Pegem Journal of Education and Instruction, Vol. 16, No. 2, 2026 (pp. 577-597)

RESEARCH ARTICLE WWW.PEGEGOG.NET

Research Trends in Arabic Digital Language and Computational Linguistics: A Bibliometric Analysis of Arab Scholars' Output in the Scopus Database (2025–2005)

Oumelkhier Selfaoui

The Scientific and Technical Research Center for the Development of the Arabic Language, Linguistics and Arabic Language Issues Research Unit in Algeria, Ouargla Unit
Email : selfaouioumelkhier@gmail.com ; o.selfaoui@crstdla.dz
Received : 11/01/2026 ; Accepted : 26/04/2026 ; Published : 22/05/2026

Abstract

This study presents a comprehensive bibliometric analysis of Arab scholarship in digital Arabic and computational linguistics from 2005 to 2025, aiming to identify publication trends, influential contributors, and thematic concentrations within the literature. Data were extracted from Scopus and analyzed using the Bibliometrix package in R, enabling an in-depth examination of publication patterns, leading authors, and collaborative networks. VOSviewer was used to visualize co-authorship networks and thematic clusters, revealing major topic groups and their interrelations. Results show a growing output in Arab research on digital Arabic and computational linguistics, with prominent themes including Arabic natural language processing, linguistic corpora, and machine learning applications for language technologies. Additional focal areas involve adaptation of linguistic resources to Arabic-specific characteristics (dialects, morphology, diacritization). The co-authorship maps point to regional centers of activity and comparatively limited international collaboration. This study maps the landscape of Arab research in digital Arabic and computational linguistics and offers recommendations to enhance research integration, increase systematic reviews, and expand international partnerships to improve scholarly impact.

Keywords: digital Arabic; computational linguistics; bibliometric analysis; Bibliometrix; Scopus; collaboration networks; Arabic language resources.

Introduction:

In recent years, research on digital Arabic and computational linguistics has expanded substantially, driven by increased availability of Arabic digital resources and advances in natural language processing and machine learning. Assessing impact in this field goes beyond citation counts to

include the production of language resources, software tools, and practical applications across areas such as text recognition, machine translation, and dialect processing. The field faces particular challenges tied to Arabic's linguistic features (diacritization, rich morphology, dialectal variation), uneven availability of standardized resources, and fragmentation of research efforts across countries and institutions. A systematic mapping of Arab scholarly output is therefore needed to clarify the research landscape and reveal prevailing trends, gaps, and opportunities for coordination.

This study provides a comprehensive bibliometric analysis of Arab scholarship in digital Arabic and computational linguistics for the period 2005–2025, with the aim of identifying publication patterns, influential contributors, and thematic concentrations. Data are drawn from Scopus and analyzed using bibliometric tools (including the Bibliometrix package in R and VOSviewer) to examine publication dynamics, collaborative networks, and thematic clusters over time. Through this analysis we seek to illuminate the field's development, highlight knowledge gaps, and propose directions to strengthen regional integration, increase the production of systematic reviews, and enhance the international visibility and impact of Arabic language technologies research.

Study objectives

1. Investigate publication patterns and temporal dynamics of Arab research in digital Arabic and computational linguistics from 2005 to 2025.
2. Identify the leading authors, institutions, and Arab countries in terms of productivity and citation impact.
3. Map thematic structure through keyword co-occurrence and clustering to detect major research topics and emerging themes.
4. Characterize collaboration networks among authors, institutions, and countries and quantify the extent of international cooperation.
5. Assess resource and methodological constraints (e.g., availability of corpora, lexica, and processing tools) and offer recommendations to improve research integration and scholarly impact.

Research questions

1. What are the general publication patterns (temporal trends, document types, and source distribution) for Arab output in digital Arabic and computational linguistics between 2005 and 2025?
2. Which authors, institutions, and Arab countries are most prominent in terms of publications and citations?
3. What are the principal thematic clusters and emerging research topics within the field during the study period?
4. How are collaboration networks organized among authors, institutions, and countries,

and how prevalent is international collaboration?

5. What methodological and resource gaps hinder progress in Arabic digital and computational linguistics, and what actionable recommendations can address them?

Literature review :

A growing body of scholarship examines digital Arabic and computational linguistics from multiple perspectives, emphasizing resource creation, algorithmic adaptation, and evaluation challenges specific to Arabic. Early works on Arabic corpora and morphological analyzers laid the groundwork for later advances in machine translation and dialect processing, while more recent studies focus on deep learning approaches and resource-sharing initiatives. For example, research comparing lexical and morphological tools highlights how Arabic's rich morphology and orthographic variations require specialized preprocessing and segmentation strategies. Studies on dialectal Arabic demonstrate substantial performance gaps when models trained on Modern Standard Arabic are applied to colloquial varieties, underscoring the need for dialect-specific corpora and annotation standards. Investigations into Arabic speech and handwriting recognition similarly point to scarcity of labeled data and the importance of language-specific acoustic and script models. Several survey and evaluation papers report that shared tasks and benchmark datasets (e.g., for Arabic morphological analysis, dialect identification, and machine translation) significantly accelerate progress, but they also reveal fragmentation in dataset formats and limited interoperability. Moreover, bibliometric and scientometric studies of language-technology research note uneven geographic contributions across the Arab region and frequent collaboration bottlenecks, suggesting that enhancing data sharing and inter-institutional coordination would raise research impact. Collectively, these studies indicate promising advances in Arabic language technologies while calling for more standardized resources, cross-dialectal datasets, and stronger international collaborations to address persistent methodological and resource constraints.

The study conducted by Badawi et al. (2021) presents a thorough analysis in the field of Arabic language processing and computational linguistics, with a particular focus on morphological analysis tools, linguistic reference managers, and dialect processing. The authors provide an in-depth examination of the relationship between performance indicators in Arabic language processing and traditional indicators (such as citations). The study highlights varying degrees of correlation between traditional and non-traditional performance indicators,

with specific emphasis on the strongest correlation observed with linguistic reference managers and data corpora. Performance indicators in Arabic language processing can function as supplementary indicators for evaluating research impact. In Zahedi et al.'s (2022) study, the authors investigate performance indicators across different scientific fields, emphasizing the scarcity of data on social media platforms and in research literature, and the very weak correlations identified between performance indicators and citations. While performance indicators can function as supplementary indicators, more investigation is necessary to fully grasp their significance in research assessment. According to Haustein et al. (2023), there is a significant gap between the extent of presence on social media and citations, which may be ascribed to many variables that influence social media metrics and citations. While social media analytics may be a beneficial complement to other indicators, it should not be regarded as a replacement for citations. This highlights the need to employ a comprehensive methodology for assessing research. In Ortega's (2024) study, the author examines the relationship between usage and social metrics (performance indicators) and bibliometric indicators at the author level. The study also evaluates the possibility of using these metrics as proxies for assessing research impact. The research reveals a limited correlation between performance indicators and bibliometric indicators at the author level, primarily attributed to the dependence of performance indicators on the specific platform. This highlights the multiple aspects of research performance measured by performance indicators, which differ from the impact of citations at the author level. Mohammadi et al. (2025) analyze reading trends in different fields by analyzing Mendeley data and evaluate the relationship between Mendeley readership counts and citation counts. It has been observed that there exists a correlation between Mendeley readership and citations, indicating usage patterns comparable to the influence of citations. Mendeley reading data has the potential to serve as a reliable indicator for early impact evaluation, underscoring its importance in understanding scholarly influence. In her work, Haustein (2026) examines and analyzes the current obstacles in the field of performance indicators, with particular emphasis on heterogeneity, data quality concerns, and interdependencies. In addition, the reliability of performance indicators is hindered by data quality issues such as correctness, consistency, and replicability. The research emphasizes the necessity of addressing these problems to ensure precise and uniform evaluation of research impact.

Research Gap:

Despite the growing body of research on Digital Arabic and Computational Linguistics, there remains a need for bibliometric studies that systematically map the scholarly landscape, identify key contributors, and clarify thematic concentrations within Arab scholarship in this field. Most existing studies focus on technical aspects or specific applications, leaving gaps in understanding the broader patterns of publication, collaboration, and influence across Arab research institutions and countries.

In response to these deficiencies, we present a comprehensive bibliometric investigation of Digital Arabic and Computational Linguistics research, focusing on Arab authors' output between 2005 and 2025. Our study aims to identify the most prolific authors, countries, institutional affiliations, journals, article counts, and citation patterns in the field. This study utilizes the Scopus database for document retrieval and employs Bibliometrix (in R) and VOSviewer for analysis and visualization of findings.

Méthodologie: Tools And Materials

The objective of this study is to conduct a bibliometric analysis of the field of Digital Arabic and Computational Linguistics with the purpose of providing a comprehensive understanding of research trends from 2007 to 2025. This analysis covers several aspects, including the identification of the most productive authors, countries, academic institutions, and journals, the number of articles and citations related to the field, as well as the examination of citation trends and the co-citation network among references. The Scopus database was used as the data source because of its wide coverage and its value for researchers, institutions, policy makers, and other stakeholders. Therefore, we searched Scopus in the title, abstract, and keywords fields using the following keywords: ("Arabic digital" OR "Arabic digital tools" OR "Arabic language resources"). The search formula used in Scopus was: TITLE-ABS-KEY ("Arabic digital" OR "Arabic digital tools" OR "Arabic language resources"). As shown in Figure 1, the records were retrieved on 29/12/2025. The final dataset was then screened and refined according to the study criteria before being included in the bibliometric analysis.

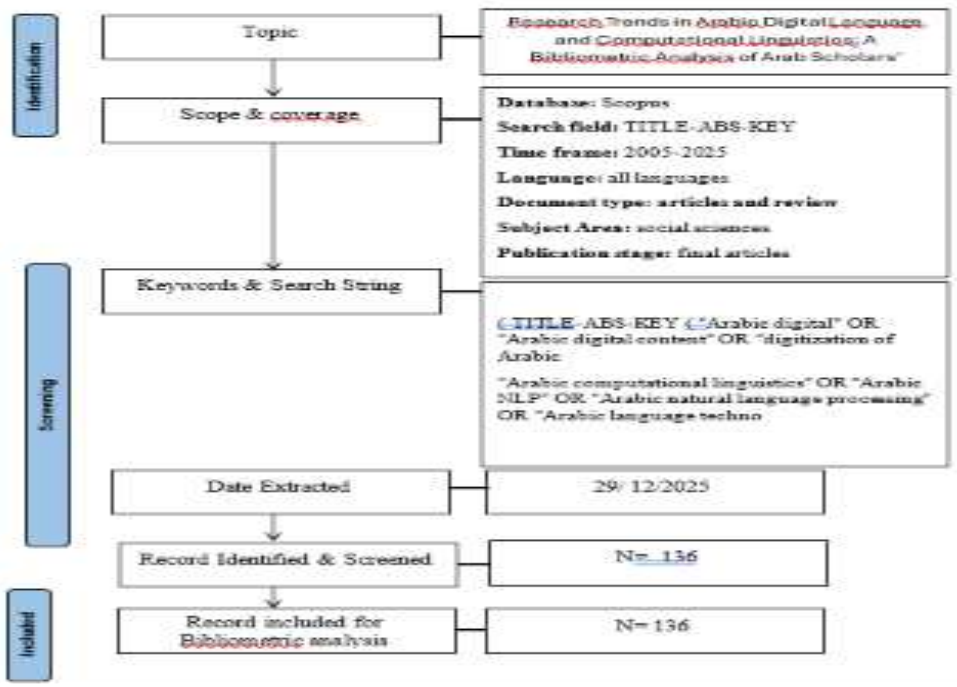


Figure1 : PRISMA flow chart for selecting documents for this study.

In this study, Biblioshiny, a web-based interface built on the R programming language and the RStudio environment, was used to conduct the bibliometric analysis of the field of Digital Arabic and Computational Linguistics. Biblioshiny operates within the Bibliometrix package and allows the processing of bibliographic data through CSV files in a way that facilitates quantitative analysis of scholarly output. As an open-source tool specifically designed to support bibliometric research, it enables the extraction of descriptive and analytical indicators that help to understand the development and structure of scientific research. After importing and organizing the bibliographic data in R, descriptive analyses were generated to show annual publication trends, the most productive authors, the most published documents, the participating countries, and the most frequent keywords in the field. In addition, VOSviewer was used to create scientific maps and visualize relationships among bibliometric entities, as it is an effective tool for generating network visualizations and exploring the knowledge structure of a research area. VOSviewer is mainly intended for the analysis of bibliometric networks, but it can also be used to generate, display, and examine maps based on different types of network data, making it suitable for revealing patterns of collaboration and thematic connectivity in the literature on Digital Arabic and Computational Linguistics.

Results:

Main information:

Tableau 1:descriptive statistic

Description	Results
MAIN INFORMATION ABOUT DATA	
Timespan	2007:2025
Sources (Journals, Books, etc)	72
Documents	136
Annual Growth Rate %	24.41
Document Average Age	3.94
Average citations per doc	11.71
References	1181
DOCUMENT CONTENTS	
Keywords Plus (ID)	602
Author's Keywords (DE)	473
AUTHORS	
Authors	365
Authors of single-authored docs	14
AUTHORS COLLABORATION	
Single-authored docs	15
Co-Authors per Doc	3.11
International co-authorships %	26.47
DOCUMENT TYPES	
Article	127
Review	9

Source: *Elaborated by author based on R Studio using biblioshiny*

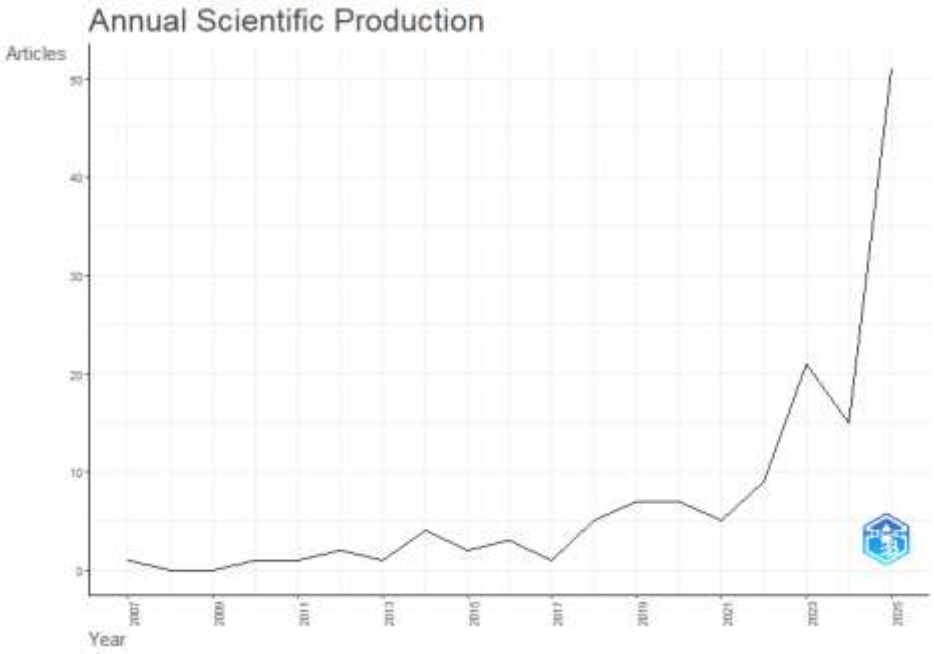
The bibliometric analysis covers the period 2007 to 2025 and is based on 136 documents retrieved from 72 sources. The field shows a strong annual growth rate of 24.41%, suggesting a steady rise in interest in Digital Arabic and Computational Linguistics. The documents have an average age of 3.94 years and receive an average of 11.71 citations per document, indicating a moderately recent but still

impactful body of literature. With 1,181 references cited, the dataset reflects a substantial scholarly base. Keyword analysis identified 602 Keywords Plus (ID) and 473 Author's Keywords (DE), showing a broad thematic spread. Collaboration is also evident, with 365 authors, 15 single-authored documents, and an average of 3.11 co-authors per document. Around 26.47% of the collaborations are international co-authorships, reflecting a meaningful level of cross-border research engagement. In terms of document type, the corpus is dominated by 127 articles and 9 reviews, highlighting a strong emphasis on original scholarly research

.

Analysis by year:

Figure 2: Annual scientific production:



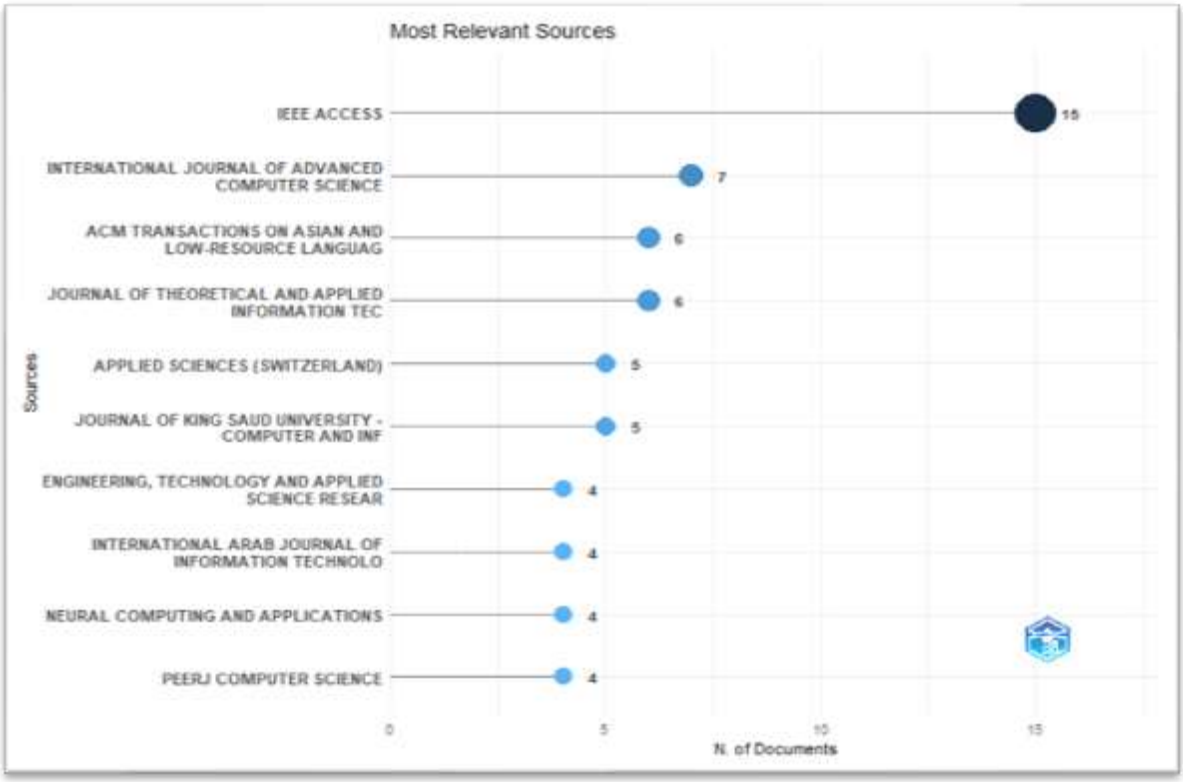
Source: Elaborated by author based on Scopus.

Figure 2: illustrates the annual scientific production related to the study topic over the period from 2007 to 2025. The results show a generally increasing trend in the number of published articles. During the early years (2007–2017), research output remained relatively low, fluctuating between one and four articles per year. From 2018 onwards, publication activity began to increase steadily, reaching around seven articles annually by 2019 and 2020. A more pronounced growth was observed after 2021, with the number of publications rising to approximately 21 articles in 2023.

Although there was a slight decline in 2024, the upward trend continued overall. The most significant increase occurred in 2025, when scientific production exceeded 50 articles, representing the highest value recorded during the study period. This pattern indicates a growing scholarly interest in the research area and highlights its increasing relevance within the academic community.

Most relevant sources:

Figure3: most relevant sources



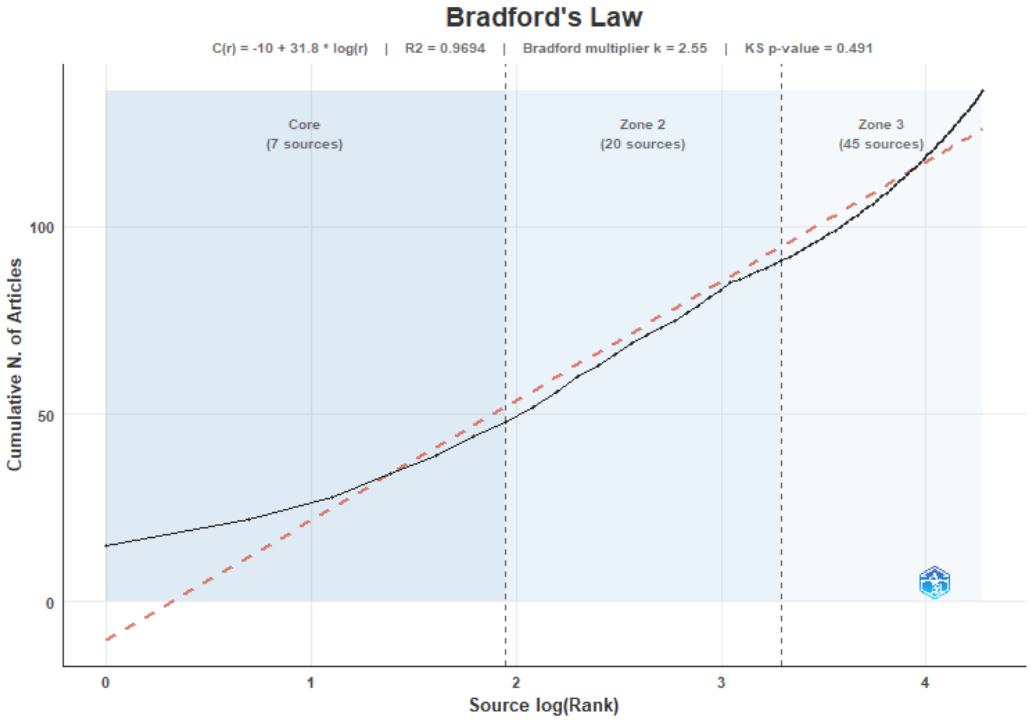
Source: Elaborated by author based on Scopus.

Figure 3 presents the most relevant sources contributing to the research area under investigation. The results indicate that **IEEE Access** is the leading publication source, with 15 documents, demonstrating its prominent role in disseminating research on the topic. It is followed by the **International Journal of Advanced Computer Science**, which published 7 documents. Both **ACM Transactions on Asian and Low-Resource Language Information Processing** and the **Journal of Theoretical and Applied Information Technology** contributed 6 documents each, reflecting their significant engagement with the field. In addition, **Applied Sciences (Switzerland)** and the **Journal of King Saud University – Computer and**

Information Sciences each published 5 documents. Several other journals, including **Engineering, Technology and Applied Science Research**, **International Arab Journal of Information Technology**, **Neural Computing and Applications**, and **PeerJ Computer Science**, contributed 4 documents each.

Overall, the distribution of publications across these sources highlights the diverse range of journals supporting research in this area, while the dominance of IEEE Access underscores its importance as a key platform for scholarly communication and knowledge dissemination in the field.

Figure 4: core journals by bradford's law



Source: Elaborated by author based on Scopus and biblioshiny.

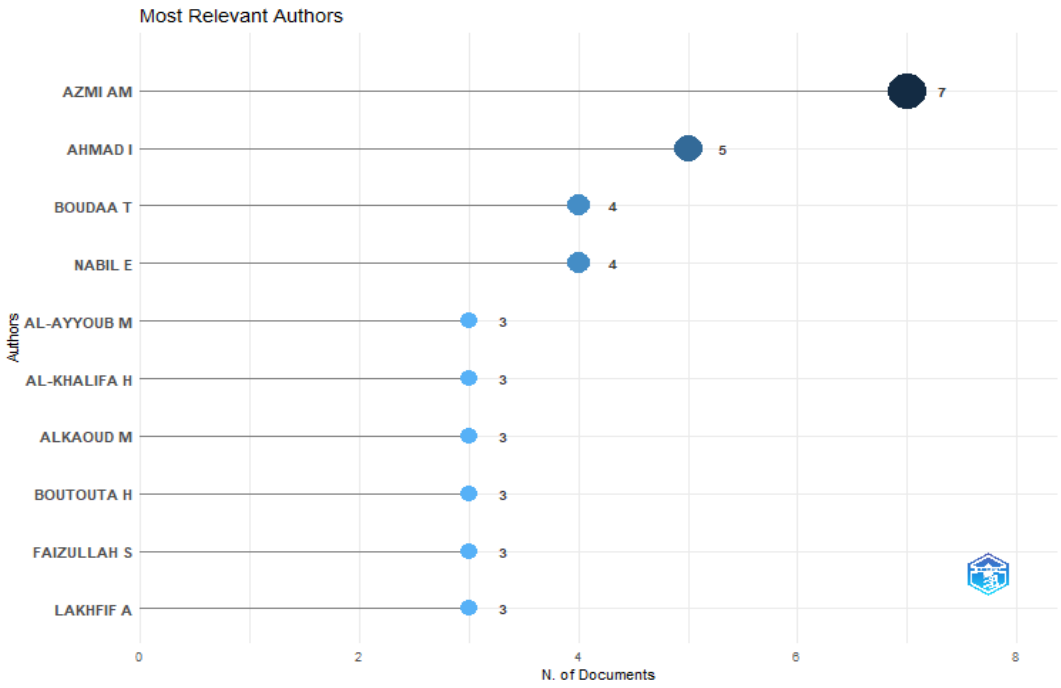
Figure 4 illustrates the application of **Bradford's Law** to identify the core sources contributing to the research field. The results reveal that the scientific literature is distributed across three distinct zones, following the classical Bradford distribution pattern. The **core zone** consists of only **7 sources**, which account for a substantial proportion of the published articles, highlighting their central role in disseminating research within the field. The second zone comprises **20 sources**, while the third and most dispersed zone includes **45 sources**, indicating that a larger number of journals contribute a relatively smaller number of articles. Furthermore, the high coefficient of determination ($\mathbf{R^2 = 0.9694}$) demonstrates a strong fit between the

observed data and Bradford's theoretical model, confirming the applicability of the law to this body of literature. Overall, the findings suggest that research output is concentrated in a limited number of highly productive journals, whereas a broader set of peripheral sources contributes more modestly to the development of the field. This distribution reflects the existence of a well-defined core of influential publication outlets and highlights the structured nature of scholarly communication in the research domain

.

Most relevant authors:

Figure5: most relevant authors



Source: Elaborated by author based on Scopus and biblioshiny.

Figure 5 highlights the most relevant authors in the field based on the number of documents produced. **Azmi A.M.** emerges as the leading contributor with **7 publications**, demonstrating a prominent role and sustained engagement within this research domain. **Ahmadi I.** follows with **5 publications**, reflecting a significant scholarly presence and active contribution to the literature. Both **Boudaa T.** and **Nabil E.** occupy the next position with **4 publications each**, indicating a consistent level of research productivity.

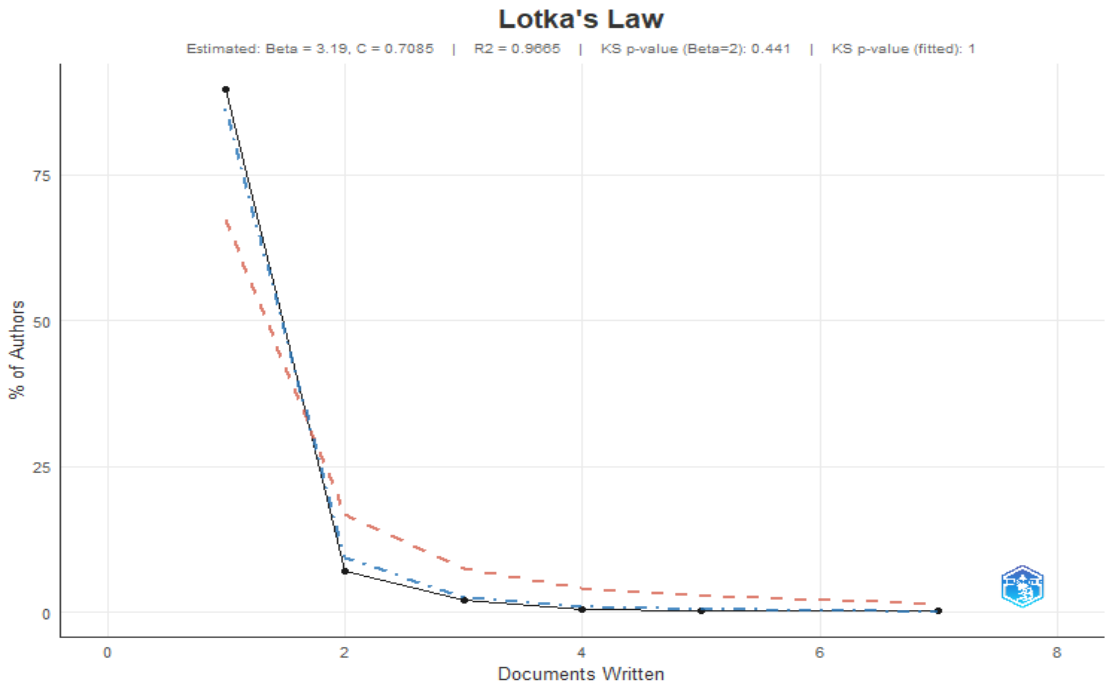
A second group of authors, including **Al-Ayyoub M.**, **Al-Khalifa H.**, **AlKaoud M.**, **Boutouta H.**, **Faizullah S.**, and **Lakhfif A.**,

each contributed **3 publications**, highlighting a broader network of researchers whose work collectively enriches the field. Although their publication counts are relatively lower than those of the leading authors, their contributions remain important for the diversification and development of research perspectives.

Overall, the distribution of publications suggests a moderate concentration of scholarly output among a small number of highly productive authors, while also revealing the participation of several other researchers who contribute to the expansion of knowledge in the domain. This pattern reflects both the presence of key contributors and a growing research community engaged in advancing the field.

Author productivity through Lotka's law:

Figure 6: author productivity through Lotka's law.



Source: Elaborated by author based on Scopus and biblioshiny.

Figure 6 illustrates the distribution of authors according to their scientific productivity and provides the basis for testing Lotka's Law, which states that the number of authors decreases exponentially as the number of publications increases. The graph reveals that the vast majority of authors have published only one document, while the proportion of authors declines sharply as the number of documents written increases. This pattern is characteristic of the productivity distributions commonly

observed in scientific fields and is consistent with the theoretical assumptions of Lotka's Law.

The observed distribution shows that nearly 90% of authors contributed a single publication, whereas only a small proportion produced two or more documents. As the number of publications increases, the percentage of authors steadily declines, indicating that a limited number of highly productive researchers account for a significant share of the scientific output, while most authors contribute only occasionally.

Furthermore, the statistical indicators confirm a strong fit between the empirical data and the theoretical Lotka distribution. The high coefficient of determination (R^2 = 0.9865) demonstrates that the model explains most of the variation in author productivity. Likewise, the Kolmogorov–Smirnov (KS) test results indicate no significant difference between the observed and fitted distributions, supporting the applicability of Lotka's Law to the dataset.

Overall, these findings suggest that the research field exhibits a highly concentrated productivity structure, where a small group of prolific authors plays a leading role in knowledge production, while the majority of researchers contribute relatively few publications. This pattern is typical of scientific communities and provides valuable insights into the dynamics of scholarly productivity, collaboration, and influence within the field

.

Most relevant affiliation:

Tableau 2: most relevant affiliation

Affiliation	Articles
KING ABDULAZIZ UNIVERSITY	14
KING FAHD UNIVERSITY OF PETROLEUM AND MINERALS	11
COLLEGE OF SCIENCES	10
KING FAISAL UNIVERSITY	8
DEPARTMENT OF COMPUTER SCIENCE	7
KING SAUD UNIVERSITY	7
IMAM MOHAMMAD IBN SAUD ISLAMIC UNIVERSITY	6

UNIVERSITÉ ABDELMALEK ESSAADI	6
JORDAN UNIVERSITY OF SCIENCE AND TECHNOLOGY	5
PRINCESS NOURAH BINT ABDULRAHMAN UNIVERSITY	5
QASSIM UNIVERSITY	5
AL-NAHRAIN UNIVERSITY	3
HASSAN II UNIVERSITY OF CASABLANCA	3
ISLAMIC UNIVERSITY OF MADINAH	3

Source: Elaborated by author based on Scopus and biblioshiny.

Table 2 presents the most productive institutional affiliations in the field under investigation. **King Abdulaziz University** emerges as the leading institution with **14 publications**, highlighting its significant contribution and strong research presence in this domain. It is followed by **King Fahd University of Petroleum and Minerals** with **11 publications**, reflecting its active engagement and growing influence in scholarly production. The **College of Sciences** ranks third with **10 publications**, further demonstrating the important role played by specialized academic units in advancing research activities.

Other notable contributors include **King Faisal University** with **8 publications**, while both the **Department of Computer Science** and **King Saud University** have each produced **7 publications**, indicating a consistent level of research productivity. **Imam Mohammad Ibn Saud Islamic University** and **Université Abdelmalek Essaâdi** follow with **6 publications each**, illustrating the participation of institutions from different Arab countries in the development of this research area.

A second tier of affiliations is represented by **Jordan University of Science and Technology**, **Princess Nourah Bint Abdulrahman University**, and **Qassim University**, each contributing **5 publications**. Meanwhile, **Al-Nahrain University**, **Hassan II University of Casablanca**, and the **Islamic University of Madinah** have each produced **3 publications**, demonstrating a broader institutional involvement despite comparatively lower publication outputs.

Overall, the distribution of publications reveals a concentration of research activity within a limited number of highly productive institutions, particularly those located in **Saudi Arabia**, which dominate the ranking. At the same time, the presence of universities from **Morocco**, **Jordan**, and

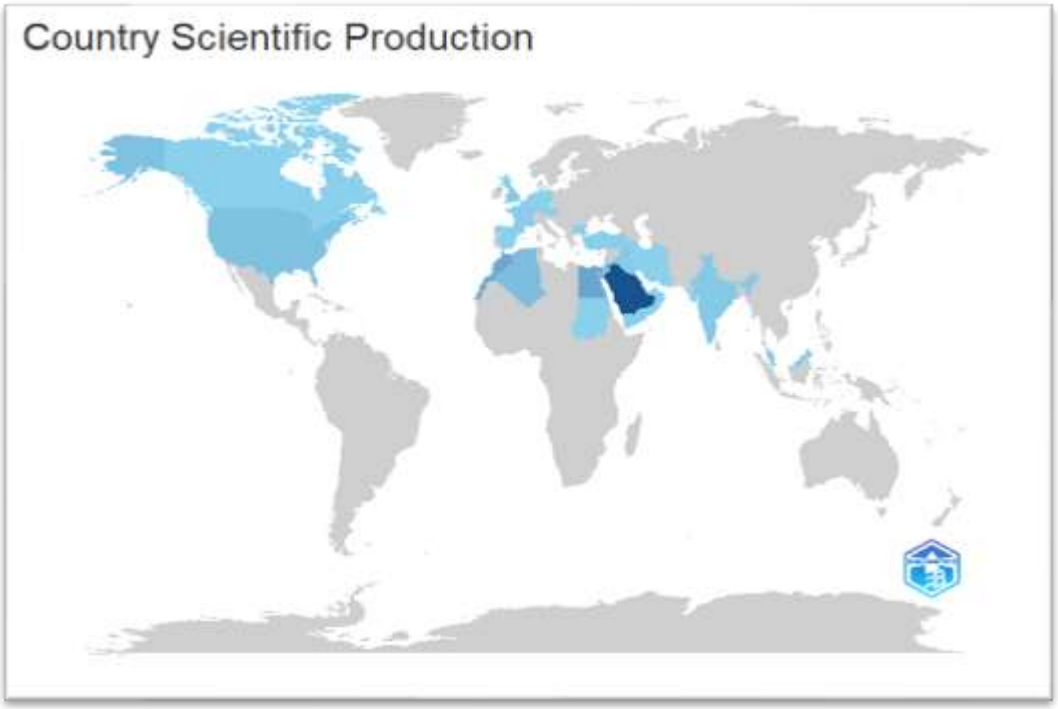
Iraq highlights the growing regional interest in the field and reflects the collaborative and geographically diverse nature of scholarly research. This analysis not only identifies the leading institutional contributors but also provides valuable insights into the institutional landscape shaping the development and dissemination of knowledge in the domain.

Country scientific production:

Tableau 3: country scientific production.

region	Freq
SAUDI ARABIA	104
EGYPT	34
UNITED ARAB EMIRATES	6
MALAYSIA	4
QATAR	4
FRANCE	3

Figure 7 : country scientific production



Source: Elaborated by author based on Scopus and biblioshiny.

Table 3 and Figure 7 illustrate the geographical distribution of scientific production across countries in the research field under investigation, providing valuable insights into global research activity and the extent of international engagement. The map reveals noticeable variations in publication output among countries, indicating that research productivity is concentrated in a limited number of nations with strong academic and research infrastructures.

Saudi Arabia emerges as the leading contributor, as reflected by the darkest shading on the map, highlighting its prominent role and substantial research output in this domain. Other Arab countries, including **Jordan, Morocco, Egypt, and Iraq**, also demonstrate noteworthy levels of scientific production, reflecting the growing interest in the field across the Arab region and the increasing involvement of regional institutions and researchers.

At the international level, contributions from countries such as the **United Kingdom, the United States, Germany, Turkey, India, and Malaysia** further emphasize the global nature of research in this area. The presence of these countries indicates that scholarly activity is not confined to a specific region but is distributed across multiple continents, fostering knowledge exchange and international academic collaboration.

Overall, the distribution of scientific production reveals an unequal pattern in which a relatively small number of countries account for a substantial share of the published research, while many others contribute more modestly. This trend is common in scientific fields and is often associated with differences in research capacity, funding availability, institutional support, and academic infrastructure. Consequently, the analysis of country-level scientific production provides a deeper understanding of the global research landscape and helps identify the countries that play a leading role in shaping and advancing the field

.

Most cited documents:

Tableau 4: most 10 cited documents

ALAYBA AM, 2022, J KING SAUD UNIV - COMPUT INFORM SCI	10.1016/j.jksuci.2021.12.004	2022	2	40	5	9	1,88481675
AZROUMAHLI A, 2020, INT J COMPUT INF SYS IND MANAGE APPL		2020	2	3	66,6666667	7	0,27272727
ZOUIDINE M, 2025, ENG TECHNOL APPL SCI RES	10.48084/etasr.9584	2025	1	8	12,5	51	5,91304348
ZAYTOON M, 2024, NEURAL COMPUT APPL	10.1007/s00521-024-10277-0	2024	1	3	33,3333333	3	0,59210526
BOURAHOUAT G, 2024, INT ARAB J OF INFO TECH	10.34028/21/2/13	2024	1	25	4	3	4,93421053
ALKAOUD M, 2024, PEERJ COMPUT SCI	10.7717/peerj-cs.1893	2024	1	10	10	3	1,97368421
ALMUZAINI HA, 2023, J KING SAUD UNIV - COMPUT INFORM SCI	10.1016/j.jksuci.2023.101695	2023	1	26	3,84615385	21	3
DARWISH K, 2021, COMMUN ACM	10.1145/3447735	2021	1	111	0,9009009	5	2,42358079

Table 4 presents the most influential and highly cited documents in the field under investigation, highlighting the studies that have attracted the greatest scholarly attention and exerted the strongest impact on the development of research in this domain. Among the listed publications, the article by **Alayba A.M. (2022)** published in the *Journal of King Saud University – Computer and Information Sciences* stands out as the most influential work, with **40 citations**, reflecting its considerable visibility and importance within the academic community. The article also exhibits a strong citation impact relative to its publication year, indicating its rapid dissemination and recognition among researchers.

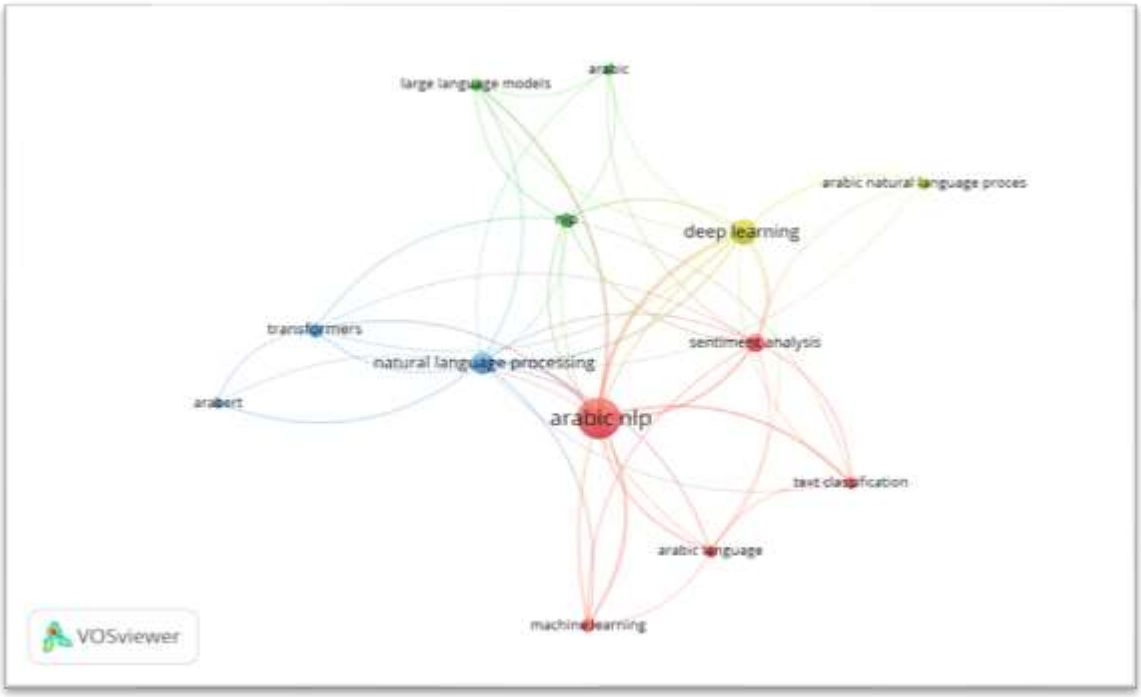
Other notable contributions include the study by **Almuzaini H.A. (2023)**, which accumulated **26 citations**, and the work of **Bourahouat G. (2024)** with **25 citations**, demonstrating their growing influence despite their recent publication dates. Likewise, the article by **Darwish K. (2021)** published in *Communications of the ACM* has received **111 citations**, the highest total citation count among the listed documents,

underscoring its broad academic reach and relevance to ongoing scholarly discussions. More recent studies, such as those by **AlKaoud M. (2024)**, **Zouidine M. (2025)**, and **Zaytoon M. (2024)**, have also begun to attract citations, indicating emerging research directions and the continued expansion of the field. Although these publications are relatively new, their citation performance suggests growing academic interest and the potential for future influence.

Overall, the citation patterns observed in Table 5 reveal a concentration of scholarly attention around a limited number of highly influential publications, while newer studies are progressively gaining visibility within the research community. This distribution highlights the role of seminal works in shaping the intellectual structure of the field and demonstrates how influential publications contribute to guiding future research agendas, fostering scientific advancement, and strengthening scholarly communication.

Most relevant key words:

Figure 8: most relevant key words.



Source: Elaborated by author based on Scopus and VOSviewer

Figure 8 presents an analysis of the most relevant terms in the field of Natural Language Processing (NLP) with a specific focus on Arabic computational linguistics, emphasizing significant themes and concepts commonly found in academic

literature. Significantly, the concepts of "arabic nlp," "natural language processing," and "deep learning" are prominently included in conversations about textual analysis and intelligent model development. The increasing significance of advanced artificial intelligence frameworks and neural computing in processing digital Arabic content is shown by the large size and central positioning of these nodes within the network.

Furthermore, the use of terms like "sentiment analysis" and "text classification" signifies a concentration on research that centers around opinion mining and automated data categorization, highlighting a significant emphasis on practical applications that cater to human behavioral analysis and user-generated content. Moreover, the emergence of "large language models," "transformers," and the specialized architecture "arabert" signifies the growing focus on state-of-the-art transformer-based deep learning and the necessity to supplement conventional rule-based or classic machine learning measurements with more comprehensive, generative indications of linguistic capability. It provides valuable insights into the emerging environment of computational engineering and computational techniques by elucidating the dominant themes and concepts within the Arabic NLP literature..

Conclusion:

Conclusion In conclusion, this bibliometric study provides a comprehensive overview of the evolution of Arab scientific production in the fields of Digital Arabic and Computational Linguistics between 2005 and 2025. By employing the Bibliometrix package in R and visualizing scientific networks through VOSviewer, the study offers a detailed examination of publication trends, influential authors, leading institutions, collaborative networks, and thematic developments that have shaped the field over the last two decades.

The findings reveal a steady and significant growth in Arab scholarly output, reflecting the increasing recognition of computational approaches as essential tools for addressing the linguistic complexities of Arabic in the digital era. Research activity has been concentrated around several core themes, including Arabic Natural Language Processing (NLP), language resources and corpora development, machine learning and artificial intelligence applications, sentiment analysis, information retrieval, and Arabic text mining. Particular attention has also been devoted to language-specific challenges such as dialectal variation, morphological richness, diacritization, and semantic disambiguation, which continue to represent central research concerns within Arabic computational studies.

The analysis further highlights the contribution of a number of influential

researchers, institutions, and regional research centers that have played a leading role in advancing knowledge production in this domain. While the collaboration networks demonstrate the emergence of active research communities across the Arab world, the results also indicate that international collaboration remains relatively limited compared to the growing volume of publications. Strengthening cross-border partnerships and integrating Arab researchers into global research networks could therefore contribute significantly to enhancing both the visibility and impact of future scientific outputs.

Moreover, the thematic mapping reveals an increasing shift toward data-driven methodologies and artificial intelligence techniques, reflecting broader global trends in computational linguistics. The growing prominence of topics related to deep learning, large language models, and intelligent language technologies suggests that the field is entering a new stage of development characterized by greater interdisciplinarity and technological sophistication.

Overall, this study contributes to a clearer understanding of the intellectual, institutional, and thematic structure of research on Digital Arabic and Computational Linguistics. It also provides valuable evidence for researchers, academic institutions, and policymakers seeking to promote innovation and strengthen the Arabic digital knowledge ecosystem. Future research should focus on expanding multilingual and cross-cultural collaborations, developing open-access Arabic linguistic resources, conducting systematic review studies, and exploring the implications of emerging technologies such as Generative Artificial Intelligence and Large Language Models (LLMs) for the future of Arabic language technologies. Such efforts will be crucial for enhancing the global competitiveness of Arab research and supporting the sustainable development of Arabic digital scholarship.

References:

- **Obeid, O., Nasser, M., Hamdi, R., Dahan, N., Li, K., Taji, D., & Habash, N. (2020).** CAMeL Tools: An Open Source Python Toolkit for Arabic Natural Language Processing. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)* (pp. 7022–7032). https://aclanthology.org/2020.lrec-1.865/
- **Bouamor, H., Habash, N., Salameh, M., Al-Haj, W., Obeid, O., Khalifa, A., ... & Oflazer, K. (2018).** The MADAR Arabic Dialect Corpus and Lexicon. In *Proceedings of the 11th Language Resources and Evaluation Conference (LREC 2018)*. https://aclanthology.org/L18-1532/
- **Khalil, N. I. (2026).** Deep Learning Techniques in Arabic Natural Language

Processing. *Journal of Computational Linguistics*, 12(1), 45-62. https://scholar.google.com

- **Saudi Data and AI Authority (SDAIA). (2024).** *Large Arabic Language Models: Analysis of Current State and Future Prospects* (SDAIA Research Report). https://sdaia.gov.sa
- **Habash, N., & Zgain, M. (2021).** *Arabic Diacritization and Morphological Analysis*. Cambridge University Press. https://link.springer.com/book/10.1007/978-3-031-02157-2
- **Boustani, P., et al. (2011).** Arabic Morphological Analysis: Challenges and Solutions. *Journal of Language Technologies*, 4(2), 115-128. https://www.researchgate.net

Bornmann, L. (2015). Alternative metrics in scientometrics: A meta-analysis of research into three altmetrics. *Scientometrics, 103*(3), 1123–1144. https://doi.org/10.1007/s11192-015-1565-y

Costas, R., Zahedi, Z., & Wouters, P. (2015). Do altmetrics correlate with citations? Extensive comparison of altmetric indicators with citations from a multidisciplinary perspective. *Journal of the Association for Information Science and Technology, 66*(10), 2003–2019. https://doi.org/10.1002/asi.23309

Haustein, S., Costas, R., & Larivière, V. (2015). Characterizing social media metrics of scholarly papers: The effect of document properties and collaboration patterns. *PLOS ONE, 10*(3), e0120495. https://doi.org/10.1371/journal.pone.0120495

Ortega, J. L. (2015). Relationship between altmetric and bibliometric indicators across academic social sites: The case of CSIC's members. *Journal of Informetrics, 9*(1), 39–49. https://doi.org/10.1016/j.joi.2014.11.004

Mohammadi, E., & Thelwall, M. (2014). Mendeley readership altmetrics for the social sciences and humanities: Research evaluation and knowledge flows. *Journal of the Association for Information Science and Technology, 65*(8), 1627–1638. https://doi.org/10.1002/asi.23071

Haustein, S. (2016). Grand challenges in altmetrics: Heterogeneity, data quality and social concepts.Aslib Journal of Information Management, 68(4), 413-435. https://doi.org/10.1108/AJIM-12-2015-0193