

A Corpus-based Model of Methodological Transformation in Arabic Studies within the Digital Humanities Era

Harmanto Raharjo^{1*}, Muhammad Arwani², Ahmad Syahid³, Mahrus As'ad⁴,
Ateng Ruhendi⁵, Aulia Rahmi⁶, Henny Septia Utami⁷

^{1,2,3,4,5,6}Islamic University of Sunan Gunung Djati Bandung, Jl. A.H. Nasution No. 105,
Cipadung, Cibiru, Bandung, West Java, Indonesia

⁷Queen's University Belfast, United Kingdom

*Correspondence Email: harmantoraharjo07@gmail.com

Doi: [10.24252/jad.v26i2a3](https://doi.org/10.24252/jad.v26i2a3)

(Submitted: 07/07/2026, Accepted: 23/09/2026, Available Online: 29/09/2026)

Abstract

The rise of the *Digital Humanities* has reshaped the methodological landscape of the humanities, including Arabic studies, through the use of digital corpora as a basis for linguistic analysis. Yet studies that specifically link corpus linguistics to the methodological transformation of Arabic studies and that offer an explicit conceptual framework remain limited. This research analyzes the position of corpus linguistics within the *Digital Humanities*, models the methodological transformation of Arabic studies that it engenders, and maps the challenges and prospects of its development. Adopting a qualitative, *library-research* design of a meta-methodological character, it combines conceptual and analytical approaches and examines reputable literature published between 2019 and 2026 through *content analysis*. As its principal contribution, the research formulates a *Three-Layer Model of Corpus-Based Methodological Transformation in Arabic Studies* linking a data-infrastructure layer, an analytical-procedure layer, and an epistemological layer together with a typology of corpus applications across four domains of Arabic studies (the Qur'an, *turāth*, modern media, and pedagogy). The model is illustrated concretely through frequency, collocation, and concordance analysis of the root م-ل-ع in the *Quranic Arabic Corpus*. The findings show that corpus linguistics drives a shift from *text-centered* toward *data-driven* studies that are more empirical and reproducible, although it still faces the limited availability of standardized corpora, morphological complexity and diglossia, academic digital literacy, and technological infrastructure. The research recommends developing representative, *open-access* Arabic digital corpora, strengthening scholars' digital literacy, and pursuing a balanced integration of classical philological methods and data-based approaches.

Keywords: Arabic Digital Corpus; Corpus Linguistics; *Digital Humanities*; Methodological Transformation Model; Arabic Studies



ملخص

أحدث التحوّل نحو الإنسانيّات الرقميّة (Digital Humanities) نقلةً في المشهد المنهجيّ للعلوم الإنسانيّة، ومنها الدراسات العربيّة، عبر توظيف المدوّنة الرقميّة أساساً لتحليل اللغة. غير أنّ الدراسات التي تربط لسانیّات المدوّنة (corpus linguistics) تحديداً بالتحوّل المنهجيّ في الدراسات العربيّة، وتقدّم إطاراً مفاهيمياً صريحاً، لا تزال محدودة. يهدف هذا البحث إلى تحليل موقع لسانیّات المدوّنة ضمن الإنسانيّات الرقميّة، ونمذجة التحوّل المنهجيّ الذي تُحدثه في الدراسات العربيّة، ورسم تحدياته وآفاقه. اعتمد البحث المنهج الكيفيّ عبر الدراسة المكتبيّة (library research) ذات الطابع المبتنا-منهجيّ، جامعاً بين المقاربتين المفاهيميّة والتحليليّة، ومحلّلاً أدبيّات محكّمة صادرة بين عاميّ ٢٠١٩ و ٢٠٢٦ عبر تحليل المضمون (content analysis). وكإسهام رئيس، صاغ البحث «نموذج الطبقات الثلاث للتحوّل المنهجيّ المستند إلى المدوّنة في الدراسات العربيّة»، الذي يربط بين طبقة البنية التحتيّة للبيانات، وطبقة الإجراء التحليليّ، والطبقة المعرفيّة (الإبستمولوجيّة)، إلى جانب تصنيف (typology) لتطبيقات المدوّنة في أربعة مجالات للدراسات العربيّة: القرآن الكريم، والتراث، والإعلام الحديث، والتعليم. وقد جرى توضيح النموذج تطبيقياً من خلال تحليل التواتر، والمصاحبة اللفظيّة، والسياقات الكشفية (concordance) لجذر (ع-ل-م) في المدوّنة القرآنيّة العربيّة (Quranic Arabic Corpus data). وأظهرت النتائج أنّ لسانیّات المدوّنة تدفع نحو الانتقال من الدراسات المتمركزة حول النصّ (text-centered) إلى الدراسات المقودة بالبيانات (-data driven) الأكثر تجريبيّةً وقابليّةً للتكرار، على الرغم مما تواجهه من محدوديّة المدوّنة العربيّة المعياريّة، وتعقيد البنية الصرفيّة والازدواجيّة اللغويّة (diglossia)، وضعف المعرفة الرقميّة الأكاديميّة، وقصور البنية التحتيّة التقنيّة. ويوصي البحث بتطوير مدوّنة عربيّة رقميّة تمثيليّة ومفتوحة الوصول (open access)، وتعزيز المعرفة الرقميّة لدى الباحثين، وتحقيق تكامل متوازن بين المناهج الفيلولوجيّة الكلاسيكيّة والمقاربات المستندة إلى البيانات.

الكلمات المفتاحية: المدونة العربية الرقمية؛ لسانيات المدونة؛ الإنسانيات الرقمية؛ نموذج التحوّل المنهجي؛ الدراسات العربية.

Abstrak

Transformasi *Digital Humanities* telah mengubah lanskap metodologis ilmu humaniora, termasuk studi Arab, melalui pemanfaatan korpus digital sebagai basis analisis kebahasaan. Namun, kajian yang secara khusus menghubungkan linguistik korpus dengan transformasi metodologi studi Arab dan menawarkan kerangka konseptual yang eksplisit masih terbatas. Penelitian ini bertujuan menganalisis posisi linguistik korpus dalam *Digital Humanities*, memodelkan transformasi metodologi studi Arab yang ditimbulkannya, serta memetakan tantangan dan prospek pengembangannya. Penelitian menggunakan pendekatan kualitatif dengan metode kepustakaan (*library research*) yang bersifat meta-metodologis, memadukan pendekatan konseptual dan analitis, serta menganalisis literatur bereputasi terbitan 2019–2026 melalui analisis isi (*content analysis*). Sebagai kontribusi utama, penelitian merumuskan *Model Tiga Lapis Transformasi Metodologi Studi Arab Berbasis Korpus* yang menautkan lapis infrastruktur data, lapis prosedur analitis, dan lapis epistemologis serta tipologi penerapan korpus pada empat ranah studi Arab (Al-Qur'an, turats, media modern, dan pedagogi). Model tersebut diilustrasikan secara konkret melalui analisis frekuensi, kolokasi, dan konkordansi atas akar ع-ل-م dalam *Quranic Arabic Corpus*. Hasil menunjukkan bahwa linguistik korpus mendorong pergeseran dari pendekatan *text-centered* menuju *data-driven studies* yang lebih empiris dan reproduktif, meski masih menghadapi keterbatasan korpus terstandarisasi, kompleksitas morfologi dan diglosia, literasi digital akademik, serta infrastruktur teknologi. Penelitian merekomendasikan pengembangan *Arabic digital corpus* yang representatif dan *open-access*, peningkatan literasi digital akademisi, serta integrasi seimbang antara metode filologis klasik dan pendekatan berbasis data.

Kata Kunci: *Arabic Digital Corpus*; *Digital Humanities*; Linguistik Korpus; Model Transformasi Metodologi; Studi Arab.

How to Cite This Article: Raharjo, Harmanto, Muhammad Arwani, Ahmad Syahid, Mahrus As'ad, Ateng Ruhendi, Aulia Rahmi, and Henny Septia Utami. "A Corpus-Based Model of Methodological Transformation in Arabic Studies Within the Digital Humanities Era". *Jurnal Adabiyah* 26, no. 2 (September 26, 2026). Accessed September 28, 2026. <https://journal.uin-alauddin.ac.id/index.php/adabiyah/article/view/70155>.

A. Introduction

The emergence of the Digital Humanities (DH) has integrated computational tools with humanistic inquiry,¹ giving rise to data-driven linguistics that prioritizes digital corpora as the primary empirical source.² This paradigm shift presents a significant methodological transition for Arabic linguistics, which has traditionally relied on philological, interpretive, and conventional textual frameworks. Precisely because these frameworks depend on the intuition and close reading of individual scholars, they reach clear limits of scale when confronted with the massive flows of digitized Arabic text now available an obstacle that empirical, corpus-based methods are designed to overcome. Yet despite growing demands for empirical rigor and reproducibility, this transition remains inadequately conceptualized, leaving a critical gap in how knowledge of the Arabic language is constructed and validated in modern scholarly discourse.³

Prior research highlights the growing integration of corpus methods into Arabic scholarship. Hizbullah et al. utilize corpus linguistics to analyze manuscript lineages and textual patterns in the Nusantara region,⁴ while Belinkov et al. leverage computationally annotated corpora to investigate the diachronic development of Arabic.⁵ In sacred-text analysis, Sawalha et al. introduced a morphologically and syntactically annotated Qur'anic dataset to facilitate advanced computational semantics.⁶ Together, these initiatives confirm the technical maturity and methodological versatility of corpus analysis across classical, historical, and Qur'anic textual domains.

¹ Eid Mohamed, "The Potential and Limits of Arabic Digital Humanities," *Journal of Cultural Analytics* 9, no. 3 (June 17, 2024): 889, doi:10.22148/001c.116818.

² Hamamah Hamamah, "To Data Driven Research and Beyond: An Overview of Corpus Linguistics Research in Indonesia," in *Proceedings of the 2nd International Conference on Language, Literature, Education, and Culture, ICOLLEC 2022, 11–12 November 2022, Malang, Indonesia* (Proceedings of the 2nd International Conference on Language, Literature, Education, and Culture, ICOLLEC 2022, 11–12 November 2022, Malang, Indonesia, Malang, Indonesia: EAI, 2023), doi:10.4108/eai.11-11-2022.2329401.

³ Karin Priem and Lynn Fendler, "Shifting Epistemologies for Discipline and Rigor in Educational Research: Challenges and Opportunities from Digital Humanities," *European Educational Research Journal* 18, no. 5 (September 2019): 610–21, doi:10.1177/1474904118820433.

⁴ Nur Hizbullah, Iin Suryaningsih, and Zaqiatul Mardiah, "Manuskrip Arab di Nusantara Dalam Tinjauan Linguistik Korpus," *Arabi: Journal of Arabic Studies* 4, no. 1 (July 1, 2019): 65, doi:10.24865/ajas.v4i1.145.

⁵ Yonatan Belinkov et al., "Studying the History of the Arabic Language: Language Technology and a Large-Scale Historical Corpus," *Language Resources and Evaluation* 53, no. 4 (December 2019): 771–805, doi:10.1007/s10579-019-09460-w.

⁶ Majdi Sawalha et al., "Morphologically-Analyzed and Syntactically-Annotated Quran Dataset," *Data in Brief* 58 (February 2025): 111211, doi:10.1016/j.dib.2024.111211.

Broadening the conceptual scope, Mohamed evaluates the capabilities and limitations of the Arabic Digital Humanities, pointing to infrastructural deficits and an overemphasis on text digitization.⁷ Complementing this, Bashir et al.'s systematic review of Qur'anic Natural Language Processing underscores the computational challenges of processing Arabic.⁸ Taken together, these studies reveal a field progressing swiftly in technical capability, yet still divided between specialized corpus construction and generalized Digital Humanities discourse.

This landscape highlights a clear conceptual divide: current literature isolates technical corpus engineering from broader Digital Humanities theory. Rarely do studies integrate these approaches to examine how corpus linguistics redefines Arabic research epistemologically namely, by transitioning from intuition-based textual authority to verifiable empirical analysis. As a result, the methodological ramifications of corpus linguistics for data validity, objectivity, and interdisciplinary integration in Arabic studies remain inadequately articulated.

This research offers novelty on two levels. Conceptually, it positions corpus linguistics not merely as a technical instrument but as a methodological approach that reconstructs the epistemology of contemporary Arabic studies. In terms of contribution, it formulates an analytical artifact a *Three-Layer Model of Corpus-Based Methodological Transformation in Arabic Studies*, together with a typology of its applications that bridges the *Digital Humanities* discourse and the methodological practice of Arabic studies.

B. Research Method

This research adopts a qualitative, meta-methodological framework to evaluate the paradigm shift driven by corpus linguistics within the Arabic Digital Humanities. Rather than deploying corpus tools as empirical instruments on a specific textual dataset, this paper treats corpus linguistics itself as the primary object of inquiry critically examining its epistemological premises, procedural mechanics, and methodological implications for contemporary Arabic studies. Given the conceptual nature of this investigation, the research design integrates theoretical synthesis and critical content analysis to map the systemic transition from conventional philological approaches to data-driven paradigms.

The research analyzes a curated body of primary and secondary scholarly literature published between 2019 and 2026. Primary sources comprise peer-reviewed journal articles, authoritative monographs, and foundational treatises directly addressing Arabic corpus linguistics, computational linguistics, Natural Language

⁷ Mohamed, "The Potential and Limits of Arabic Digital Humanities."

⁸ Muhammad Huzaifa Bashir et al., "Arabic Natural Language Processing for Qur'anic Research: A Systematic Review," *Artificial Intelligence Review* 56, no. 7 (July 2023): 6801–54, doi:10.1007/s10462-022-10313-2.

Processing (NLP), and Digital Humanities epistemology; priority was assigned to Scopus-indexed international journals and accredited national journals. Secondary sources comprise edited-volume chapters, conference proceedings, specialized project reports, and digital-archive documentation providing context on Arabic digital infrastructure. During the data-reduction stage, a source was retained when it (a) fell within the 2019–2026 window and (b) engaged directly with at least one of the study's four thematic pillars corpus linguistics, the Digital Humanities, computational linguistics/NLP, or the methodology of Arabic studies. Conversely, works falling outside this window, or bearing no substantive relation to those pillars, were excluded from the synthesis.

The retained literature was processed through a four-stage qualitative content-analysis framework. (1) Data filtering and reduction: screening sources against the inclusion criteria above. (2) Thematic categorization: grouping the literature into five conceptual domains Digital Humanities paradigms, the properties of corpus linguistics, methodological shifts in Arabic studies, structural/epistemological challenges, and technological futures. (3) Conceptual interpretation and synthesis: mapping argumentative patterns, identifying infrastructural bottlenecks, and modeling the interaction between computation and philological hermeneutics. (4) Artifact modeling: consolidating the findings into the proposed Three-Layer Model and its typology of domain applications. Throughout, sources were read comparatively and cross-checked against one another to surface convergences and tensions across the literature.

C. Results and Discussion

1. The Position of Corpus Linguistics in the Development of the Digital Humanities and the Paradigm Shift in Arabic Studies

Answering the first research question the position of corpus linguistics within the development of the *Digital Humanities* requires first understanding the evolution of that paradigm. In its early stage, the *Digital Humanities* was understood merely as the digitization of documents, archives, and academic texts into electronic form. This understanding treated technology as a mere means of storage and reproduction rather than as a new mode of thought. However, advances in computational technology, the internet, and data analysis shifted its meaning into a methodological approach that shapes how knowledge is produced, validated, and disseminated.⁹ This shift marks a transition from traditional humanities toward *computational humanities*, which draws on *big data*, statistical analysis, and artificial intelligence to understand linguistic and

⁹ Nanna Bonde Thylstrup, Mikkel Flyverbom, and Rasmus Helles, "Datafied Knowledge Production: Introduction to the Special Theme," *Big Data & Society* 6, no. 2 (July 2019): 205395171987598, doi:10.1177/2053951719875985.

cultural phenomena more systematically.¹⁰ It is within this transition that corpus linguistics occupies a strategic position: if the *Digital Humanities* supplies the epistemological framework, corpus linguistics supplies the operational procedure for realizing it.

In the context of Arabic studies, this development is marked by the increasing digitization of texts across genres classical manuscripts, *turāth* works, dictionaries, academic journals, online media, and digital Qur'anic corpora which can now be accessed and analyzed computationally.¹¹ This digitization has produced the *Arabic digital corpus*: systematically compiled bodies of Arabic-language data that support linguistic, philological, and modern language research. Their existence enables research at a scale far exceeding the manual approach that has long predominated.¹² This change of scale is not merely a matter of quantity; it alters the kinds of questions that can be posed. Patterns of lexical distribution across thousands of texts, the relative frequency of a term in a given period, or variation in the collocations of a word across contexts are questions almost impossible to answer through manual reading, yet they become tractable once data are available in structured corpus form.

The position of corpus linguistics becomes clearer still when linked to *computational linguistics*, which integrates linguistics and computer science to build technology-based systems of language analysis.¹³ In Arabic studies, this integration has developed through *Natural Language Processing* (NLP), machine translation, digital semantic analysis, the recognition of morphological patterns, and AI-based systems for text search and exploration.¹⁴ Corpus linguistics thus becomes the meeting point that connects the availability of digital corpora, as material, with computational capacity, as the tool that processes them. Without corpora, computational technology loses its object; without computation, large-scale corpora lose their analytical power.

¹⁰ Béatrice Joyeux-Prunel, "Digital Humanities in the Era of Digital Reproducibility: Towards a Fairest and Post-Computational Framework," *International Journal of Digital Humanities* 6, no. 1 (January 3, 2024): 23–43, doi:10.1007/s42803-023-00079-6.

¹¹ Syahrullah Syahrullah et al., "Corpus Coranicum and Digital Philology: A Methodological Model for Advancing Qur'anic Manuscript Studies in Indonesia," *Khazanah Theologia* 6, no. 3 (December 29, 2024): 201–22, doi:10.15575/kt.v6i3.45610.

¹² Andreas Hallberg, "An 81-Million-Word Multi-Genre Corpus of Arabic Books," *Data in Brief* 60 (June 2025): 111456, doi:10.1016/j.dib.2025.111456.

¹³ Hussein Ala'a Alkaabi, Ali Kadhim Jasim, and Ali Darroudi, "Arabic NLP: A Survey of Pre-Processing and Representation Techniques," *JCoSITTE: Journal of Computer Science, Information Technology and Telecommunication Engineering* 6, no. 2 (2025): 876–90.

¹⁴ Noor Mohamed Mohd Yousof et al., "An Intelligent NLP-Based Framework for Digital Scripts Concordance and Semantic Exploration," in *Frontiers in Artificial Intelligence and Applications*, ed. Hamido Fujita, Andres Hernandez-Matamoros, and Yutaka Watanobe (Amsterdam: IOS Press, 2025), doi:10.3233/FAIA250545.

The most fundamental consequence of this positioning is an epistemological shift. Under the older paradigm, the validity of Arabic-language research rested on textual authority and the researcher's linguistic intuition an approach that contributed greatly to the classical Arabic linguistic tradition but proved limited when confronted with large-scale data.¹⁵ By contrast, corpus linguistics places authentic data at the center of forming scholarly generalizations about Arabic usage.¹⁶ Arabic studies thereby shift from a model of *text-centered studies* toward *data-centered studies* that emphasize the objectivity of data, reproducibility, and transparency of analytical procedure. This shift also broadens the interdisciplinary character of Arabic studies through the integration of linguistics, computer science, data analysis, and artificial intelligence.¹⁷ Yet, consistent with the critical stance of this research, the dominance of a computational approach risks reducing linguistic meaning to mere statistical patterns, while the condition of *diglossia* the coexistence of Classical Arabic, *Modern Standard Arabic*, and dialectal varieties renders the claim to a single representative Arabic corpus problematic. The positioning of corpus linguistics should therefore be understood as a methodological expansion that preserves the interpretive character of the Arab-Islamic humanistic tradition.

2. Corpus Linguistics as a Methodological Approach in Arabic Studies

Corpus linguistics has developed into an important approach in modern language research that places authentic data at the center of linguistic analysis.¹⁸ What distinguishes it from a mere data-collection technique is its epistemological dimension: it fosters *evidence-based linguistics*, so that the analysis of language no longer rests entirely on the researcher's intuition but on data that can be verified computationally.¹⁹ This is the core of the methodological shift relevant to Arabic studies from reasoning centered on the interpreter's authority toward reasoning that can be re-examined.

¹⁵ Donia Gamal et al., "Implementation of Machine Learning Algorithms in Arabic Sentiment Analysis Using N-Gram Features," *Procedia Computer Science* 154 (2019): 332–40, doi:10.1016/j.procs.2019.06.048.

¹⁶ Abdelhamid Elewa, "Not Ready for Democracy: A Corpus-Assisted Discourse Analysis of the Representation of Majorities in Arabic Media Narratives," *Discourse & Society* 36, no. 6 (November 2025): 863–82, doi:10.1177/09579265251323906.

¹⁷ Bashir et al., "Arabic Natural Language Processing for Qur'anic Research."

¹⁸ Pascual Pérez-Paredes and Niall Curry, "Epistemologies of Corpus Linguistics across Disciplines," *Research Methods in Applied Linguistics* 3, no. 3 (December 2024): 100141, doi:10.1016/j.rmal.2024.100141.

¹⁹ Adrianto Ramadhan, Weningtyas Parama Iswari, and Aridah, "Corpus-Based Study on The Use of Reporting Verbs In Applied Linguistics Journal Articles Published From 2020-2024," *E3L: Journal of English Teaching, Linguistic, and Literature* 8, no. 1 (June 25, 2025): 9–26, doi:10.30872/e3l.v8i1.5242.

This methodological character rests on three properties of the data: authenticity, empiricity, and representativeness.²⁰ Authenticity means that the corpus is assembled from real language use; empiricity is evident in analysis grounded in measurable data; and representativeness requires that a corpus be designed to reflect a given range of language use. For Arabic studies, it is precisely representativeness that is most challenging given the wide span of Arabic varieties so that corpus design cannot be separated from critical consideration of genre, period, and source texts, as well as transparent reporting of corpus design.²¹ Methodologically, corpus linguistics has developed through two orientations: the *corpus-based*, which uses a corpus to verify existing theory, and the *corpus-driven*, which treats corpus data as the point of departure for building new theory.²² In Arabic studies, the *corpus-driven* orientation opens the possibility of discovering patterns unforeseen by the traditional grammatical framework, while the *corpus-based* orientation is useful for testing the rules of classical *naḥw* (syntax) and *balāghah* (rhetoric) against actual usage evidence.

In practice, corpus linguistics operates a range of techniques: frequency analysis, collocation, concordance, semantic analysis, and corpus-based discourse analysis.²³ Frequency analysis measures the occurrence of vocabulary; collocation analysis reveals words that tend to co-occur; concordance analysis displays variation in a word's use according to context; and semantic analysis identifies tendencies of meaning and lexical relations. Furthermore, corpus-based discourse analysis allows research to move beyond the level of structure toward the level of ideology, narrative, and social tendencies in language use.²⁴ This array of techniques equips Arabic studies with concrete tools for moving from normative description toward an empirical description of linguistic behavior.

Its application to Arabic studies is evident in several domains. The digitization of the Qur'anic corpus enables the identification of semantic patterns, lexical distribution, and inter-term relations more empirically, without discarding the interpretive

²⁰ Jesse Egbert, Douglas Biber, and Bethany Gray, *Designing and Evaluating Language Corpora: A Practical Framework for Corpus Representativeness*, 1st ed. (Cambridge: Cambridge University Press, 2022), doi:10.1017/9781316584880.

²¹ Brett Hashimoto and Kyra Nelson, "Recent Trends in Corpus Design and Reporting: A Methodological Synthesis," *Research in Corpus Linguistics* 12, no. 1 (2024): 59–88, doi:10.32714/ricl.12.01.03.

²² Matthew H.C. Mak, "Corpus Linguistics Will Benefit from Greater Adoption of Pre-Registration: A Novice-Friendly Split-Corpus Approach to Pre-Registration," *Applied Corpus Linguistics* 4, no. 3 (December 2024): 100111, doi:10.1016/j.acorp.2024.100111.

²³ Stefan Th. Gries, "A New Approach to (Key) Keywords Analysis: Using Frequency, and Now Also Dispersion," *Research in Corpus Linguistics* 9, no. 2 (2021): 1–33, doi:10.32714/ricl.09.02.02.

²⁴ Nur Nabilah Fauziyah and Ni Gusti Ayu Roselani, "A Corpus-Based Critical Discourse Analysis of English Discourse in The Jakarta Post," *K@ta* 27, no. 1 (June 30, 2025): 80–94, doi:10.9744/kata.27.1.80-94.

dimension of Islamic studies.²⁵ In *turāth* texts, the corpus approach helps trace the development of terms and patterns of argument within the classical Arab intellectual tradition.²⁶ In education, authentic language data enable the design of more contextual and communicative teaching materials;²⁷ indeed, several studies have already operationalized corpus tools such as AntConc, Sketch Engine, and ArabiCorpus to build Arabic-language learning materials for non-native speakers.²⁸ Corpus linguistics thus functions not merely as an analytical technique but as a methodological approach that reconstructs how research in Arabic studies is conducted shifting the fulcrum from individual interpretation toward authentic, empirical, and representative linguistic data. Nonetheless, the reliability of all these techniques still depends on the quality of corpus design, an issue examined in the following section.

3. Three-Layer Model of Corpus-Based Methodological Transformation in Arabic Studies

This section answers the second question by synthesizing the preceding discussion into a conceptual framework. The convergence of corpus linguistics and the *Digital Humanities* brings about change that is not confined to research instruments but reaches the paradigm through which Arabic is understood as an object of inquiry.²⁹ To capture the structure of this transformation, the research proposes a *Three-Layer Model of Corpus-Based Methodological Transformation in Arabic Studies*. The model asserts that methodological transformation is not a mere substitution of tools but a stratified reconfiguration in which each layer is a precondition for the layer above it.

The first layer, data infrastructure, comprises the *Arabic digital corpus* across genres manuscripts, *turāth*, the Qur'anic corpus, online media, and digital dictionaries. This layer supplies the large-scale empirical material; without adequate data availability, a paradigm shift will not materialize through a change of theoretical

²⁵ Nur Hizbullah, Fazlur Rachman, and Fuzi Fauziah, "Penyusunan Model Korpus Al-Qur'an Digital," *Jurnal Al-Azhar Indonesia Seri Humaniora* 3, no. 3 (December 20, 2017): 215, doi:10.36722/sh.v3i3.209.

²⁶ Ibid.

²⁷ Nur Hizbullah et al., "Source-Based Arabic Language Learning: A Corpus Linguistic Approach," *Humanities & Social Sciences Reviews* 8, no. 3 (June 17, 2020): 940–54, doi:10.18510/hssr.2020.8398.

²⁸ Muhammet Abazoglu and Mohammad Issa Alhourani, "The Use of Language Corpora in Teaching Arabic to Turkish Speakers Within the Framework of Computational Linguistics," *Social Sciences & Humanities Open* 12 (2025): 101947, doi:10.1016/j.ssaho.2025.101947.

²⁹ Mohammed Q. Shormani and Mohammad A. N. Alenezi, "Arabic Nominalization as Form of Pragmatic Actions in Digital Discourse: A Corpus-Based Study," *Corpus Pragmatics* 10, no. 1 (December 2026): 35, doi:10.1007/s41701-026-00237-5.

attitude alone.³⁰ *The second layer, analytical procedure*, converts data into linguistic patterns through corpus techniques (frequency, collocation, concordance, semantics, discourse analysis) reinforced by NLP and artificial intelligence; it is at this layer that Arabic-language research moves from manual reading toward the systematic processing of *massive linguistic data*.³¹ *The third layer, the epistemological*, is where the most fundamental transformation occurs: validity no longer rests solely on the strength of interpretation but on the consistency of data patterns that can be re-examined, rendering research more objective and reproducible.³²

These three layers map four interrelated axes of shift: from an intuitive to a data-based approach, from manual to digital analysis, from limited data to large-scale corpora, and from subjective to empirical and reproducible analysis. The principal enabling factor is the growth of *Arabic digital databases*, which allow the systematic identification of lexical patterns, morphological distribution, and semantic relations.³³ Consequently, research validity shifts from the interpreter's authority toward the testability of data,³⁴ and the pattern of interpretation increasingly rests on measurable evidence of language use.³⁵

To demonstrate the model's applicability, Table 1 presents a typology of corpus-linguistic applications across four principal domains of Arabic studies, together with their dominant techniques, epistemological functions, and illustrative studies. The typology shows that methodological transformation is not uniform but varies according to the character of the object from Qur'anic and *turāth* studies to modern media and pedagogy.³⁶

³⁰ Abd Aziz and Saihu Saihu, "Interpretasi Humanistik Kebahasaan: Upaya Kontekstualisasi Kaidah Bahasa Arab," *Arabiyatuna: Jurnal Bahasa Arab* 3, no. 2 (November 13, 2019): 299, doi:10.29240/jba.v3i2.1000.

³¹ Hallberg, "An 81-Million-Word Multi-Genre Corpus of Arabic Books."

³² Joseph Flanagan, "Reproducibility, Replicability, Robustness, and Generalizability in Corpus Linguistics," *International Journal of Corpus Linguistics* 30, no. 2 (October 2, 2025): 130–49, doi:10.1075/ijcl.24113.fl.

³³ Syahrullah et al., "Corpus Coranicum and Digital Philology."

³⁴ Alaa' Mohammad Smadi, "A Corpus-Based Analysis of the Lexical and Grammatical Properties of Arabic Abstracts in Social Sciences Articles," *International Journal of Linguistics* 12, no. 6 (November 24, 2020), doi:10.5296/ijl.v12i6.17849.

³⁵ Flanagan, "Reproducibility, Replicability, Robustness, and Generalizability in Corpus Linguistics."

³⁶ Gavin Brookes and Tony McEnery, "Correlation, Collocation and Cohesion: A Corpus-Based Critical Analysis of Violent Jihadist Discourse," *Discourse & Society* 31, no. 4 (July 2020): 351–73, doi:10.1177/0957926520903528.

Table 1. A Typology of Corpus-Linguistic Applications in Arabic Studies

Domain of Arabic Studies	Dominant Corpus Techniques	Epistemological Function	Illustrative Studies
Qur'anic studies	Semantic analysis, lexical distribution, concordance	Reveals patterns of meaning and inter-term relations empirically without discarding the interpretive (tafsīr) dimension	Sawalha et al. (2025); Hizbullah et al. (2017)
<i>Turāth</i> (classical texts)	Term-frequency analysis, collocation, diachronic analysis	Traces the development of terms and patterns of argument within the classical intellectual tradition	Belinkov et al. (2019); Hizbullah et al. (2019)
Modern Arabic media	Corpus-based discourse analysis, collocation	Uncovers discursive patterns, ideology, and socio-political discourse	Elewa (2025); Shormani & Alenezi (2026); Brookes & McEnery (2020)
Arabic-language pedagogy	Frequency, concordance, <i>learner corpus</i>	Enables the design of contextualized teaching materials grounded in authentic usage	Hizbullah et al. (2020); Abazoglu & Alhourani (2025)

With this framework, the research's contribution becomes explicit: the three-layer model and its typology provide a conceptual instrument for positioning, comparing, and designing corpus-based research in Arabic studies more systematically. Even so, in keeping with the study's critical stance, the model treats the epistemological layer not as a deterministic endpoint but as a space that still demands critical interpretation. The development of corpus linguistics is therefore understood not as an effort to replace the traditional approach but as an instrument that complements classical philological and interpretive methods³⁷ an integration that is a precondition for a model of Arabic-studies research that is at once objective and humanistic. Yet this entire potential is realized only if the structural challenges at the infrastructure layer can be overcome, the issue that forms the focus of the next section.

³⁷ Mohamed, "The Potential and Limits of Arabic Digital Humanities."

4. Concrete Illustration: Corpus Analysis of the Root ع-ل-م in the Qur'an

To demonstrate how the three-layer model works in concrete terms, this section offers a brief illustration not a full empirical study of the root ع-ل-م ('*l-m*, “to know”) using the *Quranic Arabic Corpus*, a morphologically annotated corpus of the Qur'an comprising 77,430 words developed by Dukes and Habash.³⁸ This corpus represents the *data-infrastructure layer*: every word has been tokenized, lemmatized, and annotated for part of speech, lemma, and root, so that it can be traced computationally and re-examined by other researchers.³⁹ The root ع-ل-م is chosen because it bears directly on the very question of epistemology the production and status of knowledge that this article addresses.

At the *analytical-procedure layer*, frequency analysis shows that the root ع-ل-م occurs 854 times in fourteen derived forms.⁴⁰ Table 2 summarizes the distribution of its principal forms. This distribution is more than a set of figures: it reveals that “knowledge” in Qur'anic discourse appears predominantly as the *act of knowing*, through the verb '*alima* (382 occurrences), and as an *attribute or superlative of knowing that is mostly ascribed to God*, through '*alīm* “All-Knowing” (163), a '*lam* (49), and '*allām* (4) together (216) exceeding the occurrence of “knowledge” as the abstract noun '*ilm* (105). Such a quantitative pattern is difficult to capture through manual reading but becomes visible through corpus aggregation.

Table 2. Frequency Distribution of the Root ع-ل-م in the Quranic Arabic Corpus

Derived form	Gloss	Frequency
Verb ' <i>alima</i> (Form I) — عَلمَ	to know	382
Nominal ' <i>alīm</i> — عَلِيم	All-Knowing	163
Noun ' <i>ilm</i> — عِلْم	knowledge	105
Noun ' <i>ālamīn</i> — عَالَمِينَ	the worlds/universe	73
Nominal a ' <i>lam</i> — أَعْلَمَ	knows best/better	49
Verb ' <i>allama</i> (Form II) — عَلَّمَ	to teach	41
Active participle ' <i>alim</i> — عَالِم	knower	18
Passive participle ma ' <i>lūm</i> — مَعْلُوم	(that which is) known	11
Six other forms	' <i>allām</i> , yata ' <i>allamu</i> , etc.	12

³⁸ Kais Dukes and Nizar Habash, “Morphological Annotation of Quranic Arabic,” in *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010)* (Seventh International Conference on Language Resources and Evaluation (LREC 2010), Valletta: European Language Resources Association (ELRA), 2010), 2530–36.

³⁹ Sawalha et al., “Morphologically-Analyzed and Syntactically-Annotated Quran Dataset.”

⁴⁰ Gries, “A New Approach to (Key) Keywords Analysis.”

Total (root ع-ل-م)	14 derived forms	854
--------------------	------------------	-----

Collocation analysis reinforces this finding: the form *‘alīm* recurrently pairs with other divine attributes, notably *ḥakīm* (“All-Wise”) in the formula *‘alīm ḥakīm*, and *samī‘* (“All-Hearing”) in *samī‘ ‘alīm* indicating that “knowledge” is systematically bound to the discourse of God’s attributes. Concordance analysis (*keyword in context*), in turn, displays patterns of use in real contexts: the rhetorical contrast between “those who know” and “those who do not know” (Q 39:9), the plea “My Lord, increase me in knowledge” (Q 20:114), and the assertion that humankind has “been given only a little knowledge” (Q 17:85). These three techniques frequency, collocation, and concordance operate on the same analytical layer and complement one another, turning raw data into meaningful linguistic patterns.

At the *epistemological layer*, this illustration shows how claims about the concept of “knowledge” in the Qur’an can be grounded in quantitative evidence that is verifiable and re-testable through the same corpus not to replace *tafsīr*, but to give it a transparent empirical foundation. Yet it must be stressed, critically, that the success of this illustration rests partly on the exceptional character of the Qur’an as a corpus: its text is fixed, standardized, and carefully annotated. Most other Arabic corpora *turāth* manuscripts and modern media alike do not possess a comparable degree of standardization, so that replicating this approach in other domains runs into the structural obstacles discussed in the following section.

5. Challenges and Prospects for the Development of Corpus Linguistics in Arabic Studies

In answer to the third question, the development of corpus linguistics in Arabic studies opens up great opportunities while leaving real methodological, technological, and epistemological challenges.⁴¹ The first and most fundamental obstacle one that strikes the infrastructure layer of the model above is the limited availability of representative, standardized Arabic corpora.⁴² To date, most Arabic corpora remain scattered across non-uniform digital formats, hindering data integration and optimal computational analysis. The problem of standardization compounds this: variation in transliteration systems, annotation schemes, storage formats, and metadata structures frequently produces technical inconsistency.⁴³ Without standardization, the claims to

⁴¹ Dukes and Habash, “Morphological Annotation of Quranic Arabic.”

⁴² Abdulaziz M. Alayba, “Arabic Natural Language Processing (NLP): A Comprehensive Review of Challenges, Techniques, and Emerging Trends,” *Computers* 14, no. 11 (November 15, 2025): 497, doi:10.3390/computers14110497.

⁴³ Fatma Zahra Besdouri, Inès Zribi, and Lamia Hadrich Belguith, “Arabic Automatic Speech Recognition: Challenges and Progress,” *Speech Communication* 163 (September 2024): 103110, doi:10.1016/j.specom.2024.103110.

objectivity and re-examinability that constitute the strength of the corpus approach risk going unmet.

The next challenge stems from the character of Arabic itself, namely its morphological complexity. Derivational structures, the root system, and syntactic variation render the processes of tokenization, stemming, part-of-speech tagging, and semantic analysis considerably more complex.⁴⁴ This complexity depresses the accuracy of Arabic NLP technology, especially for classical texts and turāth manuscripts, which differ in character from Modern Standard Arabic, and it is aggravated by diglossia, which makes a single computational model difficult to apply uniformly. Beyond technical matters, there are challenges of human resources and academic digital literacy: some Arabic-studies scholars remain more comfortable with conventional philological approaches than with technology-based methods, so that the use of corpus tools is uneven.⁴⁵ Infrastructural factors servers, software, storage, and computational access complete this chain of obstacles,⁴⁶ such that developing a Digital Humanities ecosystem demands stronger institutional support and multidisciplinary collaboration.

Nevertheless, the prospects remain considerable. Advances in artificial intelligence, big data, and NLP open opportunities for developing machine translation, semantic analysis of the Qur'an, morphological pattern recognition, and the mapping of Arabic discourse more rapidly and accurately. Another important prospect lies in corpus-based digital Arabic-language learning, which draws on authentic data to design contextual teaching materials and to support digital dictionaries and interactive learning media.⁴⁷ Furthermore, corpus linguistics opens the way to interdisciplinary research among Arabic studies, linguistics, computer science, and data science; the development of open-access corpora becomes crucial because it broadens access to data, encourages global collaboration, and strengthens transparency and reproducibility.⁴⁸ Yet consistent with this study's critical stance all of these prospects still require epistemological reflection so that a data-based approach does not reduce an Arabic language that carries historical, cultural, and spiritual dimensions. For this reason, the development of corpus

⁴⁴ Anis Charfi et al., "A Fine-Grained Annotated Multi-Dialectal Arabic Corpus," in *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)* (Varna: INCOMA Ltd., 2019), 198–204, doi:10.26615/978-954-452-056-4_023.

⁴⁵ Besdouri, Zribi, and Belguith, "Arabic Automatic Speech Recognition."

⁴⁶ Hanem El-Farahaty, Nouran Khallaf, and Amani Alonayzan, "Building the Leeds Monolingual and Parallel Legal Corpora of Arabic and English Countries' Constitutions: Methods, Challenges and Solutions," *Corpus Pragmatics* 7, no. 2 (June 2023): 103–19, doi:10.1007/s41701-023-00138-x.

⁴⁷ Alayba, "Arabic Natural Language Processing (NLP)."

⁴⁸ Abazoglu and Alhourani, "The Use of Language Corpora in Teaching Arabic to Turkish Speakers Within the Framework of Computational Linguistics."

linguistics is best understood as an effort to complement, not to replace, the classical tradition.

D. Conclusion

The rise of digital corpora marks a fundamental shift in Arabic studies, moving the discipline from an intuition-reliant, text-centered paradigm toward empirical, reproducible, data-driven inquiry. Grounded in the broader epistemological evolution of the Digital Humanities, this research demonstrates that corpus linguistics is not merely a technical instrument for textual retrieval but a framework that reconstructs how linguistic knowledge is generated and validated. As its principal contribution, the research advances the Three-Layer Model of Corpus-Based Methodological Transformation, which resolves a persistent dichotomy in current scholarship: the divide between purely technical NLP developments and broader theoretical reflection on the Arabic Digital Humanities. By explicitly linking a data-infrastructure layer, an analytical-procedure layer, and an epistemological layer, the model confirms established corpus-theoretical concerns with data representativeness and testability while contributing a structured roadmap for applying these principles across the Qur'anic, turāth, modern-media, and pedagogical domains. Its application to the root *ع-ل-م* in the Quranic Arabic Corpus further shows that quantitative aggregation frequency, collocation, and concordance provides an objective, re-testable foundation that complements rather than displaces classical tafsīr and philological hermeneutics.

Ultimately, corpus linguistics in Arabic studies must not be treated as an end in itself. Overcoming its systemic obstacles corpus non-standardization, morphological complexity, diglossia, and infrastructural disparity requires a balanced synthesis of traditional humanistic rigor and quantitative method that remains critical toward the risk of reducing meaning. Future research should test the three-layer framework empirically across broader classical archives and across each domain, ensuring that the digital transformation of Arabic studies remains at once methodologically robust and culturally grounded.

REFERENCES

- Abazoglu, Muhammet, and Mohammad Issa Alhourani. "The Use of Language Corpora in Teaching Arabic to Turkish Speakers Within the Framework of Computational Linguistics." *Social Sciences & Humanities Open* 12 (2025): 101947. doi:10.1016/j.ssaho.2025.101947.

- Alayba, Abdulaziz M. "Arabic Natural Language Processing (NLP): A Comprehensive Review of Challenges, Techniques, and Emerging Trends." *Computers* 14, no. 11 (November 15, 2025): 497. doi:10.3390/computers14110497.
- Alkaabi, Hussein Ala'a, Ali Kadhim Jasim, and Ali Darroudi. "Arabic NLP: A Survey of Pre-Processing and Representation Techniques." *JCoSITTE: Journal of Computer Science, Information Technology and Telecommunication Engineering* 6, no. 2 (2025): 876–90.
- Aziz, Abd, and Saihu Saihu. "Interpretasi Humanistik Kebahasaan: Upaya Kontekstualisasi Kaidah Bahasa Arab." *Arabiyatuna : Jurnal Bahasa Arab* 3, no. 2 (November 13, 2019): 299. doi:10.29240/jba.v3i2.1000.
- Bashir, Muhammad Huzaifa, Aqil M. Azmi, Haq Nawaz, Wajdi Zaghouani, Mona Diab, Ala Al-Fuqaha, and Junaid Qadir. "Arabic Natural Language Processing for Qur'anic Research: A Systematic Review." *Artificial Intelligence Review* 56, no. 7 (July 2023): 6801–54. doi:10.1007/s10462-022-10313-2.
- Belinkov, Yonatan, Alexander Magidow, Alberto Barrón-Cedeño, Avi Shmidman, and Maxim Romanov. "Studying the History of the Arabic Language: Language Technology and a Large-Scale Historical Corpus." *Language Resources and Evaluation* 53, no. 4 (December 2019): 771–805. doi:10.1007/s10579-019-09460-w.
- Besdouri, Fatma Zahra, Inès Zribi, and Lamia Hadrach Belguith. "Arabic Automatic Speech Recognition: Challenges and Progress." *Speech Communication* 163 (September 2024): 103110. doi:10.1016/j.specom.2024.103110.
- Bonde Thylstrup, Nanna, Mikkel Flyverbom, and Rasmus Helles. "Datafied Knowledge Production: Introduction to the Special Theme." *Big Data & Society* 6, no. 2 (July 2019): 205395171987598. doi:10.1177/2053951719875985.
- Brookes, Gavin, and Tony McEnery. "Correlation, Collocation and Cohesion: A Corpus-Based Critical Analysis of Violent Jihadist Discourse." *Discourse & Society* 31, no. 4 (July 2020): 351–73. doi:10.1177/0957926520903528.
- Charfi, Anis, Wajdi Zaghouani, Syed Hassan Mehdi, and Esraa Mohamed. "A Fine-Grained Annotated Multi-Dialectal Arabic Corpus." In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, 198–204. Varna: INCOMA Ltd., 2019. doi:10.26615/978-954-452-056-4_023.
- Dukes, Kais, and Nizar Habash. "Morphological Annotation of Quranic Arabic." In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010)*, 2530–36. Valletta: European Language Resources Association (ELRA), 2010.

- Egbert, Jesse, Douglas Biber, and Bethany Gray. *Designing and Evaluating Language Corpora: A Practical Framework for Corpus Representativeness*. 1st ed. Cambridge: Cambridge University Press, 2022. doi:10.1017/9781316584880.
- Elewa, Abdelhamid. "Not Ready for Democracy: A Corpus-Assisted Discourse Analysis of the Representation of Majorities in Arabic Media Narratives." *Discourse & Society* 36, no. 6 (November 2025): 863–82. doi:10.1177/09579265251323906.
- El-Farahaty, Hanem, Nouran Khallaf, and Amani Alonayzan. "Building the Leeds Monolingual and Parallel Legal Corpora of Arabic and English Countries' Constitutions: Methods, Challenges and Solutions." *Corpus Pragmatics* 7, no. 2 (June 2023): 103–19. doi:10.1007/s41701-023-00138-x.
- Fauziyah, Nur Nabilah, and Ni Gusti Ayu Roselani. "A Corpus-Based Critical Discourse Analysis of English Discourse in The Jakarta Post." *K@ta* 27, no. 1 (June 30, 2025): 80–94. doi:10.9744/kata.27.1.80-94.
- Flanagan, Joseph. "Reproducibility, Replicability, Robustness, and Generalizability in Corpus Linguistics." *International Journal of Corpus Linguistics* 30, no. 2 (October 2, 2025): 130–49. doi:10.1075/ijcl.24113.fl.
- Gamal, Donia, Marco Alfonse, El-Sayed M. El-Horbaty, and Abdel-Badeeh M. Salem. "Implementation of Machine Learning Algorithms in Arabic Sentiment Analysis Using N-Gram Features." *Procedia Computer Science* 154 (2019): 332–40. doi:10.1016/j.procs.2019.06.048.
- Gries, Stefan Th. "A New Approach to (Key) Keywords Analysis: Using Frequency, and Now Also Dispersion." *Research in Corpus Linguistics* 9, no. 2 (2021): 1–33. doi:10.32714/ricl.09.02.02.
- Hallberg, Andreas. "An 81-Million-Word Multi-Genre Corpus of Arabic Books." *Data in Brief* 60 (June 2025): 111456. doi:10.1016/j.dib.2025.111456.
- Hamamah, Hamamah. "To Data Driven Research and Beyond: An Overview of Corpus Linguistics Research in Indonesia." In *Proceedings of the 2nd International Conference on Language, Literature, Education, and Culture, ICOLLEC 2022, 11–12 November 2022, Malang, Indonesia*. Malang, Indonesia: EAI, 2023. doi:10.4108/eai.11-11-2022.2329401.
- Hashimoto, Brett, and Kyra Nelson. "Recent Trends in Corpus Design and Reporting: A Methodological Synthesis." *Research in Corpus Linguistics* 12, no. 1 (2024): 59–88. doi:10.32714/ricl.12.01.03.
- Hizbullah, Nur, Zakiyah Arifa, Yoke Suryadarma, Ferry Hidayat, Luthfi Muhyiddin, and Eka Kurnia Firmansyah. "Source-Based Arabic Language Learning: A Corpus Linguistic Approach." *Humanities & Social Sciences Reviews* 8, no. 3 (June 17, 2020): 940–54. doi:10.18510/hssr.2020.8398.

- Hizbullah, Nur, Fazlur Rachman, and Fuzi Fauziah. "Penyusunan Model Korpus Al-Qur'an Digital." *Jurnal Al-Azhar Indonesia Seri Humaniora* 3, no. 3 (December 20, 2017): 215. doi:10.36722/sh.v3i3.209.
- Hizbullah, Nur, Iin Suryaningsih, and Zaqiatul Mardiah. "Manuskrip Arab Di Nusantara Dalam Tinjauan Linguistik Korpus." *Arabi : Journal of Arabic Studies* 4, no. 1 (July 1, 2019): 65. doi:10.24865/ajas.v4i1.145.
- Joyeux-Prunel, Béatrice. "Digital Humanities in the Era of Digital Reproducibility: Towards a Fairest and Post-Computational Framework,." *International Journal of Digital Humanities* 6, no. 1 (January 3, 2024): 23–43. doi:10.1007/s42803-023-00079-6.
- Mak, Matthew H.C. "Corpus Linguistics Will Benefit from Greater Adoption of Pre-Registration: A Novice-Friendly Split-Corpus Approach to Pre-Registration." *Applied Corpus Linguistics* 4, no. 3 (December 2024): 100111. doi:10.1016/j.acorp.2024.100111.
- Mohamed, Eid. "The Potential and Limits of Arabic Digital Humanities." *Journal of Cultural Analytics* 9, no. 3 (June 17, 2024): 889. doi:10.22148/001c.116818.
- Mohd Yousof, Noor Mohamed, Ali Selamat, Zatul Alwani Shaffiei, Siti Nur Khadijah Aishah Ibrahim, Liyana 'Adilla Burhanuddin, and Hamido Fujita. "An Intelligent NLP-Based Framework for Digital Scripts Concordance and Semantic Exploration." In *Frontiers in Artificial Intelligence and Applications*, edited by Hamido Fujita, Andres Hernandez-Matamoros, and Yutaka Watanobe. Amsterdam: IOS Press, 2025. doi:10.3233/FAIA250545.
- Pérez-Paredes, Pascual, and Niall Curry. "Epistemologies of Corpus Linguistics across Disciplines." *Research Methods in Applied Linguistics* 3, no. 3 (December 2024): 100141. doi:10.1016/j.rmal.2024.100141.
- Priem, Karin, and Lynn Fendler. "Shifting Epistemologies for Discipline and Rigor in Educational Research: Challenges and Opportunities from Digital Humanities." *European Educational Research Journal* 18, no. 5 (September 2019): 610–21. doi:10.1177/1474904118820433.
- Ramadhan, Adrianto, Weningtyas Parama Iswari, and Aridah. "Corpus-Based Study On The Use Of Reporting Verbs In Applied Linguistics Journal Articles Published From 2020-2024." *E3L: Journal of English Teaching, Linguistic, and Literature* 8, no. 1 (June 25, 2025): 9–26. doi:10.30872/e3l.v8i1.5242.
- Sawalha, Majdi, Faisal Al-Shargi, Sane Yagi, Abdallah T. AlShdaifat, Bassam Hammo, Mariam Belajeed, and Lubna R. Al-Ogaili. "Morphologically-Analyzed and Syntactically-Annotated Quran Dataset." *Data in Brief* 58 (February 2025): 111211. doi:10.1016/j.dib.2024.111211.

- Shormani, Mohammed Q., and Mohammad A. N. Alenezi. "Arabic Nominalization as Form of Pragmatic Actions in Digital Discourse: A Corpus-Based Study." *Corpus Pragmatics* 10, no. 1 (December 2026): 35. doi:10.1007/s41701-026-00237-5.
- Smadi, Alaa' Mohammad. "A Corpus-Based Analysis of the Lexical and Grammatical Properties of Arabic Abstracts in Social Sciences Articles." *International Journal of Linguistics* 12, no. 6 (November 24, 2020). doi:10.5296/ijl.v12i6.17849.
- Syahrullah, Syahrullah, Wildan Imaduddin Muhammad, Eva Nugraha, and Aulia Raudhatul Jannah. "Corpus Coranicum and Digital Philology: A Methodological Model for Advancing Qur'anic Manuscript Studies in Indonesia." *Khazanah Theologia* 6, no. 3 (December 29, 2024): 201–22. doi:10.15575/kt.v6i3.45610.