

Arabic Text Simplification and Readability Assessment for Educational Applications: A Systematic Literature Review

Khalid Essaadani, Soumia Ziti
Mohammed V University in Rabat, Morocco

Abstract—Arabic text simplification and readability assessment support level-appropriate material selection, language learning, early literacy, teacher adaptation, and accessibility, but the degree of educational validation remains uncertain. This PRISMA 2020 systematic review searched Scopus, Web of Science Core Collection, and IEEE Xplore on 6 August 2026 and updated coverage through a targeted ACL Anthology search, a standalone-readability sensitivity search, and citation-cluster checking on 11 September 2026. Following source-identity verification, 194 records were identified, 129 were screened, 69 full texts were assessed, and 61 studies were included. An eight-criterion appraisal rated 45 studies high, 13 moderate, and 3 low quality. Thirty-four studies primarily modelled readability, while recent work increasingly used Arabic pretrained language models, ordinal objectives, graph models, large corpora, and controlled generation. Forty-two studies relied primarily on automatic or benchmark metrics; only two directly measured learner comprehension or retelling, and none measured longitudinal learning gain. Excluding low-quality studies and interpreting documented resource reuse did not alter the central conclusion. The retrieved evidence demonstrates technical feasibility and growing educationally relevant resources, but educational effectiveness remains unestablished. Priorities are independent learner and teacher evaluation, cross-domain and cross-variety validation, interoperable scales, human calibration of automatic metrics, and closed-loop diagnosis, simplification, and level verification.

Keywords—Arabic text simplification; Arabic readability assessment; educational natural language processing; systematic literature review

I. INTRODUCTION

Matching a text to its reader is a central educational requirement. Readability assessment estimates text difficulty or predicts an appropriate reader level; text simplification modifies lexical, syntactic, or discourse complexity while attempting to preserve meaning. In educational systems, the tasks are complementary: assessment can identify a mismatch, simplification can adapt the material, and reassessment can test whether the target level was reached.

Arabic complicates this cycle through rich morphology, optional diacritics, orthographic variation, lexical-frequency effects, and diglossia. Modern Standard Arabic (MSA), Classical Arabic, and spoken varieties place different demands on readers. The relevant populations range from first-language (L1) school readers to Arabic second- or foreign-language

learners (L2), university students, emergent readers, teachers, and people who need accessible text.

The central evidence problem is not whether a model improves a benchmark score, but whether the improvement transfers to a defined reader and produces a meaningful outcome. Model accuracy, annotation agreement, and linguistic judgments are necessary forms of evidence, yet they do not by themselves establish comprehension, retention, learning, teacher adoption, or sustained classroom benefit.

A. Background and Prior Reviews

Arabic readability research includes corpus-frequency formulas [3], feature-rich machine learning, leveled corpora, and neural models. Mat Daud et al. propose a formula based on reversed word-frequency ranks, while discussing length-based formulas as prior work [3]. Nassiri, Cavalli-Sforza, and Lakhouaja's survey [1] describes formula-based and machine-learning approaches to Arabic readability for L1, L2, and adult audiences, according to its published abstract.

Building on that readability-focused account, the present review jointly examines readability assessment and text simplification under an educational scope. Its contribution is an explicit connection between study-level technical findings and the kind of educational evidence they provide: resource validity, reader processing, usability, direct learner outcomes, or learning gain. PRISMA selection, criterion-level appraisal, the September 2026 search update, and documented evidence-family mapping make this connection auditable. The earlier survey's full text was unavailable for this revision; this comparison is therefore limited to its published scope and does not assert that particular procedures or topics were absent from it.

Its contributions are a combined synthesis of Arabic text simplification and readability assessment under an explicit educational scope; a reproducible search, deduplication, screening, extraction, and appraisal trail; a taxonomy of educational contexts, computational approaches, resources, and evaluation practices; an evidence-level distinction between technical performance, reader validity, and educational effectiveness; and a prioritized research agenda for learner-centred, teacher-centred, and classroom evaluation.

B. Research Questions

- RQ1: Which educational objectives, learner populations, and application contexts are addressed?

- RQ2: Which computational or analytical approaches, tools, and techniques are used?
- RQ3: Which resources, datasets, annotation schemes, and benchmarks support the field?
- RQ4: How are systems and resources evaluated, and what educational outcomes are measured?
- RQ5: Which evidence gaps and research priorities emerge across the literature?

II. REVIEW METHOD

A. Review Design and Scope

The review followed PRISMA 2020 [2]. The protocol was developed before the original screening but was not prospectively registered. The unit of inclusion was a peer-reviewed primary journal article or full conference paper examining Arabic text simplification, readability, reading level, text difficulty, or text complexity with an explicit educational, learner-facing, teacher-facing, literacy, or accessibility connection.

Table I states the operational eligibility rules. Retrieval failure was recorded separately and was not treated as a content-based exclusion.

TABLE I. ELIGIBILITY CRITERIA AND OPERATIONAL EXCLUSIONS

Criterion	Include	Exclude or record separately
Language focus	Arabic is the target language or Arabic results are separable.	Non-Arabic work, Arabic-script work in another language, or multilingual work without separable Arabic evidence.
Target phenomenon	Text simplification, readability assessment/prediction, reading level, text difficulty, or text complexity.	Adjacent tasks in which readability is only incidental.
Educational relevance	Learners, teachers, curricula, material selection, literacy, accessibility, or another explicit learner-facing use.	Domain readability with no defensible educational or learner-facing connection.
Evidence type	Original computational, corpus, system, usability, reader, learner, or pedagogical research.	Editorials, narrative surveys, perspectives, and reviews.
Publication type	Peer-reviewed journal article or full conference paper.	Thesis, abstract-only record, poster, book/chapter, standard, or non-peer-reviewed report.
Assessment	Sufficient full text to judge eligibility and extract evidence.	Unavailable reports were recorded as not retrieved; they were not assigned a substantive exclusion reason.

ACM Digital Library coverage was checked through ACM DOI-prefix metadata because the public interface triggered security verification and did not provide a reproducible export during the revision. Two ranked searches (“Arabic readability” and “Arabic text simplification”) identified no unique eligible primary report: two potentially eligible reports were already in the database set, and two Arabic-focused records were reviews. Exact queries, dates, filters, counts, clean copyable Arabic-script strings, and record-level decisions are supplied in the supplementary workbook.

C. Record Management, Retrieval, and Selection

Records were deduplicated first by normalized DOI and then by normalized title, while retaining provenance. Title and abstract records were coded Include, Maybe, or Exclude with a reason; Include and Maybe records advanced to retrieval. Additional attempts for the original 15 unavailable reports comprised publisher/DOI checks, exact-title web and institutional-repository searches, ResearchGate availability checks, and OpenAlex open-access-location verification. A

B. Information Sources and Search Strategy

Scopus, the Web of Science Core Collection, and IEEE Xplore were searched on 6 August 2026. Each English query combined Arabic identifiers with phrases for text simplification, readability assessment, reading level, text difficulty, and text complexity. Arabic-script supplementary queries returned zero records. Journal articles and full conference papers were retained through database-specific type filters; no year, country, subject, document-language, or access-status restriction was used.

To test the NLP coverage gap identified during review, the official ACL Anthology title-and-abstract metadata were searched on 11 September 2026 with the original concept blocks. A sensitivity query added the standalone term “readability,” which had been omitted originally to limit unrelated uses. The main ACL query returned 25 candidates, and the standalone-readability query returned 16 non-overlapping candidates. Five additional BAREC system papers were located by citation-cluster checking because their searchable metadata did not explicitly repeat an Arabic identifier. After removal of five overlaps with the original set and full-text screening of 41 unique reports, seven were excluded, and 34 were newly included. The standalone-readability search contributed 10 of these inclusions, demonstrating that the term materially improved recall.

source-identity audit on 16 September 2026 found that the PDF stored for ATSRA-010 [9] was a different publication by the same authors. Publisher and exact-title checks did not yield the intended full text; the linked ResearchGate PDF repeated the mismatch. ATSRA-010 was therefore reclassified as not retrieved, and its extraction and appraisal were withdrawn. The final total is 16 unretrieved reports. This was a retrieval correction, not a content-based exclusion or a new search inclusion. The reports, attempts, and topic-only coverage implications are listed in the supplement. No result, direction, or effect was imputed from an unretrieved report.

D. Data Extraction and Evidence Dependence

For every included report, the extraction table records citation, year, setting, population or educational context, Arabic variety, task, method, dataset and size, readability scale, evaluation design and metrics, principal findings, limitations, and evidence relevant to RQ1–RQ5. Publications were additionally mapped to evidence families. BAREC corpus, guidelines, demo, shared-task, and system papers form a shared

benchmark family; SAMER, DARES/SaudiTextBooks, AraEyebility, ATSC, and the L2/GLOSS research programme also contain repeated or evolving data. Documented reuse is distinguished from related research with unverified participant overlap, including the earlier 2021 eye-tracking pilot. These mappings guide interpretation but do not establish an exact count of independent datasets or participant samples.

E. Quality Appraisal

Eight equally weighted criteria were scored 0, 0.5, or 1: clear aims and educational relevance; description of data/sample/context/Arabic variety; reproducibility; claim-appropriate evaluation; design-appropriate baseline, control, agreement, or validation; clear results and suitable metrics; proximity to intended readers, learners, educators, or users; and limitations/transparency/availability. Criterion-specific anchors are provided in the supplement. For example, Q5 accepts competitive baselines or ablations in computational studies, inter-annotator agreement or held-out benchmarks in corpus studies, and comparison/control evidence in user studies. Q7 deliberately records educational-evidence proximity: direct comprehension, retelling, retention, learning gain, or authentic educator adoption scores 1; indirect human/expert/user evidence scores 0.5; no human evidence scores 0.

Totals of 6.5–8 were classified high (at least 81.25% of the maximum), 4.5–6 moderate (56.25–75%), and 0–4 low (50% or less). These pragmatic thresholds distinguish broadly complete, partially complete, and substantially limited evidence profiles. Quality was not an exclusion rule. High-quality technical studies could strongly support a model or resource claim while remaining weak evidence of educational effectiveness.

F. Review Procedure and Verification

The first reviewer conducted the full screening, data extraction, and quality appraisal. The coauthor then independently re-screened a blinded, stratified sample of 13 of 129 records (10.1%) using the same Include, Maybe, or Exclude categories and eligibility criteria. Exact three-category agreement was 7/13 (53.8%; Cohen's kappa = 0.284). Because both Include and Maybe advanced to retrieval, agreement on the operational advance-versus-exclude decision was 11/13 (84.6%; kappa = 0.683). Four differences were conservative Include-versus-Maybe judgments and did not alter progression. The first reviewer adjudicated all differences against the prespecified criteria and available full texts. ATSRA-017 was changed from Exclude to Include because the protocol explicitly includes accessibility and learner-facing information;

SUP-001 remained excluded because text difficulty was used only as a machine-translation performance variable. Extraction and quality appraisal were not independently duplicated and remain a limitation. The screening sample, reasons, agreement calculations, adjudications, complete extraction data, and item-level ratings are supplied in the supplementary materials.

G. Synthesis

A narrative thematic synthesis was organized around RQ1–RQ5. Meta-analysis was inappropriate because populations, tasks, scales, datasets, and outcomes were heterogeneous. Four categories were used: direct learner outcome (comprehension, Cloze, retelling, retention, or achievement in intended learners); reader-processing evidence (eye movements, time, or perceived difficulty without a learning-gain design); usability evidence (workflow or acceptability without comprehension); and learning gain (pre/post or longitudinal improvement attributable to an intervention). Automatic metrics, annotation reliability, expert linguistic judgments, corpus statistics, and demonstrations were classified as technical/resource evidence.

III. RESULTS

A. Study Selection

The original databases yielded 148 records. Supplementary ACL searching and citation-cluster checking identified 46 additional records, for 194 in total. Removal of 65 duplicates left 129 records for screening; 44 were excluded at title and abstract screening. Eighty-five reports were sought, 16 were not retrieved after source-identity verification, and 69 were assessed for eligibility. Eight reports were excluded after full-text assessment (seven supplementary reports outside the review scope and one master's thesis), leaving 61 included studies (Fig. 1) (Table V). The 16 unretrieved reports represented 36.4% of the 44 distinct reports sought through the original database pathway and 18.8% of all 85 reports sought after the update.

B. Study Characteristics and Evidence Structure

The 61 publications appeared from 2013 to 2026; 48 (78.7%) were published from 2022 onward. Table II summarizes the evidence base and distinguishes publication volume from evidential independence. Readability modelling was the largest primary task ($n = 34$), followed by resource/corpus work and simplification or readability-controlled generation ($n = 8$ each), tools/systems ($n = 6$), and human/pedagogical assessment ($n = 5$). The concentration of 2025 publications partly reflects 18 BAREC shared-task and overview papers that evaluate a common benchmark.

TABLE II. DISTRIBUTION OF THE INCLUDED EVIDENCE BASE

Dimension	Distribution	Total
Quality tier	High 45; moderate 13; low 3	61
Primary task	Readability modelling 34; resource/corpus 8; simplification or controlled generation 8; tool/system 6; human/pedagogical assessment 5	61
Primary evaluation	Automatic/benchmark metrics 42; resource quality 6; system/demo 5; learner/reader experiment 3; expert simplification evaluation 2; usability 1; pedagogical analysis 1; prototype 1	61
Direct educational evidence	Direct learner outcomes 2; learning-gain studies 0	2 direct; 0 gain
Major dependent families	BAREC; SAMER; DARES/SaudiTextBooks; AraEyebility; ATSC; L2/GLOSS	6 mapped families

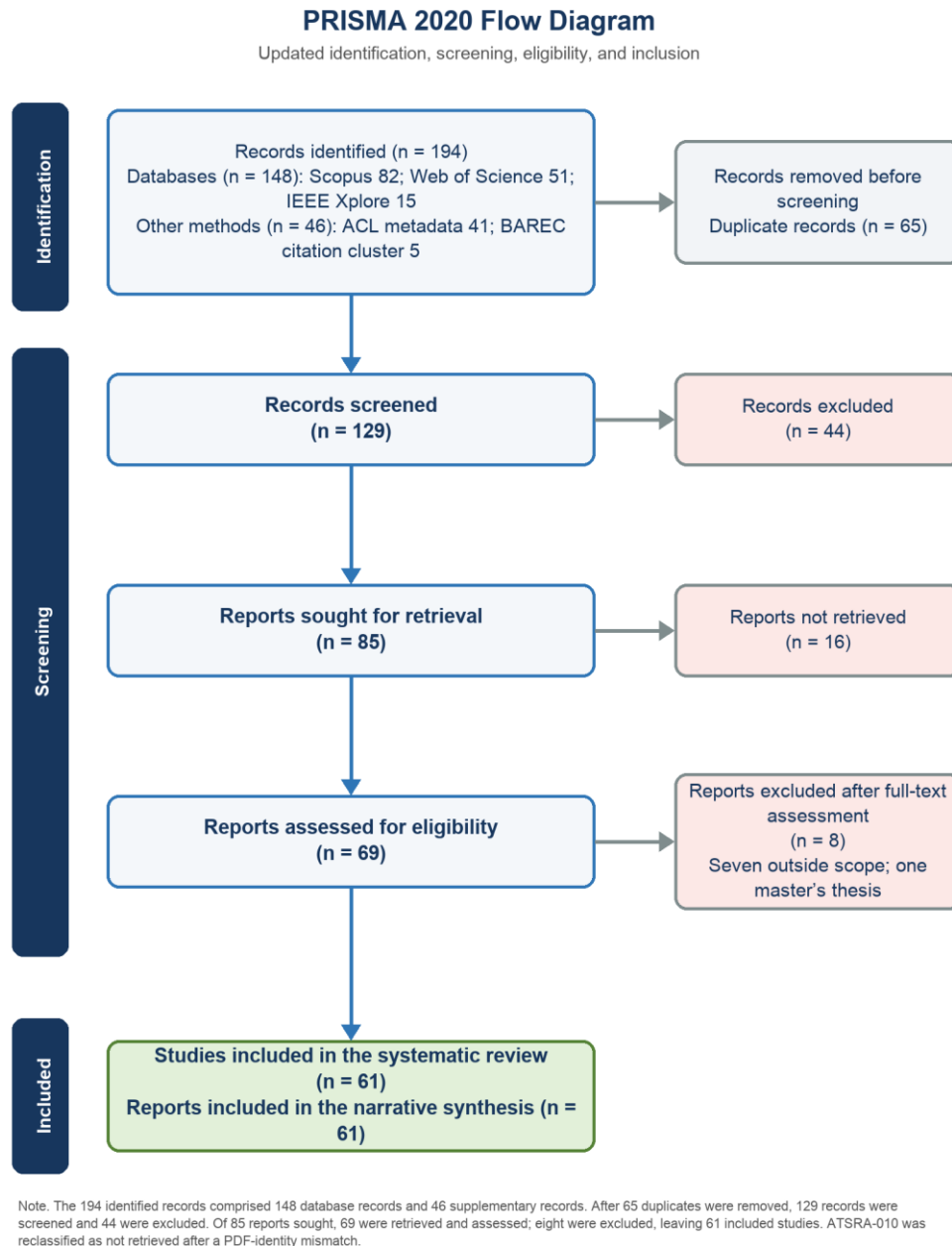


Fig. 1. PRISMA 2020 flow diagram for the updated review.

C. Quality Appraisal and Quality-Informed Synthesis

Forty-five studies were high quality, 13 moderate, and three low; the mean was 6.45 of 8. The item-level table shows a recurrent pattern: data, methods, and benchmark results are usually well reported, while Q7 is often zero because intended readers or educators are not evaluated. This does not invalidate technical findings; it limits the claim they can support.

Quality weighting was applied concretely. High-quality BAREC, SAMER, DARES, and cross-domain studies anchor claims about corpus maturity, model performance, and domain

sensitivity [4], [15], [29], [37]–[44], [55]–[63]. Moderate-quality demos and accessibility audits describe workflows or text properties but do not establish learner benefit [10], [12], [30], [34], [64]. The three low-quality reports, including the CART prototype [13], contribute historical or design context only and do not anchor any conclusion. Omitting them leaves 58 studies; both direct learner studies remain, no learning-gain study appears, and the conclusion of technical feasibility without established educational effectiveness is unchanged (Table III).

TABLE III. QUALITY-INFORMED SENSITIVITY ANALYSIS

Analysis	n	High/Moderate/Low	Direct learner outcomes	Learning gain	Interpretation
All included	61	45/13/3	2	0	Technical feasibility; effectiveness unestablished
Exclude low quality	58	45/13/0	2	0	Central conclusion unchanged
Account for evidence families	Publication total retained; dependent resources mapped	Family-level interpretation	2 independent learner samples	0	Model volume remains concentrated around shared resources

D. RQ1: Educational Objectives, Learners, and Applications

The literature proposes four principal educational uses. First, readability prediction supports material selection, curriculum grading, and passage matching for school, university, and L2 readers [7], [8], [10], [15], [22], [23], [29], [32], [35], [37], [42], [58]. Second, simplification systems and resources support lexical substitution, sentence rewriting, accessible fiction, religious text adaptation, and readability-controlled generation [5], [6], [10], [25], [29], [36], [39], [45], [56], [57]. Third, tools expose difficulty information to teachers, authors, annotators, and learners through reports, visualization, thesauri, and searchable corpora [10], [12], [16], [30], [34]. Fourth, reader studies examine textbook suitability, early literacy, variety, and cognitive processing [7], [11], [14], [18], [20], [27].

The demonstrated educational impact is much narrower than the proposed use. In the university Cloze study, 54 students completed four unseen textbook passages; passage mean scores ranged from 0.46 to 0.66, and all four passages were classified at the frustration level for that sample [7]. The study directly tests material-learner fit but is limited to one institution and textbook. In the preschool experiment, 30 Arabic-speaking children (mean age 5 years 10 months) encountered four picture-supported narratives that varied in MSA versus spoken Palestinian Arabic and in episodic complexity. Comprehension was higher for MSA, while greater episodic complexity supported longer retellings; these are outcome-specific findings, not uniform effects across all measures [27]. Its direct child evidence is valuable but limited by a small, relatively advantaged sample and four narratives.

These two reports are the only direct learner-outcome studies. The three studies counted as primary learner/reader experiments comprise those two plus one CA-MSA eye-tracking comparison of 15 matched proficient adults: 23 were recruited, 18 retained for MSA, and 15 completed both sessions [18]. Other eye-tracking papers use gaze data primarily for resource construction or prediction [11], [14], [20]. The usability study involved five graduate human-computer interaction students evaluating four existing general readability tools to inform Arabic reporting interfaces, not an Arabic learner intervention [16]. Thus, “three learner/reader experiments” and “two direct learner outcomes” describe different evidence categories and are not conflicting counts. No study measured longitudinal learning gain.

E. RQ2: Methods, Tools, and Techniques

Formula and rule-based work uses surface length, syllable structure, lexical density, word frequency, diacritics, and Arabic-specific factors such as Faseeh. The corpus-frequency

formula in [3] draws on KACSTAC (739,119,011 words); this is its source corpus, not a learner-validation sample. OSMAN evaluated an Arabic formula against English readability scores on approximately 73,000 aligned UN text units, described as sentences in its abstract and paragraphs in its methods [31]. These methods are transparent and inexpensive, but formula correlation is not equivalent to learner calibration.

Classical machine learning uses linguistic features with support-vector methods and random forests [17], [21]. SATLab instead combines character n-grams with support-vector regression and optimized ordinal thresholds, providing a non-neural BAREC baseline [60]. Feature optimization reported 94.8% L1 accuracy and 72.4% L2 SVM accuracy in its main setting; its augmented-feature experiment additionally reached 74.1% L2 accuracy with kNN (Table VII in [32]). On the Jordanian School Textbook Corpus, CAMELBERT reached in-domain QWK 0.830 but fell to 0.085 cross-domain, whereas a 127-feature XGBoost model retained the best cross-domain QWK of 0.240 [58].

Neural approaches include fastText and contextual embeddings, CNN/BiLSTM architectures, Arabic pretrained transformers, graph neural networks, ordinal heads, differentiable QWK losses, conformal decoding, ensembles, and visual text models [19], [29], [35], [38], [40]–[43], [46]–[54], [59], [61]–[63]. In the BAREC shared task, the strongest system reached QWK 0.875 at sentence level and 0.874 at document level [40], [42]. Because 18 reports share BAREC data and tracks, these results compare techniques within one family rather than supplying 18 independent validations.

Simplification methods range from lexical pipelines and embedding-based substitutions to Seq2Seq, mT5, switch transformers, machine-translation hybrids, and prompted LLM generation [5], [6], [25], [29], [36], [39], [45], [56], [57]. The 2023 embedding system reached complex-word identification F1 of 0.88; expert review found that 79% of binary-model outputs were simpler, but the multi-complexity model better preserved meaning (69%) and grammar (84%), demonstrating a real simplification-adequacy trade-off [36]. Controlled CEFR prompting reached high automatic profile alignment [56], but still lacked learner testing.

Human-centred techniques include Cloze testing, narrative comprehension and retelling, expert linguistic judgments, usability observation, and eye tracking. The later AraEyebility corpus and analyses reuse related data [11], [18], [20]. The earlier 41-participant pilot [14] belongs to the same research programme, but participant overlap with the later corpus is unverified. Small samples and documented reuse limit generalization; a shared research programme alone does not establish identical participants.

F. RQ3: Resources, Datasets, and Scales

Resource capacity has expanded substantially. BAREC contains 69,441 sentences and more than one million words across 19 levels, with QWK 0.818 in its final annotation phase [4], [44]; BAREC-10M extends the programme to approximately 10 million words and 45 subcorpora [55]. SAMER contains approximately 159,000 original-corpus words from 15 novels with word/document levels and two simplified versions [39], and its associated lexicon contains approximately 26,000 triply annotated lemmas [33]. DARES 1.0 includes 13,335 curriculum instances [37], while DARES 2.0 expands and cleans concept-centred Saudi textbook data [15]. Other resources include ATSC (500 article/simplification pairs) [6], README++ (9,757 multilingual sentences, including 1,945 Arabic) [24], AraEyebility [11], MoSAR [23], Saaq al-Bambuu (2,980 sentence pairs) [29], and the 82,512-segment, 240-textbook JSTC [58].

Interoperability remains weak. Taha/Arabi21 levels, school grades, the Common European Framework of Reference, Interagency Language Roundtable levels, Cloze thresholds, and bespoke labels differ in target population and unit of analysis. Dataset expansion therefore does not automatically produce transferable educational meaning. Crosswalks should report uncertainty and be validated with intended readers.

G. RQ4: Evaluation and Educational Outcomes

Forty-two publications primarily report automatic or benchmark metrics such as accuracy, macro/weighted F1, QWK, BLEU, SARI, METEOR, ROUGE, BERTScore, or

distance from an ordinal label. Six focus on resource quality and five on system or demo functionality. These evaluations can establish technical reproducibility, annotation reliability, or relative performance, but not comprehension or learning.

Evidence categories are exclusive at the primary-evaluation level but are interpreted alongside secondary human evidence. Direct learner outcome requires task performance by intended learners; reader processing requires cognitive or behavioral reading measures without a learning-gain design; usability concerns workflow and acceptability; learning gain requires change attributable to an intervention. The observed counts are two, three, one, and zero, respectively. Expert judgments of grammar, meaning, or annotation quality strengthen linguistic validity but are not learner outcomes.

H. RQ5: Gaps and Priorities

The dominant gap is educational validation: 59 of 61 publications lack direct learner outcomes, and no longitudinal classroom intervention was found. Other priorities are teacher-centred evaluation, external and cross-domain testing, dialect/register coverage, calibration of automatic metrics against Arabic comprehension, transparent cross-scale mapping, and independent replication beyond recurring datasets. Unretrieved reports appear to concern accessibility, early tools, feature/vectorization methods, L2 readability, class balancing, patient education, lexical simplification, and web indices; their absence may narrow historical and applied coverage, but no unobserved result is assigned to them (Table IV).

TABLE IV. EVIDENCE-BASED RESEARCH AGENDA

Priority	Evidence gap	Recommended design	Required outcomes
Critical	Direct learner evidence remains 2/61; learning-gain evidence is 0/61.	Controlled or quasi-experimental studies with intended learners.	Comprehension, retention, reading efficiency, learning gain.
Critical	No longitudinal classroom deployment.	Teacher co-design and sustained authentic-course deployment.	Adoption, workload, instructional decisions, sustainability.
High	Strong benchmark but weak cross-domain transfer.	External validation across curricula, genres, regions, and varieties.	QWK/F1 plus subgroup validity and error analysis.
High	Automatic metrics dominate 42 primary evaluations.	Calibrate metrics against Arabic human judgments and learner tasks.	Meaning, grammaticality, comprehensibility, comprehension.
High	Readability scales are not interoperable.	Multi-scale annotation and transparent probabilistic crosswalks.	Reliability, mapping uncertainty, transfer validity.
High	Publication volume clusters around shared resources.	Independent datasets, preregistered splits, and evidence-family reporting.	External replication and generalization.
High	Few closed-loop educational systems.	Diagnose difficulty, simplify, reassess target level, then test with users.	Target attainment, comprehension, teacher control.

IV. DISCUSSION

A. Principal Findings and Educational Applications

Research on Arabic readability and simplification is rapidly expanding. The updated search reveals a mature technical ecosystem of leveled lexicons, school and literary corpora, annotation tools, visual interfaces, feature-based classifiers, Arabic transformers, ordinal objectives, and controlled generators. It also strengthens—not weakens—the original central message: adding 34 ACL studies increases the amount of technical evidence without adding a direct learner-outcome or learning-gain study.

For material selection and curriculum design, practical evidence combines direct mismatch detection in the university Cloze study [7], child variety/complexity effects [27], school-corpus modelling [15], [37], [58], and textbook classification and gradation analysis [26]. For teacher authoring, tools can flag difficult words, suggest alternatives, expose levels, and search graded corpora [10], [12], [30], [34], while the five-person general-tool usability study offers interface-design guidance rather than educator-adoption evidence [16]. Adoption, workload, and decision quality remain insufficiently measured. For L2 learning, CEFR/ILR/GLOSS work shows feasible automatic leveling [8], [28], [32], [35], yet the labels require calibration to real learners.

For accessibility, the Arabic orthodontic-information audit covered 74 websites and found apparently accessible formula scores but limited reliability of the online information and no direct patient-comprehension test [64]. Other disability and patient-education reports remain unretrieved, so conclusions in these subdomains remain cautious.

Generative systems create a new educational opportunity and a new validity risk. Models can simplify, control grade or CEFR profiles, and diversify early-grade stories [5], [6], [25], [29], [36], [39], [45], [56], [57]. However, automatic level agreement can coexist with factual drift, loss of meaning, cultural mismatch, or inappropriate difficulty for a real learner. Educational deployment therefore requires human evaluation, reader testing, educator control, and post-generation reassessment.

B. Quality, Dependence, and Strength of Conclusions

The quality-informed synthesis prevents high benchmark scores from carrying inappropriate educational weight. High-quality computational studies anchor statements about what works on their datasets; direct learner reports anchor the much narrower statements about comprehension and text-reader fit. The low-quality reports are contextual only. Excluding them does not change the conclusion. Likewise, evidence-family mapping shows that the BAREC surge represents intensive comparison on one valuable benchmark, not many independent populations. Similar caution applies to AraEyebility participant subsets and successive DARES, SAMER, ATSC, and L2/GLOSS outputs.

C. Strengths and Limitations

Strengths include exact database queries, a targeted ACL update, a standalone-readability sensitivity search, citation-cluster checking, explicit record-level decisions, retrieval auditing, study-level extraction, criterion-anchored appraisal, sensitivity analysis, and evidence-family mapping. The supplementary workbook makes each main conclusion traceable to individual reports.

Limitations remain. Sixteen reports could not be retrieved, including one whose stored PDF proved to be a different publication; they represented 36.4% of reports sought through the original database pathway and 18.8% overall after the update. The source-identity error led to withdrawal of that report's extraction and appraisal. ACM coverage was checked indirectly through DOI-prefix metadata because the interface did not permit reproducible access during revision. The protocol was not registered prospectively. Heterogeneity precluded meta-analysis. Independent coauthor re-screening covered 10.1% of records and showed substantial agreement on the operational advance-versus-exclude decision, but only fair exact three-category agreement. Extraction and appraisal were not independently duplicated, so reviewer-dependent judgment may still have affected extracted classifications and quality scores. The comparison with the earlier survey is limited to its published abstract, and some participant-overlap relationships remain unverified. Finally, 2026 indexing is still evolving, and later records may alter counts.

V. CONCLUSION

Research on Arabic text simplification and readability assessment has progressed from formulas and handcrafted indicators to large, fine-grained corpora, cognitive signals, Arabic pretrained language models, graph and ordinal methods, and readability-controlled generation. These advances create credible tools for level-aware material selection, text adaptation, curriculum analysis, and educational content generation.

They do not yet establish broad educational effectiveness. Only two of 61 included publications directly measured learner outcomes, none measured longitudinal learning gain, and much of the recent model literature reuses a small number of resource families. The camera-ready conclusion is therefore bounded: technical feasibility and resource maturity are supported by the retrieved evidence; improved learning or classroom practice is not. The next phase should connect models to representative learners, teachers, curricula, varieties, and authentic use, while preserving meaning and independently verifying target-level attainment.

SUPPLEMENTARY MATERIALS

The supplementary review data, including the database search log, screening decisions, unretrieved-report audit, extraction table, study-level evidence tables, evidence-family map, quality-appraisal ratings, sensitivity analyses, evidence-category definitions, and independent coauthor verification, are publicly available in Zenodo at <https://zenodo.org/records/22829321>, DOI: 10.5281/zenodo.22829321.

DECLARATIONS

Ethics approval: Not applicable. This review analyzed published literature and collected no new human-participant data.

Funding: This research received no external funding.

Conflicts of interest: The authors declare no conflict of interest.

Data availability: The review dataset and supporting supplementary materials are publicly available in Zenodo at <https://zenodo.org/records/22829321>, DOI: 10.5281/zenodo.22829321.

Declaration on generative AI: Generative AI tools assisted spreadsheet organization, document preparation, and language refinement. The authors remain responsible for eligibility decisions, evidence interpretation, citation verification, originality, and final content.

REFERENCES

- [1] N. Nassiri, V. Cavalli-Sforza, and A. Lakhouaja, "Approaches, methods, and resources for assessing the readability of Arabic texts," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 22, no. 4, Art. no. 95, 2023. doi: 10.1145/3571510.
- [2] M. J. Page et al., "The PRISMA 2020 statement: An updated guideline for reporting systematic reviews," *BMJ*, vol. 372, Art. no. n71, 2021. doi: 10.1136/bmj.n71.

- [3] N. Mat Daud, H. Hassan, and N. Abdul Aziz, "A corpus-based readability formula for estimate of Arabic texts reading difficulty," *World Applied Sciences Journal*, vol. 21, no. SPECIAL ISSUE4, pp. 168–173, 2013.
- [4] K. N. Elmadani, N. Habash, and H. Taha-Thomure, "A Large and Balanced Corpus for Fine-grained Arabic Readability Assessment," in *Findings of the Association for Computational Linguistics: ACL 2025*, 2025, pp. 16376–16400. doi: 10.18653/v1/2025.findings-acl.842.
- [5] A. Alsughayyir, and A. Alshanqiti, "A light but efficient switch transformer model for Arabic text simplification," *International Journal of Advanced and Applied Sciences*, vol. 12, no. 4, pp. 71–78, 2025. doi: 10.21833/ijaas.2025.04.009.
- [6] A. Alshanqiti, A. Alkhodre, A. Namoun, S. Albouq, and E. Nabil, "A Transformer Seq2Seq Model with Fast Fourier Transform Layers for Rephrasing and Simplifying Complex Arabic Text," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 2, pp. 888–898, 2023. doi: 10.14569/IJACSA.2023.01402101.
- [7] A.H.H. Ahmad, and Y.N. Hasan, "Applied Arabic Textbook Readability Level for University Students at Al-Balqa Applied University," *Theory and Practice in Language Studies*, vol. 14, no. 1, pp. 186–191, 2024. doi: 10.17507/tpls.1401.22.
- [8] N. Nassiri, A. Lakhouaja, and V. Cavalli-Sforza, "Arabic L2 readability assessment: Dimensionality reduction study," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 6, pp. 3789–3799, 2022. doi: 10.1016/j.jksuci.2020.12.021.
- [9] N. Nassiri, A. Lakhouaja, and V. Cavalli-Sforza, "Arabic readability assessment for foreign language learners," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2018, pp. 480–488. doi: 10.1007/978-3-319-91947-8_49.
- [10] R. Hazim, H. Saddiki, B. Alhafni, M. Al Khalil, and N. Habash, "Arabic Word-level Readability Visualization for Assisted Text Simplification," in *EMNLP 2022 - 2022 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Demonstrations Session*, 2022, pp. 242–249. doi: 10.18653/v1/2022.emnlp-demos.24.
- [11] I. Baazeem, H. Al-Khalifa, and A. Al-Salman, "AraEyebility: Eye-Tracking Data for Arabic Text Readability," *Computation*, vol. 13, no. 5, Art. no. 108, 2025. doi: 10.3390/computation13050108.
- [12] K. Altarbouch, K.N. Elmadani, O. Obeid, H. Taha-Thomure, and N. Habash, "BAREC Demo: Resources and Tools for Sentence-level Arabic Readability Assessment," in *EMNLP 2025 - 2025 Conference on Empirical Methods in Natural Language Processing, Proceedings of the System Demonstrations*, 2025, pp. 950–959. doi: 10.18653/v1/2025.emnlp-demos.74.
- [13] A. H. Ishraf Shuib, N. A. Aziz, N. M. Daud, and H. Hasan, "CART — An Arabic teaching-learning tool for non-native Arabic speakers," in *The 5th International Conference on Information and Communication Technology for The Muslim World (ICT4M)*, 2014, pp. 1–4. doi: 10.1109/ICT4M.2014.7020585.
- [14] I. Baazeem, H. Al-Khalifa, and A. Al-Salman, "Cognitively driven arabic text readability assessment using eye-tracking," *Applied Sciences (Switzerland)*, vol. 11, no. 18, Art. no. 8607, 2021. doi: 10.3390/app11188607.
- [15] S. Almujaivel, D. Premasiri, T. Ranasinghe, M. El-Haj, and R. Mitkov, "Complex Concept-Based Readability Estimation from Arabic Curriculum," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 24, no. 11, Art. no. 128, 2025. doi: 10.1145/3770070.
- [16] H. Saddiki, V. Cavalli-Sforza, and K. Bouzoubaa, "Enhancing Visualization in Readability Reports for Arabic Texts," in *Procedia Computer Science*, 2017, pp. 241–247. doi: 10.1016/j.procs.2017.10.115.
- [17] E. Salesky and W. Shen, "Exploiting Morphological, Grammatical, and Semantic Correlates for Improved Text Difficulty Assessment," in *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications*, 2014, pp. 155–162. doi: 10.3115/v1/W14-1819.
- [18] H. Al-Khalifa, "Eye Movement Patterns as Indicators of Text Complexity in Arabic: A Comparative Analysis of Classical and Modern Standard Arabic," *Journal of Eye Movement Research*, vol. 18, no. 4, Art. no. 30, 2025. doi: 10.3390/jemr18040030.
- [19] M.A. Ouassil et al., "Hybrid convolutional neural network-bidirectional long short-term memory model for Arabic sentence readability assessment," *IAES International Journal of Artificial Intelligence*, vol. 15, no. 3, pp. 2849–2862, 2026. doi: 10.11591/ijai.v15.i3.pp2849-2862.
- [20] I. Baazeem, H. Al-Khalifa, and A. Al-Salman, "Integrating Linguistic and Eye Movements Features for Arabic Text Readability Assessment Using ML and DL Models," *Computation*, vol. 13, no. 11, Art. no. 258, 2025. doi: 10.3390/computation13110258.
- [21] S. Berrichi, N. Nassiri, A. Mazroui, and A. Lakhouaja, "INTERPRETING THE RELEVANCE OF READABILITY PREDICTION FEATURES," *Jordanian Journal of Computers and Information Technology*, vol. 9, no. 1, pp. 36–52, 2023. doi: 10.5455/jjeit.71-1667559201.
- [22] A. Basheer, S. Alawneh, T. Bostami, S. Daraghme, and A. Hawash, "Mapping Arabic Text Studying Material With The Most Suitable Student Grade," in *2022 ASU International Conference in Emerging Technologies for Sustainability and Intelligent Systems (ICETIS)*, 2022, pp. 259–267. doi: 10.1109/ICETIS55481.2022.9888935.
- [23] N. Nassiri, V. Cavalli-Sforza, and A. Lakhouaja, "MoSAR: Modern standard Arabic readability corpus for L1 learners," in *ACM International Conference Proceeding Series*, 2019. doi: 10.1145/3372938.3372961.
- [24] T. Naous, M.J. Ryan, A. Lavrouk, M. Chandra, and W. Xu, "README++: Benchmarking Multilingual Language Models for Multi-Domain Readability Assessment," in *EMNLP 2024 - 2024 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, 2024, pp. 12230–12266. doi: 10.18653/v1/2024.emnlp-main.682.
- [25] S.S. Al-Thanyyan, and A.M. Azmi, "Simplification of Arabic text: A hybrid approach integrating machine translation and transformer-based lexical model," *Journal of King Saud University - Computer and Information Sciences*, vol. 35, no. 8, Art. no. 101662, 2023. doi: 10.1016/j.jksuci.2023.101662.
- [26] S. Mohamed, "Text classification and gradation in Arabic textbooks," *Language Learning Journal*, vol. 52, no. 6, pp. 629–649, 2024. doi: 10.1080/09571736.2023.2213695.
- [27] K. Kwar, E. Saiegh-Haddad, and S. Armon-Lotem, "Text complexity and variety factors in narrative retelling and narrative comprehension among Arabic-speaking preschool children," *First Language*, vol. 43, no. 4, pp. 355–379, 2023. doi: 10.1177/01427237221149800.
- [28] H. Saddiki, K. Bouzoubaa, and V. Cavalli-Sforza, "Text readability for Arabic as a foreign language," in *Proceedings of the 2015 IEEE/ACS 12th International Conference of Computer Systems and Applications (AICCSA)*, 2015, pp. 1–8. doi: 10.1109/AICCSA.2015.7507232.
- [29] N. Khallaf, S. Sharoff, and R. Soliman, "Towards Arabic Sentence Simplification via Classification and Generative Approaches," in *Proceedings of the Seventh Arabic Natural Language Processing Workshop*, 2022, pp. 43–52. doi: 10.18653/v1/2022.wanlp-1.5.
- [30] N. Al-Twaresh, A. Al-Dayel, H. Al-Khalifa, M. Al-Yahya, S. Alageel, N. Abanmy, et al., "MADAD: A Readability Annotation Tool for Arabic Text," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 2016.
- [31] M. El-Haj, and P. Rayson, "OSMAN A Novel Arabic Readability Metric," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 2016.
- [32] H. Saddiki, N. Habash, V. Cavalli-Sforza, and M. Al Khalil, "Feature Optimization for Predicting Readability of Arabic L1 and L2," in *Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications*, 2018. doi: 10.18653/v1/W18-3703.
- [33] M. Al Khalil, N. Habash, and Z. Jiang, "A Large-Scale Leveled Readability Lexicon for Standard Arabic," in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 2020.

- [34] Z. Jiang, N. Habash, and M. Al Khalil, "An Online Readability Leveled Arabic Thesaurus," in Proceedings of the 28th International Conference on Computational Linguistics: System Demonstrations, 2020. doi: 10.18653/v1/2020.coling-demos.11.
- [35] N. Khallaf, and S. Sharoff, "Automatic Difficulty Classification of Arabic Sentences," in Proceedings of the Sixth Arabic Natural Language Processing Workshop, 2021.
- [36] Y. SalahEldin, and C. Sabty, "Simplify: Automatic Arabic Sentence Simplification using Word Embeddings," in Proceedings of ArabicNLP 2023, 2023. doi: 10.18653/v1/2023.arabincnlp-1.35.
- [37] M. El-Haj, S. Almujaivel, D. Premasiri, T. Ranasinghe, and R. Mitkov, "DARES: Dataset for Arabic Readability Estimation of School Materials," in Proceedings of the Workshop on DeTermIt! Evaluating Text Difficulty in a Multilingual Context @ LREC-COLING 2024, 2024.
- [38] J. Liberato, B. Alhafni, M. Khalil, and N. Habash, "Strategies for Arabic Readability Modeling," in Proceedings of the Second Arabic Natural Language Processing Conference, 2024. doi: 10.18653/v1/2024.arabincnlp-1.5.
- [39] B. Alhafni, R. Hazim, J. D. Pinos Liberato, M. Al Khalil, and N. Habash, "The SAMER Arabic Text Simplification Corpus," in Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), 2024.
- [40] M. Basem, M. Younes, S. Ahmed, and A. Moustafa, "IMSA at BAREC Shared Task 2025: Ensembling Arabic Transformers for Readability Assessment," in Proceedings of The Third Arabic Natural Language Processing Conference: Shared Tasks, 2025. doi: 10.18653/v1/2025.arabincnlp-sharedtasks.42.
- [41] M. Saeed, R. Waly, and A. Ashraf Hussein, "AMAR at BAREC Shared Task 2025: Arabic Meta-learner for Assessing Readability," in Proceedings of The Third Arabic Natural Language Processing Conference: Shared Tasks, 2025. doi: 10.18653/v1/2025.arabincnlp-sharedtasks.45.
- [42] K. N. Elmadani, B. Alhafni, H. Taha, and N. Habash, "BAREC Shared Task 2025 on Arabic Readability Assessment," in Proceedings of The Third Arabic Natural Language Processing Conference: Shared Tasks, 2025. doi: 10.18653/v1/2025.arabincnlp-sharedtasks.34.
- [43] P. Elchafei, M. Osama, M. Rageh, and M. Abu-Elkheir, "GNNinjas at BAREC Shared Task 2025: Lexicon-Enriched Graph Modeling for Arabic Document Readability Prediction," in Proceedings of The Third Arabic Natural Language Processing Conference: Shared Tasks, 2025. doi: 10.18653/v1/2025.arabincnlp-sharedtasks.36.
- [44] N. Habash, H. Taha-Thomure, K. N. Elmadani, Z. Zeino, and A. Abushmaes, "Guidelines for Fine-grained Sentence-level Arabic Readability Annotation," in Proceedings of the 19th Linguistic Annotation Workshop (LAW-XIX-2025), 2025. doi: 10.18653/v1/2025.law-1.30.
- [45] S. Altammami, "Leveraging AI to Bridge Classical Arabic and Modern Standard Arabic for Text Simplification," in Proceedings of the New Horizons in Computational Linguistics for Religious Texts, 2025.
- [46] A. Amel Nait Djoudi, P. Bellot, and A. G. Chifu, "LIS at BAREC Shared Task 2025: Multi-Scale Curriculum Learning for Arabic Sentence-Level Readability Assessment Using Pre-trained Language Models," in Proceedings of The Third Arabic Natural Language Processing Conference: Shared Tasks, 2025. doi: 10.18653/v1/2025.arabincnlp-sharedtasks.51.
- [47] F. Mohamed Emad Elden, "MorphoArabia at BAREC Shared Task 2025: A Hybrid Architecture with Morphological Analysis for Arabic Readability Assessment," in Proceedings of The Third Arabic Natural Language Processing Conference: Shared Tasks, 2025. doi: 10.18653/v1/2025.arabincnlp-sharedtasks.41.
- [48] A. Abdou, "mucAI at BAREC Shared Task 2025: Towards Uncertainty Aware Arabic Readability Assessment," in Proceedings of The Third Arabic Natural Language Processing Conference: Shared Tasks, 2025. doi: 10.18653/v1/2025.arabincnlp-sharedtasks.44.
- [49] M. Ayesh, "PalNLP at BAREC Shared Task 2025: Predicting Arabic Readability Using Ordinal Regression and K-Fold Ensemble Learning," in Proceedings of The Third Arabic Natural Language Processing Conference: Shared Tasks, 2025. doi: 10.18653/v1/2025.arabincnlp-sharedtasks.47.
- [50] A. Alhassan, A. Mohamed, and M. Elamin, "Phantoms at BAREC Shared Task 2025: Enhancing Arabic Readability Prediction with Hybrid BERT and Linguistic Features," in Proceedings of The Third Arabic Natural Language Processing Conference: Shared Tasks, 2025. doi: 10.18653/v1/2025.arabincnlp-sharedtasks.49.
- [51] B. Sapirstein, "Pixels at BAREC Shared Task 2025: Visual Arabic Readability Assessment," in Proceedings of The Third Arabic Natural Language Processing Conference: Shared Tasks, 2025. doi: 10.18653/v1/2025.arabincnlp-sharedtasks.48.
- [52] S. Ahmad, "Qais at BAREC Shared Task 2025: A Fine-Grained Approach for Arabic Readability Classification Using a pre-trained model," in Proceedings of The Third Arabic Natural Language Processing Conference: Shared Tasks, 2025. doi: 10.18653/v1/2025.arabincnlp-sharedtasks.43.
- [53] S. Trigui, "STBW at BAREC Shared Task 2025: AraBERT-v2 with MSE-SoftQWK Loss for Sentence-Level Arabic Readability," in Proceedings of The Third Arabic Natural Language Processing Conference: Shared Tasks, 2025. doi: 10.18653/v1/2025.arabincnlp-sharedtasks.50.
- [54] A. M. Nazzal, "ZAI at BAREC Shared Task 2025: AraBERT CORAL for Fine Grained Arabic Readability," in Proceedings of The Third Arabic Natural Language Processing Conference: Shared Tasks, 2025. doi: 10.18653/v1/2025.arabincnlp-sharedtasks.37.
- [55] K. N. Elmadani, A. Mahmoud Wizani, H. Taha Thomure, and N. Habash, "A Large and Balanced Multi-Domain Arabic Corpus Annotated for Morphology, Syntax, and Readability," in Proceedings of the Fifteenth Language Resources and Evaluation Conference, 2026. doi: 10.63317/45f2o6t8piyi.
- [56] N. Rabihi, C. Qwaider, and T. Briscoe, "Can LLMs Control Readability? A Multi-Dimensional Evaluation Framework for CEFR-Controlled Arabic Generation," in Proceedings of the Joint Workshop on Readability and Text Simplification (READIXTSAR) @ LREC 2026, 2026. doi: 10.63317/48v7mxywgfja.
- [57] H. Khalid, S. Shapsough, and I. Zuolkernan, "Noise Steering for Controlled Text Generation: Improving Diversity and Reading-Level Fidelity in Arabic Educational Story Generation," in Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2026), 2026. doi: 10.18653/v1/2026.bea-1.41.
- [58] S. Alzubi, and R. Reynolds, "Transformer-based readability classifiers are worse than you think: Evidence from cross-domain Arabic readability assessment," in Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2026), 2026. doi: 10.18653/v1/2026.bea-1.52.
- [59] A. Bahloul, "Syntaxa at BAREC Shared Task 2025: BERTnParse - Fusion of BERT and Dependency Graphs for Readability Prediction," in Proceedings of The Third Arabic Natural Language Processing Conference: Shared Tasks, 2025. doi: 10.18653/v1/2025.arabincnlp-sharedtasks.35.
- [60] Y. Bestgen, "SATLab at BAREC Shared Task 2025: Optimizing a Language-Independent System for Fine-Grained Readability Assessment," in Proceedings of The Third Arabic Natural Language Processing Conference: Shared Tasks, 2025. doi: 10.18653/v1/2025.arabincnlp-sharedtasks.40.
- [61] S. Ibrahim, M. Rafiul Biswas, M. Bessghaier, and W. Zaghouani, "MarsadLab at BAREC Shared Task 2025: Strict-Track Readability Prediction with Specialized AraBERT Models on BAREC," in Proceedings of The Third Arabic Natural Language Processing Conference: Shared Tasks, 2025. doi: 10.18653/v1/2025.arabincnlp-sharedtasks.39.
- [62] N. Rabihi, "Noor at BAREC Shared Task 2025: A Hybrid Transformer-Feature Architecture for Sentence-level Readability Assessment," in Proceedings of The Third Arabic Natural Language Processing Conference: Shared Tasks, 2025. doi: 10.18653/v1/2025.arabincnlp-sharedtasks.46.

- [63] S. Sibae, O. Nacar, Y. Alhabashi, A. Ammar, and W. Boulila, "ANLPers at BAREC Shared Task 2025: Readability of Embeddings Training Neural Readability Classifiers on the BAREC Corpus," in Proceedings of The Third Arabic Natural Language Processing Conference: Shared Tasks, 2025. doi: 10.18653/v1/2025.arabicnlp-sharedtasks.38.
- [64] M. S. Alassaf, H. H. Hamadallah, A. Almuzaini, A. M. Aloufi, K. N. Al-Turki, A. S. Khoshhal, M. A. Alsulaimani, and R. Eshky, "Assessment of the Arabic patient-centered online information about orthodontic pain: A quality and readability assessment," PLOS ONE, vol. 19, no. 5, Art. no. e0303308, 2024. doi: 10.1371/journal.pone.0303308.

APPENDIX A: REPRODUCIBLE SEARCH COVERAGE

The full database-specific strings, filters, dates, and counts are supplied in the Search Log sheet of the supplementary workbook. The ACL update replicated the original Arabic and concept blocks in official title-and-abstract metadata, then ran a distinct sensitivity search using standalone "readability." Arabic-script strings are provided as plain UTF-8 cells to avoid bidirectional corruption in typeset syntax. The BAREC cluster check and ACM DOI-prefix verification are documented separately.

TABLE V. REPRESENTATIVE STUDY-LEVEL EVIDENCE (FULL 61-STUDY TABLE IN SUPPLEMENT)

Study/family	Population, data, or design	Principal finding	Main limitation
ATSRA-007	54 university students; four textbook passages; Cloze	All passages at frustration level; passage means 0.46–0.66	One institution/textbook
ATSRA-077	30 preschool children; four MSA/Palestinian Arabic narratives	Variety and complexity affected comprehension/retelling; MSA higher in tested design	Small advantaged sample; four stories
AraEyebility family	Later corpus/analyses share data; CA–MSA comparison n = 15	Gaze features and register contrasts reveal processing differences	Overlap with the earlier 41-reader pilot is unverified
BAREC family	69,441 sentences; 1M+ words; 19 levels; 18 shared-task/overview papers	Best QWK 0.875 sentence/0.874 document; annotation QWK 0.818	Common benchmark; no learner outcome
SAMER family	Approximately 26,000 lemmas; 159,000 original-corpus words	Supports word/fragment models and two target simplifications	Linked resources; literary-domain limits
DARES family	13,335 school-text instances; grades 1–12	Coarse weighted/macro F1 0.91/0.79; fine 0.68/0.55	Saudi curriculum; later version overlaps
ATSRA-004	ATSC, 500 article/simplification pairs	SARI≈0.73; BLEU≈0.65; METEOR≈0.68	Automatic metrics; no learner test
ACL-007	4,258 word records; 8/69/765 sentence-pair sets; expert ratings	Identification F1 0.88; simplicity–meaning–grammar trade-off	Experts, not intended learners
ACL-029	JSTC, 82,512 segments from 240 textbooks; DARES/BAREC external test	CAMeLBERT QWK 0.830 in-domain, 0.085 cross-domain; XGBoost cross-domain 0.240	Shows severe domain dependence
L2/GLOSS family	Repeated learner-oriented corpora and feature studies	Main L1/L2 accuracy 94.8%/72.4%; augmented L2 kNN 74.1%	Different feature/model settings; no new comprehension sample