# A Silver Standard Arabic Corpus for Segmentation and Validation

Hussein Awdeh, Adelle Abdallah, Gilles Bernard
LIASD
University of Paris 8
2 rue de la Liberté, 93526 Saint-Denis cedex – France
hussein.awdi85@gmail.com, adelle.abdallah@gmail.com,
gilles.bernard@iedparis8.net

Mazen El Sayed, Mohammad Hajjar
Faculty of Technology
Lebanese University
Hisbeh Street - Saida – Lebanon
Mazen_elsayed@yahoo.fr
mohammadhajjar@ul.edu.lb

***Abstract*—The Arabic Natural Language Processing applications suffer from the deficiency of both Arabic corpus and gold standard corpus. Defined as a collection of written or spoken texts stored on a computer, a corpus is written either in a single language, Monolingual Corpus or in several languages, Multilingual Corpus. A corpus is considered as the most important sources for semantic and syntaxic analysis in the domain of natural language processing.**

**Our study aims to build a New Silver Arabic Corpus collected from a set of Newspaper Articles morphologically analyzed. It contains 18,167,183 words in total incorporating six categories, Religion, Economy, Culture, Sports, Local and International News. It is encoded namely in UTF-8 encoding and XML. This silver corpus can be used as an accurate reference for validation and learning in the syntaxic analysis mainly for the word segmentation and part of speech tagging.**

***Keywords—Arabic Language, Arabic Natural Language Process, Validation, Information Retrieval, Silver standard corpus.***

## I. INTRODUCTION

Being one of the six official languages of the united nations ever since 1973, the Arabic language lies among the world's six major ones. It is the language of the Holy Quran and spoken by more than 273 million people around the world. Arabic dialects consist of several branches: the classical (Language of Quran), the modern standard (used in newspapers, books...) and the local dialects (vary considerably among different countries).

Despite all the above and the increasing need for Arabic corpus, the Arabic corpus remains deficient to support various Arabic linguistic researches. For instance, the majority of Arabic corpora are limited in sources, types, and genres or even not freely available.

Though it is a major world language, it still under-represented in corpus linguistic. Due to this lack, one of the aims of this paper is to build a new free Arabic Multipurpose corpus which we call "Silver Multipurpose Arabic Corpus" with a large size, collected over many years from multiple sources, covering all types and genres. It will be freely available for the researcher working on different Arabic Natural Language Processing techniques, and mainly for the validation and evaluation in the unsupervised methods, and in the domain of learning for the supervised methods.

## II. RELATED WORKS

This part of paper reviews the existing Arabic Corpora as we pass by the different types of the textual Arabic corpus.

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, POS tagging, semantic and syntactic information.
- Lexicon: words lists and lexical database.
- Miscellaneous corpora: multipurpose corpus (Q/A, summaries…)

Our study will be limited to textual monolingual corpora, divided into freely and commercially available corpora.

### *A. Freely available arabic corpora*

In this section we cite 14 freely available raw text, annotated and miscellaneous corpora. They cover mostly the news domain. We list it according to their size.

1- Adjir Corpora [1]: made by Abdelali, collected from Arabic daily newspapers distributed between the years 2004 and 2005. It's distributed in multiple text files. Every file is compiled and cleansed. It contains regarding 113 million words in total.
2- King Saud University Corpus of Classical Arabic (KSUCCA) [2]: made by Alrabiah, collected from classical Arabic texts dating between seventh and eleventh century covering completely different genres of texts like science, religion, literature, sociology, biography and linguistic. Every genre is split into sub genres and contains in total regarding fifty million words.
3- Tashkeela Corpus [3]: made by Zarrouki, collected from freely revealed texts in ancient books largely from Islamic classical books (while the foremost half is collected from Shamela Library) and therefore the Modern Standard Arabic texts crawled from the Internet using web crawling process. The books had been rewritten and vocalized by volunteers manually, to confirm that words are vocalized. It contains seventy five million of fully vocalized words.
4- Open Source Arabic Corpora (OSAC) [4]: made by Moataz Saad, collected from multiple websites like BBC Arabic, CNN Arabic and several other sources. It includes 22429 text documents divided into ten categories like Economics, History,

Copyright © 2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR-WS.org/Vol-2622/paper12.pdf

Entertainments, Education and Family, Religion, Sports, Health, Astronomy, Law, Stories and cook recipes, and they were encoded with UTF-8. It contains regarding twenty two million words.

5- Al Watan Corpus [5]: made by M. Abbas, collected from al watan newspaper articles in Oman. It contains regarding 20000 articles talking about the six following topics "categories": Culture, Religion, Economy, Local News, International News and sports. It contains about 10 million words.
6- Al Khaleej Corpus [6]: made by M. Abbas, collected from thousands of articles, which had downloaded from an online newspaper "Akhbar El Khaleej". This corpus contains more than five hundred articles which correspond to nearly 3 million words talking about International and local news, Economy and sports.
7- King Abdulaziz City for Science and Technology Arabic Corpus (KACST) [7]: made by Al-Thubaity et al, collected from a diversity of publishing media, such as manuscripts, newspapers, books, magazines, scientific periodicals, etc. The corpus contains seven hundred million words beginning from the pre-Islamic era to the modern era. It contains about 869800 files, with 732,780,509 words, out of which, 7,464,396 are unique.
8- Kalimat Corpus [8]: made by el-haj and koulali is a miscellaneous corpus, collected from the Arabic newspaper Alwatan, summarized into 2,057 multi document system summaries, NER annotated, POS tagged and full morphologically analyzed. It contains about 20,291 articles, with 18,167,183 words and fall into six categories (culture, economy, international news, local news, religion and sports).
9- SACS Corpus (Saudi Arabian National Computer Science Conference) [9]: created by Abu Salem, collected from the proceedings of the Saudi Arabian National Computer Science Conference. It contains 46,968 words tagged with title, authors, sources and abstract.
10- Al-Raya Corpus [10]: made by Hasnah, collected from the articles of Al-Raya newspaper. It contains 187 articles and 219,978 words over 30,096 unique words.
11- Contemporary Arabic Corpus (CAC) [11]: made by Al-Sulaiti and Atwell, collected from newspapers, emails and websites from 1990 to 2004. It contains 842,684 words and balanced in topics. It is tagged in xml language.
12- The International Corpus of Arabic (ICA) [12]: made by Alansary, nagi and adly, collected from the articles of Arabic newspapers sites, blogs and forums, electronic books and academic research papers and dissertations. It contains 70,022 articles with 80 million words over about 1,272,766 unique words. It includes texts in eleven categories like strategic, national and social sciences, sports, religion, literature, bibliography and others.
13- Arabic Modern Standard Corpus [13]: made by Abdalali collected from newspaper articles from different Arabic countries. It contains 102,134 articles with about 113 million words.
14- University of Jordan Arabic Corpus (UJAC) [14]: made by researchers from Jordan University, collected from 15 Arabic newspapers and other resources from 19 arabic countries. It contains 61,037 articles with 7,522,941 words over 70, 7385 unique words. It is tagged in XML.

### *B. Commercially available arabic corpora*

In this section we cite 5 commercially available corpora. They are monolingual text corpora and annotated corpora. They cover the news domain. We list it according to their size.

1- LDC Corpus (Arabic Newswire) [15]: made by Graff and walker at the University of Pennsylvania's LDC, collected from the articles of the Agency France Press newswire published between 1994 and 2000. It contains 383,872 documents containing 76 million tokens over 666,094 unique words.
2- An-Nahar Newspaper Text Corpus [16]: collected from an-Nahar newspaper from 1995 to 2000, stored as hypertext Mark-up Language (HTML) files. It contains 45000 articles and 24 million words. Each article includes information such as title, type, date, country and page.
3- Al-Hayat Arabic Corpus [17]: made by al-Hayat Arabic corpus, collected from the al-Hayat Arabic newspaper and organized into seven domains in accordance with al-Hayat's subject tags: General, car, computer, news, economics, science and sport. The corpus is cleaned by removing the mark-ups, numbers, special characters and punctuation. It contains about 42,591 articles with 18,639,264 unique words.
4- Nemlar Corpus [18]: it is produced by the Nemlar project. It contains about 500000 words collected from 13 different categories such as political news and debate, 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. It is provides in four versions: raw, fully vowelized, with Arabic lexical analysis and with Arabic POS-tags.
5- Arabic Gigaword Corpus [19]: made by Graff. It is collected from four distinct Arabic newswire (Agency France Press, Al-hayat, Annahar and Xinhua news agency). It contains 1,256,719 articles and 391619 Kwords. The corpus was encoded with utf-8 and written in SGML.

TABLE I. AVAILABLE ARABIC CORPORA

| Author | Corpus | Words | Category | Sources | F/C? |
|---|---|---|---|---|---|
| Abdelali [1] | Adjir Corpora | 113,000,000 | monolingual text corpus | Arabic daily newspapers distributed between the years 2004 and 2005 | F |
| Alrabiah, Salman et Atwell [2] | King Saud University Corpus of Classical Arabic (KSUCCA) | 50,000,000 | monolingual text corpus | classical Arabic texts dating between seventh and eleventh century | F |
| Zarrouki, Balla [3] | Tashkeela Corpus | 75,000,000(fully vocalized words) | monolingual text corpus | Islamic classical books from Shamela Library, Modern Standard Arabic texts crawled from the Internet | F |
| Saad, Ashour [4] | Open source Arabic Copora (OSAC) | 22,000,000 | monolingual text corpus | websites such as BBC Arabic, CNN Arabic and several other sources | F |
| Abbas, Smaili et Berkani [5] | Al Watan Corpus | 10,000,000 | monolingual text corpus | al watan newspaper articles in Oman | F |
| Abbas, Smaili et Berkani [6] | Al Khaleej Corpus | 3,000,000 | monolingual text corpus | online newspaper “Akhbar El Khaleej” | F |
| Al-Thubaity [7] | King Abdulaziz City for Science and Technology Arabic Corpus | 732,780,509 | online searchable corpus | publishing media, such as manuscripts, newspapers, books, magazines, scientific periodicals | F |
| El-Haj, Koulali [8] | Kalimat Corpus | 18,167,183 | Multipurpose Arabic Corpora | Arabic newspaper Alwatan | F |
| Abu Salem [9] | SACS Corpus | 46,968 | monolingual text corpus | proceedings of the Saudi Arabian National Computer Science Conference | F |
| Hasnah [10] | Al-Raya Corpus Arabic | 219,978 | monolingual text corpus | the articles of Al-Raya newspaper | F |
| Al-Suleiti, Atwell [11] | Contemporary Arabic Corpus | 842,684 | monolingual text corpus | newspapers, emails and websites from 1990 to 2004 | F |
| Alansary, et Nagi [12] | The International Corpus of Arabic (ICA) | 80,000,000 | Online searchable corpus | articles of Arabic newspapers sites, blogs and forums, electronic books and academic research papers and dissertations | F |
| Abdalali, Cowi et Soliman [13] | Arabic Modern Standard Corpus | 113,000,000 | monolingual text corpus | newspaper articles from different Arabic countries | F |
| Hammo, Al-Shargi, Yagi et Obeid [14] | University of Jordan Arabic Corpus (UJAC) | 7, 522,941 | monolingual text corpus | 15 Arabic newspapers and other resources from 19 arabic countries | F |
| Graff, Walker [15] | LDC Corpus | 76,000,000 | monolingual text corpus | the articles of the Agency France Press newswire published between 1994 and 2000 | C |
| ELRA [16] | An-Nahar Newspaper Text Corpus | 144,000,000 | monolingual text corpus | an-Nahar newspaper from 1995 to 2000 | C |
| University Essex [17] | Al-Hayat Arabic Corpus | 18,639,624 | monolingual text corpus | al-Hayat Arabic newspaper | C |
| ALP team [18] | Nemlar Corpus | 500,000 | monolingual text corpus | political news and debate, 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 | C |
| Graff [19] | Arabic Gigaword Corpus first edition | 391,619 | monolingual text corpus | four distinct Arabic newswire (Agency France Press, Al-hayat, Annahar and Xinhua news agency) | C |
| Graff, Chen, Kong et Maeda [21] | ArabicGigaword Corpus 2 edition | 481,906 | monolingual text corpus | | |
| Graff [20] | Arabic Gigaword Corpus 3 edition | 576,799 | monolingual text corpus | | |
| Parker et al. | Arabic Gigaword | 848,469 | monolingual | | C |

| Author | Corpus | Words | Category | Sources | F/C? |
|---|---|---|---|---|---|
| | Corpus fourth edition | | text corpus | | |
| Parker et al. | Arabic Gigaword Corpus five edition | 1,077,382,000 | monolingual text corpus | | |

## III. SILVER ARABIC CORPUS (SAC)

Despite of the Arabic language corpora cited above, there is serious lack of high quality Arabic Corpora compared to other languages like English.

The existing corpora still has many limitations. Few numbers of corpora are freely available with huge lack of freely tagged corpora. And the tagged corpora doesn't fit our need for word segmentation evaluation. Moreover, existing corpora are narrowly small in size, limited in source types and genres. And most of these trials don't reach the increasing need to have a reliable, updateable and well-structured corpus.

To overcome these problems, it was necessary to build a new free, tagged and reliable corpus named SAC (Silver Arabic Corpus) for Arabic word segmentation evaluation.

Our goal is to build a gold arabic corpus serving as a language resource for NLP researchers in the field of evaluation for syntaxic analysis purpose. We began with building our silver Arabic Corpus called "**SAC**". It consists of a collection of texts annotated and enriched with linguistic information. In our work we used the Arabic Stanford POS tagger [28], our arabic grammar rules based tool for word segmented and our xml annotator.

### *A. SAC Building steps*

#### 1- Collecting web documents

The first version of SAC consists of arabic articles collected from alwatan arabic newspapers (based on alkalimat corpus) corrected and cleaned manually and covers six topics such as culture, economy, religion, sports, local and international news.

We use the alkalimat corpus as reference for many reasons. It is a multipurpose corpus like our corpus, tagged, compiled and freely available. But it is developped for specific targets such as document system summaries, NER annotated and POS tagged and its structure differ from our structure to use for word segmentation. Although, we notice some words not compiled and untagged. The table below shows the statistics.

TABLE II. SILVER ARABIC CORPUS CATEGORIES

| Categories | Words Count | Articles Count |
|---|---|---|
| International News | 855,945 | 2,035 |
| Religion | 1,555,635 | 3,860 |
| Culture | 1,359,210 | 2,782 |
| Local news | 1,460,462 | 3,596 |
| Sports | 9,813,366 | 4,550 |
| Economy | 3,122,565 | 3,468 |
| **Total** | **18,167,183** | **20,291** |

F: Free, C: Commercial

#### 2- Creation process steps

The process of creating SAC was applied to the entire data.

Firstly, we need to know the POS of each word, thus we apply Stanford POS tagger to the document collection. The Stanford POS Tagger is a java implementation of a log-linear part-of-speech tagger. it is a supervised system depending on different trained models for many languages including Arabic. The accurate model for Arabic was trained using the Arabic Tree-bank p1-3 corpus based on maximum entropy. The POS tagger identifies 33 part of speeches such as: Noun (NN), Plural Noun (NNS), Proper Noun (NNP), Verb (VB), Adjective (JJ). The tagger reached an accuracy of 96.50% [29].

Secondly, we clean the POS tagged document by removing the repetitive words.

Finally, we apply our rule system based on the POS tagged corpus, to make our word segmentation by using our java tool and set of grammar rules [27].

We classify the word segmentation rules system based on [27][31][3] into three main classes:

- Rules based on punctuation marks.
- Rules based on coordinating conjunctions.
- Rules based on certain connector words like (لكن, اذا)

### *B. Arabic word segmentation rules [27][31][3]*

1. Prepositions that are prefixed to the word like (ب, ل) will be separated from the word during segmentation, as the example:

   بكتاب = ب كتاب

2. The definition Lam will be separated from the word during segmentation As the following example:

   لمعرفٌ = ل معرفٌ

3. The linked pronouns will be separated from the word during segmentation, as the following example:

   كتابكما = كتاب كما

4. For the noun, the Alf noun mosana (ان المثنى) , Ouaou Jamaat (ون الجماعة), At (ات), (except ta'ta2nith ة) will be separated from the word during segmentation as the following examples:

   معلمة = معلمة

   معلمان = معلم ان

   معلمتان = معلمة ان

   معلمون = معلم ون

   معلمات = معلم ات

5. For the verbs that are conjugated in the past, the Ta Ta"nith (تاء التأنيث) , Ouaou Jamaat (واو الجماعة), Noun nissoua (نون النسوة), will be separated from the word during segmentation as the following examples:

لعبت = لعب ت

ذهبوا = ذهب وا

ذهبن = ذهب ن

6. For the verbs that are conjugated in the present, the Ya (ي) , Ta (ت), Alf noun mosana (ان المثنى) Ouaou Jamaat(ون الجماعة), will be separated from the word during segmentation as the following examples:

يلعب = ي لعب

تلعب = ت لعب

يذهبان = ي ذهب ان

يذهبون = ي ذهب ون

Finally, we clean the POS tagged document by removing the tags data and the repetitive words, then we apply our xml annotator.

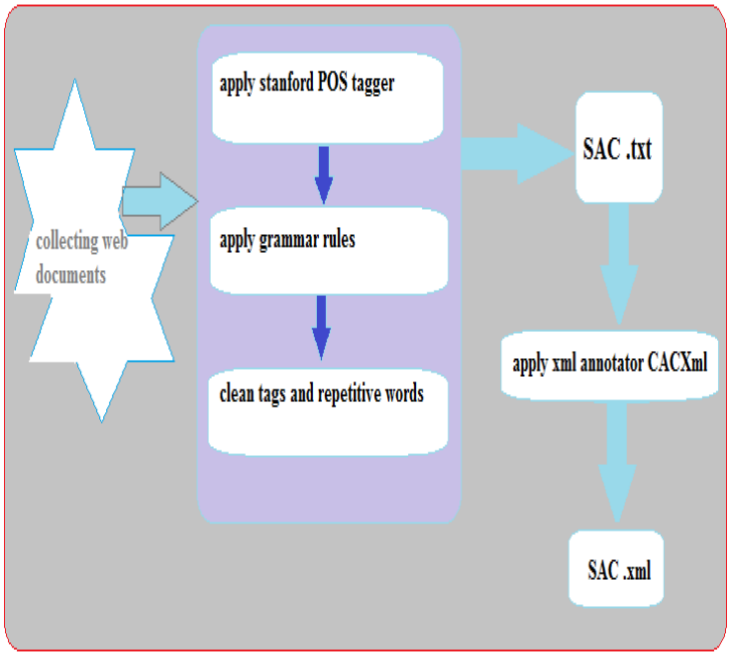
### C. *XML annotation*

We need to have the data in a specific format, thus the last task is adding the linguistic information to the compiled corpus. We developed our own tool **CACXml** (java based tool )that converts the corpus into xml tagged corpus.

As a word consists of a sequence of morphemes in the pattern **prefix*-stem-suffix*,** we generate our xml structure to suit this sequence of morphemes:

```
<Seg>
<Word>w</Word>
<Prefixes>
    <Prefix>p1</Prefix>
    <Prefix>p1</Prefix>..
</Prefixes>
<Stem>stm</Stem>
<Suffixes>
    < Suffix>s1</ Suffix>
    < Suffix>s1</ Suffix>..
</Suffixes >
</Seg>
```

The developed tagged corpus contains four fields: the field for the word by itself and three fields for the morphemes of the words: prefix, stem and suffix.




SCHEMA I. SAC BUILDING STEPS

## IV. ANAYSIS AND COMPARISON

Building a corpus is not limited to a data collection process but it consists to compile and prepare this data to be used in different natural language processing applications. In the last decade, the Arabic language researchers gained more interests. Several universities and organizations builder Arabic corpora using multiple factors.

A standard arabic corpus was the aim of this paper. This corpus must be in large size, representative covering different text genres, and with a specific new structure used by Arabic NLP applications.

The first factor to consider in building a corpus is the size of this corpus. Being a multipurpose corpus, it is preferable to be a long corpus covering a large features of linguistic information. The corpus sizes in different languages, for the same topics, still bigger than the arabic corpus. The largest freely corpus is the KACST Corpus (Al-Thubaity, 2014) having 700 million words with about 1.5 million articles extracted from ten different sources covering the pre-Islamic era to the present day (a period covering more than 1,500 years). It mainly contains both classical and modern arabic. The largest commercially is the Arabic GigaWord corpus created by an institution like the LDC covering a period of ten years. It has 3.3 million articles, and 1.077 billion words. Our corpus is 18,167,183 words with 20,291 articles covering periods.

The second factor is the topics included in the corpus. Most of the cited corpus cover multiple topics as well as our corpus that covers the mainly topics that makes it representative. It is multitopic corpus covering different categories like culture, economy, religion, sports, local and international news.

The third factor is the price of corpus. The commercial corpus are difficult to reach by the arabic linguistic research. So, our aim was to build a freely available corpus.

The last factor is the structure of corpus. The arabic language has a complex morphology, that's why we need a

well-structured corpus. Most of the available corpora are not tagged, and the tagged corpora do not meet our needs. So we make our own structure for the segmentation of arabic word. In such a way, the representation of SAC in xml format makes it easily exploitable for any syntaxic program.

## V. Conclusion

In order to resolve problems related to the lack of Arabic tagged corpora, this work presents a Silver Standard Corpus (SAC) building process. Thus, we provide the built SAC for free to help advancing the work 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 syntaxic domain. We provide Silver Arabic Corpus for free including the articles, annotated text, entities and summaries to help advancing the work on Arabic NLP. The corpus can be downloaded directly from: https://gitlab.com/Data-Liasd-papers/silver-arabic-corpus. The Silver Arabic Corpus (SAC) can be used by researchers as standards corpus and or baselines to test and evaluate their Arabic tools. In addition, we welcome any amendments to the corpus by other researchers.

In our future researches, we expect to enrich the built Silver Standard Corpus (SAC) to support more words syntaxic information and use an xml tagger to build our Gold Arabic Corpus GAC..

## 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.

## References

[1] A. Abdelali, "Adjir Corpora,", http://aracorpus.e3rab.com/, 2005.

[2] M. Alrabiah, A. Salman and E. Atwell, "King Saud University Corpus of Classical Arabic KSUCCA," In Proceedings of WACL'2 Second Workshop on Arabic Corpus Linguistics, Lancaster University, UK. The University of Leeds , 2013.

[3] T. Zerrouki, A. Balla, "Tashkeela: Novel corpus of Arabic vocalized texts, data for auto-diacritization systems," The National Computer Science Engineering School (ESI), Algiers, Algeria, January 2017.

[4] M. Saad, W. Ashour, "Open Source Arabic Corpora OSAC," 6th ArchEng International Symposiums, 6th International Conference on Electrical and Computer Systems (EECS'10), Nov 25-26, 2010.

[5] M. Abbas, K. Smaili and D. Berkani, "Al Watan Corpus," Evaluation of Topic Identification Methods on Arabic Corpora. JDIM, 9(5), 185-192, 2011.

[6] M. Abbas, K. Smaili and D. Berkani, "Al Khaleej Corpus," Comparison of topic identification methods for Arabic language.Paper presented at the Proceedings of International Conference on Recent Advances in Natural Language Processing, (RANLP.), 2005.

[7] A. Al-Thubaity, "King Abdulaziz City for Science and Technology Arabic Corpus KACST," In Journal Language Resources and Evaluation Volume 49 Issue 3 Pages 721-751, Springer-Verlag New York, September 2015.

[8] M. El Haj, R. Koulali, "Kalimat a Multipurpose Arabic Corpus," at the Second Workshop on Arabic Corpus Linguistics (WACL-2), 2013.

[9] H. Abu Salem, "SACS Corpus," Saudi Arabian National Computer Science Conference.

[10] A. Hasnah, "Al-Raya Corpus," Full Text Processing and Retrieval: Weight Ranking, Text Structuring, and Passage Retrieval for Arabic Documents. Ph.D. Dissertation, Illinois Institute of Technology, 1996.

[11] L. Al-Sulaiti, E. Atwell, "Contemporary Arabic Corpus CAC," In the Proceedings of the CL, Corpus Linguistics Conference, 2005.

[12] S. Alansary, M. Nagi, "The International Corpus of Arabic ICA," The International Corpus of Arabic: Compilation, Analysis and Evaluation. ANLP, 2014.

[13] A. Abdelali, J. Cowie and H. Soliman, "Arabic Modern Standard Corpus," In the workshop on computational modeling of lexical acquisition, the split meeting. Croatia, July 2005.

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

[15] D. Graff, K. Walker, "LDC Corpus," Arabic newswire part 1. Linguistic Data Consortium, Philadelphia. LDC catalog number LDC2001T55, from: https://catalog.ldc.upenn.edu/LDC2001T55, 2001.

[16] ELRA, "An-Nahar Newspaper Text Corpus," European Language Resources Association, ELRA Catalog number ELRA-W0027, from: http://catalog.elra.info/product_info.php?products_id=767, 2001.

[17] University Essex, "Al-Hayat Arabic Corpus," European Language Resources Association, ELRA Catalog number ELRA-W0030, from: http://catalog.elra.info/product_info.php?products_id=632, 2001.

[18] ALP team, "Nemlar Corpus," European Language Resources Association, ELRA Cat-alog number ELRA-W0042, retrieved on, from:http://catalog.elra.info/product_info.php?prod-ucts_id=873, 2003.

[19] D. Graff, "Arabic Gigaword," Linguistic Data Consortium, Philadelphia, LDC catalog number LDC2003T12, retrieved on: 10/25/2015,from: https://catalog.ldc.upenn.edu/LDC2003T12, 2003.

[20] D. Graff, "Arabic Gigaword Third," Edition. Linguistic Data Consortium, Philadel-phia, LDC catalog number LDC2007T40, from: https://catalog.ldc.upenn.edu/LDC2007T4017, 2007.

[21] D. Graff, K. Chen, J. Kong, and K. Maeda, "Arabic Gigaword Second Edition," Linguistic Data Consortium, Philadelphia. LDC catalog number LDC2006T02., retrieved on: 10/25/2015, from: https://catalog.ldc.upenn.edu/LDC2006T0218, 2007.

[22] W. Zaghouani, "Critical Survey of the Freely Available Arabic Corpora," In Carnegie Mellon University Qatar Computer Science, Workshop on Free/Open-Source Arabic Corpora and Corpora Processing Tools Conference, 2014.

[23] D. Namly, R. Tajmout, K. Bouzoubaa and L. Abouenour, "A Gold Standard Corpus for Arabic Stemmers Evaluation," 28th IBIMA Conference,Seville, Spain, Nov 2016.

[24] A. Abdelali, J. Cowie, H. Soliman, "Building a modern standard Arabic corpus," Workshop on Computational Modeling of Lexical Acquisition. The Split Meeting, Croatia, 25th to 28th of july 2005.

[25] D. Mostefa, J. Abualasal, M. Gzawi, O. Asbayou and R. Abbes, "a hybrid Arabic Error Correction System," The Second Workshop on Arabic Natural Language ProcessingAssociation for Computational Linguistics, Beijing, China, July 2015.

[26] I. Zeroual, A. Lakhouaja, "Arabic Corpus Linguistics: Major Progress, but Still a Long Way to Go," In Shaalan K., Hassanien A., Tolba F. (eds) Intelligent Natural Language Processing: Trends and Applications. Studies in Computational Intelligence, vol 740. Springer, Cham, 2018.

[27] H. Awdeh, A. abdallah, "Guide de segmentation des mots Arabes", https://gitlab.com/Data-Liasd-papers/guide-de-segmentation, 2018.

[28] K. Toutanova, D. Klein et C. Manning, "Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger," In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000), pp. 63-70, 2000.

[29] K. Toutanova, D. Klein et C. Manning, and Y Singer. "Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network," In Proceedings of HLT-NAACL 2003, pp. 252-259, 2003.

[30] M. Mansour, "The Absence of Arabic Corpus Linguistics: A Call for Creating an Arabic National Corpus", International Journal of Humanities and Social Science, June 2013.

[31] M. Altantawy, N. Habash, O. Rombow, I. Saleh, "Morphological Analysis and Generation of Arabic Nouns: A Morphemic functional Approach Handbook of Natural Language Processing" Second Edition, Center for Computational Learning systems, Culombia University, New York, USA.

[32] A. S. Talha ibnu William, "Les règles de grammaire", du premier Livre de Médine: Prentice Hall, 2008