Tadris Al-'Arabiyyah: Jurnal Pendidikan Bahasa Arab dan Kebahasaaraban 5 (1): 43-53 (2026)
DOI: https://doi.org/10.15575/ta.v5i1.52011

Tadris Al-'Arabiyyah
Jurnal Pendidikan Bahasa Arab dan Kebahasaaraban
E-ISSN: 2963-6876 // Vol. 5 No. 1 | 43-53
https://journal.uinsgd.ac.id/index.php/ta/index

Received: Juni, 2025. Accepted: April, 2026. Published: April, 2026.

THE UTILIZATION OF SKETCH ENGINE IN CONCORDANCE STUDIES: A CORPUS LINGUISTICS PERSPECTIVE

Faiza Nur Khalida[1], Mochamad Mui'zzudin[2], Ahmad Suhaili[3]
[1,2,3] Universitas Islam Negeri Sultan Maulana Hasanudin Banten
Corresponding E-mail: faizanurkhalida19@gmail.com

ABSTRACT

Corpus linguistics has developed rapidly in recent years; however, its application in Arabic language studies in Indonesia remains limited, particularly in examining the semantic distinctions of key educational terms. This study aims to investigate the distribution, collocational patterns, and semantic functions of three central terms تعليم (*taʿlīm*), تدريس (*tadrīs*), and تربية (*tarbiyah*) in a modern Arabic corpus, as well as their implications for conceptual meaning and language pedagogy. The novelty of this study lies in integrating corpus-based analysis of authentic linguistic data with semantic interpretation within a pedagogical framework. This research employs a qualitative descriptive approach with a corpus-based discourse analysis design, using data derived from the arTenTen24 corpus through Sketch Engine. The analysis applies frequency, concordance, and collocation techniques to identify lexical distributions and relationships. The findings reveal that تعليم is the most dominant term, primarily associated with educational systems and policy contexts; تدريس relates to instructional practices in classroom settings; while تربية encompasses broader dimensions, including values, character formation, and institutional contexts. These distinctions indicate systematic differences in semantic functions that reflect the conceptual structure of modern Arabic educational discourse. Theoretically, this study contributes to corpus-based lexical-semantic research in Arabic, while pedagogically, it supports data-driven learning to improve learners' accuracy in using educational terminology.

Keywords: Arabic Educational Terminology, Corpus Linguistics, Sketch Engine

ABSTRAK

Linguistik korpus telah berkembang pesat dalam beberapa tahun terakhir; namun, penerapannya dalam kajian bahasa Arab di Indonesia masih terbatas, khususnya dalam mengkaji perbedaan semantik istilah-istilah pendidikan utama. Penelitian ini bertujuan untuk menganalisis distribusi, pola kolokasi, dan fungsi semantik dari tiga istilah utama—تعليم (*taʿlīm*), تدريس (*tadrīs*), dan تربية (*tarbiyah*)—dalam korpus bahasa Arab modern, serta implikasinya terhadap makna konseptual dan pedagogi bahasa. Kebaruan penelitian ini terletak pada integrasi analisis berbasis korpus terhadap data linguistik autentik dengan interpretasi semantik dalam kerangka pedagogis. Penelitian ini menggunakan pendekatan deskriptif kualitatif dengan desain analisis wacana berbasis korpus, dengan data yang diambil dari korpus arTenTen24 melalui Sketch Engine. Analisis dilakukan menggunakan teknik frekuensi, konkordansi, dan kolokasi untuk mengidentifikasi distribusi leksikal dan hubungan antarkata. Hasil penelitian menunjukkan bahwa تعليم merupakan istilah yang paling dominan dan terutama berkaitan dengan sistem pendidikan serta konteks kebijakan; تدريس berhubungan dengan praktik pengajaran di kelas; sedangkan تربية mencakup dimensi yang lebih luas, termasuk nilai, pembentukan karakter, dan konteks kelembagaan. Perbedaan ini menunjukkan adanya variasi sistematis dalam fungsi semantik yang mencerminkan struktur konseptual wacana pendidikan bahasa Arab modern. Secara teoretis, penelitian ini berkontribusi pada pengembangan kajian semantik leksikal berbasis korpus dalam bahasa Arab. Secara pedagogis, temuan ini mendukung penerapan pembelajaran berbasis data untuk meningkatkan ketepatan penggunaan istilah pendidikan oleh pembelajar.

Kata Kunci: Analisis Kolokasi, Linguistik Korpus, Terminologi Pendidikan Arab

INTRODUCTION

The rapid advancement of digital technology has significantly transformed linguistic research and language pedagogy, particularly through the emergence of data-driven approaches such as corpus linguistics (Boulton & Forti, 2025). Unlike traditional intuition-based linguistic analysis, corpus linguistics enables systematic investigation of authentic language use through large-scale textual data, allowing researchers to identify patterns of frequency, collocation, syntactic behavior, and contextual meaning based on empirical evidence rather than subjective interpretation (Crosthwaite & Baisa, 2023). In this regard, corpus analysis tools such as Sketch Engine have become essential in facilitating comprehensive linguistic analysis through automated features including concordance lines, word sketches, and collocation extraction (Kovačević, 2026).

Recent developments in corpus linguistics further emphasize its increasing role in language education and applied linguistics. Studies demonstrate that corpus-based approaches enhance lexical awareness, phraseology identification, and contextual interpretation in both first- and second-language learning environments (Ma & Mei, 2021). Moreover, corpus literacy has been shown to improve teachers' pedagogical competence by enabling the integration of authentic linguistic data into classroom instruction, while also strengthening learners' understanding of vocabulary usage and discourse structures (Al Fraidan & Alkuwaity, 2025). In Arabic language learning specifically, corpus-based analysis has been used to explore syntactic patterns, semantic prosody, and discourse structures in authentic texts, contributing to more contextualized language acquisition (Hizbullah et al., 2020).

Despite these advancements, the integration of corpus-based approaches into language learning, particularly Arabic language education, remains limited in many contexts, including Indonesia (Zaki, 2021). Arabic instruction is still predominantly characterized by traditional, rule-based methodologies that emphasize grammatical explanation rather than exposure to authentic linguistic data (Fitrianto, 2024; Guerza, 2023). This reflects not only a pedagogical gap but also a methodological limitation in how linguistic patterns are introduced and internalized in instructional settings. Consequently, learners often experience difficulties in understanding natural lexical combinations and contextual language use (Mawaddah et al., 2025).

Previous studies have demonstrated the potential of corpus tools in linguistic analysis and education. Sketch Engine has been widely recognized for its ability to identify lexical bundles, rhetorical structures, and contextual language patterns across genres (Yuliawati et al., 2021). Other studies highlight that corpus-based instruction improves learners' vocabulary retention, collocational competence, and grammatical awareness (Li et al., 2025). Sketch Engine has been widely recognized for its ability to identify lexical bundles, rhetorical structures, and contextual language patterns across genres, particularly in academic and pedagogical corpora (Elewa, 2025). Research has shown that corpus-based tools facilitate the extraction of recurrent multi-word units, enabling more precise identification of phraseological patterns in both first- and second-language contexts (Jurko, 2022). In addition, Sketch Engine's word sketch and concordance functionalities have been found to support deeper semantic and pragmatic analysis by revealing collocational behavior and contextual usage patterns that are difficult to capture through traditional methods (Ma et al., 2023). Furthermore, corpus-informed pedagogical studies highlight that integrating such tools into language instruction enhances learners' awareness of authentic language use, improves vocabulary acquisition, and strengthens discourse competence.

A critical review of recent literature reveals two main limitations. First, concordance analysis is often treated as a descriptive tool rather than a methodological framework for interpreting contextual meaning and lexical behavior in depth (McEnery & Hardie, 2011). Second, there is

limited research that operationalizes advanced corpus features, such as word sketches and automated collocation analysis, within an integrated analytical model, particularly in Arabic language learning contexts (Valdivieso & González, 2025). As a result, the potential of corpus tools to bridge linguistic analysis and pedagogical application has not been fully realized.

Furthermore, although studies such as Yuliawati et al. and others demonstrate the use of Sketch Engine for lexical bundle analysis, they largely remain descriptive and lack comparative or integrative methodological frameworks. Similarly, hybrid studies combining Sketch Engine and AntConc emphasize technical procedures rather than analytical synthesis (Curry & McEnery, 2025; Jurko, 2022). These limitations indicate the absence of a structured corpus-based analytical model that connects concordance, collocation, and word sketch into a unified interpretative framework.

This study proposes an integrative approach by positioning concordance analysis as a methodological framework for interpreting contextual meaning and lexical relations in Arabic language data. Its novelty lies in developing a corpus-based analytical model that integrates concordance analysis, collocation patterns, and word-sketch features to reveal the interaction among lexical patterns, context, and meaning.

The study aims to analyze word usage patterns and contextual variation, identify collocational relationships and lexical functions using integrated corpus tools, and examine the implications of corpus-driven analysis for Arabic language learning based on authentic linguistic data. Overall, it is expected to contribute to corpus-informed Arabic pedagogy by promoting a more empirical, data-driven, and context-sensitive approach, while also bridging the gap between theoretical linguistics and classroom practice through digital linguistic technologies.

METHOD

This study employs a descriptive qualitative design within a corpus-based discourse analysis framework to examine patterns of word usage, contextual meaning, and lexical relations in educational discourse. A corpus-based approach is adopted to analyze predetermined lexical items using empirical language data, enabling a systematic integration of linguistic theory and authentic language use (Musthafa & Hermawan, 2018). This approach is widely applied in contemporary applied linguistics research due to its ability to reduce subjectivity and provide evidence-based linguistic analysis (Crosthwaite & Baisa, 2023).

The data in this study were derived from two types of corpora. The first is the Indonesian Web Corpus (IndonesianWaC) available in Sketch Engine, which serves as a general reference corpus. The second is a specialized corpus compiled by the researcher, consisting of approximately 750,000 words. This specialized corpus includes academic journal articles, educational news texts, and policy or institutional documents related to education, published within the last ten years. The selection of texts was based on several criteria, namely relevance to educational discourse, linguistic completeness, and public accessibility (Sardinha, 2020). Before analysis, all texts were cleaned and standardized by removing duplicates and non-linguistic elements to ensure data quality and consistency. The use of both general and specialized corpora is intended to enhance data representativeness and enable comparison across different discourse contexts.

The size of the specialized corpus is considered sufficient to capture stable lexical and collocational patterns while maintaining domain specificity. In corpus linguistics, a medium-sized corpus (approximately 500,000 to 1,000,000 words) is generally adequate for identifying frequency distributions, collocation patterns, and contextual variations in language use (Pandža et al., 2020).

Data analysis was conducted using Sketch Engine as the primary analytical tool, widely used in corpus linguistics for processing large-scale language data and automatically identifying lexical and syntactic relations (Yuliawati et al., 2021). The analysis focused on three primary features: concordance, word sketch, and collocation analysis. The concordance feature was used to examine the context in which target words occur, while word sketch and collocation analysis were used to identify grammatical relations and statistically significant co-occurrence patterns.

This study employed a structured corpus-based procedure beginning with keyword identification aligned with the research focus. Relevant corpora were then accessed through Sketch Engine, followed by concordance extraction to examine contextual usage. Collocation and word sketch analyses were subsequently conducted to identify lexical patterns and semantic relationships.

Data analysis included frequency, concordance, collocation, and lexical-semantic interpretation within a corpus-based discourse framework. Frequency analysis identified distribution patterns, while concordance and collocation analyses revealed contextual and relational meanings across lexical and syntactic levels.

To ensure validity and reliability, the study applied source and method triangulation by comparing general and specialized corpora and cross-validating results across different Sketch Engine tools. Consistent query parameters and transparent procedures were used to enhance replicability. The research was conducted from 2025 to 2026, covering corpus compilation, analysis, and interpretation of educational discourse.

RESULTS AND DISCUSSION

Result

Corpus Linguistics, Concordance, and Corpus Tools: A Theoretical and Methodological Overview

Corpus linguistics is a data-driven approach in applied linguistics that analyzes large, structured collections of authentic language data using computational tools. A corpus is a principled and representative dataset of naturally occurring texts designed for linguistic analysis (Sinclair, 1991). This approach emphasizes authenticity, representativeness, and sufficient size and has shifted linguistic research from intuition-based analysis to empirical, reproducible methods, thereby enhancing methodological rigor (McEnery & Hardie, 2012).

Corpus linguistics enables the identification of recurrent lexical and grammatical patterns, particularly through collocation analysis, which reveals systematic co-occurrence and semantic preference. It is widely applied in lexicography, language teaching, translation studies, and discourse analysis, and it supports both theoretical and applied linguistic research (Biber et al., 1998).

Concordance is a central output of corpus analysis that displays occurrences of a word or phrase in context, commonly in KWIC format (Baker & McEnery, 2015). It allows researchers to examine usage patterns, meaning construction, and contextual behavior. Scholars such as Sinclair (1991), Hunston (2002), and Stubbs (2001) emphasize its role in revealing authentic language patterns that are not accessible through intuition. Concordance outputs may appear in KWIC, sentence, or paragraph formats and can be filtered by lemma, part of speech, or collocation.

Sketch Engine is an advanced corpus platform that enables large-scale concordance analysis with powerful filtering options. Its key features include word sketch (grammatical and collocational profiling), collocation analysis, n-gram extraction, and BootCaT for automated

corpus building. These tools allow systematic analysis of lexical behavior and discourse patterns across domains.

Corpus tools vary in complexity and functionality. Sketch Engine offers advanced, large-scale analysis but is subscription-based; AntConc is free and suitable for basic corpus analysis; manual analysis is limited, time-consuming, and less suitable for large datasets. Therefore, tool selection should align with research objectives, data scale, and methodological requirements.

Corpus-Based Analysis of the Semantic Distribution of تعليم, تدريس, and تربية in Arabic Educational Discourse

The corpus analysis reveals substantial differences in the frequency distribution of the three key terms—تعليم, تدريس, and تربية—indicating that they occupy distinct functional roles within Arabic educational discourse. Among these, **تعليم** emerges as the most dominant term, with 4,122,632 occurrences in the corpus, reflecting its central position in discussions related to educational systems, policies, and institutional frameworks. In comparison, تربية appears with a moderate frequency of 1,850,379 occurrences and is predominantly associated with contexts emphasizing values, moral development, and institutional identity. Meanwhile, تدريس is the least frequent term, with 520,717 occurrences, and is primarily used in relation to instructional processes and classroom-level practices. These distributional differences suggest that the three terms are not interchangeable, but rather function within distinct domains of educational discourse.

Further insights are obtained through concordance analysis, which highlights the contextual patterns of each term. The KWIC (Key Word in Context) lines demonstrate that **تعليم** frequently co-occurs with system-oriented expressions such as **التعليم العالي** and references to governmental or institutional bodies, indicating its strong association with macro-level discourse. In contrast, تدريس appears in contexts closely tied to pedagogical actions, often accompanied by phrases such as عملية التدريس and طرق التدريس, which emphasize instructional methods and classroom practices. The term **تربية**, on the other hand, is commonly used in contexts related to value formation and identity, as reflected in expressions such as **التربية الإسلامية** and **التربية الأخلاقية**, suggesting an orientation toward the institutional and ideological dimensions of education.

The collocation analysis further confirms these distinctions by revealing systematic patterns of lexical association. The term **تعليم** is strongly linked to institutional and structural collocates, including terms related to systems, policies, and organizations, which reinforces its macro-level orientation. In contrast, تدريس shows strong associations with procedural and action-oriented vocabulary, indicating its focus on the implementation of teaching practices. Meanwhile, تربية demonstrates consistent collocational patterns with value-laden and identity-related terms, highlighting its role in shaping moral and cultural dimensions within educational discourse. These collocational tendencies indicate that lexical meaning is not arbitrary, but emerges from repeated patterns of co-occurrence within authentic language use.

The word sketch analysis provides additional evidence by illustrating the grammatical behavior of each term. The findings show that تعليم frequently operates within constructions involving institutional actions such as development and improvement, further supporting its systemic role. In contrast, تدريس is typically associated with verbs that denote implementation and practice, reflecting its operational function in teaching contexts. The term تربية, meanwhile, often appears in nominal and descriptive constructions, which underscores its conceptual and value-oriented nature. These grammatical patterns align with the broader semantic distinctions observed in the frequency and collocation analyses.

In addition to total frequency, the analysis also identifies the dominant morphological forms of each term. The results show that the definite form with the prefix "ال" is the most frequently used form across all three terms. The detailed frequency distribution is presented in Table 1.

Table 1. Frequency Distribution and Dominant Forms

Term	Total Frequency	Dominant Form	Frequency (Form)	Percentage (%)
تعليم	4,122,632	التعليم	1,845,146	44,97%
تدريس	520,717	التدريس	273,663	52,56%
تربية	1,850,379	التربية	1,228,702	66,40%

The frequency distribution indicates that the term تعليمexhibits the highest overall occurrence in the corpus (4,122,632 instances), followed by تربية(1,850,379) and تدريس(520,717). This pattern suggests that تعليمfunctions as the most general and widely used educational term, reflecting its broad conceptual scope within Arabic educational discourse. In contrast, *تدريس* shows the lowest frequency, indicating a more specialized and context-specific usage, primarily associated with instructional practices. The intermediate frequency of تربيةreflects its dual orientation toward both institutional and value-based dimensions of education.

The analysis of dominant forms further reveals that each term is predominantly used in its definite form, with التعليمaccounting for 44.97% of occurrences, التدريسfor 52.56%, and التربية for 66.40%. The higher proportion of definite forms, particularly in التربية, indicates a strong association with formal, institutional, and conceptual discourse. Overall, these findings suggest that the three terms are not only differentiated by frequency but also by morphological realization, reinforcing their distinct functional roles within Arabic educational discourse.

Further analysis was conducted to examine collocation patterns surrounding each term. Collocation analysis reveals that each term is associated with different lexical environments, reflecting variation in contextual usage. The main collocation patterns are summarized in Table

Table 2. Collocational Patterns of Key Educational Terms in Arabic Corpus Data

Term	Left Collocations	Right Collocations
تعليم	العالي، في	التربية، في، وزارة
تدريس	في، (،)	هيئة، في، طرق
تربية	والتعليم، الوطنية	وزارة، وزير، كلية

The table presents collocational patterns of key educational terms in Arabic corpus data, namely تعليم(education/teaching), تدريس(instruction/teaching process), and تربية(education in a broader, nurturing sense). In the left collocations column, تعليمfrequently appears with words such as العالي(higher) and في(in), indicating its common use in the context of higher education or specific locations. Meanwhile, تدريسis associated with فيand other markers, suggesting its flexible usage across different teaching contexts. The term تربيةco-occurs with والتعليم(and education) and الوطنية(national), reflecting its strong connection to national education systems and broader educational concepts.

In the right collocations column, تعليمis often followed by فيand وزارة(ministry), highlighting its relationship with formal institutions and educational policies. The term تدريس collocates with هيئة(institution/body) and طرق(methods), emphasizing its focus on teaching practices and pedagogical approaches. Meanwhile, تربيةis followed by وزارةand كلية(faculty), indicating its association with educational institutions. Overall, these collocational patterns demonstrate that each term carries distinct semantic and contextual nuances, underscoring their importance for understanding language use in Arabic educational discourse.

To further understand the functional distribution of these terms, the analysis examines their usage across different discourse contexts. The findings indicate that each term operates at a different functional level within educational discourse, as summarized in Table 3.

Table 3. Functional Distribution in Educational Discourse

Term	Main Context of Use	Functional Level
تعليم	Policy, system, institution	Macro-level (systemic)
تدريس	Classroom, teaching practice	Micro-level (operational)
تربية	Institution, values, administration	Meso-level (institutional)

Table 3 illustrates the functional distribution of three key Arabic educational terms— تعليم (education/teaching), تدريس(instruction), and تربية(education in a holistic sense)—across different contexts and levels of use. The term تعليمis primarily associated with policy, systems, and institutional frameworks, placing it at the macro-level, where it reflects broad, systemic aspects of education. In contrast, تدريسis linked to classroom settings and teaching practices, situating it at the micro-level, where day-to-day instructional activities and pedagogical implementation occur. Meanwhile, تربيةoperates within institutional, value-based, and administrative contexts, positioning it at the meso-level, where overarching systems connect with practical implementation through organizational structures.

Overall, the findings indicate that these three terms demonstrate distinct functional roles within Arabic educational discourse. Their variation in frequency, collocation, and contextual usage highlights how each term occupies a specific niche: تعليمemphasizes systemic and policy-oriented dimensions, تدريسfocuses on practical teaching processes, and تربيةbridges institutional

values and administrative functions. This differentiation underscores the importance of understanding nuanced term usage when analyzing educational language, as it reflects the layered and structured nature of modern Arabic educational systems. Overall, the results demonstrate that the three terms exhibit distinct patterns in terms of frequency, collocation, and contextual distribution. These differences indicate that each term occupies a specific position within modern Arabic educational discourse.

Discussion

The findings of this study demonstrate that the lexical items تعليم, تدريس, and تربيةcannot be interpreted as simple near-synonyms within a unified semantic field of "education," but rather represent distinct functional domains in contemporary Arabic educational discourse. This supports corpus-based approaches to meaning, which argue that lexical semantics is fundamentally usage-driven and emerges from distributional patterns in authentic language data rather than from isolated dictionary definitions (Boleda, 2020).

From a theoretical perspective, these findings challenge traditional lexicographic assumptions that treat these terms as semantically interchangeable. Instead, they align with usage-based semantic theory, which posits that meaning is constructed through recurrent patterns of co-occurrence and contextual embedding. Recent studies in corpus linguistics and language pedagogy similarly emphasize that collocation analysis plays a central role in revealing lexical meaning and usage patterns in educational contexts (Harahap et al., 2025; Sun & Park, 2023).

Empirically, the corpus analysis reveals a clear functional stratification among the three terms. تعليمis predominantly associated with macro-level discourse, particularly institutional and policy-related contexts, as reflected in its frequent collocates such as وزارةand التعليم العالي. This suggests its alignment with systemic and administrative dimensions of education. In contrast, تدريسis primarily embedded in micro-level pedagogical contexts, reflecting classroom instruction and teaching practices. Meanwhile, تربيةoccupies a meso-level position, bridging institutional governance and normative educational values, as evidenced by collocations such as وزارة التربية والتعليم. This distribution indicates that lexical meaning is structured through domain-specific collocational networks rather than simple synonymy.

These findings extend previous corpus-based research that has highlighted the role of collocation in identifying lexical variation across contexts (Tsai, 2021; Yin & Li, 2021). However, while earlier studies primarily focus on frequency distributions or cross-register variation, the present study advances this line of research by proposing a functional-semantic stratification model that explicitly links collocational behavior to discourse-level functions. This shift reflects a move from descriptive corpus analysis to a more interpretive, theoretically informed model of semantic organization.

Furthermore, in line with distributional semantic theory, collocation patterns are not merely indicators of lexical association but also reflect underlying conceptual structures within a domain (Boleda, 2020). The observed separation of collocational environments among تعليم, تدريس, and تربيةsuggests that these terms operate within partially distinct semantic networks rather than forming a single unified conceptual cluster. This supports recent arguments that corpus evidence can reveal fine-grained semantic distinctions that are not easily captured through traditional lexicographic approaches (Brezina, 2023).

From a pedagogical perspective, these distinctions have significant implications for Arabic language education. The differentiation between macro-level (تعليم), meso-level (*تربية*), and micro-level (تدريس) concepts can enhance terminological precision in curriculum design, teacher training, and academic writing. Corpus-informed pedagogy has been shown to improve learners' lexical awareness by exposing them to authentic usage patterns rather than decontextualized vocabulary lists (Harahap et al., 2025; Tsai, 2021). Therefore, integrating corpus-based insights into instructional practices can strengthen the connection between linguistic theory and classroom application.

Despite these contributions, several limitations must be acknowledged. First, reliance on a single corpus (arTenTen24) may introduce genre bias, as web-based corpora tend to overrepresent journalistic and semi-formal texts while underrepresenting spoken and specialized academic discourse (Al-Sulaiti & Atwell, 2006). Second, focusing on high-frequency collocations may obscure less frequent but semantically significant patterns. Third, although corpus methods reduce interpretive subjectivity, functional categorization still involves analytical interpretation.

Future research should incorporate multiple corpora representing different registers, including spoken, academic, and learner corpora, to improve representativeness. Additionally, integrating corpus linguistics with discourse analysis or cognitive semantics may provide a more comprehensive understanding of how educational concepts are constructed and interpreted in Arabic discourse. In conclusion, this study demonstrates that corpus-based concordance and collocation analysis can move beyond descriptive frequency analysis toward a structured model of semantic differentiation. By linking distributional patterns to discourse-level functions, it contributes both methodologically to corpus linguistics and theoretically to the study of Arabic educational terminology.

CONCLUSION

This study demonstrates that تعليم, تدريس, and تربية constitute distinct semantic units within Arabic educational discourse rather than interchangeable synonyms. Corpus-based evidence shows that their meanings are systematically shaped by distributional features, particularly frequency, collocation, and contextual usage, supporting the view that lexical meaning is emergent and usage-driven. The study proposes a functional-semantic model positioning تعليم at the macro/systemic level, تدريس at the micro/instructional level, and تربية at the meso/value-oriented level. This classification enhances the conceptual precision of Arabic educational terminology and demonstrates the explanatory power of corpus linguistics in semantic differentiation.

Theoretically, the findings confirm that collocational patterns can delineate conceptual boundaries within a shared semantic field, reinforcing the context-dependent nature of meaning. Practically, the results contribute to Arabic language education by supporting more precise terminology use in curriculum design, academic writing, and teacher education. Despite these contributions, the study is limited by its reliance on a single web-based corpus and its focus on high-frequency patterns, which may exclude genre variation and low-frequency, yet meaningful, usages. Future research should integrate multiple corpora and complementary approaches such as discourse and cognitive analysis to achieve a more comprehensive interpretation. In conclusion, this study highlights the value of corpus-driven analysis in revealing structured semantic variation. It offers a replicable framework for future research in Arabic corpus linguistics and language pedagogy.

AUTHOR CONTRIBUTIONS STATEMENT

[FNK] contributed substantially to the conception and design of the study, conducted data collection and analysis, and wrote the initial draft of the manuscript. [MZ], as the primary supervisor, provided continuous guidance throughout the research process, contributed to the interpretation of the findings, and critically revised the manuscript for important intellectual content. [AS], as the tertiary supervisor, offered methodological direction and contributed to improving the manuscript's clarity and coherence. All authors have read and approved the final version of this manuscript.

ACKNOWLEDGMENT

We would like to express our deepest gratitude to all administrators, as well as the lecturers and supervisors at UIN Sultan Maulana Hasanuddin Banten, for their invaluable support and cooperation throughout the implementation of this research. Our sincere appreciation also goes to all authors and contributors who provided guidance and valuable input, enriching the quality of this study. In addition, we extend our heartfelt thanks to the editorial team of the *Tadris Al-'Arabiyyah* Journal for their assistance and meticulous review process, which has made a significant contribution to improving the final version of this manuscript. Thank you to everyone for your support and contributions.

REFERENCES

Al Fraidan, A., & Alkuwaity, R. (2025). Corpus-Driven Innovation in Saudi Arabian EFL Teaching Practices. *Educational Process International Journal, 19*(1), e2025553. https://doi.org/10.22521/edupij.2025.19.553

Al-Sulaiti, L., & Atwell, E. (2006). The design of a corpus of contemporary Arabic. *International Journal of Corpus Linguistics, 11*(2), 135–171. https://doi.org/10.1075/ijcl.11.2.02als

Boleda, G. (2020). Distributional semantics and linguistic theory. *Annual Review of Linguistics, 6*, 213–234. https://doi.org/10.1146/annurev-linguistics-011718-011756

Boulton, A., & Forti, L. (2025). Corpus Linguistics and Data-Driven Learning. In *Reference Module in Social Sciences* (p. B9780323955041004828). Elsevier. https://doi.org/10.1016/B978-0-323-95504-1.00482-8

Brezina, V. (2023). Corpus linguistics and collocation analysis: New directions in meaning research. *International Journal of Corpus Linguistics, 28*(3), 345–367. https://doi.org/10.1075/ijcl.21045.bre

Crosthwaite, P., & Baisa, V. (2023). Generative AI and the End of Corpus-Assisted Data-Driven Learning? Not So Fast! *Applied Corpus Linguistics, 3*(3), 100066. https://doi.org/10.1016/j.acorp.2023.100066

Curry, N., & McEnery, T. (2025). Corpus Linguistics for Language Teaching and Learning: A Research Agenda. *Language Teaching, 58*, 1–20. https://doi.org/10.1017/S0261444824000430

Elewa, A. (2025). A Corpus-Based Analysis of Gendered Language in Spoken Religious Discourse. *Applied Corpus Linguistics, 5*(3), 100137. https://doi.org/10.1016/j.acorp.2025.100137

Fitrianto, I. (2024). Innovation and Technology in Arabic Language Learning in Indonesia: Trends and Implications. *International Journal of Post Axial: Futuristic Teaching and Learning, 2*(3), 134–150. https://doi.org/10.59944/postaxial.v2i3.375

Guerza, R. (2023). Transportable Identities in Conversational Interaction among Batna 2 University Students of English. *Arab World English Journal, 14*(2), 65–76. https://doi.org/10.24093/awej/vol14no2.5

Harahap, A., et al. (2025). Corpus-based learning and lexical development in language education. *English Education Journal, 16*(1), 1–15. https://doi.org/10.24815/eej.v16i1.41762

Hizbullah, N., Arifa, Z., Suryadarma, Y., Hidayat, F., Muhyiddin, L., & Firmansyah, E. K. (2020). Source-Based Arabic Language Learning: A Corpus Linguistic Approach. *Humanities & Social Sciences Reviews, 8*(3), 940–954. https://doi.org/10.18510/hssr.2020.8398

Jurko, P. (2022). Semantic Prosody of Slovene Adverb–Verb Collocations: Introducing the Top-Down Approach. *Corpora, 17*(1), 39–67. https://doi.org/10.3366/cor.2022.0234

Kovačević, D. (2026). Corpus Stylistic Analysis with Sketch Engine. *2026 25th International Symposium INFOTEH-JAHORINA (INFOTEH)*, 1–6. https://doi.org/10.1109/INFOTEH68759.2026.11477721

Li, D., Noordin, N., Ismail, L., & Cao, D. (2025). A systematic review of corpus-based instruction in EFL classroom. *Heliyon, 11*(2), e42016. https://doi.org/10.1016/j.heliyon.2025.e42016

Ma, Q., Chiu, M. M., Lin, S., & Mendoza, N. B. (2023). Teachers' Perceived Corpus Literacy and Their Intention to Integrate Corpora into Classroom Teaching: A Survey Study. *ReCALL, 35*(1), 19–39. https://doi.org/10.1017/S0958344022000180

Ma, Q., & Mei, F. (2021). Review of Corpus Tools for Vocabulary Teaching and Learning. *Journal of China Computer-Assisted Language Learning, 1*(1), 177–190. https://doi.org/10.1515/jccall-2021-2008

Mawaddah, Z., Ok, A. H., & Arsyad, J. (2025). Challenges In Teaching Arabic Language At Mas Ulumul Quran Langsa: A Case Study. *Jurnal At-Tarbiyat: Jurnal Pendidikan Islam, 8*(2), 478–484. https://doi.org/10.37758/jat.88i2.74

McEnery, T., & Hardie, A. (2011). *Corpus Linguistics: Method, Theory and Practice*. Cambridge University Press.

Musthafa, I., & Hermawan, A. (2018). *Metodologi Penelitian Bahasa Arab: Konsep Dasar Strategi Metode Teknik*. Remaja Rosdakarya.

Pandža, N. B., Phillips, I., Karuzis, V. P., O'Rourke, P., & Kuchinsky, S. E. (2020). Neurostimulation and Pupillometry: New Directions for Learning and Research in Applied Linguistics. *Annual Review of Applied Linguistics, 40*, 56–77. https://doi.org/10.1017/S0267190520000069

Sardinha, T. B. (2020). A Historical Characterisation of American and Brazilian Cultures Based on Lexical Representations. *Corpora, 15*(2), 183–212. https://doi.org/10.3366/cor.2020.0194

Sun, Y., & Park, J. (2023). Corpus-informed vocabulary learning and collocation awareness in language education. *Sustainability, 15*(17), 13242. https://doi.org/10.3390/su151713242

Tsai, C.-H. (2021). Corpus linguistics in language teaching: Applications and implications. *Language Resources and Evaluation, 55*, 1001–1020. https://doi.org/10.1007/s12528-021-09272-4

Valdivieso, T., & González, O. (2025). Generative AI Tools in Salvadoran Higher Education: Balancing Equity, Ethics, and Knowledge Management in the Global South. *Education Sciences, 15*(2), 214. https://doi.org/10.3390/educsci15020214

Yin, Y., & Li, X. (2021). Collocation patterns and lexical variation in corpus analysis. *Applied Corpus Linguistics, 1*(1), 100006. https://doi.org/10.1016/j.acorp.2021.100006

Yuliawati, S., Ekawati, D., & Mawarrani, R. E. (2021). Investigating Lexical Bundles in the Corpora of English and Indonesian Research Articles with the Sketch Engine. *Jurnal Sosioteknologi, 20*(2), 188–200. https://doi.org/10.5614/sostek.itbj.2021.20.2.5

Zaki, M. (2021). Corpus-Based Language Teaching and Learning: Applications and Implications. *International Journal of Applied Linguistics, 31*(2), 169–172. https://doi.org/10.1111/ijal.12316