of poor retrieval because the assumed language model is incorrect.

The authors of this work are therefore developing a multilingual Arabic/English corpus of common computer science vocabulary as the first step to studying multilingual features in both queries and documents in scientific Arabic documents.

4. BUILDING THE CORPUS

The corpus has been collected both automatically and manually. In the automatic gathering process, the WebReaper Web crawler (WebReaper, 2010) was used. WebReaper has the ability to download pages at a given main URL and then follow a recursive process in downloading other linked pages. The Web crawler was initiated with some selected URLs that contain books, lectures, articles and discussions on common computer science. The choice was mainly governed by the availability of computer science documents and the respecting of authors' copyrights.

A manual collection of data was also considered. A group of 50 students at different academic levels at an Arabic university were asked to collect documents on common computer science topics. Some students downloaded documents from specific websites while others submitted their own queries to some search engines. Some students extracted documents from their academic reports and graduation projects. The collected documents were merged into a single pool. Duplicates were removed and a total size of 2.4 GB of raw Web-based data plus the extracted documents, from students' academic reports, was obtained.

Although the expansion of the corpus is still in progress, the process was characterized by two major challenges. First, many Arabic documents were found to be images in pdf format. This phenomenon is very common when conversion tools are used to convert Arabic documents to pdf files. In such cases, a contact was held with book's authors, in many cases, in order to provide a plain text version. Second, issues related to respecting copyright and intellectual properties were raised. Thus, an iterated process of contacting books' authors was carried before collecting documents.

5. CORPUS PROCESSING

After gathering the collection, documents were processed. At first, documents in different formats (shtml, html, doc, pdf..etc) were converted to HTML. The process was iterated and several applications were employed to perform this phase, (HTML parsers, Adobe Acrobat Reader, etc). During this step, tags, symbols, images and special characters, like ®, were removed. Only the raw text was retained. The new formatted HTML documents were saved in a common encoding, which is Unicode. Along with this step, each document was tagged with a special tag for referencing purposes, namely the name of the student who downloaded the document and his academic level if the document is downloaded manually - otherwise the phrase 'automatically downloaded' was used.

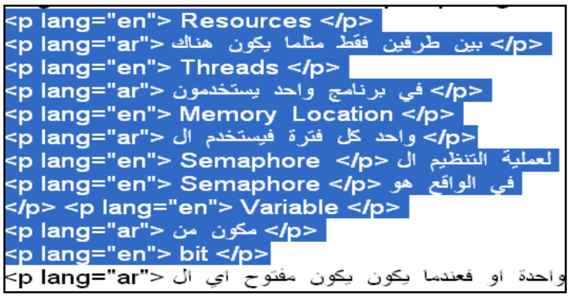
Run-on words between Arabic and English were also categorized and fixed as much as possible. The run-on words (Buckwalter, 2002) problem in multilingual collections occurs when the preceding word ends with a non-connecting letter. For instance, the word الSemaphore (meaning: the semaphore) is a run-on word because it is a concatenation between the Arabic definite ال (meaning: the) and the English word semaphore. In multilingual documents this is a severe problem because it may cause an IR system to stem such run-on words with the wrong stemmer. Along with this step, a normalization process was carried out for Arabic documents and Arabic parts in multilingual documents to render different forms of some letters with a single Unicode representation. Normalization in Arabic is usually performed in order to control the orthographic variations, which is very common in Arabic (Tayli and Al-Salamah, 1990). The problem makes exact matching inadequate for Arabic retrieval and may cause invalid stemming of words. Therefore, in Arabic IR some letters are unified into a single letter. These cases are well-known in Arabic because there are only a few Arabic letters that have different spelling variants in glyphs. Thus, noramilzation that has been performed in the corpus for Arabic words includes: replacing HAMZA (أ،إ) and MADDA (آ) with bare ALIF (ا); replacing final un-dotted YAA (ى) with dotted YAA (ي); replacing final TAA MARBOOTA (ة) with HAA (ه); and replacing the sequence ىء with ئ. Diacritical marks were also removed. Kasheeda, the Arabic stylistic elongation of some words for cosmetic writing, was also normalised by removing the letters included purely for elongation (e.g., التجميـــــع becomes التجميع). English documents and English parts in multilingual documents were also normalized in terms of case-sensitivity.

Regional variants in the collection were kept although a significant proportion of Arabic technical terms were found to be inconsistent and in different regional variants. Table 1 shows a sample of these regional variations in the collection. Academies of Arabic Language across the Arabic world need to unify their terminologies when a new technical term is Arabicized.

Table 1: Some regional variants.

English Term	Arabic Term	English Term	Arabic Term
Linked List	القائمة المتصلة	Object Oriented Programming	البرمجة الشئية
	السلسلة المتصلة		البرمجة الكائنية
	اللائحة المترابطة		البرمجة موجهة الأهداف
Deadlock	الجمود	Normalization	التبسيط
	الإقفال		
	التقاطع		التطبيع
	الإستعصاء		

In order to prepare the text for multilingual indexing later, every word/phrase/portion/ paragraph - depending on how much a document is mixed - in documents was marked with a language tag attribute using a simple language identifier. This would help to identify the correct stemmer during the indexing phase. So if a given document is in a monolingual language, the attribute “lang” is added to the body tag of the html file, e.g. <body lang =”en”> ; otherwise the “lang” attribute is added to a paragraph tag <p> in order to show that this portion is in a specific language , e.g. <p lang= “ar”>. The former is used for monolingual documents while the latter is used for multilingual documents. Figure 1 shows a multilingual document after being processed. Arabic is read from right to left. Thus, insertion of English words sometimes makes sentences appear a little confused.



```
<p lang="en"> Resources </p>
<p lang="ar"> بين طرفين فقط مثلما يكون هناك </p>
<p lang="en"> Threads </p>
<p lang="ar"> في برنامج واحد يستخدمون </p>
<p lang="en"> Memory  Location </p>
<p lang="ar"> واحد كل فترة فيستخدم ال </p>
<p lang="en"> Semaphore  </p> لعملية التنظيم ال
<p lang="en"> Semaphore </p> في الواقع هو
</p> <p lang="en"> Variable </p>
<p lang="ar"> مكون من </p>
<p lang="en"> bit </p>
<p lang="ar"> واحدة او فعندما يكون يكون مفتوح اي ال
```

Figure 1: A processed document.

6. CORPUS STATISTICS

In order to obtain the essential information needed for the corpus analysis, the *Lucene* IR system[5] was used. Lucene is an experimental information retrieval system that has been extensively used in previous editions of the CLEF, NTCIR and TREC joint evaluation experiments. Lucene is a high-performance, full-featured text search engine library written entirely in Java. Lucene has the ability to index and retrieve files in Unicode. The size of index in Lucene is roughly 20-30% compared to the size of text to be indexed. Lucene has a very good diagnostic tool known as Luke[6] that is able to access an index that is created by Lucene. Through Luke it is possible to: browse documents; display frequent terms; and optimize the index. Thus, using both Lucene and Luke all documents in the corpus were indexed and simple statistics about the numbers of words in the collection were extracted. Table 2 shows these statistics. From the table, it is observed that the average number of words per document is relatively high. This is typically true when it is compared with standard collections such as AP (Associated Press newswire documents – from TREC disks 1-3), which contains 242,918 documents. The average number of words per document in the AP collection is 474 (Croft et al., 2010).

This is considerably shorter when compared to the average number of words in Table 2 (approximately 3 times larger), bearing in mind the big difference in sizes. This is because AP is a news collection. In such collections, the general purpose vocabulary is predominant in most documents. This is not the case in scientific collections. Another important observation from Table 2 is that although the data has being collected arbitrarily, monolingual Arabic documents are very rare, at least in terms of common computer science

[5] http://www.lucene.com
[6] http://www.getopt.org/luke

Table 2: Collection's summary.

Description	Language(s) in documents	Number of words	Total
Number of words	English words	2,194,651	3,071,003
	Arabic words	876,352	
Number of distinct words	Distinct words in English	68,615	99,430
	distinct words in Arabic	30,815	
Number of documents	Monolingual English documents	1397	2,232
	monolingual Arabic documents	26	
	multilingual (both Arabic and English) documents	809	
Average number of words per document		1,376	

7. CORPUS ANALYSIS

Implementation of statistical tests on a corpus is an important process in understanding its nature in terms of validity and adequacy to serve as a test-bed. Such statistical tests would help in estimating how terms are distributed across documents and whether their distribution is skewed or not. Among the several possible statistical tests, the following were applied.

7.1 Zipf's Law

Zipf's law is a commonly used model for the distribution of words in a collection. Given a corpus in a natural language, Zipf's law states that the frequency of any word (*f*) is inversely proportional to its rank (*r*). Alternatively, the frequency of a word (*f*) times its rank (*r*) is approximately a constant (*k*):

$$r \ * f = k \qquad (1)$$

Ideally, when *log(f)* is drawn against *log(r)* in a graph, a straight line with a slope of -1 is obtained. The intuition in Zipf's law is that frequency decreases very rapidly with rank.

Figure 2 shows the three Zipf's curves applied to the corpus. In the figure each language is analyzed separately, along with analyzing the entire corpus together. Curves are quite accurate and clearly show that frequencies decrease rapidly with ranks, meaning that frequencies of the most common words are inversely proportional to their ranks. There are no skewed frequencies. In addition, the predicted relationship of curves indicates that they improve as the size of data increases.

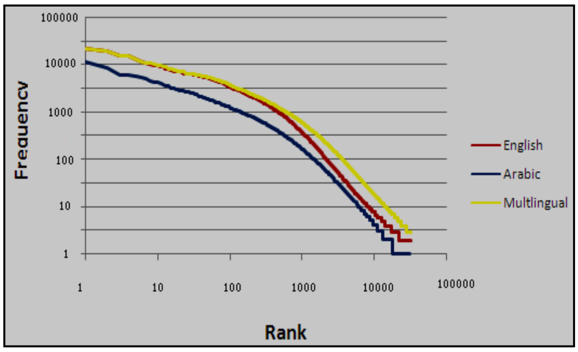


Figure 2: Zipf curves for the corpus.

7.2 Token-to-Type Ratio

Token-to-Type Ratio, known as TTR, is a lexical variety measure for text usually used to evaluate the richness of collections and their appropriateness for use in a specific task, i.e. in IR (WordSmith Tools, 2011). Thus, the measure reflects mainly sparseness of data (Schmitt, 2002). The TTR is computed as the number of occurrences divided by the distinct words. Therefore, lower ratios are expected for more distinct words. The TTR is informative only if we are dealing with a corpus comprising lots of equal-sized text segments (WordSmith Tools, 2011). Contrarily, if we are dealing with texts of different lengths then the TTR will not help much. Therefore, different and equal text length(s) for both Arabic and English are used. This was done by accumulating words at these points regardless of positions inside documents.

Table 3 shows the TTR ratios for both Arabic and English in the corpus while Figure 3 shows TTR curves. Both regional variants across scientific terminology and Arabic morphology affect the obtained TTR. It is clear that the lexical variety in the corpus is suitable and has high sparseness as more words occur only once.

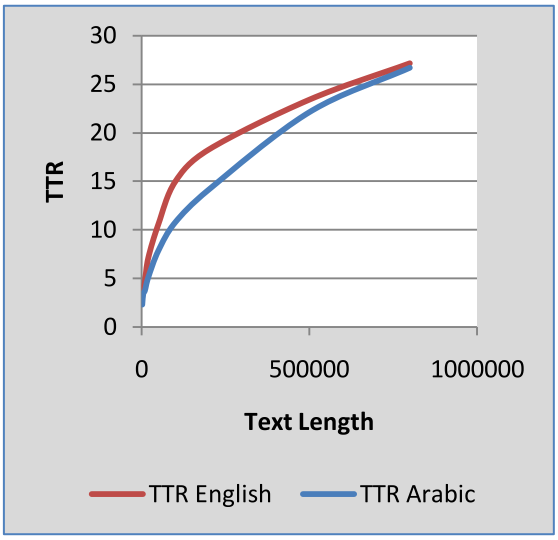


Figure 3: Token type ratio curve.

Table 3: Token-to-type ratios for different lengths.

Text Size	Arabic		English	
	Distinct words	TTR	Distinct words	TTR
2000	879	2.28	602	3.32
5000	1475	3.39	1427	3.50
10000	2673	3.74	2111	4.74
20000	3919	5.10	2797	7.15
50000	6379	7.84	4714	10.60
100000	9299	10.75	6678	14.97
200000	14155	14.13	10941	18.28
500000	22604	22.12	21349	23.42
800000	29951	26.71	29441	27.17

7.3 Heap's Law

Heap's law is another predication model that is used to predict vocabulary growth (Manning et al.,2008; Croft, 2009) and it is used to estimate the vocabulary size as a function of a collection size. Thus, Heap's law states that the relationship between the size of the corpus and the size of the vocabulary is:

$$v = k * N^{\beta} \qquad (2)$$

where *v* is the vocabulary size for a corpus of size N words and *k* and *β* are parameters. A typical value of k is $10 \leq k \leq 100$ and $\beta \approx 0.5$. Thus, Heap's law predicts that new words increase very rapidly when the corpus is small and would continue to increase, but at a slower rate, as the corpus size increases. Figure 4 shows a plot of vocabulary growth for the corpus. On the same figure, the Heap's curve with k = 50 and β =0.455 is also illustrated. The curve is a good fit. As examples for this prediction's accuracy: in the first 20,609 words in the corpus, Heap's law estimates that the number of the distinct words will be 4,591, whereas the actual value is 4,803; in the first 181,796 words, Heap's law predicts 12,361, whereas the actual number is 12,724. From this comparison of Heap's law with the corpus, it is concluded that vocabulary growth at present is a good fit.

However, as the corpus grows steadily, with future gathering of data, it is estimated that Heap's curve will become inaccurate at some points, unless a randomization accumulation of documents is performed. This is because the collection is multilingual and scientific. Consider that the first 20,000 documents are in English, whereas the second 10,000 documents are in Arabic or multilingual. In such cases it is estimated that the first 20,000 monolingual English documents will be accurate if their growth is estimated by Heap's law but after the occurrence of Arabic documents or multilingual documents the growth of the vocabulary will increase rapidly because most words are in Arabic and indeed different from the accumulated vocabulary of English. Thus, for future experiments, it is better to consider applying Heap's law for each language in the multilingual collection, separately. Another option is to randomize accumulation of documents.

In addition, there is another issue that was noticed in the implementation of Heap's law. Scientific documents are usually very varied in their length. Along with this fact, their vocabulary may be totally different from one field to another, e.g., information retrieval field vs. human computer interaction field. These two characteristics may affect the document growth substantially.

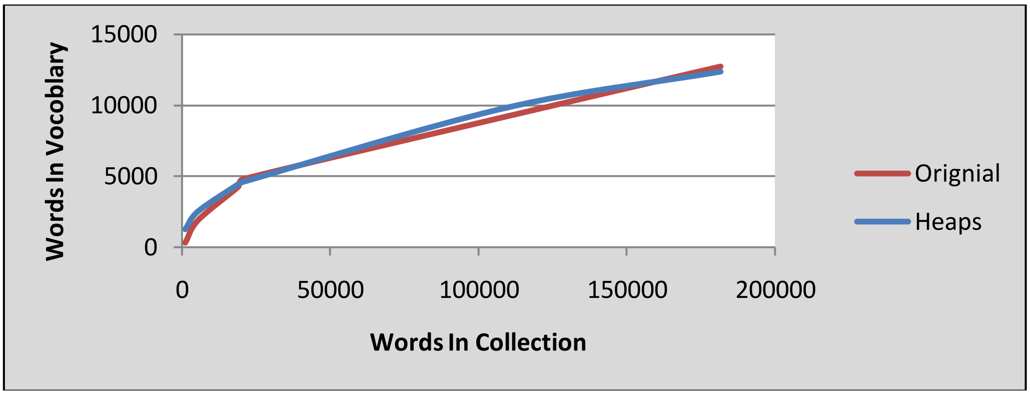


Figure 4: Vocabulary growth in the collection

.

8. CONCLUSION

Pages available on the WWW are rich resources for building a significant multilingual corpus of common computer science vocabulary. The nature of the corpus and the diversity in languages, plus its scientific characteristic, make it interesting to investigate. Such a scientific multilingual corpus would serve as a test-bed for studying the feature of multilinguality in both Arabic documents and queries in order to devise new techniques for weighting, indexing and retrieval of such documents. The most significant difficulties that may slow down building scientific corpora are: obtaining permission from authors to avoid the intellectual property and copyrights issues; the efforts needed to clean up documents, especially the Arabic ones; and the assessment of relevance judgments of documents, which will be considered later. In this work, the sample corpus collecting and analysis has been presented.

The corpus was validated in terms of terms frequencies, sparseness and growth, using statistical tests. There is thus no reason to believe that the data is imbalanced at present.

Future work will focus on extending the corpus in terms of size. Other investigations in terms of multilinguality also will be considered.

8. REFERENCES

Abdelali, A., 2006. *Improving Arabic Information Retrieval Using Local Variations in Modern Standard Arabic*, PhD Dissertation, New Mexico Institute of Mining and Technology

Cleverdon, C. W., 1970. Progress in documentation evaluation tests of information retrieval systems, *Journal of Documentation*, Volume (26), pp. 55-67

Dunlop M.D., Van Rijsbergen C.J., 1993. Hypermedia and free text retrieval, *Information Processing & Management*, Elsevier Ltd., Volume (29), pp. 287-298

Gey, Fredric C., Noriko, K., Carol, P., 2005. Language Information Retrieval: the way ahead, *Journal of Information Processing and Management*, Elsevier, Volume (41), pp. 415 - 431

Hmeidi, I., Kanaan, G., Evens, M., 1997. Design and Implementation of Automatic Indexing for Information Retrieval with Arabic Documents. *Journal of the America Society for Information Science*, Volume (48), pp. 867-881

Lin, W.,Chen, H., 2003. Merging Mechanisms in Multilingual Information Retrieval, *Advances in Cross-Language Information Retrieval LNCS*, Springer-Verlag, Volume (2785), pp . 175-186

McEnery, T., Xiao, R., Tono, Y., 2006. *Corpus-based language studies: an advanced resource book*, Routledge, USA, 1

Rogati, M. Yang, Y. 2004. *Resource Selection for Domain Specific Cross-Lingual IR*, In Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval SIGIR'04, ACM, USA

WordSmith Tools.2011, Type/Token Ratios and the Standardised Type/Token ratio, Available at: http://www.lexically.net/downloads/version5/HTML/index.html?type_token_ratio_proc.htm/**,** Last accessed 17 -1- 2011