Published at the 6th *International Conference on Electrical and Computer Systems (EECS'10)*, Nov 25-26, 2010, Lefke, North Cyprus.

OSAC: Open Source Arabic Corpora

Motaz K. Saad
Faculty of Information Technology
Islamic University of Gaza
Gaza, Palestine
e-mail msaad@iugaza.edu.ps

Wesam Ashour
Computer Engineering Department
Islamic University of Gaza
Gaza, Palestine
e-mail washour@iugaza.edu.ps

***Abstract*—Arabic Linguistics is promising research field. The acute lack of free public accessible Arabic corpora is one of the major difficulties that Arabic linguistics researches face. The effort of this paper is a step towards supporting Arabic linguistics research field. This paper presents the complex nature of Arabic language, pose the problems of: (1) lacking free public Arabic corpora, (2) the lack of high-quality, well-structured Arabic digital contents. The paper finally presents *OSAC*, the largest free accessible that we collected.**

Keywords: Arabic Language, Arabic corpora, Arabic Digital contents.

1. INTRODUCTION

Arabic Language is the 5th widely used languages in the world. It is spoken by more than 422 million people as a first language and by 250 million as a second language [8]. Arabic has 3 forms; Classical Arabic (*CA*), Modern Standard Arabic (*MSA*), and Dialectal Arabic (*DA*). *CA* includes classical historical liturgical text, MSA includes news media and formal speech, and *DA* includes predominantly spoken vernaculars and has no written standards. Arabic alphabet consists of the following 28 letters (أ ب ت ث ج ح خ د ذ ر ز س ش ص ض ط ظ ع غ ف ق ك ل م ن ه و ي) in addition, the Hamza (ء). There is no upper or lower case for Arabic letters like English letters. The letters (أ و ي) are vowels, and the rest are constants. Unlike Latin-based alphabets, the orientation of writing in Arabic is from right to left.

Table 1: Diacritics

Double Constant	No Vowel	Nunation			Vowel		
بّ /bb/	بْ /b/	بٍ /bin/	بٌ /bun/	بً /ban/	بِ /bi/	بُ /bu/	بَ /ba/

حقوق الانسان
حقـــــوق الانســـــان
حقـــــــــــوق الانســـــــــــان
حقـــــــــــــــوق الانســـــــــــــــان

Fig. 1: Tatweel (kasheeda)

The Arabic script has numerous diacritics, including I'jam (إعجام), consonant pointing, and tashkil (تشكيل), supplementary diacritics. The latter include the ḥarakat (حركات, singular haraka حركة), vowel marks. The literal meaning of tashkil is "forming". As the normal Arabic text does not provide enough information about the correct pronunciation, the main purpose of tashkil (and ḥarakat) is to provide a phonetic guide or a phonetic aid; i.e. show the correct pronunciation (double the word in pronunciation or to act as short vowels). The ḥarakat, which literally means "motions", are the short vowel marks [7]. Arabic diacritics include Fatha, Kasra, Damma, Sukūn, Shadda, and Tanwin. The pronunciations of diacritics aforementioned are presented in Table 1. Arabic words may also have Tatweel or kasheeda as shown in figure 1.

Arabic words have two genders, masculine (مذكر) and feminine (مؤنث); three numbers, singular (مفرد), dual (مثنى), and plural (جمع); and three grammatical cases, nominative (الرفع), accusative (النصب), and genitive (الجر). A noun has the nominative case when it is subject (فاعل); accusative when it is the object of a verb (مفعول); and the genitive when it is the object of a preposition (مجرور بحرف جر). Words are classified into three main parts of speech, nouns (اسماء) (including adjectives (صفات) and adverbs (ظروف)), verbs (افعال), and particles (ادوات).

Despite Arabic language is widespread, there is acute luck of well-structured and high quality Arabic digital contents. There are also a lack of free and Public Arabic corpora. This paper presents OSAC, open source Arabic corpora that cover different text genres which can be used in the future as a benchmark.

The rest of this paper is organized as follows: section 2 presents the complexity of Arabic language, section 3 talks about the problem of lacking of Arabic corpora and Arabic digital contents, describes corpora building steps, and presents the collected corpora, and finally section 4 draw the conclusion.

2. COMPLEXITY OF ARABIC LANGUAGE

Arabic is a challenging language for a number of reasons:

- Orthographic (الاملاء) with diacritics is less ambiguous and more phonetic in Arabic, certain combinations of characters can be written in different ways [7].

- Arabic language has short vowels which give different pronunciation. Grammatically they are required but omitted in written Arabic texts [7].
- Arabic has a very complex morphology as compare to English language [1, 9, 12, 13].
- Synonyms are widespread. Arabic is a highly inflectional and derivational language [1, 8, 12, 13].
- Lack of publically freely accessible Arabic Corpora [3, 4, 5, 6, 13].
- Lack of Arabic digital contents [11, 13].

In the following, we shall discuss these points in details.

Word meanings: It is possible to identify the different meanings associated with a word, due to one word may have more than one meaning in different contexts.. Table 2 shows the Arabic word (قلب) which has 3 meaning as a noun.

Table 2: The meaning of word (قلب) as a noun

Word meaning	Sentence
core	في **قلب** الاحداث
heart	اجرى عملية **قلب** مفتوح
center, middle	الكرة في **قلب** الملعب

Variations in lexical category: One word may have more than lexical category (noun, verb, adjective, etc.) in different contexts as shown in Table 3. Morphological analysis of a given corpus includes investigating word frequency of a word as a lexical category.

Table 3: The Lexical Category of word (عين)

Word meaning	Word Category	Sentence
Ain	Proper-Noun	**عين** جالوت
wellspring	Noun	**عين** الماء
eye	Noun	**عين** الانسان
delimitate/be delimitate	Verb/passive Verb	**عين** وزيرا للخارجية

Synonyms: Languages have many words that are considered synonymous. Through a given corpus, the researchers can use morphological analysis tools to know synonyms of a word, the frequency of each word of those synonyms and which one of them is more common. Examples of synonyms in Arabic are (بذل منح اعطى وهب) which means (give), (اسرة عائلة) which means (family), and (فصل صف) which means (classroom).

The word form according to its case: The form of some Arabic words may change according to their case modes (nominative, accusative or genitive). For instance, the plural of word (مسافر) which means (traveler) may be in the form (مسافرون) in the case of nominative (مرفوعة) and the form (مسافرين) in the case of accusative/genitive (منصوبة/مجرورة). Arabic light stemming can handle these cases.

Morphological characteristics: An Arabic word may be composed of a stem plus affixes and clitics. The stem consists of a consonantal root (جذر صحيح) and a pattern morpheme (اصغر كلمة ذات معنى). The affixes include inflectional markers (علامات او حركات اعرابية) for tense, gender, and/or numbers. The clitics include some prepositions (حروف جر), conjunctions (حروف العطف), determiners (محددات), possessive pronouns (ضمائر الملكية) and pronouns (ضمائر). The clitics attached to the beginning of a stem are called proclitic and the ones attached to the end of it are called enclitics. Most Arabic morphemes are defined by three consonants, to which various affixes can be attached to create a word. For example, from the tri-consonant "ktb" (كتب), we can inflect (يصرف) several different words concerning the idea of writing as (wrote كَتب), (book كِتاب), (the book الكِتاب), (books كُتُب), (he writes يكْتُب), (writer كاتِب), (library مكْتَبَة). Moreover an Arabic word may correspond to several English words. Another example is the Arabic word (وبنفوذها) and its equivalence in English "and with her influences". This makes segmentation of Arabic textual data different and more difficult than Latin languages.

Affixes set in Arabic are shown in Table 4, and Arabic patterns (الأوزان) and roots are shown in Table 5. The word (علم) may give various meanings by adding different affixes (prefixes, infixes, or suffixes) as shown in Table 6. Other morphological variations example is the word (يذهب) which means (go) are presented in Table 7.

Table 4: Affix set in Arabic Language

Affixes in Arabic	Examples
Prefixes of length 3	ولل ، وال ، كال ، بال
Length 2 prefixes	ال ، لل
Length 1 prefixes	ل ، ب ، ف ، س ، و ، ى ، ت ، ن ، ا
Length 3 suffixes	تمل ، همل ، تان ، تين ، كمل
Length 2 suffixes	ون ، ات ، ان ، ين ، تن ، كم ، هن ، نا ، يا ، ها ، تم ، كن ، ني ، وا ، ما ، هم
Length 1 suffixes	ة ، ه ، ي ، ك ، ت ، ا ، ن

Table 5: Arabic Patterns and Roots

Arabic Pattern and roots (الأوزان)	Examples
Length 4 pattern	فاعل فعلة فعال مفعل
Length 5 pattern and length 3 roots	تفاعل افتعل افعال فعالة فعلان فعولة تفعلة تفعيل مفعلة مفعول فاعول فواعل مفاعل مفعيل افعلة فعائل منفعل مفتعل فاعلة مفاعل فملاع يفتعل تفتعل فعلالي انفعل
Length 5 pattern and length 4 roots	تفعلل افعلل مفعلل فعللة فعلان فعالل
Length 6 pattern and length 3 roots	استفعل مفاعلة افتعال افعوعل انفعال مستفعل
Length 6 pattern and length 4 roots	افتعلل افعلال متفعلل

Stemming usually used to convert words to root form, it dramatically reduces the complexity of Arabic language morphology by reducing the number of feature / keywords in corpora. The reason for using stemming as feature / keywords reduction technique is that all morphology of words mostly has the same context meaning, but the case is not always true. Table 8 shows some of these cases. There is another approach for morphology reduction that just removes

affixes and does not convert the word to bas/root form. This approach is called light stemming [13, 14].

Table 6: Versions of the word (علم) and its meaning when adding affixes

Meaning	Suffix	Infix	Prefix	Word
Scientific	ية	***	***	علمية
Learned us	تنا	***	***	علمتنا
His science	ه	***	***	علمه
Scientists	اء	***	***	علماء
Teaching	***	ي	ت	تعليم
Sciences	***	و	***	علوم
Informative	يه	ا	است	استعلامية

Table 7: Morphological variation of word (ذهب)

verb	time	# of participants	Gender of subjects
ذهب	Past	1	Male
ذهبت	Past	1	Female
ذهبا	Past	2	Male
ذهبتا	Past	3	Female
ذهبوا	Past	3 or more	Male
ذهبن	Past	3 or more	Female
يذهب	Present	1	Male
تذهب	Present	1	Female
سيذهب	Future	1	Male
ستذهب	Future	1	Female
سيذهبوا	Future	3 or more	Male
سيذهبن	Future	3 or more	Female

Table 8: Different meaning of morphology of the same root in Arabic

Meaning	Root	Word
Class room **Apartheid**	فصل	**الفصل** الدراسي **الفصل** العنصري
Goes out of house **Graduate from university**	خرج	**يخرج** من البيت **تخرج** من الجامعة
The fisherman twist the cord **The student argued with the teacher**	جدل	**جدل** الصياد الحبل **جادل** الطالب المدرس
He focuses the arrow **The man lost his mind**	صوب	انه **يصوب** السهم فقد الرجل **صوابه**

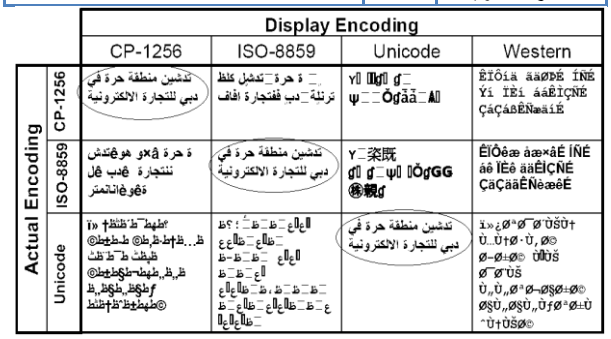

Fig. 2: Arabic Encoding Problem

Table 9: Unicode vs. cp-1256 Arabic windows encoding

Unicode	CP-1256 Arabic windows
Becoming the standard more and more	Commonly used
2-byte characters	1-byte characters
Widely supported input/display	Widely supported input/display
Supports extended Arabic characters	Minimal support for extended Arabic characters
Multi-script representation	bi-script support (Roman/Arabic)
Supports presentation forms (shapes and ligatures)	Tri-lingual support: Arabic, French, English (ala ANSI)

Encoding Problem: Arabic Language has display Problems (encoding issues) because it has different encoding according to machine platform. Figure 2 shows the problem of using incorrect encoding where all circled cells are displayed correctly while the other cells are displayed incorrectly. Text preprocessing, mining, and information retrieval with incorrect encoding may lead to incorrect results. Table 9 presents the characteristics of two common Arabic encoding systems; *Unicode* and code page 1256 *CP-1256* Arabic windows.

3. ARABIC CORPORA

Corpus-based approaches to language have introduced new dimensions to linguistic description and various applications by permitting some degree of automatic analysis of text. The identification, counting and sorting of words, collocations and grammatical structures which occur in a corpus can be carried out quickly and accurately by computer, thus greatly reducing some of the human drudgery sometimes associated with linguistic description and vastly expanding the empirical basis [3, 4]. Linguistic research has become heavily reliant on text corpora over the past ten years. Text data mining is a multidisciplinary field involving information retrieval, text analysis, information extraction, clustering, categorization and linguistics. Text mining is becoming of more significance, and efforts have been multiplied in studies to provide for fetching the increasingly available information efficiently [3, 4].

Due to the increasing need of an Arabic corpus to represent the Arabic language and because of the trials to build an Arabic corpus in the last few years were not enough to consider that the Arabic language has a real, representative and reliable corpus, it was necessary to build such an Arabic corpus to support various linguistic research on Arabic [3, 4]. Thus, one of the difficulties that encountered Arabic Language researches is the lack of publicly available Arabic corpus [3, 4, 5, 6]. Arabic corpus problem was posed by [3, 4, 5, 6]. A survey by [3, 4] confirms that existing corpora are too narrowly limited in source-type and genre, and that there is a need for a freely-accessible Corpus of Contemporary Arabic (*CCA*) covering a broad range of text-types. Due to the Arabic language lacking of corpora, it is difficult to display textual content and quantitative data of Arabic.

Al-Nasray et. al. [3, 4] discussed three axes in their paper; the 1st axes is a survey of the importance of corpora in language studies e.g. lexicography, grammar, semantics, Natural Language Processing and other areas. The 2nd axis demonstrates how the Arabic language lacks textual resources, such as corpora and tools for corpus analysis and the effected of this lack on the quality of Arabic language applications. There are rarely successful trials in compiling Arabic corpora, therefore, the 3rd axis presents the technical design of the International Corpus of Arabic (*ICA*), a newly established representative corpus of Arabic that is intended to cover the Arabic language as being used all over the Arab world. The corpus is planned to

support various Arabic studies that depends on authentic (اصيلة) data, in addition to building Arabic Natural Language Processing Applications.

International Corpus of Arabic (*ICA*) is a big project initiated by Bibliotheca Alexandrina (*BA*). *BA* is one of the international Egyptian organizations that play a noticeable role in disseminating culture and knowledge, and in supporting scientific research. *ICA* is a real trial to build a representative Arabic corpus as being used all over the Arab world to support research on Arabic [3, 4]. *ICA* corpus has been analyzed by Al-Nasry et. al. in [4], they shed light on the levels of corpus analysis e.g. morphological analysis, lexical analysis, syntactic analysis and semantic analysis. Al-Nasry also demonstrates different available tools for Arabic morphological analysis (*Xerox, Tim Buckwalter, Sakhr and RDI*). The morphological analysis of *ICA* includes: selecting and describing the model of analysis, pre-analysis stage and full text analysis stages. *ICA* is not publically available now and it expected to be released soon [3, 4].

3.1 Arabic Digital Content

Yet Arabic is the fastest-growing language on the internet, with Arabic-speaking internet users increasing 2,298 per cent from 2000-2009, according to the Internet World Statistics Report *internetworldstats.com*. The number of internet users in the Middle East and North Africa (Mena) region has leapt from 3.2 million users in 2000 to 60.25 million in 2009 and it is estimated that at least another 55 million new users will come online in the next five years. If mobile internet users are included, that figure soars even further to 150 million.

The content problem is of both quantity and quality. There is a lack of high-quality, well-structured websites managed by companies creating digital content for Arabic-speaking users. For example, if you search in English for a specific mobile phone model, you will land on a specialized portal with specifications, reviews and photos. In Arabic, you will probably end up in a forum where a question is being asked about that phone. It is unlikely in Arabic searches that the first page of results would not have a forum. There is a regional need for real local content and generally users in the region prefer Arabic today.

However, while Arabic content may have had a growth spurt in the past year, the content that has grown is still primarily user-generated and often machine translated. There is still a lack of original, localized, high-quality content.

3.2 Creating Arabic contenet online

There are many Arabic digital content enrichment initiatives. United Nations Economic and Social Commission for Western Asia - *ESCWA* released a project in 2007 to develop the industry of Arabic digital content. Wiki Arabi is a project initiated by King Abdulaziz City for Science and Technology (KACST) within the framework of King Abdullah's Initiative for Arabic Content. The project aims to improve the Arabic content on Wikipedia by promoting translation of high quality articles i different subject areas, including Nanotechnology, Biotechnology and Public Health. It aims to translate 2,000 articles within these areas in its first phase.

Major web players are looking to boost Arabic-language content online in a bid to meet demand from a rapidly growing Arab audience [11]. The Arab world has been facing a digital conundrum for the past few years – not enough users online creating content in Arabic; not enough content in Arabic to push internet penetration [11]. Although there are more than 422 million Arabic speakers worldwide and Arabic is the seventh-most popular language on the web, less than one per cent of all online content is in Arabic and there is just a 17.5 per cent internet penetration across the region's population.

Google has been working on several initiatives to help increase Arabic-language content. It tied up with Wikipedia after observing the Arabic portal of the online encyclopedia carried 120,000 pages compared with the 2 million pages of its Catalan equivalent. This is despite the disproportionate number of potential Arabic-speaking users, 422 million, compared with 6 million Catalan speakers [11]. About 10 million words have now been translated into Arabic from English on the site and 6 million from Arabic to English [11].

The search giant has also been educating small businesses to build their own websites using Google Sites – or to at least put their business directory information on Google Maps. It has built Ejabat, a user-generated question and answer system, which now has 600,000 questions and 2 million answers from 300,000 registered users [11]. With 20-25 per cent of Mena users in the past year being completely new to the web and a third of them under the age of 18, Google launched educational video site Ahlan (google.com/intl/ar/ahlanonline) to introduce users to the world of online learning. Within three months there were 1.2 million views of the Ahlan training videos.

US giant internet portal Yahoo, meanwhile, took a big leap into the Arabic content arena in 2009 when it acquired Maktoob, the region's largest community site. Maktoob is currently the 157th biggest site on the internet, according to web information company Alexa's listings (alexa.com). This makes it the 2nd most popular Arabic site behind Google Saudi Arabia at number 104 and way ahead of the third Arabic site in the world rankings, sports site Koora. Maktoob was founded in 2000 as the world's 1st free Arabic/English email service, but discussion forums quickly became its biggest traffic and content driver, with the women's forum one of the largest. Other popular areas include games, matrimonial, blogs and sports.

3.3 Bulding Arabic Corpora

Different corpora are available in English. Reuter's collections of news stories are popular and typical example. The Linguistic Data Consortium (*LDC*) provides two non-free Arabic corpora, the Arabic

NEWSWIRE and Arabic *Gigaword* corpus. Both corpora contain newswire stories.

There is a need for a freely-accessible corpus of Arabic. There are no standard or benchmark corpora. Thus, all researchers conduct their researches on their own compiled corpus. Arabic language is highly inflectional and derivational language which makes text mining / Information Retrieval a complex task. In Arabic text mining research field, there are some published experimental results, but these results came from different datasets, it is hard to compare classifiers because each research used different datasets for training and testing [15]. Sebastiani stated at [15] "We have to bear in mind that comparisons are reliable only when based on experiments performed by the same author under carefully controlled conditions".

One of the aims of this paper is to compile representative Arabic corpora that cover different text genres which can be in the future as a benchmark. Therefore, three different datasets were compiled covering different genres and subject domains.

Corpus sizes for the same topics written in Arabic and other different languages are not the same. In fact, the size of the corpus extracted from the French newspaper "*Le monde*" from the period of 4 years, is 80 million words [1, 2]. Moreover, the size of corpus extracted from the period of almost 7 years of Associated French Press (*AFP*) Arabic Newswire, and released in 2001 by *LDC* is 76 million tokens [1, 2]. This gap between the two sizes is justified by the compact form of the Arabic words. Formally speaking, the English word "write" is equivalent to one Arabic word "كتب". But the group "He writes", made up of two words, and also corresponds to one Arabic word "يكتب". And the Arabic equivalent of the sentence "He will write" is the only one word "سيكتب". Moreover, the word "سيكتبه" amounts to the group of words "He will write it". Another example is the Arabic word (وينفوذها) and its equivalence in English (4 words) "and with her influences". This makes segmentation of Arabic textual data different and more difficult than Latin languages. This gives an explanation of the gap between the two corpuses size, if we make into consideration the difference of data extraction period [1, 2]. On the other hand, the required amount of storage (disk or RAM) for Arabic corpus is twice of English corpus for the same number of characters for both corpora because Arabic characters require 2 bytes to be saved in Unicode format. This implies that feature/keyword reduction for Arabic text is necessary to consider storage limit.

Corpora Building Steps involves compiling and labeling text documents into corpus. We collect web documents from internet using the open source offline explorer, *HTTrack*. The process also includes converting corpus *html/xml* files into *UTF-8* encoding using "*Text Encoding Converter*" by *WebKeySoft*. The final step is to strip/remove html/xml tags as shown in Figure 3. We developed a Java program that strip / remove html/xml tags. The program is available publically at [10].

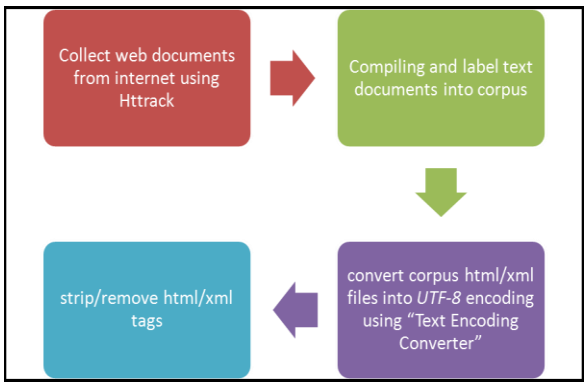

Fig. 3: Corpora building steps

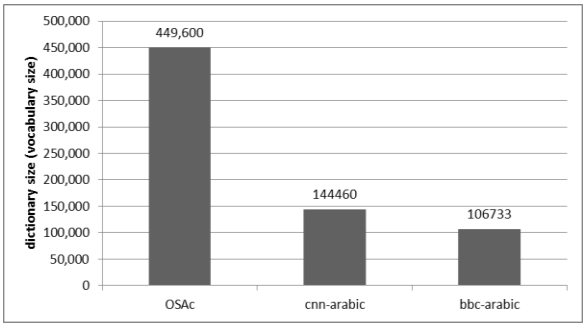

Fig 4: Dictionary size (# of keywords) for each corpus in *OSAC*

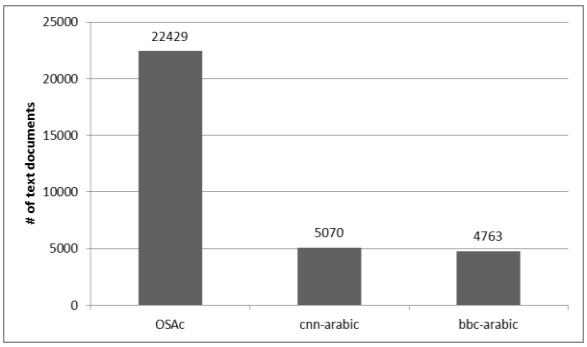

Fig 5: Number of text documents for each corpus in *OSAC*

***BBC* Arabic corpus**: We collected *BBC* Arabic corpus from *BBC* Arabic website bbcarabic.com, the corpus includes 4,763 text documents. Each text document belongs to 1 of 7 categories (Middle East News 2356, World News 1489, Business & Economy 296, Sports 219, International Press 49, Science & Technology 232, Art & Culture 122). The corpus contains 1,860,786 (1.8M) words and 106,733 district keywords after stopwords removal.

***CNN* Arabic corpus**: We collected *CNN* Arabic corpus from *CNN* Arabic website cnnarabic.com, the corpus includes 5,070 text documents. Each text document belongs to 1 of 6 categories (Business 836, Entertainments 474, Middle East News 1462, Science & Technology 526, Sports 762, World News 1010). The corpus contains 2,241,348 (2.2M) words and 144,460 district keywords after stopwords removal.

***OSAc* corpus**: We collected *OSAC* Arabic corpus from multiple websites as presented in Table 10, the corpus includes 22,429 text documents. Each text document belongs to 1 of 10 categories (Economics,

History, Entertainments, Education & Family, Religious and Fatwas, Sports, Heath, Astronomy, Low, Stories, Cooking Recipes). The corpus contains about 18,183,511 (18M) words and 449,600 district keywords after stopwords removal.

All collected corpora were converted the corpus to *utf-8* encoding, html tags were removed. The corpora are available publically at [10]. *OSAC* were used by Saad [13] to address the impact of text preprocessing on the Arabic text classification.

Table 10: *OSAC* corpus

Category	# of text docs	Sources
Economic	3102	bbcarabic.com - cnnarabic.com - aljazeera.net - khaleej.com - banquecentrale.gov.sy
History	3233	www.hukam.net تاريخ الحكام - altareekh.com التاريخ - moqatel.com islamichistory.net تاريخ الاسلام
Education and family	3608	saaid.net صيد الفوائد - نصائح للسعادة naseh.net - المربي الاسرية almurabbi.com
Religious and Fatwas	3171	*CCA* corpus - *EASC* corpus moqatel.com - islamic-fatwa.com شبكة الفتاوى الشرعية saaid.net صيد الفوائد -
Sport	2419	bbcarabic.com - cnnarabic.com - khaleej.com
Health	2296	dr-ashraf.com العيادة الالكترونية - *CCA* corpus - *EASC* corpus - W corpus - kids.jo صحة الطفل - arabaltmed.com العلاج البديل العربي
Astronomy	557	arabastronomy.com الفلك العربي - alkawn.net الكون نت - bawabatalfalak.com بوابة الفلك المغربية - nabulsi.com الفلك – موسوعة النابلسي - www.alkoon.alnomrosi.net -
Low	944	lawoflibya.com القانون الليبي qnoun.com قانون كوم
Stories	726	*CCA* corpus - kids.jo قصص الاطفال - saaid.net صيد الفوائد
Cooking Recipes	2373	aklaat.com - fatafeat.com
TOTAL	**22,429**	

4. CONCLUSION

Linguistic research has become heavily reliant on text corpora over the past ten years. Due to the increasing need of an Arabic corpus to represent the Arabic language and because of the trials to build an Arabic corpus in the last few years were not enough to consider that the Arabic language has a real, representative and reliable corpus, it was necessary to build *OSAC* to contribute supporting various linguistic research on Arabic.

Arabic language has complex morphology. The lack of well structured, high quality Arabic digital contents and the lack of the free accessible Arabic corpora were one of the major obstacles to Arabic linguistics research field. This paper is a step towards tackling these obstacles by collecting the largest free accessible Arabic corpus, *OSAC*, which contains about 18M words and about 0.5M district keywords.

In the future works, we shall work on extending and elaborating *OSAC*. Elaborations include performing extensive corpus analysis and tag them with Part of speech tags. We also open the door for other researchers and contributors to elaborate the open source corpora.

REFERENCES

[1]. Abbas M., Smaili K., Berkani D.: *Comparing TR-Classifier and KNN by using Reduced Sizes of Vocabularies*. The 3rd Int. Conf. on Arabic Language Processing, CITALA2009, Mohammadia School of Engineers, Rabat, Morroco. 2009.

[2]. Abdelali, A., Cowie, J., Soliman, H.: *Building a modern standard corpus, Workshop on Computational Modeling of Lexical Acquisition*. The Split Meeting, Split, 2005

[3]. Al-Ansary, S. Nagi, M., Adly N.: *Building an International Corpus of Arabic (ICA): Progress of Compilation Stage.* Bibliotheca Alexandrina. 2008.

[4]. Al-Ansary, S., Nagi, M., Adly N.*: Towards analyzing the International Corpus of Arabic: Progress of Morphological Stage.* Bibliotheca Alexandrina. 2008.

[5]. Al-Sulaiti L, Atwell E.: *Designing and developing a corpus of contemporary Arabic.* Int. Journal of Corpus Linguistics. pp.: 1 – 36. 2006.

[6]. Al-Sulaiti L, Atwell E.: *Designing and developing a corpus of contemporary Arabic.* Proc. of the 6th TALC conference, 2004.

[7]. Arabic diacritics - Wikipedia, the free encyclopedia, http://en.wikipedia.org/wiki/Arabic_diacritics

[8]. Arabic language - Wikipedia, the free encyclopedia, http://ar.wikipedia.org/wiki/لغة_عربية

[9]. Khoja S., Garside R.: *Stemming Arabic text.* Computer Science Department, Lancaster University, Lancaster, UK, 1999.

[10]. Motaz K. Saad: *Open Source Arabic Language and Text Mining Tools.* 2010. http://sourceforge.net/projects/ar-text-mining

[11]. Locke S., The *push for Arabic content*, http://www.meed.com/sectors/telecoms-and-it/telecoms/the-push-for-arabic-content-online/3007704.article Issue No 28 9-15 July 2010.

[12]. Saad M. K., Ashour W., *Arabic Text Classification Using Decision Trees*, Proceedings of the 12th international workshop on computer science and information technologies CSIT'2010, Moscow – Saint-Petersburg, Russia, 2010.

[13]. Saad M. K., *The Impact of Text Preprocessing and Term Weighting on Arabic Text Classification*, MSc. Thesis Dissertation, Computer Engineering Dept., Islamic University of Gaza, Palestine, 2010.

[14]. Saad M, K., and Ashour W., *Arabic Morphological Tools for Text Mining*, 6th ArchEng Int. Symposiums, EEECS'10 the 6th Int. Symposium on Electrical and Electronics Engineering and Computer Science, European University of Lefke, Cyprus, 2010.

[15]. Sebastiani, F.: *Machine learning in automated text categorization.* ACM Computing Surveys, 34(1), 1–47. 2002.