

A Benchmark for Arabic Lemma-Level Readability Assessment

Khalid N. Elmadani,[†] Mostafa Saeed,[†] Hanada Taha-Thomure,[‡] Nizar Habash[†]

[†]Computational Approaches to Modeling Language Lab, New York University Abu Dhabi

[‡]Zai Arabic Language Research Centre, Zayed University

{khalid.nabigh,mostafa.saeed,nizar.habash}@nyu.edu, Hanada.Thomure@zu.ac.ae

Abstract

We present **BAREC-LexBench**, a benchmark for Arabic lemma-level readability assessment, designed to evaluate readability lexicons and computational models independently of their granularity levels. We build the benchmark through a large-scale pairwise comparison task, where annotators determine which of two Arabic lemmas is easier to read and understand. The benchmark consists of 36,000 annotated pairs covering both same- and cross-part-of-speech comparisons among open-class lemmas: adjectives, nouns, proper nouns, and verbs. The annotations were produced by expert Arabic educators with an average inter-annotator agreement of 77.7% (0.56 Kappa $\sim$ moderate agreement), and we analyze disagreement patterns to better understand where lemma-level readability judgments are most challenging. We benchmark a wide range of approaches, from readability lexicons and frequency-based lexicons to trainable models and large language models (LLMs). Finally, we demonstrate downstream uses of the benchmark by constructing a ranked readability lexicon covering open-class lemmas (**BAREC-LexRank**) and evaluating its contribution to sentence-level readability assessment. To support research and education, we make the dataset, ranked lexicon, and benchmark results openly available.¹

1 Introduction

Text readability plays a central role in comprehension, learning, and accessibility (DuBay, 2004; Klare, 1963). In educational and NLP applications, modeling readability enables aligning texts with readers’ linguistic and cognitive abilities, text simplification, and content recommendation (Allington et al., 2015; Barber and Klauda, 2020). While readability has been extensively studied in English and other major languages, Arabic readability research

Lex A	Lex B	POS	Easier
مُتْرَم concluded	مُفْرَم minced	adj	B
شام Syria	كُلْثُوم Kulthoum	noun_prop	A
كَابِد suffer	رَغِب desire	verb	B
قَطِيعَة breaking	إِزْغَام compulsion	noun	Equal

Table 1: Example lemma pairs, with their POS tags, and easier readability annotation.

remains limited due to the lack of fine-grained lexical and corpus-level resources.

The *SAMER* lexicon (Al Khalil et al., 2018) introduced readability lexicons for Arabic. It maps lemmas to one of five readability levels. Although this resource has been valuable for readability modeling (Liberato et al., 2024) and text simplification (Alhafni et al., 2024), its coarse five-level scale limits its ability to capture gradual lexical acquisition across the 19 educational levels defined by Taha/Arabi21 (Taha-Thomure, 2017) and BAREC (Habash et al., 2025) frameworks.

Our contributions are as follows:

- We **design a large-scale annotation framework** for Arabic lemma-level readability assessment using pairwise comparisons, where expert annotators select the easier lemma between two options.
- We **construct BAREC-LexBench, a benchmark dataset** of 36K annotated pairs spanning same- and cross-part-of-speech comparisons among open-class lemmas.
- We **evaluate diverse modeling approaches**, including readability lexicons, frequency-based resources, classical ML, BERT-based models, and LLMs.
- We **demonstrate downstream applications** by building **BAREC-LexRank**, a ranked readability lexicon and assessing its impact on sentence-level readability prediction.

¹<https://barec.camel-lab.com/>

2 Related Work

Arabic readability research has evolved from hand-crafted formulas to corpus-based and neural approaches. Early work relied mainly on surface features such as word and sentence length (Al-Dawsari, 2004; Al-Khalifa and Al-Ajlan, 2010), while later models incorporated POS tags, morphology, diacritization, and pretrained language models (Forsyth, 2014; Saddiki et al., 2018; El-Haj and Rayson, 2016; Liberato et al., 2024). These studies highlight the importance of linguistic features for Arabic readability, especially given Arabic’s rich morphology and orthographic variation.

Several graded resources have been developed for Arabic. The Taha/Arabi21 framework (Taha-Thomure, 2017) defines a 19-level readability scale for children’s books based on qualitative and quantitative criteria, but it is difficult to apply at large scale. Other efforts have explored CEFR-based Arabic leveling and vocabulary profiling (Habash and Palfreyman, 2022; Naous et al., 2024; Soliman and Familiar, 2024), as well as readability datasets based on educational materials such as DARES (El-Haj et al., 2024; Almujaivel et al., 2025). The SAMER project (Al Khalil et al., 2020) introduced a five-level Arabic readability lexicon, later used to build a manually annotated Arabic text simplification corpus (Alhafni et al., 2024).

This paper is part of the broader BAREC project, which develops resources for Arabic readability across the Arab world. Building on the 19-level BAREC Annotation Framework (Habash et al., 2025), the BAREC Corpus (Elmadani et al., 2025b) provided a large-scale, manually annotated sentence-level readability dataset. To support practical use, BAREC Demo (Altarbouch et al., 2025) introduced a web-based system for sentence-level Arabic readability assessment. The BAREC Shared Task 2025 (Elmadani et al., 2025a) encouraged community engagement for Arabic readability assessment. The BAREC-10M Corpus (Elmadani et al., 2026) extended this line of work with a much larger automatically labeled corpus.

This work complements these resources by focusing on lemma-level readability. Unlike SAMER, which assigns discrete five-level labels, our benchmark uses pairwise comparisons to evaluate readability distinctions independently of a particular scale. Unlike BAREC corpus, which targets sentence-level readability, this resource provides a lexical benchmark and a ranked lemma lexicon.

Set ID	Split	Lex #	Lex/SAMER Level	Pairs
<i>Same-POS: Adjective, Noun, Verb, Proper Noun</i>				
Set 1-4	Train	500	100	3,600
Set 5	Tune	125	25	900
Set 6	Dev	125	25	900
Set 7-8	Test	250	50	1,800
Subtotal per POS category				7,200
Subtotal for four POS categories				28,800
<i>Cross-POS</i>				
Set 1-4	Train	480	96	3,600
Set 5	Tune	120	24	900
Set 6	Dev	120	24	900
Set 7-8	Test	240	48	1,800
Subtotal				7,200
Total Pairs				36,000

Table 2: Organization of lemma (Lex) sets and pair sampling across POS categories.

3 Dataset Creation

We construct **BAREC-LexBench**, a benchmarking dataset for Arabic lemma-level readability, designed to be independent of any specific granularity or level definitions. The task is formulated as a **pairwise comparison**: given two lemmas, the annotator chooses which one is easier to read and understand. We start from the **SAMER lexicon** (Al Khalil et al., 2020), classifying lemmas into open- and closed-class categories by part of speech (POS). Our benchmark therefore focuses on open-class lemmas - adjectives, nouns, proper nouns, and verbs - which account for 99% of lexical entries. It covers same- and cross-POS lemma comparisons. Although closed-class lemmas are not included in the benchmark, we annotate them separately following the BAREC 19-level framework (Habash et al., 2025), thereby extending the resource to cover all parts of speech.

3.1 Dataset Sampling

Our setup consists of five POS categories; four Same-POS categories covering comparisons between lemmas from the same POS (adjectives (Adj), nouns (Noun), verbs (Verb), and proper nouns (PropN)), and one Cross-POS category covering comparisons between lemmas from the different POS (Cross).

Same-POS Setup For each POS, we select 1,000 lemmas evenly distributed across the five SAMER levels (200 per level), divided into eight disjoint sets of 125 lemmas. Each set yields 900 lemma pairs (1800 total lemma appearances). The aver-

age lemma appearance in each set is 14.4 (1800 appearances / 125 lemmas). We ensure each lemma appears at least 14 times -twice with each SAMER level and the rest with random lemmas.

Cross-POS Setup Similar to the Same-POS setup, Cross-POS lemmas are divided into eight disjoint sets. For each set, we select 30 lemmas from each of the corresponding Same-POS sets (120 total lemmas), while ensuring balanced selection across the five SAMER levels (6 lemmas per POS per SAMER level). Each set yields 900 lemma pairs (1800 total lemma appearances), ensuring each lemma appears 15 times -once with each combination of the three other parts-of-speech and the five SAMER levels.

This results in 7.2K pairs per POS category and 36K total across all categories (see Table 2).

3.2 Annotation Team and Process

Six native Arabic educators annotated the dataset using a Google Sheets interface (see Appendix B). For each lemma pair, annotators selected: (1) first lemma is easier, (2) second lemma is easier, (3) equally easy, or (4) I don’t know. Due to the inherent subjectiveness of the task, our annotation guidelines are general: annotators are asked to select the lemma they expect learners to find easier due to earlier encounter, spelling simplicity, or other pedagogically grounded factors, but the specific reasoning is left to their judgment. The “equal” option is reserved for genuinely equal difficulty pairs. The 7,200 pairs per POS category are shuffled and split into 14 batches of 500 pairs and one smaller batch of 200 pairs. Annotating one batch tasks an hour on average. Each batch is annotated by two annotators. The average pairwise agreement is 77.7% with 0.56 Kappa $\sim$ moderate agreement (Cohen, 1960). All disagreements were resolved by a third annotator. After consolidation, 99.5% of all lemma pairs are annotated with one lemma being easier than the other, the rest are equally easy.

3.3 Data Analysis

Table 3 analyzes annotator disagreements by the absolute SAMER-level difference between lemma pairs. Although disagreements occur in 22% of all pairs, they are most frequent when the lemmas have similar readability levels: 32% for pairs with the same SAMER level, decreasing steadily to 6% for pairs four levels apart. This suggests that disagreements mainly reflect fine-grained readability

SAMER Difference	# Pairs	Disagreements	
		#	%
0	7,144	2,287	32%
1	11,573	3,109	27%
2	8,640	1,681	19%
3	5,746	695	12%
4	2,897	174	6%
All	36,000	7,946	22%

Table 3: Disagreements by SAMER level difference between the lemma pairs.

distinctions, while pairs with larger readability gaps are easier for annotators to judge consistently.

Qualitative Analysis of Disagreements To better understand the sources of annotator disagreement, we randomly sampled 100 lemma pairs from the disagreement cases and asked annotators to explain the reasons behind their decisions. We then analyzed the comments and grouped the disagreements into seven categories: world knowledge, revision, dialect similarity, structure, concept, register, and ambiguity. Table 4 summarizes the resulting categories. The most frequent source of disagreement is world knowledge, where annotators differed in their familiarity with names, places, historical figures, or culturally specific references.

3.4 Closed Classes

We annotate closed-class lemmas separately from the benchmark dataset, following the BAREC 19-level framework (Habash et al., 2025). This decision is motivated by the small number of closed-class lemmas (368 lemmas), many of which already have explicit levels in BAREC. The annotation was carried out through discussions among three annotators until agreement was reached.

4 Experiments

We formulate the pairwise comparison task as a three-way classification problem, where the input is a pair of lemmas and the output is one of three labels: (1) the first lemma is easier, (2) the second lemma is easier, or (3) both are equally easy. We explore four modeling approaches: readability-based, frequency-based, trainable models, and LLMs.

4.1 Readability Lexicons

Each lemma is assigned a readability level from a given lexicon. For any pair, the lemma with the lower level is considered easier; if levels match, the pair is labeled “equal.”

Category	Main idea	Count	Example Pair
World Knowledge	Familiarity with names, places, or history.	31	بِرَازِيل <i>Brazil</i> PropN < > نَحَاس <i>Nahhas</i> PropN
Revision	Judgment changed after review.	23	وَارِد <i>imported</i> Adj < > مُنْتَقَى <i>selected</i> Adj
Dialect Similarity	Easy because it resembles local usage.	14	إِشْوَار <i>bracelet</i> Noun < > مُتَنَبِّح <i>leaning</i> Noun
Structure	Difficulty from sound or word structure.	10	مَضْرُوب <i>multiplier</i> Noun < > وَثْبَة <i>leap</i> Noun
Concept	Concept is abstract, technical, or cultural.	10	أَتَمَر <i>yield</i> Verb < > ضَاعَف <i>double</i> Verb
Register	Commonness or formal/literary status differs.	9	وَافَق <i>agree</i> Verb < > نَزَعَ <i>divest</i> Verb
Ambiguity	Different meaning or grammatical reading.	3	كَاد <i>almost</i> Verb < > سَاح <i>travel/melt</i> Verb

Table 4: Qualitative categories of annotator disagreement based on a random sample of 100 disagreement cases.

SAMER Lexicon The SAMER Lexicon assigns each lemma one of five readability levels based on the average of three annotators. We use two variants: *SAMER-5* (rounded average, 5 levels) and *SAMER-13* (unrounded, 13 levels), with *SAMER-5* as the primary baseline.

BAREC-based Lexicons Using the 19-level BAREC Corpus (Elmadani et al., 2025b), we lemmatize all words with Saeed and Habash’s (2025) best-performing system and assign each lemma the level of the easiest sentence in which it appears. Lemmas that do not appear in the corpus are assigned the hardest level, 19. We repeat the same procedure using the BAREC-10M Corpus (Elmadani et al., 2026), which is ten times larger than the BAREC Corpus but labeled automatically.

4.2 Frequency Lexicons

Each lemma receives a frequency value, with higher frequency indicating easier readability. Lemmas not in the list receive zero frequency.

CAMELBERT Frequency List Based on dediacritized word counts from the CAMELBERT-mix corpus (Khalifa et al., 2021), each lemma inherits the frequency of its dediacritized form.

BAREC-based Frequency Lists We construct two frequency lists following the same preprocessing procedure used for the *BAREC-based Lexicons*. For each lemma, we compute its frequency in the lemmatized version of each corpus, producing two frequency lists from the BAREC and BAREC-10M Corpora.

4.3 Trainable Models

We train classifiers to predict the three labels.

Classical ML We evaluate classical ML models using frequency features from CAMELBERT-MSA, DA, CA, and Mix (*ArFreq*), together with SAMER-based readability features from the three individual annotators, SAMER-5, and SAMER-13 (*ArSAMER*). Furthermore, we test a range of orthographic features, including lemma length (*ArLength*), character frequency, and diacritic frequency (*ArCharFreq*), which are known to interact with readability in Arabic (Taha-Thomure, 2017). We additionally represent the root (*ArRoot*) and abstract pattern (*ArAbsPattern*) as character-level features.

We also add English gloss-based features to the feature pool. The motivation is that if certain concepts are introduced to learners at similar readability levels across languages, then readability-related properties of an English gloss may provide additional information about the Arabic lemma meaning. Specifically, we use the frequency of the English gloss based on dictionary word frequencies derived from the Google Books Ngram dataset (*GlossFreq*),² as well as the gloss readability level from Maddela and Xu’s (2018) manual readability annotations of the 15K most frequent words in the Google Books Ngram dataset (*GlossRL*). We also include the number of words in the gloss (*GlossWordCount*), motivated by the intuition that more specialized or specific lemmas may require longer expressions to convey their meaning in another language. Finally, we add surface-level gloss features, including character (*GlossCharCount*) and syllable counts (*GlossSyllableCount*).

We report the best-performing configuration and provide feature-level analysis.

²<https://www.kaggle.com/datasets/wheelercode/dictionary-word-frequency>

BERT Finetuning We fine-tune the top three Arabic BERT-based models identified by Inoue et al. (2021): AraBERTv02 (Antoun et al., 2020), MARBERTv2 (Abdul-Mageed et al., 2021), and CAMeLBERT-MSA (Inoue et al., 2021). The input to all models is formatted as:

[CLS] *Lemma_1* [SEP] *Lemma_2* [SEP]

The experimental setup is available at Appendix C.

4.4 Large Language Models (LLMs)

We frame the task as multiple choice (first lemma, second lemma, or equal) (see Appendix D.2 for the prompts). We evaluate six open-weights LLMs: three Arabic (*ALLaM-7B-Instruct* (Bari et al., 2025), *Fanar-2-27B-Instruct* (TEAM et al., 2026), and *Jais-2-70B-Chat* (Anwar et al., 2025)), and three multilingual models (*Qwen3-30B-A3B-Instruct* (Team, 2025), *Aya-Expanse-32B* (Dang et al., 2024), and *Llama-3.3-70B-Instruct* (Grattafiori et al., 2024)), and two proprietary models: *Gemini-3.1-Pro* (Team et al., 2025) and *GPT-5.5-Pro* (Singh et al., 2026).

We test variations in input features, prompt language, lemma diacritization, and few-shot settings, reporting the best configuration for each model and analyzing prompt sensitivity for each LLM.

5 Results

Table 5 shows the performance of all models across *Dev* and *Test* sets for all POS categories.

5.1 Lexicon-based Models

SAMER-13 consistently outperforms *SAMER-5*, thanks to its finer granularity capturing small readability differences within the same SAMER level. The BAREC-based and BAREC-10M-based lexicons and frequency lists achieve moderate performance, limited by the coverage of the Corpora and the accuracy of automatic lemmatization system. The *BAREC-based Lexicon* outperforms the *BAREC-10M-based Lexicon*, suggesting that manual annotation quality is more important than corpus size for deriving readability levels. In contrast, the larger *BAREC-10M* corpus yields a stronger frequency list, where no readability annotations are required. *CAMeLBERT Frequency List* performs well overall but is less effective for nouns and verbs, as adjectives and proper nouns often appear in their lemma form, making word frequency lookup more accurate for these classes.

5.2 Trainable Models

Classical ML achieves the second-best *Test* performance. This highlights the complementarity of lexical frequency, SAMER-13, and orthographic features in modeling readability. Readability is largely governed by measurable lexical and orthographic properties that classical ML methods capture effectively, especially when supported by annotated resources like the SAMER lexicon.

Classical ML Sensitivity to Different Features

Table 6 presents an ablation study of the effect of different feature groups on Classical ML performance. Overall, removing all gloss-based features gives the best performance, suggesting that the English gloss features do not add substantial information beyond the Arabic features. Across POS categories, the best-performing feature setting varies without a clear pattern. However, the SAMER-based feature group is the most important individual feature set: removing it causes the largest drop in performance across all settings. This indicates that the model relies primarily on existing readability signals from SAMER, while the remaining features provide complementary information that further improves performance. Appendix E shows the results of different Classical ML techniques.

5.3 Large Language Models

Among all the models, *GPT-5.5-Pro* is the strongest system. The other LLMs perform competitively but remain below the *Classical ML* model, with most results clustered in the mid-70% range. The top Arabic-based model is *Jais-2-70B-Chat*, while the top multilingual open-weights model is *Aya-Expanse-32B*.

Prompt-Sensitivity Analysis To analyze LLM sensitivity to prompt design, we evaluate each model under 24 prompt configurations. The configurations vary four factors: input information (*Lex*, *Lex+POS*, and *Lex+POS+Gloss*), prompt language (Arabic or English), lemma diacritization (diacritized or dediacritized), and number of shots (zero-shot or five-shot). Table 7 reports the average performance of each model when each parameter value is fixed, marginalizing over the remaining factors. Within the tested prompts and model versions, *GPT-5.5-Pro* achieves the strongest and most stable performance, obtaining the best results across most prompt settings and the lowest standard deviation.

Model	Dev						Test					
	All	Adj	Noun	Verb	PropN	Cross	All	Adj	Noun	Verb	PropN	Cross
<i>Readability Lexicons</i>												
SAMER-5	66.6	67.1	66.0	68.2	64.3	67.4	67.7	69.9	68.4	67.1	67.3	66.1
SAMER-13	<u>73.0</u>	<u>72.9</u>	<u>73.7</u>	<u>74.7</u>	<u>69.8</u>	<u>73.8</u>	<u>74.8</u>	<u>77.2</u>	<u>76.1</u>	<u>74.4</u>	<u>73.0</u>	<u>73.2</u>
BAREC-based Lexicon	65.0	70.6	67.7	68.1	53.8	64.8	66.2	68.4	68.0	67.2	61.1	66.2
BAREC-10M-based Lexicon	63.5	64.9	63.0	60.6	64.0	65.0	65.2	65.4	62.2	65.4	67.2	65.6
<i>Frequency Lexicons</i>												
CAMeLBERT Freq.	<u>66.6</u>	<u>70.2</u>	66.9	59.0	<u>67.6</u>	<u>69.2</u>	<u>68.2</u>	70.8	67.1	62.4	<u>72.2</u>	<u>68.6</u>
BAREC-1M Freq.	63.8	67.7	<u>67.1</u>	<u>70.4</u>	51.0	63.0	65.8	70.4	<u>67.9</u>	67.9	58.9	63.8
BAREC-10M Freq.	65.3	65.2	65.6	68.7	64.4	62.6	68.1	<u>71.9</u>	67.5	<u>68.7</u>	66.4	65.8
<i>Trainable Models</i>												
Classical ML	<u>80.4</u>	82.0	<u>79.6</u>	<u>79.6</u>	<u>79.8</u>	81.3	<u>81.1</u>	83.5	<u>81.9</u>	<u>78.4</u>	<u>82.2</u>	<u>79.4</u>
AraBERTv02	75.3	78.6	71.4	73.1	78.1	75.4	75.6	76.8	75.8	70.9	80.3	74.0
CAMeLBERT-msa	75.0	77.9	74.3	70.2	78.0	74.3	75.0	77.0	73.1	71.6	78.8	74.4
MARBERTv2	72.1	74.2	72.0	64.9	75.0	74.2	72.8	76.0	72.5	65.6	76.4	73.4
<i>Large Language Models (LLMs)</i>												
ALLaM-7B-Instruct	72.9	75.8	73.9	70.3	71.4	73.1	72.0	74.1	69.9	68.3	76.8	71.0
Fanar-2-27B-Instruct	74.4	77.0	77.6	71.3	76.0	69.9	75.0	78.4	74.8	72.1	78.7	71.2
Qwen3-30B-A3B-Instruct	74.6	77.6	77.6	73.7	73.1	71.1	74.6	77.2	73.7	71.3	77.9	72.9
Aya-Expansive-32B	76.5	76.2	76.0	75.1	79.6	75.6	75.8	75.7	74.0	76.3	78.2	74.7
Jais-2-70B-Chat	75.6	77.7	74.3	74.0	76.3	75.6	75.8	77.7	74.2	73.6	78.6	75.2
Llama-3.3-70B-Instruct	73.9	75.3	79.1	68.1	75.7	71.3	74.5	77.2	73.8	69.9	80.8	70.8
Gemini-3.1-Pro	74.5	76.3	70.9	79.1	77.0	69.1	75.0	78.0	72.3	76.7	77.2	70.7
GPT-5.5-Pro	81.5	<u>81.0</u>	82.8	85.1	81.2	<u>77.2</u>	81.7	<u>83.2</u>	82.2	81.3	82.4	79.6

Table 5: Accuracy (%) of all models on the *Dev* and *Test* sets across parts of speech. PropN = Proper Nouns. **Bold** indicates the best performance for each POS, and underline indicates the best model in each group.

The number of shots shows the clearest effect in our experiments, with few-shot prompting improving most models. Among the Arabic-centric LLMs, *ALLaM* benefits most from Arabic prompts, while *Fanar* shows a smaller gain. *Jais*, however, performs almost identically with Arabic and English prompts, suggesting that Arabic-centric models do not uniformly benefit from Arabic prompting under the tested settings. Other prompt factors show model-dependent effects. Because LLM behavior can change with model updates and prompt design choices, these results should be interpreted as a snapshot of the evaluated models and configurations rather than as fixed conclusions about each model. Appendix D.1 shows the results across all configurations.

5.4 Error Analysis

We analyzed the 4,500 Dev pairs to compare annotator agreement and SAMER lexicon consistency against the best model’s (*GPT-5.5-Pro*) accuracy. SAMER agreement reflects whether lexical distances align with gold labels; distance 0 is *Ambivalent*. Annotator agreement distinguishes first-round consensus (higher agreement) from tie-break cases. Table 8 shows that model accuracy rises with annotator/SAMER agreement, reaching 94.8% in 56.2%

of cases. The lowest accuracy (36.6%) occurs when both SAMER and annotators disagree, highlighting the hardest lemma pairs for humans and models.

6 Downstream Uses

In this section, we present two downstream applications of our benchmark dataset: constructing a ranked readability lexicon covering all open-class lemmas (**BAREC-LexRank**), and improving sentence-level readability assessment.

6.1 Ranked Readability Lexicon

We use the *Classical ML* model as a pairwise comparison function to rank all open-class lemmas. Although this model is not the best-performing system on the *Dev* set, it is substantially cheaper to run while performing within 1.1% of the best model.

Ranking Algorithm We first generate the required features for all lemmas (see §4.3). We then apply a comparison-based sorting algorithm, where the *Classical ML* model determines which lemma in each pair is easier. Since the model is not perfect, its pairwise predictions are not guaranteed to be globally consistent; it produces cyclic preferences such as $\text{LexA} < \text{LexB} < \text{LexC} < \text{LexA}$. As a result, the final ranking will depend on the ini-

Dropped Features	Best Classical ML Performance					
	All	Adj	Noun	Verb	PropN	Cross
None	80.3	81.6	79.4	79.8	79.9	80.8
All Ar	61.9	66.4	58.4	57.4	66.6	60.8
All Gloss	80.4	82.0	79.6	79.6	79.8	81.3
ArFreq	79.0	80.0	77.8	79.3	79.9	78.0
ArSAMER	70.2	74.3	67.6	63.0	75.8	70.2
ArLength	80.2	81.2	79.7	79.8	79.1	81.0
ArCharFreq	80.0	81.1	79.7	80.1	78.9	80.4
ArRoot	80.3	81.7	79.2	79.7	79.7	81.1
ArAbsPattern	80.3	81.4	80.1	79.1	80.2	80.6
GlossFreq	80.4	82.4	79.9	79.7	79.4	80.7
GlossRL	80.3	81.0	79.3	79.7	80.1	81.4
GlossWordCount	80.2	81.8	79.3	79.7	79.6	80.8
GlossCharCount	80.3	81.2	79.4	79.7	79.8	81.2
GlossSyllableCount	80.3	81.7	79.4	79.3	80.2	80.8

Table 6: Ablation results for the Classical ML model on the *Dev* set. Each row reports the best performance after removing one feature group; *None* corresponds to the full feature set. **Bold** indicates the best performance for each POS.

tial ordering of the lemmas. To reduce this effect, we repeat the ranking process using 20 different random initializations and assign each lemma its average rank across the 20 runs.

We use the same procedure to generate two additional variants of the ranked lexicon. The first variant, **BAREC-LexComplete**, ranks all lemmas, including closed-class lemmas, to produce a complete ranked resource. The second variant, **BAREC-LexSAMER**, preserves the SAMER level ordering: we use the same sorting algorithm, but the *Classical ML* model is used only to break ties between lemmas from the same SAMER level. We include both variants in the release and report their downstream results in Appendix H.

Results We evaluate BAREC-LexRank intrinsically on our benchmark dataset. For each lemma pair in the *Dev* and *Test* sets, we directly compare the two lexicon ranks and select the higher-ranked lemma as easier. BAREC-LexRank achieves an accuracy of 80.0% on *Dev* and 81.1% on *Test*, ranking third after *GPT-5.5-Pro* and *Classical ML*. In the next section, we evaluate the lexicon extrinsically by measuring its impact on sentence-level readability assessment.

6.2 Sentence-Level Readability Assessment

In this section, we explore several ways of using BAREC-LexRank to improve sentence-level readability assessment.

Setup We use the BAREC Corpus (Elmadani et al., 2025b) for training and evaluation. The corpus contains 69,441 sentences annotated with 19 readability levels following the BAREC annotation framework (Habash et al., 2025). We compare our methods against the best system reported in their paper, *AraBERTv2+D3Tok+Reg*.

Zero-Shot Lexicon-Based Prediction As a first technique, we use BAREC-LexRank for sentence-level readability prediction without task-specific training; we refer to this model as **LexRank-ZS**. The method combines the ranked open-class lemma lexicon with the closed-class lemma annotations, which are already mapped to the 19 BAREC levels.

For the ranked lexicon, we map lemma ranks to the 19-level BAREC scale using monotonic functions. The mappings are motivated by the strong correlation between frequency and readability: easier lemmas tend to be more frequent. We therefore assign smaller vocabularies to lower levels and larger vocabularies to higher levels, approximating the skewed distribution expected under Zipf’s law. The full set of mappings and their results are described in Appendix F. For each sentence with a mixture of open- and closed-class lemmas, we assign it the level of the hardest lemma.

Classical ML Models with Lexicon-Based Features As a second technique, we use BAREC-LexRank and closed-class annotations to extract sentence-level features for training Classical ML models; we refer to this family of models as **LexRank-ML**. For each sentence, we compute 11 interpretable features summarizing lexical difficulty, including sentence length, the proportion of lemmas in the easiest and hardest rank percentiles, and the level of the hardest closed-class lemma. The full feature list is provided in Appendix G. This setup evaluates how much readability signal can be captured from lexicon-derived features alone, without contextual representations.

BERT with Lexicon-Based Sentence Features As a third technique, we augment the BERT readability model with the same 11 lexicon-based sentence features from Section 6.2. We concatenate these features with the final [CLS] representation and pass the combined vector to the prediction head. The resulting model, *AraBERTv2+D3Tok+Reg (+LexRank)*, is fine-tuned end-to-end for readability prediction. This setup tests whether explicit lexical difficulty features provide complementary

Model	Inputs			Prompt Lang.		Lex Diacritization		# Shots		STD
	Lex	Lex+POS	Lex+POS+Gloss	Ar	En	Diac.	Dediac.	Few	Zero	
ALLaM-7B-Instruct	66.7	63.2	65.5	69.3	61.0	66.4	63.8	63.5	66.7	5.5
Fanar-2-27B-Instruct	70.4	66.7	68.5	69.7	67.4	68.5	68.6	73.0	64.1	5.3
Qwen3-30B-A3B-Instruct	69.5	66.8	68.5	69.4	67.2	68.3	68.3	72.8	63.7	5.0
Aya-Expansive-32B	64.2	64.7	64.5	62.4	66.6	64.2	64.8	74.3	54.7	11.1
Llama-3.3-70B-Instruct	67.3	66.0	70.3	68.8	66.9	66.9	68.9	65.3	70.4	3.9
Jais-2-70B-Chat	73.8	71.1	70.6	71.8	71.9	72.2	71.5	73.0	70.7	2.3
Gemini-3.1-Pro	65.7	66.7	64.9	63.2	68.3	66.6	64.9	71.0	60.5	6.1
GPT-5.5-Pro	79.5	80.8	80.7	80.1	80.5	81.0	79.7	80.4	80.2	1.0

Table 7: Prompt-sensitivity analysis for LLMs. Each model is evaluated under 24 prompt configurations, varying the input information, prompt language, lemma diacritization, and number of shots. Each column reports the average accuracy when the corresponding parameter value is fixed. The final column reports the standard deviation across all 24 configurations.

Annotator	SAMER Category			
	Agree	Ambivalent	Disagree	
First Round	94.8 (56.2)	75.2 (14.0)	50.8 (8.2)	
Tie-Break	81.7 (10.2)	52.7 (6.1)	36.6 (5.3)	

Table 8: The best model’s (*GPT-5.5-Pro*) accuracy (%) on the *Dev* set. Values in parentheses indicate the percentage of cases.

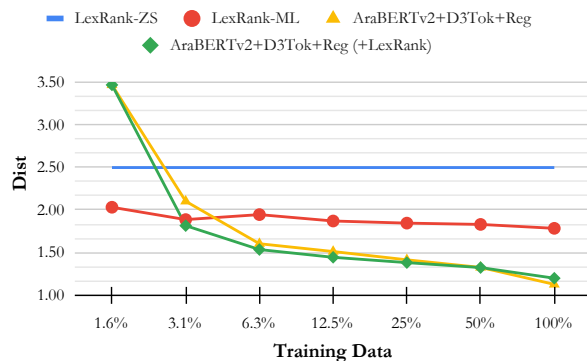


Figure 1: Learning-curve results on the BAREC *Dev* set using average absolute distance (**Dist**); lower is better.

information beyond BERT’s contextual representation.

Results We evaluate the three techniques described above and compare them with the best model from Elmadani et al. (2025b), *AraBERTv2+D3Tok+Reg*. To examine whether the benefit of the ranked lexicon depends on the amount of available training data, we conduct a learning-curve experiment using 1.6%, 3.1%, 6.3%, 12.5%, 25%, 50%, and 100% (54,845 sentences) of the training set. We evaluate all models on the BAREC *Dev* set using the average absolute distance between the predicted and gold readability levels (**Dist**).

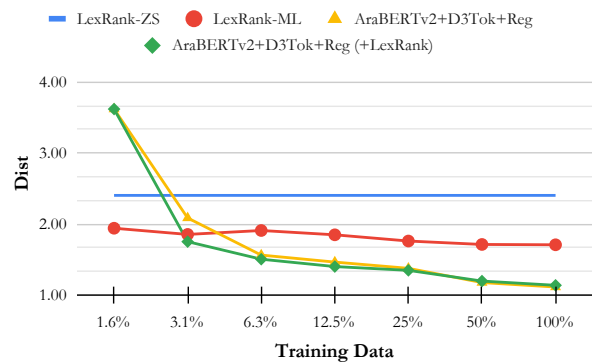


Figure 2: Learning-curve results on the BAREC *Test* set using average absolute distance (**Dist**); lower is better.

Learning Curve Figure 1 shows the learning-curve results for sentence-level readability assessment on the BAREC *Dev* set. The zero-shot *LexRank-ZS* model achieves an average distance of 2.5 from the gold labels without using any sentence-level training data, showing that the ranked lexicon alone provides a useful readability signal. In the lowest-resource setting, *LexRank-ML* is the best-performing model, while both BERT-based models struggle to learn effectively from this amount of data. At 3.1% of the training data, *LexRank-ML* still outperforms *AraBERTv2+D3Tok+Reg*; however, adding *LexRank* features allows *AraBERTv2+D3Tok+Reg (+LexRank)* to become the best model. As the amount of training data increases, the BERT-based models improve substantially, while *LexRank-ML* benefits only modestly from additional data. The benefit of adding *LexRank* features to BERT is most clear in low-resource settings: *AraBERTv2+D3Tok+Reg (+LexRank)* outperforms the baseline BERT model when using 3.1% to 25% of the training data. This suggests that the ranked

Train-Unseen		3.1% Training Data				100% Training Data				
LexRank-Covered Lemmas	Dev Sentences		Base	+LexRank	Δ	Dev Sentences		Base	+LexRank	Δ
	#	%				#	%			
1	1,278	17%	2.23	2.23	0.00	1,867	26%	1.15	1.18	-0.04
2	1,149	16%	1.94	1.85	0.09	889	12%	1.08	1.15	-0.07
3	866	12%	1.95	1.76	0.19	403	6%	0.94	0.96	-0.02
4	738	10%	1.93	1.59	0.35	211	3%	1.00	1.03	-0.03
5	628	9%	1.97	1.61	0.37	79	1%	0.73	0.77	-0.04
6+	1,869	26%	2.22	1.55	0.68	92	1%	0.89	0.93	-0.04
All	6,528	89%	2.08	1.77	0.31	3,541	48%	1.08	1.12	-0.04

Table 9: Seen-word analysis on the BAREC *Dev* set. *Train-Unseen LexRank-Covered Lemmas* indicates the number of words in a sentence whose lemmas are absent from the training subset but covered by *LexRank*. *Base* refers to *AraBERTv2+D3Tok+Reg*, and *+LexRank* refers to *AraBERTv2+D3Tok+Reg (+LexRank)*. Δ is the improvement from adding *LexRank*; positive values indicate lower average distance.

lexicon is especially useful in low-resource settings, either as a standalone model or as complementary signal for neural models.

Seen-Word Analysis We analyze the performance gap between *AraBERTv2+D3Tok+Reg (+LexRank)* and *AraBERTv2+D3Tok+Reg* in two settings: using 3.1% of the training data, where adding *LexRank* yields the largest improvement, and using the full training set, where the baseline no longer benefits from *LexRank*. Our goal is to understand why *LexRank* becomes less useful as more training data is available. To do so, we compare the vocabulary coverage of each training subset against the BAREC *Dev* set. We focus on *Dev*-set sentences that contain words whose lemmas are absent from the training subset but are covered by *LexRank*, since these are cases where the lexicon can provide readability information not observed during training.

Table 9 analyzes when *LexRank* helps the BERT-based readability model by grouping *Dev* sentences according to the number of words whose lemmas are absent from the training data but covered by *LexRank*. In the 3.1% training setting, such cases are very common, covering 89% of the *Dev* set, and adding *LexRank* reduces the average distance from 2.08 to 1.77. The improvement becomes larger as the number of train-unseen lemmas increases, reaching 0.68 for sentences with six or more such words. This suggests that *LexRank* is especially useful when the training data does not expose the model to many of the lemmas appearing at evaluation time. With the full training set, the coverage gap is much smaller: only 48% of *Dev* sentences contain train-unseen lemmas covered by *LexRank*, and adding *LexRank* no longer improves perfor-

mance. Although these sentences still contain train-unseen lemmas covered by *LexRank*, the baseline BERT model has access to enough supervised data to learn strong lexical and contextual readability signals. In this setting, the lemma-level features become mostly redundant and may add slight noise, since sentence readability also depends on contextual and syntactic factors beyond individual lemma difficulty. This explains why the lexicon helps most in low-resource settings but becomes less useful once the model has seen enough lexical evidence during training.

7 Conclusions and Future Work

We introduce **BAREC-LexBench**, the first benchmark for Arabic lemma-level readability assessment, designed to evaluate lexicons and computational models independently of specific granularity levels. Our large-scale annotation effort covers 36K lemma pairs across same- and cross-part-of-speech comparisons. Our analysis further shows that annotator disagreements are concentrated among lemma pairs with similar readability levels, highlighting the difficulty of fine-grained lexical readability judgments. Through evaluation, we show trainable models, especially Classical ML and stronger LLMs, achieve the best performance. Finally, we demonstrate downstream uses of the benchmark by constructing **BAREC-LexRank**, a ranked readability lexicon, and applying it to sentence-level readability assessment. BAREC-LexRank is especially useful in low-resource settings, where it provides lexical readability information not available from limited training data. This benchmark establishes a foundation for future research on continuous-scale Arabic readability lexicons.

Acknowledgments

The BAREC project is supported by the Abu Dhabi Arabic Language Centre (ALC) / Department of Culture and Tourism, UAE. This research is in part with support from Google.org and the Google Cloud Research Credits program for the Gemini Academic Program. We also acknowledge the support of the High Performance Computing Center at New York University Abu Dhabi. Finally, we are deeply grateful to our outstanding annotation team: Mirvat Dawi, Reem Faraj, Rita Raad, Sawsan Tanir, and Adel Wizani, Samar Zeino, and Zeina Zeino.

Limitations

A notable limitation of this work is the inherent subjectivity of readability assessment, which can lead to disagreements among annotators despite our best efforts to maintain consistency. Additionally, the current version of the benchmark does not capture all the open-class lemmas, it includes 15% of adjectives, 5% of nouns, 11% of verbs, and 18% of proper nouns, in favor of ensuring extensive evaluation per lemma (at least 14 appearances per lemma). Finally, while readability measures can offer valuable educational and linguistic insights, they could also be misused for profiling purposes; this is not our intention, and we discourage it.

Ethics Statement

The annotation process is conducted with transparency and fairness, with multiple annotators involved to mitigate biases and ensure reliability. All annotators are paid fair wages for their contribution. The benchmark dataset will be openly accessible to promote transparency, reproducibility, and collaboration in Arabic language research.

We used AI writing assistance within the scope of “Assistance purely with the language of the paper” described in the ACL Policy on Publication Ethics.

References

- Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021. [ARBERT & MARBERT: Deep bidirectional transformers for Arabic](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7088–7105, Online. Association for Computational Linguistics.
- M Al-Dawsari. 2004. The assessment of readability books content (boys-girls) of the first grade of intermediate school according to readability standards. *Sultan Qaboos University, Muscat*.
- Hend S Al-Khalifa and Amani A Al-Ajlan. 2010. Automatic readability measurements of the Arabic text: An exploratory study. *Arabian Journal for Science and Engineering*, 35(2 C):103–124.
- Muhamed Al Khalil, Nizar Habash, and Zhengyang Jiang. 2020. [A large-scale leveled readability lexicon for Standard Arabic](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3053–3062, Marseille, France. European Language Resources Association.
- Muhamed Al Khalil, Hind Saddiki, Nizar Habash, and Latifa Alfalasi. 2018. A Leveled Reading Corpus of Modern Standard Arabic. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Miyazaki, Japan.
- Bashar Alhafni, Reem Hazim, Juan David Pineros Liberato, Muhamed Al Khalil, and Nizar Habash. 2024. [The SAMER Arabic text simplification corpus](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16079–16093, Torino, Italia. ELRA and ICCL.
- Richard L Allington, Kimberly McCuiston, and Monica Billen. 2015. What research says about text complexity and learning to read. *The Reading Teacher*, 68(7):491–501.
- Sultan Almujaivel, Damith Premasiri, Tharindu Ranasinghe, Mo El-Haj, and Ruslan Mitkov. 2025. [Complex concept-based readability estimation from arabic curriculum](#). *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 24(11).
- Kinda Altarbouch, Khalid N. Elmadani, Ossama Obeid, Hanada Taha, and Nizar Habash. 2025. [BAREC demo: Resources and tools for sentence-level Arabic readability assessment](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 950–959, Suzhou, China. Association for Computational Linguistics.
- Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. [AraBERT: Transformer-based model for Arabic language understanding](#). In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15, Marseille, France. European Language Resource Association.
- Mohamed Anwar, Abdelhakim Freihat, George Ibrahim, Mostafa Awad, Abdelrahman Atef Mohamed Ali Sadallah, Gurpreet Gosal, Gokul Ramakrishnan, Sarath Chandran, Biswajit Mishra, Rituraj Joshi, Ahmed Frikha, Etienne Goffinet, Abhishek Maiti, Ali El Filali, Sarah Al Barri, Samujjwal Ghosh, Rahul Pal, Parvez Mullah, Awantika Shukla, Sajid Siddiki, Samta Kamboj, Onkar Pandit, Sunil Sahu,

- Abelrahman El Badawy, Amr Mohamed, Ahmad Chamma, Evan Dufraisse, Abdelaziz Bounhar, Dani Bouch, Hadi Abdine, Guokan Shang, Fajri Koto, Yuxia Wang, Zhuohan Xie, Ali Mekky, Rania Hosam Elmoahady Elbadry, Sarfraz Ahmad, Momina Ahsan, Omar Emad Mohamed El-Herraoui, Daniil Orel, Hasan Iqbal, Kareem Mohamed Naguib Abdelmohsen Fahmy Elzeky, Mervat Abassy, Kareem Ali, Saadeldine Eletter, Farah Atif, Nurdaulet Mukhituly, Haonan Li, Xudong Han, Aaryamonvikram Singh, Zain Quraishi, Neha Sengupta, Larry Murray, Avraham Sheinin, Joel Hestness, Natalia Vassilieva, Hector Ren, Zhengzhong Liu, Michalis Vazirgiannis, and Preslav Nakov. 2025. Jais 2: A family of Arabic-centric open large language models. Technical report, IFM.
- Amelia T. Barber and Susan L. Klauda. 2020. [How reading motivation and engagement enable reading achievement: Policy implications](#). *Policy Insights from the Behavioral and Brain Sciences*, 7(1):27–34.
- M Saiful Bari, Yazeed Alnumay, Norah A. Alzahrani, Nouf M. Alotaibi, Hisham Abdullah Alyahya, Sultan AlRashed, Faisal Abdulrahman Mirza, Shaykhah Z. Alsubaie, Hassan A. Alahmed, Ghadah Alabduljabbar, Raghad Alkhathran, Yousef Almushayqih, Raneem Alnajim, Salman Alsubaihi, Maryam Al Mansour, Saad Amin Hassan, Dr. Majed Alrubaiyan, Ali Alammari, Zaki Alawami, Abdulmohsen AlThubaity, Ahmed Abdelali, Jeril Kuriakose, Abdalghani Abujabal, Nora Al-Twairish, Areeb Alowisheq, and Haidar Khan. 2025. [ALLam: Large language models for arabic and english](#). In *The Thirteenth International Conference on Learning Representations*.
- J Cohen. 1960. A coefficient of agreement for nominal scales. educational and psychosocial measurement, 20, 37–46.
- John Dang, Shivalika Singh, Daniel D’souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, Acyr Locatelli, Hangyu Lin, Dwarak Talupuru, Bharat Venkitesh, David Cairuz, Bowen Yang, Tim Chung, Wei-Yin Ko, Sylvie Shang Shi, Amir Shukayev, Sammie Bae, Aleksandra Piktus, Roman Castagné, Felipe Cruz-Salinas, Eddie Kim, Lucas Crawhall-Stein, Adrien Morisot, Sudip Roy, Phil Blunsom, Ivan Zhang, Aidan Gomez, Nick Frosst, Marzieh Fadaee, Beyza Ermis, Ahmet Üstün, and Sara Hooker. 2024. [Aya expanse: Combining research breakthroughs for a new multilingual frontier](#). Preprint, arXiv:2412.04261.
- William H DuBay. 2004. The principles of readability. *Online Submission*.
- Mahmoud El-Haj and Paul Rayson. 2016. [OSMAN — a novel Arabic readability metric](#). In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pages 250–255, Portorož, Slovenia. European Language Resources Association (ELRA).
- Mo El-Haj, Sultan Almujaivel, Damith Premasiri, Tharindu Ranasinghe, and Ruslan Mitkov. 2024. [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*, pages 103–113, Torino, Italia. ELRA and ICCL.
- Khalid N. Elmadani, Bashar Alhafni, Hanada Taha, and Nizar Habash. 2025a. [BAREC shared task 2025 on Arabic readability assessment](#). In *Proceedings of The Third Arabic Natural Language Processing Conference: Shared Tasks*, pages 239–252, Suzhou, China. Association for Computational Linguistics.
- Khalid N. Elmadani, Nizar Habash, and Hanada Taha-Thomure. 2025b. [A large and balanced corpus for fine-grained Arabic readability assessment](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 16376–16400, Vienna, Austria. Association for Computational Linguistics.
- Khalid N. Elmadani, Adel Mahmoud Wizani, Hanada Taha Thomure, and Nizar Habash. 2026. [A large and balanced multi-domain arabic corpus annotated for morphology, syntax, and readability](#). In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 11761–11775, Palma, Mallorca, Spain. European Language Resources Association (ELRA).
- Jonathan Forsyth. 2014. Automatic readability prediction for modern standard Arabic. In *Proceedings of the Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT)*.
- Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac’h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2024. [The language model evaluation harness](#).
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, et al. 2024. [The llama 3 herd of models](#). Preprint, arXiv:2407.21783.
- Nizar Habash and David Palfreyman. 2022. [ZAEBCU: An annotated Arabic-English bilingual writer corpus](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 79–88, Marseille, France. European Language Resources Association.
- Nizar Habash, Hanada Taha-Thomure, Khalid N. Elmadani, Zeina Zeino, and Abdallah Abushmaes. 2025. [Guidelines for fine-grained sentence-level Arabic readability annotation](#). In *Proceedings of the 19th Linguistic Annotation Workshop (LAW-XIX-2025)*, pages 359–376, Vienna, Austria. Association for Computational Linguistics.
- Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2021. [The interplay](#)

- of variant, size, and task type in Arabic pre-trained language models. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 92–104, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.
- Salam Khalifa, Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2021. [Camel Arabic Frequency Lists](#).
- G.R. Klare. 1963. *The Measurement of Readability*. Iowa State University Press.
- Juan Liberato, Bashar Alhafni, Muhamed Khalil, and Nizar Habash. 2024. [Strategies for Arabic readability modeling](#). In *Proceedings of The Second Arabic Natural Language Processing Conference*, pages 55–66, Bangkok, Thailand. Association for Computational Linguistics.
- Mounica Maddela and Wei Xu. 2018. [A word-complexity lexicon and a neural readability ranking model for lexical simplification](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3749–3760, Brussels, Belgium. Association for Computational Linguistics.
- Tarek Naous, Michael J Ryan, Anton Lavrouk, Mohit Chandra, and Wei Xu. 2024. [ReadMe++: Benchmarking multilingual language models for multi-domain readability assessment](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12230–12266, Miami, Florida, USA. Association for Computational Linguistics.
- Ossama Obeid, Nasser Zalmout, Salam Khalifa, Dima Taji, Mai Oudah, Bashar Alhafni, Go Inoue, Fadhil Eryani, Alexander Erdmann, and Nizar Habash. 2020. [CAMEL tools: An open source python toolkit for Arabic natural language processing](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 7022–7032, Marseille, France. European Language Resources Association.
- F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.
- Hind Saddiki, Nizar Habash, Violetta Cavalli-Sforza, and Muhamed Al Khalil. 2018. Feature optimization for predicting readability of arabic l1 and l2. In *Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications*, pages 20–29.
- Mostafa Saeed and Nizar Habash. 2025. [Lemmatization as a classification task: Results from arabic across multiple genres](#). *Preprint*, arXiv:2506.18399.
- Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, Akshay Nathan, Alan Luo, Alec Helyar, Aleksander Madry, Aleksandr Efremov, Aleksandra Spyra, Alex Baker-Whitcomb, Alex Beutel, Alex Karpenko, Alex Makelov, Alex Neitz, Alex Wei, Alexandra Barr, Alexandre Kirchmeyer, Alexey Ivanov, Alexi Christakis, Alistair Gillespie, Allison Tam, Ally Bennett, Alvin Wan, Alyssa Huang, Amy McDonald Sandjideh, Amy Yang, Ananya Kumar, Andre Saraiva, Andrea Vallone, Andrei Gheorghe, Andres Garcia Garcia, Andrew Braunstein, Andrew Liu, Andrew Schmidt, Andrey Mereskin, Andrey Mishchenko, Andy Applebaum, Andy Rogerson, Ann Rajan, Annie Wei, Anoop Kotha, Anubha Srivastava, Anushree Agrawal, Arun Vijayvergiya, Ashley Tyra, Ashvin Nair, Avi Nayak, Ben Eggers, Bessie Ji, Beth Hoover, Bill Chen, Blair Chen, Boaz Barak, Borys Minaiev, Botao Hao, Bowen Baker, Brad Lightcap, Brandon McKinzie, Brandon Wang, Brendan Quinn, Brian Fioca, Brian Hsu, Brian Yang, Brian Yu, Brian Zhang, Brittany Brenner, Callie Riggins Zetino, Cameron Raymond, Camillo Lugaresi, Carolina Paz, Cary Hudson, Cedric Whitney, Chak Li, Charles Chen, Charlotte Cole, Chelsea Voss, Chen Ding, Chen Shen, Chengdu Huang, Chris Colby, Chris Hallacy, Chris Koch, Chris Lu, Christina Kaplan, Christina Kim, CJ Minott-Henriques, Cliff Frey, Cody Yu, Coley Czarnecki, Colin Reid, Colin Wei, Cory Decareaux, Cristina Scheau, et al. 2026. [Openai gpt-5 system card](#). *Preprint*, arXiv:2601.03267.
- Rasha Soliman and Laila Familiar. 2024. Creating a CEFR Arabic vocabulary profile: A frequency-based multi-dialectal approach. *Critical Multilingualism Studies*, 11(1):266–286.
- Hanada Taha-Thomure. 2017. *Arabic Language Text Leveling* (معايير هنادا طه لتصنيف مستويات النصوص العربية). Educational Book House (دار الكتاب التربوي للنشر والتوزيع).
- FANAR TEAM, Ummar Abbas, Mohammad Shahmeer Ahmad, Minhaj Ahmad, Abdulaziz Al-Homaid, Anas Al-Nuaimi, Enes Altinisik, Ehsaneddin Asgari, Sanjay Chawla, Shammur Chowdhury, Fahim Dalvi, Kareem Darwish, Nadir Durrani, Mohamed Elfeky, Ahmed Elmagarmid, Mohamed Eltabakh, Asim Ersoy, Masoomali Fatehkia, Mohammed Qusay Hashim, Majd Hawasly, Mohamed Hefeeda, Mus’ab Husaini, Kevin Isufaj, Soon-Gyo Jung, Houssam Lachemat, Ji Kim Lucas, Abubakr Mohamed, Tasnim Mohiuddin, Basel Mousi, Hamdy Mubarak, Ahmad Musleh, Mourad Ouzzani, Amin Sadeghi, Husrev Taha Sencar, Mohammed Shinoy, Omar Sinan, and Yifan Zhang. 2026. [Fanar 2.0: Arabic generative ai stack](#). *Preprint*, arXiv:2603.16397.
- Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, et al. 2025. [Gemini: A family of highly capable multimodal models](#). *Preprint*, arXiv:2312.11805.
- Qwen Team. 2025. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

A License

We list below the licenses of the data and tools used in this work, all of which are employed in accordance with their intended use.

- SAMER Lexicon (Al Khalil et al., 2020): Non Commercial Use - NYUAD License.³
- Maddela and Xu's (2018) Word-Complexity Lexicon: MIT License
- BAREC Corpus (Elmadani et al., 2025b): Creative Commons Attribution Share Alike 4.0
- BAREC-10M Corpus (Elmadani et al., 2026): Creative Commons Attribution Share Alike 4.0
- CAMELBERT Frequency List (Khalifa et al., 2021): Creative Commons Attribution Share Alike 4.0
- CAMEL Tools (Obeid et al., 2020) and CAMELBERT (Inoue et al., 2021): MIT License
- ALLaM-7B-Instruct (Bari et al., 2025): Apache license 2.0
- Fanar-2-27B-Instruct (TEAM et al., 2026): Apache license 2.0
- Jais-2-70B-Chat (Anwar et al., 2025): Apache license 2.0
- Qwen3-30B-A3B-Instruct (Team, 2025): Apache license 2.0
- Aya-Expanse-32B (Dang et al., 2024): Creative Commons Attribution Non Commercial 4.0
- Llama-3.3-70B-Instruct (Grattafiori et al., 2024): LLAMA 3.3 COMMUNITY LICENSE AGREEMENT⁴
- AraBERTv2+D3Tok+Reg (Elmadani et al., 2025b): MIT License.

B Annotation Interface

Figure 3 shows the Google Sheet interface used for annotation. Annotators are given the lemma pairs and their parts-of-speech. From a drop-down menu, they select the number of the easier lemma, "0" for equal, or "X" for I don't know.

C Experimental Setup

We use the *Tune* set for hyperparameter tuning and prompt engineering. Classical machine learn-

³<https://camel.abudhabi.nyu.edu/samer-readability-lexicon/>

⁴<https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct/blob/main/LICENSE>

المفردة 1 Lex	قسم الكلام POS 1	المفردة 2 Lex	قسم الكلام POS 2	المفردة الأسهل Easier Lemma
أَشْرَفَ	فعل Verb	ناسا	اسم علم Proper Noun	1
ذَكَا	فعل Verb	حَنَاشَات	اسم علم Proper Noun	2
رَاجِبَ	صفة Adjective	حَطَّطَ	فعل Verb	1
دَامِغَ	صفة Adjective	طَلَّلَ	اسم Noun	0
حُقُوقِيَّ	صفة Adjective	ذَكَا	فعل Verb	1
مَعْسُولَ	صفة Adjective	سَيِّدِنِي	اسم علم Proper Noun	1
حَصِينَ	صفة Adjective	تَلَاَحَقَ	فعل Verb	2
أَنْثُرُوبُولُوجِيَا	اسم Noun	سَيِّمُونُوفَ	اسم علم Proper Noun	X
شَبِكِيَّ	صفة Adjective	تَطْفَلَ	فعل Verb	1
حَقَّ	صفة Adjective	رَتَّبَ	فعل Verb	2
لَمْ	فعل Verb	عَلَّمَ	اسم علم Proper Noun	2
إِخْرَاجَ	اسم Noun	عَنَّثَر	اسم علم Proper Noun	2
تَنَزَّعَ	فعل Verb	إِرْتِيْسِنَا	اسم علم Proper Noun	1
أَمَرَ	فعل Verb	إِلْيَاسَ	اسم علم Proper Noun	1

Figure 3: A screenshot of the Google Sheet interface used for annotation.

ing models are trained using scikit-learn (Pedregosa et al., 2011),⁵ BERT-based models are fine-tuned with the HuggingFace Transformers library (Wolf et al., 2019),⁶ and large language models (LLMs) are evaluated using the LM Evaluation Harness framework (Gao et al., 2024).⁷

For BERT fine-tuning, we use the dediacritized form of lemmas as input. We use CAMEL Tools (Obeid et al., 2020) for dediacritization.⁸ Training is performed on a single NVIDIA A100 GPU with 40GB of memory. We train for three epochs with a batch size of 16 and a learning rate of 2e-5. We select the model checkpoint that achieves the lowest loss on the *Tune* set. We experiment all trainable models with a single seed.

All open-weights LLM experiments are conducted on a two NVIDIA A100 GPUs with 80GB of memory. Benchmarking all six LLMs on the *Dev* set takes approximately four hours for each prompt configuration. *Gemini-3.1-Pro* costs \$11.16 on average per prompt configuration, while *GPT-5.5-Pro* costs \$67.52 on average per prompt configuration.

D Large Language Models (LLMs)

D.1 Experiments

We evaluate four experimental configurations across all LLMs:

- **Input information:** Lemma only, Lemma+POS, or Lemma+POS+Gloss
- **Prompt language:** English or Arabic
- **Lemma diacritization:** Diacritized vs. Dedicritized
- **Number of shots:** Zero-shot vs. Few-shot (5 examples)

We conduct 24 experiments per model, covering all possible combinations. Table 10 shows the performance of all LLMs under different configurations.

D.2 Prompts

The prompt changes based on two factors: the language and the input feature. Figures 4, 5, 6, 7, 8, and 9 show the different prompt templates depending on the input feature and the prompt language.

⁵scikit-learn v1.4.2

⁶Transformers v4.43.4

⁷lm-evaluation-harness v0.4.0

⁸CamelTools v1.5.5

Input Info	Prompt Lang.	diac/ dediac	Shots	ALLaM (7B)	Fanar2 (27B)	Qwen3 (30B)	Aya (32B)	Jais2 (70B)	Llama3.3 (70B)	Gemini 3.1 Pro	GPT 5.5 Pro
Lex	EN	diac	Zero Few	70.8 60.2	64.0 73.9	65.0 72.5	59.5 76.0	74.4 74.4	70.1 60.7	64.8 74.5	80.5 80.9
		dediac	Zero Few	68.1 56.0	65.6 74.1	65.7 73.2	60.9 76.3	74.4 73.6	72.9 62.8	62.1 71.9	78.3 78.7
	AR	diac	Zero Few	72.9 68.4	68.2 73.3	65.5 73.6	48.5 76.5	71.8 75.6	71.3 62.7	59.4 69.4	80.5 80.6
		dediac	Zero Few	70.1 67.4	69.9 74.4	65.8 74.6	39.8 76.3	72.6 74.0	73.9 64.3	55.4 68.1	77.9 78.2
Lex+POS	EN	diac	Zero Few	59.4 59.6	60.8 72.1	59.2 71.5	53.9 75.1	70.0 72.7	65.9 61.4	65.0 74.2	81.3 81.4
		dediac	Zero Few	58.2 54.1	60.4 73.2	59.3 71.3	58.1 75.1	71.0 73.0	68.8 63.6	63.4 74.2	80.7 80.3
	AR	diac	Zero Few	70.0 69.1	61.1 73.4	63.7 73.3	48.5 75.0	69.8 74.2	68.8 64.1	59.7 70.4	81.5 81.3
		dediac	Zero Few	69.4 65.4	58.9 73.5	62.7 73.5	57.6 74.1	65.5 72.8	70.3 64.9	57.6 69.2	79.7 80.1
Lex+POS+Gloss	EN	diac	Zero Few	61.9 63.2	62.3 70.4	63.6 71.6	60.9 71.3	69.6 70.5	68.9 68.2	63.4 73.0	81.1 81.1
		dediac	Zero Few	60.5 59.9	61.5 70.9	62.8 70.6	61.7 70.1	69.2 69.6	69.6 70.2	61.5 72.2	80.7 80.7
	AR	diac	Zero Few	70.7 70.9	69.4 73.1	65.3 74.5	50.8 73.6	70.2 73.2	71.3 69.0	57.6 68.3	80.7 80.6
		dediac	Zero Few	68.5 68.3	67.2 73.3	66.3 73.4	55.5 72.1	70.3 72.2	73.2 71.7	56.5 66.5	80.2 80.4

Table 10: Accuracy (%) of LLMs on the *Dev* set under different prompt configurations. **Bold** indicates the best performance for each LLM.

You are evaluating Arabic lexical readability for general readers.
Given two Arabic lemmas, determine which one is easier for a general Arabic-speaking audience to read and understand.
Prefer words that are:

- common in everyday language,
- familiar to non-specialists,
- simple and widely recognized.

Less readable words are often:

- rare,
- archaic,
- literary,
- highly formal,
- technical,
- or specialized.

Choose "equal" only if both lemmas are extremely similar in familiarity and difficulty.
Which Arabic lexical entry is easier for a general audience to read and comprehend?

A. Lemma: {{Lex1}}
B. Lemma: {{Lex2}}
C. Both are equally easy.

Answer with only A, B, or C.
Answer:

Figure 4: The Lemma only English prompt.

أنت تقيّم سهولة المفردات العربية بالنسبة للقارئ العام.
ستُعطى مفردتين عربيتين، ومهمتك هي تحديد أيهما أسهل قراءةً وفهماً لدى عامة الناطقين بالعربية.
فضّل الكلمات التي تكون:

- شائعة في الاستعمال اليومي،
- مألوفة لغير المتخصصين،
- بسيطة وسهلة التعرف.

الكلمات الأصعب غالباً تكون:

- نادرة،
- تراثية،
- أدبية،
- رسمية جداً،
- تقنية،
- أو متخصصة.

اختر أمتساويتان فقط إذا كانت الكلمتان متقاربتين جداً في السهولة والشيوع.
أي المدخلين المعجميين العربيين أسهل قراءةً وفهماً للقارئ العام؟

A. {{Lex1}} الكلمة:
B. {{Lex2}} الكلمة:
C. متساويتان في السهولة.
أجب فقط بـ A أو B أو C.
الإجابة:

Figure 5: The Lemma only Arabic prompt.

You are evaluating Arabic lexical readability for general readers.
Given two Arabic lemmas, determine which one is easier for a general Arabic-speaking audience to read and understand.
Prefer words that are:

- common in everyday language,
- familiar to non-specialists,
- simple and widely recognized.

Less readable words are often:

- rare,
- archaic,
- literary,
- highly formal,
- technical,
- or specialized.

Choose "equal" only if both lemmas are extremely similar in familiarity and difficulty.
Which Arabic lexical entry is easier for a general audience to read and comprehend?

A. Lemma: {{Lex1}}
Part of Speech: {{POS1}}

B. Lemma: {{Lex2}}
Part of Speech: {{POS2}}

C. Both are equally easy.

Answer with only A, B, or C.
Answer:

Figure 6: The Lemma+POS English prompt.

أنت تقيّم سهولة المفردات العربية بالنسبة للقارئ العام.
 ستُعطي مفردتين عربيتين، ومهمتك هي تحديد أيهما أسهل قراءةً وفهماً لدى عامة الناطقين بالعربية.
 فضّل الكلمات التي تكون:

- شائعة في الاستعمال اليومي،
- مألوفة لغير المتخصصين،
- بسيطة وسهلة التعرف.

الكلمات الأصعب غالباً تكون:

- نادرة،
- تراثية،
- أدبية،
- رسمية جداً،
- تقنية،
- أو متخصصة.

اختر آمتساويتان فقط إذا كانت الكلمتان متقاربتين جداً في السهولة والشيوع.
 أي المدخلين المعجميين العربيين أسهل قراءةً وفهماً للقارئ العام؟

A. {{Lex1}} الكلمة:
 قسم الكلام: {{POS1}}

B. {{Lex2}} الكلمة:
 قسم الكلام: {{POS2}}

C. متساويتان في السهولة.

أجب فقط بـ A أو B أو C.

الإجابة:

Figure 7: The Lemma+POS Arabic prompt.

You are evaluating Arabic lexical readability for general readers.
Given two Arabic lemmas, determine which one is easier for a general Arabic-speaking audience to read and understand.
Prefer words that are:

- common in everyday language,
- familiar to non-specialists,
- simple and widely recognized.

Less readable words are often:

- rare,
- archaic,
- literary,
- highly formal,
- technical,
- or specialized.

Use the English glosses only to clarify meaning. Judge the Arabic lemmas themselves. Choose "equal" only if both lemmas are extremely similar in familiarity and difficulty.

Which Arabic lexical entry is easier for a general audience to read and comprehend?

A. Lemma: {{Lex1}}
Part of Speech: {{POS1}}
Meaning: {{Gloss1}}

B. Lemma: {{Lex2}}
Part of Speech: {{POS2}}
Meaning: {{Gloss2}}

C. Both are equally easy.

Answer with only A, B, or C.

Answer:

Figure 8: The Lemma+POS+Gloss English prompt.

أنت تقيّم سهولة المفردات العربية بالنسبة للقارئ العام.
ستُعطي مفردتين عربيتين، ومهمتك هي تحديد أيهما أسهل قراءةً وفهماً لدى عامة الناطقين بالعربية.
فضّل الكلمات التي تكون:

- شائعة في الاستعمال اليومي،
- مألوفة لغير المتخصصين،
- بسيطة وسهلة التعرف.

الكلمات الأصعب غالباً تكون:

- نادرة،
- تراثية،
- أدبية،
- رسمية جداً،
- تقنية،
- أو متخصصة.

استخدم المعنى الإنجليزي فقط لتوضيح المقصود.
قيّم سهولة الكلمة العربية نفسها.
اختر متساويتان فقط إذا كانت الكلمتان متقاربتين جداً في السهولة والشيوع.
أي المدخلين المعجميين العربيين أسهل قراءةً وفهماً للقارئ العام؟

A. {{Lex1}} الكلمة:
{{POS1}} قسم الكلام:
{{Gloss1}} المعنى:

B. {{Lex2}} الكلمة:
{{POS2}} قسم الكلام:
{{Gloss2}} المعنى:

C. متساويتان في السهولة.

أجب فقط بـ A أو B أو C.

الإجابة:

Figure 9: The Lemma+POS+Gloss Arabic prompt.

E ML Experiments & Features Analysis

Table 11 presents the full set of Classical ML experiments across different feature-ablation setups. The feature groups include Arabic-side features, such as frequency, SAMER-based readability levels, lemma length, character frequency, root, and abstract pattern, as well as English gloss-based features, such as gloss frequency, gloss readability level, gloss word count, gloss character count, and gloss syllable count. For each setup, the table indicates which feature groups are included and reports the corresponding *Dev*-set accuracy for each Classical ML model.

F Zero-Shot Lexicon Mapping Functions

Let $N = 40,687$ be the number of lemmas in the ranked lexicon. Each lemma has a rank $r \in \{1, \dots, N\}$, where lower ranks correspond to easier lemmas and higher ranks correspond to harder lemmas. We define a monotonic mapping

$$f : \{1, \dots, N\} \rightarrow \{1, \dots, 19\}$$

such that

$$r_i \leq r_j \Rightarrow f(r_i) \leq f(r_j).$$

We consider several deterministic mappings.

Linear mapping.

$$f_{\text{linear}}(r) = \left\lceil 19 \cdot \frac{r}{N} \right\rceil.$$

Logarithmic mapping.

$$f_{\log}(r; a) = \left\lceil 19 \cdot \frac{\log(1 + ar/N)}{\log(1 + a)} \right\rceil, \quad a > 0.$$

Power-law mapping.

$$f_{\text{power}}(r; \gamma) = \left\lceil 19 \cdot \left(\frac{r}{N} \right)^\gamma \right\rceil, \quad 0 < \gamma < 1.$$

Table 12 compares different monotonic mappings for converting the ranked lexicon into the 19-level BAREC scale, with and without closed-class lemma annotations. The results show that adding closed-class information consistently improves performance for most mappings, confirming that closed-class lemmas provide important readability signals at the sentence level. Among the tested mappings, the power mapping with $\gamma = 0.5$ performs best, achieving a distance of 2.50 when combined with closed-class annotations.

G Lexicon-Based Sentence Features

For the classical machine learning experiments described in Section 6.2, we extract 11 sentence-level features from the ranked lemma lexicon and the closed-class readability annotations.

The extracted features are as follows:

- **word_count**: the number of words in the sentence.
- **mean_lemma_rank**: the mean lemma rank.
- **max_lemma_rank**: the maximum lemma rank in the sentence.
- **std_lemma_rank**: the standard deviation of lemma ranks.
- **lemma_ratio_top_1pct**: the proportion of sentence lemmas whose ranks fall in the top 1% of the ranked lexicon.
- **lemma_ratio_top_2pct**: the proportion of sentence lemmas whose ranks fall in the top 2% of the ranked lexicon.
- **lemma_ratio_top_5pct**: the proportion of sentence lemmas whose ranks fall in the top 5% of the ranked lexicon.
- **lemma_ratio_top_10pct**: the proportion of sentence lemmas whose ranks fall in the top 10% of the ranked lexicon.
- **lemma_ratio_bottom_50pct**: the proportion of sentence lemmas whose ranks fall in the bottom 50% of the ranked lexicon.
- **lemma_ratio_closed_classes**: the proportion of words in the sentence that correspond to closed-class lemmas.
- **max_rl_based_on_closed_classes**: the maximum readability level assigned to any closed-class lemma in the sentence.

H Ranked Lexicon Variants

In addition to the main ranked lexicon (BAREC-LexRank), we construct two variants to support different use cases. The first variant is **BAREC-LexComplete**, which applies the same comparison-based sorting procedure to all lemmas, including closed-class lemmas.

The second variant is **BAREC-LexSAMER**. This variant preserves the original SAMER readability-level ordering while adding finer-grained distinctions within each level. Specifically, we apply the same sorting algorithm, but restrict

Features		Setups													
ArFreq	✓		✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
ArSAMER	✓		✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
ArLength	✓		✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓
ArCharFreq	✓		✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓
ArRoot	✓		✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓
ArAbsPattern	✓		✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓
GlossFreq	✓	✓		✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓
GlossRL	✓	✓		✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓
GlossWordCount	✓	✓		✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓
GlossCharCount	✓	✓		✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓
GlossSyllableCount	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓
GradientBoosting	80.3	61.9	80.4	79.0	70.2	80.2	80.0	80.3	80.3	80.4	80.3	80.2	80.3	80.3	80.3
ExtraTrees	79.7	61.1	79.9	79.0	70.2	79.7	79.5	80.0	79.8	79.7	80.0	79.9	79.7	79.8	79.8
RandomForest	79.6	60.4	79.7	78.7	70.2	79.5	79.5	79.3	79.5	79.9	79.6	79.5	79.4	79.8	79.8
AdaBoost	77.6	60.5	78.3	77.4	67.9	77.5	78.4	77.3	77.6	78.1	78.4	77.6	78.0	77.5	77.5
DecisionTree	70.5	54.4	71.9	71.2	61.8	70.9	70.2	70.8	70.9	70.5	70.5	70.7	71.2	70.9	70.9
LogisticRegression	60.6	56.5	67.6	56.1	61.3	60.4	60.9	60.6	60.7	65.4	60.6	60.5	60.6	60.6	60.6
KNN	55.0	55.0	64.4	56.0	55.0	55.0	55.0	55.0	55.0	64.4	55.0	55.0	55.0	55.0	55.0
MultinomialNB	51.5	49.7	63.0	50.6	51.5	51.5	51.5	51.5	51.5	62.9	51.5	51.5	51.5	51.5	51.5
SVC	53.0	53.0	59.9	53.0	53.0	53.0	53.0	53.0	53.0	59.9	53.0	53.0	53.0	53.0	53.0

Table 11: Classical ML performance across feature ablation setups on the *Dev* set. Checkmarks indicate the feature groups included in each setup.

Mapping	Parameter	Dist	Dist (+Closed Classes)
Linear	–	4.52	2.90
Log	$a = 3.0$	3.45	2.60
Log	$a = 9.0$	3.21	2.70
Log	$a = 19.0$	3.29	2.95
Log	$a = 49.0$	3.67	3.48
Power	$\gamma = 0.25$	4.03	3.99
Power	$\gamma = 0.33$	3.28	3.17
Power	$\gamma = 0.5$	2.90	2.50
Power	$\gamma = 0.75$	3.59	2.59
Power	$\gamma = 1.0$	4.52	2.90
Closed Classes	–	4.04	–

Table 12: Zero-shot lexicon-based readability prediction results on the BAREC *Dev* set.

the *Classical ML* model to comparing lemmas that belong to the same SAMER level.

We evaluate both variants on sentence-level readability assessment using BAREC learning-curve experiments. Figures 10 and 11 show the *Dev* and *Test* results for the *BAREC-LexComplete*, respectively. Figures 12 and 13 show the corresponding results for the *BAREC-LexSAMER*.

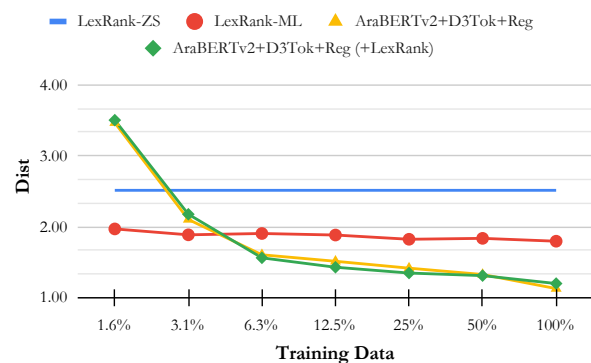


Figure 10: Learning-curve results on the BAREC *Dev* set using *BAREC-LexComplete*. Performance is measured by average absolute distance (**Dist**); lower is better.

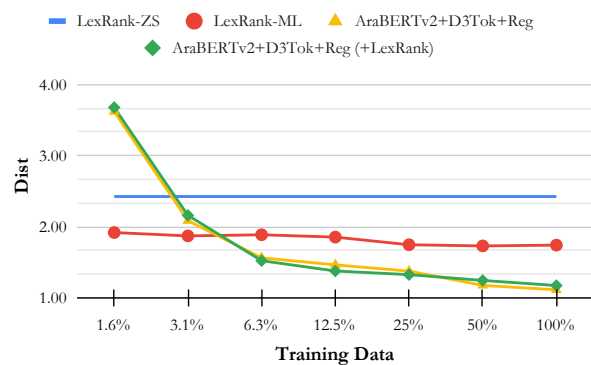


Figure 11: Learning-curve results on the BAREC *Test* set using *BAREC-LexComplete*. Performance is measured by average absolute distance (**Dist**); lower is better.

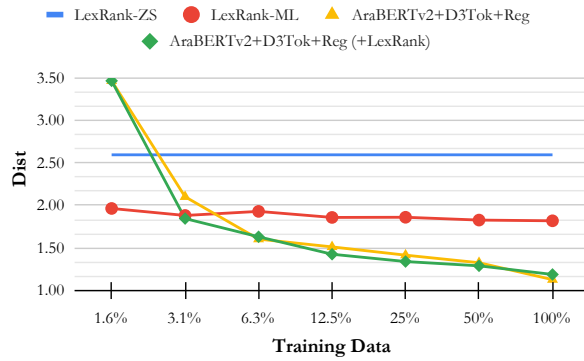


Figure 12: Learning-curve results on the BAREC *Dev* set using *BAREC-LexSAMER*. Performance is measured by average absolute distance (**Dist**); lower is better.

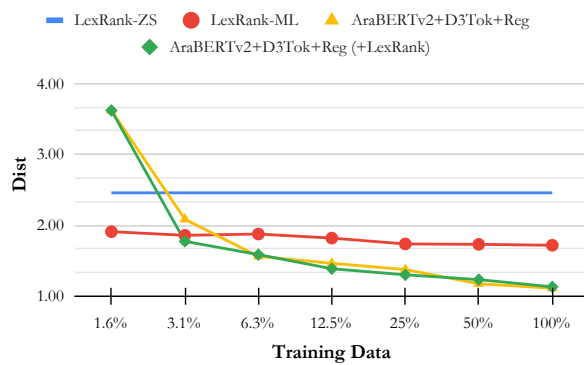


Figure 13: Learning-curve results on the BAREC *Test* set using *BAREC-LexSAMER*. Performance is measured by average absolute distance (**Dist**); lower is better.