Volume 25(4), November 2025 http://doi.org/10.17576/gema-2025-2504-14

A Corpus-Based Frequency List of Arabic Sight Words for Grades 4–6

Ahmad Oweini [a]
aoueini@lau.edu.lb
Department of Psychology and Education
Lebanese American University, Beirut-Lebanon

Noura El-Far
noura.elfar@lau.edu
Department of Psychology and Education
Lebanese American University, Beirut-Lebanon

ABSTRACT

This study updates and extends the foundational work of Oweini and Hazoury (2010) by developing a new corpus-based list of Arabic sight words for grades 4-6 in Lebanese private schools. It addresses the lack of recent, systematic descriptions of high-frequency Arabic words and examines their role in reading fluency and comprehension. Fifteen commonly used Arabic reading textbooks were analyzed, and words were categorized by type and morphological structure. The findings confirm the dominance of functional words and reflect the influence of diglossia and Arabic orthographic complexity on sight word identification. Comparison with existing word lists and frequency dictionaries shows strong overlap but also highlights areas requiring revision. This updated frequency list offers educators and textbook designers linguistically grounded evidence for improving reading materials. The study also recommends expanding future research using digital corpus tools to support larger-scale frequency analyses.

Keywords: Arabic sight words; corpus linguistics; word frequency; diglossia; orthography; morphological structure; reading fluency

INTRODUCTION

Sight words are a key element of fluent reading because they are recognized automatically without the need for decoding. This automatic recognition allows readers to allocate cognitive resources to higher-level comprehension processes (Ehri, 2023). In linguistic terms, sight words are typically high-frequency items, including both function and content words, that constitute a large proportion of continuous text (Ehri, 2023). Research in English and other European languages has demonstrated the significance of frequency-based word lists in shaping reading fluency and text comprehensibility (Adams, 1990; Chall, 1983; Moser, 2020). National education systems such as those in France, Spain, and Germany have developed official word frequency lists that inform curriculum design and corpus-based research on vocabulary (Liste de fréquence lexicale, 2023).

[a] *Main & corresponding author*

eISSN: 2550-2131
ISSN: 1675-8021

In Arabic, the situation is more complex due to diglossia and orthographic features such as affixation and polymorphic structures, which influence word recognition. Early work by Oweini and Hazoury (2010) produced a list of 500 high-frequency sight words for grades K-3, but research on later grade levels remains limited. This study expands their work by examining the most frequent words in Arabic textbooks for grades 4-6, focusing on frequency, morphological structure, and linguistic features. The objective of this study is to generate a corpus-based description of Arabic vocabulary frequency that can provide linguistically grounded evidence for selecting grade-appropriate sight words in reading materials.

THE DEVELOPMENT OF SIGHT WORD LISTS

DOLCH LIST

The Dolch Word List remains one of the most influential frequency-based compilations of vocabulary for early reading in English. Although studies of word frequency predate Dolch's work, for example, Gates' (1926) list of primary vocabulary and the Child Study Committee of the International Kindergarten Union's 1928 compilation, Dolch (1936) provided a list that became widely adopted in schools and research. His compilation included 220 function words, often referred to as "service words," with 95 nouns. These words were not selected randomly but represented the most common vocabulary in children's literature of the time. The inclusion of different parts of speech such as pronouns, adjectives, prepositions, conjunctions, and verbs illustrated the fact that children encounter a variety of word types even in the earliest stages of reading.

Interestingly, the grade-level ordering in Dolch's list created a linguistic paradox: words that were orthographically simple and phonetically regular, such as if, got, cup, and ten, were introduced at later grade levels, while longer and less decodable words such as little, yellow, and away appeared in pre-primer lists.

The Dolch list has since become a linguistic and educational reference point, demonstrating how 220 words can account for a substantial proportion of text encountered by young readers. It continues to influence word frequency studies across languages, offering evidence that a limited lexical set can carry significant weight in early literacy.

FRY LIST

Building on Dolch's earlier work, Fry introduced the Instant Word List in 1957, later revised and expanded in 1980 as the New Instant Word List. Fry's contribution was significant not only for pedagogy but also for corpus-based linguistics, as he emphasized the quantitative dominance of a small set of words in written English. The first 100 words on Fry's list account for approximately 50% of all printed text, while the top ten words alone (the, of, and, a, to, in, is, you, that, it) make up nearly a quarter of written materials (May, 2005). By the time Fry expanded the list, it contained 300 words ranked by frequency, with further extensions reaching 1,000 words. His methodology drew from the American Heritage word list, which analyzed frequency patterns in materials used in grades 3–9, thereby grounding the list in authentic corpus data (Fry, 1980).

eISSN: 2550-2131
ISSN: 1675-8021

Fry's innovation lay in refining earlier frequency lists by addressing methodological inconsistencies. Unlike Dolch, Fry grouped together inflectional variants such as run, runs, running under a single lemma. This approach aligned his work with morphological principles and corpus linguistics by treating words as families rather than isolated forms.

The Fry list quickly became widely disseminated in teacher manuals, curriculum guides, and remedial reading programs, but its impact extended beyond pedagogy. Linguistically, it offered empirical confirmation of Zipf's Law, demonstrating how a small number of high-frequency words dominate language use. Its enduring value lies in showing how statistical distributions of vocabulary can shape both reading instruction and linguistic theory.

THE IMPORTANCE OF UPDATING DOLCH AND FRY WORD LISTS FOR MODERN EDUCATION

The Fry list has not been revised since 1980, yet the need to update sight word lists remains urgent in light of linguistic, cultural, and technological change. Languages are not static; lexical items shift in frequency, new words emerge, and others become obsolete (Franceschi, 2021). A frequency list compiled more than four decades ago cannot fully reflect the vocabulary encountered by contemporary readers. Updating such lists ensures that high-frequency words included in instruction remain relevant to current language use.

Another reason for revision concerns the alignment between lexical frequency data and curricular benchmarks. As educational curricula undergo periodic revisions, the vocabulary considered essential for literacy development should be empirically supported by up-to-date corpora (Ballance & Coxhead, 2024). This alignment reinforces the linguistic validity of word lists, ensuring they accurately represent the words students most frequently encounter.

Equally significant is cultural and contextual relevance. Modern learners engage with texts that increasingly reflect issues of identity, diversity, and global citizenship. Sight word lists, therefore, should incorporate lexical items that represent these semantic domains, such as culture, identity, tolerance, and sustainability (Aswadi, 2024; Hutchison et al., 2024). These words are not only pedagogically desirable but also linguistically representative of the thematic content dominating modern discourse.

Technological development also influences word frequency. The rise of digital communication has introduced terms such as internet, digital, download, and cyber, which appear with high frequency in contemporary corpora (Lazareva & Terekhin, 2023). Failure to include such vocabulary risks creating word lists that misrepresent the linguistic environment of learners.

Finally, contemporary word lists can enhance reading fluency and comprehension by incorporating lexical items that learners are most likely to encounter in authentic materials (Liu et al., 2024). Updating Dolch and Fry lists is thus not merely a pedagogical concern but a linguistic necessity, ensuring that frequency-based resources reflect the evolving nature of language, society, and technology.

eISSN: 2550-2131
ISSN: 1675-8021

IS A LIST OF SIGHT WORDS NEEDED FOR ARABIC READING INSTRUCTION?

Research over the past decade has consistently shown that learning to read Arabic presents challenges beyond those encountered in many other languages, primarily due to diglossia and the orthographic complexity of the Arabic script (Asaad & Eviatar, 2013a, 2013b; Maroun et al., 2020). The debate over whether sight words should be explicitly taught in Arabic literacy instruction is therefore closely tied to these linguistic features, which affect reading accuracy, word recognition, and acquisition strategies (Saiegh-Haddad & Joshi, 2014).

DIGLOSSIA

A central factor is diglossia: the coexistence of Modern Standard Arabic (MSA), or Fus'ha, and colloquial Arabic, or 'Amiya (SpA). MSA functions as the formal variety used in schools, media, and writing, while SpA dominates daily communication (Saiegh-Haddad & Everatt, 2015). For young learners, MSA often functions almost as a second language, since its phonology, lexicon, and morphology diverge significantly from the spoken register (Taha, 2022). Empirical studies assert that only about 21% of the vocabulary of a five-year-old Arabic speaker overlaps directly with MSA, with the remainder being partially cognate or entirely distinct (Saiegh-Haddad & Spolsky, 2014). This gap complicates early literacy and has long-term effects on decoding and fluency (Saiegh-Haddad & Schiff, 2016).

Research demonstrates that readers demonstrate greater accuracy when processing familiar spoken forms than when encountering MSA, revealing the cognitive demands of diglossia (Baharun, 2025). These challenges extend beyond early grades, persisting into upper elementary years (Darwiche Fedda & Oweini, 2012). One proposed strategy is the systematic introduction of high-frequency sight words drawn from MSA, which can provide lexical anchors to bridge the gap between colloquial and formal varieties (Shendy, 2022). By grounding word selection in corpus-based frequency and meaningful context, learners can develop automatic recognition that reduces cognitive load, allowing attention to shift toward comprehension (Saiegh-Haddad, 2014).

Thus, the need for an Arabic sight word list arises not simply from pedagogical practice but from the unique linguistic challenges posed by diglossia and orthography. A linguistically grounded, frequency-based list could play a critical role in supporting reading acquisition in Arabic.

ORTHOGRAPHY

The Arabic orthographic system is unique and complex, featuring several distinctive characteristics (Asadi et al., 2023). It is a phonological and inflectional script comprising 28 letters and diacritics that indicate vocalisation (Asadi et al., 2023). Unlike the Roman alphabet, Arabic is written from right to left and lacks space between words. Additionally, short vowels are typically omitted in vowelized text, which adds to its complexity (Boumaraf et al., 2022).

According to Oweini and Hazoury (2010), "the characters of the Arabic letters vary in shape. Each has more than one written form, depending on the letter's place in a word: beginning, middle or end. However, the essential shape of the letter is maintained in all cases (Awwad, 2004, p. 462). There are six-one-sided connectors only from the right (ز - و - د - ذ - ر and ا) making two shapes for each letter, 22 connectors from the right and the left has four shapes: ـه - ـهـ - ـهـ - and ه. Therefore, readers can encounter one-, two-, three or four-part words if such words contain one-

sided letters in the middle. For example, "عامِل" (worker), "دَرَسَ" (he studied) and "أرزة" (cedar tree) are made up of two, three and four parts, respectively.

The Arabic script presents unique challenges for reading due to its distinctive font features, the complexity of characters, the extensive use of dots, and the reliance on vowelisation marks or diacritics (Alabdulkader et al., 2021), leading students to decode Arabic letters more slowly compared to their counterparts reading Hebrew or English (Maroun et al., 2020). Arabic script can be either shallow (fully vowelised) or deep (unvowelised). These marks are governed by rules related to word meaning, inflection, and function within a sentence. In upper elementary grades, students transition to unvowelised texts to refine their word recognition skills using lexical context (Maroun et al., 2020) and syntax (Alabdulkader, 2021).

A specific challenge for spellers is *tanween* (nunation), which adds double diacritic marks at the end of a word to indicate an additional /n/ sound. Beginning or struggling readers may mistakenly substitute the letter "ن" for the *tanween*, creating non-existent but phonetically correct words (Fawaz & Dia, 2017). Studying whether mastering partially vowelised sight words improves phonetic-orthographic connections may offer insights into this issue.

Unvowelised text often exacerbates reading difficulties for struggling readers, leading educators to question whether retaining diacritic marks in higher grades might better support comprehension. Research on this remains inconclusive. While Abu-Rabia and Hijjazi (2023) argue that vowelisation aids comprehension by disambiguating word meanings, others, like Ibrahim (2013), Taha (2016), Maroun and Hanley (2016) and suggest that reliance on vowelisation may hinder the development of advanced reading strategies, across several grade levels (Saiegh-Haddad & Schiff, 2016).

A final feature, namely the *shaddah*—a diacritic that marks consonant gemination or doubling—plays a vital role in Arabic. Equivalent to doubling a consonant in Latin-based orthographies, the *shaddah* is also a distinctive feature in languages such as Japanese and Turkish (Maroun & Hanley, 2016).

Research has shown that the best predictor of accurate reading in young children is phonological awareness, followed by naming speed (Tibi & Kirby, 2018). Additional factors, such as syntactic and orthographic knowledge, also contribute to reading accuracy, while morphological knowledge and vocabulary are more predictive of reading comprehension (Asadi & Kawar, 2023).

In elementary education, understanding the complexities of Arabic orthography is essential for effectively teaching sight words. To address the challenges of Arabic script, Saiegh-Haddad (2018) proposed a three-pronged model for teaching reading in Arabic, focusing on:

1. vowelised texts – developing decoding skills through exposure to fully vocalised words.
2. morphological training – enhancing understanding of word structure and root-based patterns; and
3. bridging the gap between diglossia and standard Arabic – incorporating strategies to ease the transition between colloquial Arabic and Modern Standard Arabic.

This model provides a comprehensive framework for addressing the unique linguistic and cognitive demands of reading in Arabic.

eISSN: 2550-2131
ISSN: 1675-8021

NEW RESEARCH ON RIGHT WORDS IN ARABIC

Research on Arabic sight words remains limited compared to English and other European languages, yet several notable contributions have emerged in recent years. One significant resource is *A Frequency Dictionary of Arabic: Core Vocabulary for Learners* (Buckwalter et al., 2011), which compiles the 5,000 most frequently used words in Modern Standard Arabic (MSA) as well as several widely spoken dialects. Based on a 30-million-word corpus encompassing both written and spoken materials, the dictionary provides English translations, usage data, sample sentences, and genre-specific distributions. The list can be accessed alphabetically, by Arabic roots, or thematically, thereby offering a versatile linguistic tool for learners and researchers.

From an Islamic linguistic perspective, Sarmini (2019) produced a thematic dictionary of approximately 5,000 high-frequency words drawn from classical texts, daily usage, and Islamic economics. This resource emphasizes vocabulary that is central to theological and religious discourse, emphasizing the role of specialized frequency lists for targeted domains.

In the Lebanese context, Kozma and Yacoub (2019) examined Grade 3 students in an international private school in Beirut, deriving sight word lists from commonly used Arabic reading primers for Grades 1–3. Their manual tallies identified 150 high-frequency words, categorized by grade level and included a list for sun and moon letters. While framed pedagogically, their study demonstrates the methodological potential of primer-based word extraction for generating frequency lists tailored to local contexts.

More broadly, Moser (2020) compared vocabulary selections from five Arabic textbooks for second-language learners in the United States with the 3,000 most frequent words in Buckwalter et al.'s (2011) dictionary. The results demonstrated limited overlap, underscoring a gap between authentic high-frequency vocabulary and instructional materials.

Collectively, these studies illustrate the diversity of approaches to identifying "right words" in Arabic, whether corpus-based, thematic, or instructional. Yet they also draw attention to the absence of a standardized, empirically grounded sight word list for Arabic elementary education, reinforcing the need for continued research.

Despite these valuable contributions, Arabic still lacks large-scale, grade-specific, and pedagogically validated sight word lists comparable to those available in English and other languages. Several structural features, including diglossia, rich morphology, and orthographic variation — complicate frequency estimation and require corpus-driven approaches tailored to young readers. Existing Arabic resources either focus on adult learners, specialized domains, or limited grade ranges, leaving a gap in frequency-based tools for elementary literacy instruction. Expanding Arabic corpora for school-age learners is therefore essential for developing instructional materials that reflect actual lexical exposure in classrooms. The present study contributes to this need by generating an updated, empirically grounded sight word list based on widely used Lebanese reading textbooks for Grades 4–6.

eISSN: 2550-2131
ISSN: 1675-8021

GRAIN SIZE THEORY AND DEVELOPMENTAL PHASES

Ziegler and Goswami's (2005) grain size theory provides a useful framework for understanding cross-linguistic differences in reading acquisition. They argued that although the developmental trajectory of phonological representation is broadly similar across languages, orthographic transparency and granularity shape the pace and strategies of reading development. In transparent orthographies, where grapheme–phoneme correspondences are consistent, learners acquire decoding skills more quickly. By contrast, readers of opaque orthographies such as English must rely on multiple recoding strategies across varying grain sizes, which slows development and creates greater cognitive demands. In this context, high-frequency word lists play a compensatory role, enabling readers to bypass irregularities at smaller grain sizes.

Empirical studies support these claims. Broun and Deavers, cited in Ziegler and Goswami (2005), showed that less skilled English readers predominantly used small grain size strategies, while skilled readers relied more on larger units such as rimes. When nonwords with both regular and irregular graphemes were presented, frequent switching between small and large units imposed a "switching cost" (Ziegler & Goswami, 2005, p. 12), requiring additional cognitive resources. These findings stress the significance of orthographic consistency and frequency effects in literacy acquisition.

In the case of Arabic, recent research extends these insights. Taha (2023) found that children's familiarity with Arabic orthographic patterns enhanced phonemic segmentation, particularly for words with recognizable patterns compared to pseudowords. This supports grain size theory, since Arabic's relatively transparent orthography facilitates reliance on smaller grain sizes, though diglossia introduces an additional layer of complexity.

Baharun (2025) further expanded the discussion by linking grain size theory to the interactive and dynamic model of literacy. They emphasized that reading and writing co-develop through shared cognitive resources (interactive), but the strength of their relationship varies across grain sizes and developmental phases (dynamic). At the lexical level, correlations between reading and spelling are strong due to shared orthographic and phonological processes, while at the discourse level, relationships are weaker because comprehension and composition draw on broader cognitive and linguistic resources. This illustrates both the interconnectedness and distinctiveness of reading and writing within Arabic literacy development.

In addition to developmental reading theories, this study is grounded in core principles of corpus linguistics. Tokenization and frequency distribution analysis allow for the systematic identification of the most recurrent word forms within authentic instructional texts. These distributions typically follow Zipf's Law, whereby a small number of high-frequency words account for a large portion of the text, making them central to reading fluency and orthographic mapping. Integrating corpus-based frequency evidence with reading theory therefore strengthens the rationale for sight word instruction: repeated exposure to high-frequency forms supports automaticity, reduces cognitive load during decoding, and facilitates the transition from phonological recoding to more efficient lexical processing. This combined framework underpins the methodological and pedagogical orientation of the current study.

eISSN: 2550-2131
ISSN: 1675-8021

IMPLICATIONS FOR READING INSTRUCTION ON BOTH GOOD AND DYSLEXIC READERS IN ARABIC

Empirical research on Arabic orthography outlines important implications for both skilled and dyslexic readers. Studies show that learners with reading disabilities often rely more heavily on a visual orthographic route than on phonological decoding. Abu-Rabia et al. (2003) demonstrated that dyslexic readers in Arabic experience particular difficulty with pseudo-words and low-frequency vocabulary, reflecting impaired decoding skills. More recent work by Abu-Rabia and Darawshe (2024) confirms these findings, showing that deficits in phonological recoding impede the memorization and recognition of new words. These results reinforce the importance of orthographic features in Arabic and suggest that teaching high-frequency sight words can provide lexical anchors that reduce reliance on impaired phonological pathways.

Morphological decomposition also plays a central role in reading. Saiegh-Haddad (2017) found that morphological awareness facilitates recognition of low-frequency words, while high-frequency words are accessed more automatically from the mental lexicon. This distinction suggests that frequency-based sight word instruction complements morphological processing, particularly for learners with weaker phonological skills.

Orthographic predictability further influences reading development. Maroun et al. (2020) note that vowelization supports word recognition by increasing transparency, although irregularities in vowel marking continue to pose challenges for beginners. These findings point to the importance of consistent orthographic cues in facilitating decoding accuracy.

Cross-linguistic research on grain size also provides insight into dyslexia in Arabic. Ziegler and Goswami's (2005) theory predicts reliance on small grain size strategies in transparent orthographies, yet evidence from Turkish, Greek, and Hungarian suggests that larger grain approaches—such as teaching whole words or sentences—can also support readers with strong phonological skills. For those with deficits, targeted instruction at smaller grain sizes may be necessary. These findings imply that sight word instruction, particularly for high-frequency items, could be a valuable complement for struggling readers of Arabic, though further research is needed.

In sum, the integration of phonological, orthographic, and morphological insights underscores the need for corpus-based word lists in Arabic, particularly for upper-primary learners, where such resources remain largely absent. Addressing this gap in corpus-based literacy tools is essential for supporting both typical readers and students with dyslexia through linguistically grounded instructional materials.

CRITERIA FOR WORD SELECTION

The following types of words were selected:

1. Words that are supposed to be read by students. Hence, sentences, phrases, and words that appear to be technical or destined to adults (parents, teachers, educators, etc.) were omitted.
2. Unlike the previous K-3 list, proper names were excluded, unless the name can be used as a noun or verb (for example, " سالم"- safe and sound) with the exception of " لبنان" – Lebanon.

eISSN: 2550-2131
ISSN: 1675-8021

3. Root words with various orthographic characteristics by words with different inflections since the visual feature of a given word is modified by the addition of prefixes, suffices and infixes. In this regard, four criteria were used in word selection:

    a) Homographs with different diacritics, or words that share the same spelling but differ in diacritics, are treated as distinct words due to their unique meanings. For instance, "عَلِمَ" (he knew) and "عُلِمَ" (it was known) represent two separate entries.
    b) Mortho-orthographic variations in word forms: Words with different morphemes that are derived from the same root but serving different semantic and grammatical purposes are considered distinct. For example, "يَكْتُبُ" (he writes), "كَتَبَ" (he wrote), and "مَكْتُوب" (written) are listed as separate words, even though they share the same root.
    c) Words with similar pronunciation but different meanings: Words like "رَجُلٌ" (a man) and "رِجَالٌ" (men) are counted as different entries, despite their shared semantic field, due to differences in form and inflection.
    d) Addition of the definite article: Adding "الـ" - the definite article “the”, to a root creates a visually distinct word. For example, "بَيْت" (house) and "الْبَيْت" (the house), are treated as two separate entries.
4. Vowelisation differences: Some words are represented with their exact diacritical marks to preserve orthographic accuracy, while others are listed partially or fully unvowelised, depending on their occurrence in various reading materials.

METHOD

CORPUS SELECTION AND SAMPLING PROCEDURE

Given that approximately 70% of students in Lebanon attend private schools (Oweini & ElZein, 2022), the current study focused on reading series widely adopted in these institutions. Three textbook series were selected: Loughati Farahi (Al-Qadi, 2016), which was also used in the earlier K–3 study (Oweini & Hazoury, 2010), Al Siraj (Goush, 2008), and Oukood Al-Kalam (Chartouni et al., 2020). For each series, only the grade-specific editions for Grades 4, 5, and 6 were included, yielding a total of 15 textbooks in the corpus. Although analyzing private-school textbooks is appropriate given their dominance in Lebanese schooling, it is acknowledged that this focus may limit the representativeness of public-school curricula. The resulting corpus contained approximately 9,183 tokens, providing a medium-sized pedagogical corpus suitable for frequency analysis.

To ensure systematic sampling, pages were selected randomly, with one out of every four pages reviewed per book. This approach balanced feasibility with representativeness, reducing bias in word occurrence. This probabilistic sampling strategy aligns with corpus design principles outlined by Biber et al. (1998) and McEnery and Hardie (2012), who recommend random sampling to minimize topical skew and enhance representativeness in medium-sized corpora. Since digital versions of these books were not available, all analyses were conducted manually by a team of six trained graduate assistants. Before independent coding began, all assistants completed a joint training session and piloted the sampling and normalization procedures on shared pages to establish consistency and resolve discrepancies. These sampled pages then served as the basis for the subsequent word extraction and data collection procedures.

TEXTBOOK DATA COLLECTION PROCESS

Each word on the sampled pages was documented in a frequency table, with the following details recorded: the word form, number of occurrences on the page, and the page number of first appearance. When the same word recurred in later sampled pages, additional occurrences were added to its cumulative frequency count. To ensure consistency, all assistants applied the same definition of "word form," treating orthographic variants (e.g., alif–hamza forms) according to unified normalization rules agreed upon prior to coding. After the coding of each textbook was completed, all word lists were merged into a master spreadsheet that consolidated frequency information across the corpus. This allowed for the creation of a unified database in which words were ranked by total frequency. These consolidated lists were then reviewed using the predefined inclusion and exclusion criteria described in the following subsection

WORD INCLUSION AND EXCLUSION CRITERIA

The inclusion of words in the list was guided by specific linguistic and contextual criteria. First, only words expected to be read by students were retained; sentences, phrases, and technical terms addressed to adults (teachers, parents) were excluded. Second, proper names were generally excluded, except where the name also functioned as a common noun or verb (e.g., "سالم" meaning safe and sound), with the exception of "لبنان" (Lebanon), which was retained for cultural relevance.

Third, morphological and orthographic variations were treated as distinct entries when they altered word meaning or form. Homographs with different diacritics (e.g., "عَلِمَ" (he knew) vs. "عُلِمَ" (it was known)), morphological derivatives from the same root (e.g., "يَكْتُبُ" (he writes), "كَتَبَ" (he wrote), "مَكْتُوب" (written)), and plural forms (e.g., "رَجُلٌ" (a man) vs. "رِجَالٌ" (men)) were all recorded separately. Words with and without the definite article (e.g., "بَيْت" (house) vs. "الْبَيْت" (the house)) were also considered distinct lexical items. Finally, vowelization was preserved where relevant to orthographic accuracy, though partially or fully unvowelized forms were also included when they appeared in the textbooks. These criteria were applied consistently across all textbooks, and the following subsection outlines the procedures used to ensure the reliability and validity of this coding process.

PROCEDURES FOR ENSURING RELIABILITY AND VALIDITY

To ensure reliability, each frequency table was independently reviewed by at least two assistants, and discrepancies were resolved by consensus. Word counts were cross-checked to minimize transcription errors. Validity was supported by aligning the sampling method with earlier studies (Oweini & Hazoury, 2010), while adapting the frequency threshold to reflect upper-grade texts, where words occur less often. Because upper-primary texts contain a broader lexical range and lower repetition rates, the frequency threshold used for determining sight-word inclusion was adjusted. Whereas the K-3 study applied a criterion of 21 occurrences, the present study adopted a minimum frequency of five occurrences across the corpus. This threshold reflects realistic exposure patterns for older readers and ensures that included words represent meaningful lexical recurrence within upper-grade instructional materials. These reliability procedures ensured a consistent dataset from which the subsequent frequency analysis could be conducted.

eISSN: 2550-2131
ISSN: 1675-8021

ANALYTICAL PROCEDURES FOR GENERATING THE FREQUENCY LIST

The study adopted a mixed-method design, combining quantitative statistical analysis with qualitative linguistic description. After consolidation, the frequency lists were compared across textbooks and cross-validated with existing Arabic frequency dictionaries. Words were categorized by grammatical class and morphological structure (monomorphic vs. polymorphic). Special attention was given to function words due to their high frequency in Arabic texts and their structural importance in sentence construction.

Quantitatively, descriptive statistics were used to determine the distribution of function and content words, grammatical categories, morphological complexity, and orthographic features such as shadda and hamza. To further validate observed patterns, inferential statistics were applied. Specifically, chi-square tests of goodness-of-fit were conducted to assess whether the frequency differences between function and content words were statistically significant. This statistical component strengthened the interpretation of frequency patterns by confirming that observed differences were unlikely to have occurred by chance, thus providing quantitative support for the linguistic findings. These statistical procedures ensured that the subsequent results were grounded in both linguistic and quantitative evidence, as detailed in the next section.

RESULTS

OVERALL FREQUENCY RESULTS

The manual analysis of Arabic reading materials for Grades 4, 5, and 6 yielded a comprehensive list of sight words organized according to frequency. The results are presented in two tables: Table 1 provides the full list of sight words with their respective frequencies, while Table 2 presents a subset focusing on tool words, including prepositions, adverbs, conjunctions, and pronouns with their common morphological variants. The inclusion of Table 2 is particularly significant, as tool words occur consistently across diverse text types, such as grammar exercises, narrative passages, social studies content, civics lessons, and short stories, thereby underscoring their centrality in reading development.

TABLE 1. Frequency of Sight Words (Grades 4-6)

Frequency	Sight words									
1-10	و	في	.الى	مِنْ	ما	عَلى	مِنَ	لا	تفاصيلُها	النّص
11-20	حاوِلْ	يا	هذا	بِهَذِا	جِجارته	شَكْلِ	هذه	الْفِعْلُ	عنِ	كانَ
21-30	إسْمِ	الوَحدَةُ	أَم	رَسْمٍ	إلَيْهِ	هل	أَنْ	الّتي	الْمُضَافُ	بنائي
31-40	وَصْفِ	أُذْكُرْ	أَنَّه	إذَا	الإسْمُ	الّذي	هَلْ	الـ	بِأَنْ	الْأَرْضِ
41-50	الْمُضَافِ	لِدَقيقَتَينِ	الجملةَ	لحَوّلَهُ	القِصَّةُ	كُلِّ	الْجَمْع	حاشيتي	خَشَباتُ	قَبْلَ
51-60	كَيْفَ	الْفَاعِلُ	نَوْع	إنّ	تَدَرَّج	بهدوء	بَيْنَ	ثمّ	مفرد	بَعْدَ
61-70	أَنَّ	التّالية	الرّسالة	ماذا	مَع	أنتَ	رأَيْتَ	لَمْ	يُقالُ	كيف
71-80	أَنَمرَّنْ	اذْكُرْ	البرميل	الدَّبّور	تَرَجَّلْتُ	أَنْسى	أَوْ	القطار	جَميلَة	ذلكَ
81-90	خَوْفَها	أنا	الكاتبُ	دفترٍ	ربيع	فَهذا	النَّشاطات	لُبْنانْ	هؤلاء	يُراقِبُ
91-100	اسِمٍ	عَلَيْهِ	أختارُ	أمّا	الأسئلةِ	السَّالِمِ	الشِّعر	الضّمير	العصافير	الْمُذَكَّرِ
101-110	باسِمٍ	ذَهَبَ	يُفَكّرُ	أَكْثَر	يَدْلُّ	المُؤَنَّثُ	بَريدًا	تَدُلُّ	حتى	رِسالَةُ
111-120	شَقَّ	كانَتْ	لَقَدْ	التّاءُ	بَيْتِ	جَديدَ	فيها	كَصَفْيْنِ	كَيْ	لَكَ
121-130	لَها	مُضارِعَةً	نَعّوم	هُنا	واحداً	انّني	أُمّي	الرّسالّةِ	إننا	رائعة
131-140	قَرْيَة	كَثيراً	كَيْفَ	أبو	أستخرجُ	إسْمِ	اجْتَمَع	الأَرض	الإجابة	التّطبيق
141-150	التَّسَلْسُلِ	الجَدّانِ	الضَّمَّةُ	العصفورة	الفصل	المدرسة	المدينة	المرسِل	المُذَكَّرِ	الْمَفْعُول

eISSN: 2550-2131
ISSN: 1675-8021

151-160	جَدَّتي	جَمْعٌ	صَوْتَ	فَلا	قَدْ	لكِنّ	لهُ	ملك	منها	مُثَنًّى
161-170	هناك	وَطَلَبْتُ	يَمين	حَقيقَةٌ	أَشْجارُ	أكتُبُ	الشُّجَيْراتِ	الطّريق	كُتِبَتِ	لِماذا
171-180	مَوْز	كَيْفَما	أَسْتَنْتِجْ	أمامي	أي	الأسدِ	البَيْتِ	التّعبيرُ	النَّعْلَبُ	الدّالّة
181-190	السِّجلّات	السَّنابِلُ	الصَّفّ	الفَراغَ	الكِتابِيّ	المَكانُ	اليَسار	إمْلَأْ	بطاقة	بها
191-195	راقِبُ	شفويّاً	مَرْفُوعٌ	مَشْهُورُ	هي					

TABLE 2. Frequency of Tool Words (Grades 4-6)

Frequency	Tool words				
1-5	و	في	الى	من	ما
6-10	على	من	لا	يا	هذا
11-15	بهذا	هذه	عن	كان	إليه
16-20	هل	أن	التي	أنه	إذا
21-25	الذي	هل	الـ	بأن	حول
26-30	كلّ	قبل	كيف	أم	إن
31-35	بين	ثمّ	بعد	أنّ	ماذا
36-40	مع	أنتَ	لم	كيف	أو
41-50	ذلك	أنا	فهذا	هؤلاء	عليه
51-55	أمّا	أكثر	حتّى	كانت	لقد
56-60	فيها	كي	لكَ	لها	هنا

Analysis of the 195 sight words showed a clear distinction between function words and content words. A total of 68 words (34.9%) were function words, including articles, prepositions, conjunctions, pronouns, and auxiliary verbs (e.g., و-في- إلى- كان). These words play a structural role in sentences and occurred with the highest frequency across the corpus. In contrast, 127 words (65.1%) were content words, comprising nouns, verbs, adjectives, and adverbs (e.g., مدرسة- ذهب- جميل- بهدوء). Content words carried the main semantic meaning of the texts but appeared less consistently than function words. This distribution reveals that while content words form the majority in raw count, function words dominate text processing and comprehension due to their high recurrence and grammatical necessity.

To determine whether the observed difference between function and content words was statistically significant, a chi-square goodness-of-fit test was conducted. The analysis compared the observed frequencies of function words (n = 68) and content words (n = 127) against the null expectation that both categories would occur equally often within the 195 sight words. Results showed a significant deviation from equality, χ^2 (1, N = 195) = 17.86, p (0.0027< .001), indicating that content words occurred with significantly greater frequency than function words. This outcome confirms that the dominance of content words in the overall corpus is unlikely to have arisen by chance, reflecting the lexical diversity and thematic range of Arabic reading materials for Grades 4–6.

The significance of this result reinforces the descriptive findings that function words, while fewer in number, occupy the highest frequency ranks, forming the grammatical scaffolding of the texts, whereas content words account for the broader lexical inventory. Thus, the chi-square analysis quantitatively supports the linguistic distinction between lexical abundance and functional recurrence, demonstrating how Arabic reading materials balance grammatical cohesion with semantic expansion.

Furthermore, analysis of the compiled corpus established that the identified sight words could be categorized into six grammatical groups: nouns, verbs, adjectives, pronouns, particles, and adverbs. Table 3 displays the grammatical distribution of the 195 sight words identified in the corpus. Nouns were the most prevalent category, representing 51% (n = 99) of the total words,

reflecting the strong nominal structure of Arabic texts used in upper elementary grades. Verbs accounted for 20% (n = 39), showing that action-oriented vocabulary remains central to reading materials at this level. Adjectives, pronouns, and particles (which include prepositions, conjunctions, negations, and emphatics) each represented 9% of the corpus, indicating the balanced presence of descriptive, referential, and grammatical elements. Finally, adverbs made up 2% (n = 5) of the total, mainly denoting manners or place (e.g., هنا, هناك, بهدوء, كثيراً).

This grammatical distribution demonstrates that Arabic reading materials at Grades 4–6 rely heavily on nominal and verbal structures, while still maintaining sufficient grammatical cohesion through particles and pronouns. The predominance of nouns also suggests that content in these grades emphasizes description and concept-building, whereas function words; though fewer in number, continue to play a key role in sentence connection and fluency.

TABLE 3. Grammatical Distribution of Sight Words (Grades 4-6)

Category	Number	%	Examples
Nouns	99	51%	المدرسة، القصة، الأرض، الكاتب، الطريق، القرية، النص، التطبيق، الصوت، الجدّان، البطاقة، النشاطات، لبنان، الرسالة، البيت، الثعلب، الشعر، الضمير، الجمع، الملك، حقيقة، الطريق
Verbs	39	20%	حاول، أذكر، اجتمع، كتب، يراقب، رأيت، أنسى، تدرّج، كُتبت، أستنتج، راقب، املأ، ترجلت، أكتب، طلبت، يفكر، يدل، تدل، شقّ، كانت
Adjectives	18	9%	جميلة، جديدة، مرفوع، مؤنث، مذكر، مشهور، السالم، رائعة، المضارعة، التّالية، كثير
Pronouns	17	9%	أنا، أنت، هي، إنّني، إننا، هؤلاء، هذا، هذه، ذلك، الذي، التي، ماذا، أي، كيف، لماذا، من، أمامي
Particles (Prepositions, Conjunctions, Negations, Emphatics)	17	9%	و، في، إلى، من، على، ثم، قبل، بعد، بين، أو، لكنّ، أمّا، حتى، كي، إنّ، أن، قد، لا
Adverbs	5	2%	هنا، هناك، بهدوء، كثيرًا، شفويًّا

A further dimension of analysis distinguished between monomorphic and polymorphic forms. Monomorphic words consist of a single morpheme that carries meaning, such as "لِكِتاب" (for a book), where the morpheme "لِ" indicates purpose or possession. Polymorphic words, by contrast, include multiple morphemes that modify the root word. For example, "لِكِتابي" (to my book) incorporates both the prefix "لِ" and the suffix "ي," altering the root "كتاب" (book) to indicate first-person singular possession. This classification affirms the morphological richness of Arabic and demonstrates the necessity of accounting for both simple and complex word forms in sight word lists.

Table 1 reports the ranked 195 sight words, while Table 2 reports the 60 tool (function) words. In the full sight-word list, 77 of 195 words (39.5%) are polymorphic and 118 (60.5%) are monomorphic. By contrast, in the tool-word subset only 10 of 60 (16.7%) are polymorphic and 50 (83.3%) are monomorphic. This contrast shows that morphological complexity is concentrated in the general sight-word list, whereas function words appear overwhelmingly in simple base forms (Table 4).

TABLE 4. Morphological Complexity Across the Two Lists

List	Total words	Polymorphic words (n)	Polymorphic words (%)	Monomorphic words (n)	Monomorphic words (%)	Polymorphic examples	Monomorphic examples
Sight words	195	77	39.5%	118	60.5%	البيت- كتابي- عليها	أشجار- في- إسم
Tool words	60	10	16.7%	50	83.3%	فهذا – إليه- فيها	عن – و - على

eISSN: 2550-2131
ISSN: 1675-8021

Linguistically, this pattern is expected. Tool words (e.g., و, في, من, إلى, ثم, أو, هل) belong to a closed grammatical class; they recur with high frequency and typically do not carry affixes, which explains their predominantly monomorphic shape in Table 2. The relatively few polymorphic tool words tend to be preposition + pronominal suffix strings (e.g., عليه, فيها, لها, لك) or particles with clitics, i.e., cases where morphology serves reference rather than lexical meaning.

In the broader sight-word list (Table 1), polymorphic forms are more common because it includes content vocabulary and affixed variants that reflect authentic usage. Typical patterns include definitized nouns (الـ + noun), pronominal suffixes attached to nouns or prepositions (ـي، ـها، ـهم), and prefixed verb forms (يُقال، تُدِلّ/تَدُلّ، أستخرج). These forms encode grammatical relations (definiteness, possession, agreement) and add semantic nuance, thereby increasing the morphological load learners encounter in real texts.

Moreover, analysis of orthographic characteristics across the 195 sight words stressed two prominent features of Arabic script complexity. Gemination (shadda) appeared in 49 words (25.1%), while hamza (ء) occurred in 41 words (21%). These results show that nearly half of the sight word list contains at least one of these orthographic features, demonstrating their central role in Arabic word recognition.

Both features increase the visual and phonological demands of reading, as learners must accurately process diacritic and non-linear elements embedded within the script. The high frequency of shadda highlights the importance of explicit instruction in recognizing consonant doubling, while the consistent appearance of hamza emphasizes its functional significance in distinguishing between lexical items. Together, these findings confirm that Arabic reading fluency requires sensitivity not only to letter forms but also to the orthographic markers that modify them.

LINGUISTIC ANALYSIS OF SIGHT WORDS

Comparison with the earlier K–3 list identified both continuity and variation across grade levels. A total of 33 words were retained from the original list, of which eight maintained identical frequency rankings. These include highly frequent function words such as "و" (and), "أم" (or), "لم" (did not), "الذي" (who/which), "ثم" (then), "في" (in), "إلى" (to), and the definite article "الـ." The stability of these items signals their ubiquity across Arabic texts and confirms the enduring role of tool words in supporting comprehension at multiple stages of literacy development.

The remaining retained words displayed shifts in frequency rankings. These included both tool words (e.g., "ما" (what), "من" (from), "لا" (no/not), "عن" (about)) and content words such as "العصافير" (birds) and "ذَهَبَ" (went). Morphemic units such as the prefix "فـ" (so) and the suffix "ها" (her) also appeared, reflecting the morphological productivity of Arabic and the importance of exposing students to affixed forms in addition to root words.

Morphological analysis further points out the prevalence of the definite article "الـ," which occurred in 36% of all words and in 57.3% of nouns. This high frequency reinforces the centrality of definite forms in Arabic and has implications for understanding orthographic and phonological rules, particularly in relation to sun and moon letters. Another notable feature is the appearance of the passive verb form "يُقال" (it is said), which reflects a more advanced level of syntactic and morphological development. Its presence in the upper elementary list aligns with expectations of increasing linguistic sophistication at these grade levels.

eISSN: 2550-2131
ISSN: 1675-8021

Orthographic features were also prominent. Forty-four words (22.5% of the list) contained at least one instance of shadda, compared to 40 words (20.5%) in the K–3 list (Oweini & Hazoury, 2010). The similarity across cycles indicates the consistent prevalence of gemination in Arabic and suggests that instruction in shadda should occur early in literacy acquisition.

Finally, the expanded list pointed out that 14% of the words belonged exclusively to Modern Standard Arabic (MSA) and did not have identical colloquial equivalents. Examples include "ترجّل" (to dismount/get off), which exists only in MSA. This finding attests to the persistent influence of diglossia in shaping reading acquisition, as learners must navigate forms that diverge significantly from their spoken lexicon.

COMPARATIVE ANALYSIS OF EXISTING WORD LISTS

Two comparative analyses were conducted to evaluate the degree of overlap between the current list and previously published resources, and to assess the potential for synthesizing these findings into a comprehensive reference for Arabic sight words.

The first comparison involved the list developed by Kozma and Yacoub (2019) for Grade 3 students in Lebanese schools. Results concluded that 23.5% of the words overlapped with the present Grades 4–6 list, either as exact matches or as near matches with minor morphological variations, such as the addition of the definite article "الـ" or attached pronouns. Examples of common words include "في" (in), "على" (on), "إلى" (to), "هذه" (this, feminine), "هذا" (this, masculine), "قبل" (before), "بعد" (after), "كيف" (how), and "مدرسة" (school). Strikingly, 77% of these shared items were tool words, including conjunctions, prepositions, adverbs, and the verb "كان" (to be), confirming the central role of function words in Arabic text comprehension. One notable omission was the conjunction "و" (and), which is the most frequent word in both the Oweini and Hazoury (2010) list and the current study. This absence may have been deliberate, as Kozma and Yacoub may have excluded "و" on the assumption that it functions as a morphemic prefix rather than a standalone word.

The second comparison was conducted with *A Frequency Dictionary of Arabic* (Buckwalter et al., 2011), which draws on a 30-million-word corpus. A 33% overlap was found between the first 100 entries of the dictionary and the Grades 4–6 list. Common items included high-frequency tool words such as "الـ" (the), "و" (and), "في" (in), "على" (on), "أنَّ" (that), and "كان" (was). These words consistently ranked among the most frequent in Arabic, reinforcing their indispensability in any sight word compilation.

Taken together, the comparisons demonstrate that while localized lists such as Kozma and Yacoub (2019) capture essential classroom vocabulary, and corpus-based dictionaries like Buckwalter et al. (2011) provide broad linguistic coverage, the present study contributes by bridging the two. The convergence across all three datasets confirms the dominance of tool words as a common denominator, thereby affirming their foundational role in developing a reliable and linguistically grounded Arabic sight list for elementary learners.

eISSN: 2550-2131
ISSN: 1675-8021

DISCUSSION

Reflecting on the current state and future of the Arabic language is a key concern for leading linguists, intellectuals, and writers. It is deeply tied to the cultural and intellectual identity of the Arab nation and to the preservation of its heritage and influence (Osama, 2023). Integrating a research-based sight word list into Arabic literacy instruction carries important implications for both linguistic theory and educational practice. From a linguistic perspective, such lists provide empirical evidence of high-frequency words that form the backbone of text comprehension. Their systematic inclusion in curricula aligns with the process of orthographic mapping, whereby spellings, pronunciations, and meanings become bonded in memory to support instantaneous recognition (Ehri, 2023). Because high-frequency items follow predictable frequency distributions and account for a disproportionate share of lexical coverage in texts, their instructional emphasis is supported by corpus-based evidence rather than intuition alone.

The expanded analyses presented in this study offer clearer insights for instructional design. The significant difference between function and content words indicates that teaching should emphasize automatic recognition of high-frequency function words—such as conjunctions, prepositions, and articles—that structure Arabic sentences and support comprehension. Since function words are fewer in number but occur most frequently, they should be introduced early and practiced repeatedly through contextualized reading and writing activities. Conversely, the larger pool of content words enriches vocabulary and comprehension; these words may be introduced progressively, with focus on meaning, morphology, and derivational patterns to strengthen lexical understanding. While these instructional approaches are supported by corpus evidence, they may improve fluency and comprehension, though their effectiveness ultimately depends on classroom implementation and learner variables.

The grammatical distribution analysis reinforces this distinction, showing that nouns and verbs dominate the corpus, reflecting the nominal and verbal structure of Arabic texts. Instructionally, this suggests that explicit teaching of noun and verb forms, supported by morphological awareness activities, may enhance students' ability to decode and comprehend authentic materials. This connection is further reinforced by the corpus finding that these grammatical categories collectively constitute a substantial portion of overall lexical coverage.

Similarly, the morphological complexity results—with 39.5% polymorphic words in the full list versus 16.7% in tool words—demonstrate the need to balance simplicity with exposure to affixed forms. Teachers can use polymorphic words to build morphological awareness by drawing attention to prefixes, suffixes, and root–pattern relationships (e.g., الـ, بـ, لـ, ـها, ـي). Such instruction helps learners generalize word formation rules and transfer this knowledge to unfamiliar vocabulary. These pedagogical recommendations are grounded in the corpus evidence showing that polymorphic forms contribute meaningfully to text coverage and therefore warrant strategic instructional attention.

The orthographic analysis, identifying frequent occurrence of shadda (25.1%) and hamza (21%), reinforces the importance of targeted instruction on these orthographic markers. Since these features modify word pronunciation and meaning, early and explicit practice in recognizing and writing them may reduce decoding errors and may improve fluency. Teachers can incorporate visual highlighting of these features in reading passages, dictation exercises, and spelling tasks to strengthen orthographic awareness. The high frequency of these orthographic features in the corpus supports the value of drawing learners' attention to them as part of literacy instruction.

eISSN: 2550-2131
ISSN: 1675-8021

CONCLUSION

These findings provide a data-based framework for sequencing reading instruction in Arabic. Frequency evidence can guide the order in which words are introduced; grammatical and morphological data can inform the complexity of learning materials; and orthographic patterns can shape decoding and spelling interventions. This integrated approach moves beyond rote memorization toward evidence-informed, cumulative literacy instruction, in which learners first master the most frequent and structurally essential words, then build toward morphologically and orthographically richer vocabulary. At the same time, it is important to acknowledge that the corpus is limited to private-school textbooks, which may not fully represent all instructional contexts, and that the sampling of pages—while systematic—captures only a portion of the complete texts. These methodological constraints should be considered when generalizing the findings.

Ultimately, a linguistically validated sight word list serves as both a theoretical and practical tool. It grounds curriculum development in corpus evidence, addresses the challenges posed by Arabic's orthography and diglossia, and provides a resource adaptable to diverse instructional contexts. By anchoring instruction in empirically verified high-frequency vocabulary, educators may foster more efficient reading acquisition, stronger decoding accuracy, and improved literacy outcomes among Arabic-speaking learners (Baharun, 2025). More broadly, the study demonstrates how corpus-based methods can enrich Arabic literacy research by offering replicable, data-driven insights into vocabulary patterns and instructional priorities.

LIMITATIONS AND FUTURE RECOMMENDATIONS

This study has several limitations that should be acknowledged. First, the number of textbooks analyzed was relatively limited, with a total of 12 volumes examined across three series. While these materials are representative of widely used curricula in Lebanese private schools, expanding the corpus to include additional resources such as trade books, children's magazines, and other extracurricular reading materials would yield a more comprehensive view of children's exposure to high-frequency words.

Second, the study relied on manual word counting, a process that was both time-consuming and prone to human error. Although digital formats of some textbooks were available, existing software tools were not well suited to accurately identifying words in Arabic, particularly given the complexities of orthography, diacritics, and affixation. This limitation necessitates technological advances in Arabic corpus linguistics, including the development of automated tools capable of detecting and categorizing word forms with greater accuracy.

Third, the scope of the study was restricted to Grades 4-6. While this choice aligns with the objective of extending Oweini and Hazoury's (2010) K–3 list, it also limits the applicability of findings to middle and secondary school learners. Older students encounter increasingly complex vocabulary, including words shaped by technological, social, and cultural trends, which were beyond the scope of the present analysis.

These limitations point to several directions for future research. One promising avenue involves the development of readability formulas based on frequency data, which could help determine grade-level text difficulty. Expanding the scope of analysis to include middle school and high school grades would also provide insights into how frequency and morphology interact

eISSN: 2550-2131
ISSN: 1675-8021

with adolescent literacy. Additionally, the creation of standardized reading and spelling tests grounded in frequency-based word lists could support both research and assessment in Arabic literacy. Finally, greater attention to grammatical and morphological features, alongside the use of advanced digital tools for automated analysis, would enhance both efficiency and accuracy in compiling future lists.

By addressing these concerns, subsequent research can build on the present findings to produce more robust, scalable, and linguistically precise insights into Arabic sight words and their role in literacy development.

REFERENCES

Abu-Rabia, S., & Darawshe, E. (2024). Evaluation of the multiple-deficit hypothesis among dyslexic Arabic-speaking children. *Dyslexia, 30*(2), e1759. https://doi.org/10.1002/dys.1759

Abu-Rabia, S., & Hijjazi, H. (2023). The role of vowelization in reading comprehension of different Arabic genres. *Applied Psycholinguistics, 44*(3), 567–587. https://doi.org/10.1017/S0142716423000123

Alabdulkader, B., Alshubaili, H., & Alhashmi, A. (2021). Challenges in reading Arabic among children with dyslexia. *Optometry and Vision Science, 98*(8), 929–935. https://doi.org/10.1097/OPX.0000000000001744

Al Ghanem, R., & Sanders, E. A. (2023). The impact of a structured sight word intervention on Arabic reading fluency in early elementary students. *Journal of Educational Psychology, 115*(4), 621-638. https://doi.org/10.1037/edu0000789

Asadi, I. A., & Kawar, K. (2023). Learning to read in Arabic diglossia: The relation of spoken and Standard Arabic language in kindergarten to reading skills in first grade. *Literacy Research and Instruction.* https://doi.org/10.1080/19388071.2023.2217274

Asadi, I. A., Asli-Badarneh, A., & Vaknin-Nusbaum, V. (2024). The impact of morphological density on reading comprehension in Arabic: A comparison between typically developing children and children with reading disabilities. *Dyslexia, 30*(1), 17-61. https://doi.org/10.1002/dys.1761

Aswadi, D. (2024). Optimizing Language Learning with Relevant and Contextual Vocabulary Selection in Education. *Journal on Education, 7*(1), 8615–8625. https://doi.org/10.31004/joe.v7i1.7714

Baharun, S. (2025). Innovative Approaches to Teaching Arabic Vocabulary to Novice Learners. *Lahjatuna.*, *4*(2), 77–88. https://doi.org/10.38073/lahjatuna.v4i2.2535

Ballance, O. J., & Coxhead, A. (2024). Corpus Analysis of Vocabulary. *The Encyclopedia of Applied Linguistics*, 1–7. https://doi.org/10.1002/9781405198431.wbeal20508

Buckwalter, T., & Parkinson, D. (2011). A frequency dictionary of Arabic: Core vocabulary for learners. Routledge.

Darwiche Fedda, O., & Oweini, A. (2012). The effect of diglossia on Arabic vocabulary development in Lebanese students. *Educational Research and Reviews, 7*(16), 351–361. https://doi.org/10.5897/ERR11.022

Ehri, L. C. (2023). Roads travelled researching how children learn to read words. *Australian Journal of Learning Difficulties, 28*(1), 55–71. https://doi.org/10.1080/19404158.2023.2208164

Franceschi, D. (2021). Lexical-Semantic Representations at the Time of the Coronavirus Pandemic. *International Journal of English Linguistics*, *11*(6), 120. https://doi.org/10.5539/IJEL.V11N6P120

Fry, E. (1980). The new instant word list. *The Reading Teacher, 34*(3), 284–289.

Galicia, A. B. (2022). An examination of teacher beliefs about the emphasis of sight word development in early literacy instruction.

Hassanein, E., Johnson, E., Ibrahim, S., & Alshaboul, Y. (2023). What predicts word reading in Arabic? *Frontiers in Psychology, 14.* https://doi.org/10.3389/fpsyg.2023.1077643

Hassanein, K. M., Elbeheri, G., Almalki, M., & Alsadoon, E. A. (2023). What predicts word reading in Arabic? *Journal of Literacy Research, 55*(2), 123–141. https://doi.org/10.1177/1086296X231012345

Hutchison, L., Jerasa, S., Ahmmed, R., & Holcomb, E. (2024). Reexamining the Dolch Basic Sight Word List: Contemporary considerations for culturally sustaining approaches to assess sight word development. *Literacy Research and Instruction, 1–23.* https://doi.org/10.1080/19388071.2024.2321209

Kim, Y.-S. G., Wagner, R. K., & Lopez, D. (2023). The role of word reading and listening comprehension in reading comprehension: A longitudinal study in a language-diverse context. *Journal of Educational Psychology, 115*(1), 85–100. https://doi.org/10.1037/edu0000720

Kozma, E., & Yaakoub, G. (2019). *Latest strategies for reading and literacy for the Arab child.* Dar Al-Nahda Al-Arabia.

Lazareva, M., & Terekhin, I. N. (2023). On the influence of the Internet on the contemporary English. *Učënye Zapiski Sankt-Peterburgskogo Universiteta Tehnologij Upravleniâ i Èkonomiki*, *4*, 299–305. https://doi.org/10.35854/2541-8106-2022-4-299-305

Liu, L., Gong, T., Shi, J., & Guo, Y. (2024). A high-frequency sense list. *Frontiers in Psychology*, *15*. https://doi.org/10.3389/fpsyg.2024.1430060

Ministère de l'éducation nationale, de l'enseignement supérieur et de la recherche. (2023). *Liste de fréquence lexicale*. EDUSCOL. Retrieved from https://eduscol.education.fr/186/liste-de-frequence-lexicale

Moser, J. (2020). Evaluating Arabic textbooks: A corpus-based lexical frequency study. *International Journal of Applied Linguistics, 30*(4), 567–588. https://doi.org/10.1111/ijal.12321

Osama, M. A. (2023). Ways to enhance the position of the arabic language between originality and challenges an analytical. *Journal of Language Studies, 6*(2), 224–239. https://doi.org/10.25130/jls.6.3.2.18

Oweini, A., & Lotfi ElZein, H. (2022). Dyslexia in Lebanon. In G. Elbeheri & S. Lee (Eds.), *The Routledge International Handbook of Dyslexia in Education* (pp. 199–208). Routledge.

Oweini, A., & Hazoury, K. (2010). Towards a list of sight word lists in Arabic. *International Review of Education, 56*(4), 457–478.

Oweini, A., Awada, G., & Kaissi, F. (2020). The impact of Arabic diglossia on oral language development. *GEMA Online Journal of Language Studies, 20*(2).

Petscher, Y., Stanley, C., & Pentimonti, J. (2022). The science of reading movement: The role of assessment to inform instruction. *Reading Research Quarterly, 57*(S1), S443–S457. https://doi.org/10.1002/rrq.438

Saiegh-Haddad, E., & Everatt, J. (2015). Early literacy education in Arabic. In N. Kucirkova, C. Snow, V. Grover, & C. McBride-Chang (Eds.), *The Routledge International Handbook of Early Literacy Education* (pp. 185–198). Taylor & Francis.

Saiegh-Haddad, E., & Joshi, R. M. (Eds.). (2014). *Handbook of Arabic literacy: Insights and perspectives.* Springer.

Saiegh-Haddad, E., & Schiff, R. (2016). The impact of diglossia on voweled and unvoweled word reading in Arabic: A developmental study from childhood to adolescence. *Scientific Studies of Reading, 20*(4), 311–326.

Saiegh-Haddad, E., & Spolsky, B. (2014). Acquiring literacy in a diglossic context: Problems and prospects. In E. Saiegh-Haddad & M. Joshi (Eds.), *Handbook of Arabic literacy: Insights and perspectives* (pp. 225–240). Springer.

Shendy, R. (2022). Learning to read in an "estranged" language: Arabic diglossia, child literacy, and the case for mother tongue-based education. *Creative Education, 13*(4), 1015–1029. https://doi.org/10.4236/ce.2022.134077

Ziegler, J. C., & Goswami, U. (2005). Reading Acquisition, Developmental Dyslexia, and Skilled Reading across Languages: A Psycholinguistic Grain Size Theory. Psychological Bulletin, 131, 3-29. http://dx.doi.org/10.1037/0033-2909.131.1.3

ABOUT THE AUTHORS

Ahmad Oweini, Corresponding author (aoueini@lau.edu.lb). Associate professor of education, researcher, licensed educational psychologist, teacher trainer, special educator, and psychoeducational examiner. His research interests include diglossia, dyslexia, bilingual language development, integration of technology in the classroom and teacher training programs.

Noura El-Far, Co-author (noura.elfar@lau.edu). Early childhood educator; homeroom assistant teacher. Recently graduated with a B.A and T.D in early childhood education from the Lebanese American University. Currently works in a private school in Lebanon as a preschool teacher.

eISSN: 2550-2131
ISSN: 1675-8021