Arabic Corpus of Library and Information Science: Design and Construction

Ayman Eddakrouri

General Education Department, College of Humanities, Effat Library and Cultural Museum, Effat University, Saudi Arabia

aeddakrouri@effatuniversity.edu.sa

Abstract: ***This paper addresses the principal considerations in creating the Arabic Corpus of Library and Information Science, a specialized Arabic corpus on the academic genre. This discusses ten phases of creation: the rationale of the Arabic Corpus of Library and Information Science, types of texts, resources of texts, legal approval, data collection, refining texts, revising texts, saving texts, coding texts, and finally, the size of the Arabic Corpus of Library and Information Science (357,485 tokens). Collecting texts of the articles was the longest and most challenging phase of building the corpus. Especially when we encounter files in PDFs or images that are difficult to read 100% correctly by various software. This challenge has been overcome by considering several factors that have been clarified at this stage. The Arabic Corpus of Library and Information Science can play a significant role in addressing the salient features of the academic genre, including keywords identification, lexico-grammatical patterns, themes, topics, and index terms used in the genre of Library and Information Science. Furthermore, the steps of creating the Arabic Corpus of Library and Information Science can guide in building other corpora for any genre or language.***

Keywords: ***Arabic Corpora, Arabic Natural Language Processing, Information Retrieval Systems, Indexing Arabic Texts, Arabic Information Extraction, Academic Genre.***

1 INTRODUCTION

Corpora are tools that provide sufficient authentic texts/data needed for Natural Language Processing, Applied Linguistics, Indexing, Data Mining, Information Retrieval, and Machine Translation [1] [2] [3]. As a result, using these tools has increased exponentially worldwide since the nineteenth of the last century. However, benefiting from these tools is highly disappointing in the Arab world [4].

Furthermore, most creators of corpora gave a particular concern to the language of media (e.g., online newspapers, magazines, and newswire agencies) rather than other genres, such as academic texts. Therefore, the present study attempts to partially fill this gap by creating an open/free source Arabic corpus that can be consulted for any empirical research purposes rather than relying on researchers' intuition. To fulfill this aim, the researcher tried to build an Arabic corpus for Library and Information Science (LIS), abbreviated by ArCLIS. The creation of ArCLIS is inspired by the general steps of constructing specialized corpora suggested by Atkins, Clear, and Ostler, MacMullen [5], and Blecha [6].

The main goal that the author of this research aspires to achieve is to bridge this gap and contribute to the Arabic Academic Corpus Linguistics field, which has not received sufficient attention [7]. One of the critical steps to be taken is to build a corpus that will facilitate studies of Arabic academic writing. Hence, this paper describes the processes we applied to create the Arabic corpus for Library and Information Science. Specifically, it presents the decisions we made regarding design considerations for the corpus and the subsequent steps we followed to build it.

Some may ask: Why do we need a specific Arabic corpus from the scholarly studies of Library and Information Science? Apart from the gap described above, we are excited to compile such a corpus for the following reasons:

1) Scholarly articles are among the strongest and finest features of the literary movement in any language. In other words, scientific studies are characterized by their artistic excellence, production quality, and occupation of a prominent place on the cultural and scientific level [8].

2) One aspect of this prominence is also the dramatic increase in the number of scholarly articles available in electronic form in recent years: the output of collective regional publications for the Middle East, including the Arab World, is reported to have grown over the past four decades from 7,665 papers in 1981 to more than of 150,000 papers in 2019. This 20-fold absolute growth can also be seen on the back of expanding global output. Countries have invested in research as an essential part of economic policy [9]. This growing Arab scholarly production would attract research and analytical attention to it, put it in the correct organization, and study it from all aspects, especially linguistic ones.

3) It is too expensive and time-consuming to survey a whole population in a research study. Sampling is the best way to proceed with the research [10]. Here, the researcher chose the scholarly articles of Library and Information Science as a purposive sample since it is one of the scholarly and professional disciplines he is interested in.

The above figures indicate the growing role of Arab scholarly production and its place in the Arab scene. However, there is an apparent absence of a quantitative approach and a systematic corpus in the aforementioned studies, which indicates an inability to explore different aspects of big data from Arab studies. This likely emerged due to the inaccessibility of a computerized data set of Arab intellectual production. Therefore, the compilation of such Arabic scholarly articles will

allow researchers to enhance their studies and provide them, through the use of linguistics techniques, with valuable results unexpected when relying solely on human intuition.
Hence, this study can answer the following central question: What steps of constructing an Arabic corpus for the academic genre, namely Library and Information Science?

2 LITERATURE REVIEW

We cannot deal with the whole literature on Arabic corpora in general. Such a premise needs separate work to cover the entire literature in this area. Therefore, we will focus here only on that which deals with building Arabic corpora.
While surveying the literature on building Arabic corpora, we found a significant shortage compared with literature on creating English or French corpora. The Arabic corpora we have identified serve various functions and vary in size, content, accessibility, and annotation technique. These corpora are either general or specialized. We can claim that online newspapers make up the vast majority of texts in specialized Arabic. Until recently, far too little attention has been paid to setting up Arabic corpora other than online news media, especially in the research and academic genre [11] [12] [13].
Goweder and De Rock [14] introduced one of these few amounts of literature when they described how to create an Arabic corpus consisting of 18,5 million words collected from Arabic news of Alhayah newspaper. Initially written in HTML, this corpus' text represented 42,591 articles and covered seven categories (General, News, Economics, Sports, Computers, Internet, Sciences and Technology, and Cars and Business). The authors shaped the outline of characteristics and features of data and how researchers can represent it in Arabic corpora. The authors also attempted to identify the differences between establishing an Arabic corpus and an English one.
Four years after Goweder and De Rock's study, three researchers suggested another model for an Arabic corpus, or rather a set of Arabic sub-corpora for Modern Standard Arabic (MSA) [1]. The researchers aimed to investigate Stylistics and lexical and semantic vocabulary variation from one Arab state to another. The researchers retrieved and collected texts from some Arabic newspapers that made their articles available on the Web to fulfill this aim. Meanwhile, they excluded (Arabic) PDF files that cannot be processed or analyzed electronically. Thus, the sample was settled to Alahraam (Egypt), Alra`y Al'aam (Kuwait), Alwatan (Oman), the Algerian News Agency, Safir (Lebanon), Aljazeera (Saudi Arabia), Almaghreb Alyoum (Morocco), Petra (Jordan), Alraya (Qatar), Tishreen (Syria), and the Iraqi News Agency. Finally, the three researchers used URSA program to crawl the Websites of these newspapers.
Al-Sulaiti and her colleague Atwell [15] also argued that the Arabic language needs a project that provides free access to authentic texts as in other languages, e.g., English, French, and German. Their viewpoint was to build an Arabic corpus that does not include Standard Arabic only but also Modern Arabic used in daily-life communications all over the Arab World. Therefore, the researchers collected written and spoken texts representing the modern regional Arabic varieties. They aimed to teach researchers, linguists, and learners of Arabic how to use Modern Arabic and its new lexical items. Thus, they collected texts from different resources, i.e., magazines, newspapers, news agencies, and emails. The final corpus consisted of 843,000 tokens.
One year later, the first statistical measure for assessing Arabic corpora depended on Zipf's Law was introduced by Benajiba and Rosso [16]. They applied this statistical measure to four Arabic corpora from different topics and genres to characterize their features: complexity, variety, and word frequency distribution.
AbdelRaouf and his colleagues [17] presented a paper describing an Arabic corpus of 6 million tokens extracted from various genres and language resources (Websites, online newspapers, chat rooms, Arabic dictionaries, old Arabic books, academic literature, and the Holy Qur`an). The authors aimed to offer a ready-made Arabic corpus for any research use.
In addition, Hamada [18] suggested an approach for designing Arabic corpora addressing their importance in resolving linguistic problems, lexicography, and teaching languages. The researcher also signaled obstacles in building Arabic corpora, how to encounter these challenges, and how researchers can collect raw texts representing the infrastructure for any corpus. Besides, the researcher referred to the processes of annotation and coding.
Mansour [4] also asserted the significant role of corpora in Natural Language Processing and Analysis and Linguistic Statistics. Additionally, he spotlighted the absence of studying Arabic corpora. In this regard, Mansour suggested practical solutions via a model for an Arabic corpus entitled 'Arabic National Corpus (ANC).' The researcher proposed four creation stages: planning for the ANC, data collection, computing the ANC, and analyzing the ANC. Also, he recommended collaboration between national and regional organizations to let this project come to light.
In the same year, Almeman and Lee [19] introduced a scheme for a multi-dialectal Arabic corpus on the Web to be the primary resource for Arabic texts. Their approach included five steps of the creation: collecting data and multi-dialectical phrases from different Arab states, assigning the collected words per every Arabic dialect, calculating numbers of tokens for every Webpage and link, specifying Webpages for downloading texts using Bing API, and lastly, refining and normalizing data. The final size of ANC reached 50 million tokens.
One of the most recent studies on building Arabic corpora was introduced by Abu Elkhair [20] when he constructed an Arabic corpus consisting of 3,303,723 tokens collected from news published on ten Arabic Websites. Abu Elkhair used two programs to extract data from the websites: MetaProducts Offline Explorer Pro and Visual Web Ripper. One added value distinguished in Abu Elkhair's corpus was marking up its data by adding metadata fields using SGML and XML.
Alfraidi and his colleagues [21] recently introduced the Saudi Novels Corpus, a useful linguistic and stylistic research tool that contains around 3,000,000 tagged words gathered from 53 novels written by different writers and covers the period

from 1930 to 2019. They outlined the steps they took and the choices they made when building the corpus. They outlined and made clear the design requirements, data collection techniques, annotation process, and encoding procedures. They also gave some preliminary findings from the content analysis of the corpus. The research represented a step toward trying to close the gap between corpus linguistics and Arabic literary texts.
To sum up, the field of building Arabic Corpora is still tiny, with a reasonable number of significant contributions in recent years. These contributions can be found in fundamental principles, such as the creation criteria applied and the language resources consulted. Improvements in the Arabic corpora capacities process and the vast quantities of Arabic data that are becoming available have contributed to these fair achievements. However, these factors also continue to drive the demand for faster and more accurate searches, especially in genres other than media, i.e., academic texts.

3 METHODOLOGY AND DATA

Despite building some Arabic corpora that have already been used or will be used, there is no specialized corpus on the academic literature. Therefore, this study attempts to partially fill this gap by investigating the existing literature and, thus, collecting authentic texts that reflect the actual use of Arabic academic literature.

4 POPULATION OF THE STUDY

The target population or the primary genre of the study is represented in the full texts of electronically published literature on Library and Information Science. The researcher selected this field because it is one of his professional and academic interests, enabling him to easily access this literature compared to any other field.

5 PILOT STUDY AND SAMPLING

The data collection procedure is a critical issue that data compilers must address, and it is usually the starting point for building a corpus [22] [23] [24]. However, we piloted our study before collecting the data to provide paramount precision during the ArCLIS corpus construction.
The aims of this pilot study were [25] [26] [27]:

- *Assess the feasibility of the approach that is intended to be used in our larger-scale study.*
- *Verify the suitability and processability of the collected texts.*
- *Distribute and organize the whole procedures in a way that guarantees flexibility in performing all processes without overlapping procedures for one another.*

The pilot study was initiated to investigate the availability of Arabic texts on LIS, which guarantees the representativeness and the balance between different types of literature. One critical notice that dramatically affected this study's data selection and sampling was the lack of full-text published literature, i.e., books, theses, dissertations, conference papers, etc., compared to Arabic journal articles. So, the researcher decided to select the latter as his study's primary sample.
Accordingly, the researcher selected a random sample of various articles from the primary sample before gathering the full texts of the journal articles. Whereas these articles were saved in Notepad files to be effectively processed by any software later.

6 DATA COLLECTION AND BUILDING THE CORPUS

After piloting and verifying the data sampling, the data/text collection was executed, and then the practical procedures for creating the ArCLIS were described. It should be noted that any corpus must be created through a series of meticulous and consecutive processes and steps. As a result, this section depicts the sub-stages of ArCLIS development. Fig. 1 illustrates these sub-stages.

FIGURE 1: SUB-STAGES OF ArCLIS CONSTRUCTION

A. Identifying the aim and rationale of the corpus

This research aims to develop a specialized corpus that reflects the language used in the Arabic academic genre (i.e., LIS). This corpus will be made available for future research and/or linguistic analysis (e.g., Arabic academic phrasebanks, Arabic register, Arabic academic writings, annotation, tagging, etc.).

B. Specifying types of texts

Types of data are essential to be addressed by data collectors, and it is also a critical issue that must be identified from the beginning to build any corpus. The success of this step is usually determined by whether the data is available in a format that allows for a convenient and quick process. Because most of our collected data was unavailable on the Web in machine-readable formats, no spoken texts have been dealt with, only written ones.

It should also be noted that we collected data in a practical manner. As a result, the corpus makes no claim to adhere to a strict sampling framework [28]. We had to use this method of data collection because we ran into a significant stumbling block early in the project. We encountered difficulties in obtaining the texts and converting them into a machine-readable format, i.e., not all of the articles we obtained were easily machine-readable [29]. This issue is not unique to our case, as it is a common challenge in the construction of Arabic corpora in general [30]. Hence, the texts included were contingent on their availability. In other words, we only included texts that we could easily obtain in the appropriate format for further computations. As a result, there was no personal preference (e.g., a focus on specific writers, journals, or countries) or consistent sampling framework used in the selection of included articles (this also supports point C of 6.3 ASSIGNING RESOURCES of TEXTS). This method was beneficial in our case because it allowed us to manage our time during the project period. We collected as much relevant data as possible in a reasonable amount of time.

C. Assigning resources of texts

This corpus depended only on one language resource as the primary sample of LIS genre: journal articles. This sample guarantees confronting three main criteria whenever any corpus is built. These criteria are:

1) Variation: journal articles here introduce sufficient variation in research and writing.

2) Comprehensiveness and representativeness: journal articles here were characterized by their relative coverage of all primary subjects and sub-topics in LIS. Additionally, these journal articles mainly represented a particular field of social sciences and humanities.

3) Balance: *[2]* The criteria mentioned above guarantee balance and non-bias in the suggested corpus as the primary requirement in creating any corpus in general.

Besides, the difficulties in collecting the journal articles were significantly less than in collecting other types of literature, i.e., books, theses, dissertations, papers, etc.

D. Obtaining the legal approval

There is no need for any legal approval or consent to benefit from copyrighted articles since the law allows access to make use of journal articles without getting back to their authors or even having an excuse; based on the exceptions reported in copyright laws, some of which are:

1) Use the publication (texts of literature in [ArCLIS]) for purely educational purposes.

2) Create one copy of the publication for non-commercial personal use.

3) Data mining for publications is processed for digitizing and indexing their contents so that the texts can be processed by specialized software programs [31].

E. Collecting texts of the articles

Unlike other corpus-based studies [32], this phase was the longest and most challenging phase of building ArCLIS because the whole collection process was carried out manually without any harvesting tool. Whereas the researcher relied on three resources to retrieve and collect the texts. These resources are:

1) Web search engines: to retrieve texts from the Web.

2) Arabic bibliographic database: to survey the Arabic literature on LIS.

3) Colleagues: ask and prompt them to provide all available electronic files of literature on LIS.

It is worth mentioning that the researcher paid great attention in this phase to two available formats of texts only, which are:

1) Web pages.

2) Microsoft Word files.

This is attributed to the easiness and readiness of processing Arabic texts made available in these two formats rather than any other ones, e.g., PDF and image files. It is challenging to process these two formats in Arabic using any software program [33]. So, some electronic journals were excluded from the inclusion in ArCLIS: Journal of King Fahd National

Library (Kingdom of Saudi Arabia), the Journal of Library and Information (Libya), the Information Science Journal (Morocco), and the Iraqi Journal for Information (Iraq).
The corpus contains the full text of each processable article collected. This method has the advantage of increasing the likelihood of detecting the most linguistic characteristics [11]. If the corpus had only included a subset of each text, it would have missed many salient features, especially when targeting a specific genre like our academic one.

F. Refining texts

This stage aims to customize and prepare the texts to perform later on by processing and analysis. In order to fix the errors we found and delete mistakenly inserted undesired HTML texts, we had to thoroughly analyze the whole text by carefully reading their words and comparing them with the original copy if they were published in a web page format, for instance. This step was critical since it made the data fed into the corpus accurate, despite being time-consuming and effort-intensive.
A manual data refining task was carried out following the data's cleansing and acquisition in an editable, machine-readable format. We manually edited each text in this phase to remove any extraneous elements. In particular, we eliminate images, charts, forms, some tables, cataloging data, ISSN, and authors' biography. Thus, we were sure that the information entered in each text file only contained the full texts of the articles by eliminating these components.
Another manual refining step was implemented by polishing the texts and making them in the best machine-readable format. The research depended on the "Replacement" capability in Microsoft Word. This would pave the way for it to work flexibly and highly efficient with any text analysis program that can later be used on ArCLIS corpus by avoiding any obstacles that may hinder the program's work. This step entailed the following manual work: eliminating extra spaces, diacritics, and Kasheeda, and detaching numbers, punctuations, and symbols from words if needed.
After that, the collected texts were revised and double-checked to verify their conformity to the original texts. As such, we ensured that each word in the corpus was assigned to its most appropriate token.

G. Saving and encoding the texts

After collecting, customizing, and preparing the texts, they were all saved in separate Notepad files (with .txt extension) to facilitate their processing and analysis procedures.
Equally importantly, the saved texts were encoded in Unicode after the previous phase step. Most software that accepts processing Arabic texts well applies this encoding system.

7 ARCLIS CORPUS STATISTICS & PRELIMINARY RESULTS

Here, in this part of the study, we present some statistical data that can be extracted after designing and creating ArCLIS corpus. Table 1 presents the principle information about ArCLIS corpus, which comprises 357,485 tokens distributed over 674 articles from 7 journals.
Table 1 also shows the data distributions for the numbers and percentages of articles, tokens, and the size of each journal. Expectedly, there are differences and imbalances between these elements. This is normal to occur because not all or even some journals need to be equal in the number of articles or even their size. Thus, we saw this big difference from one journal to another.
One proviso that needs mentioning and repeating here relates to ArCLIS corpus is that all journal articles were collected; however, some journals were not due to the lack of electronically available articles. This may be methodologically accepted since the availability of articles and the lack of others are related to what extent these articles are made available in a machine-readable format, not to consider some of them and ignore others purposively. At the same time, this situation may lend randomness to the sample selection, especially if we admit that the randomness in collecting data is considered one quality of corpora [34].

1 CONCLUDED REMARKS

Despite the lack of Arabic corpora, compared to English, there is an increasing interest in building Arabic corpora in recent times, especially those concerned with newswire. There is still a gap (this was highlighted in the introduction of the study) in the Arabic corpora of other genres, especially the academic ones. In this study, we investigated ways to create and design an Arabic corpus on the academic genre, consisting of 674 scientific articles that were collected from journals in LIS, which are available in digital form. In order to create a clear corpus, we cleared and normalized the obtained files appropriately after the collection stage. This also includes ways to trim the collected text to remain in the best machine-readable format.
ArCLIS corpus is one of the very rare corpora specialized in the academic genre and Arabic. Even though ArCLIS corpus already covers a wide range of journal articles, progressive growth of the corpus by including other literature (books, theses, dissertations, etc.) within LIS remains a further goal. Thus, ArCLIS corpus could accurately study the LIS genre and its terminology, keywords, knowledge profiles, themes, authors' interests, regional concerns, and more.
ArCLIS corpus is now ready for future research work or linguistic analysis that would enrich the academic domain. In addition, it is ready to conduct POS tagging and annotation of these texts, which can produce accurate statistical indicators and results that reflect the academic writing used by Arab authors. ArCLIS corpus can also be used to make stylistic and qualitative studies, identify keywords, linguistic patterns, the linguistic environment associated with words, and lexical studies.

TABLE 1

SKELETON OF ARCLIS CORPUS

#	Journal	Articles #	Articles %	Tokens #	Tokens %	Size in megabyte	Size %	URL
1	Cyberarians Journal	175	25.9	90,516	25.3	16.7	33.4	http://www.journal.cybrarians.info/
2	Alarabiya 3000 = Arabic Journal 3000	215	31.8	76,531	21.4	10.2	20.4	http://www.arabcin.net
3	Almajalla Alordoniyya lillmaktabaat wa alma'loomaat = The Jordanian Journal for Libraries and Information (previously, Resaalit almaktaba = The Library Mission)	73	10.8	58,884	16.4	5.61	11.2	http://www.jlia.org/ar/2012-05-07-20-09-49/2012-05-07-20-10-46.html
4	E'lam (stand for the Arabic Federation of Libraries and Information (AFLI)	54	8	38,444	10.7	4.54	9.1	http://arab-afli.org/main/content.php?alias=مجلة_أعلم
5	Dirasaat alma'loomaat = Information Studies	50	7.4	35,277	9.8	6.07	12.1	http://journals.ksiscs.com.sa/index.php/ijs
6	Diraasaat Arabiya fi elmaktabaat wa 'elm alma'loomaat = Arabic Studies of Library and Information Science	45	6.6	33,731	9.4	4.92	9.8	http://www.mandumah.com/Humandexjournals
7	Maktabaat dot net = Libraries .net	62	9.1	24,102	6.7	1.84	3.6	http://www.mandumah.com/Humandexjournals
	Sum	674		357,485		49.88		

It should be noted that there is still no Arabic corpus specialized in the academic genre to be compared to the current study. However, related studies, such as one on Arabic dialects, were created by exploiting the web as a corpus. This study concluded that the texts collected from websites to obtain fair texts. These corpora include four main Gulf dialects, Levantine, Egyptian, and North African, which gave a result of 14.5 million, 10.4 million, 13 million, and 10.1 million tokens, respectively, and the sum of the distinct types in all corpora is more than 2 mega types [35]. ArCLIS corpus is much smaller than this corpus used in this study due to the qualitative and quantitative differences between the two corpora and the readiness and availability of texts used in both corpora. We also find the same observations in other Arabic news-based corpora, such as KACST Arabic Corpus (1,182,515,633 words) [36], Leeds Arabic Internet Corpus (317,000,000 words) [37], International Corpus of Arabic (100,000,000 words) [38], and arabiCorpus (100,000,000 words) [39].
Finally, it is worth mentioning that after conducting this study, the researcher successfully implemented a project he had always sought throughout his MA and Ph.D. studies: establishing an Arabic phrasebank for academic writing.

REFERENCES

[1] A. Abdelali, J. Cowie and H. Soliman, "Building A Modern Standard Arabic Corpus," in *Workshop on Computational Modeling of Lexical Acquisition, The Split Meeting*, Croatia, (2005).

[2] F. Rizk, " التوظيف اللغوي للمدونة المتوازية بين اللغتين اليونانية والعربية دراسة تطبيقية على مسرحية "هيكابي" ليوربيديس مشروع "بيرسيدس"أنموذجاً" = The Linguistic Employment of the Parallel Coprpus of Greek and Arabic: An Empirical Study on the Play "Hikapi" by Euripid," *Egyptian Journal of Language Engineering,* vol. 5, no. 2, pp. 78-98, (2018).

[3] M. Elsaadany and S. Alansary, "A Tool for Measuring Linguistic Variations in Machine Translation: A Corpus Based Study," *Egyptian Journal of Language Engineering,* vol. 6, no. 2, pp. 29-43, (2019).

[4] M. Mansour, "The Absence of Arabic Corpus Linguistics: A Call for Creating an Arabic National Corpus," *International Journal of Humanities and Social Science,* vol. 3, no. 12, pp. 81-90, (June 2013).

[5] W. MacMullen, "Requirements Definition and Design Criteria for Test Corpora in Information Science," *SILS Technical Report,* no. https://sils.unc.edu/sites/default/files/general/research/ TR-2003-03.pdf, (2003).

[6] J. Blecha, *Building Specialized Corpora, MA Thesis,* Brno: Masaryk University, Faculty of Arts, Department of English and American Studies, (2012).

[7] N. Mohamad, M. Baharun, Z. Ramli, A. A. Rahman and A. Saifuzzaman, "Academic Language Register In Arabic Articles From Al-Majallah Al-Urduniyyah Fi Al-Ulum Al-Tarbawiyyah," *Arabi Journal of Arabic Learning,* vol. 5, no. 2, pp. 459-474, (June 2022).

[8] N. Al-Mansour, "Teaching Academic Writing to Undergraduate Saudi students: Problems and Solutions - A King Saud University perspective," *Arab World English Journal,* vol. 8, no. 3, p. 94– 107, (2015).

[9] J. Adams, J. Ouahi, D. Pendlebury and M. Szomszor, "Global Research Report: The Changing research landscape of the Middle East, North Africa and Turkey," Institute for Scientific Information (ISI), Philadelphia, (2021).

[10] W. Cochran, Sampling Techniques, 3rd ed. ed., New York: Wiley and Sons, (1977), pp. 259-61.

[11] A. Al-Thubaity, "A 700M+ Arabic Corpus: KACST Arabic Corpus Design and Construction," *Lang. Resour. Eval.,* vol. 49, no. 3, p. 721–751, (2015).

[12] A. Shoufan and S. Alameri, "Natural Language Processing for Dialectical Arabic: A Survey," in *Proceedings of the Second Workshop on Arabic Natural Language Processing*, Beijing, China, (2015).

[13] W. Zaghouani, "Critical Survey of the Freely Available Arabic Corpora," in *Proceedings of International Conference on Language Resources and Evaluation (LREC'2014), OSACT Workshop*, Reykjavik, Iceland, (2014).

[14] A. Goweder and A. De Roeck, "Assessment of a Significant Arabic Corpus," in *In: the Arabic NLP Workshop at ACL/EACL*, http://www.abdelali.net/ref/ACL-EACL%202001_goweder.pdf, (2001).

[15] L. Al-Sulaiti and E. Atwell, "The design of a corpus of Contemporary Arabic," *International Journal of Corpus Linguistics,* vol. 11, no. 1, p. 1–36, (2006).

[16] Y. Benajiba and P. Rosso, "Towards a measure for Arabic corpora quality," in *Proceeding of International Colloquium on Arabic Language Processing, CITALA-2007*, http://www.researchgate.net/publication/228972993_Towards_a_measure_for_arabic_corpora_quality/file/9fcfd51421b952e785.pdf, (2007).

[17] A. AbdelRaouf and e. al., "Building a multi-modal Arabic corpus (MMAC)," *IJDAR,* vol. 13, p. 285–302, (2010).

[18] S. Hamada, "Nahu mahaj arabi moqtarah li-tassmeem almoodawanaat alloghaweeya = Towards an Arabic suggested approach for designing corpora," (2011). [Online]. Available: http://www.globalarabnetwork.com/science-a-it/2784-2011-04-04-14-49-07.

[19] K. Almeman and M. Lee, "Automatic Building of Arabic Multi Dialect Text Corpora by Bootstrapping Dialect Words," in *1st International Conference on Communications, Signal Processing, and their Applications (ICCSPA)*, Sharjah, (2013).

[20] I. Abu El-Khair, "Abu El-Khair Corpus: A Modern Standard Arabic Corpus," *International Journal of Recent Trends in Engineering & Research (IJRTER),* vol. 2, no. 11, pp. 5-13, (November 2016).

[21] T. Alfraidi and e. al., "The Saudi Novel Corpus: Design and Compilation," *Applied Sciences,* vol. 12, no. 6648, 2022.

[22] G. Kennedy, An Introduction to Corpus Linguistics, Abingdon, UK: Routledge, (1998).

[23] S. Kübler and H. Zinsmeister, Corpus Linguistics and Linguistically Annotated Corpora, London: Bloomsbury Publishing, (2015).

[24] H. Salama and S. Alansary, "Lexical Growth in Egyptian Arabic Speaking Children: A corpus Based Study," *Egyptian Journal of Language Engineering,* vol. 4, no. 1, pp. 29-34, (2017).

[25] T. Baker, Doing Social Research, New York: McGraw-Hill Inc., (1994).

[26] J. Frankland and M. Bloor, "Some issues arising in the systematic analysis of focus group material," in *Developing Focus Group Research: Politics, Theory & Practice*, London, Sage, (1999).

[27] I. Holloway, Basic Concepts for Qualitative Research, Oxford: Blackwell Science, (1997).

[28] T. McEnery and A. Hardie, Corpus Linguistics: Method, Theory and Practice, Cambridge: Cambridge University Press, (2011).

[29] A. Abdel-Kareem, A. Hussein, E. Shokry and O. A. El-Din, "ID Card Recognition Based on Arabic OCR System," *Egyptian Journal of Language Engineering,* vol. 1, no. 2, pp. 35-49, (2014).

[30] P. Maiwald, "Exploring a Corpus of George MacDonald's Fiction," *North Wind: A Journal of George MacDonald Studies,* vol. 30, no. 5, pp. 50-84, (2011).

[31] P. Samuelson, "Justifications for Copyright Limitations and Exceptions," in *Copyright Law in an Age of Limitations and Exceptions*, Cambridge University Press, (2017), pp. 12 - 59.

[32] M. Abdo, A. Youssef and N. Sarhan, "Analyzing Judgment in Bipolar Depression Patients' Narratives Using Syntactic Patterns: A Corpus-Based Study," *Egyptian Journal of Language Engineering,* vol. 6, no. 1, pp. 1-11, (2019).

[33] M. Tayyab, A. Hussain, M. Alshara, S. Khan, R. Alotaibi and A. Baig, "Recognition of Visual Arabic Scripting News Ticker From Broadcast Stream," *IEEE Access,* vol. 10, pp. 59189-59204 , (2022).

[34] N. Az-Zouheri, Qamoos mosstalahaat almaaloomaatiyya wa alloghawiyaat alhisaabiyya: engleezi-arabi with masaared bi elengleeziya wa alarabiyya = Dictionary of Informatics and Computational Linguistics Terms: English-Arbic with lists in English and Arabic, Beirut: Lebanon Library, (2003).

[35] K. Almeman and M. Lee, "Automatic building of Arabic multi dialect text corpora by bootstrapping dialect words," in *The First International Conference on Communications, Signal Processing, and their Applications (ICCSPA'13)*, Sharjah, (February 2013).

[36] King Abdulaziz City for Science and Technology, "KACST Corpus," [Online]. Available: https://corpus.kacst.edu.sa/. [Accessed 15 January 2023].

[37] L. University, "Leeds Arabic Internet Corpus: Querying Internet corpora," [Online]. Available: http://corpus.leeds.ac.uk/internet.html. [Accessed 15 January 2023].

[38] B. Alexandrina, "International Corpus of Arabic," [Online]. Available: https://www.bibalex.org/ica/en/. [Accessed 15 January 2023].

[39] D. Parkinson, "arabiCorpus," Brigham Young University, [Online]. Available: https://arabicorpus.byu.edu/. [Accessed 15 January 2023].

[40] P. Samuelson, "Justifications for Copyright Limitations & Exceptions," [Online]. Available: https://www.law.berkeley.edu/files/Justications_for_Copyright_Limitations_and_Exceptions_Pamuela_Samuelson.pdf.

[41] N. Az-Zouheri, Qamoos mosstalahaat almaaloomaatiyya wa alloghawiyaat alhisaabiyya: engleezi-arabi with masaared bi elengleeziya wa alarabiyya = Dictionary of Informatics and Computational Linguistics Terms: English-Arbic with lists in English and Arabic, Beirut: Lebanon Library, (2003).

[42] A. Lepage, "Overview of exceptions and limitations to copyright in the digital environment," *e-Copright Bulletin,* no. http://unesdoc.unesco.org/images/0013/001396/139696E.pdf, (2003).

BIOGRAPHY

Dr. Ayman Eddakrouri is Assistant Professor and Director of Effat Library & Cultural Museum at Effat University, Jeddah. He holds three academic degrees; MA in Library and Information Science (LIS) from Cairo University, MA in Applied Linguistics (AL) from the American University in Cairo, and PhD in LIS and AL from Helwan University. His main areas of interest are Computational Linguistics, Corpus Linguistics, Managing Academic & Digital Libraries and Museums, and Teaching Arabic for Native & non-Native Speakers. Dr. Eddakrouri is a member of the American International Consortium of Academic Libraries (AMICAL), the Arab Federation of Library and Information, and the Editorial Secretary of Arabic Journal of Library & Information Science. He established the Arabic Language Bank (ALB), an online reference that provides researchers with ready-made academic writings. ALB has recently been awarded Jeddah Award for Innovation. Dr. Eddakrouri is the Chair of the Higher Committee for Promoting Arabic Research at Effat University. Dr. Eddakrouri also worked at the American University in Cairo, the British University in Egypt, the Egyptian Ministry of Higher Education, and other reputable institutions, either as an instructor of Arabic Language & Culture, librarian, evaluator of Information Retrieval Systems, or consultant of learning management systems.

الذخيرة اللغوية العربية لعلوم المكتبات والمعلومات: التصميم والبناء

أيمن الدكروري

قسم التعليم العام، كلية عفت للعلوم الإنسانية، مكتبة عفت والمتحف الثقافي، جامعة عفت، جدة، المملكة العربية السعودية
aeddakrouri@effatuniversity.edu.sa

المستخلص

تتناول هذه الدراسة الاعتبارات الرئيسة لإنشاء الذخيرة اللغوية العربية لعلوم المكتبات والمعلومات (ArCLIS Corpus). وهي ذخيرة لغوية عربية متخصصة في النوع الأكاديمي. ويناقش هذا عشر مراحل للإنشاء،؛ وهي: الأساس المنطقي للذخيرة اللغوية العربية لعلوم المكتبات والمعلومات، وأنواع النصوص، ومصادر النصوص، والموافقة القانونية، وجمع البيانات، وتنقيح النصوص، ومراجعة النصوص، وحفظ النصوص، وترميز النصوص، وأخيرًا حجم الذخيرة اللغوية العربية لعلوم المكتبات والمعلومات (357485 هيكل كلمات). وكانت مرحلة جمع نصوص المقالات هي الأطول والأكثر صعوبة في بناء هذه الذخيرة اللغوية، خاصة مع ملفات البي دي إف والصور التي يصعب معالجتها بشكل صحيح 100٪ بواسطة برامج المعالجة المختلفة. إلا أنه تم التغلب على هذا التحدي من خلال مراعاة عدد من العوامل التي تم توضيحها في هذه المرحلة. هذا، ويمكن أن تلعب الذخيرة اللغوية العربية لعلوم المكتبات والمعلومات دورًا مهمًا في تحديد الكلمات الرئيسة، والأنماط المعجمية النحوية، والموضوعات، والمصطلحات الكشفية المستخدمة في نوع علوم المكتبات والمعلومات. علاوة على ذلك فإنه يمكن لخطوات إنشاء الذخيرة اللغوية العربية لعلوم المكتبات والمعلومات أن ترشد الباحثين في بناء ذخائر لغوية أخرى لأي نوع أو لغة.

الكلمات المفتاحية

الذخائر اللغوية العربية، معالجة اللغة العربية الطبيعية، نظم استرجاع المعلومات، تكشيف النصوص العربية، استخلاص المعلومات العربية.