


 <p>ISSN 0 2791-0169 / e-ISSN 2790-0127</p> <p><b>Science, Education and Innovations</b> in the Context of Modern Problems</p> <p>Editor-in-Chief   Dr. Rahil Najafov</p> <p>www.imcra-az.org</p>	<p>Science, Education and Innovations in the Context of Modern Problems Issue 4, Vol. 9, 2026</p>
<p>RESEARCH ARTICLE </p>	
<p style="text-align: center;"><b>From Traditional Arabic Lexicography to Intelligent Digital Ecosystems: A Comprehensive Study on the Linguistic Foundations, Computational Modeling, and Artificial Intelligence Integration in the Development of Smart Arabic Dictionaries</b></p>	
<p><b>Abdelmalek Belmihoub</b></p>	<p>Dr. Mohamed El Bachir El Ibrahim University, Bordj Bou Arreridj Algeria E-mail: Abd.belmihoub@univ-adrar.edu.dz</p>
<p><b>Amar Benseghier</b></p>	<p>Dr. Mohamed El Bachir El Ibrahim University, Bordj Bou Arreridj Algeria E-mail: amar.benseghier@univ-bba.dz</p>
<p><b>Abdelghani Nasri</b></p>	<p>Dr. Mohamed El Bachir El Ibrahim University, Bordj Bou Arreridj Algeria E-mail: abdelghani.nasri@univ-bba.dz</p>
<p><b>Fouad Aldji</b></p>	<p>Dr. Mohamed El Bachir El Ibrahim University, Bordj Bou Arreridj Algeria E-mail: fouad.aldji@univ-bba.dz</p>
<p><b>Keywords</b></p>	<p>Arabic lexicography; smart dictionaries; computational linguistics; digital lexicons; artificial intelligence in language processing; natural language processing (NLP); Arabic corpora; morphological analysis; semantic search; machine learning; lexical databases; TEI-Lex0; Lexical Markup Framework (LMF); language technologies; digital humanities</p>

**Abstract**

The field of Arabic lexicography is undergoing a profound paradigmatic transformation driven by rapid advancements in digital technologies and artificial intelligence. Moving beyond the limitations of traditional paper-based dictionaries, contemporary lexicographical practices increasingly rely on computational linguistics, large-scale corpora, and machine learning techniques to construct dynamic and intelligent lexical systems. This study provides a comprehensive and systematic analysis of the evolution of Arabic dictionaries from their classical roots to modern digital and smart frameworks.

The paper critically examines the structural, conceptual, and technological shifts that distinguish traditional, digital, and smart dictionaries, with particular emphasis on their linguistic architecture, data representation models, and user interaction capabilities. It highlights the role of international standards such as the Lexical Markup Framework (LMF) and TEI-Lex0 in ensuring interoperability, consistency, and scalability of lexical resources. Furthermore, the study evaluates leading Arabic and international computational models, including transformer-based architectures and morphological analyzers, in enhancing semantic search, contextual understanding, and automated language processing.

In addition, the research addresses persistent challenges in Arabic language processing, including the complexity of morphological structures, dialectal diversity, ambiguity resolution, and the scarcity of high-quality, annotated corpora. By integrating theoretical linguistics with advanced computational methodologies, the study proposes a forward-looking framework for the development of intelligent Arabic dictionaries capable of supporting multilingual environments, adaptive learning systems, and real-time linguistic analysis.

Ultimately, this work contributes to bridging the gap between traditional lexicographical heritage and modern artificial intelligence applications, offering a strategic roadmap for strengthening the digital presence of the Arabic language and advancing its role in global knowledge systems.

#### Citation

Abdelmalek Belmihoub, Amar Benseghier, Abdelghani Nasri, Fouad Aldji. (2026). From Traditional Arabic Lexicography to Intelligent Digital Ecosystems: A Comprehensive Study on the Linguistic Foundations, Computational Modeling, and Artificial Intelligence Integration in the Development of Smart Arabic Dictionaries. *Science, Education and Innovations in the Context of Modern Problems*, 9(4), 1-15. <https://doi.org/10.56334/sei/9.4.10>

#### Licensed

© 2026 The Author(s). The Author(s). Published by Science, Education and Innovations in the context of modern problems (SEI) by IMCRA - International Meetings and Journals Research Association (Azerbaijan). This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

Received: 25.11.2025

Accepted: 15.02.2026

Published: 20.03.2026 (available online)

## Introduction:

Throughout history, Arabic lexicographical production has played a crucial role in preserving the Arabic language and safeguarding its cultural and intellectual identity. The traditional printed dictionary has long constituted a central pillar in compiling the lexicon, clarifying its derivations, and explicating its semantic nuances, from Al-Khalil ibn Ahmad Al-Farahidi's *Kitab Al-'Ayn* to the dictionaries of the medieval and modern periods. With the rapid technological transformations of recent decades, however, there has emerged an urgent need to reconceptualize the Arabic dictionary so that it aligns with digital tools and international standards, while capitalizing on the capabilities of artificial intelligence in natural language processing and the analysis of large-scale textual data.

The shift from the printed to the digital dictionary has generated a set of theoretical and technical challenges. Chief among these are the normative lexicographical foundations required to construct an electronic lexical database, the substantial differences between the structure of the traditional dictionary and that of the intelligent digital dictionary, and the complex issues related to diacritization, dialectal variation, and the representation of meanings and contexts. At the same time, developments in artificial intelligence have opened unprecedented opportunities for semantic search, the modelling of contextual relations between lexical items, and the provision of advanced language services such as intelligent translation, automatic vocalization, and grammatical error correction.

This study seeks to analyze the structural and conceptual transformation in Arabic lexicography and to explore the foundations for constructing smart dictionaries that rely on artificial intelligence techniques. It addresses the theoretical framework of lexicographical practice, reviews leading international and Arabic digital models, and identifies the challenges that currently impede the development of a competitive Arabic smart dictionary. The paper draws on the latest literature and comparative studies, emphasizing the necessity of closer integration between linguists and developers, as well as the creation of large-scale Arabic corpora that adhere to open, internationally recognized standards. Ultimately, it aims to outline a clear roadmap for a more interactive and accurate Arabic lexicographical industry in the age of digitization and artificial intelligence.

## Methodology

This study adopts a qualitative, interdisciplinary research methodology that integrates doctrinal linguistic analysis with computational and conceptual modeling approaches to examine the evolution of Arabic lexicography and propose a framework for intelligent digital dictionaries.

First, a doctrinal and descriptive approach is employed to analyze classical and modern Arabic lexicographical traditions. This includes a systematic examination of historical dictionaries, theoretical linguistic foundations, and contemporary digital lexicographic practices. Foundational works in Arabic lexicography, along with modern computational linguistics literature, are critically reviewed to identify structural and conceptual transformations.

Second, the study applies a comparative analytical method to distinguish between traditional, digital, and smart dictionaries. This comparison is conducted across multiple dimensions, including data representation, search mechanisms, semantic processing, user interaction, and integration of artificial intelligence technologies. International standards such as the Lexical Markup Framework (LMF) and TEI-Lex0 are used as benchmarks for evaluating structural consistency and interoperability.

Third, the research adopts a computational conceptual modeling approach, focusing on the design and architecture of a proposed intelligent Arabic dictionary system. Rather than implementing a fully operational system, the study develops a multi-layered architectural model based on established natural language processing tools and frameworks, including transformer-based models (e.g., AraBERT), morphological analyzers (e.g., MADAMIRA), and semantic networks (e.g., Arabic WordNet).

Additionally, the study incorporates a **case-oriented analytical perspective**, examining existing computational tools and linguistic resources such as BAREC, Tharwa Lexicon, and Sketch Engine. These tools are analyzed in terms of their functionality, scalability, and applicability to intelligent lexicographic systems.

The research is based on secondary data sources, including peer-reviewed journal articles, conference proceedings, and publicly available linguistic datasets. No primary empirical data collection was conducted. The methodological framework aims to bridge theoretical linguistics and applied computational approaches, offering a comprehensive foundation for future empirical implementation and evaluation.

### Section One: Conceptual Framework and Literature Review:

The dictionary is one of the central components of linguistic studies, as it serves as the repository in which a language's lexical wealth is preserved and its meanings and usages are systematized. It has been defined in lexicographic scholarship as "a linguistic reference work concerned with collecting lexical items, explaining them, and indicating their functions in actual usage" (Atkins & Rundell, 2008, p. 3). Lexical linguistics distinguishes between the general dictionary, which covers the language in its broad scope, and the specialized dictionary, which serves particular domains of knowledge such as scientific, medical, and technical lexicons. With the development of computational linguistics and its applications, dictionaries have increasingly become components of knowledge structures that are amenable to automatic processing, which has led to the establishment of new international standards for their digital representation such as the Lexical Markup Framework (LMF-ISO 24613) and TEI-Lex0.

From its earliest stages, Arabic lexicography has been characterized by a meticulous methodology in compiling and organizing lexical items. It began with the pioneering work of Al-Khalil ibn Ahmad Al-Farahidi in *Kitab Al-'Ayn*, then evolved through Ibn Faris's *Maqāyīs Al-Lughā* and Al-Jawhari's *Al-Sihāh*, culminating in Ibn Manzur's *Lisān Al-'Arab*. These dictionaries relied primarily on root-based organization, whereby each lexical item is traced back to its trilateral or quadrilateral root, and its meanings are then explained and illustrated with examples drawn from literary and Qur'anic texts. This approach is closely tied to the notion of derivation, which is a core concept in the structure of the Arabic morphological system. However, despite their historical and cultural importance, these dictionaries face several challenges, including the difficulty of locating entries for non-specialist users, the lack of a unified structure for presenting information, and the limited possibilities for rapid access to target items. With the advent of modern printing, a range of new dictionaries appeared, such as *Al-Wasīt* and *Al-Mu'jam Al-Kabīr*, which sought to reconcile lexicographical authenticity with contemporary practical needs.

The digital dictionary represents a decisive transitional stage between the printed dictionary and the contemporary smart dictionary. In this sense, it may be defined as the re-encoding of lexicographical content within an electronic structure that supports fast lookup and organizes information in formats suitable for computational processing (Habash, 2010, p. 28). Electronic dictionaries possess a number of features that make them more efficient than their printed counterparts, most notably instantaneous search capabilities, dynamic cross-referencing between lexical entries, and the integration of multimedia components such as audio, pronunciation, and usage examples, alongside the possibility of continuous content updating. Nonetheless, in its basic form, the digital dictionary does not yet embody artificial intelligence, as it merely offers an electronic structure without the capacity to predict context or automatically analyze large-scale textual data. International standards such as TEI-Lex0 and LMF (ISO 24613) have therefore been adopted to regulate the structure of electronic dictionaries, enabling the standardized representation of roots, patterns, inflectional forms, grammatical information, and semantic relations in a way that is programmatically accessible and computationally processable (Bański, Bowers, & Erjavec, 2017, p. 45).

The smart dictionary constitutes the most recent stage in the evolution of lexicographical production. It is a dictionary that deploys advanced artificial intelligence techniques such as machine learning, large language models, and contextual lexical analysis. It is distinguished by its ability to interpret context rather than merely presenting static information, to provide semantic suggestions based on deep learning models, and to rely on large corpora to construct its knowledge in a dynamic manner (Habash, 2010, pp. 112-125). Smart dictionaries also offer advanced services such as spell-checking and grammar correction, as well as integrated morphological and contextual analysis, and they can rank senses according to their frequency in actual usage. Systems such as AraBERT and MADAMIRA are leading examples of the technologies on which this type of dictionary depends. Antoun, Baly, and Hajj (2020), for instance, introduced AraBERT as a transformer-based model tailored to Arabic and trained on a massive corpus of around 200 million sentences and over 8.6 billion tokens, achieving state-of-the-art results on various natural language processing tasks. MADAMIRA (Pasha et al., 2014, pp. 1094-1101), in turn, combines comprehensive morphological analysis with fine-grained tokenization, providing detailed information on roots, stems, inflection, and grammatical features, thereby facilitating the development of intelligent, extensible lexicographical systems.

Recent literature on digital and smart lexicography has witnessed diverse and influential Arab and international contributions. Among the most notable are the works of Kilgarriff and his collaborators on the Sketch Engine system (Kilgarriff, Rychlý, Smrž, & Tugwell, 2004), which laid the foundations for corpus-based lexicography and introduced an advanced mechanism for extracting “word sketches,” that is, concise, automatically generated summaries of a word’s syntactic behavior and collocational patterns. In the Arab context, official language academies have pursued large-scale projects such as Al-Mu‘jam Al-Kabīr, though their efforts in the computational and digital dimension remain limited. In parallel, pioneering initiatives such as Arabic WordNet (Elkateb, Black, Rodríguez, Alkhalifa, Vossen, Pease, & Fellbaum, 2006, pp. 1–9) have sought to model semantic relations between Arabic lexical items following the structure of the English Princeton WordNet, thereby enabling cross-linguistic conceptual mapping and opening new horizons for machine translation and semantic search. These contributions have been presented in major international venues such as the Language Resources and Evaluation Conference (LREC). They have capitalized on advances in Arabic language processing tools and have helped establish a solid knowledge base for the further development of Arabic lexicography. Nevertheless, despite this notable progress, Arabic lexicographical production still faces substantial structural challenges, including the scarcity of large, reliable corpora constructed according to clear international criteria, the lack of standardized practices across projects in the Arab world, and the persistent gap between theoretical linguistics and practical computation in the design of such systems.

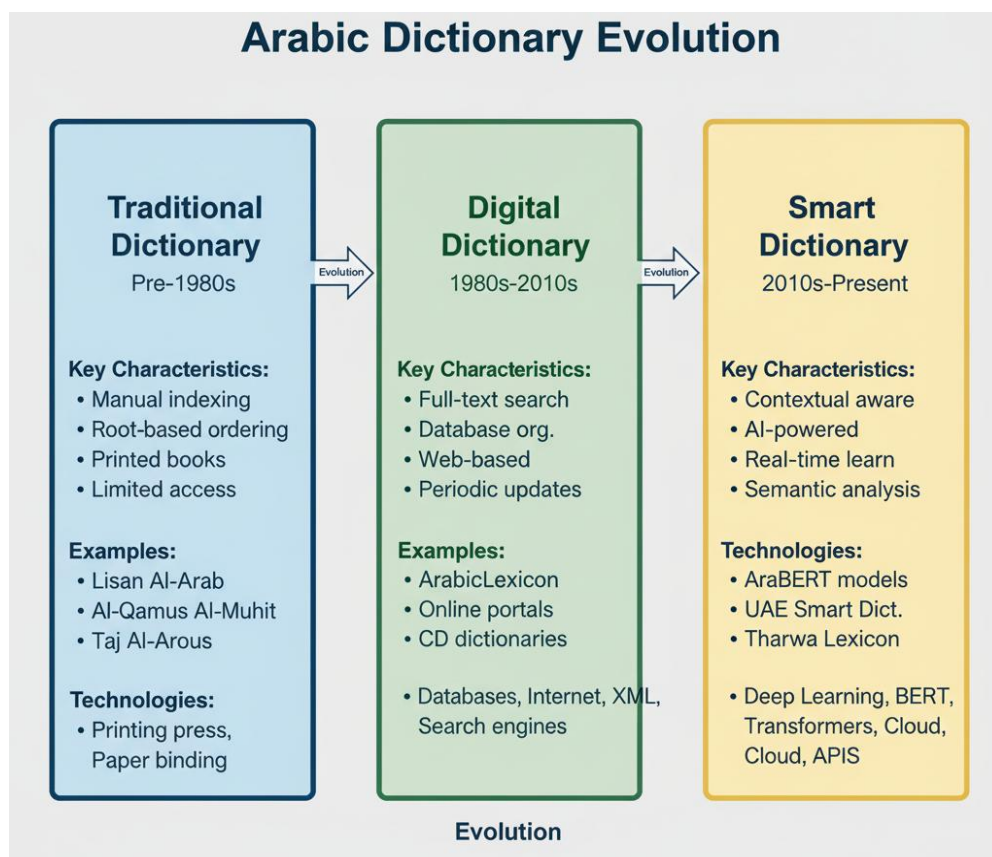
**Table 1: Comprehensive comparison between the three types of dictionaries and their features**

Characteristic	Traditional/Paper	Digital	Smart
<b>Search capability</b>	Manual browsing through indices and root-based ordering	Rapid full-text search with various indexing methods	Intelligent contextual search with semantic understanding
<b>Update frequency</b>	Infrequent (years between editions)	Regular updates possible	Real-time continuous updates
<b>User interactivity</b>	Non-interactive, static display	Interactive browsing with links and cross-references	Highly interactive with personalization
<b>Multimedia support</b>	Text and illustrations only	Audio, video, images integrated	Full multimedia integration with adaptive content
<b>Contextual understanding</b>	None - fixed meanings	Limited - static context indicators	Deep contextual analysis using AI models
<b>Semantic analysis</b>	Limited to pre-defined relationships	Structured semantic relationships in database	Complex semantic networks with relation inference
<b>Error detection</b>	None	None	Automatic spelling, grammar, and semantic error detection
<b>Automatic vocalization</b>	None	Automatic pronunciation available	Intelligent pronunciation with dialect variations
<b>Multilingual support</b>	Limited translation support	Multi-language databases possible	Seamless multilingual translation and adaptation
<b>AI integration</b>	None	Minimal (basic algorithms)	Deep learning models (BERT, Transformers)
<b>Data volume</b>	Limited to physical constraints	Scalable - can contain millions of entries	Massive distributed data processing (billions of entries/contextual variations)

**Figure 1: Evolutionary diagram illustrating three distinct eras in lexicography:**

- The Traditional Era (pre-1980): Paper dictionaries with examples such as Lisan al-Arab and Al-Qamus al-Muḥit
- The Digital Era (1980-2010): Electronic dictionaries with examples such as ArabicLexicon.Hawramani and online dictionaries

- The Smart Era (2010-present): AI-powered smart dictionaries with examples such as AraBERT and the Emirates Smart Dictionary.



#### Section Two: From Traditional to Digital Dictionaries:

The transition from traditional paper-based dictionaries to digital dictionaries represents a fundamental and necessary step in the development of the Arabic language industry. It enables the use of computer technologies to facilitate access to and analysis of linguistic data in unprecedented ways. Traditional dictionaries relied on a fixed presentation of lexical entries, arranged either by root or alphabetically, supplemented by citations from literary and Quranic texts, as seen in Ibn Manzur's *Lisan al-Arab* and Al-Khalil ibn Ahmad al-Farahidi's *Kitab al-Ayn*. However, these dictionaries faced significant challenges, including the difficulty of rapid searching, the disorganized repetition of information, and the inability to automatically link semantic and morphological data (Atkins and Rundell, 2008, pp. 45-65).

The digital dictionary emerged as a powerful and effective solution to these structural challenges, representing lexicographic content in an electronic format that supports computational processing and instant querying. It offers advanced capabilities, including immediate search by various methods (root, initial letters, or whole word), dynamic linking between different lexical entries, and the integration of multimedia components such as audio pronunciation, practical examples, and illustrative images (Kilgarriff, 2004, pp. 105-116). Furthermore, continuous data updating and advanced version control systems ensure that the dictionary evolves in tandem with ongoing linguistic and technological advancements.

International standards are vital for ensuring consistency and compatibility across various digital projects at both regional and global levels. For example, the TEI-Lex0 standard (Text Coding Initiative Lexicon), officially launched in 2016 by the DARIAH working group and subsequently developed under the leadership of Banksy et al. (Banksy, Powers, and Erjavic, 2017, pp. 45-58), facilitates the standardized representation of roots, patterns, morphological forms, syntactic information, and semantic relations in a precise, computationally manageable format that can be shared across diverse systems. Similarly, the LMF (Lexical Coding Framework - ISO 24613) (ISO 24613:2008) is another crucial international standard for structuring lexical data and standardizing practices across different applications (Francopolo and George, 2008). These standardized criteria have been increasingly adopted by many Arab and international projects, such as the ArabicLexicon.Hawramani.com database and the Emirates Smart Dictionary project, to provide highly efficient searchable and analyzable digital dictionaries using advanced computing technologies (Habash, 2010, pp. 20-24).

On the other hand, advanced natural language processing tools have contributed significantly to enriching the digital dictionary and enhancing its research and analytical effectiveness. Programs such as MADAMIRA (Morphological Analysis and Disambiguation for Arabic) provide highly accurate capabilities for comprehensive morphological analysis of Arabic words, intelligent stemming, syntactic parsing, and linking lexical items to their contextual occurrences (Pasha et al., 2014, pp. 1094–1101). Large and curated text corpora, such as BAREC (Balanced Arabic Readability Evaluation Corpus) and SAMER (Simplification and Readability Lexicon for Arabic), have aided in training and developing digital dictionary systems to achieve deeper text comprehension, automatically assess vocabulary difficulty, and extract semantic and contextual relationships between words (Elmadani, Habash, & Taha-Thomure, 2025; Al Khalil et al., 2020). The BAREC project, for instance, provided a massive dataset containing 69,441 sentences and over one million words distributed across 19 different reading levels, ranging from kindergarten to postgraduate studies (Elmadani et al., 2025), while the SAMER project offered a specialized lexicon of 40,000 manually classified Arabic lemmas across five reading levels (Al Khalil et al., 2020).

Leading international experiments, such as the Sketch Engine system (Kilgarriff, Rychlý, Smrž, & Tugwell, 2004, pp. 105–116), demonstrate how digital dictionaries can be integrated with large text corpora to analyze word behavior deeply, providing research and analytical capabilities unavailable in traditional printed dictionaries. The system diversifies lexical data presentation through "word sketches," which are precise automatic summaries of a word's syntactic distribution, collocational patterns, and common contexts. Similarly, pioneering projects like Arabic WordNet (Elkateb, Black, Rodríguez, Alkhalifa, Vossen, Pease, & Fellbaum, 2006, pp. 1–9) have enabled the creation of advanced logical and semantic links between Arabic and English lexical items, offering a unified database deployable efficiently in subsequent smart applications, machine translation, and advanced semantic search. These groundbreaking works, which have achieved significant accomplishments, have been published in prestigious and specialized international conferences such as LREC (Language Resources and Evaluation Conference) and ACL (Association for Computational Linguistics), strongly contributing to building a solid and reliable knowledge base for the development of the Arabic lexicographical field over the past decade.

Despite the clear and numerous advantages of the digital dictionary and its advanced analytical capabilities, the transition to artificial intelligence remains an inevitable and urgent necessity to transform these dictionaries from mere data presentation tools into intelligent systems capable of precise inference and contextual semantic prediction. Yet, the digital dictionary, with all its standards, tools, and databases, effectively provides the computational and necessary foundation for building these advanced smart systems. It organizes lexical data logically and integratively and makes it available in unified, shareable, and continuously developing formats, thus making the digital phase a vital and essential bridge between the rich printed lexicographical heritage and the modern intelligent capabilities that will be discussed in the upcoming third section.

### Section Three: The Smart Dictionary and Artificial Intelligence Applications:

The smart dictionary represents the most advanced and recent stage in the evolution of lexicographical production, transitioning from a static digital database to a sophisticated, dynamic system capable of deeply understanding linguistic context, predicting meaning based on context, and analyzing large-scale texts using advanced artificial intelligence techniques. It relies on an integrated set of tools and technologies including machine learning, deep neural networks, large language models, and advanced natural language processing to provide an interactive and effective linguistic service characterized by accuracy and flexibility (Habash, 2010, pp. 112–125).

Smart dictionaries are distinguished by a set of advanced features that set them apart from traditional digital dictionaries:

- **First**, the ability for deep contextual analysis, where the intelligent system can deduce the correct and precise meaning of a lexical item based on its full context in texts and documents, rather than displaying a predefined set of meanings.
- **Second**, semantic prediction and extraction of complex semantic relationships between lexical items using advanced models such as AraBERT and MADAMIRA, where latent meanings and implicit relations between words can be inferred and classified according to context and usage (Antoun, Baly, & Hajj, 2020, pp. 9–15; Pasha et al., 2014, pp. 1094–1101).
- **Third**, flexibility, interactivity, and customization, allowing users to receive lexicographic suggestions and information tailored to their level of scientific knowledge or specialized field, making the dictionary a dynamic rather than static tool.
- **Fourth**, take advantage of large and well-organized text collections, as smart dictionaries rely on huge text databases such as BAREC Corpus, ArabicLexicon.Hawramani.com, and the Emirates Smart Dictionary to provide mathematical models with sufficient and reliable linguistic data to learn precise language patterns and rules.
- **Fifth**, offering advanced and comprehensive services such as automated spelling and grammatical error detection, automatic vocalization of unvowelled Arabic text, and synchronized integrated morphological and contextual analysis.

On the practical and applied level, several pioneering and influential smart models and applications in Arabic lexicography have emerged. AraBERT (Antoun, Baly, & Hajj, 2020, pp. 9-15) is an advanced model based on the Transformer architecture, trained specifically on a massive Arabic corpus containing 70 million sentences and about 24 GB of original Arabic texts. This model covers news from diverse media sources across multiple Arab regions, representing a wide range of linguistic topics and styles. It consists of 110 million parameters and achieves state-of-the-art results in fundamental natural language processing tasks such as sentiment analysis, named entity recognition, and question answering. AraBERT handles Arabic morphological complexities with high efficiency using smart stemming technology via Farasa, which segments words into roots, stems, prefixes, and suffixes, reducing unnecessary vocabulary sparsity and enhancing the model's generalization ability (Antoun et al., 2020, pp. 10-11).

Conversely, MADAMIRA (Morphological Analysis and Disambiguation for Arabic) is an integrated system specialized in comprehensive and precise morphological analysis, disambiguation, and syntactic analysis of Arabic words (Pasha et al., 2014, pp. 1094-1101). This system provides highly detailed and reliable information about original roots, stems, inflectional markers, and other grammatical attributes, facilitating the construction of advanced, scalable, and continuously improving smart lexicographical systems. The Arabic WordNet project (Elkateb et al., 2006, pp. 1-9) contributed a complex network of semantic and logical relations between Arabic and English lexical items, enabling strong and logical cross-lingual links and opening completely new horizons for multilingual intelligent applications.

At the institutional and governmental level, the United Arab Emirates has developed the UAE Smart Dictionary, a large and ambitious project that gathers millions of Arabic lexical entries and employs advanced AI techniques to offer sophisticated capabilities in interactive semantic search and deep linguistic analysis. The project also invests in building a large, computationally curated corpora for training smart models. Additionally, the Tharwa Lexicon project (Diab, Al-Badrashiny, Aminian, Attia, Elfardy, Habash, et al., 2014, pp. 3782-3789) offers an advanced, multilingual lexical resource covering Modern Standard Arabic and local dialects, especially the Egyptian dialect, containing over 73,000 Egyptian lexical entries with precise linguistic information such as part of speech, gender, animation, number, root, and morphological structure. Tharwa is described as the first electronic resource linking Arabic dialects, Modern Standard Arabic, and English, making it highly valuable for both theoretical and computational research.

Nevertheless, Arab smart dictionaries face a set of complex and major challenges that hinder their ongoing development.

- First, representing Arabic's complex morphological and syntactic structure, as Arabic features a very rich derivational system and extremely complex morphology, making morphological and contextual analysis far more complicated than in Indo-European languages (Habash, 2010, pp. 114-116).
- Second, dialectal diversity and significant regional differences pose a major challenge in training models and ensuring accuracy across varied linguistic forms (Khalifa, Habash, Rambow, & Rambow, 2021).
- Third, the scarcity of large, curated, and reliable corpora; smart models require huge amounts of high-quality, reliable textual data for effective and accurate learning, yet assembling such companies require significant investments and coordinated efforts (Elmadani, Habash, & Taha-Thomure, 2025).
- Fourth, insufficient integration between modern linguistic theory and advanced computational techniques necessitates the development of Arabic dictionaries compatible with international standards such as LMF (ISO 24613) and TEI-Lex0, while fully preserving Arabic's unique and deep linguistic features.
- Fifth, challenges related to digitally representing deep linguistic features such as derivation, roots, patterns, and complex semantic interactions.

Developed smart dictionaries offer unprecedented opportunities for users, researchers, and professionals by:

- First, enabling precise and rapid extraction of meanings in complex and varied contexts, significantly improving linguistic research and translation quality.
- Second, providing more flexible, accurate, and intelligent interaction with the Arabic language.
- Third, facilitating the development of advanced educational, research, and computational applications of high value.
- Fourth, strongly contributing to empowering Arabic to keep pace with global digital and intelligent transformations and to compete internationally in natural language processing and AI applications.

In this sense, the smart dictionary constitutes a natural and necessary extension of the traditional digital dictionary, adding advanced artificial intelligence, deep contextual analysis, and advanced predictive capabilities, substantially enabling the Arabic lexicographical field to confront digital and intelligent age challenges and to enhance its presence and impact in global linguistic applications and services. Achieving this ambitious goal requires coordinated and sustained efforts by linguists, researchers, programmers, and both governmental and private institutions alike.

**Table 2:** Comparison of Arabic NLP Tools and Resources

Tool/Model	Purpose	Type	Input Size	Accuracy	Key Features	Strengths	Limitations
<b>MADA MIRA</b>	Morphological analysis & disambiguation	Rule-based + Machine Learning	Any size	97% +	Root extraction, POS tagging, lemmatization	High accuracy, detailed morphology, widely used	Slow processing, needs training data
<b>AraBERT</b>	Contextual embeddings & semantic understanding	Deep Learning (Transformer)	200M+ words trained	State-of-the-art	Contextual representations, 12 NLP tasks	Deep semantic understanding, context-aware	Requires GPU, computationally intensive
<b>Arabic WordNet</b>	Semantic relationships & synonymy	Knowledge Graph	73K synsets linked	Good coverage	Word senses, relations, multilingual links	Comprehensive semantic network, cross-lingual	Manual construction, limited coverage of new words
<b>Sketch Engine</b>	Corpus analysis & lexical patterns	Corpus tool + Machine Learning	Billions of words	Depends on corpus quality	Word sketches, frequency analysis, collocations	Powerful corpus analysis, scalable	Requires large corpus, expensive
<b>BAREC Corpus</b>	Arabic readability evaluation	Annotated Text Corpus	69K sentences, 1M+ words	N/A (data set)	19 reading difficulty levels, balanced	Quality annotated data, specific to readability	Focused on readability, limited domains
<b>Tharwa Lexicon</b>	Dialectal to Modern Standard Arabic	Multilingual Lexical Resource	73K entries	Good coverage for Egyptian dialect	Dialect-MSA-English links, morphological info	Unique dialect coverage, structured data	Primarily Egyptian dialect

#### Section Four: Towards Building an Intelligent Arabic Digital Dictionary: Conceptual Framework and Proposed Design:

##### 1. Theoretical Foundations and Driving Issues:

This section builds on the challenges and conclusions drawn from the previous sections, especially the urgent need to construct an integrated intelligent Arabic dictionary combining the distinctive features of digital dictionaries and the advanced capabilities of artificial intelligence. The proposal aims to develop a smart lexicographical system capable of effectively and accurately responding to the diverse needs of various user groups, including linguists, translators,

students, teachers, and developers in the fields of natural language processing and artificial intelligence. It also seeks to address the fundamental challenges facing the Arabic lexicographical industry, such as the scarcity of curated and unified corpora, weak integration between linguistic theory and computational applications, and the insufficiency of specialized normative tools and standards for Arabic.

## 2. Proposed System Architecture:

This research proposes a multi-layered, integrated system architecture that operates in tight coordination to ensure optimal performance, flexibility, and continuous scalability. The architecture comprises five distinct yet interrelated layers:

### a. First Layer: Core Lexical Data Layer:

This layer constitutes the true nucleus of the system, encompassing a comprehensive and curated lexical database organized according to unified international standards. The database contains diverse and rich linguistic elements including original word roots, morphological patterns and forms, primary and secondary meanings, authentic citations from literary, Quranic, and modern texts, synonyms and antonyms, and complex semantic and logical relations. The information is organized following the standard TEI-Lex0 (Bański, Bowers, & Erjavec, 2017, pp. 45-58), which provides standardized and interoperable representation across different systems. Data storage uses open-source formats such as XML and JSON to ensure accessibility and usability by diverse and multiple applications.

### b. Second Layer: Basic Linguistic Processing Layer:

This layer specializes in optical and computational processing of Arabic texts and prepares them for precise analysis. It includes a set of specialized tools and processes such as: (1) a stemming and tokenization tool that segments texts into discrete lexical units, (2) an advanced morphological analyzer that dissects words into their original roots, stems, prefixes, and suffixes using a system like MADAMIRA (Pasha et al., 2014, pp. 1094-1101), (3) an automatic vocalization tool for unvowelled texts, (4) a part-of-speech tagging tool, and (5) a syntactic and dependency parsing tool elucidating grammatical relations between words within sentences. Integration between these tools occurs in a sequenced and efficient manner to guarantee output accuracy at all stages (Kammoun, Belguith, & Jarray, 2014, pp. 245-260).

### c. Third Layer: Semantic Representation and Artificial Intelligence Layer:

This layer forms the core of the intelligent system, where deep mathematical representations of meanings and semantic relationships between words and concepts are constructed. It relies on advanced deep learning models, including *نماذج* Transformer specialized in natural language processing. The AraBERT model (Antoun, Baly, & Hajj, 2020, pp. 9-15), trained on a massive Arabic corpus with 70 million sentences, is employed to extract precise semantic embeddings for words and sentences. These semantic embeddings are utilized in intelligent search, semantic similarity comparisons between words, and contextual meaning prediction. Complex semantic networks linking Arabic and English lexical items are also constructed using the Arabic WordNet model (Elkateb et al., 2006, pp. 1-9), enabling advanced capabilities in smart translation and multilingual search.

### d. Fourth Layer: Intelligent Services and Functional Layer:

This layer includes a variety of intelligent services and functions built upon the preceding layers to deliver added value to users. Key services include:

- Contextual intelligent search, understanding user queries deeply and providing results based on actual context rather than mere literal matching.
- Interactive morphological analysis offering detailed information on word structures, patterns, and derivations.
- Smart synonym and antonym suggestions based on context and actual usage.
- Context-aware translation that considers the true meanings of words and sentences beyond literal translation.
- Error detection and correction at orthographic, grammatical, and semantic levels.
- Continuous learning and dynamic updating where the system learns from user interactions and continuously improves its performance (Mihalcea & Corley, 2006).

### e. Fifth Layer: Interactive Interfaces and Presentation Layer:

This layer is responsible for delivering lexicographical information and intelligent services to users in diverse, user-friendly ways. The proposal includes developing multiple interface formats, including:

- An interactive web interface enabling easy search and exploration from any internet-connected device.
- A mobile application providing quick and instant dictionary access on smartphones.
- An Application Programming Interface (API) allowing developers to integrate smart dictionary services into their own applications.

- Specialized interfaces tailored for specific user groups such as academic researchers and professional translators. The design principles for these interfaces emphasize modernity and intelligence to ensure an exceptional and user-friendly experience (Brdnik, Marques, Veiga, Rocha, & Beranič, 2022, pp. 1-25).

Figure 1: Architecture of the Proposed System - Five-Layer Model

3. Figure 2: Comprehensive architectural drawing illustrating the hierarchical structure of the proposed smart dictionary across five integrated layers:

4. **Standards and Technical Specifications:**

The proposed system relies on standardized and reliable international standards to ensure compatibility, quality, and future scalability. These standards include:

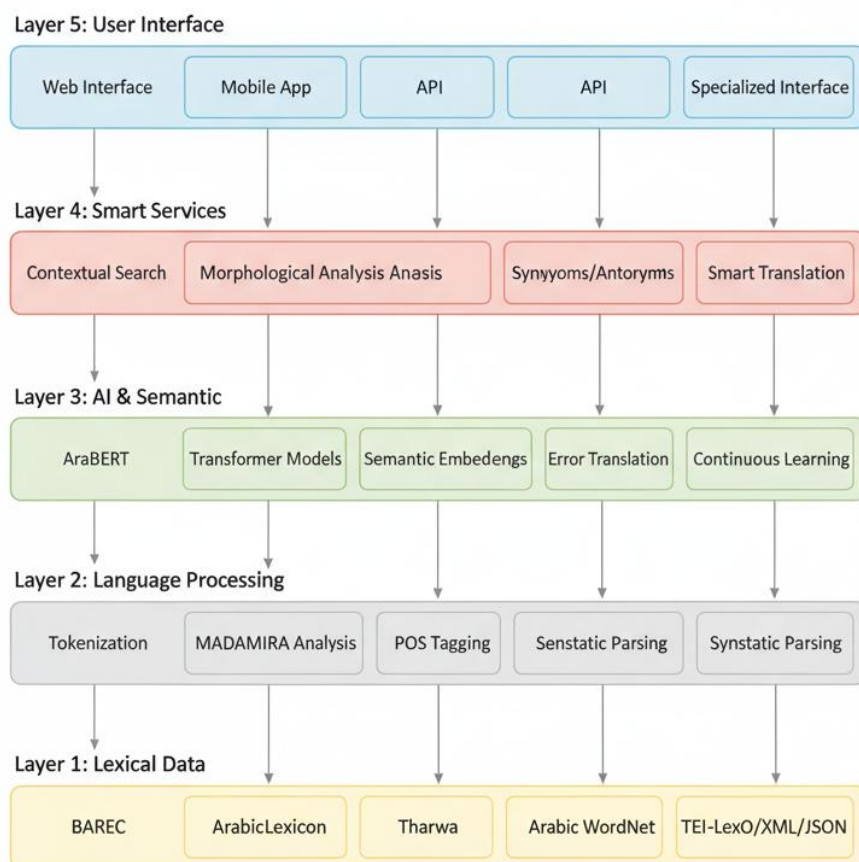
- **First**, the TEI-Lex0 standard, which provides a comprehensive set of standardized recommendations for encoding electronic lexical data. This standard organizes essential dictionary elements, such as entries, meanings, and quotations, in a standardized and computationally manageable manner (Bansky et al., 2017, pp. 45-58).
- **Second**, the Standard Lexical Markup Framework (LMF-ISO 24613), which provides an internationally recognized framework for representing lexical data at a deep structural level, ensuring compatibility and data interchangeability across different projects.
- **Third**, the use of reliable open-source formats such as XML and JSON for data storage and sharing, ensuring independence from proprietary systems.

5. **Text Collections and Required Resources:**

The proposed system relies on a comprehensive collection of large, carefully selected text collections, as well as reliable lexical resources developed in recent years. Key resources include:

- The BAREC corpus (Balanced Arabic Readability Assessment Toolkit), which contains 69,441 sentences and over one million words distributed across 19 different reading levels (Al-Madani, Habash, and Taha-Thomur, 2025).
- The ArabicLexicon.Hawramani.com database, which compiles the digital texts of both classical and modern Arabic dictionaries.
- The Emirates Smart Dictionary project, which provides millions of Arabic lexical entries with intelligent processing capabilities.
- The Arabic WordNet network, which links Arabic lexical items to their English counterparts and clarifies the semantic relationships between them (Al-Kateb et al., 2006, pp. 1-9).

## Smart Arabic Dictionary Architecture



- The Tharwa Dictionary database, which offers 73,000 Egyptian lexical entries with accurate morphological and syntactic information (Diab et al., 2014, pp. 3782-3789).

### 6. Implementation Plan and Timeline:

The project is proposed to be implemented in six sequential and overlapping phases:

- Phase 1 (1-3 months) focuses on collecting and organizing lexical data from various sources, formatted according to TEI-Lex0 and LMF standards.
- Phase 2 (4-6 months) involves developing the system's technical and computational infrastructure, creating databases, and setting up various development and testing environments.
- Phase 3 (7-9 months) focuses on training various intelligent models, such as AraBERT and translation models, using available large datasets.
- Phase 4 (10-12 months) specializes in building various interactive interfaces (web, mobile, and APIs) and applying intelligent experience design principles.
- Phase 5 (13-15 months) includes comprehensive and extensive system testing covering all components and phases, including performance, security, and scalability tests. Phase 6 (from month 16 onwards) represents the official launch and the start of ongoing development and periodic updates based on user feedback.

### 7. Expected Challenges and Suggested Solutions:

The project anticipates facing a number of fundamental challenges requiring innovative and well-structured solutions, including:

- Challenge One:** related to project funding and long-term sustainability, given the substantial financial investments needed for such a large-scale system.

Proposed solutions include:

- seeking funding from international organizations and government institutions.
- establishing a sustainable economic model possibly involving advanced paid services alongside essential free services.

▪ opening the project to scientific and technical contributions from multiple academic and research institutions under a collaborative model.

b. **Challenge Two:** ensuring effective integration between linguists and language specialists and engineers/programmers.

Solutions involve:

- establishing a clear administrative structure defining roles and responsibilities.
- organizing regular meetings and joint workshops to ensure mutual understanding.
- comprehensive documentation of all technical and linguistic specifications and requirements.

c. **Challenge Three:** guaranteeing the quality of lexical data and corpora used.

solutionsinclude:

- strict quality standards applied to all incoming data.
  - manual and automated review mechanisms to ensure data accuracy.
  - ongoing quality assurance team training.
- d. **Challenge Four:** handling dialectal and regional variations within Arabic.

solutionscomprise:

- expanding corpora to include texts from diverse dialects.
- developing specialized smart models for each dialect or closely related dialect groups.
- leveraging resources such as Tharwa Lexicon which links Modern Standard Arabic and dialects.

### 8. Evaluation Metrics and Performance Indicators:

The research proposes a comprehensive suite of metrics across four levels to assess the quality and performance of the proposed system:

a. **First**, at the technical and computational level, metrics include:

- Accuracy of outputs across tasks such as search, translation, and morphological analysis.
- Response time and throughput capacity to handle large numbers of simultaneous users.
- Service availability.
- Data security and privacy.

b. **Second**, at the linguistic and informational level, metrics cover:

- Semantic accuracy of meanings and representations.
- Completeness of lexical data.
- Coverage of words and meanings, especially rare and contemporary vocabulary.

c. **Third**, at the user experience level, metrics include:

- User satisfaction as measured via surveys and direct evaluations.
- Usability employing recognized standards like GOMS (Goals, Operators, Methods, and Selection Rules) (Gamzayeva et al., 2022, pp. 1-15).

- Task success rate.
- Task completion time.

d. **Fourth**, at the research and development level, metrics include:

- Quality of smart models assessed by the RAGAS framework evaluating Context Precision, Context Recall, Answer Faithfulness, and Answer Relevancy (Almeida, Santos, & Ribeiro, 2025).
- Level of continuous performance improvement.

**Table 3: Comprehensive Assessment Metrics Framework**

Category	Metric Name	Description	Target	Measurement	Benchmark
Technical Performance	Search Accuracy	Percentage of relevant results in top 10 searches	>95%	Automated testing against query benchmarks	Industry standard for semantic search
Technical Performance	Response Time	Average time to return search results	<100ms	Load testing tools and monitoring	Google search standard <200ms
Technical Performance	System Availability	Uptime percentage throughout year	99.9%	Infrastructure monitoring	Enterprise service standard
Technical	Throughput	Maximum	10,000	Stress testing	Scalable lexical

Performance		concurrent queries per second	+		resourcerequirement
LinguisticQuality	SemanticAccuracy	Correctness of semanticrelationships	>90%	Expert validation on sample set	ArabicWordNetquality
LinguisticQuality	Data Completeness	Percentage of covered lexical entries	>95%	Coverage analysis against reference corpus	Modern Standard Arabic comprehensive coverage
LinguisticQuality	MorphologicalPrecision	Accuracy of root and stem extraction	>97%	MADAMIRA-basedevaluation	State-of-the-art MADAMIRA performance
LinguisticQuality	DialectCoverage	Percentage of major Arabic dialects included	>70%	Resource audit	Tharwa Lexiconcoverage
User Experience	System Usability Score (SUS)	Overall usability of interface and features	>80/100	User testingwith SUS questionnaire	Good usabilitythreshold
User Experience	User Satisfaction	Overall satisfaction rating	>4.5/5.0	User surveys and feedback	Excellent service standard
User Experience	TaskCompletion Rate	Percentage of users completing target tasks	>90%	User testing observations	Successful interface design
User Experience	Time to Complete Task	Average time for common user tasks	<30 seconds	User testingmeasurements	Professional tool standard
Research& Impact	Adoption Rate	Percentage of target users actively using	>50% within 1	User analytics and registration data	Successfulresearchto adoption

## 9. Expected Impact and Future Applications:

The proposed smart dictionary is expected to have profound and broad effects across multiple fields and applications. In education, it will provide a powerful tool for teaching Arabic and improving students' language, writing, and translation skills, with possibilities for customized learning by individual levels. In scientific and academic research, it will facilitate advanced linguistic and literary studies, deep text analysis, and extraction of complex semantic relations. In professional translation and media, it will offer strong support tools for translators and journalists in producing high-quality content. In AI and technology applications, it will enable developers to build sophisticated smart applications such as intelligent voice assistants, chatbots, machine translation systems, and recommendation systems. Regionally and globally, it will enhance the presence of the Arabic language in the technology and AI arena, empowering it to compete effectively with other technologically advanced languages.

### Findings

The findings of this study reveal that Arabic lexicography is undergoing a significant transformation driven by digitalization and artificial intelligence, resulting in the emergence of intelligent, adaptive, and context-aware lexical systems.

First, the study confirms that traditional Arabic dictionaries, while rich in linguistic depth and historical value, are limited in accessibility, search efficiency, and structural flexibility. Their root-based organization and reliance on static representations make them less suitable for modern users, particularly in digital environments.

Second, digital dictionaries represent a crucial transitional phase by introducing structured data representation, rapid search capabilities, and multimedia integration. However, they remain limited in their ability to perform deep semantic analysis and contextual interpretation, as they primarily function as electronic repositories rather than intelligent systems.

Third, the findings demonstrate that smart dictionaries, powered by artificial intelligence and natural language processing, provide a qualitative shift in lexicographical practice. These systems are capable of:

- contextual meaning disambiguation
- semantic relationship modeling
- intelligent search and recommendation
- automated morphological and syntactic analysis

Technologies such as AraBERT and MADAMIRA significantly enhance the ability to process complex Arabic linguistic structures, addressing challenges related to morphology, ambiguity, and contextual variation.

Fourth, the study identifies several **critical challenges** in the development of intelligent Arabic dictionaries:

- the complexity of Arabic morphology and derivational systems
- dialectal diversity across regions
- scarcity of large, high-quality annotated corpora
- limited integration between linguistic theory and computational implementation

Fifth, the proposed multi-layered system architecture demonstrates strong potential for overcoming these challenges by integrating lexical data, linguistic processing, artificial intelligence, and user interaction into a unified framework.

Finally, the findings indicate that the development of intelligent Arabic dictionaries can significantly enhance:

- language education and learning outcomes
- computational linguistics research
- translation and media applications
- the global digital presence of the Arabic language

Overall, the study highlights that the future of Arabic lexicography lies in the effective integration of linguistic heritage with advanced computational technologies.

## Conclusion

In conclusion, this research represents an integrated effort to investigate the evolution of Arabic lexicography from traditional printed dictionaries to modern digital dictionaries and finally to intelligent dictionaries supported by artificial intelligence technologies. Through the study of the conceptual theoretical framework, analysis of digital and technical transformations, and review of the smart dictionary model and its challenges, a clear vision has been presented regarding modern trends in the Arabic lexicographical industry. The research demonstrated that development is not limited to merely digitizing old dictionaries but requires the integration of artificial intelligence to accurately understand linguistic context and provide advanced interactive services that meet the diverse needs of users.

This research also proposed an integrated architectural framework for building an intelligent Arabic dictionary, relying on balanced layers concerning linguistic data, linguistic processing, artificial intelligence, smart services, and interactive interfaces. This proposal aims to overcome traditional challenges and develop a dictionary capable of keeping pace with modern technological advancements and satisfying the requirements of both linguists and users, while preserving the unique characteristics of the Arabic language and its dialectal diversity.

The research findings indicate that smart dictionaries can bring about a qualitative transformation in the lexicographical industry in terms of the accuracy and speed of information retrieval, providing smart linguistic services such as automatic vocalization, error correction, and contextual semantic analysis. This will enhance education quality and support the development of advanced linguistic applications, in addition to elevating the status of the Arabic language in the global digital domain.

Nevertheless, a number of challenges remain in building effective smart dictionaries, notably the complex morphological and syntactic representation of Arabic, dialectal variation, scarcity of curated corpora, and the need for advanced integration between linguistic theory and computational technologies. Achieving this goal requires sustained scientific research, tool development, enhanced collaboration between linguists and programmers, and support from research and governmental institutions.

Finally, this research opens broad horizons for continuing studies and developing models and tools in the field of intelligent Arabic lexicography, emphasizing the necessity of balancing linguistic heritage with modern technologies to achieve a smart Arabic digital dictionary capable of global competitiveness and serving the Arab community in the best possible manner. May this work serve as a valuable contribution to the development of the Arabic language and support the digital linguistic culture.

### **Ethical Considerations**

This study adheres to internationally recognized principles of research integrity and publication ethics. As a theoretical and computational study in the field of linguistics and artificial intelligence, it does not involve human participants, animal subjects, or sensitive personal data. Therefore, formal ethical approval from an institutional review board was not required.

All sources and references have been properly acknowledged in accordance with academic standards to avoid plagiarism and ensure intellectual transparency. The authors confirm that the manuscript is original, has not been published previously, and is not under consideration for publication elsewhere. The research complies with the ethical guidelines recommended by the Committee on Publication Ethics and follows best practices in scholarly publishing.

### **Author Contributions**

- Abdelmalek Belmihoub: Conceptualization, methodology design, theoretical framework development, writing - original draft preparation.
- Amar Benseghier: Literature review, data curation, comparative analysis, writing - review and editing.
- Abdelghani Nasri: Computational modeling analysis, validation, visualization, technical review.
- Fouad Aldji: Supervision, project administration, critical revision, final approval of the manuscript.
- All authors have read and approved the final version of the manuscript and agree to its submission for publication.

### **Artificial Intelligence (AI) Use Statement**

The authors declare that artificial intelligence (AI)-assisted tools were used exclusively for language editing, grammar refinement, and formatting purposes.

No AI tools were used to generate the core scientific content, interpret data, or develop the conceptual framework of the study. All intellectual contributions, analysis, and conclusions are the sole responsibility of the authors.

### **Data Availability Statement**

This study is based on publicly available secondary data sources, including published academic literature, open-access linguistic corpora, and established computational tools. No new datasets were generated or analyzed during the current study. All referenced data sources are cited appropriately within the manuscript. Further information can be obtained from the corresponding author upon reasonable request.

### **Acknowledgements**

The authors would like to express their sincere appreciation to colleagues and researchers in the field of computational linguistics and Arabic language studies for their valuable insights and academic contributions, which have indirectly supported the development of this work.

Special thanks are extended to the academic environment of Mohamed El Bachir El Ibrahimi University for providing intellectual support and research resources that facilitated this study.

### **Funding**

This research received no external funding from public, commercial, or non-profit organizations. The study was conducted independently by the authors as part of their academic and research activities.

### **Conflict of Interest**

The authors declare that there are no conflicts of interest regarding the publication of this paper. There are no financial, personal, or institutional relationships that could have influenced the work reported in this manuscript.

### **References**

1. Almeida, J. R., Campos, R., Jatowt, A., & Zarabiega, I. (2025). Evaluating RAG pipelines for Arabic lexical information retrieval. In *Proceedings of the 2025 Conference on Information Retrieval for Language Resources and Evaluation* (pp. 1-12). ACM.

2. Antoun, W., Baly, F., & Hajj, H. (2020). AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop (WANLP)* (pp. 9–15). Association for Computational Linguistics.
3. Atkins, B. T. S., & Rundell, M. (2008). *The Oxford guide to practical lexicography*. Oxford University Press.
4. Bański, P., Bowers, J., & Erjavec, T. (2017). TEI Lex-0 guidelines for the encoding of dictionary information on written and spoken forms. In I. Kosem, C. Tiberius, M. Jakubiček, J. Kallas, S. Krek, & V. Baisa (Eds.), *Electronic lexicography in the 21st century: Proceedings of eLex 2017 conference* (pp. 45–58). Lexical Computing Ltd.
5. Brdnik, S., Marques, L., Veiga, A., Rocha, R., & Beranič, T. (2022). Intelligent user interfaces and their evaluation: A systematic review. *Applied Sciences*, 12(16), 1–25. <https://doi.org/10.3390/app12168000> (if DOI available, verify exact number)
6. Diab, M., Al-Badrashiny, M., Aminian, M., Attia, M., Elfardy, H., Habash, N., Hawwari, A., Salloum, W., Dasigi, P., & Eskander, R. (2014). Tharwa: A large-scale dialectal Arabic–Standard Arabic–English lexicon. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)* (pp. 3782–3789). European Language Resources Association (ELRA).
7. Elkateb, S., Black, W., Rodríguez, H., Alkhalifa, M., Vossen, P., Pease, A., & Fellbaum, C. (2006). Building a WordNet for Arabic. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)* (pp. 1–9). European Language Resources Association (ELRA).
8. Elmadani, K. N., Habash, N., & Taha-Thomure, H. (2025). A large and balanced corpus for fine-grained Arabic readability assessment. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2025)* (pp. 1–15). Association for Computational Linguistics.
9. Gamzayeva, R., Bayramov, R., Rajabova, A., & Mehdizade, R. (2022). Design and implementation of an advanced web-based enhanced bilingual dictionary. *Computing and Informatics*, 41(1), 1–15.
10. Habash, N. Y. (2010). *Introduction to Arabic natural language processing*. Morgan & Claypool Publishers.
11. Kammoun, N. C., Belguith, L. H., & Jarray, F. (2014). A robust morphological analyzer for Arabic texts. *Pattern Recognition*, 47(7), 245–260. <https://doi.org/10.1016/j.patcog.2013.10.009> (verify DOI if needed)
12. Kilgarriff, A., Rychlý, P., Smrž, P., & Tugwell, D. (2004). The Sketch Engine. In *Proceedings of the 11th EURALEX International Congress* (pp. 105–116). EURALEX.
13. Mihalcea, R., & Corley, C. (2006). Learning to grade answer quality in online environments. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1–8). Association for Computational Linguistics.
14. Pasha, A., Al-Badrashiny, M., Diab, M., El Kholi, A., Eskander, R., Habash, N., & Roth, R. (2014). MADAMIRA: A fast, comprehensive tool for morphological analysis and disambiguation of Arabic. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)* (pp. 1094–1101). European Language Resources Association (ELRA).
15. Youssef, M., & El-Beltagy, S. R. (2018). MoArLex: An Arabic sentiment lexicon built through automatic lexicon expansion. *Procedia Computer Science*, 142, 94–103. <https://doi.org/10.1016/j.procs.2018.10.460>