Jordanian Journal of Computers and Information Technology (JJCIT), Vol. 10, No. 04, December 2024.

# PROCESSING TOOLS FOR CORPUS LINGUISTICS: A CASE STUDY ON ARABIC HISTORICAL CORPUS

Bassam Hammo[1] and Sane Yagi[2]

*(Received: 10-May-2024, Revised: 2-Jul.-2024, Accepted: 29-Jul.-2024)*

## ABSTRACT

*This paper explores the development, design and reconstruction of a Historical Arabic Corpus (HAC), which covers more than 1600 years of uninterrupted language use. The study emphasizes the technical aspects followed to enhance the system and provide a usable concordancer, along with simple experiments conducted on the corpus and the concordancer. Arabic has a rich literary and cultural heritage spanning thousands of years. The inclusion of digital resources and the advancement in natural language processing (NLP) technology have made Arabic historical corpora increasingly crucial for researchers and learners worldwide. By integrating HAC and its tools into Arabic language learning, learners can delve deeper into vocabulary and culture and gain valuable insights that improve their language skills and understanding of Arabic. This combination of human guidance and NLP technology makes learning an engaging and enjoyable experience, offering a dynamic and authentic way to master the Arabic language.*

## KEYWORDS

*Historical Arabic corpus, Corpus tools, Concordancer, Learning Arabic, Data normalization, Semantic shifting.*

## 1. INTRODUCTION

The Arabic language is recognized for its linguistic richness. It has a rich and extensive history and is one of the most widely spoken languages in the world. Literature and historical texts have been produced from the early Islamic period to the modern era, providing a valuable resource for researchers and language learners. Understanding and analyzing Arabic requires a comprehensive and systematic approach as a language deeply rooted in literature, religion and daily life. One vital tool for linguistic research in the Arabic language is the compilation of text corpora and dictionaries [1].

A language corpus (plural corpora) is a collection of texts systematically organized and annotated for linguistic analysis. Constructing an Arabic-language corpus provides a fundamental resource for linguistic, cultural and historical studies. The following are a few benefits of what a dedicated linguistic Arabic corpus can do.

1. Providing linguists and researchers with a vast and varied dataset, enabling in-depth analyses of language patterns, syntax and semantics. It also facilitates investigations into language evolution and usage across different regions.
2. Offering educators and language learners an extensive range of authentic materials representing the language in diverse contexts helps develop effective teaching methodologies, curriculum design and language-proficiency assessments.
3. Contributing to preserving cultural heritage and becoming a repository of cultural expressions, idioms, philosophies and social standards.
4. Providing an essential tool for developing applications and advancing research in areas like natural language processing (NLP), machine learning (ML), deep learning (DL), data mining (DM) and large language models (LLM). For instance, building robust language models, sentiment-analysis tools and machine translation systems relies on the availability of high-quality linguistic data.

Historical corpora, another type of text corpora, are extensive collections of written texts compiled and organized for educational and research purposes. They provide researchers and learners access to various historical documents, allowing them to study the evolution of a language and gain insights into a particular cultural and social history from different eras. These corpora contain many historical documents, such as religious texts, legal texts, scientific works and other genres. Arabic historical corpora are essential resources for researchers interested in studying the history of the Arab world and

[1] B. Hammo is with the Department of Computer Information Systems, KASIT, The University of Jordan and with the Department of Software Engineering, Princess Sumaya University for Technology, Amman, Jordan. Email: b.hammo@ju.edu.jo, b.hammo@psut.edu.jo

[2] S. Yagi is with the Department of Foreign Languages, University of Sharjah, United Arab Emirates. Email: syagi@sharjah.ac.ae

Islamic civilization.

In this study, building upon the groundwork described in [2]-[3], we continue the efforts to compile and refine a historical corpus of Arabic texts named HAC, spanning 1600 years of language evolution. The previous attempts focused on introducing the essential tools required for assembling HAC and implementing NLP techniques to preprocess the text, ultimately organizing it into a structured eXtensible Markup Language (XML) schema, allowing for search and analysis. The HAC corpus is a valuable resource for linguists and is designed for Arabic-language learners studying at the University of Jordan. It enables them to delve into the rich linguistic heritage of millions of textual instances.

The corpus is designed to serve multiple purposes, including supporting Arabic-language learners and advancing research in NLP, data mining and other computational fields. The extensive dataset and the developed tools are tailored to the needs of both educational and research-oriented users. Researchers in NLP and other fields benefit from the structured and annotated data for developing and testing their models. The need for a new concordancer tool is due to the limitations of existing tools when applied to historical Arabic texts. Many of these tools are not optimized for the Arabic language's unique morphological and orthographic complexities, particularly in a historical context. Existing tools often fail to handle search variations critical for accurate historical-corpora analysis. The new concordancer tool addresses these gaps by offering enhanced morphological analysis capabilities, adapting to historical-text variations and providing advanced customization options. This tool is essential for researchers and linguists to perform accurate searches and analyses on historical texts, making it a significant advancement in Arabic-corpus linguistics.

The remainder of this paper is structured as follows: Section 2 presents background information on corpus linguistics. Section 3 reviews pertinent prior research. Section 4 presents the methodology and technical processes that we have followed to reconstruct the HAC corpus. Section 5 showcases experiments demonstrating the practical utility of our efforts. Finally, Section 6 concludes the work and outlines future-research avenues.

## 2. BACKGROUND

### 2.1 A Brief Introduction to Corpora and Corpus Linguistics

Corpora encompass extensive and structured collections of texts from diverse sources, accommodating written texts such as books, articles, websites and transcriptions of spoken language. These electronically stored and processed repositories are often augmented with annotations and metadata, such as Part-of-speech (POS) tags and morphological information, to enhance their utility for specific research objectives [2]-[4]. Corpora are vital resources for linguistic and computational linguistic research and the study of language use in real-world contexts. They are fundamental to applications in lexicography [5], translation [6], language learning and teaching [7] and data mining [8].

Historical corpora are particularly significant for historical linguists. They comprise texts spanning the entire history of a language or specific eras. These corpora illuminate the evolution of languages over time, revealing shifts in word meanings and grammatical structures. They play a significant role in compiling historical dictionaries [1], offering insights into semantic changes and providing illustrative quotations for word senses.

Whether we are designing a contemporary or historical corpus, details regarding sources' authorship, publication date, genre and broader context should be documented to enrich the corpus with valuable insights that aid researchers in interpreting and categorizing linguistic data.

Corpora and their associated tools facilitate language understanding and learning, offering valuable resources for educators and learners [9]. These tools are essential in language acquisition, contributing to vocabulary expansion, contextual learning and a deeper understanding of grammar and syntax. For instance, adaptive learning systems can utilize corpus data to individualize learning experiences [10]. These platforms customize language learning based on individual strengths and weaknesses and optimize the learning process for each student. Other essential tools might include the following [11]:

#### *Tools for the Creation of the HAC*

1. Language-analysis software: We utilized a concordancer, Khoja's stemmer and Stanford part-of-

speech tagger to analyze the language in HAC, ensuring accurate annotation and tagging of the historical texts.
2. Metadata-management tools: Tools used to manage and catalog information about each text in the corpus, including author, publication date and title, facilitating easy search and retrieval.
3. Corpus-management software: We employed a database-management system to manage the HAC, allowing for efficient storage, retrieval and updates to the corpus data.
4. Computational tools for analysis: NLP tools were integral in analyzing the HAC, enabling tasks, such as tokenization and light stemming, which are crucial for handling the linguistic complexity of Arabic texts.

### *Tools for Exploitation of the HAC*

1. Concordancer: Researchers and language students can use the concordancer to search and analyze specific words or phrases within HAC, helping understand historical usage patterns and contexts.
2. Visualization tool: Visualizes linguistic data to uncover patterns and trends within the HAC, which are valuable for linguistic and historical research.

## 2.2 The Historical Arabic Corpus (HAC)

The initial historical Arabic corpus (HAC) was initiated in 2015. It was constructed using a corpus-builder system developed in Java to compile and encode its data into an XML schema automatically. Interested readers who want to learn more about the corpus-builder system are referred to [2]-[3]. The input to the corpus builder was a text document encoded in UTF-8 with its meta-data [2]. We integrated a stemmer and a part-of-speech (POS) tagging modules in the corpus-builder system to build the final corpus. We adapted an Arabic stemmer developed by Khoja [12] to extract a root, stem and morphological pattern for each word. For POS tagging, we employed the Stanford Part-Of-Speech Tagger [13]. Both tools were popular when the project was started, providing essential functionalities required for processing Arabic text.

Since then, newer tools such as Farasa [14] and Computational Approaches to Modeling Language Lab (CAMeL) [15] have been developed, offering enhanced capabilities in Arabic NLP. These tools could be considered for future updates and improvements to HAC, potentially increasing the accuracy and efficiency of text processing.

### *Project Goals and Implementation*

The primary goal of the HAC project is to create a comprehensive, searchable database of historical Arabic texts. This involved several key objectives:

1. Elucidating the conceptual and technical refinements applied to the HAC corpus, including the normalization procedures employed to enhance its consistency and coherence.
2. Outlining the technical methodologies employed in constructing a database and searchable indices, incorporating various simplified and normalized tokens to facilitate efficient information retrieval.
3. Presenting the design of a new concordancer tool developed with user-friendly interfaces, providing researchers with a platform to experiment with and analyze the corpus.
4. Experimenting with the HAC corpus and the enhanced concordancer.

To achieve these goals, the following steps were undertaken:

1. A robust database architecture was designed to facilitate efficient storage and retrieval of text data. This included restructuring the data-storage system to handle the large volume of text and metadata.
2. Advanced search algorithms were implemented to enable precise and fast data retrieval. This included the development of custom-search interfaces tailored to the needs of researchers.

## 2.3 The Concordancer

A concordancer is a software tool used in linguistics and language analysis to identify and analyze the frequency, distribution and usage of words and phrases in a text corpus. It helps determine the context in which a word or phrase appears and displays the lines of text containing the word or phrase and its neighboring words. This enables researchers to study how words and phrases are used in different contexts and how they are related. The following points highlight a few of the benefits of using a concordancer.

1. Teachers can utilize a concordancer in language teaching to help students understand the usage of words and phrases in contexts and to develop their vocabulary and grammar skills [16]-[19].
2. Translators can use a concordancer to identify the most appropriate translation of a word or phrase in a given context [20]-[21].
3. Linguists can use a concordancer to study language use and patterns, such as the distribution of words and phrases across different genres, periods or social groups [22]-[23].
4. Researchers in fields, such as literature, history and sociology, can use a concordancer to analyze text to identify patterns and trends in the data [24].

### 2.4 Initial Challenges and Limitations

At the early stage of constructing HAC, a few problems were raised. The major problem was associated with the scalability of data storage, while the second one was related to searching and retrieving the data effectively. The primary users of the system were students from the Linguistic Department at the University of Jordan. They know little about computers. As the volume of data started growing, the response time of inquiring data from the XML database turned out to be very slow and the system's GUI was not user-friendly. Therefore, it was obvious that the original data structure and the GUI design were unsophisticated and needed to be revised and enhanced. This study aims to solve this problem by replacing the XML schema with a sophisticated relational database-management system running on the server side with optimized queries and a friendlier concordancer system.

Another issue that we are still striving to solve is that HAC needs to be balanced and requires more historical Arabic text in digital format, which makes it, in its current state, unrepresentative of the genres and eras that it should cover [2]. As part of our continuous-improvement efforts, we added further five million terms to HAC, bringing the total number of terms to 50 million terms.

## 3. LITERATURE REVIEW

The development of Arabic corpora is still in its early stages [25]. Initially, Arabic corpora were mainly created through manual efforts or basic tools that compile texts into XML format, often accompanied by metadata annotations [26]. The UAM Corpus Tool, developed by the Universidad Autónoma de Madrid [27], is a comprehensive software suitable for corpus linguistics research. It offers functionalities for corpus compilation, annotation and analysis, including concordancing, collocation analysis and statistical-processing tools. It utilized XML as its underlying data-storage format and facilitated cross-layer searching, semi-automatic tagging, statistical reporting and visualization of tagged data.

Later, plenty of contemporary Arabic corpora were designed with a range of structures and annotations [28]-[29]. Another avenue encompassed Quranic and Hadith (Prophet Mohammad's traditions) corpora [30]-[33]. Other scholars focused on designing tools for the Arabic language, such as the work of [34], where the authors proposed iSPEDAL, an enhanced electronic dictionary for the Arabic language. A corpus and a set of tools to experiment with contemporary Arabic were introduced in [35]. The corpus included editorials of newspapers collected from different countries, Arab countries' constitutions, dictionaries and the Holy Quran, in addition to news from sports, technology and politics.

The development of Arabic corpora has faced challenges, but significant progress has been made. Resources such as the Multilingual Annotated Standard Dataset of Educational Resources (MASADER) [36] and the Linguistic Data Consortium (LDC) catalog [37] have played a vital role in advancing Arabic corpus linguistics. MASADER provides a comprehensive catalog of Arabic-language resources, including datasets and tools for various NLP tasks. The LDC catalog includes extensive Arabic-language resources such as speech and text corpora, lexicons and annotated linguistic data. Building on these foundations, the HAC corpus aims to provide additional resources and tools tailored to historical Arabic texts.

English historical corpora, comprising samples of texts from earlier eras, are instrumental in studying language variation, changes and development. Examples of well-known English historical corpora accessible through the web are given in Table 1. For instance, the Helsinki Corpus of English Texts is a structured multi-genre corpus spanning Old, Middle and Early Modern English periods, offering insights into linguistic forms, structures and lexemes across different epochs. Similarly, ARCHER (A Representative Corpus of Historical English Registers) presents a multi-genre corpus of British and American English, covering the period from 1600 to 1999.

Table 1. Examples of well-known projects encompassing English historical corpora.

| **English historical corpora** | **URL address (*Accessed on April 20, 2024*)** |
|---|---|
| The British National Corpus (BNC)) | http://www.natcorp.ox.ac.uk/corpus/ |
| The Penn Treebank (PTB) | https://catalog.ldc.upenn.edu/LDC99T42 |
| Helsinki Corpus of English Texts | https://varieng.helsinki.fi/CoRD/corpora/HelsinkiCorpus/ |
| ARCHER: A Representative Corpus of Historical English Registers | https://www.projects.alc.manchester.ac.uk/archer/ |

For Arabic historical corpora, the most relevant examples are the King Saud University Corpus of Classical Arabic (KSUCCA) and the King Abdul-Aziz City for Science and Technology (KACST) Arabic corpus [2]. KSUCCA, although supposedly encompassing classical Arabic texts from the pre-Islamic era until 1100 C.E., it lacks comprehensive coverage and evidence of representativeness [38]. Similarly, while the KACST Arabic corpus aims to be a comprehensive resource spanning various periods and domains, efforts are ongoing to enhance its representativeness and balance [39].

Other examples of well-known Arabic historical corpora projects accessible through the web are given in Table 2. The projects highlight the growing efforts to digitize, preserve and make accessible Arabic historical datasets, enabling researchers to investigate the rich history of the Arabic world. These projects are just a few examples of the wide range of ongoing research on Arabic historical corpora and there may be additional recent works since our knowledge cut-off date in December 2023.

Concordancing tools are crucial in linguistic analysis, aiding language learners and researchers in vocabulary acquisition, collocation identification and grammatical comprehension. While English boasts numerous concordancing tools, Arabic offerings are relatively limited. Earlier works included AntConc [40], aConCorde [41], AraConc [42].

While these tools serve essential functions, there remains a need for further research to develop sophisticated schemas accompanied by tools tailored to handle morphological annotation and facilitate automated Arabic-corpora construction. Moreover, advancement in Arabic NLP is slower than in English due to a scarcity of freely available corpora, lexicons and sophisticated machine-readable dictionaries, underscoring the need for concerted efforts to advance research in this area.

Table 2. Examples of well-known projects encompassing Arabic historical corpora.

| **Arabic historical corpora** | **URL address**<br>**(*Accessed on April 20, 2024*)** |
|---|---|
| The Digital Library of the Middle East (DLME) from Stanford Libraries represents a platform combining data collections from various cultural heritage institutions worldwide. It offers free and open access to the rich cultural legacy of the Middle East and North Africa. | https://dlmenetwork.org/library |
| The Qatar Digital Library (QDL) is a massive online repository that offers access to a diverse collection of historical documents related to the Gulf and Middle East. The collection includes manuscripts, maps, photographs and archival materials that provide insights into the Arabic world's social, cultural and political history. | https://www.qdl.qa/en |
| Al-Maktaba al-Shamela is a digital library that hosts a vast collection of classical Arabic texts, including religious, historical, literary and scientific data. It offers a comprehensive platform for accessing and searching thousands of Arabic manuscripts and books [43]. | https://shamela.ws/ |
| The King Saud University Corpus of Classical Arabic (***KSUCCA***) is a 50 million tokens annotated corpus of Classical Arabic texts from the period of pre-Islamic era (7th Century CE) until the fourth Hijri century (11th Century). | https://sourceforge.net/projects/ksucca-corpus/ |

This study aims to refine the HAC corpus through various technical improvements, including using normalization procedures to enhance consistency and coherence, the creation of a database with optimized queries for efficient information retrieval and the development of a user-friendly concordancer tool. By accomplishing these objectives, this research provides a solution to current challenges and contributes to the progress of Arabic-corpus linguistics and language-learning methodologies.

The refined HAC corpus and its accompanying tools will be a valuable resource for linguists, researchers, educators and learners, enabling them to conduct detailed analyses of Arabic-language patterns, syntax and semantics and provide authentic materials for language acquisition. This research represents a significant step towards bridging the gap between historical Arabic corpora and presentday language-learning needs, delivering a dynamic platform for studying and mastering the Arabic language.

Studying authentic historical Arabic texts can provide valuable insights into classical Arabic grammar, vocabulary and stylistic conventions not commonly found in modern language-learning materials. While primarily beneficial for researchers and linguists, historians and those interested in digitizing Arabic cultural heritage can also benefit from the HAC corpus. However, it may have limited direct application for beginners learning Arabic as a foreign language. The HAC corpus primarily serves the needs of historians, linguists, researchers of the Arabic language and those interested in digitizing Arabic cultural heritage. Here is how each party may benefit from HAC:

1. HAC aims to preserve and provide access to historical Arabic texts, which are valuable for researchers studying the evolution of the Arabic language, linguistic variations over time and historical events documented in Arabic sources.
2. Linguists can use the corpus to analyze language usage, semantic changes and syntactic structures in historical contexts, aiding in understanding how the language has evolved and adapted across different periods of history.
3. Digitizing historical Arabic texts preserves cultural heritage and promotes awareness and appreciation of Arabic literature and history.

### *Differentiation from Other Arabic Historical Corpora*

1. Scope and coverage: The HAC corpus has eight genres and around 50 million words distributed among predefined eras spanning 1600 years from the pre-Islamic era to the twenty-first century.
2. Accessibility and tools: An extensive dataset and a concordancer tailored to the needs of both educational and research-oriented users. Researchers in NLP and other fields benefit from the structured and annotated data for developing and testing their models.

## 4. Research Methodology

The methodology we employed in this study is depicted in Figure 1. It incorporates four stages: (1) data collection, (2) data pre-processing, (3) constructing the HAC database and indices and (4) experimentation with HAC and the concordancer. In the following subsections, we discuss each stage in more detail.

### 4.1 Data Collection

The HAC corpus was planned to include all primary classical Arabic text material available online, using the automated tools described in [2]. We searched the internet because of the limited resources for free digitized Arabic text and found Al-Maktaba al-Shamela, a free and open-source digital library (available at https://shamela.ws/) [43]. It has a wide range of religious, historical, literary and scientific data, making it an excellent platform for accessing and searching thousands of Arabic manuscripts and books.

To assemble a comprehensive corpus, we carefully considered the issue of textual representation before collecting the corpus data. Our primary concern was that the collection should cover all periods of Arabic history as recommended by Arabic literary historians [44]. This approach helps us understand the development and evolution of Arabic literature over time.

From previous work, we collected over 50 million tokens [2]-[3]. We classified the data into different eras every 100 years, beginning from Classical Arabic (pre-Islamic times) and ending with Modern Standard Arabic (MSA) of the current century. In addition, the corpus data was categorized into primary and secondary sources based on their representation of the language used during the time of authorship [2]. Primary texts, such as poetry, literary prose and non-fiction, offer insight into contemporary language practices without commenting on older texts. In contrast, secondary texts, like Quran exegesis and critical analyses of ancient poetry, provide commentary reflecting the language usage of the commentator's era, shedding light on linguistic customs from earlier times.

Similar to the works [2]-[3], genres conventionally influence the language used in a text. They are considered significant factors when representing texts. Consequently, we categorized the texts into eight genres: Dictionaries, Literary Prose, Poetry, History, Philosophy, Religion, Science and Thought. Apart from the era, genre and primary/secondary categorization, we gathered general information about the texts, such as document title and author, to compile them into the corpus. We are also working on another text distribution based on regions and varieties of Arabic dialects, which we plan to include in future work. Table 3 shows various text examples from the HAC corpus based on historical principles and annotations regarding genre, author and era.

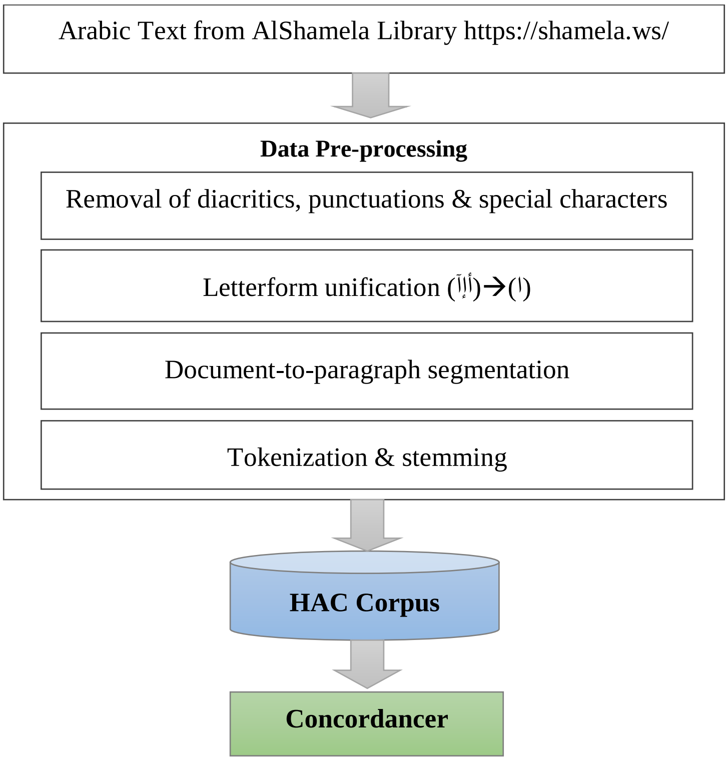

Arabic Text from AlShamela Library https://shamela.ws/
Data Pre-processing
Removal of diacritics, punctuations & special characters
Letterform unification (آأإ)→(ا)
Document-to-paragraph segmentation
Tokenization & stemming
HAC Corpus
Concordancer

Figure 1. Methodology flow diagram.

Table 3. A sample of HAC data resources (collected from https://shamela.ws/).

| | Title | Author | Era |
|---|---|---|---|
| **Dictionaries** | العين | الفراهيدي | 700-800 |
| | الصحاح تاج اللغة وصحاح العربية | الجوهري | 900-1000 |
| | مفردات القرآن | الأصفهاني | 1100-1200 |
| | مختار الصحاح | الرازي | 1200-1300 |
| | القاموس المحيط | الفيروزابادي | 1400-1500 |
| | تاج العروس 1-3 | مرتضى الزبيدي | 1700-1800 |
| **Poetry** | ديوان امرئ القيس | امرؤ القيس | < 600 |
| | شعر زهير بن أبي سلمى | زهير بن أبي سلمى | < 600 |
| | ديوان حسان بن ثابت | حسان بن ثابت | 600-700 |
| | ديوان كزهر اللوز أو أبعد | محمود درويش | 1900-2000 |
| | الأعمال الشعرية الكاملة لابراهيم طوقان | ابراهيم طوقان | 1900-2000 |
| **Philosophy** | تهافت الفلاسفة | الغزالي | 1000-1100 |
| | تلخيص الخطابة | ابن رشد | 1100-1200 |
| | حي بن يقظان | ابن طفيل | 1100-1200 |
| | المختصر في المنطق | محمد بن محمد ابن عرفة | 1300-1400 |
| | أهل المدينة الفاضلة | الفارابي | 1500-1600 |
| | تاريخ الفلسفة الحديثة | يوسف مكرم | > 2000 |
| **Religion** | القرآن الكريم | كلام الله عز وجل | 600-700 |
| | صحيح البخاري | أحاديث الرسول – البخاري | 600-700 |
| | الموطأ | الإمام مالك | 700-800 |
| | سنن أبي داوود | أبو داوود | 800-900 |
| | روضة العقلاء | ابن حبان | 900-1000 |
| | الأذكار | النووي | 1200-1300 |
| | تفسير الجلالين | جلال الدين المحلي و جلال الدين السيوطي | 1400-1500 |
| | شرح مسند أبي حنيفة | الامام القاري | 1500-1600 |

## 4.2 Data Pre-processing

Arabic is a highly derivational and inflectional language. To handle the different ways in which Arabic text can be represented, we applied several normalization techniques described in the works [45]-[47]. These techniques utilize the indices for efficient search through the database, while the content of the texts in the database should be preserved to maintain its originality and integrity.

Data pre-processing starts with converting a text document $D_i$ into the UTF-8 universal encoding, which represents every character in the Unicode character set, including Arabic characters. Further, a set of tasks is applied to extract the following information for each word $w_j$ in $D_i$: word's root, pattern, part of speech and stem, but stop words were not removed.

It is noteworthy to mention that in our approach, we strictly maintain the integrity of the original text. To improve search capabilities, we applied pre-processing steps, such as tokenization, stemming and root extraction to parallel text versions, not the original. These parallel versions were used exclusively for search and retrieval, ensuring that the original text remains unaltered and can be accessed in its original form. This methodology allows us to provide efficient search functionality while preserving the authenticity of the historical documents.

The pre-processing steps were handled automatically and included the following tasks:

1. ***Normalization*:** Building a corpus requires normalization before exploring its content. Arabic-text normalization usually involves removing punctuation, stripping numbers out, removing diacritical marks, …etc. Root extraction, for example, is essential for effective searching and frequency-based analysis, so that words such as (كاتب, *writer*), (مكتبة, *library*) and (مكتب, *office*) can all be correlated to the third person singular root (كتب, *he writes*). Sometimes, normalization also includes stemming words, so that words such as (الكاتبون, *writers*), (والكاتبون and *writers*) and (الكاتبات, "*female*" *writers*) can all be stemmed to (كاتب, *writer*) and hence are not considered different words, as they all represent the same concept. Indices might also be normalized to prepare a textual database for searching. For instance, if a search for (لَعِب, *he played*) is intended to match the words (لَعِبْ, *play*) and (لُعَبْ, *toys*), then the text would be normalized by removing the diacritical marks to be all represented by one token (لعب). A set of steps is applied to reduce the number of extracted terms. They include:
   a) The removal of nonletters and special characters.
   b) The removal of non-Arabic letters.
   c) The replacement of initial أ،إ،آ with bare alef ا.
   d) The replacement of knotted ة (*ta marbuta*) with ه *(ha)*.
   e) The replacement of ending dotless yeh (alef maksura, ى) with yeh ي.
   f) The removal of leading proclitic particles, such as definite article, prepositions and conjunctions, trailing haa (ه), trailing Yeh_Yeh_Noon (يين), trailing Waw_Noon (ون), trailing Haa_Alef (ها), trailing pronominal enclitics used for dual and masculine plural forms (هم، هما).
   g) The removal of single-tone letters, such as Waw (و) and those produced by the above normalization steps.

2. ***Splitting documents into paragraphs*:** This process breaks a text document $D_i$ into $n$ paragraphs at the boundaries of paragraphs.

3. ***Tokenization*:** This process analyzes the paragraphs and splits them into individual token (word) streams. The boundaries of words, such as whitespaces and punctuation marks, are determined in this process.

4. ***Stemming and root extraction:*** A shallow stemming approach was applied to remove common affixes (i.e., prefixes and suffixes) from each word to extract its stem. This helps simplify words for frequency-based analysis and searching. For example, the words (الكاتبون, writers), (والكاتبون and writers) and (الكاتبات, female writers) would be stemmed to (كاتب, writer). Meanwhile, Khoja's algorithm was utilized to extract the roots of words. This process involves a deeper morphological analysis to identify the core set of letters that convey the fundamental meaning of the word. As an example: The words (كاتب, writer), (مكتبة, library) and (مكتب, office) would all be correlated to the root (كتب, write).

In this study, we addressed ambiguity across clitics and stems, such as 'وجد,' which can mean "he found" or "and grandfather," using a simple approach based on basic linguistic rules and lexical analysis applied in the stemmer. Due to resource limitations when the project began in 2015, we did not implement advanced techniques, such as contextual analysis or machine learning, for disambiguation. This limitation requires further investigation in future studies.

5. ***Part-of-speech tagging***: The process of assigning a part of speech to each word in a sentence, such as a noun, verb, adjective and more. It is crucial in determining how sentences are constructed from smaller units. POS tagging is widely used in syntactic and semantic analysis of sentences. We used the Stanford tagger, an open-source package written in Java programming language, to assign part of speech tags to words [10]. The package includes two trained tagger models for English and tagger models for Arabic, Chinese, French and German.

## 4.3 Motivation for New Data Storage

When designing HAC, we considered a portable and adaptable storage structure that is readable by humans and computers. Therefore, an XML schema was developed, including metadata tags for each document and annotation text for each token's morphology. Each token was stored in a single tag, along with its annotation attributes, such as the root, morphological pattern, POS tag and stem [2]-[3]. However, as the XML corpus grows, searching becomes slower, but its accessibility by text editors and portability make it appealing. Interested readers could refer to [2] to learn about the corpus structure.

We switched from utilizing the XML schema to a database to manage and manipulate the data in HAC. The change is because databases ensure scalability to handle large volumes of data with high levels of integrity and reliability and offer structured storage with the capability to define tables, relationships and constraints. In addition, databases provide powerful query capabilities, allowing for efficient data retrieval. We redesigned the database using the MS SQL-Server database-management system running on the server side to make it accessible through the web. For a more productive search, we ended up with three relations, as shown in Figure 2. They are as follows:

1. ***The Genre Table***: Stores information about each document *D* in the corpus and has the following attributes:
   - *docid*: a unique primary key assigned to each document composed of a genre, era and sequence number.
   - *path*: the actual path to the document.
   - *title*: document's title.
   - *author*: document's author.
   - *year*: document's year.
   - *era*: the era to which the document belongs
   - *category*: stores one of two values: primary or secondary.
   - *region* & *variety*: to be used in the future to store the region of the document and the dialect of that period.
2. ***The Paragraph Table***: Stores the document content after being split into paragraphs where a record has the following attributes:
   - *docid*: the document *id*.
   - *lineid*: a number assigned to a paragraph extracted from each document *D*.
   - *context*: the actual text of a paragraph.
3. ***The Posting Table***: Stores and tracks the occurrences of words, roots, …etc., associated with each text line in a document *D*. It has the following attributes:
   - *postid*: a unique number assigned to each post.
   - *docid*: the document ID from where the posting originated.
   - *lineid*: the number of the paragraph from where the posting originated.
   - *word*: the word in a text line after being normalized as described earlier.
   - *stem*: the word's stem after being processed as described earlier.
   - *root*: the word's processed root.
   - *tag*: the part of speech tag assigned to the word within the context.
   - *pattern*: an annotation assigned to each word based on the position of the three consonants (فعل) and the affixes. Patterns help in understanding the meaning of words.
   - *TF*: term frequency. Counts the occurrences of a word within a document *D*.
   - *DF*: document frequency. Counts the occurrences of a word within the entire corpus. TF and

DF are used to assign a weight for each token in the corpus.

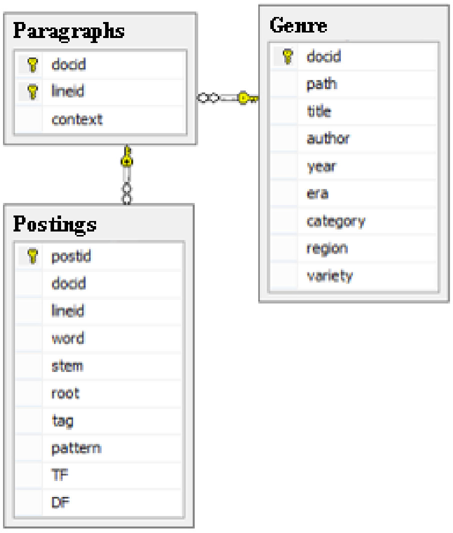


Figure 2. Database schema of the HAC corpus.

### 4.4 The Design of a New Concordancer

The primary purpose of a concordancer is to retrieve and display a text from the corpus in short contexts. It should allow the user to observe how a term is used within the context and how it might develop semantically within a period. Linguistic students at the University of Jordan heavily influenced HAC's concordancer. The initial design did not meet the users' satisfaction and a new, friendly design was in demand. The new concordancer includes alphabetical listings of all words in the historical corpus classified based on eras, genres and two main categories: primary or secondary. Searching through the concordancer shows where the terms (words, stems or roots) occur throughout all text.

The need for a new concordancer is emphasized by the limitations of existing tools when applied to historical Arabic texts. Many of these tools are not optimized for the Arabic language's unique morphological and orthographic complexities, particularly in a historical context. Existing tools lack the flexibility to analyze texts across different historical genres and eras. Our new concordancer is tailored specifically for the HAC corpus to address these gaps by offering enhanced morphological search capabilities, adapting to historical text variations and providing advanced customization options. This tool is essential for researchers and linguists to perform searches and analyses on historical texts, making it a significant advancement in Arabic-corpus linguistics. The concordancer provides the following functions:

1. Creating your word lists (vocabulary table) and producing concordances.
2. Searching for collocations and learning about a word's usage within neighboring words.
3. Counting word frequencies based on different eras and genres.
4. Discover a writer's stylistic traits by searching through authors.
5. Learning about all root derivatives and seeing each within the text's context.
6. Exploring results of searches to Excel sheets for further offline processing and analysis.

To illustrate the present interface of the concordancer, Figures 3 and 4 present screenshots from the new rendered version. Figure 3 shows the main components of the concordancer. The search selection tab has five options: word, stem, root, pattern and POS tag. The advanced tab allows a user to search for neighboring words around the word under search. If a search is performed on a root, a root-derivative list can be loaded to see all words derived from this root. A paragraph-up and paragraph-down offer an option to retrieve the previous and the following paragraphs for a selected paragraph from the grid. These functions are beneficial for linguists and researchers to understand and clarify the meaning of a word. Another search option is "Author," where a user can filter the search results by a particular author to learn about his/her writing style, for example. Finally, the statistics tab provides many HAC-related statistics and the user can export all his/her findings to an Excel sheet for further processing. Figure 4 shows a tracking of the Arabic root (عدل) meaning (modify, alter or adjust) in the literary prose genre in the era (700-800).

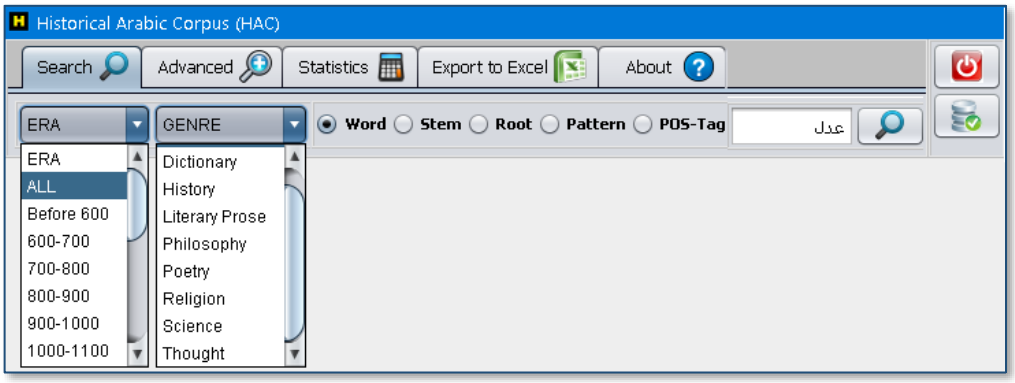
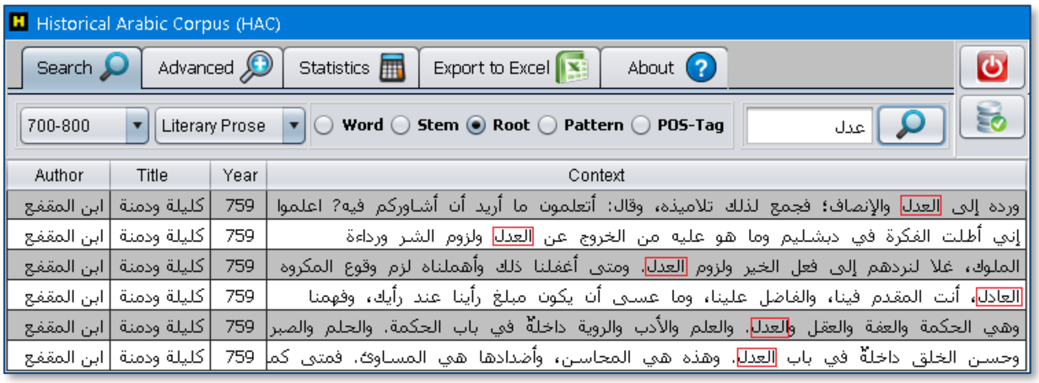


Figure 3. The concordancer and its functions.



| Author | Title | Year | Context |
|---|---|---|---|
| ابن المقفع | كليلة ودمنة | 759 | ورده إلى العدل والإنصاف؛ فجمع لذلك تلاميذه، وقال: أتعلمون ما أريد أن أشاوركم فيه؟ اعلموا |
| ابن المقفع | كليلة ودمنة | 759 | إني أطلت الفكرة في دبشليم وما هو عليه من الخروج عن العدل ولزوم الشر ورداءة |
| ابن المقفع | كليلة ودمنة | 759 | الملوك، غلا لنردهم إلى فعل الخير ولزوم العدل. ومتى أغفلنا ذلك وأهملناه لزم وقوع المكروه |
| ابن المقفع | كليلة ودمنة | 759 | العادل، أنت المقدم فينا، والفاضل علينا، وما عسى أن يكون مبلغ رأينا عند رأيك، وفهمنا |
| ابن المقفع | كليلة ودمنة | 759 | وهي الحكمة والعفة والعقل والعدل. والعلم والأدب والروية داخلةٌ في باب الحكمة. والحلم والصبر |
| ابن المقفع | كليلة ودمنة | 759 | وحسن الخلق داخلةٌ في باب العدل. وهذه هي المحاسن، وأضدادها هي المساوئ. فمتى كم |

Figure 4. The concordancer tracking the Arabic root (عدل) in Literary Prose in the period 700-800.

## 5. Experiments and Discussion

### 5.1 Corpus Statistics

In this experiment, we analyzed HAC statistically to gain insights into the prominent words and their frequencies. The corpus comprises texts from various historical sources, providing a rich resource for understanding linguistic patterns and historical themes.

The HAC corpus has eight genres and around 50 million words, as depicted in Table 4. Table 5 shows the distribution of words among pre-defined eras spanning 1600 years from the pre-Islamic era to the twenty-first century. As one might notice from Table 4, most of the words (82%) fall under two genres: 49% in Literary Prose and 33% in History. Table 5 shows that 14% of the words were found in the (after 2000) era, while three eras have equal distributions, each representing 11%. Unfortunately, HAC is not yet balanced and still unrepresentative regarding genres and eras that it is supposed to accommodate. Our ambition is to create a representative and balanced corpus for the future.

Table 4. The Historical Arabic Corpus (HAC) (source: https://shamela.ws/).

| Genre | Number of documents | Pct. | Number of paragraphs | Pct. | Number of words | Pct. | Number of distinct words | Pct. |
|---|---|---|---|---|---|---|---|---|
| Dictionaries | 9 | 1.6% | 183,059 | 5.2% | 2,605,962 | 5.2% | 320,484 | 13.2% |
| History | 143 | 24.7% | 1,163,937 | 33.0% | 16,479,162 | 32.9% | 642,901 | 26.4% |
| Literary Prose | 362 | 62.5% | 1,732,353 | 49.1% | 24,547,664 | 49.0% | 961,749 | 39.5% |
| Philosophy | 12 | 2.1% | 39,782 | 1.1% | 586,055 | 1.2% | 73,020 | 3.0% |
| Poetry | 11 | 1.9% | 18,213 | 0.5% | 252,471 | 0.5% | 68,491 | 2.8% |
| Religion | 11 | 1.9% | 166,477 | 4.7% | 2,383,475 | 4.8% | 137,240 | 5.6% |
| Science | 29 | 5.0% | 208,087 | 5.9% | 3,010,912 | 6.0% | 197,685 | 8.1% |
| Thoughts | 2 | 0.3% | 15,444 | 0.4% | 220,687 | 0.4% | 34,880 | 1.4% |
| *Total* | 579 | | 3,527,352 | | 50,086,388 | | 2,436,450 | |

Table 5. HAC's word distribution in different eras.

| Era | Word (Count) | Pct. | Era | Word (Count) | Pct. |
|---|---|---|---|---|---|
| Before 600 | 46,856 | 2% | 1300-1400 | 3,563,990 | 7% |
| 600-700 | 1,001,678 | 2% | 1400-1500 | 3,508,167 | 7% |
| 700-800 | 1,147,709 | 8% | 1500-1600 | 1,055,423 | 2% |
| 800-900 | 4,034,476 | 11% | 1600-1700 | 857,038 | 2% |
| 900-1000 | 5,647,041 | 11% | 1700-1800 | 762,241 | 2% |
| 1000-1100 | 5,653,135 | 10% | 1800-1900 | 1,188,169 | 2% |
| 1100-1200 | 5,153,226 | 7% | 1900-2000 | 5,416,067 | 11% |
| 1200-1300 | 3,638,343 | 1% | After 2000 | 7,085,705 | 14% |
| Unknown | 327,124 | 2% | *Total* | 50,086,388 | |

## 5.2 Experimenting with the Historical Arabic Corpus (HAC)

In this experiment, we inquired about the top 100 words and their frequencies in HAC. Table 6 gives the words and their frequencies in the corpus, while Table 7 gives a sample of the top 10 frequent words in each genre. The analysis revealed the following observations:

1. The top words mainly consist of common Arabic verbs, conjunctions and prepositions, reflecting their frequent usage in texts.
2. Words such as "قال" (said) and "قد" (had) indicate narrative and temporal aspects, suggesting a focus on describing events and actions in historical narratives.
3. By examining the top 10 frequent words in each genre, we can observe distinct lexical patterns characteristic of the respective genres, as shown in Table 7. For example, in the historical genre, words related to events, places and individuals are predominant, while in poetry, words associated with emotions, manners, war and environment are more prevalent. In religion, words related to spirituality and faith are predominant, whereas in the scientific genre, words associated with the human body and empirical observations may be more common.
4. Comparing the top frequent words across genres enables researchers to identify similarities and differences in linguistic usage and thematic emphasis. This comparative analysis can highlight genre interrelations and provide a deeper understanding of conventions.

Table 6. Top 100 words and their frequencies in the HAC corpus.

| Rank | Word | Freq. | Rank | Word | Freq. | Rank | Word | Freq. | Rank | Word | Freq. |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | من | 1,297,204 | 26 | لم | 135,987 | 51 | وفي | 80,375 | 76 | تعالى | 51,155 |
| 2 | في | 1,207,616 | 27 | الذي | 135,724 | 52 | كما | 78,988 | 77 | فلما | 51,053 |
| 3 | بن | 642,586 | 28 | ابو | 133,470 | 53 | سنة | 78,459 | 78 | كانت | 50,997 |
| 4 | على | 582,057 | 29 | فقال | 129,347 | 54 | غير | 78,258 | 79 | رسول | 50,858 |
| 5 | ان | 557,484 | 30 | قد | 127,867 | 55 | منه | 77,777 | 80 | لها | 50,631 |
| 6 | الى | 446,508 | 31 | وقال | 124,817 | 56 | مع | 75,880 | 81 | مثل | 48,512 |
| 7 | الله | 428,560 | 32 | وقد | 121,080 | 57 | وسلم | 74,301 | 82 | لما | 48,211 |
| 8 | ما | 394,041 | 33 | حتى | 117,852 | 58 | فيها | 74,218 | 83 | ايضا | 47,242 |
| 9 | قال | 354,285 | 34 | وكان | 116,497 | 59 | الدين | 70,727 | 84 | الملك | 42,944 |
| 10 | عن | 300,407 | 35 | هو | 114,798 | 60 | فإن | 70,501 | 85 | يقال | 42,791 |
| 11 | لا | 295,900 | 36 | ابي | 113,068 | 61 | الناس | 69,598 | 86 | قوله | 42,607 |
| 12 | كان | 209,242 | 37 | فيه | 111,886 | 62 | اليه | 69,441 | 87 | عمر | 42,592 |
| 13 | عليه | 200,995 | 38 | كل | 110,544 | 63 | علي | 68,289 | 88 | حدثنا | 42,573 |
| 14 | او | 191,781 | 39 | ومن | 110,453 | 64 | يكون | 67,655 | 89 | احمد | 42,484 |
| 15 | له | 191,766 | 40 | انه | 109,243 | 65 | وإن | 65,896 | 90 | لي | 41,093 |
| 16 | هذا | 187,287 | 41 | اي | 106,782 | 66 | عنه | 64,941 | 91 | ليس | 40,320 |
| 17 | ولا | 183,938 | 42 | محمد | 106,217 | 67 | ولم | 63,116 | 92 | فلا | 38,521 |
| 18 | ثم | 182,465 | 43 | التي | 103,839 | 68 | عند | 62,283 | 93 | هي | 38,154 |
| 19 | ذلك | 177,197 | 44 | بين | 102,981 | 69 | يقول | 59,388 | 94 | قول | 37,469 |
| 20 | اذا | 169,685 | 45 | هذه | 102,636 | 70 | اهل | 57,243 | 95 | قبل | 37,227 |
| 21 | ابن | 167,491 | 46 | بعد | 87,101 | 71 | وهي | 55,793 | 96 | اخر | 37,201 |
| 22 | عبد | 154,277 | 47 | صلى | 84,441 | 72 | منها | 55,364 | 97 | بني | 36,283 |
| 23 | وهو | 149,265 | 48 | يا | 84,040 | 73 | بعض | 54,259 | 98 | فإنه | 35,996 |
| 24 | به | 143,352 | 49 | بها | 82,465 | 74 | يوم | 54,068 | 99 | لأن | 35,972 |
| 25 | الا | 142,705 | 50 | وما | 81,155 | 75 | فى | 52,022 | 100 | شيء | 35,805 |

Table 7. Sample of top 10 words in each genre in the HAC corpus.

| History | Literary Prose | Philosophy | Poetry | Religion | Science | Thought |
|---|---|---|---|---|---|---|
| عمر | الزمان | الإنسان | السيف | الله | العين | المجتمع |
| السلطان | الهوى | العقل | الإبل | محمد | الرأس | النموذج |
| الحسن | القلب | النفس | الكأس | صلى | المعدة | العلمانية |
| مدينة | الحب | الفلسفة | الوغى | وسلم | الأدوية | الإنسانية |
| توفي | القصيدة | الحياة | الليالي | حدثني | الغذاء | الصهيونية |
| إبراهيم | الأمير | المذهب | الفؤاد | سمعت | البدن | الجماعات |
| صاحب | ديوان | المنطق | المكارم | أخرجه | الحرارة | المادي |
| عثمان | الفرزدق | خلدون | الأعداء | هريرة | القلب | النظام |
| دمشق | ليلى | أفلاطون | المنية | صحيح | السموم | الإمبريالية |
| بغداد | جارية | الموت | الرماح | رواية | الشيخوخة | الاقتصادي |

### 5.3 Measuring the Linguistic Richness of HAC's Text

Zipf's law, named after the American linguist George Kingsley Zipf [48], who proposed it in the 1930s, can be a helpful tool for identifying important words or concepts in the HAC corpus and measuring its text's lexical richness. Zipf's law, also known as the "law of word frequencies," is an empirical law that describes the statistical distribution of word frequencies in a corpus of natural-language text. The law states that the frequency of a word in a given corpus of text is inversely proportional to its rank in the frequency table. For instance, the second most common word in the corpus will occur approximately a half as often as the most common word, the third most common word will occur approximately one-third as often as the most common word and so forth. Equation (1) can mathematically express the law.

$$f(w) = \frac{\mathrm{k}}{\mathrm{r}} \tag{1}$$

where f(w) is the frequency of the word w, r is its rank in the frequency table and k is a constant. Similar to English, the law was also observed in Arabic. To show if HAC's content complies with Zipf's law, we inquired about the top 1000 words and their frequencies per each of the eight genres, as shown in Figure 5. We visualized the distribution of words using a log-log scatter chart showing the frequency of each word in the collection plotted against its rank. To assess whether the data aligns with Zipf's law, we applied a linear trendline to find the best-fit straight line. The reliability of this line is highest when the R-squared ($R^2$) value is close to one. In the case of the entire corpus, the $R^2$ value was 0.997 and for each of the eight genres, $R^2$ was close to 0.998. This finding indicates a close adherence of the corpus to Zipf's law.

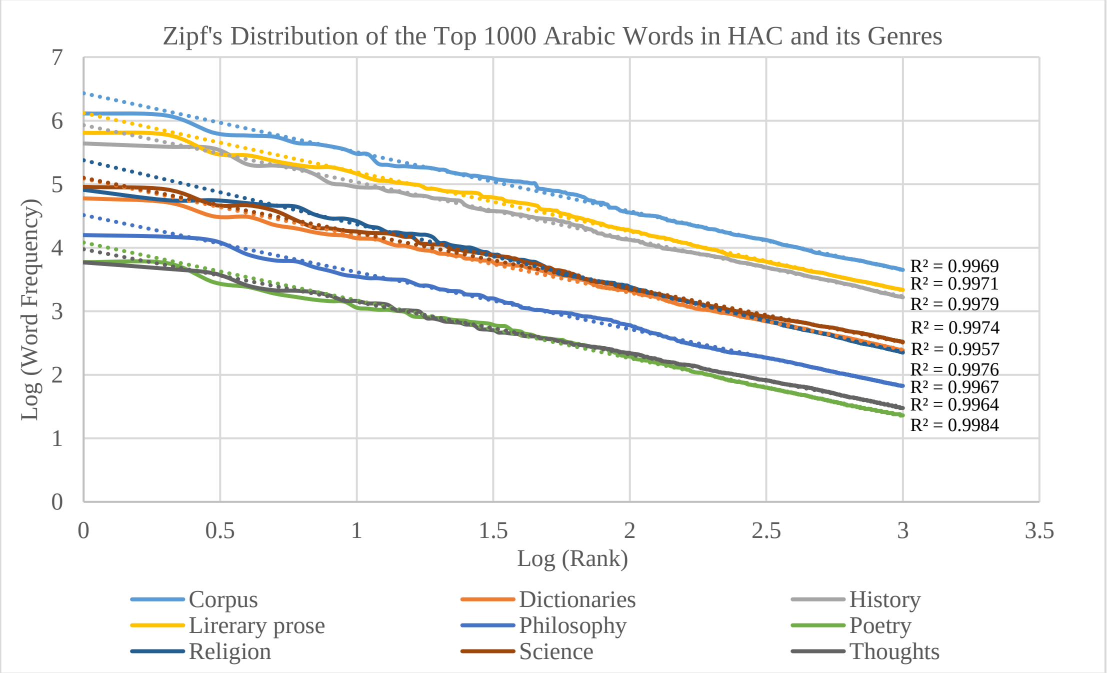


Figure 5. Zipf's distribution of the top 1000 words in the entire HAC corpus and per each of its eight genres.

## 5.4 Experimenting with the Concordancer

The potential behind developing the new concordancer was to enable students and researchers to explore new research questions and uncover hidden patterns and relationships. In addition, it allows researchers to gain new insights into the history and culture of the Arab world. The concordancer can be used in various fields, including language teaching, translation, linguistic research and text analysis. In general, it allows researchers to (1) explore new research questions and uncover hidden patterns and relationships and (2) gain new insights into the history and culture of the Arab world.

### 5.4.1 Studying Semantic Change

The phenomenon of semantic change, also referred to as meaning change, is a widespread trend in which the meanings of words undergo alterations over time, either by acquiring new senses or losing old ones, supplanting default senses, shifting in terms of word prototype, narrowing or expanding category boundaries, undergoing pejoration or amelioration and bleaching [3], [49].

Like other languages, Arabic has undergone significant changes over the centuries and many words have shifted meaning to adapt to the needs of every era. Almarwaey and Ahmad [50] suggested that social, economic and political life might influence a word's meaning. Other studies indicated that many words might disappear over time and it becomes necessary to alter the original meanings of dictionaries and language books [1], [51].

### *Methodology*

We conducted a comprehensive corpus analysis to identify and illustrate semantic changes, focusing on high-frequency terms and their historical context. Our approach was designed to ensure a representative selection of terms grounded in systematic criteria.

### *a) Selection Criteria*

1. Frequency and distribution: We identified high-frequency terms across different genres and historical periods within the corpus. Terms were selected based on their consistent presence and significant occurrences, ensuring that they were well-represented across various texts.
2. Diverse representation: The selected terms, كـافر "*kafir*," حجـاب "*hijab*," لحن "*lahn*," and فتنـة "*fitnah*," were chosen to cover a range of semantic fields, including religious, social, cultural and linguistic aspects. This diversity ensures a holistic view of semantic evolution.
3. Historical coverage: We ensured that the chosen terms had a documented presence from early historical eras to contemporary times. This allows for a comprehensive analysis of their semantic trajectories over centuries.

### *b) Analytical Process*

We utilized the concordancer tool to track these terms' usage and semantic shifts over time. This involved several steps:

1. Root-based search: To capture all derivatives and inflections of the selected terms, we conducted searches using their root forms. This approach ensures that variations of each term are included in the analysis, providing a complete picture of their usage.
2. Frequency calculation: The concordancer tool calculated the frequency of each term across different genres and eras, as shown in Tables 8 and 9. This quantitative analysis highlights trends in the popularity and contextual usage of the terms.
3. Contextual analysis: We examined specific instances of the terms in various texts to understand their evolving meanings. Table 10 presents examples of these developments, extracted from different genres and historical periods.

### *c) Experiment Results and Discussion*

The analysis revealed significant semantic shifts for each of the four terms, as follows:

- كافر (*kafir*): Originally meaning "covering" or "concealing," it evolved to signify "disbelief" in the Islamic era, with further contextual variations in the modern period.
- حجـاب (*hijab*): From its early cultural and social dimensions, it has become a contemporary symbol of Islamic identity and women empowerment.
- فتنة (*fitnah*): Transitioned from "temptation" to "political chaos" or "conflict" in modern usage.

- لحن (*lahn*): Shifted from indicating "errors" in speech to representing "melodies" and "music."

Table 10 outlines a few examples of the historical development of the four terms extracted from HAC using the concordancer. As one might notice, the highest distribution of the studied terms was under three main genres: History, Literary Prose and Religion. Let's start with the root *kfr* "كـفـر," which historically meant covering or concealing. The word and its derivatives were spotted in three paragraphs from the year 600 up to the 7th Century, as depicted in Table 9. However, examining the word's derivatives in the Islamic era (7th to 14th Century), it has been spotted in 12,900 paragraphs and has come to signify disbelief or rejection of faith. Over the centuries (19th to mid-20th), the word has been spotted in 3967 paragraphs and was employed in various contexts, including religious discussions, legal matters and cultural discourse. In the contemporary era (Mid-20th to present), the word interpretation can vary widely depending on the cultural, religious and political context.

Table 8. The concordance's statistical search results of four Arabic terms across different genres.

| Genre | كفر | حجب | لحن | فتن |
|---|---|---|---|---|
| History | 6145 | 4568 | 744 | 4188 |
| Literary Prose | 6588 | 7560 | 2575 | 3569 |
| Philosophy | 236 | 39 | 20 | 26 |
| Poetry | 56 | 116 | 47 | 29 |
| Religion | 3330 | 348 | 59 | 624 |
| Science | 559 | 563 | 56 | 352 |
| Thought | 45 | 11 | 2 | 19 |

Table 9. The concordance's statistical search results of four Arabic terms across different eras.

| Era | كفر | حجب | لحن | فتن |
|---|---|---|---|---|
| before 600 | 3 | 9 | 2 | 0 |
| 600-700 | 1252 | 220 | 30 | 411 |
| 700-800 | 457 | 174 | 43 | 211 |
| 800-900 | 1294 | 1442 | 320 | 526 |
| 900-1000 | 1607 | 1770 | 591 | 682 |
| 1000-1100 | 961 | 1790 | 417 | 572 |
| 1100-1200 | 1789 | 1428 | 392 | 809 |
| 1200-1300 | 1627 | 1175 | 266 | 668 |
| 1300-1400 | 1755 | 1133 | 248 | 746 |
| 1400-1500 | 2158 | 1080 | 283 | 583 |
| 1500-1600 | 435 | 614 | 31 | 224 |
| 1600-1700 | 260 | 285 | 51 | 120 |
| 1700-1800 | 369 | 273 | 49 | 221 |
| 1800-1900 | 256 | 265 | 139 | 358 |
| 1900-2000 | 1297 | 999 | 646 | 1169 |
| after 2000 | 2414 | 957 | 282 | 1742 |

Table 10. Samples of the development of four Arabic words from HAC from different genres and eras.

| Root | Genre | Era | Text Extracted from the concordance |
|---|---|---|---|
| **كفر** | Poetry | 600-700 | الثغر: الطريق في الجبل. **الكافر**: الليل الذي يستر كل ما يقع عليه... |
| | Poetry | 600-700 | في ليلةٍ **كَفَرَ** النُّجومَ غَمَامُهَا (8) _ (1) السري: النهر الصغير... |
| | Literary prose | 1000-1100 | و**الكفر** مجتمع على الإيمانِ وضاقت الطرق بكثرة الرماح وأهل **الكفر** ... |
| | Literary prose | 1400-1500 | يَتَبَدَّلِ **الْكُفْرَ** بِالْإِيمانِ فَقَدْ ضَلَّ سَواءَ السَّبِيلِ «2» . فقلت: يا شيخ... |
| | History | 1500-1600 | واقتتلوا مَعَهم وَقتل جمَاعَة من **الْكفَّار** وَاسْتشْهدَ ثَلَاثَة من المماليك الْخَواص... |
| **حجب** | Poetry | 900-1000 | (جعل ابن حزم **حاجبا** .. سُبحانَ من جعلَ ابنَ حزم **يحجب**) وقال آخر: (**إحتجب** الكاتب... |
| | History | 1000-1100 | **الحاجب**: وهو الذي يقف على باب القاضي، **ليحجب** عنه الناس أثناء النظر في الدعاوي... |
| | History | 1900-2000 | الالتزام التام **بالحجاب** الإسلامي حيث أن المحادثات بينهما كانت تتم بواسطة امرأة تُندب لهذا الأمر،... |
| | History | after 2000 | وعملت حكومته على إلغاء **حجاب** المرأة وأمرت بالسفور... |
| **فتن** | Literary prose | 800-900 | بك والصبر عنك ما لا يكون يا غزالاً بلحظه **يفتن** النا س وفي طرفه... |
| | Literary prose | 800-900 | **الفتنة** في هذا الموضع: النعمة واللذة. ومنه قول الله جل وعز: (إنما أموالكم وأولادكم **فتنة**... |
| | Literary prose | 1900-2000 | ما يسمى **بالفتن** وثورات الطامحين والمنشقين عن طاعة قرطبة، ويكفي أن... |
| | History | after 2000 | على المسلمين باب **الفتنة** إلى اليوم. وهذا الورع الجاهل نلاحظه اليوم في تصرفات بعض المسلمين... |
| **لحن** | Literary prose | 800-900 | وَمِمَّنْ كَانَ لَا **يُلَحِّن** أَلْبَتَّهُ حَتَّى كَأَنَّ لِسَانه أَعْرَابِيّ فَصِيح أَبُو زَيْد اَلنَّحْوِيّ وَأَبُو سَعِيد... |
| | Literary prose | 800-900 | وهي الايجاز والابتعاد عن **اللحن**، ووضوح المعنى واللفظ، وعدم اللجوء الى الزخرفة البيانية،... |
| | Literary prose | 1900-2000 | وفي فمي **لحن** وشهد وراح فالراح في البيت الأخير على العكس من الراح... |
| | Literary prose | after 2000 | إيقاع واحد، والجمع: أناشيد. وإن كان الإنشاد للشعر قد يصحبه **تلحين** وحسن إيقاع،... |

Similarly, Table 10 shows the root *hjb* “حجب” across different eras, from its conceptual origins in the pre-Islamic era to its various cultural and social dimensions in later eras. In its original meaning, *hajaba* meant “to hide” and the role of *hajeb* was prestigious in the Islamic-Caliphate periods. The contemporary era has seen a revival in the use of the *hijab* as a symbol of Islamic identity, modesty, faith and empowerment for many women. Initially, the root *ftn* “فتن” meant “temptation”. However, the meaning of *fitna* in contemporary Arabic usage has shifted to be associated with political chaos or conflict over time. Some other words have undergone semantic elevation, moving from negative meanings to ones that are now significant and positive. For example, the root *lhn* “لحن” was used to signify errors or discord in speech and it has transformed to refer to the sweetness of melodies and music.

These are just a few examples of the many Arabic words that have changed meaning over time. These changes in meaning can be challenging for researchers and scholars working with historical Arabic texts, as they may need to consider the historical context to interpret the definition of a word or phrase accurately. Additionally, they may need to be aware of these changes in meaning when comparing historical texts to contemporary usage.

These findings, supported by the data in Tables 8, 9 and 10, illustrate the comprehensive corpus analysis and ensure that our study provides a robust historical corpus and tools to advance research in Arabic linguistics and NLP.

## 6. Conclusion and Future Work

The Historical Arabic Corpus (HAC) and the developed tools provide an excellent resource for extracting historical semantic knowledge. The importance of HAC lies in its rich and extensive history, spanning over 1600 years and providing a unique perspective on the development of the historical Arabic context. These texts are a valuable resource for scholars and researchers seeking to study the Arabic language and to understand the Islamic world's cultural, social and political contexts and have contributed significantly to the field of history.

The HAC corpus can be a valuable resource for both native speakers and foreign learners of Arabic-language learning. For native speakers, HAC offers opportunities to explore classical Arabic texts, deepen their linguistic proficiency and engage with cultural heritage. On the other hand, foreign learners can utilize HAC to enhance their understanding of classical Arabic vocabulary and immerse themselves in historical contexts. Instructors can incorporate HAC into lesson plans by assigning readings, conducting comparative analyses between modern and historical Arabic texts and guiding discussions on linguistic evolution. Learners can independently utilize HAC for vocabulary expansion, comprehension tasks and research projects on specific historical periods.

We have created a collection of tools for handling and experimenting with HAC. The corpus builder incorporates a stemmer and a tagger to annotate and manipulate documents and save them in a database. We tokenized and normalized the corpus words into an indexer for efficient searching. We also created an easy-to-use concordancer to assist in searching and extracting linguistic knowledge from HAC, as well as helping in compiling dictionary entries for a hypothetical historical dictionary.

Our goal is to improve HAC by providing more accurate annotation, enlarging the corpus to represent Arabic more thoroughly, optimizing and adding features to the search engine and the concordancer, responding to the needs of linguists and offering more flexibility to meet their satisfaction.

## Acknowledgments

This research was completed while the first author was on sabbatical leave from The University of Jordan for the academic year 2021/2022 to Princess Sumaya University for Technology, Amman, Jordan. The authors would like to thank Omaima Ismail for her initial contribution to the HAC project.

## References

[1] R. Laatar, C. Aloulou and L. Hadrich Belguith, "Towards a Historical Dictionary for Arabic Language," International Journal of Speech Technology, vol. 25, no. 1, pp. 29-41, 2022.

[2] B. Hammo, S. Yagi, O. Ismail and M. Abushariah, "Exploring and Exploiting a Historical Corpus for Arabic," Language Resources & Evaluation, vol. 50, pp. 839-861, DOI:10.1007/s10579-015-9304-9, 2016.

[3] O. Ismail, S. Yagi and B. Hammo, "Corpus Linguistic Tools for Historical Semantics in Arabic," International Journal of Arabic-English Studies, vol. 15, pp. 135-152, 2014.

[4] E. Al-Thwaib, B. H. Hammo and S. Yagi, "An Academic Arabic Corpus for Plagiarism Detection: Design, Construction and Experimentation," Int. Journal of Educational Technology in Higher Education, vol. 17, no. 1, DOI:10.1186/s41239-019-0174-x, 2020.

[5] A.F. Mukhamadiarova, "Application of Corpus-based Technologies in the Formation of Lexical and Grammatical Skills in German," Perspectives of Science and Education, vol. 53, pp. 247-259, DOI:10.32744/pse.2021.5.17, 2021.

[6] S. P. Cheng, "University Students' Perceived Benefits and Difficulties Related to Corpus-assisted Translation," Compilation and Translation Review, vol. 16, no. 1, pp. 81-132, 2023.

[7] A. Boulton, "Data-driven Learning: Taking the Computer out of the Equation," Language Learning, vol. 60, no. 3, pp. 534-572, 2010.

[8] L. Zhao, W. Kong and C. Wang, "Electricity Corpus Construction Based on Data Mining and Machine Learning Algorithm," Proc. of the 2020 IEEE 5th Information Technology and Mechatronics Engineering Conf. (ITOEC), pp. 1478-1481, Chongqing, China, 2020.

[9] A. O'keeffe, M. McCarthy and R. Carter, From Corpus to Classroom: Language Use and Language Teaching, DOI: 10.1017/CBO9780511497650, Cambridge University Press, 2007.

[10] T. Wambsganss, T. Kueng, M. Soellner and J. M. Leimeister, "ArgueTutor: An Adaptive Dialog-based Learning System for Argumentation Skills," Proc. of the 2021 CHI Conf. on Human Factors in Computing Systems, pp. 1-13, DOI: 10.1145/3411764.3445781, May 2021.

[11] Essential Corpus Tools, [Online], Available: https://corpus-analysis.com/, Last visited in June 2024.

[12] S. Khoja, "Khoja's Stemmer," [Online], Available: http://zeus.cs.pacificu.edu/shereen/research.htm, 2015. Accessed April 2024.

[13] K. Toutanova, D. Klein, C. Manning and Y. Singer, "Feature-rich Part-of-speech Tagging with a Cyclic Dependency Network," Proc. of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2003), pp. 252-259, 2003.

[14] A. Abdelali, K. Darwish, N. Durrani and H. Mubarak, "Farasa: A Fast and Furious Segmenter for Arabic," Proc. of the 2016 Conf. of the North American Chapter of the Association for Computational Linguistics: Demonstrations, pp. 11-16, San Diego, California, 2016.

[15] O. Ossama, N. Zalmout, S. Khalifa, D. Taji, M. Oudah, B. Alhafni, G. Inoue, F. Eryani, A. Erdmann and N. Habash, "CAMeL Tools: An Open Source Python Toolkit for Arabic Natural Language Processing," Proc. of the 12th Language Resources and Evaluation Conf., LREC, pp. 7022-7032, Marseille, France, 2020.

[16] J. Zare, S. Karimpour and K. Aqajani Delavar, "Classroom Concordancing and English Academic Lecture Comprehension: An Implication of Data-driven Learning," Computer Assisted Language Learning, vol. 36, nos. 5-6, pp. 885-905, DOI:10.1080/09588221.2021.1953081, 2023.

[17] J. Zare, S. Karimpour and K. A. Delavar, "The Impact of Concordancing on English Learners' Foreign Language Anxiety and Enjoyment: An Application of Data-driven Learning," System, vol. 109, p. 102891, DOI:10.1016/j.system.2022.102891, 2022.

[18] V. Mohammadi and N. Mohit, "Student and Teacher Attitude toward Using Concordancing in Learning and Teaching Preposition Collocations: Issues and Options," Journal of Language Horizons, vol. 5, no. 2, pp. 139-166, 2021.

[19] I. Kazaz, "Alternative Vocabulary Assessment: Using Concordance Line Activities for Testing Lexical Knowledge," Int. Online Journal of Education and Teaching, vol. 7, no. 3, pp. 1221-1238, 2020.

[20] A. T. Shawaqfeh and M. A. Khasawneh, "Incorporating Corpus Linguistics Tools in the Training and Professional Development of Lecturers in Translation Studies," Studies in Media and Communication, vol. 11, no. 7, p. 260, DOI:10.11114/smc.v11i7.6379, 2023.

[21] M. del Mar Sánchez Ramos, "Teaching English for Medical Translation: A Corpus-based Approach," Iranian Journal of Language Teaching Research, vol. 8, no. 2, pp. 25-40, 2020.

[22] O. J. Ballance and A. Coxhead, "How Much Vocabulary is Needed to Use a Concordance?" Int. Journal of Corpus Linguistics, vol. 25, no. 1, pp. 36-61, DOI:10.1075/ijcl.17116.bal, 2020.

[23] S. Un-udom and N. Un-udom, "A Corpus-based Study on the Use of Reporting Verbs in Applied Linguistics Articles," English Language Teaching, vol. 13, no. 4, pp. 162-169, 2020.

[24] M. Bednarek and G. Carr, "Computer-assisted Digital Text Analysis for Journalism and Communications Research: Introducing Corpus Linguistic Techniques That Do not Require Programming," Media International Australia, vol. 181, no. 1, pp. 131-151, DOI: 10.1177/1329878X20947124, 2021.

[25] A. Eddakrouri, "Arabic Corpus of Library and Information Science: Design and Construction," Egyptian Journal of Language Engineering, vol. 10, no. 1, pp. 1-9, DOI:10.21608/ejle.2023.183529.1040, 2023.

[26] S. Khoja, "An RSS Feed Analysis Application and Corpus Builder," Proc. of the 2nd Int. Conf. on Arabic Language Resources and Tools, pp. 115-118, Cairo, Egypt, 2009.

[27] M. O'Donnell, "The UAM Corpus Tool: Software for Corpus Annotation and Exploration," Proc. of Bretones Callejas, Carmen M. et al. (eds.) Applied Linguistics Now: Understanding Language and Mind, Almería: Universidad de Almería, pp. 1433-1447, 2008.

[28] S. Alansary, M. Nagi and N. Adly, "Towards Analyzing the International Corpus of Arabic (ICA): Progress of Morphological Stage," Proc. of the 8th Int. Conf. on Language Engineering, Cairo, Egypt, 2008.

[29] M. Attia, P. Pecina, L. Tounsi, A. Toral and J. van Genabith, "Lexical Profiling for Arabic," Proc. of eLex 2011, pp. 23-33, 2011.

[30] K. Dukes and N. Habash, "Morphological Annotation of Quranic Arabic," Proc. of the 7th Int. Conf. on Language Resources and Evaluation (LREC'10), pp. 2530-2536, Valletta, Malta, 2010.

[31] M. Boella, F. Romani, A. Al-Raies, C. Solimando and G. Lancioni, "The SALAH Project: Segmentation and Linguistic Analysis of Ḥadīṯ Arabic Texts," Proc. of Information Retrieval Technology, Part of the Book Series Lecture Notes in Computer Science, vol. 7097, pp. 538-549, DOI: 10.1007/978-3-642-25631-8_49, Springer, Berlin, Heidelberg, 2011.

[32] A. Sharaf and E. Atwell, "QurAna: Corpus of the Quran Annotated with Pronominal Anaphora," Proc. of the 8th Int. Conf. on Language Resources and Evaluation (LREC'12), pp. 130-137, Istanbul, Turkey, 2012.

[33] S. Altammami, E. Atwell and A. Alsalka, "Constructing a Bilingual Hadith Corpus Using a Segmentation Tool," Proc. of the 12th Language Resources and Evaluation Conf., pp. 3390-3398, Marseille, France, 2020.

[34] M. Hajjar, A. Al-Hajjar, K. Zreik and P. Gallinari, "An Improved Structured and Progressive Electronic Dictionary for the Arabic Language: iSPEDAL," Proc. of the 5th Int. Conf. on Internet and Web Applications and Services (ICIW), pp. 489-495, Barcelona, Spain, 2010.

[35] B. Hammo, F. Al-Shargi, S. Yagi and N. Obeid, "Developing Tools for Arabic Corpus for Researchers," Proc. of the 2nd Workshop on Arabic Corpus Linguistics (WACL-2), Lancaster University, UK, 2013.

[36] Z. Alyafeai, M. Masoud, M. Ghaleb and M. S. Al-shaibani, "Masader: Metadata Sourcing for Arabic Text and Speech Data Resources," Proc. of the 13th Language Resources and Evaluation Conf., pp 6340–6351, European Language Resources Association, Marseille, France, 2022.

[37] The Linguistic Data Consortium, [Online], Available: https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/arabic.pdf. (Last visited in June 2024).

[38] M. Alrabiah, A. Al-Salman and E. Atwell, "The Design and Construction of the 50 Million Words KSUCCA," Proc. of the 2nd Workshop on Arabic Corpus Linguistics (WACL-2), Lancaster University, UK, 2013.

[39] A. O. Al-Thubaity, "A 700 M + Arabic Corpus: KACST Arabic Corpus Design and Construction," Language Resources and Evaluation, vol. 49, pp. 721-75, DOI: 10.1007/s10579-014-9284-1, 2015.

[40] L. Anthony, "AntConc: Design and Development of a Freeware Corpus Analysis Toolkit for the Technical Writing Classroom," Proc. of Professional Communication Conf. (IPCC 2005), pp. 729-737, Limerick, Ireland, 2005.

[41] A. Roberts, L. Al-Sulaiti and E. Atwell, "aConCorde: Towards an Open-source, Extendable Concordancer for Arabic," Corpora, vol. 1, no. 1, pp. 39-60, 2006.

[42] R. Abbès and J. Dichy, "AraConc, an Arabic Concordance Software Based on the DIINAR.1 Language Resource," Proc. of the 6th Int. Conf. on Informatics and Systems, pp. 127-134, 2008.

[43] Y. Belinkov, A. Magidow, M. Romanov, A. Shmidman and M. Koppel, "Shamela: A Large-scale Historical Arabic Corpus," Proc. of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH), pp. 45–53, Osaka, Japan, 2016.

[44] A. Hourani, A History of the Arab Peoples: Updated Edition, ISBN: 9780571288014, London: Faber and Faber, 2013.

[45] L. Larkey and M. Connell, "Arabic Information Retrieval at UMass in TREC-10," Proc. of the 10th Text Retrieval Conference (TREC-10), pp. 562-570, Maryland, USA, 2010.

[46] R. H. Al Mahmoud, B. H. Hammo and H. Faris, "Cluster-based Ensemble Learning Model for Improving Sentiment Classification of Arabic Documents," Natural Language Engineering, pp. 1-39, DOI: 10.1017/S135132492300027X, 2023.

[47] R. H. AlMahmoud and B. H. Hammo, "SEWAR: A Corpus-based N-gram Approach for Extracting Semantically-related Words from Arabic Medical Corpus," Expert Systems with Applications, vol. 238, p. 121767, DOI: 10.1016/j.eswa.2023.121767, 2014.

[48] G. K. Zipf, "The Meaning-Frequency Relationship of Words," The Journal of General Psychology, vol. 33, no. 2, pp. 251-256, DOI:10.1080/00221309.1945.10544509, 1945.

[49] N. N. Hanifah, "The Origin of Arabic Lexicography: Its Emergence and Evolution," HuRuf Journal: Int. Journal of Arabic Applied Linguistic, vol. 1, no. 2, pp. 238-251, DOI: 10.30983/huruf.v1i1.4932, 2021.

[50] A. O. Almarwaey and U. K. Ahmad, "Semantic Change of Hijab, Halal and Islamist from Arabic to English. 3L: Language, Linguistics, Literature," The Southeast Asian Journal of English Language Studies, vol. 27, no. 2, pp. 161-176, DOI: 10.17576/3L-2021-2702-12, 2021.

[51] R. Laatar, A. Rhayem, C. Aloulou and L. H. Belguith, "Towards a Historical Ontology for Arabic Language: Investigation and Future Directions," Proc. of the Int. Conf. on Intelligent Systems Design and Applications, Part of the Book Series: Lecture Notes in Networks and Systems, vol. 418, pp. 1078-1087, Cham: Springer International Publishing, December 2021.

## ملخص البحث:

تسْتكشف هذه الورقة تطوير وتصميم وإعادة بناء مجموعة نصوص تاريخية باللّغة العربية (HAC) تُغطّي فترة زمنية تتجاوز 1600 سنة من الاستخدامات اللُّغوية غير المنقطعة. وتؤكّد الورقة الجوانب التّقنية المتّبعة لتحسين النّظام، وتعرض التّجارب الّتي تمّ إجراؤها على مجموعة النّصوص. هذا مع الإشارة إلى أنّ اللُّغة العربية تمتلك إرثاً أدبياً وثقافياً يمتدّ لآلاف السّنين. وقد جعل شمول المصادر الرّقمية بالإضافة إلى التّقدّم في تقنيات معالجة اللّغات الطّبيعية مجموعات النّصوص التّاريخية باللّغة العربية ذات أهمّية متزايدة للباحثين والمتعلمين في جميع أنحاء العالم.

ومن خلال دمْج مجموعة النُّصوص باللُّغة العربية وأدواتها في تعلُّم اللغة العربية، يُمْكن للمعلِّمين أنِ يغوصوا على نحوٍ أعمق في المصْطَلحات اللُّغوية والثّقافة العربية ويكتسبوا نظرةً ثاقبةً في سبيل تحسين المهاراتٍ اللغوية وفهْم اللُّغة العربية.

والجدير بالذّكر أنّ دمْج الإرشادات البشرية وتقنيات معالجة اللُّغات الطّبيعية من شأنه أنِ يجعل تعلُّم اللُّغة أمراً مُمتعاً، وذلك من خلال تقديم طريقةٍ ديناميكية وأصيلة لإتقان اللُّغة العربية.