Overview of Arabic Sentence Corpora

Hussein Awdeh[1][a], Adelle Abdallah[1][b], Gilles Bernard[1][c] and Mohammad Hajjar[2]
[1]*LIASD Lab, Paris 8 University, 2 rue de la Liberté 93526 Saint-Denis, Cedex, France*
[2]*Faculty of Technology, Lebanese University, Hisbeh Street, Saida, Lebanon*

Keywords: Arabic Sentence Corpora, Arabic Language, Supervised Learning, Arabic Natural Language Process, Information Retrieval, Standard Corpus.

Abstract: The Arabic corpus, specifically the gold standard corpus is an important part of The Arabic Natural Language Processing. Described as a very large collection of texts stored on a computer, a corpus is considered as the most important source for semantic and syntax research and it can be a single language, a monolingual Corpus, or a multilingual Corpus. Then, an easy access to available corpora is highly needed in the Natural Language process (NLP) research community especially for language such as Arabic. Currently, there is no easy way to access to a comprehensive and updated list of available Arabic corpora. Our study in this paper, aims to present the results of a recent survey conducted to identify the list of the available Arabic corpora classified into categories and their resources.

1 INTRODUCTION

Arabic is one of the six most important languages on Earth. It is used and spoken by more than 273 million people today, is the language of the Quran, and has a rich history. There is much to be learned and preserved using natural language processing techniques as is commonly done with English, French and other languages. However, there is no central, freely-accessible, standard corpus that researchers can use for analysis. This paper compiles a survey of existing Arabic corpora, and discusses resources for building such corpora, as one step toward the goal of having a standard corpus.

The existing corpus with limited access does not support a growing interest and diversity in Arabic language research. The majority of Arabic corpora are limited in sources, types, and genres or are not freely available, and the high costs of building or licensing corpora could be an obstacle for many young researchers or even some institutions in several parts of the world.

Unfortunately, Arabic can still be considered a relatively poor resource language when compared to other languages such as French and English. They are generally: not easily found, sustainable by language data providers for fees, exclusively reserved for subscribers, or are expensive. Therefore, having free access to an Arabic standard corpus is clearly a desirable goal.

Based on the above, we conducted a survey of the existing Arabic corpora in order to compare the most famous ones, and to present the resources used for their building. Section 2 and 3 discuss the types of text corpora and the most relevant available free and commercial corpora. Then, in the section 4, we list the diverse ranges of sources for corpora. Section 5 addresses steps in our data pre-processing pipeline that are needed to prepare source material for NLP.

2 TYPES OF TEXT CORPORA

A corpus can be arranged into a wide range of categories by the metadata, resources, and/or its dependency on other corpus. It can fall into multiple categories according to the criteria if it satisfies these criteria.

This part of the paper reviews the different types of corpus, their specifications, and their

[a] https://orcid.org/0000-0002-2805-4444
[b] https://orcid.org/0000-0001-5837-8688
[c] https://orcid.org/0000-0002-4587-4209

Awdeh, H., Abdallah, A., Bernard, G. and Hajjar, M.
Overview of Arabic Sentence Corpora.
DOI: 10.5220/0010651200003063
In *Proceedings of the 13th International Joint Conference on Computational Intelligence (IJCCI 2021)*, pages 285-292
ISBN: 978-989-758-534-0; ISSN: 2184-2825

subcategories. Additionally, we present, at the end of this part, about 60 available Arabic corpuses distributed according to their categories (Table 1).

There are many different types of textual corpora:

- Raw text corpora – plain text with no additional information written in one language (Monolingual Corpus) or in multiple languages (Multilingual Corpus).
- Annotated corpora – text tagged with linguistic information such as named entity recognition, Error-Annotated Corpora, and Miscellaneous Annotated Corpora.
- Speech Corpora – recording transcribed data and audio.
- Handwriting Recognition Corpora – annotated and scanned documents.
- Miscellaneous corpora – multipurpose corpus (Q/A, summaries...).
- Lexicon – words lists and lexical database.

2.1 Raw Text Corpora

- Monolingual corpus: it is the most common type of corpus. It contains text in a single language. A wide range of researchers use this type of corpus for different tasks, like checking the right use of a word, detecting frequent patterns, or looking into the most characteristic word combinations. It is normally tagged for parts of speech. Sketch Engine contains many monolingual corpora in many languages.
- Parallel corpus: it contains two monolingual corpora, each one is the translation of the other, and for example, a story and its translation are used to build a parallel corpus. Both languages need to be aligned. The user would be able to look up for all instances of a word or expression in one language, and the outcomes will be shown together with the corresponding sentences in the other language. The user could see how the searched word or phrase is translated.
- Multilingual corpus: it is very close to a parallel corpus. It contains texts in several languages which are all translations of the same text and are aligned in the same way as parallel corpora. If the user chooses only two languages, a multilingual corpus behaves as a parallel corpus. The user can likewise choose to work with one language to utilize it as monolingual corpora.
- Comparable corpus: it is at least two monolingual corpora which are related to the same topic. Unlike multilingual corpus, texts are not translations of each other, and therefore, are not aligned. An example of comparable corpora in Sketch Engine is various corpora produced from Wikipedia or CHILDES corpora.
- Learner corpus: it is built by electronic collections of a language, and is used to study the mistakes and problems learners have when learning a foreign language.
- Diachronic corpus: it is a corpus that includes texts from various epochs. It can help researchers to study the language variation and development.
- Specialized corpus: it includes texts which are restricted to one or more topic fields, domains etc. These corpuses are used to study the usage of the specialized language.
- Multimedia corpus: it includes texts that are enriched and enhanced with audio, visual resources, or other type of multimedia content.
- Web-based Corpora: it can be very useful for concordance and frequency studies, provided the variety and big scale of these corpora (73 M words KACST corpus, 317M words for Leeds and 100M words for ICA and Parkinson corpus). Furthermore, the variety and the text type of the Arabic languages covered are large, which make these corpuses very appropriate for all genres of Arabic linguistic studies (Quranic Arabic, classic Arabic, newswire, books etc.).
- Dialectal Corpora: this Corpus (Almeman and Lee 2013) and the Tunisian Dialect Corpus (Graja et al.) are the most valuable dialectal corpora, particularly due to the fact that Arabic dialect processing research is a relatively recent activity and there is a great need for these services.

2.2 Annotated Corpora

Annotated corpora are very helpful in developing systems and software based on supervised algorithms, and the free access of resources can allow young researchers to train and develop systems at a reduced cost.

- Named Entity Corpora: Many of these corpuses have been mentioned by their respective writers at major NLP conferences that bring attention to these services. Their annotation structure is compatible with the

XML annotation requirements developed by big evaluation campaigns such as the Automated Information Extraction (ACE) Assessment Campaign.

- Error-Annotated Corpora: it can be very helpful for building automatic spelling correction tools and corpus-based error research.
- Miscellaneous Annotated Corpora: it contains a wide range of text genres (news, conversational telephony, weblogs, newsgroups, broadcasts, chat shows) like OntoNotes corpus (Weischedel et al. 2013) which is created in three languages (English, Chinese, and Arabic), and like the Quranic Arabic Corpus which is an annotated linguistic resource consisting of 77,430 words of Quranic Arabic.

2.3 Speech Corpora

The Arabic Speech Corpora made by Almeman and Lee (2013) is the only freely available speech corpus for Arabic.

2.4 Handwriting Recognition Corpora

Despite the small number of the available handwriting recognition corpora, they are mostly used commercially. They can be used for different NLP tasks from Optical Character Recognition (OCR) to writer identification.

2.5 Miscellaneous Corpora

The Miscellaneous Corpora such as question answering Ben Ajiba et al. (2007) and Trigui et al. (2010), plagiarism detection Bensalem et al. (2013), document summarization El-Haj et al. (2010) and El-Haj and Rayson (2013), are useful for a multitude of NLP related tasks.

2.6 Lexicon

- Lexical Databases: recently, many attempts have been made to develop various lexical resources for Arabic. Fortunately, several of them are accessible, such as version 1.0 of the well-known Buckwalter morphological analyser (Buckwalter, 2002). Many essential projects have been translated from English to Arabic such as the Arabic VerbNet (Mouss, 2010) and the Arabic WordNet (Elkateb, 2006).
- Words Lists: the Arabic words lists corpus can be used in the spell checking systems, or can be integrated with the lexicons of systems and tools to improve their performances. They can also be used by lexicographers to study various aspects of the Arabic language such as the Arabic MSA word count list.

Table 1: The distribution of Arabic corpora by types.

Corpus	Author	Words count
Monolingual Corpora		
Ajdir Corpus	Abdelali	113,000,000
KSU Corpus	Alrabiah	50,000,000
OSAC	Saad	18,183,511
Alwatan	Abbas	10,000,000
Tashkeela	Zarrouki	6,149,726
Al Khaleej	Abbas	3,000,000
KACST	Al-Thubaity	2,000,000
Arabic Words	Al-Saadi	1,500,000
Corpus of Contemporary Arabic	Al-Suleiti	842,684
CRI KACST Arabic Corpus	Alkanhal	235,000
Arabic Learners Written Corpus	Farwaneh	50,000
Multilingual Corpora		
Silver Arabic corpus	Awdeh	18,000,000
UN Corpus(Arabic portion)	Rafalovitc	2,721,463
Hadith Standard	Bounhas	2,500,000
MEEDAN Translation Memory	Meedan	1,000,000
EGYPT Translation Toolkit	CLSP/JHU	80,000
Dialectal Corpora		
Arabic Multi Dialect Text Corpora	Almeman, Lee	2,000,000
Tunisian Dialect Corpus	Graja, al.	3,403
Web-based Corpora		
KACST Arabic corpus	Al-Thubaity	732,780,509
Leeds Arabic Internet	Leeds	317,000,000
International Corpus of Arabic	Alansary, al.	100,000,000
Arabic Corpus	Parkinson	100,000,000
QURANY	Abbas N.	78,000
Quranic Text mining Dataset	Sharaf, al.	24,000
Named Entity Corpora		
JRC-Names	Steinberger	230,000
ANERCorp	Ben Ajiba	150,000
AQMAR Named Entity Corpus	Mohit, al.	74,000
Named Entities List	Attia, al.	45,202
ANERGazet	Ben Ajiba	14,000

Table 1: The distribution of Arabic corpora by types (cont.).

Corpus	Author	Words count
Error-Annotated Corpora		
Qatar Arabic language Bank(QALB)	Habash, al.	2,000,000
Arabic Learner Corpus	Alfifi, al.	282,000
KACST Error Corrected Corpus	Alkanhal, al.	65,000
Miscellaneous Annotated Corpora		
Kalimat Corpus	el-haj	18,167,183
OntoNotes Release 5.0	Weischede	300,000
The Quranic Arabic Corpus	Duke	77,430
AQMAR Arabic Wiki. Supersense Corpus	Schneider, al.	65,000
Khoja POS tagged	Khoja, al.	51,700
Arabic Wikipedia Dependency Corpus	Mohammed	36,000
AnATAr Corpus	Mezghani	18,895
Lexical Databases List		
BAMA 1.0 English-Arabic Lexicon	Buckwalter	82,158
Arabic-English Learner's Dictionary	Salmone	74,000
Unitex Arabic Package	Doumi, al.	50,407
ARALEX Online	Boudelaa	37,494
AraComLex Arabic Lexical Database	Attia, al.	30,000
Arabic VerbNEt	Mousser	23,341
Arabic WordNet	Elkateb, al.	18,957
NOOJ Arabic Dictionary	Mesfar	10,000
Qamoose	ArabEyes	N.A
List of Words Lists		
Word Count of Modern Standard Arabic	Attia, al.	1,000,000,000
Arabic Wordlist for Spellchecking	Attia, al.	9,000,000
Multiword Expressions	Attia, al.	34,658
Arabic Unknown Words	Attia, al.	18,000
Arabic Stop words	Zarrouki	13,000
Obsolete Arabic Words	Attia, al.	8,400
Arabic Broken Plurals	Attia, al.	2,562
		Files count
Arabic Speech Corpora	Almeman	67,132
Handwriting Recognition Corpora		
QUWI Handwritings Dataset	Al-Maadeed, al	1,000
Writer Identification Contest for Arabic Scripts Data set	Hassaïne, Maadeed	200
AHDB Data Set	Al-Maadeed	100
ICDAR2011 competition Data set	Al-Maadeed, al	50

3 AVAILABLE ARABIC CORPUS

In this section, we present the findings of our research for the most relevant available Arabic corpora distributed into freely and commercially available corpora (Awdeh, 2019).

3.1 Freely Available Arabic Corpora

The 15 freely available raw texts, annotated and miscellaneous corpora which are cited below, are the most relevant Arabic corpus, and cover the news domain.

- Silver Arabic corpus SAC: collected between the years 2017 and 2020 from Alwatan Arabic newspapers by Awdeh and Abdallah. It covers around 20,291 articles, distributed into 6 categories (Culture, Religion, Economy, Local News, International News and Sports), and covers more than 18 million words (Awdeh, 2019).
- Adjir Corpora: collected between the years 2004 and 2005 from Arabic daily newspapers by Abdelali. It is stored in several text files which are compiled and cleaned up, and covers a minimum of 113 million words (Abdelali, 2005).
- King Saud University Corpus of Classical Arabic (KSUCCA): collected between seventh and eleventh century from classical Arabic texts dating by Alrabiah. It includes various categories of texts such as science, literature, religion, sociology, biography and linguistics, and covers a minimum of 50 million words (Alrabiah, 2013).
- Open Source Arabic Corpora: collected from several websites like CNN Arabic, BBC Arabic, and several other sources by Moataz Saad. It is stored in 22429 text documents, distributed into 10 categories (Economics, History, Entertainments, Education and Family, Religion, Sports, Health, Astronomy, Law, Stories and Cook recipes), and covers around 22 million words (Saad, 2010).
- Al Watan Corpus: collected in Oman from al watan newspaper articles by M. Abbas. It contains regarding 20000 articles, distributed into 6 categories (Culture, Religion, Economy, Local News, International News and sports), and covers about 10 million words (Abbas, 2011).
- Tashkeela Corpus: collected from Shamela Library and the freely revealed texts in ancient books largely from Islamic classical books by

Zarrouki. It had been manually rewritten and vocalized by volunteers, and covers about 75 million vocalized words (Zerrouki, 2017).

- Al Khaleej Corpus: collected from online newspaper "Akhbar El Khaleej" by M. Abbas. It contains more than five hundred articles, distributed into 3 categories (International and local news, Economy and sports), and covers about 3 million words (Abbas, 2005).
- King Abdulaziz City for Science and Technology Arabic Corpus: collected from a diversity of publishing media by Al-Thubaity and al. It contains more than 869800 files, distributed into several categories (manuscripts, newspapers, books, magazines, scientific periodicals, etc.), and covers more than 700 million words; 7,464,396 of which are unique (Al-Thubaity, 2015).
- Contemporary Arabic Corpus: collected between 1990 and 2004 from newspapers, emails and websites by Al-Sulaiti and Atwell. It is tagged in xml language and it covers more than 842,684 words (Al-Sulaiti,, 2005).
- Kalimat Corpus: collected from the Arabic newspaper Alwatan by el-haj and koulali, summed up into 2,057 multi document system summaries, NER annotated, POS tagged and full morphologically analyzed. It contains more than 20,291 articles, distributed into six categories (culture, economy, international news, local news, religion and sports), and covers about 18,167,183 million words (El Haj, 2013).
- SACS Corpus: collected from the proceedings of the Saudi Arabian National Computer Science Conference by Abu Salem. It covers 46,968 words tagged with title, authors, sources and abstract (Abu Salem).
- The International Corpus of Arabic: collected from electronic books, academic research papers, and articles of newspapers sites by Alansary. It contains 70,022 articles, distributed into eleven categories (strategic, national and social sciences, sports, religion, literature, bibliography and others), and covers more than 80 million words; 1,272,766 of which are unique (Alansary, 2014).
- Al-Raya Corpus: collected from the articles of Al-Raya newspaper by Hasnah. It contains about 187 articles and 219,978 words, over 30,096 of which are unique words (Hasnah, 1996).
- Arabic Modern Standard Corpus: collected from newspaper articles from different Arabic countries by Abdalali. It covers 102,134 articles with about 113 million words (Abdelali, 2005).
- University of Jordan Arabic Corpus: collected from 15 Arabic newspapers and other resources from 19 Arabic countries by researchers from Jordan University. It is tagged in XML, and contains 61,037 articles with 7,522,941 words, and over 70, 7385 of which are unique words (Hammo, 2013).

3.2 Commercially Available Arabic Corpora

The 5 monolingual text, and annotated corpora, which is cited below, are commercially Arabic corpus, and covers the news domain.

- LDC Corpus (Arabic Newswire): collected from the articles of the Agency France Press newswire published between 1994 and 2000 by Graff and Walker at the University of Pennsylvania's LDC. It covers more than 76 million words, 666,094 of which are unique, distributed into 383,872 files (Graff, 2001).
- An-Nahar Newspaper Text Corpus: collected from an-Nahar newspaper from 1995 to 2000, stored as hypertext Mark-up Language (HTML) files. It covers about 45 hundred articles and 24 million words (ELRA, 2001).
- Al-Hayat Arabic Corpus: collected from the al-Hayat Arabic newspaper. It contains 42,591 articles, distributed into several categories (General, Car, Computer, News, Economics, Science and Sport), and covers around 42,591 articles with 18,639,264 unique words (University Essex, 2001).
- Nemlar Corpus: collected from 13 different categories (political news, Islamic text, phrases of common words, broadcast news, business, Arabic literature, general news, interviews, scientific press, sports press, dictionary entries explanation and legal domain text) by Nemlar project. It is provided four versions: raw, fully vowelized, with Arabic lexical analysis, and with Arabic POS-tags, and covers more than 500000 words (ALP team, 2003).
- Arabic Gigaword Corpus: collected from four distinct Arabic newswire (Agency France Press, Al-hayat, Annahar and Xinhua news agency) by Graff. It is encoded with utf-8 and written in SGML, and covers about 1,256,719 articles words with 391619 words (Graff, 2003).

4 CORPUS RESOURCES

The Arabic corpus contains a diverse range of sources. Some of these sources are divided into other sub sources:

- Press: Arabic daily newspapers, magazine (general, specialized) and electronic press.
- Online article
- Books
- Academic source

In press, texts were collected from principal Arabic daily newspapers and magazines from different countries in the word (Table 2).

Table 2: The principal daily Arabic press in the word.

Country	Newspaper/Magazine
Arab newspapers/magazines published inside the Arab world	
Lebanon	Al-Akbar
	An-Nahar
Saudi Arabia	Al-Jazyra
	Hedaya
	Al-Watan
U.A.E "Dubai"	CNN
Oman	al watan newspaper
Qatar	Al-Rayah
Iraq	Kol El-Iraq
	Etthad
Tunisia	Assabah News Agency
Egypt	Ahram
	Akhbar
	Gomhoria
	Akhbar-elryadah
	Elssyasa Al-Dawli
Bahrain	Akhbar El Khaleej
Jordan	Addustoor
Palestine	Alhayat.Algadidah
	Al-Quds
	Al a'maan
Arab newspapers/magazines published outside the Arab world	
China	Xinhua news agency
France	Agency France Press
England	Al-Hayat
	Al-Quds Al-Arabic
	Asharq Al-Awsat

The Arabic corpus also contains a wide range of written categories. Some of these categories are divided into other subcategories (Table 3).

Table 3: List of categories and subcategories.

Categories	Subcategories
Literature	Poetry, novels, short stories, child stories
Religion	Islam, Christianity, other religions
Humanities	History, psychology, philosophy
Sciences	Biology, geology, physics , chemistry, economy, sociology, law, politics
Applied Sciences	Technology, engineering, agriculture
Biography	
Arts	
Sports	

5 DATA PREPROCESSING PIPELINE

Because the texts covered in a corpus will be edited with a word-processing program, they keep computerized texts in a file used via word-processing applications (such as documents with the extension.DOC in Microsoft Word).

Additionally, many Arabic corpuses were tagged into SGML (Standard Generalized Mark-up Language) like the Modern Standard Arabic Corpus and TREC corpora, XML (Extensible Mark-up Language) which is used in the LDC corpora, and/or GSON (Google Search Organization Network) like the Gold Arabic corpus GAC, to facilitate the utilization by other researchers and programs.

Also, the Arabic corpus transcoded with two or more separate encoding schemes, such as windows cp-1256 for Arabic language, and/or with UTF-8, would be of great benefit for researchers in the area of Arabic information retrieval, and Natural language Processing (NLP).

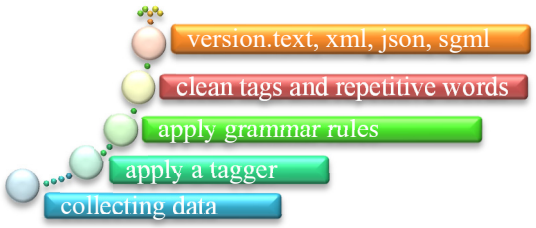

Figure 1: Corpus building steps.

Following data collection (Figure 1) from different resources, we have to compile and prepare it in order to build a corpus that can be used in NLP applications.

Many Arabic language researchers, universities and organizations are developing Arabic corpus utilizing several factors.

The first factor is the size of the corpus, the bigger corpus is more efficiently for study. The corpus sizes in different languages, for the same topics, still bigger than the Arabic corpus. The Arabic GigaWord corpus, which is generated by an institution like the LDC and cover a period of ten years is the largest commercially available corpus, with 3.3 million articles, and 1.077 billion words, while the largest free corpus available is KACST Corpus created by team from King Abdulaziz City for Science and Technology with about 1.5 million articles and 700 million words.

The second factor is the categories included in the corpus. Most of the cited corpus covers multiple categories that make it well representative.

The third factor is the price of corpus. The commercial corpus is difficult to reach by the Arabic linguistic research. So, our aim is to survey a wide range of freely available Arabic corpus in this paper.
The last factor is the structure of corpus. The Arabic language has a complex morphology, that's why a well-structured corpus is efficient. Unfortunately, a few of the available corpora are tagged in XML and/or SGML, and/or GSON.

6 CONCLUSION

In order to solve problems related to the lack of available Arabic corpora, we have presented in our study the survey of available Arabic corpus distributed into free and commercial Arabic corpus, their categories, subcategories, the different resources and the steps needed for building an Arabic corpus.

Our purpose is to encourage the usage of free available corpora, in particular for those who lack financing and cannot afford membership or high fees to purchase an Arabic corpus from a language data center.

Additionally, we listed the types of Arabic corpora and the distribution about 60 Arabic corpora according to their types to help advancing works on Arabic NLP researchers in the field of evaluation and validation of their unsupervised learning tools and for the learning in the supervised learning tools in the syntax domain.

In our forthcoming researches, we hope to enrich our Silver Arabic Standard Corpus (Awdeh, 2019) and build a new Gold Arabic Corpus GAC and then to create a frame work that can facilitate the use of our corpus (SAC). Eventually, we are very open to any suggestions regarding the corpus by our colleagues.

ACKNOWLEDGEMENTS

This work has been done as a part of the project "Analyse sémantique de textes Arabes utilisant l'ontologie et WordNet", supported by the Lebanese University and Paris 8 University.

REFERENCES

Abbas, M., Smaili, M., and Berkani, D. (2011). Al Watan Corpus. In Evaluation of Topic Identification Methods on Arabic Corpora.

Abbas, M., Smaili, M., and Berkani, D. (2005). Al Khaleej Corpus. *In the Proceedings of International Conference* on Recent Advances in Natural Language Processing, (RANLP.).

Abdelali, A. (2005). Adjir Corpora, on http://aracorpus.e3rab.com.

Abdelali, A., Cowie, J., and Soliman, H. (2005). Arabic Modern Standard Corpus. *In the workshop on computational modelling of lexical acquisition*. Croatia.

Abu Salem, H., SACS Corpus. *In Saudi Arabian National Computer Science Conference*. Saudi Arabian.

Alansary, S., Nagi, M. (2014). The International Corpus of Arabic ICA. *In the International Corpus of Arabic: Compilation, Analysis and Evaluation*.

ALP team (2003). Nemlar Corpus. *In European Language Resources Association, ELRA Catalogue number ELRAW0042* on http://catalog.elra.info/product info.php?products?id=873.

Alrabiah, M., Salman, A., and Atwell, A. (2013). King Saud University Corpus of Classical Arabic. In Proceedings of WACL'2 Second Workshop on Arabic Corpus Linguistics. Lancaster University, UK.

Al-Thubaity, A. (2015). King Abdulaziz City for Science and Technology Arabic Corpus KACST. *In Journal Language Resources and Evaluation* on Springer-Verlag New York.

Al-Sulaiti, L., Atwell E. (2005). Contemporary Arabic Corpus CAC. *In the Proceedings of the CL, Corpus Linguistics Conference*.

Awdeh,, H., Abdallah, A., Gernard, G., Hajjar, M. (2019). A Silver Standard Arabic Corpus for Segmentation and Validation SAC. *In the international conference on Big Data and Cyber Security BDCSIntell'2019* on the University of Versailles Saint-Quentin-en-Yveline. France.

Buckwalter, J. (2002). Clinical Orthopaedics and Related Research®. *In the Clinical Orthopaedics and Related Research*.

El Haj, M., Koulali, R. (2013). Kalimat a Multipurpose Arabic Corpus. *In the Second Workshop* on Arabic Corpus Linguistics (WACL-2).

ELRA (2001). An-Nahar Newspaper Text Corpus. *In Language Resources Association* on

http://catalog.elra.info/productinfo.php?products?id=767. Europe.

Graff, D. Chen, K., Kong, J., and Maeda, K. (2015). Arabic Gigaword Second Edition. *In Linguistic Data Consortium,* Philadelphia.

Graff, D. Chen, K., Kong, J., and Maeda, K. (2011). Arabic Gigaword Fifth Edition. *In Linguistic Data Consortium LDC*. Philadelphia.

Graff, D. (2007). Arabic Gigaword Third Edition. *In Linguistic Data Consortium, Philadelphia, LDC catalog number LDC2007T40* on https://catalog.ldc.upenn.edu/LDC2007T4.

Graff, D. Chen, K., Kong, J., and Maeda, K. (2009). Arabic Gigaword Fourth Edition. *In Linguistic Data Consortium LDC catalog number LDC2009T30* on https://catalog.ldc.upenn.edu/LDC2009T30. Philadelphia.

Graff, D. (2003), Arabic Gigaword. *In Linguistic Data Consortium, LDC catalog number LDC2003T12* on https://catalog.ldc.upenn.edu/LDC2003T12. Philadelphia.

Graff, D., Walker, K. (2001). LDC Corpus. *In Arabic newswire part 1. Linguistic Data Consortium*, on https://catalog.ldc.upenn.edu/LDC2001T55. Philadelphia.

Hammo, B., Al-Shargi, F., Yagi, S., Obeid, N. (2013). University of Jordan Arabic Corpus UJAC. *In the Second Workshop* on Arabic Corpus Linguistics (WACL-2). UK.

Hasnah, A. (1996). Al-Raya Corpus. *In full Text Processing and Retrieval: Weight Ranking, Text Structuring, and Passage Retrieval for Arabic Document. Ph.d.* on Dissertation, Illinois Institute of Technology.

Mansour, M. (2013). The absence of Arabic corpus linguistics: a call for creating an Arabic national corpus. *International Journal of Humanities and Social Science, 3(12).*

Saad, M., Ashour, W. (2010). Open Source Arabic Corpora OSA. *In 6th International Conference* on Electrical and Computer Systems (EECS'10).

University Essex (2001). Al-Hayat Arabic Corpus. *In European Language Resources Association, ELRA Catalog number ELRA W0030* on http://catalog.elra.info/product2info.php?products ?id=632.

Zerrouki, T., Balla, A. (2017). Tashkeela: Novel corpus of Arabic vocalized texts, data for auto-diacritization systems. *In the National Computer Science Engineering School (ESI)*. Algiers.