IInTÆ
Universidad Nacional de Catamarca - Facultad de Humanidades
Instituto de Investigación en Teorías del Arte y Estética
Revista Rigel – ISSN 2525-1945.

Linguistic Corpora and Heritage Text Processing: Challenges and Prospects in the Era of Artificial Intelligence

Abdelkader Lakehal

University of Mohamed Boudiaf - M'Sila, Algeria
Email: abdelkader.lakehal@univ-msila.dz

Samir Djaballah

University of Mohamed Boudiaf - M'Sila, Algeria
Email: samir.djaballah@univ-msila.dz

Received: 11/01/2026 | Accepted: 20/03/2026 | Published: 11/08/2026

Abstract

This study examines the role of linguistic corpora in the automated processing of classical Arabic heritage texts within the context of recent developments in computational linguistics, artificial intelligence, and large language models. It highlights the linguistic corpus as a foundational digital infrastructure that enables the representation, analysis, and preservation of Arabic textual heritage. Adopting a descriptive-analytical and comparative framework, the study discusses the major challenges facing heritage text processing, including optical and handwritten text recognition, the absence of diacritics, semantic ambiguity, historical terminological shifts, and the multiplicity of manuscript copies. It also explores the contribution of deep learning models, Transformer architectures, automatic diacritization systems, and Arabic digital platforms such as Falak and Siwar in enhancing the accuracy of heritage digitization and linguistic analysis. The paper concludes that building verified, balanced, and well-annotated heritage corpora is a strategic necessity for preserving Arabic linguistic identity, supporting digital sovereignty, and enabling future artificial intelligence applications to engage with Arabic heritage in an authentic and scientifically reliable manner.

Keywords: linguistic corpora; Arabic heritage texts; computational linguistics; artificial intelligence; large language models; deep learning; automatic diacritization; digital humanities.

1. Introduction

Over the last decade, the field of computational linguistics has experienced a profound structural paradigm shift. This scholarly domain has transitioned from a reliance on constrained formal rules toward the expansive frameworks of statistical modeling and deep neural networks.

Central to this evolution is the "Linguistic Corpus," which occupies a foundational role. It serves not merely as a passive textual repository but as a dynamic substrate that supplies artificial intelligence algorithms with the high-fidelity training datasets required to encapsulate the diversity and nuances of natural language.

The automated processing of classical Arabic heritage texts possesses a dual-dimensional significance. Firstly, it constitutes a formidable technical challenge, stemming from the intricate structural and diachronic complexities inherent in the Arabic linguistic system. Secondly, it represents a critical civilizational mandate; facilitating access to this intellectual wealth for global researchers and future generations is an ethical obligation toward the preservation of collective human heritage.

Given the recent proliferation of Large Language Models (LLMs), there is an exigent need to re-examine the methodologies underlying the construction of heritage corpora. Such a re-evaluation is essential to guarantee the precision of information retrieval, the rigor of linguistic analysis, and the authenticity of representation within digital frameworks.

1.1. Conceptual Framework and Significance of Linguistic Corpora in Heritage Studies

The Linguistic Corpus is defined within computational linguistics literature as an automated bank comprising texts selected according to rigorous criteria of representativeness. These are classified by specific chronological eras or epistemic domains and stored in a digital format that facilitates high-efficiency computational processing.

The pioneering Algerian linguist Abderrahmane Hadj-Salah provided a foundational contribution to establishing this concept within Arabic studies when he proposed the "Arabic Linguistic Corpus" project at the 1986 Amman Conference. He asserted that the corpus is not a passive repository but a dynamic infrastructure aimed at the codification and automated processing of the Arabic tongue (Hadj-Salah, 2007). In the same vein, Mohamed El-Hannach devoted particular attention to the applied dimension of these corpora within the context of Arabic language engineering and programming, noting the close correlation between the quality of heritage corpora and the accuracy of the computational applications built upon them (El-Hannach, 2015). Furthermore, Benzina and Dreem traced the applied trajectory of this project and its developmental stages, demonstrating that commitment to the standard of representativeness and textual balance represents the backbone of any heritage corpus aspiring to scientific reliability (Benzina & Dreem, 2019, p. 22).

The primary significance of linguistic corpora in the field of heritage studies is manifested in their ability to transition philology from traditional impressionistic methodologies to quantitative and qualitative statistical approaches. They enable researchers to study the diachronic semantic evolution of vocabulary and monitor the stylistic criteria governing classical Arabic writing with a precision that manual examination cannot achieve (Al-Shahat, 2022, p. 222). Moreover, verified

and tagged corpora constitute the essential raw material for training Large Language Models (LLMs) specialized in Arabic heritage, ensuring these models maintain a contextual presence in juridical, philosophical, and rhetorical fields that are absent in general-purpose models.

Based on the aforementioned, the corpus becomes a semantic bridge connecting the static paper manuscript with evolving Generative Artificial Intelligence applications, opening promising horizons for the restoration and analysis of linguistic memory within its authentic contexts. The function of the corpus is only completed when its methodological foundations are governed alongside its instrumental techniques. Without a heritage corpus that is rigorously vocalized and scientifically verified, Artificial Intelligence lacks the tools to differentiate between the semantic loads of words across successive eras, as well as the capacity to comprehend the deep referential contexts in which those texts were formulated.

2. Materials and Methods

To systematically investigate the intersection between classical Arabic heritage texts and modern computational architectures, this research employs a multi-tiered descriptive-analytical and comparative research design. The methodological framework is organized into three interrelated stages:

• Corpus Design Criteria Formulation: Establishes rigorous criteria for heritage corpus construction based on diachronic representativeness, structural balance, morphological vocalization, and philological verification.

• Processing Pipeline Modeling: Analyzes the sequential stages of the automated heritage text processing pipeline, spanning Optical Character Recognition (OCR) and Historical Handwritten Text Recognition (HTR), neural automated diacritization, lexical disambiguation, and Named Entity Recognition (NER).

• Technological and Institutional Benchmarking: Conducts an evaluative comparative assessment of recent deep learning architectures (e.g., Transformer encoders, AraModernBERT, HATFormer) and institutional platform initiatives (such as KSAA's Falak and Siwar platforms, and the Zinki project), evaluating their performance against error metrics and contextual capacities.

3. Results

The investigation yields crucial empirical and analytical results regarding the structural obstacles facing Arabic heritage processing and the corresponding technological solutions enabled by state-of-the-art artificial intelligence models.

3.1. Challenges in the Processing of Heritage Texts

The automated processing of Arabic heritage texts represents one of the most intricate domains in contemporary computational linguistics. The complexities inherent in the physical medium of the text intersect with deep-seated structural linguistic obstacles, rendering any endeavor toward automated digitization contingent upon addressing this multi-dimensional and integrated complexity.

• Optical and Handwritten Text Recognition (OCR/HTR): Ancient manuscripts frequently suffer from paper degradation, overlapping lines, and an absence of regular inter-word spacing. These issues are compounded by the diversity of calligraphic styles—ranging from Kufic to Naskh

and Thuluth—which often lack a consistent baseline. Consequently, traditional models exhibit significant variance in the accuracy of textual content extraction (ICESCO Journal of the Arabic Language, 2024, p. 272). Despite advancements in machine learning algorithms, a persistent gap remains between the precision of human paleographers and automated software, necessitating the development of computer vision models that are more flexible in navigating the curvatures and decorative intertwinements of Arabic script (Qatar National Library, 2024, p. 5).

• Absence of Diacritics and Semantic Ambiguity: Automatic diacritization constitutes a linguistic challenge with profound roots. The absence of short vowels (harakat) in the majority of heritage texts obstructs precise contextual understanding, as a single orthographic representation may carry entirely contradictory meanings depending on its grammatical position and stylistic context. In his proposal for the Corpus project, Abderrahmane Hadj-Salah noted that diacritization is not merely a technical utility but an epistemic problem related to understanding the deep structure of Arabic and its derivational system (Hadj-Salah, 2007). This difficulty is exacerbated by a scarcity of digitized, diacritized, and verified linguistic resources—a structural barrier that continues to impede the training of high-precision algorithms on heritage data.

• Semantic and Terminological Evolution: Lexical and semantic challenges pertaining to the phenomenon of historical semantic shift are among the most formidable. A word within a heritage text often bears a cognitive load that differs radically from its established meaning in modern dictionaries, or even from its usage in heritage texts of a different epoch. Constructing comprehensive digital historical dictionaries requires the aggregation of massive quantities of linguistically tagged data—precisely categorized by period and domain—followed by the deployment of semantic tagging algorithms specifically designed to accommodate the unique characteristics of Arabic, its expansive metaphors, and its derivational richness (Al-Shahat, 2022, p. 227).

• Proliferation of Manuscript Copies and Textual Variances: The phenomenon of manuscript proliferation—wherein multiple copies of a single text exhibit variances through interpolation, omission, or orthographic corruption (tashif)—further compounds this complexity. This necessitates the construction of a flexible, standardized corpus capable of integrating these discrepancies into a multi-layered data architecture without compromising the integrity of the original semantic essence (Benzina & Dreem, 2019, p. 25).

3.2. Artificial Intelligence and Deep Learning: A Paradigm Shift in Heritage Processing

The implementation of Transformer architectures and Attention Mechanisms has precipitated a qualitative leap in the comprehension of the protracted and intricate contexts inherent in classical texts. This represents a significant advancement over Recurrent Neural Networks (RNNs), which predominated in earlier eras but were constrained by an inherent inability to encompass long-range dependencies within extensive textual segments (Chan et al., 2024, p. 1). This transition constitutes more than a mere technical upgrade; it redefines the parameters of possibility regarding the stylistic analysis of ancient authors and the extraction of their multifaceted semantic networks across thousands of pages.

• Augmented Contextual Capacity (AraModernBERT): This revitalized capacity is exemplified by the AraModernBERT model, which leverages the ModernBERT architecture and utilizes Transtokenized Initialization to achieve a context window of up to 8,192 tokens. In the practical application of heritage studies, this enables the model to process entire chapters rather than fragmented sentences, thereby comprehending comprehensive arguments rather than isolated indices. This advancement is particularly consequential for processing voluminous juridical and exegetical works characterized by convoluted stylistic structures (Elshehy et al., 2026, p. 2).

• The Sukkoun System: The Sukkoun system serves as an advanced instrumental foundation for resolving lexical ambiguity and restoring diacritics in heritage texts. This system operates within a broader mission framed by the King Salman Global Academy for the Arabic Language (KSAA), wherein Large Language Models (LLMs) are deployed to address the complexities of automated diacritization and semantic retrieval, achieving a level of precision that progressively approximates the proficiency of expert human philologists (KSAA, 2026b).

• Historical Handwritten Text Recognition (HATFormer): In the domain of Historical Handwritten Text Recognition (HTR), the HATFormer system represents a mature embodiment of Transformer architectures applied to the complexities of manuscript Arabic script, such as overlapping ligatures and the allomorphic variation of characters based on their position. This system attained a Character Error Rate (CER) of 8.6% on the largest public dataset of historical Arabic script—a 51% improvement over preceding models. This metric reflects a genuine methodological maturity in treating the Arabic manuscript as a complex document rather than a flat textual image (Chan et al., 2024, p. 2).

• Automation of Linguistic Analysis: Deep learning techniques have facilitated what may be termed the "automation of linguistic analysis." Artificial Intelligence is no longer merely a tool for initial digitization; it has become an active collaborator in rigorous philological investigation. It extracts named entities, classifies poetic and prosaic styles, and distinguishes between overlapping discursive layers within a single text. Recent studies in Arabic Natural Language Processing (ANLP) suggest that integrating these technologies with human linguistic oversight represents the optimal model for engaging with heritage, particularly in instances where meaning is contingent upon a cultural and civilizational context that the machine cannot perceive in isolation (Alayba, 2025, p. 497).

3.3. Pioneering Initiatives and Arabic Digital Platforms

• King Abdullah Initiative and KSAA Platforms: The contemporary landscape is presided over by the platforms of the King Salman Global Academy for Arabic Language (KSAA), which embody a comprehensive paradigm for deploying technology in service of linguistic heritage. Notably, the Falak linguistic corpus platform serves as a prolific repository for the study of Arabic linguistic phenomena across both normative and usage-based dimensions. It provides sophisticated analytical instruments, such as concordancers, frequency lists, and collocations, rendering it a foundational pillar for researchers in both Arabic processing and computational lexicography (KSAA, 2024a). Furthermore, Falak supplies LLMs with high-quality Arabic data

exceeding one billion words, markedly enhancing contextual comprehension and generation (King Abdullah Initiative, 2025, p. 15).

• The Siwar Platform: The Siwar lexicographical platform undertakes the task of digitizing the entire lexicographical industry. It facilitates queries across dozens of published dictionaries and provides Application Programming Interfaces (APIs) that enable developers to seamlessly integrate heritage and contemporary lexicons into external applications. This transforms lexicographical labor from an isolated effort into an open, collaborative ecosystem where humans and machines coexist (KSAA, 2024b). These platforms are augmented by the "Arabic Intelligence Center," which coordinates specialized technical projects for the verification and analysis of heritage (KSAA, 2026b).

• The Zinki Project: The Zinki project represents a highly influential research contribution to remapping what historians term the Islamic "Republic of Letters." Its methodology achieves advanced levels of Named Entity Recognition (NER)—identifying figures, locations, genealogies, and institutions. These entities are then utilized to construct interactive historical relationship maps that manifest networks of narration and epistemic transmission across centuries (Al-Zaidi, 2025, p. 229; Zinki Research Group, 2024, p. 8).

Table 1

Synthesis of Heritage Text Processing Challenges and AI-Driven Technological Solutions

Heritage Processing Challenge	Linguistic / Technical Impact	AI Technological Resolution
Physical script degradation & calligraphic variability	High Character Error Rate in OCR/HTR; loss of baseline alignment	HATFormer Vision-Transformer architecture (CER reduced to 8.6%)
Absence of short vowels (harakat)	Pervasive syntactic ambiguity and semantic misinterpretation	Sukkoun Neural Diacritization System & LLM contextual disambiguation
Diachronic semantic and terminological shifts	Cognitive mismatch between classical semantics and modern lexicons	Historical corpus stratification & Siwar open lexicographical APIs
Manuscript multiplicity & textual variances (tashif)	Fragmented textual transmission and risk of corrupted readings	Multi-layered flexible data schemas & automated variant collators
Long-range rhetorical and syntactic dependencies	Truncated context windows leading to fragmented discourse parsing	AraModernBERT with Transtokenized Initialization (8,192 token window)
Named entity extraction across historical chronicles	Inability to map dense biographical and genealogical networks	Zinki Project NER algorithms & historical network graphing

Note. Synthesized from study findings, empirical model benchmarks, and platform documentation.

4. Discussion

4.1. The Siwar Platform and Digital Sovereignty

The scholarly impact of the Siwar platform is further deepened as it empowers researchers—via its Open APIs—to construct specialized digital heritage lexicons for precise epistemic fields, such as the terminology of Kalam philosophy, Sufic discourse, and the biographical dictionaries of Hanbali jurisprudence. In this context, "Siwar" evolves beyond an advisory tool into an independent platform for knowledge production.

This solidifies what critical literature describes as "Digital Sovereignty" over the Arabic language—the reclamation of the right for Arabic-speaking communities to define their language, preserve its concepts, and frame its cognitive instruments independently of dominant Western technological frameworks (KSAA, 2026a, p. 10). A heritage lexicon built upon "Siwar" is not merely a digital dictionary but an act of identity reclamation involving the researcher, the developer, and the machine in unison.

These integrated initiatives reflect a steadfast strategic vision aimed at transforming Arabic heritage from static, neglected texts into intelligent data that fuels global knowledge engines. In an era where the power of nations is partially measured by their capacity to train and nourish Large Language Models with authentic national data, these initiatives become indispensable pillars of Digital Sovereignty. They ensure that future Artificial Intelligence applications are not constructed upon corpora that fail to reflect the spirit and depth of this heritage, but rather upon authentic Arabic data derived from the nation's civilizational wealth.

4.2. Future Horizons and Anticipated Transformations: The Automated Digital Philologist

The future of linguistic corpora and the automated processing of heritage texts promises the emergence of what may be termed the "Automated Digital Philologist." This conceptual AI system transcends the mere physical digitization of text; it performs critical comparative analyses across multiple manuscript copies and proposes scholarly emendations based on an expansive inferential repertoire of authorial styles and scribal tendencies from specific historical epochs. Al-Zaidi (2025) explored these prospects extensively, noting that the forthcoming phase will empower automated systems to "reconstruct" damaged passages and textual lacunae with a degree of precision that rivals specialized human researchers (p. 232).

In the domain of immersive technologies, there is a burgeoning trend toward integrating Augmented Reality (AR) and Virtual Reality (VR) with digital linguistic corpora. This would enable researchers to navigate a heritage document within a virtual environment, supplemented by an instantaneous digital analytical layer providing diacritization, semantic commentary, and the historical context for each term simultaneously.

Furthermore, Generative Artificial Intelligence (GenAI) is poised to play a pivotal role in the "digital restoration" of deteriorated manuscripts. Computer vision algorithms will possess the capability to reconstruct missing pages and bridge textual gaps by utilizing deep modeling of the author's idiosyncratic style, guided by the surrounding textual environment.

On a structural level, linguistic corpora will gravitate toward achieving Semantic Interoperability through the adoption of Linked Data technologies and international authority standards, such as the Virtual International Authority File (VIAF). This evolution will facilitate the global integration of heritage information across research centers, transforming the Arabic linguistic legacy into a central node within the global network of human knowledge, rather than a closed repository restricted by geographical or institutional boundaries (Kamel Shahin, 2024).

5. Conclusion

The construction of linguistic corpora for heritage texts does not constitute an academic or technical luxury; rather, it represents a strategic imperative for safeguarding linguistic and

civilizational identity within a rapidly accelerating digital era. Successive advancements in the field of Artificial Intelligence have demonstrated that traditional challenges—ranging from the absence of diacritics and manuscript proliferation to calligraphic complexity—can be surmounted, even if not entirely resolved, through the deployment of specialized deep learning models nourished by precise and verified heritage corpora.

However, these endeavors will only achieve their full potential if established upon a robust methodological foundation, sustained by the collaborative efforts of linguistic and heritage experts working in tandem with engineers and programmers. The synergy of these competencies, supported by ambitious institutional projects such as the Falak and Siwar platforms and the Arabic Intelligence Center, is the essential mechanism for transitioning our heritage from the confines of library shelves to the vanguard of global knowledge engines. Thus, the linguistic legacy of this nation shall become an active contributor to the civilization of the digital future.

References

Al-Omran, F. M. (2026). Pre-Islamic poetry between heritage commentaries and automated analysis: A linguistic study using deep learning techniques. Bank Journal, 6(1), 1–20.

Al-Shahat, A. (2022). Linguistic corpora and their role in modern Arabic lexicography (p. 222). Dar Al-Wafaa.

Al-Zaidi, W. K. (2025). Uses of artificial intelligence in deciphering ancient scripts and restoring damaged historical manuscripts. Proceedings of the Sixth International Scientific Conference on the History of Science among Arabs, 220–235.

Alayba, A. M. (2025). Arabic natural language processing (NLP): A comprehensive review of challenges, techniques, and emerging trends. Computers, 14(11), 497. https://doi.org/10.3390/computers14110497

Benzina, S., & Dreem, N. (2019). The Arabic linguistic corpus project: Reality and prospects. Journal of Linguistics, 4(1), 15–30.

Chan, A., Mijar, A., Saeed, M., Wong, C.-W., & Khater, A. (2024). HATFormer: Historic handwritten Arabic text recognition with transformers (arXiv:2410.02179). arXiv. https://doi.org/10.48550/arXiv.2410.02179

Conference on Arabic Language Computing and Linguistic Data Enrichment. (2024). Proceedings of the Academy's First International Conference. King Salman Global Academy for Arabic Language.

El-Hannach, M. (2015). The use of the Arabic language in information technology. Journal of Linguistic Communication, Fez.

Elshehy, O., Nacar, O., Djamai, A., Ragab, M., Al Jallad, K., et al. (2026). AraModernBERT: Transtokenized initialization and long-context encoder modeling for Arabic (arXiv:2603.09982). arXiv. https://doi.org/10.48550/arXiv.2603.09982

Hadj-Salah, A. (2007). The Arabic linguistic corpus project and its scientific and applied dimensions. Journal of the Arabic Language, 3(1), 9–30.

ICESCO Journal of the Arabic Language. (2024). Prospects of artificial intelligence in the service of the Arabic language: Manuscripts as a model. Islamic World Educational, Scientific and Cultural Organization, (12), 272.

Kamel Shahin, S. (2024). The Arabic guide for integrating community memory institutions: Libraries, archives, museums, and digital transformation requirements. Sultan Qaboos University.

King Abdullah Initiative for Arabic Content. (2025). Report on artificial intelligence and Arabic linguistic data: Toward one billion words for training large models (p. 15). King Abdullah Initiative.

King Salman Global Academy for Arabic Language. (2024a). Falak platform for linguistic corpora. https://falak.ksaa.gov.sa

King Salman Global Academy for Arabic Language. (2024b). Siwar platform for linguistic lexicons. https://siwar.ksaa.gov.sa

King Salman Global Academy for Arabic Language. (2026a). Digital sovereignty and heritage lexicon APIs: The Siwar platform as a model (p. 10). Arabic Intelligence Center.

King Salman Global Academy for Arabic Language. (2026b). Report of the KSAA-2026 joint mission for automated diacritization. Arabic Intelligence Center. https://arai.ksaa.gov.sa/arsharedTask2026/

Qatar National Library. (2024). Report on the digitization of Arabic manuscripts and challenges of automated recognition (p. 5). Qatar National Library. https://www.qnl.qa/ar/about/annual-report

Zinki Research Group. (2024). Mapping the Islamic Republic of Letters: Named entity extraction and relational networks in classical Arabic texts (Working Paper Series, p. 8). Zinki Research Group.