International Journal of Innovative Technologies in Social Science

e-ISSN: 2544-9435

Operating Publisher
SciFormat Publishing Inc.
ISNI: 0000 0005 1449 8214

2734 17 Avenue SW,
Calgary, Alberta, T3E0A7,
Canada
+15878858911
editorial-office@sciformat.ca

ARTICLE TITLE	CRSTDLA-ARAFREQ: AN ARABIC LEXICAL DATABASE FOR SPEECH-LANGUAGE ASSESSMENT IN ALGERIAN PRIMARY SCHOOLS
DOI	https://doi.org/10.31435/ijitss.1(49).2026.4975
RECEIVED	14 January 2025
ACCEPTED	10 April 2025
PUBLISHED	20 January 2026
LICENSE	The article is licensed under a **Creative Commons Attribution 4.0 International License.**

CRSTDLA-ARAFREQ: AN ARABIC LEXICAL DATABASE FOR SPEECH-LANGUAGE ASSESSMENT IN ALGERIAN PRIMARY SCHOOLS

Kahina Lettad
Scientific and Technical Research Centre for the Development of the Arabic Language, Algeria

Amina Saadedine
Scientific and Technical Research Centre for the Development of the Arabic Language, Algeria

Fouzia Badaoui
Scientific and Technical Research Centre for the Development of the Arabic Language, Algeria

Mahdia Benaissa
Scientific and Technical Research Centre for the Development of the Arabic Language, Algeria

Radia Baba
Scientific and Technical Research Centre for the Development of the Arabic Language, Algeria

Nouara Badi
Cheikh Mohamed Bachir el Ibrahimi Higher Teacher Training College, Kouba, Algiers, Algeria

Ratiba Khorta
University of Algiers 2 – Abu El Kacem Saad Allah, Algeria

Asma Benchouk
Speech Therapist, University of Algiers 2 – Abu El Kacem Saad Allah, Algeria

ABSTRACT

Lexical databases are increasingly recognized as indispensable resources for both speech-language pathology and education. This paper presents CRSTDLA-Arafreq, the first Modern Standard Arabic lexical database specifically constructed from Algerian textbooks (Preschool to Grade 3), integrating written and oral frequencies as well as student productions. While such tools enable clinicians to design sensitive assessments and support teachers in selecting age-appropriate materials, Algeria has lacked a reference database tailored to its specific school and clinical contexts.

To address this gap, the CRSTDLA-Arafreq project was launched at the Scientific and Technical Research Center for the Development of the Arabic Language. Conducted over three years by a multidisciplinary team, the project provides a robust, evidence-based lexical resource grounded in authentic Algerian primary school corpora to support inclusive education and validated diagnostic instruments.

KEYWORDS

Lexical Databases, Modern Standard Arabic, Speech-Language Assessment, Primary Education, Educational Corpora, CRSTDLA-Arafreq- Scientific and Technical Research Center for The Development of The Arabic Language

CITATION

Kahina Lettad, Amina Saadedine, Fouzia Badaoui, Mahdia Benaissa, Radia Baba, Nouara Badi, Ratiba Khorta, Asma Benchouk. (2026) CRSTDLA-Arafreq: An Arabic Lexical Database for Speech-Language Assessment in Algerian Primary Schools. *International Journal of Innovative Technologies in Social Science*. 1(49). doi: 10.31435/ijitss.1(49).2026.4975

1. Introduction

Lexical databases are now recognized as essential tools in applied linguistic research. By providing structured repositories of linguistic data, they enable a systematic description and analysis of language across multiple dimensions. In education and speech-language pathology, their value is particularly significant: they offer access to scientifically validated material that supports the design of reliable diagnostic tools and pedagogical resources adapted to learners' needs.

In Algeria, and more generally in the Arab world, the lack of such lexical databases represents a major challenge. Clinical assessments of language disorders often rely on imported tools developed for other languages and cultural contexts, which limits both their validity and their applicability. Likewise, the creation of educational content is frequently based on intuition rather than on objective data regarding the frequency, complexity, or structural properties of Arabic words.

Against this backdrop, this study addresses two operational research questions:

1. How are lexical frequencies and morphological patterns distributed across Algerian textbooks and student productions for the 5–9 age group, and how do these patterns evolve across grade levels?
2. To what extent can this context-specific lexical distribution serve as a calibrated normative basis for designing more accurate speech-language assessment tools and age-appropriate pedagogical materials?

To answer these questions, this study introduces the *CRSTDLA-Arafreq* database. This tool constitutes the first lexical repository in Modern Standard Arabic specifically derived from the Algerian educational ecosystem, covering levels from preschool to the 3rd year of primary school. By triangulating data from textbooks, oral school discourse, and written productions by the students themselves, *CRSTDLA-Arafreq* provides a unique developmental perspective on the Arabic lexicon in the Algerian school context.

2. Problem Statement and Challenges

The central problem lies in the critical lack of linguistic resources tailored to Arabic, and more specifically to the school-based Arabic used by children. While many countries have long benefited from lexical databases that support research, clinical practice, and education, the Arab world still lags behind in this field.

The consequences of this gap are manifold:

In speech-language pathology, the absence of adapted tests prevents reliable diagnosis of disorders such as dyslexia and dysorthography.

In education, the lack of objective data hampers the design of textbooks and exercises that are progressive and aligned with pupils' actual age and linguistic level.

At the social level, these shortcomings contribute to school failure and dropout, thereby exacerbating educational inequalities.

3. Objectives

The *CRSTDLA-Arafreq* project seeks to address the lack of lexical resources in Modern Standard Arabic by developing a database specifically designed for Algerian primary school contexts. Its objectives are threefold:

Clinical Objective: To provide speech-language pathologists with a reliable lexical resource that enables the design of standardized, culturally relevant diagnostic tools for identifying language and learning disorders such as dyslexia and dysorthography.

Educational Objective: To offer teachers and curriculum designers objective data—such as word frequency, morphological complexity, and syllabic structure—that can guide the development of pedagogical materials adapted to learners' age and linguistic level.

Scientific Objective: To establish a reference corpus of school-based Arabic that can serve as a foundation for linguistic, psycholinguistic, and educational research, while fostering innovation in digital tools for language learning and clinical practice.

By achieving these objectives, the project aspires to support more inclusive education, improve clinical assessment, and contribute to reducing inequalities in language learning and literacy acquisition.

4. The Theoretical Framework and State of the Art

4.1. International Lexical Databases

Lexical databases have been available for several decades in languages such as English, French, German, and Spanish, where they are widely used in research, education, and speech-language pathology to analyze word frequency, morphological structure, and lexical usage. In contrast, Arabic remains under-resourced, with existing initiatives being limited and fragmented, often focusing on literary or press corpora rather than on children's language and school-based usage.

Most lexical databases are derived from adult corpora, despite the fact that children represent a central population in educational and clinical contexts. Developmentally appropriate assessment therefore requires child-oriented oral and written corpora. The German database ChildLex (Schroeder et al., 2015), based on a large, annotated corpus of children's literature, illustrates the value of such resources by providing reliable linguistic norms for children aged 6 to 12.

The importance of lexical databases in applied contexts has been well established. Paul Nation (2001) demonstrated that frequency-based resources support the graded selection of linguistic items for assessment and instruction, while Joan Bybee (2006) showed that frequency of use is a strong predictor of lexical acquisition. These findings underscore the relevance of frequency-based databases for sensitive and valid language assessment.

Several languages have developed robust child-focused lexical resources using rigorous methodologies. The French database Manulex, derived from primary school corpora, and the Spanish database EsPal, which provides extensive psycholinguistic norms, exemplify this trend. In the Italian context, the DILLo database—DILLo—was specifically designed for speech-language pathology and allows clinicians to generate customized word lists based on multiple linguistic criteria.

Despite this international progress, comparable resources remain scarce for Arabic, particularly those targeting children's school-based language. Within this context, CRSTDLA-Arafreq constitutes a significant contribution by addressing a critical gap in Arabic lexical resources and aligning with international standards in educational and clinical linguistics.

4.2 Comparative Analysis of Child Lexical Databases

The table 1 highlights both convergences and divergences among major child-oriented lexical databases and clearly situates CRSTDLA-Arafreq within the international landscape while underscoring its specific contributions.

In terms of language and educational scope, databases such as Manulex, ChildLex, and DILLo primarily target primary school pupils and are grounded in school-based materials or children's literature. CRSTDLA-Arafreq extends this scope by covering a younger population, from preschool to Grade 3, which is particularly relevant for early identification, assessment, and intervention in language and literacy disorders.

Regarding source materials, most international databases rely mainly on textbooks or written child-directed texts. In contrast, CRSTDLA-Arafreq adopts a more comprehensive corpus design by integrating textbooks, oral language data, and children's written productions. This triangulated approach enhances ecological validity and provides a more faithful representation of children's actual language use in school contexts.

With respect to linguistic annotations, existing databases predominantly focus on frequency measures, orthographic or morphological information, and general psycholinguistic variables, as illustrated by EsPal. CRSTDLA-Arafreq, however, incorporates annotation layers that are specifically tailored to the structural properties of Arabic, including roots, patterns (schemes), syllabication, and phonetic transcription. These features are central to Arabic lexical organization and are crucial for analyzing reading and spelling acquisition, particularly in clinical contexts.

Overall, the comparative analysis demonstrates that CRSTDLA-Arafreq does not merely replicate existing international models but rather adapts and extends them to meet the linguistic, educational, and clinical specificities of Arabic. As such, it fills a significant gap in child-focused Arabic lexical resources while aligning with internationally established standards in educational and clinical linguistics.

Table 1. Comparison of CRSTDLA-Arafreq with international children's lexical databases

Database	Language	Target Population	Source Material	Key Annotations
Manulex	French	Grades 1–5	Textbooks	Frequency, orthographic neighbors, phonology.
ChildLex	German	Grades 1–6	Children's books	Frequency, lemma-based statistics, POS tagging.
EsPal	Spanish	Adults/Children	Diverse corpora	Subtitles, literature, morphological structure.
DILLo	Dutch	Grades 1–6	Textbooks & media	Frequency, age of acquisition, semantics.
CRSTDLA-Arafreq	Arabic (MSA)	Preschool–Grade 3	Textbooks + Oral + Student Writing	Roots, Patterns (Schemes), Syllabication, Phonetic transcription.

4.3. Limitations for Arabic

The Arabic language presents a set of linguistic and technological specificities that make the construction of reliable and pedagogically relevant lexical databases particularly challenging:

- Complex derivational morphology: Arabic relies on a root-and-pattern morphological system, where triliteral or quadriliteral roots are embedded into patterns to generate a wide range of derived forms. These forms are further enriched with affixes, proclitics, and enclitics, which are often fused into the written word, substantially increasing lexical diversity and the complexity of computational processing.
- Orthographic ambiguity: the omission of short vowels and diacritics in standard Arabic writing leads to significant morphosyntactic ambiguities. For instance, the string علم can be interpreted differently depending on context: /ʾalam/ ("flag"), /ʾilm/ ("science"), /ʿallama/ ("to teach"), or /ʾulima/ ("was learned").
- Technological limitations of OCR systems: current Optical Character Recognition (OCR) software remains underperforming for Arabic. These systems struggle to accurately process Arabic's cursive script, ligatures, and diacritic positioning, making the digitization and automatic transcription of school textbooks and other pedagogical materials particularly difficult. This limitation hinders the creation of comprehensive and machine-readable textual corpora.

Several existing lexical databases attempt to address these challenges, yet none of them are specifically tailored to school-related vocabulary or to the needs of speech-language pathology:

Aralex: a lexical database for Modern Standard Arabic, developed from a 40-million-word contemporary corpus. It provides token frequencies, patterns, n-grams, and is freely available under a GNU-like license (Boudelaa & Marslen-Wilson, 2010).

Kalmasoft: a platform that develops specialized databases for Natural Language Processing (NLP), including morphological lexicons (roots, inflected forms), loanword data, and resources for building morphological analyzers and POS taggers (Kalmasoft, 2025).

ArabLEX (Database of Arabic General Vocabulary – DAG): an extensive resource containing nearly 88 million word forms and more than 30,000 lemmas, with phonemic transcriptions, full diacritization, orthographic variants, and rich morpho-phonological annotations. While highly valuable for NLP applications, its size, scope, and cost make it less suited for school-oriented or clinical purposes.

While these resources provide valuable linguistic data, they fall short in addressing the specific needs of speech-language pathology and education in Arabic-speaking contexts. Aralex offers robust frequency information and structural patterns but is built from general adult corpora, which limits its relevance for child-centered assessments. Kalmasoft, although rich in morphological tools for computational linguistics, is designed primarily for NLP applications rather than for clinical or pedagogical use. ArabLEX, with its extensive coverage and advanced annotations, is a powerful research tool, yet its large scale and high cost make it inaccessible and impractical for the design of diagnostic materials or age-appropriate pedagogical resources.

In contrast, the *CRSTDLA-Arafreq* project distinguishes itself by focusing explicitly on school-related lexical data and by integrating clinical, educational, and computational perspectives within a multidisciplinary team. Its design emphasizes age-appropriate frequency norms, morphological and phonological annotations, and usability for both clinicians and educators. By situating itself at the intersection of speech-language pathology, pedagogy, and language technology, *CRSTDLA-Arafreq* aims to fill a critical gap and to provide a scientifically validated, accessible, and context-sensitive lexical database for Arabic.

5. Methodology and Description of the *CRSTDLA-Arafreq* Database

The *CRSTDLA-Arafreq* lexical database represents an innovative and pioneering resource for the analysis of lexical frequency in Algerian primary school textbooks. It was developed through a rigorous multi-step methodological process designed to ensure both the reliability and representativeness of the data collected.

The corpus consists of the official textbooks used from preschool to the third year of primary education (ages 5 to 9), thus covering a critical period of linguistic and cognitive development. By Grade 3, most children have developed the foundations of morphological processing, including the recognition of roots, derivational patterns, and inflectional morphology. This is particularly relevant for Arabic, a morphologically rich and root-based language, where the ability to manipulate roots and patterns is essential for reading fluency, spelling, and written expression.

Furthermore, difficulties in lexical acquisition or morphological awareness at this stage are strong predictors of later reading and writing disorders, including dyslexia and dysorthographia (Carlisle, 2000; Kirby et al., 2012). Hence, early identification and support between ages 5 and 9 are crucial for effective intervention.

For these reasons, the *CRSTDLA-Arafreq* database deliberately targets this developmental window, ensuring that the lexical and morphological data collected are both pedagogically relevant and clinically meaningful for speech-language pathology and early literacy instruction.

Each word was extracted, classified, and analyzed with regard to its morphological structure (root, pattern, affixes), frequency of occurrence, syllabic segmentation, and phonetic transcription.

The primary aim of this resource is to provide a reference tool that allows for a precise evaluation of children's lexical exposure and that informs pedagogical practices in the teaching of Arabic. At the same time, the database offers a robust empirical foundation for the development of diagnostic tests in speech-language pathology, where the selection of validated linguistic items is essential to achieve reliable assessments.

By analyzing both the frequency and distribution of words within the Algerian school context, *CRSTDLA-Arafreq* contributes to a deeper understanding of lexical development trajectories in Arabic-speaking children and provides new perspectives on educational and clinical practices. Thus, it goes beyond the function of a simple word list and constitutes a strategic scientific infrastructure at the intersection of linguistics, pedagogy, and speech-language pathology.

6. Methodological Design of the *CRSTDLA-Arafreq* Database

CRSTDLA-Arafreq is an original computerized lexical resource, designed to analyze and document the lexical exposure of Algerian schoolchildren between the ages of 5 and 9 (preschool to Grade 3). The database includes several key fields, such as graphic word form, frequency of occurrence, oral frequency, and written frequency.

6.1. Development Stages of the Database

a) Corpus constitution

An exhaustive corpus was compiled, encompassing the entire set of official textbooks used in Algerian primary schools from preschool to Grade 3. This corpus forms the foundation of the database and faithfully reflects the linguistic environment to which children are exposed.

b) Frequency analysis based on images

The database integrates an innovative approach by exploiting illustrations with multiple possible labels, frequently present in school textbooks. These images allow the collection of different correct denominations for the same concept, thereby reflecting the lexical variability among learners. The frequency of each denomination is measured, providing insights into early vocabulary development and the links between visual representations and word retrieval.

c) Analysis of written productions

Children's written compositions were collected and analyzed according to themes derived from the textbooks. This process highlights the frequency with which words are retrieved in production tasks, offering an additional perspective on how children internalize and mobilize vocabulary learned in class.

6.2. Educational Levels Covered

The analysis covers four key educational levels:
Preschool (5 years)
Grade 1 (6 years)
Grade 2 (7 years)
Grade 3 (8 years)
Grade 4 (9 years)

This developmental window is crucial as it represents the transition from emergent literacy to vocabulary enrichment, in line with the framework proposed by Adams (1990). Across these levels, textbooks share similar thematic content but introduce an increasingly rich lexicon, which makes it possible to closely track the impact of pedagogical progression on lexical acquisition.

7- *CRSTDLA-Arafreq* Survey Stages

7.1. Stage 1 – Corpus Collection

The *CRSTDLA-Arafreq* corpus was compiled from 19 Arabic-language school textbooks covering several subjects (reading, science, Islamic studies, civic education, history, geography, mathematics). These textbooks targeted levels ranging from preschool to Grade 3, distributed as follows:

2 books for preschool
4 books for Grade 1
4 books for Grade 2
9 books for Grade 3

Because no reliable OCR system exists for Arabic, all texts were manually transcribed and entered into a specialized software program in order to analyze terminological redundancy and calculate lexical frequency.

Frequency analysis, a central concept in corpus linguistics, makes it possible to evaluate the accessibility, versatility, and conceptual complexity of words (Nagy & Anderson, 1984). These data are crucial for adapting teaching materials to learners' abilities and for selecting relevant vocabulary in speech-language assessments.

7.2. Stage 2 – Frequency Analysis through Picture Naming

A set of 200 recurring illustrations in the textbooks was selected (objects, places, foods, everyday life scenes). Each image could elicit multiple lexical denominations. For example, a picture of a house might be named as follows: baytun «بيت», dārun «دار», manzilun «منزل».

Methodology: individual presentation of pictures to students, who were asked to spontaneously name what they saw.

Sample: 640 students, distributed by grade level (CP, G1, G2, G3) and by region (East, Center, West: Annaba, Algiers, Tlemcen, Mostaganem, Bouira, Béjaïa, Ouargla, south: ouargla).

This approach identified dominant lexical preferences and also highlighted regional variations in a multilingual context. It thus provides a valuable sociolinguistic map for unerstanding the dynamics of lexical acquisition in Standard Arabic.

7.3. Stage 3 – Analysis of Written Compositions

To assess vocabulary use in productive situations, eight linguistic integration tasks were designed based on the themes of Algerian primary school textbooks. These tasks aimed to stimulate autonomous written production while covering the full lexical repertoire prescribed in the national curriculum.

Each task corresponded to a specific lexical field and reflected realities close to the child's sociocultural environment. This design ensured a balanced representation of the prescribed semantic fields, allowing us to measure both the richness of students' vocabulary use and its adequacy to the curriculum.

Sample: 100 Grade 3 students from several regions (Annaba, Algiers, Tlemcen, Mostaganem, Bouira, Béjaïa).

Procedure: each student produced 8 texts (total = 800 written productions), which were manually transcribed and analyzed with lexical software.

Analysis: word frequency was calculated using the following formula, which evaluates both redundancy and lexical diversity:

$$\text{Frequency(w)} = \frac{\text{Occurrences of word (or n} - \text{gram)w}}{\text{Total number of n} - \text{grams in the corpus}}$$

The results show that integration tasks foster vocabulary enrichment and improve writing skills, while highlighting pedagogical areas for improvement (explicit teaching of vocabulary, consolidation of acquired knowledge, development of written expression).

8. Contributions of the Methodology

The combination of the three stages (corpus → picture naming → written productions) provides the *CRSTDLA-Arafreq* database with a unique methodological robustness. This approach makes it possible to:

- Identify lexical frequency patterns in school textbooks,
- Integrate regional sociolinguistic variations,
- Assess the actual mobilization of vocabulary by students.

This integrative methodology establishes *CRSTDLA-Arafreq* as a scientific and pedagogical reference tool, useful both for linguistic research and for clinical practice in speech-language pathology as well as for the didactics of Arabic.

Through this method, we were able to construct a precise lexical profile of students' written productions, identifying the most frequently used words while also evaluating lexical diversity. The analysis highlighted the impact of integration tasks on vocabulary enrichment and on the improvement of writing skills.

Furthermore, this stage allowed us to outline pedagogical directions for improvement, particularly regarding the explicit teaching of vocabulary, the consolidation of linguistic knowledge, and the strengthening of written expression in the school context.

9. Results

The development of the lexical database for school-aged children was structured around three major methodological stages, each contributing to the reliability and representativeness of linguistic data drawn from the Algerian educational context. The project pursues a dual ambition: to make a significant contribution to applied linguistic research and to improve the teaching of Arabic—particularly in the field of speech-language pathology—by providing reliable indicators for assessing children's lexical and morphological skills.

The outcome of this work is a structured digital lexical database that integrates detailed morpho-lexical information. For each entry, multiple descriptors are encoded, including:

Standard orthographic form (unvocalized, as it appears in textbooks)

Vocalized form (with diacritics, useful for phonological and speech-language applications)

Syllable count, as an indicator of phonological complexity

Grammatical gender (masculine/feminine)

Derivational morphology: prefixes, suffixes, triliteral or quadriliteral root, and morphological pattern (e.g., faʿala, mafʿūl, tafrīl)

Oral frequency (from the picture-naming task)

Written frequency (from students' written compositions)

Overall orthographic frequency (derived from textual data)

The *CRSTDLA-Arafreq* database contains a total of 3472 unvocalized words (as they typically appear in school textbooks) and 3599 fully vocalized words (with diacritics), providing a balanced representation for educational and clinical use. The database encompasses 1,134 unique roots and 1,506 distinct morphological patterns (awzan), reflecting the rich structure of the Arabic language.

Words are distributed across various grammatical types, highlighting its linguistic diversity: singular masculine words account for 1162 entries, singular feminine for 618, present tense verbs for 268, broken plurals for 267, and past tense verbs for 124. This comprehensive coverage makes *CRSTDLA-Arafreq* particularly suitable for linguistic research, reading and spelling studies, and educational applications.

Beyond its descriptive scope, this database offers a valuable contribution to understanding lexical and morphological acquisition processes. It also provides a solid reference for longitudinal research on the development of linguistic skills in Arabic.

Furthermore, *CRSTDLA-Arafreq* opens new perspectives for improving Arabic vocabulary instruction, addressing documented gaps in various educational contexts.

In sum, the *CRSTDLA-Arafreq* project represents a strategic scientific and pedagogical innovation, poised to transform how Arabic is taught, assessed, and rehabilitated within the Algerian context.

538	ID	كلمات دون التشكيل	كلمات مشكلة	مقاطع	النوع	الجذر	السابقة	اللاحقة	الوزن	النسخ الصوتي	تواتر لفظي تحضيري	تواتر لفظي 1	تواتر لفظي 2	تواتر لفظي3	تواتر كتابي 3	تواتر ش الإملائي	تواتر ش الإملائي	تواتر ش الإملائي	تواتر ش الإملائي 3
539	535	باخرة	بَاخِرَةٌ	4	مفرد مؤنث	بخر		ة	فَاعِلَةٌ	biāxiratun	11,1111	12,0968	12,3529	17,7419			0,15361	0,01087	
540	536	بارد	بَارِدٌ	3	مفرد مذكر	برد			فَاعِلٌ	bāridun	0,6944	5,6452	1,1765	2,6882			0,15361	0,01087	
541	537	بائع	بَائِعٌ	3	مفرد مذكر	بيع			فَاعِلٌ	bāʾiʿun	25,0000	25,0000	1,1765	53,2258			0,15361	0,01087	
542	538	بائعة	بَائِعَةٌ	4	مفرد مؤنث	بيع		ة	فَاعِلَةٌ	bāʾiʿatun	0,6944	0,0000					0,15361	0,01087	
543	539	بائعات	بَائِعَاتٌ	4	جمع مؤنث	بيع		ات	فَاعِلَاتٌ	bāʾiʿāt		0,0000	2,9412	0,5376			0,15361	0,01087	
544	540	بائعون	بَائِعُونَ	4	جمع مذكر	بيع		ون	فَاعِلُونَ	bāʾiʿūn		0,0000	0,5882	4,3011			0,15361	0,01087	
545	541	بائعين	بَائِعِينَ	4	جمع مذكر	بيع		ين	فَاعِلِينَ	bāʾiʿaīn		0,8065	0,5882	1,0753			0,15361	0,01087	
546	543	بحار	بَحَّارٌ	3	مفرد مذكر	بحر			فَعَّالٌ	baḥḥārun	1,3889	6,4516	8,2353	4,8387			0,15361	0,01087	
547	544	بحر	بَحْرٌ	2	مفرد مذكر	بحر			فَعْلٌ	baḥrun	10,4167	16,9355	23,5294	8,6022			0,15361	0,01087	
548	545	بحيرة	بُحَيْرَةٌ	4	مفرد مذكر	بحر		ة	فُعَيْلَةٌ	buḥayratun	9,7222	8,0645	6,4706	13,4409			0,15361	0,01087	

Fig. 1. *Illustration of the CRSTDLA-Arafreq database corpus*

10. Discussion and Conclusion

The *CRSTDLA-Arafreq* database represents a pioneering initiative in the field of Arabic language resources, addressing a long-standing gap in both educational and clinical contexts. Unlike existing lexical databases—often designed for computational linguistics or broad language modeling—*CRSTDLA-Arafreq* focuses specifically on the school-aged population and on the educational materials that shape their linguistic exposure. This focus ensures that the data collected are directly relevant to classroom practice, speech-language pathology, and child language research.

From a scientific perspective, the integration of multiple data sources—school textbooks, picture-naming tasks, and children's written productions—provides a robust methodological framework. This triangulation not only enhances the reliability of frequency measures but also offers unique insights into the interplay between lexical input (exposure through manuals) and lexical output (children's active use of vocabulary). The lexical gap observed in the CRSTDLA-Arafreq database highlights a lexical competition between the formal vocabulary of textbooks and the learners' accessible lexicon during picture-naming tasks. Even when specific items (e.g., ʾibhāmun) show high curricular frequency, their low production rate in response to visual stimuli suggests a fragile referential anchoring.

This phenomenon stems from the presence of 'competing signifiers': when faced with a polysemous visual stimulus, learners tend to favor terms that offer greater cognitive economy or those rooted in their vernacular register, rather than using precise academic language. Consequently, the lexical gap index measures more than just a lack of knowledge; it tracks the extent to which academic vocabulary has—or has not—attained dominance over spontaneous speech.

For clinicians, a significant negative gap on images with multiple potential labels serves as a tool to evaluate the child's ability to inhibit lexical distractors. An inability to retrieve the textbook-sanctioned term despite repeated exposure becomes a vital indicator for identifying lexical access deficits or poor lexical organization—both of which are core markers in the early diagnosis of dyslexia and dysorthography.

From a clinical perspective, the database constitutes an essential tool for developing standardized diagnostic tests in Arabic. By providing reliable indicators of lexical frequency, phonological complexity, and morphological structures, *CRSTDLA-Arafreq* allows clinicians to select linguistically valid items for the assessment and rehabilitation of children with dyslexia, dysorthography, or other language-related difficulties.

From an educational perspective, the database highlights the lexical trajectories that accompany the progression from preschool to the third year of primary school. This information can inform the design of more effective pedagogical materials, helping educators to balance frequency, diversity, and complexity in the vocabulary presented to learners.

Ultimately, *CRSTDLA-Arafreq* is more than a lexical inventory—it is a strategic scientific infrastructure at the crossroads of linguistics, pedagogy, and speech-language pathology. Its potential impact extends beyond Algeria, as it lays the groundwork for comparable initiatives across the Arab world, where the lack of tailored lexical resources has long hindered advances in language education and clinical practice.

By bridging this gap, *CRSTDLA-Arafreq* not only contributes to the advancement of applied linguistics in Arabic but also represents a transformative step toward equity in education and clinical assessment for Arabic-speaking children.

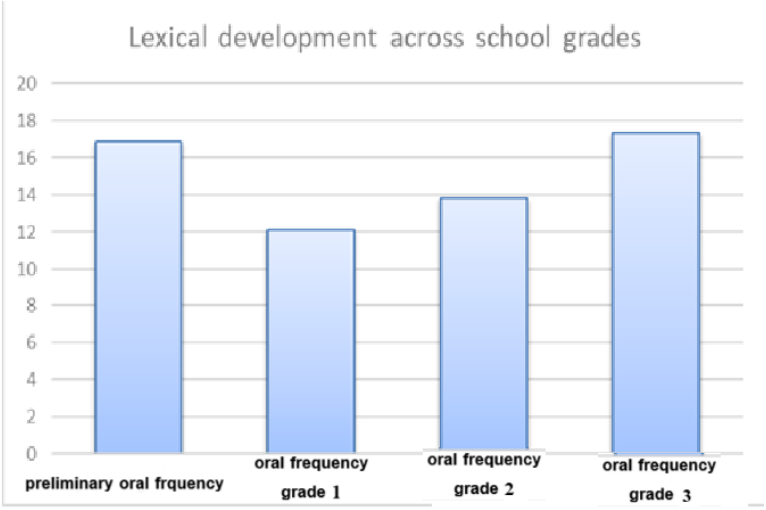

***Fig. 2**. Lexical development across school grades (oral frequency)*

The graph shows lexical development across school grades based on oral frequency measures. Initially, the preliminary oral frequency is high at around 17 units, indicating a strong oral vocabulary before formal schooling begins. In grade 1, the oral frequency drops to about 12, suggesting a transitional phase where students may be adapting to new learning environments or focusing on other linguistic or cognitive skills. From grade 1 to grade 3, oral frequency rises steadily again, reaching approximately 17 at grade 3, which indicates vocabulary expansion as students progress in school.

This pattern suggests that lexical development is not strictly linear; there is a dip in early schooling followed by growth. This could be due to increased complexity of academic content and vocabulary learning demands over time. The initial decline might also reflect shifts in learning focus during early education stages.

Overall, the data highlight the importance of targeted vocabulary support especially in early grades to mitigate the initial decline and promote continuous lexical growth throughout schooling. This trend underscores the dynamic nature of oral vocabulary development during the first years of formal education.

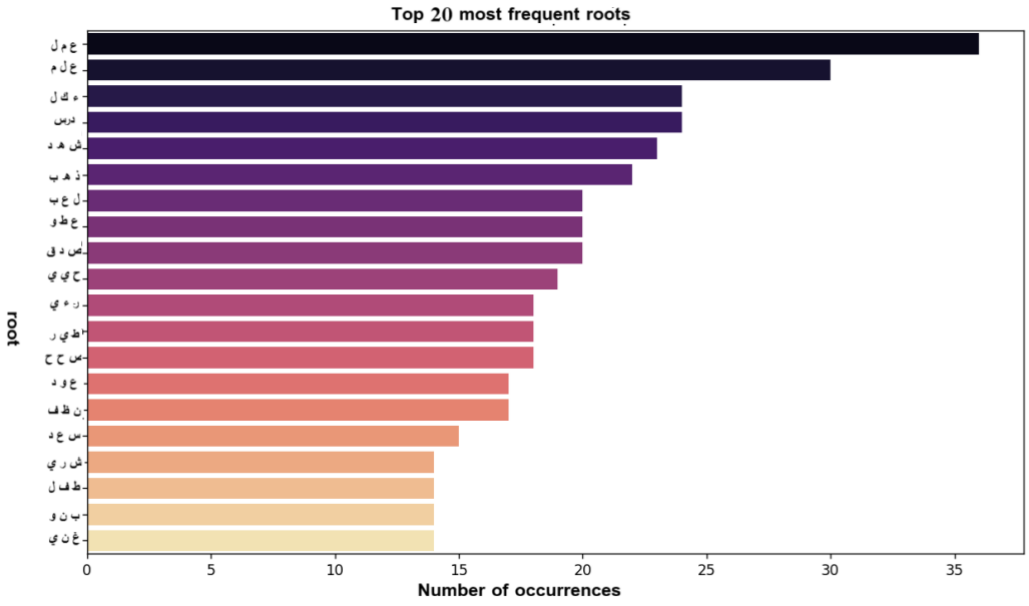

***Fig. 3.** Number of occurrences of root in database CRSTDLA-Arafreq*

The most frequent root is "ع م ل" with about 35 occurrences, indicating its dominant usage in the corpus. It is closely followed by the root "ع ل م" with around 30 occurrences, and then the root "ء ك ل" with nearly 25 occurrences. The distribution of occurrences gradually decreases from the top towards the less frequent roots, which cluster around 14-15 occurrences. This slow yet steady decline shows that some roots are highly productive and central, while others have a more moderate but still significant presence.

10.1. Linguistic interpretation

The quantitative analysis of the CRSTDLA-Arafreq reveals a strongly structured hierarchy driven by differential root productivity within the Arabic morphological system. Root frequencies display a clear asymmetrical distribution, with a small number of highly productive triliteral roots accounting for a substantial proportion of lexical occurrences.

High-productivity roots and morphological centrality. The root ع م ل (ʿ-m-l, action/work) emerges as the most productive root, with approximately 35 occurrences, followed by ع ل م (ʿ-l-m, knowledge/science; ~30 occurrences) and ء ك ل (ʾ-k-l, to eat; ~25 occurrences). These roots exhibit high derivational yield, generating multiple lemmas across verbal, nominal, and adjectival paradigms (e.g., verbal forms, verbal nouns, agentive and instrumental patterns). Their high frequency and broad morphological dispersion position them as the structural core of the school-based and everyday Arabic lexicon, illustrates their semantic importance in everyday discourse and their derivative richness. These roots serve as semantic bases for many words morphologically constructed using the patterns identified in the previous chart.

Long-tail distribution and lexical specialization. Beyond this core, root frequencies decline progressively toward a cluster centered around 14–15 occurrences, consistent with a long-tail distribution. This pattern reflects the transition from a small set of morphologically central roots with high paradigm expansion to a larger set of less productive roots with restricted derivational scope. In Arabic, this shift corresponds to movement from general-purpose lexical items toward more semantically and morphologically specialized forms, which are typically acquired later and are more closely associated with academic language demands.

Overall, the distribution highlights the role of root-and-pattern morphology in structuring lexical availability: lexical frequency is not evenly distributed across roots but is instead conditioned by the capacity of specific roots to support extensive derivational networks. This structural asymmetry has direct implications for lexical acquisition, written production, and the progressive enrichment of academic vocabulary.

The most frequent roots are predominantly triliteral roots that are highly productive and form a wide variety of words through morphological derivation in Arabic. For example, "ء ك ل" relates to writing and is very central in the Arabic lexicon, well documented as one of the fundamental roots. The high frequency of roots like "ع م ل" (to work), "د ر س" (to study), or "ق ل ب" (heart, or sometimes change) illustrates their semantic importance in everyday discourse and their derivative richness. These roots serve as semantic bases for many words morphologically constructed using the patterns identified in the previous chart.

11. Conclusions

This ranking of roots by frequency highlights their central role in Arabic morphology and lexicon. It emphasizes the importance of frequent triliteral roots as semantic pivots for word formation.

Enhanced Analysis of Morphological Patterns (*Figure 4*) The analysis of morphological patterns in the *CRSTDLA-Arafreq* database provides a critical layer of understanding regarding the "cost" of lexical production.

Structural Primacy and Morphological Complexity Dominance of Basic Forms: The pattern فَعَلَ (fa‘ala) is the most frequent (~65 occurrences), representing the most productive and foundational morphological form in the corpus. Causative and Intensive Weights: The high frequency of أَفْعَلَ (’af‘ala) (~50) and فَعَّلَ (fa‘‘ala) (~48) highlights the importance of derived verbal forms that carry specific semantic loads (causativity, intensity).

Nominal Richness: The presence of patterns such as أَفْعَال (’af‘āl) and فَاعِل (fā‘ilun) indicates a morphosyntactic richness that bridges the gap between basic verbs and derived nouns.

Clinical and Developmental Interpretation To make this data sufficient for a high-level interpretation, it should be linked to the child's cognitive effort:

The Complexity Threshold: The "clear drop" after the first three patterns suggests a threshold where morphological complexity increases. In clinical practice, an inability to mobilize patterns around the 18-occurrence mark reflects a failure to master the morphological "long tail" of the language.

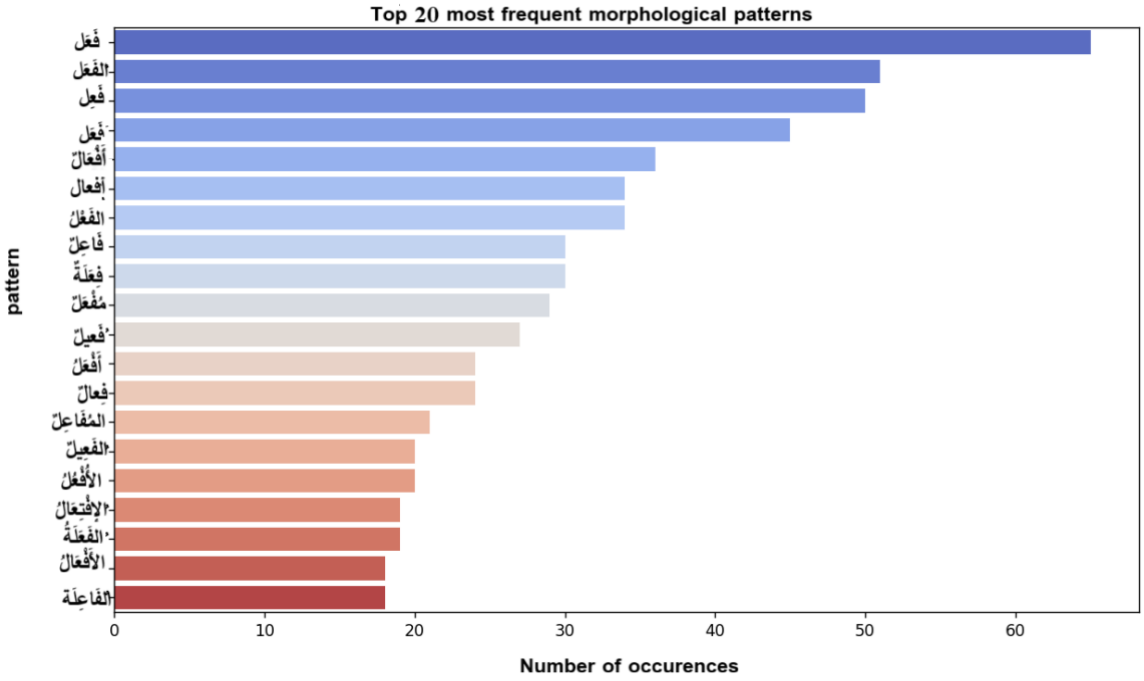

Fig. 4. *Number of occurrences of patten in database CRSTDLA-Arafreq*

Inhibition of Distractors: When naming an image, a child might substitute a complex pattern (e.g., إِفْعَال) with a simpler one (فَعَلَ). The lexical gap index tracks this "morphological simplification," which is a key marker of dysorthography and lexical organization difficulties.

Alignment with Cognitive Models of Written Production

Analysis of morphological patterns in the CRSTDLA-Arafreq database can be interpreted through the lens of cognitive models of written production, including Levelt's modular and sequential models and their extensions to writing (Ellis; Rapp).

Written production involves a sequence of interdependent processes: semantic activation, lexical selection, morphological encoding, phonological encoding, and graphemic conversion. Morphological patterns serve as a key interface between the abstract lexicon and written output.

Highly frequent, simple patterns, such as فَعَلَ, correspond to strongly automated lexical representations. Their activation requires minimal cognitive effort, accounting for their high predictability and frequent use in writing. More complex derived patterns (e.g., أَفْعَلَ, فَعَّلَ, إِفْعَال) require greater cognitive resources, both for selecting the appropriate pattern and for managing competing alternatives, reflecting the natural increase in cognitive load associated with less frequent morphological forms.

Overall, the observed distribution of morphological patterns illustrates the functional organization of the cognitive system in writing: highly frequent patterns follow the most automated pathways, while peripheral, less frequent patterns reflect points of higher cognitive demand, highlighting the strategies learners naturally adopt to manage writing efficiently.

12. Database Architecture

The originality of *CRSTDLA-Arafreq* lies in its "360-degree" view of the learner's lexicon. The preliminary results highlight a significant divergence between the institutional lexicon (textbooks) and the mobilized lexicon (student productions).

12.1. The Morphological Matrix

Unlike standard frequency lists, *CRSTDLA-Arafreq* provides a matrix where each word is linked to its root and derivational pattern.

- **Lexical Exposure:** Identification of the most frequent roots (e.g., k-t-b, q-r-') and patterns (e.g., fa'ala, mufā'ala) across levels.

• **Morphological Complexity Index:** Each word is assigned a weight based on its length, number of affixes, and phonological structure, allowing for the first time an objective measurement of "textbook difficulty" in Algeria.

12.2. Frequency Mismatch (Input vs. Output)

The triangulation reveals that some high-frequency words in textbooks (Input) are rarely used or are frequently misspelled by students in written productions (Output). This "Lexical Gap" is a core finding of the project, providing a scientific basis for revising pedagogical materials.

12.3. Operational Applications: Bridging Research and Practice

CRSTDLA-Arafreq is not merely a linguistic repository; it is a strategic tool designed for three specific user groups.

12.4. For Speech-Language Pathologists (Clinical Practice)

The database enables the creation of evidence-based assessment batteries:

• **Target Selection:** Clinicians can select test items based on real school exposure rather than intuition.

• **Diagnostic Precision:** By knowing the frequency of a word in the Algerian curriculum, the SLP can distinguish between a lack of exposure and a true language disorder (Specific Language Impairment/Dyslexia).

• **Graded Therapy:** Providing word lists ranked by morphological complexity for progressive rehabilitation.

12.5. For Educators and Curriculum Designers (Pedagogy)

• **Textbook Optimization:** Aligning the introduction of new vocabulary with the cognitive and morphological development of children aged 5–9.

• **Controlled Literacy Instruction:** Designing reading primers that prioritize high-frequency roots and simple patterns to reduce the initial cognitive load of literacy acquisition.

12.6. For Researchers (Linguistics & Psychology)

• **Normative Data:** Serving as a reference for longitudinal studies on Arabic lexical acquisition.

• **Psycholinguistic Experiments:** Providing controlled stimuli (frequency, length, syllabic structure) for studies on word recognition and reading latency in the Arab world.

13. Conclusion: A Strategic Infrastructure

In conclusion, *CRSTDLA-Arafreq* constitutes the first scientific infrastructure in Algeria that systematically links the Modern Standard Arabic of schoolbooks with the real-world performance of students. By providing a quantified map of the primary school lexicon, it moves the field from intuitive practice to data-driven decision-making in both education and speech-language pathology.

14. Future Directions

While the *CRSTDLA-Arafreq* project has established a strong foundation, several avenues remain open for further development and expansion:

Extension to Higher Grade Levels

The current database focuses on preschool to third grade. Extending the corpus to include materials from later primary and secondary education would provide a more comprehensive picture of lexical development across the school trajectory.

Integration of Regional and Dialectal Variations

Although the project has accounted for sociolinguistic variation, a systematic inclusion of dialectal Arabic data could enrich the database, offering resources not only for Modern Standard Arabic (MSA) but also for the varieties children encounter in everyday communication.

Digital Tools and NLP Applications

Linking *CRSTDLA-Arafreq* with Natural Language Processing (NLP) technologies would enhance its usability, allowing automated analyses of frequency, morphological parsing, and orthographic error detection. Such integration could support the development of intelligent tutoring systems and digital assessment tools.

Cross-linguistic Comparisons

Establishing connections with databases like Manulex (French), ChildLex (German), or DILLo (Italian) would allow comparative studies on lexical acquisition across languages. This could foster international collaborations and highlight specificities of Arabic lexical development.

Clinical Validation and Test Construction

Further research should focus on validating the database for clinical use, particularly in the design of standardized, culturally adapted assessment tools for children with dyslexia, dysgraphia, or other learning difficulties.

Open Access and Educational Outreach

Making the database available as an open-access platform for teachers, clinicians, and researchers would maximize its impact. Complementary training modules and user-friendly interfaces could help practitioners incorporate *CRSTDLA-Arafreq* into daily practice.

REFERENCES

1. Adams, M. J. (1990). Beginning to read: Thinking and learning about print. MIT Press.
2. https://mitpress.mit.edu/9780262510769/
3. Beccaria, R., Cristiano, A., Pisciotta, F., & Usardi, N. (2019). DILLo: A lexical database for speech-language therapy in Italian. Proceedings of the 12th Language Resources and Evaluation Conference (LREC), 5051–5055. https://aclanthology.org/L18-1366/
4. Boudelaa, S., & Marslen-Wilson, W. D. (2010). ARALEX: A lexical database for Modern Standard Arabic. Behavior Research Methods, 42(2), 481–487. https://doi.org/10.3758/BRM.42.2.481
5. Bybee, J. (2006). From usage to grammar: The mind's response to repetition. Language, 82(4), 711–733.
6. https://www.jstor.org/stable/4490266
7. Carlisle, J. F. (2000). Awareness of the structure and meaning of morphologically complex words: Impact on reading. Reading and Writing: An Interdisciplinary Journal, 12(3–4), 169–190.
8. https://doi.org/10.1023/A:1008131926604
9. Duchon, A., Perea, M., Sebastián-Gallés, N., Martí, A., & Carreiras, M. (2013). EsPal: One-stop shopping for Spanish word properties. Behavior Research Methods, 45(4), 1246–1258.
10. https://doi.org/10.3758/s13428-013-0326-1
11. Kirby, J. R., Deacon, S. H., Bowers, P. N., Izenberg, L., Wade-Woolley, L., & Parrila, R. (2012). Children's morphological awareness and reading ability. Reading and Writing, 25(2), 389–410.
12. https://doi.org/10.1007/s11145-010-9276-5
13. Lété, B., Sprenger-Charolles, L., & Colé, P. (2004). MANULEX: A lexical database of French elementary school reading textbooks. Behavior Research Methods, Instruments, & Computers, 36(1), 156–166.
14. https://doi.org/10.3758/BF03195560
15. Nagy, W. E., & Anderson, R. C. (1984). How many words are there in printed school English? Reading Research Quarterly, 19(3), 304–330. https://doi.org/10.2307/747823
16. Nation, I. S. P. (2001). Learning vocabulary in another language. Cambridge University Press.
17. https://doi.org/10.1017/CBO9781139524759
18. Schroeder, S., Würzner, K.-M., Heister, J., Geyken, A., & Kliegl, R. (2015). childLex: A lexical database of German read by children. Behavior Research Methods, 47(4), 1085–1094.
19. https://doi.org/10.3758/s13428-014-0528-1