Proc. of the International Conference on Electrical, Computer and Energy Technologies (ICECET)
9-10 December 2021, Cape Town-South Africa

A Gold Multipurpose Arabic Corpus (GAC)

Hussein Awdeh, Adelle Abdallah, Youssef Zaki, 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,
youssefzaki@gmail.com, gilles.bernard@iedparis8.net

Mohammad Hajjar
Faculty of Technology
Lebanese University
Hisbeh Street - Saida – Lebanon
mohammadhajjar@ul.edu.lb

2021 International Conference on Electrical, Computer and Energy Technologies (ICECET) | 978-1-6654-4231-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICECET52533.2021.9698621

***Abstract*—A corpus is a large collection of spoken or written one or more language, it is collected from different resources, then structured, stored, and treated automatically by special algorithm to reach the desired goal, this is why it is used by researchers in different domain such as grammar, semantics, lexicography, natural Language Processing and other language studies. Therefore, building a corpus was still a challenge for many researchers in those fields for many years. Our study in this paper aims to build a new gold standard Arabic corpus due to the lack of successful trials in compiling Arabic corpora. The corpus produced by our team, is a text corpus, collected from a set of Arabic Newspaper articles morphologically analyzed from eight Arabic countries, and it contains more than 18 million words in total, covering six categories (Religion, Economy, Culture, Sports, Local and International News). It was encoded with UTF-8 encoding and marked with two mark-up languages: JSON and XML. In the hope that this corpus can be used as an accurate reference for segmentation and validation and learning in the syntax 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

Arabic is one of the world's six main languages since 1972. More than 273 million people speak Arabic, and it is the language of the Holy Quran. There are three forms for Arabic dialects: the colloquial Arabic (or al-'ammiyya) which is widespread among different countries, the classical Arabic which is the language of the Quran and the modern standard (or al-fusha) used in newspapers and books.

Recently, the Natural language processing (NLP), including Information Retrieval, Machine Translation and other Natural Language-related disciplines, is showing more interest in the Arabic language [20]. But the Arabic corpus is still deficient to support a large variety of Arabic linguistic researches. Actually, most of the Arabic corpora are limited in sources, types, genres, and not freely.

Although Arabic is one of the most widespread languages in the world, it is still underrepresented in linguistic corpora, which makes it more difficult for researchers worldwide to build it as a corpus.

Due to these deficiencies, the whole point of this study is to create a new free Arabic Multipurpose Corpus, called "Gold Arabic Corpus GAC", collected over several years from various resources, with a large size, and covering different categories and types. GAC will be available for free to help researchers in Arabic NLP, supervised learning, and unsupervised algorithms in order to evaluate them and validate their results. The remainder of this paper is organized as follows: Section 2 discusses the most relevant available free and commercial Arabic corpora. Then, in Section 3, we list the types of the text. Section 4 provides our gold Arabic corpus and its characteristics. Section 5 compares our corpus and other corpora. Section 6 describes the different steps for creation a corpus.

II. RELATED WORKS

Hereunder is the summary of the available Arabic corpora including the different types of the textual Arabic corpus.

The Arabic corpus is divided into two sections, freely and commercially available corpora. Most of the existing corpora are relatively limited in categories, small in size, expensive, and require further researches and studies.

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…).

Table 1 shows the result of our survey conducted on raw text available for free, annotated and miscellaneous corpora according to their size, categories and sources, focusing on the most important work about 15 free corpuses related to the categories. In addition, it shows nine commercially monolingual text corpora and annotated available corpora, that are covered mostly the Arabic newspapers. More details about corpus can be found in a previous article entitled "A Silver Standard Arabic Corpus for Segmentation and Validation" [17].

978-1-6654-4231-2/21/$31.00 ©2021 IEEE

TABLE I. AVAILABLE ARABIC CORPORA

Corpus	Author	Words	Category	F/C?
SAC [17]	Awdeh, Abdallah	18,000,000	multipurpose Arabic Corpora	F
Adjir [3]	Abdelali	113,000,000	monolingual text corpus	F
KSUCCA [25]	Alrabiah, Salman, Atwell	50,000,000	monolingual text corpus	F
OSAC [29]	Saad, Ashour	22,000,000	monolingual text corpus	F
Al Watan [28]	Abbas, Smaili, Berkani	10,000,000	monolingual text corpus	F
Tashkeela [33]	Zarrouki, Balla	75,000,000	monolingual text corpus	F
Al Khaleej [30]	Abbas, Smaili, Berkani	3,000,000	monolingual text corpus	F
KACST [1]	Al-Thubaity	732,780,509	online searchable corpus	F
CAC [24]	Al-Suleiti, Atwell	842,684	monolingual text corpus	F
Kalimat [26]	El-Haj, Koulali	18,167,183	Multipurpose Arabic Corpora	F
SACS [19]	Abu Salem	46,968	monolingual text corpus	F
ICA [32]	Alansary, Nagi	80,000,000	Online searchable corpus	F
Al-Raya [7]	Hasnah	219,978	monolingual text corpus	F
Arabic Modern Standard [4]	Abdalali, Cowi, Soliman	113,000,000	monolingual text corpus	F
UJAC [8]	Hammo, Al-Shargi	7, 522,941	monolingual text corpus	F
LDC [14]	Graff, Walker	76,000,000	monolingual text corpus	C
An-Nahar Newspaper [15]	ELRA	144,000,000	monolingual text corpus	C
Al-Hayat [34]	University Essex	18,639,624	monolingual text corpus	C
Nemlar [6]	ALP team	500,000	monolingual text corpus	C
Arabic Gigaword Corpus 1th ed. [13]	Graff	391,619	monolingual text corpus	C
ArabicGigaword 2th ed. [11]	Graff, Chen, Kong, Maeda	481,906	monolingual text corpus	
ArabicGigaword 3th ed. [12]	Graff	576,799	monolingual text corpus	
ArabicGigaword 4th ed.	Graff, Chen, Kong, Maeda	848,469	monolingual text corpus	
Arabic Gigaword, 5th ed.	Graff, Chen, Kong, Maeda	1,077,382,000	monolingual text corpus	

F: Free, C: Commercial

III. TYPES OF TEXT CORPORA

There are many types for corpus, each one of them meets the specific needs and interests of researchers in different domains (Table 2).

TABLE II. TYPES OF CORPUS

Main types	Subtypes
Raw Text Corpora	Monolingual corpus Parallel corpus Multilingual corpus Comparable corpus Learner corpus Diachronic corpus Specialized corpus Multimedia corpus Web-based Corpora Dialectal Corpora
Annotated Corpora	Annotated Corpora Named Entity Corpora Error-Annotated Corpora Miscellaneous Annotated Corp.
Speech Corpora	
Handwriting Recognition	
Miscellaneous Corpora	
Lexicon	Lexical Databases Words Lists

For further information on the types of corpus and their distribution, consider reviewing the article entitled “Overview of Arabic Sentence Corpora” [16].

IV. GOLD ARABIC CORPUS (GAC)

Arabic data available online is the appropriate resources for building a large corpus that can be used in language studies. Many researchers in Information Retrieval, Machine Translation and Arabic Language processing in general benefit from the data provided in online like Arabic newspapers, Arabic magazines, and others.

Nevertheless, the existing tagged corpora are not complete in term of segmentation and validation for our needs and besides the majority of them are not free, limited in sources, types and genres, small in size, and misspelled. This is why the existing Arabic corpora still has some limitations. In order to resolve these issues, a new free, tagged and reliable corpus called GAC (Gold Arabic Corpus) for the standard Arabic word segmentation and evaluation might be a solution.

To carry out the gold Arabic corpus for serving NLP, building the Silver Arabic Corpus “SAC” [17] was the start. First, data collection was done from alwatan Arabic newspapers, and then treated, and manually modified into a gold, free, tagged, and reliable corpus with a specific structure.

It contains around 20,291 articles, organized into six categories (Culture, Religion, Economy, Local News, International News and Sports), covers more than 18 million words, and it consists of a collection of texts annotated and enriched with linguistic information. In our work, we used Stanford Arabic POS tagger [23], our XML annotator, our JSON annotator, and the word segmentation Arabic grammar regulations [18].

This gold corpus is maintained in a particular format (prefix*-stem-suffix*), and the CACXml tool is developed to convert the corpus into a corpus with xml tags.

According to the Arabic word structure (prefix*- stem - suffix*), our xml structure matches this sequence of morphemes as follows:

```
<Segment>
     <Word> يخرجونها </Word>
     <Prefixes>
      <Prefix>ي </Prefix>
     </Prefixes>
      <Stem> خرج </Stem>
     <Suffixes>
      <Suffix> ون  </ Suffix>
      <Suffix>ها</ Suffix >
     </Suffixes>
</Segment>
```

The tagged corpus formed includes four fields: the word field (before segmentation), and the word after segmentation which contains three fields:

Prefixes – stem – suffixes
ي – ذهب – ون

In parallel, the JavaScript Object Notation (JSON) structure is built for many reasons:

- To enrich our gold corpus
- Faster, and very easy to use
- The wide range of supported browser compatibility with Schema Support
- Can be used by web services and other connected applications
- Restful web services use JSON extensively as the format for the data inside requests and responses.

```
{
“Segment” [
     {
          “Word”:”يخرجونها”,
          “Prefixes”: [“ي”],
          “Stem”:”خرج”,
          “Suffixes”: [“ون”, “ها”]
          ]
     },
     {
          “Word”:”يلعبون”,
          “Prefixes”: [“ي”],
          “Stem”:”لعب”,
          “Suffixes”: [“ون”]
     }
          ]   }
```

A. Corpus Resources

The Gold Arabic corpus profits from a wide range of Arabic resources, such as Arabic newspaper, Arabic summaries, and other as shown below (Table 3).

TABLE III. SOURCES OF CATEGORIES

Sources
Omani newspaper Alwatan
Extractive single-document system summaries
Multi-document system summaries
Named Entity Recognized articles
Part of speech tagged articles
Morphologically analyze articles

The articles of data collection for the gold corpus fall into six categories: religion, economy, culture, international news, local news, and sports (Table 4).

TABLE IV. COLLECTION STATISTICS

Categories	Number of Words
Religion	1,555,635
Economy	3,122,565
Culture	1,359,210
International News	855,945
Local news	1,460,462
Sports	9,813,366
Total	**18,167,183**

B. Metadata and Encoding

The GAC corpus is still stored in text files in order to expand its use by other researchers and programs.

In addition, this Arabic corpus is tagged into JSON to facilitate and expand its use by others researchers and programs.

This corpus is encoded with UTF-8, as this encoding scheme will be great benefit to researchers in the field of Arabic information retrieval and NLP.

V. ANAYSIS AND COMPARISON

The gold corpus is considered among the minority corpus classified as standard Arabic corpus. It is certainly required for the supervised training and assessment of systems doing NLP. It is characterized by:

- The wide range of Arabic resource
- Covering multiple categories which makes it well representative
- Tagging into Xml and JSON
- Manual corrected
- Specific structure
- Free of cost
- Considered as a large corpus containing about 18,167,183 words with 20,291 articles covering periods.

Several factors are taken into consideration by organizations, universities and Arabic language researchers upon creating Arabic corpus such as size, categories, price and structure of corpora. These factors are respected by the gold Arabic corpus.

Regarding to the size, the bigger corpus is more efficient in the study and results. The gold corpus contains more than 18,000,000 words with 20291 articles covering periods.

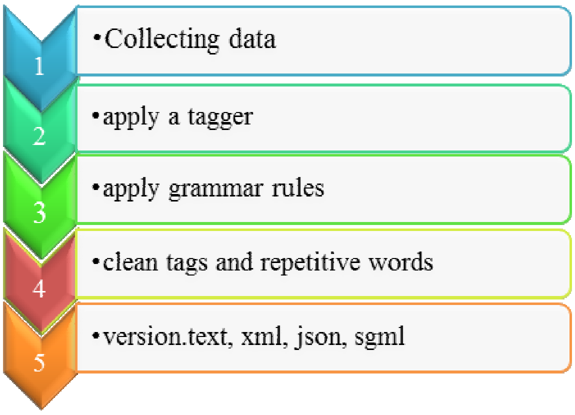
Concerning categories, this corpus covers multiple categories which make it well representative (Table 4).

Contrary to the commercial corpus, the gold corpus is available for free for researchers in different Arabic Natural Language Processing.

VI. METHODOLOGY

Following data collection (Schema 1) from different Arabic resources like alwatan Arabic newspapers and alkalimat corpus, we apply our grammar rules [18], and system based on the POS tagged corpus to split the Arabic words (prefixes-stem-suffixes), then we clean manually our corpus by removing the repetitive words, and fixing some words segmentation.

SCHEMA I. CORPUS BUILDING STEPS


1
•Collecting data
2
•apply a tagger
3
•apply grammar rules
4
•clean tags and repetitive words
5
•version.text, xml, json, sgml

VII. CONCLUSION

In order to deal with problems attributed to the lack of Arabic tagged corpora, this article has addressed a Gold Arabic Standard Corpus (GAC) building process that can be used in the evaluation and validation of the supervised and unsupervised learning tools in the syntaxic field, and it is free of cost to support the Arabic NLP researchers.

The use of the Gold Arabic Corpus promotes further work on the Arabic NLP for free including the papers, annotated texts, entities and summaries. Researchers can use it to test and assess their Arabic tools as standard corpus or baselines in supervised and/or unsupervised machine learning field, and it can use to check the result of Arabic word segmentation tools.

In the future, working on creating an online frame work that can facilitate the freely use of our Arabic corpus GAC will be the next goal. Eventually, any suggestion regarding the corpus by colleagues and researchers is very welcome.

ACKNOWLEDGEMENTS

This work was part of the Lebanese University initiative ’Semantic Analysis of Arabic Texts Using Ontology and Word Net’, supported by the Lebanese University and Paris 8 University.

REFERENCES

[1] 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.

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

[3] A. Abdelali, “Adjir Corpora,”, http://aracorpus.e3rab.com/, 2005.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

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

[10] 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.

[11] 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.

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

[13] 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.

[14] 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.

[15] 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.

[16] H. Awdeh, G. Gernard, M. Hajjar, “Overview of Arabic Sentence Corpora,” In 13th International Conference on Neural Computation Theory and Applications NCTA, Setubal, Portugal, 2021.

[17] H. Awdeh, A. Abdallah, G. Gernard, M. Hajjar, “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, 2019

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

[19] H. Abu Salem, “SACS Corpus,” Saudi Arabian National Computer Science Conference.

[20] I. Guellil, H. Saâdane, F. Azouaou, B. Gueni, Nouvel D, “Arabic natural language processing: an overview,” In Journal of King Saud University - Computer and Information Sciences. 2019.

[21] 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.

[22] 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.

[23] 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.

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

[25] 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.

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

[27] 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.

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

[29] 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.

[30] 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.

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

[33] 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.

[34] 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.

[35] 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.