A Large and Balanced Corpus for Fine-grained Arabic Readability Assessment

Khalid N. Elmadani,[†] Nizar Habash,[†] Hanada Taha-Thomure[‡]
[†]Computational Approaches to Modeling Language Lab, New York University Abu Dhabi
[‡]Zai Arabic Language Research Centre, Zayed University
{khalid.nabigh,nizar.habash}@nyu.edu, Hanada.Thomure@zu.ac.ae

Abstract

This paper introduces the Balanced Arabic Readability Evaluation Corpus (**BAREC**),[1] a large-scale, fine-grained dataset for Arabic readability assessment. **BAREC** consists of 69,441 sentences spanning 1+ million words, carefully curated to cover 19 readability levels, from kindergarten to postgraduate comprehension. The corpus balances genre diversity, topical coverage, and target audiences, offering a comprehensive resource for evaluating Arabic text complexity. The corpus was fully manually annotated by a large team of annotators. The average pairwise inter-annotator agreement, measured by Quadratic Weighted Kappa, is 81.8%, reflecting a high level of substantial agreement. Beyond presenting the corpus, we benchmark automatic readability assessment across different granularity levels, comparing a range of techniques. Our results highlight the challenges and opportunities in Arabic readability modeling, demonstrating competitive performance across various methods. To support research and education, we make **BAREC** openly available, along with detailed annotation guidelines and benchmark results.[2]

1 Introduction

Text readability impacts understanding, retention, reading speed, and engagement (DuBay, 2004). Texts above a student's readability level can lead to disengagement (Klare, 1963). Nassiri et al. (2023) highlighted that readability and legibility depend on both external features (e.g., production, fonts) and content. Text leveling in classrooms helps match books to students' reading levels, promoting independent reading and comprehension (Allington et al., 2015). Developing readability models is crucial for improving literacy, language learning, and academic performance.

Readability levels have long been a key component of literacy teaching and learning. One of the most widely used systems in English literacy is Fountas and Pinnell (Fountas and Pinnell, 2006), which employs qualitative measures to classify texts into 27 levels (A to Z+), spanning from kindergarten to adult proficiency. Similarly, Taha-Thomure (2017)'s system for Arabic has 19 levels from Arabic letters أ A to ق Q. These fine-grained levels are designed for pedagogical effectiveness, ensuring young readers experience gradual, measurable progress, particularly in early education (K–6) (Barber and Klauda, 2020). A key advantage is that they can be easily mapped to coarser levels with fewer categories, which may be more efficient for broader applications in readability research and automated assessments.

In this paper we present the Balanced Arabic Readability Evaluation Corpus (**BAREC**) – a large-scale fine-grained readability assessment corpus across a broad space of genres and readability levels. Inspired by the Taha/Arabi21 readability reference (Taha-Thomure, 2017), which has been instrumental in tagging over 9,000 children's books, **BAREC** seeks to establish a standardized framework for evaluating sentence-level[3] Arabic text readability across 19 distinct levels, ranging from kindergarten to postgraduate comprehension.

Our contributions are: (a) **a large-scale curated corpus** with 69K+ sentences (1M+ words) spanning diverse genres; and (b) **benchmarking of automatic readability assessment** models across multiple granularities, including both fine-grained (19 levels) and collapsed tiered systems (e.g., five-level and three-level scales) to support various research and application needs, aligning with previous Arabic readability frameworks (Al Khalil et al., 2018; Al-Khalifa and Al-Ajlan, 2010).

[1] بارق *bAriq* is Arabic for 'very bright and glittering'.

[2] http://barec.camel-lab.com

[3] We use *sentence* to refer to any standalone text segment, including phrases and single words (e.g., book titles).

Findings of the Association for Computational Linguistics: ACL 2025, pages 16376–16400
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

2 Related Work

Automatic Readability Assessment Automatic readability assessment has been widely studied, resulting in numerous datasets and resources (Collins-Thompson and Callan, 2004; Pitler and Nenkova, 2008; Feng et al., 2010; Vajjala and Meurers, 2012; Xu et al., 2015; Xia et al., 2016; Nadeem and Ostendorf, 2018; Vajjala and Lučić, 2018; Deutsch et al., 2020; Lee et al., 2021). Early English datasets were often derived from textbooks, as their graded content naturally aligns with readability assessment (Vajjala, 2022). However, copyright restrictions and limited digitization have driven researchers to crowdsource readability annotations from online sources (Vajjala and Meurers, 2012; Vajjala and Lučić, 2018) or leverage CEFR-based L2 assessment exams (Xia et al., 2016).

Arabic Readability Efforts Arabic readability research has focused on text leveling and assessment across various frameworks. Taha-Thomure (2017) proposed a 19-level system for children's books based on qualitative and quantitative criteria. Other efforts applied CEFR leveling to Arabic, including the KELLY project's frequency-based word lists, manually annotated corpora (Habash and Palfreyman, 2022; Naous et al., 2024), and vocabulary profiling (Soliman and Familiar, 2024). El-Haj et al. (2024) introduced DARES, a readability assessment dataset collected from Saudi school materials. The SAMER project (Al Khalil et al., 2020) developed a lexicon with a five-level readability scale, leading to the first manually annotated Arabic parallel corpus for text simplification (Alhafni et al., 2024). Automated readability assessment has also been explored through rule-based and machine learning approaches. Early models relied on surface-level features like word and sentence length (Al-Dawsari, 2004; Al-Khalifa and Al-Ajlan, 2010), while later work incorporated POS-based and morphological features (Forsyth, 2014; Saddiki et al., 2018). The OSMAN metric (El-Haj and Rayson, 2016) leveraged script markers and diacritization, and recent efforts (Liberato et al., 2024) achieved strong results using pretrained models on the SAMER corpus.

Building on these efforts, we curated the **BAREC** corpus across genres and readability levels, and manually annotated it at the sentence-level based on an adaptation of Taha/Arabi21 guidelines (Taha-Thomure, 2017), offering finer-grained control and a more objective assessment of textual variation.

3 BAREC Corpus Annotation

In this section, we summarize the guidelines and annotation process. For more details, see Habash et al. (2025). In the next section, we discuss corpus selection and statistics.

3.1 BAREC Guidelines

We present below a summarized version of the **BAREC** annotation guidelines. A detailed account of the adaptation process from Taha-Thomure (2017)'s guidelines is in Habash et al. (2025).

Readability Levels The readability level system of Taha-Thomure (2017) uses the Abjad order of Arabic letters for 19 levels: **1-alif**, **2-ba**, **3-jim**, through to **19-qaf**. This system emphasizes a finer distinction in the lower levels, where readability is more varied. The **BAREC** pyramid (Figure 1) illustrates the scaffolding of these levels and their mapping to, guidelines components, school grades, and three collapsed versions of level size 7, 5, and 3. All four level types (19-7-5-3) are fully aligned to easy mapping from fine-grained to coarse-grained levels. We present results for these levels in Section 6.

Readability Annotation Principles The guidelines focus on readability and comprehension, considering the ease of reading and understanding for independent readers. The evaluation does not depend on grammatical analysis or rhetorical depth but rather on understanding basic, literal meanings. Larger texts may contain sentences at different readability levels, but we focus on sentence-level evaluation, ignoring context and author intent.

Textual Features Levels are assessed in six key dimensions. Each of these specify numerous linguistic phenomena that are needed to qualify for being ranked in a harder level. Annotators assign each sentence a readability level based on its most difficult linguistic phenomenon. The Cheat Sheet used by the annotators in Arabic and its translation in English are included in Appendix A.

1. **Spelling**: Word length and syllable count affect difficulty.
2. **Word Count**: The number of unique words determines the highest level for easier levels.
3. **Morphology**: We distinguish between simple and complex morphological forms including the use of clitics and infrequent inflectional features, such as the dual.

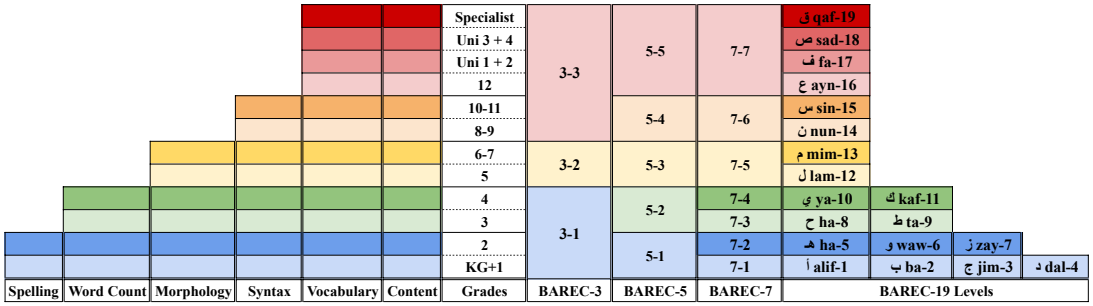

Figure 1: The **BAREC** *Pyramid* illustrates the relationship across **BAREC** levels and linguistic dimensions, three collapsed variants, and education grades.

RL	Arabic Sentence/Phrase	Translation
1-alif	**أَرْنَبٌ**	**Rabbit**
3-jim	أنا أحب **اللون** الأحمر.	I love **the** color red.
5-ha	القطة تستريح على السرير **وتستمتع بأشعة الشمس الدافئة**.	The cat rests on the bed **and enjoys the warm sunshine**.
7-zay	**الأصدقاء** يحتفلون بعيد ميلاد صديقهم بكعكة وهدايا رائعة.	**Friends** celebrate their friend's birthday with cake and amazing gifts.
10-ya	وَسَأَلْتُكَ هَلْ **كُنْتُمْ** تَتَّهِمُونَهُ بِالْكَذِبِ قَبْلَ أَنْ يَقُولَ مَا قَالَ فَذَكَرْتَ أَنْ لاَ،	I asked you whether **you were** accusing him of lying before he said what he said, and you said no.
12-lam	لا أحد يجمع هذه الزهور معًا في باقة، فهي منتشرة جدًّا — **حتى إنه كان من المعروف عنها أنها تنمو بين أحجار الرصف، وتنبثق في كل مكان مثل الحشائش الضارة** — وتحمل اسمًا قبيحًا جدًّا وهو «زهور الكلاب» أو «الهندباء البرية».	No one puts these flowers together in a bouquet, they are so common—**they have even been known to grow between paving stones, and spring up everywhere like weeds**—and they have the very unsightly name of "dog-flowers" or "dandelions."
14-nun	حيث إن هذه الزيادة في **الجسيمات المشحونة** تشير إلى خروج المركبة من نطاق تأثير **الرياح الشمسية** الذي يسمى **الغلاف الشمسي** (والذي يعتبر حسب بعض التعاريف حدود **المجموعة الشمسية**).	This increase in **charged particles** indicates the spacecraft's departure from the influence of the **solar wind**, which is called **the heliosphere** (which, according to some definitions, is the border of the **solar system**).
16-ayn	ويذهب المؤرخون إلى أن **النابغة الذبياني** كان من **المُحَكَّمين**، تقام له في هذه الأسواق قبة يذهب إليها الشعراء ليعرضوا شعرهم، فمن أشاد به **ذاع صيته**، وتناقلت شعره **الركبان**.	Historians assert that **Al-Nabigha Al-Dhubyani** was one of the **arbiters**. In these markets, a dome is erected for him where poets go to present their poetry. Whomever he praised, **his fame spread**, and his poetry circulated among the **caravans**.
19-qaf	كأن **حدوج المالكية غدوةً خلايا سفين بالنواصف من دد**	As if **the camel saddles of the Malikiyya caravan leaving the Dadi valley were great ships**

Table 1: Representative subset of examples of the 19 **BAREC** readability levels, with English translations, and readability level reasoning. Underlining is used to highlight the main keys that determined the level.

4. **Syntax**: Specific sentence structure and syntactic relation constructions are identified as pivotal for certain levels.
5. **Vocabulary**: The complexity of word choices is key, with higher levels introducing more technical and classical literature vocabulary.
6. **Content**: The required prior knowledge and abstraction levels are considered for higher levels.

The **BAREC** pyramid (Figure 1) illustrates which aspects are used (broadly) for which levels. For example, spelling criteria are only used up to level **7-zay**, while syntax is used until level **15-sin**, and word count is not used beyond level **11-kaf**.

Problems and Difficulties Annotators are encouraged to report any issues like spelling errors, colloquial language, or problematic topics. Difficulty is noted when annotations cannot be made due to conflicting guidelines.

A few representative examples for each level are provided in Table 1. A full set of examples with explanations of leveling choices is in Appendix A.3.

3.2 Annotation Team and Process

Annotation Team The **BAREC** annotation team comprised six native Arabic speakers, all of whom are experienced Arabic language educators. Among the team members, one individual (A0) brought prior experience in computational linguistic annotation projects, while the remaining five (A1-5) possessed extensive expertise in readability leveling, gained through their involvement in the Taha/Arabi21 project.

Annotation Process The annotation process began with A0, who led sentence-level segmentation and initial text flagging and selection. We followed the Arabic sentence segmentation guidelines by Habash et al. (2022a). Subsequently, A1-5 were tasked with assigning readability labels to the indi-

Domain	Readership Group	#Documents		#Sentences		#Words	
Arts & Humanities	**Foundational**	562	29%	24,978	36%	274,497	26%
Arts & Humanities	**Advanced**	478	25%	15,285	22%	222,933	21%
Arts & Humanities	**Specialized**	327	17%	10,179	15%	155,565	15%
STEM	**Foundational**	27	1%	533	1%	12,879	1%
STEM	**Advanced**	85	4%	1,948	3%	48,501	5%
STEM	**Specialized**	68	4%	2,199	3%	49,265	5%
Social Sciences	**Foundational**	44	2%	2,270	3%	26,692	3%
Social Sciences	**Advanced**	168	9%	5,463	8%	110,226	11%
Social Sciences	**Specialized**	163	8%	6,586	9%	138,813	13%
Arts & Humanities		1,367	71%	50,442	73%	652,995	63%
STEM		180	9%	4,680	7%	110,645	11%
Social Sciences		375	20%	14,319	21%	275,731	27%
	Foundational	633	33%	27,781	40%	314,068	30%
	Advanced	731	38%	22,696	33%	381,660	37%
	Specialized	558	29%	18,964	27%	343,643	33%
		1,922	100%	69,441	100%	1,039,371	100%

Table 2: Summary statistics of the **BAREC** Corpus.

vidually segmented texts. The annotation was done through a simple Google Sheet interface. A1-5 received folders containing annotation sets, comprising 100 randomly selected sentences each. The average annotation speed was around 2.5 hours per batch (1.5 minutes/sentence).

Before starting the annotation, all annotators received rigorous training, including three pilot rounds. These rounds provided opportunities for detailed discussions of the guidelines, helping to identify and address any issues. 19 shared annotation sets (100 sentence each) were included covertly to ensure quality and measure inter-annotator agreement (IAA). Finally, we conducted a thorough second review of the corpus data, resulting in every sentence being checked twice for the first phase (10,658 sentences) before continuing to finish the 69,441 sentences (1M words).

In total, the annotators annotated 92.6K sentences, 25% of which is not in the final corpus: 3.3% were deemed problematic (typos and offensive topics); 11.5% were part of the second round of first phase annotation; and 10.3% were part of the IAA efforts, not including their unification. We report on IAA in Section 6.1.

4 BAREC Corpus

Corpus Selection In the process of corpus selection, we aimed to cover a wide educational span as well as different domains and topics. We collected the corpus from 1,922 documents, which we manually categorized into three domains: **Arts & Humanities**, **Social Sciences**, and **STEM** (details in Appendix C.2) and three readership groups: **Foundational**, **Advanced**, and **Specialized** (details in Appendix C.3). Table 2 shows the distribution of the documents, sentences, and words across domains and groups. The distribution across readership levels aligns with the corpus's educational focus, with a higher-than-usual proportion at foundational levels. Variations across domains reflect differences in the availability of texts and reader interest (more Arts & Humanities, less STEM). The corpus uses documents from 30 different resources. All selected texts are either out of copyright, within the fair-use limit, or obtained in agreement with publishers. The decision of selecting some of these resources is influenced by the fact that other annotations exist for them. Around 25% of all sentences came from completely new sources that were manually typed to make them digitally usable. All details about the resources are available in Appendix C.

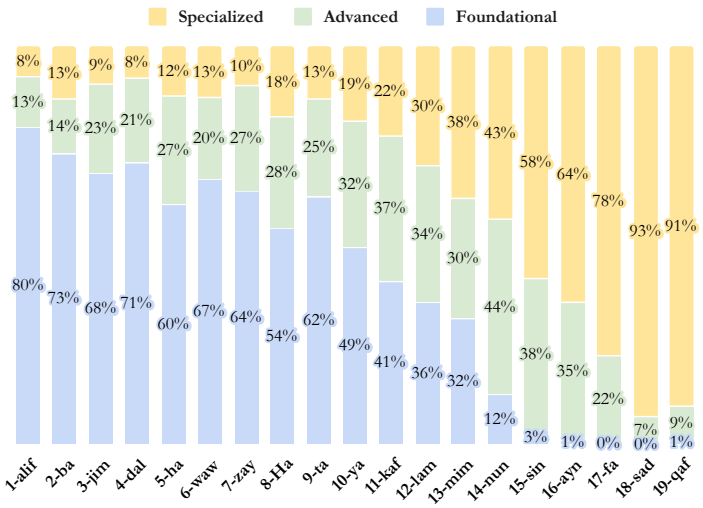

Figure 2: The distribution of the readership groups across **BAREC** levels.

	#Documents		#Sentences		#Words	
Train	1,518	79%	54,845	79%	832,743	80%
Dev	194	10%	7,310	11%	101,364	10%
Test	210	11%	7,286	10%	105,264	10%
All	**1,922**	**100%**	**69,441**	**100%**	**1,039,371**	**100%**

Table 3: **BAREC** Corpus splits.

Readability Statistics Figure 2 shows the distribution of the three readership groups across all readability levels. As expected, foundational texts strictly dominate the lower levels up to **9-ta**, then the presence of advanced and specialized texts starts increasing gradually till the highest level. Specialized texts dominate the highest levels, while the middle levels (**10-ya** to **14-nun**) include a mix of the three groups with a slight advantage for advanced texts.

Corpus Splits We split the corpus into **Train** (≃**80%**), **Dev** (≃**10%**), and **Test** (≃**10%**) at the document level. Sentences from IAA studies are divided between all splits. However, We will release the IAA studies as a special set as they provide multiple references from different annotators for each example.[2] Also, if other annotations exist for a resource (e.g., CamelTB (Habash et al., 2022b) and ReadMe++ (Naous et al., 2024)), we follow the existing splits. Table 3 shows the corpus splits in the level of documents, sentences, and words. More details about the splits across readability levels, domains, and readership groups are available in Appendix B.

5 Experiments

5.1 Metrics

In this paper, we define the task of Readability Assesment as an ordinal classification task. We use the following metrics for evaluation.

Accuracy (Acc^{19}) The percentage of cases where reference and prediction classes match in the 19-level scheme. We addition consider three variants, **Acc**^{7}, **Acc**^{5}, **Acc**^{3}, that respectively collapse the 19-levels into the 7, 5, and 3-level schemes discussed in Section 3.

Adjacent Accuracy ($\pm$1 Acc^{19}) Also known as off-by-1 accuracy. It allows some tolerance for predictions that are close to the true labels. It measures the proportion of predictions that are either exactly correct or off by at most one level.

Average Distance (Dist) Also known as Mean Absolute Error (MAE), it measures the average absolute difference between predicted and true labels.

Quadratic Weighted Kappa (QWK) An extension of Cohen's Kappa (Cohen, 1968; Doewes et al., 2023) measuring the agreement between predicted and true labels, but applies a quadratic penalty to larger misclassifications, meaning that predictions farther from the true label are penalized more heavily.

We consider Quadratic Weighted Kappa as the primary metrics for selecting the best system.

Input Variant	Example
Original	فَإِلَى شَرْقِ الْبَحْرِ يَقْبَعُ مَمَرُّ الذَّهَبِ وَالْفِضَّةِ.
Word	فإلى شرق البحر يقبع ممر الذهب والفضة .
Lex	إلى شرق بحر قبع ممر ذهب فضة .
D3Tok	ف+ إلى شرق ال+ بحر يقبع ممر ال+ ذهب و+ ال+ فضة .
D3Lex	ف+ إلى شرق ال+ بحر قبع ممر ال+ ذهب و+ ال+ فضة .

Table 4: Example sentence in different input variants.

5.2 Input Variants

In morphologically rich languages, affixation, compounding, and inflection convey key linguistic information that influences readability. Human annotators consider morphological complexity when assessing readability, but standard tokenization may obscure these cues. Segmenting sentences into morphological units helps preserve structural patterns relevant to readability prediction.

We generate four input variants using CamelTools morphological disambiguation to identify top choice analysis in context (Obeid et al., 2020).[4] For the **Word** variant, we simply tokenize the sentences and remove diacritics and kashida using CAMeL Tools (Obeid et al., 2020). For **Lex**, we replace each word with its predicted Lemma. For **D3Tok**, we tokenize the word into its base and clitics form; and for **D3Lex**, we replace the base form in D3Tok with the lemma. All variants are dediacritized. Table 4 shows an example of a sentence and the corresponding input variants.

5.3 Fine-Tuning

We fine-tuned the top three Arabic BERT-based models according to Inoue et al. (2021) (AraBERTv02 (Antoun et al., 2020), MARBERTv2 (Abdul-Mageed et al., 2021), CamelBERT-msa (Inoue et al., 2021)). We also added AraBERTv2 to our experiments due to the possible matching between its pre-training data (morphologically segmented sentences by Farasa (Darwish and Mubarak, 2016)) and the different input variants.

5.4 Loss Functions

Since readability levels exhibit a natural ordering, we explore loss functions that account for the distance between predicted and true labels (Heilman et al., 2008). In addition to standard cross-entropy loss (**CE**), we experiment with Ordinal Log Loss (**OLL**) (Castagnos et al., 2022), Soft Labels Loss (**SOFT**) (Bertinetto et al., 2020), Earth Mover's Distance-based loss (**EMD**) (L. Hou, 2017), and Regression using Mean Squared Error (**Reg**) as these have been previously used for ordinal classification tasks. OLL, SOFT, and EMD incorporate a distance matrix D into their formulations to penalize predictions proportionally to their distance from the true label. For simplicity, we define the distance between any two adjacent levels as one, setting $D(i, j) = |i - j|$ for labels i and j. For regression, we round the final output to the nearest readability level to ensure predictions align with the 19 levels.

5.5 Hyper-parameters

For all experiments, we use a learning rate of 5×10^{-5}, a batch size of 64, and train for six epochs on an NVIDIA V100 GPU. After training, we select the best-performing epoch based on evaluation loss. For Ordinal Log Loss (OLL), we experiment with different values of the weighting parameter α, choosing from {0.5, 1, 1.5, 2}. Similarly, for Soft Labels Loss (SOFT), we evaluate different values of the smoothing parameter β, selecting from {2, 3, 4, 5}. The training of the models in this paper took approximately 20 hours.

5.6 Procedure

Our experiments involve three main variables: the pretrained model, the input variant, and the loss function. Our goal is to determine the optimal combination of these three factors. Due to the large number of experiments required, we divide the process into two stages. In **Stage 1**, we train all combinations of pretrained models and input variants using cross-entropy loss. We then select the best combination based on a majority vote from our primary evaluation metrics (Acc, Acc $\pm$1, Dist, and QWK). In **Stage 2**, we take the best combination of pretrained model and input variant from the first stage and train models using all the different loss functions.

6 Results

6.1 Inter-Annotator Agreement (IAA)

In this section, we report on 16 IAA studies, excluding the three pilots and first two IAAs, which overlapped with annotator training.

Pairwise Agreement The average pairwise exact-match over 19 **BAREC** levels between any two annotators is only 61.1%, which reflects the

[4] CamelTools v1.5.5: Bert-Disambig+calima-msa-s31 db.

task's complexity. Allowing a fuzzy match distance of up to one level raises the match to 74.4%. The overall average pairwise level difference is 0.94 levels. The average pairwise Quadratic Weighted Kappa 81.8% (substantial agreement) confirms most disagreements are minor (Cohen, 1968; Doewes et al., 2023).

Unification Agreement After each IAA study, the annotators discussed and agreed on a unified readability level for each sentence. On average, the exact match between the annotators and the unified level (Acc^{19}) was 71.7%, reflecting the difficulty of the task. However, the high average ± 1 Acc^{19} (82.3%), low Distance (0.65), and strong Quadratic Weighted Kappa (88.1%) suggest that most disagreements between annotators and the unified labels were minor. For more detailed results on IAA, see (Habash et al., 2025).

6.2 Stage 1 Results

Table 5 presents the results of stage 1, where we evaluate different combinations of pretrained models and input variants using cross-entropy loss. Based on the all metrics, we observe that the AraBERTv02 and AraBERTv2 models generally achieve higher performance across multiple input variants.

Among input variants, the Word and D3Tok representations tend to yield better results compared to Lex and D3Lex. Specifically, AraBERTv2 with the D3Tok input achieves the best scores in all metrics. Notably, AraBERTv2 is the only model that benefits from the D3Tok and D3Lex inputs compared to the Word input, showing an improvement across all metrics. We argue that this occurs because AraBERTv2 is the only model in this set that was pretrained on segmented data, making it more compatible with morphologically segmented input. These results suggest that both the choice of input variant and the pretrained model significantly impact performance.

Based on all metrics, we select AraBERTv2 with the D3Tok input as the best-performing combination. In stage 2, we evaluate it with different loss functions. The confusion matrix for this model is available in the Appendix D.1.

6.3 Stage 2 Results

Table 6 presents the results of stage 2, where we use the best model from stage 1 to evaluate different loss functions. Among all the loss functions evaluated, Cross-Entropy (CE) achieves the highest exact accuracy (Acc^{19}) at 56.6%, indicating that it performs best when predicting the exact readability level. In contrast, other loss functions show stronger performance on metrics that consider the ordinal nature of readability levels. Notably, Regression achieves the highest ± 1 accuracy at 73.1% and the best Quadratic Weighted Kappa (QWK) at 84.0%, suggesting it excels at predicting levels close to the gold label, despite being the worst in terms of exact accuracy. These findings support that loss functions designed for ordinal or continuous labels—such as EMD, OLL, and Regression—are more effective on evaluation metrics that reward proximity to the correct label, even if they underperform on strict accuracy. More results for other loss functions are in Appendix D.2.

6.4 Ensemble Results

Table 7 presents results from Stage 1, where AraBERTv2 is evaluated with four different input variants, and Stage 2, where it is trained using the two best-performing loss functions. It also includes results from two ensemble strategies applied across all six models to assess whether combining predictions can further improve performance. We also include an oracle combination, which represents an upper bound on performance. This allows us to estimate the maximum potential gain achievable through ensembling.

Ensemble To further improve performance, we experiment with ensemble methods. We define the **Average ensemble**, where the final prediction is the rounded average of the levels predicted by the six models, and the **Most Common ensemble**, where the final prediction is the predicted levels' mode.

The results show that the Average ensemble performs better in terms of Distance, indicating that it tends to stay closer to the correct label. However, it struggles with exact accuracy (Acc), as averaging can blur distinctions between classes. On the other hand, the Most Common ensemble achieves higher Acc but can sometimes be misled by an incorrect majority, leading to greater deviation from the correct label.

Oracle We also report an **Oracle Combination**, where we assume access to the best possible prediction from the six models for each sample. This serves as an upper bound on model performance. The Oracle results are significantly higher than those of individual models and are comparable to

Input	Model	Acc^{19}	±1 Acc^{19}	Dist	QWK
Word	CamelBERT-msa	54.4%	68.7%	1.20	79.1%
	MARBERTv2	53.3%	68.0%	1.20	79.1%
	AraBERTv02	55.8%	69.2%	1.17	79.2%
	AraBERTv2	51.6%	65.9%	1.32	76.3%
Lex	CamelBERT-msa	48.3%	64.4%	1.34	77.1%
	MARBERTv2	50.1%	64.9%	1.31	77.0%
	AraBERTv02	48.8%	65.4%	1.30	78.5%
	AraBERTv2	50.1%	65.4%	1.29	77.7%
D3Tok	CamelBERT-msa	54.8%	68.2%	1.21	78.2%
	MARBERTv2	54.0%	68.5%	1.20	78.9%
	AraBERTv02	54.8%	68.1%	1.22	78.2%
	AraBERTv2	**56.6%**	**69.9%**	**1.14**	**80.0%**
D3Lex	CamelBERT-msa	51.1%	65.5%	1.29	78.0%
	MARBERTv2	51.6%	65.7%	1.28	78.0%
	AraBERTv02	53.3%	68.1%	1.24	78.2%
	AraBERTv2	53.2%	67.1%	1.24	78.6%

Table 5: Results comparing different combinations of models and input variants on **BAREC** Dev set. **Bold** are the best results on each metric.

Loss	Acc^{19}	±1 Acc^{19}	Dist	QWK
CE	**56.6%**	69.9%	1.14	80.0%
EMD	55.3%	70.3%	**1.11**	81.2%
OLL2	35.2%	70.3%	1.25	82.0%
OLL15	47.3%	71.1%	1.13	82.8%
OLL1	50.8%	71.5%	1.12	81.7%
OLL05	53.1%	68.8%	1.18	79.7%
SOFT2	55.8%	69.8%	1.15	80.0%
SOFT3	56.4%	69.9%	1.14	80.1%
SOFT4	56.4%	69.9%	1.15	79.6%
SOFT5	56.2%	69.5%	1.17	79.3%
Reg	43.1%	**73.1%**	1.13	**84.0%**

Table 6: Loss functions comparisons on **BAREC** Dev set. We use AraBERTv2 model and D3Tok input with all loss function. **Bold** are the best results on each metric.

human annotators' agreement with the unified labels (see section 6.1). This suggests that while individual models are still far from human-level performance, ensembling has the potential to push results closer to human agreement. More oracle combinations are provided in Appendix D.4. We also include more results on the impact of training granularity on readability level prediction in Appendix D.3

Finally, table 8 shows the results on the test set. We note that the trends observed in the development set persist in the test set, further validating our findings.

6.5 Error Analysis

To assess the errors in our best-performing model, we analyzed error patterns in the inter-annotator portion of the development (DEV) set. Each sentence in this subset had five human annotations, which we compared to the model's prediction.

We grouped sentences by the level of annotator agreement, from full agreement (5 out of 5 annotators) down to minimal agreement (1 out of 5). Full 5-way agreement accounts for 25% of the data. With each reduction in agreement – to 4, 3, 2, and finally 1 annotator – the cumulative coverage increases to 50%, 61%, 72%, and 87%, respectively. In other words, in 87% of the cases, the model prediction can be meaningfully compared to at least some level of human consensus.

The remaining 13% fall outside this range. In 1% of these, the model's prediction was within the span of human annotations but did not exactly match any of them. In 3%, the prediction was above the maximum annotation, and in 9%, it was below the minimum. We manually reviewed these out-of-range cases and found that the annotators were generally correct. We speculate that the model's errors arise from limited training data, lack of con-

Input	Loss	Acc^{19}	±1 Acc^{19}	Dist	QWK	Acc^{7}	Acc^{5}	Acc^{3}
Word	**CE**	51.6%	65.9%	1.32	76.3%	61.6%	67.2%	74.0%
Lex	**CE**	50.1%	65.4%	1.29	77.7%	60.6%	66.3%	74.9%
D3Tok	**CE**	**56.6%**	69.9%	1.14	80.0%	65.9%	70.3%	76.5%
D3Lex	**CE**	53.2%	67.1%	1.24	78.6%	63.6%	69.0%	75.3%
D3Tok	**EMD**	55.3%	70.3%	**1.11**	81.2%	65.2%	70.0%	76.4%
D3Tok	**Reg**	43.1%	**73.1%**	1.13	**84.0%**	61.1%	67.8%	75.9%
Average		46.9%	72.5%	**1.11**	83.4%	64.0%	70.3%	**77.2%**
Most Common		56.3%	70.0%	1.13	80.4%	**66.3%**	**70.9%**	76.9%
Oracle Combo		75.2%	87.4%	0.50	93.8%	83.2%	85.7%	89.1%

Table 7: Results comparing different loss function, ensemble methods, and oracle performance on **BAREC** Dev set. **Bold** are the best results across individual models and across ensembles.

Input	Loss	Acc^{19}	±1 Acc^{19}	Dist	QWK	Acc^{7}	Acc^{5}	Acc^{3}
Word	**CE**	51.1%	65.1%	1.31	76.2%	60.7%	65.6%	72.2%
Lex	**CE**	51.2%	66.2%	1.23	78.5%	61.1%	66.2%	74.4%
D3Tok	**CE**	55.9%	70.0%	1.12	80.2%	65.1%	69.4%	75.2%
D3Lex	**CE**	53.7%	67.9%	1.17	79.5%	63.8%	69.1%	74.8%
D3Tok	**EMD**	54.9%	71.4%	**1.02**	83.7%	64.9%	69.0%	75.2%
D3Tok	**Reg**	41.4%	**73.5%**	1.11	84.4%	59.4%	65.3%	72.8%
Average		46.0%	73.4%	1.06	**84.5%**	63.6%	69.4%	**75.8%**
Most Common		**56.2%**	70.4%	1.07	81.3%	**65.9%**	**70.0%**	75.6%
Oracle Combo		75.9%	87.8%	0.46	94.7%	83.5%	85.7%	88.9%

Table 8: Results comparing different loss function, ensemble methods, and oracle performance on **BAREC** Test set. **Bold** are the best results across individual models and across ensembles.

textual understanding, or insufficient modeling of linguistic features. For example, the obscure word عصامة ʿSAmħ 'tightly wound head dress' may be misinterpreted as the feminine form of the proper name عصام ʿSAm 'Esam', much like connecting كريم krym 'Kareem' with كريمة krymħ 'Kareema'. However, عصامة ʿSAmħ is not a plausible proper name. This remains speculative, as our model is not inherently interpretable.

7 Conclusions and Future Work

This paper presented the **Balanced Arabic Readability Evaluation Corpus (BAREC)**, a large-scale, finely annotated dataset for assessing Arabic text readability across 19 levels. With over 69K sentences and 1 million words, it is the largest Arabic corpus for readability assessment, covering diverse genres, topics, and audiences, to our knowledge. High inter-annotator agreement ensures reliable annotations. Through benchmarking various readability assessment techniques, we highlighted both the challenges and opportunities in Arabic readability modeling, demonstrating promising performance across different methods.

Looking ahead, we plan to expand the corpus, enhancing its size and diversity to cover additional genres and topics. We also aim to add annotations related to vocabulary leveling and syntactic treebanks to study less-explored genres in syntax. Future work will include analyzing readability differences across genres and topics. Additionally, the tools we have developed will be integrated into a system to help children's story writers target specific reading levels.

The **BAREC** dataset, its annotation guidelines, and benchmark results, will be made publicly available to support future research and educational applications in Arabic readability assessment.

Acknowledgments

The **BAREC** project is supported by the Abu Dhabi Arabic Language Centre (ALC) / Department of Culture and Tourism, UAE. We acknowledge the support of the High Performance Computing Center at New York University Abu Dhabi. We are deeply grateful to our outstanding annotation team: Mirvat Dawi, Reem Faraj, Rita Raad, Sawsan Tannir, and Adel Wizani, Samar Zeino, and Zeina Zeino. Special thanks go to Abdallah Abushmaes, Karin Aghadjanian, and Omar Al Ayyoubi of the ALC for their continued support. We would also like to thank the Zayed University ZAI Arabic Language Research Center team, in particular Hamda Al-Hadhrami, Maha Fatha, and Metha Talhak, for their valuable contributions to typing materials for the project. We also acknowledge Ali Gomaa and his team for their additional support in this area. Finally, we thank our colleagues at the New York University Abu Dhabi Computational Approaches to Modeling Language (CAMeL) Lab, Muhammed Abu Odeh, Bashar Alhafni, Ossama Obeid, and Mostafa Saeed, as well as Nour Rabih (Mohamed bin Zayed University of Artificial Intelligence) for their helpful conversations and feedback.

Limitations

One notable limitation is the inherent subjectivity associated with readability assessment, which may introduce variability in annotation decisions despite our best efforts to maintain consistency. Additionally, the current version of the corpus may not fully capture the diverse linguistic landscape of the Arab world. Finally, while our methodology strives for inclusivity, there may be biases or gaps in the corpus due to factors such as selection bias in the source materials or limitations in the annotation process. We acknowledge that readability measures can be used with malicious intent to profile people; this is not our intention, and we discourage it.

Ethics Statement

All data used in the corpus curation process are sourced responsibly and legally. 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 corpus and associated guidelines are made openly accessible to promote transparency, reproducibility, and collaboration in Arabic language research.

References

Ahmed Abdelali, Kareem Darwish, Nadir Durrani, and Hamdy Mubarak. 2016. Farasa: A fast and furious segmenter for Arabic. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 11–16, San Diego, California.

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.

Abbas Mahmoud Al-Akkad. 1938. *Sarah*. Hindawi.

Imam Muhammad al Bukhari. 846. *Sahih al-Bukhari*. Dar Ibn Khathir.

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.

Bayan Al-Safadi. 2005. *Al-Kashkoul: selection of poetry and prose for children* (الكشكول: مختارات من الشعر والنثر للأطفال). Al-Sa'ih Library (مكتبة السائح).

A. Alfaifi. 2015. *Building the Arabic Learner Corpus and a System for Arabic Error Annotation*. Ph.D. thesis, University of Leeds.

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.

Shatha Altammami, Eric Atwell, and Ammar Alsalka. 2019. The arabic–english parallel corpus of authentic hadith. *International Journal on Islamic Applications in Computer Science And Technology-IJASAT*.

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.

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.

Luca Bertinetto, Romain Mueller, Konstantinos Tertikas, Sina Samangooei, and Nicholas A. Lord. 2020. Making Better Mistakes: Leveraging Class Hierarchies With Deep Networks . In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12503–12512, Los Alamitos, CA, USA. IEEE Computer Society.

Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, and Kemal Oflazer. 2018. The MADAR Arabic dialect corpus and lexicon. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

François Castagnos, Martin Mihelich, and Charles Dognin. 2022. A simple log-based loss function for ordinal text classification. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4604–4609, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213.

Kevyn Collins-Thompson and James P. Callan. 2004. A language modeling approach to predicting reading difficulty. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 193–200, Boston, Massachusetts, USA. Association for Computational Linguistics.

Kareem Darwish and Hamdy Mubarak. 2016. Farasa: A new fast and accurate Arabic word segmenter. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1070–1074, Portorož, Slovenia. European Language Resources Association (ELRA).

Tovly Deutsch, Masoud Jasbi, and Stuart Shieber. 2020. Linguistic features for readability assessment. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–17, Seattle, WA, USA → Online. Association for Computational Linguistics.

Afrizal Doewes, Nughthoh Arfawi Kurdhi, and Akrati Saxena. 2023. Evaluating quadratic weighted kappa as the standard performance metric for automated essay scoring. In *Proceedings of the 16th International Conference on Educational Data Mining*, pages 103–113, Bengaluru, India. International Educational Data Mining Society.

William H DuBay. 2004. The principles of readability. *Online Submission*.

Kais Dukes, Eric Atwell, and Nizar Habash. 2013. Supervised collaboration for syntactic annotation of quranic arabic. *Language resources and evaluation*, 47(1):33–62.

Matthias Eck and Chiori Hori. 2005. Overview of the IWSLT 2005 evaluation campaign. In *Proceedings of the Second International Workshop on Spoken Language Translation*, Pittsburgh, Pennsylvania, USA.

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 Almujaiwel, 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.

Mo El-Haj and Saad Ezzini. 2024. The multilingual corpus of world's constitutions (MCWC). In *Proceedings of the 6th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT) with Shared Tasks on Arabic LLMs Hallucination and Dialect to MSA Machine Translation @ LREC-COLING 2024*, pages 57–66, Torino, Italia. ELRA and ICCL.

Lijun Feng, Martin Jansche, Matt Huenerfauth, and Noémie Elhadad. 2010. A comparison of features for automatic readability assessment. In *Coling 2010: Posters*, pages 276–284, Beijing, China. Coling 2010 Organizing Committee.

Jonathan Forsyth. 2014. Automatic readability prediction for modern standard Arabic. In *Proceedings of the Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT)*.

Irene C Fountas and Gay Su Pinnell. 2006. *Leveled books (k-8): Matching texts to readers for effective teaching*. Heinemann Educational Books.

Nizar Habash, Muhammed AbuOdeh, Dima Taji, Reem Faraj, Jamila El Gizuli, and Omar Kallas. 2022a. Camel treebank: An open multi-genre Arabic dependency treebank. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2672–2681, Marseille, France. European Language Resources Association.

Nizar Habash, Muhammed AbuOdeh, Dima Taji, Reem Faraj, Jamila El Gizuli, and Omar Kallas. 2022b. Camel treebank: An open multi-genre Arabic dependency treebank. In *Proceedings of the Thirteenth*

Language Resources and Evaluation Conference, pages 2672–2681, Marseille, France.

Nizar Habash and David Palfreyman. 2022. ZAEBUC: 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)*, Vienna, Austria. Association for Computational Linguistics.

Michael Heilman, Kevyn Collins-Thompson, and Maxine Eskenazi. 2008. An analysis of statistical models and features for reading difficulty prediction. In *Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications*, pages 71–79, Columbus, Ohio. 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.

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.

G.R. Klare. 1963. *The Measurement of Readability*. Iowa State University Press.

Fajri Koto, Haonan Li, Sara Shatnawi, Jad Doughman, Abdelrahman Sadallah, Aisha Alraeesi, Khalid Almubarak, Zaid Alyafeai, Neha Sengupta, Shady Shehata, Nizar Habash, Preslav Nakov, and Timothy Baldwin. 2024. ArabicMMLU: Assessing massive multitask language understanding in Arabic. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5622–5640, Bangkok, Thailand. Association for Computational Linguistics.

D. Samaras L. Hou, C.P. Yu. 2017. Squared earth mover's distance loss for training deep neural networks on ordered-classes. In *NIPS workshop on Learning on Distributions, Functions, Graphs and Groups*.

Bruce W. Lee, Yoo Sung Jang, and Jason Lee. 2021. Pushing on text readability assessment: A transformer meets handcrafted linguistic features. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10669–10686, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Portorož, Slovenia.

Farah Nadeem and Mari Ostendorf. 2018. Estimating linguistic complexity for science texts. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 45–55, New Orleans, Louisiana. 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.

Naoual Nassiri, Violetta Cavalli-Sforza, and Abdelhak Lakhouaja. 2023. Approaches, methods, and resources for assessing the readability of arabic texts. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 22(4).

Ossama Obeid, Nasser Zalmout, Salam Khalifa, Dima Taji, Mai Oudah, Bashar Alhafni, Go Inoue, Fadhl 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.

Emily Pitler and Ani Nenkova. 2008. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 186–195, Honolulu, Hawaii. Association for Computational Linguistics.

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.

Eli Smith and Cornelius Van Dyck. 1860. *New Testament (Arabic Translation)*.

Eli Smith and Cornelius Van Dyck. 1865. *Old Testament (Arabic Translation)*.

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. 2007. *Poems and News* (أشعار وأخبار). Educational Book House (دار الكتاب التربوي للنشر والتوزيع).

Hanada Taha-Thomure. 2017. *Arabic Language Text Leveling* (معايير هنادا طه لتصنيف مستويات النصوص العربية). Educational Book House (دار الكتاب التربوي للنشر والتوزيع).

Toshiyuki Takezawa, Genichiro Kikui, Masahide Mizushima, and Eiichiro Sumita. 2007. Multilingual spoken language corpus development for communication research. In *International Journal of Computational Linguistics & Chinese Language Processing, Volume 12, Number 3, September 2007: Special Issue on Invited Papers from ISCSLP 2006*, pages 303–324.

Ibn Tufail. 1150. *Hayy ibn Yaqdhan*. Hindawi.

Unknown. 12th century. *One Thousand and One Nights*.

Sowmya Vajjala. 2022. Trends, limitations and open challenges in automatic readability assessment research. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5366–5377, Marseille, France. European Language Resources Association.

Sowmya Vajjala and Ivana Lučić. 2018. OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 297–304, New Orleans, Louisiana. Association for Computational Linguistics.

Sowmya Vajjala and Detmar Meurers. 2012. On improving the accuracy of readability classification using insights from second language acquisition. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 163–173, Montréal, Canada. Association for Computational Linguistics.

Menglin Xia, Ekaterina Kochmar, and Ted Briscoe. 2016. Text readability assessment for second language learners. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 12–22, San Diego, CA. Association for Computational Linguistics.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297.

A BAREC Annotation Guidelines Cheat Sheet and Examples

A.1 Arabic Original

مستوى بارق	صف	ACTFL	عدد كلمات	تهجئة وإملاء	تصريف واشتقاق	تراكيب نحوية	مفردات	فكرة ومحتوى
أ	روضة-1	مبتدئ أدنى	1	• كلمات من مقطع واحد أو مقطعين	• الفعل المضارع المفرد	• كلمة واحدة	• اسم جنس • اسم علم (متداول بسيط تركيبيا) • ضمير منفصل • مفردات متطابقة مع العامية - سامر I • الأرقام (العربية أو الهندية) 1-10	• فكرة مباشرة وصريحة وحسية. • لا رمزية في النص.
ب	1	مبتدئ أدنى	≤2	• كلمات من 3 مقاطع		• جملة اسمية (هو يلعب) • إضافة حقيقية (باب البيت) • صفة وموصوف (باب كبير)	• فعل • صفة • مفردات متشابهة مع العامية - سامر I • العدد الأصلي بالأحرف • الأسماء الخمسة: أبو، أخو	
ج		مبتدئ متوسط	≤4	• كلمات من 3 مقاطع	• سوابق: ال التعريف • سوابق: واو العطف • لواحق: ضمير المتكلم المفرد المتصل	• بدل كل: (صديقي أحمد) • بدل إشارة: (هذا البيت)	• مفردات فصيحة شائعة - سامر I • اسم الإشارة المفرد • الأرقام (العربية أو الهندية) 11-100	
د		مبتدئ متوسط	≤6	• كلمات تستخدم مد الألف (أ)	• الفعل المضارع الجمع • سوابق: حروف جر متصلة • ظرف منون	• جملة فعلية بدون مفعول به • جار ومجرور	• حروف الجر	
ه	2	مبتدئ أعلى	≤8	• كلمات من 4 مقاطع	• لواحق: ضمير متصل مفرد أو جمع • المثنى (في الأسماء والصفات) • جمع المؤنث السالم	• جملة فعلية مع مفعول به واحد اسم • جمل معطوفة • أدوات استفهام أساسية: ماذا، متى، من، أين، ما، كيف • صيغة التعجب "ما أفعل"	• العدد الترتيبي • الأرقام (العربية أو الهندية) 101-1,000 • اسم اشارة مثنى، جمع	• المحتوى من حياة القارئ. • لا رمزية في النص.
و		مبتدئ أعلى	≤9	• كلمات من 5 مقاطع	• الفعل الماضي المفرد والجمع • جمع مذكر سالم	• جملة فيها فعلين (مثلا جملة فعلية مفعولها أن المصدرية)	• مفردات فصيحة - سامر I	
ز	3	متوسط أدنى	≤10	• كلمات من 6+ مقاطع • أفعال/أسماء معتلة الآخر	• الفعل الماضي المثنى •الفعل المضارع المثنى • فعل الأمر المفرد • لواحق: ضمير المثنى المتصل • جمع التكسير • واو القسم (والله)	• مفعول فيه (ظروف زمان ومكان) • حال • أداة الاستفهام هل	• مفردات فصيحة شائعة - سامر II	• بعض الرمزية أو عدم التصريح المباشر بكل المقصود في الجملة
ح		متوسط أدنى	≤11		• فعل الأمر الجمع • نون النسوة في الأسماء والأفعال • سوابق أخرى: سين الاستقبال، واو الاستئناف، فاء العطف • أدوات ربط (ثم، حتى، أو، أم، لكن، أمّا)	• المفعول المطلق • المفعول لأجله • المفعول معه • جملة فعلية تتعدى إلى مفعولين	• مفردات فصيحة - سامر I و سامر II • أحرف النفي • الأرقام (العربية أو الهندية) 1,001-1,000,000	• بعض الرمزية يحتاج معها القارئ إلى مساعدة من يشرح له المقصود من الفكرة
ط		متوسط أوسط	≤12		• فعل الأمر للمثنى • أداة الاستفهام: أ (أسمعتَ؟) • باء القسم • القسم: أداة القسم والمقسم به وجواب القسم.	• المنادى	• مفردات تصف حالات مزاجية وشعورية إيجابية وسلبية مثل الفرح، السعادة، الغضب، الأسف، الحسرة	• هناك شيء من الرمزية على مستوى الحدث في الجملة يدركها القارئ بنفسه أو من خلال معارفه السابقة
ي	4	متوسط أوسط	≤15		• المبني للمجهول	• إن وأخواتها • كان وأخواتها • خبر مقدم / مبتدأ مؤخر • العنعنة/السند • رُبَّ (حرف جر شبيه بالزائد) • جملة الصلة وجملة الصفة • جملة الحال وجملة المفعول به	• أسماء الوصل المفردة • (قد – لقد) • (مِمّا – عمّا – ممّ – علامَ – فيمَ – إلامَ - بمَ...)	
ك		متوسط أعلى	≤20		• المشتقات العاملة (مثلا اسم الفاعل)	• جملة أسمية خبرها جملة أسمية • إضافة لفظية (طويل القامة)	• أسماء الوصل المثنى والجمع	• هناك درجة من الرمزية وحاجة للمعرفة السابقة كي يُفهم المقصود من الجملة
ل	5	متقدم أدنى			• التصغير	• جمل اعتراضية (تفسير، دعاء) • استثناء • حصر • بدل (مثلا بدل بعض أو اشتمال) • تمييز	• مفردات فصيحة - سامر III • اسم الفعل (مثلا آمين) • الأرقام (العربية أو الهندية > 1,000,000 • ذو • (بل - بلى - أجل - قط)	
م	6-7	متقدم أوسط			• نون التوكيد • تاء القسم	• الجمل شرطية (مركبة - عادية) • حرف الجزم لما	• كلمات تصف حالات نفسية عميقة مثل الاكتئاب، الضياع، الاستغفار النفسي • استخدام كلمات منحوتة غير متداولة (مثلا هجرع للخفيف الأحمق مشتقة من هرع و هجع) • الرموز (ش.م.)	• أفكار رمزية ومعنى باطن خاصة على صعيد البعد النفسي للشخصيات أو الأحداث. • تعابير ثقافية محلية قد لا يفهمها من لا يشترك في نفس الثقافة.
ن	8-9	متقدم أعلى				• التوكيد المعنوي • المدح والذم • جملة أن المصدرية في محل رفع مبتدأ • صيغة التعجب "أفعل به من"	• مفردات فصيحة - سامر IV • مفردات قانونية، علمية، دينية، سياسية،... غير متخصصة/عامة • فو - حمو	
س	10-11	متقن أدنى				• تراكيب غير متداولة فيها التباس يحتاج إلى التشكيل الإعرابي لفكه	• المفردات المتخصصة التي لا تكفي معرفة الكلمة وحدها لفهمها، وإنما يحتاج إلى معرفة الفكرة/المفهوم لفهمها • الترخيم في أسماء العلم (مثلا أفاطم؟)	• أفكار رمزية، مجردة، علمية، أو شعرية وتحتاج إلى معارف لغوية ومعرفية سابقة للبناء عليها لأجل فهمها
ع	12	متقن أوسط					• مفردات فصيحة - سامر V • مفردات متخصصة ومفردات عربية عالية غير شائعة كثيرا في الفضاء العام. • مفردات في الغالب بعيدة عن اللهجات العامية.	
ف	جامعة 1-2	متقن أعلى					• مفردات علمية وتراثية غير متداولة اليوم وغير مألوفة لغير المتخصص المبتدئ	
ص	جامعة 3-4	متفوق					• مفردات علمية وتراثية غير متداولة اليوم وغير مألوفة لغير المتخصص	
ق	متخصص	متميز					• مفردات علمية وتراثية غير متداولة اليوم وغير مألوفة لغير المتخصص الباحث	
هناك صعوبة	هذا الوسم يستخدم في حالة وجود صعوبة في تقييم المستوى، المفضل استخدام هذا الوسم حتى نتمكن كفريق عمل أن نجد حلا (مثلا بتعديل المعايير أو إضافة تفاصيل شرحية لها)							
هناك مشكلة	بصورة عامة، نستخدم هذا الوسم للجمل الحاوية على:			• أخطاء إملائية (مثلا همزات، تاء مربوطة، ألف مقصورة/ياء) • أخطاء في التشكيل • ركاكة لغوية (أمية، عامية، ترجمة سيئة من لغة أجنبية) • مواضيع غير لائقة (عنصرية، حيازية، تنمرية، إباحية، إلخ) • جمل وعبارات معظمها مكتوب بلغات غير العربية أو بغير الخط العربي	ولكن في الحالات التالية نوسم الجمل ونضيف أحد الحروف التالية في عامود الملاحظات: • خطأ في همزة الوصل/همزة القطع >> (أ) • كلمات خادشة >> (ع) • الخطأ في التشكيل في بداية الجملة >> (ت) • الياء غير المنقوطة في آخر الكلمة >> (ي)		

A.2 English Translation

BAREC Level	Grade	ACTFL	Word Count	Spelling/Pronunciation	Morphology	Syntax	Vocabulary	Idea/Content
1-alif	Pre1-1	Novice Low	1	• One-syllable and two-syllable words	• Singular imperfective verb	• One word	• Common noun • Proper noun (frequent and simple) • Personal pronouns (non-clitics) • Vocabulary identical to dialectal form - SAMER I • Numbers (Arabic or Indo-Arabic) 1-10	• Direct, explicit, and concrete idea. • No symbolism in the text.
2-ba	1	Novice Low	≤2	• Three-syllable words			• Verb • Adjective • Vocabulary similar to dialectal form - SAMER I • Spelled cardinal numbers • The five nouns: *Abw (father)*, *Axw (brother)*	
3-jim		Novice Mid	≤4		• Prtoclitic: Definite article *Al+* • Proclitic: Conjunction *wa+* • Enclitic: First Person Singular pronoun	• Apposition (full) • Demonstratives	• Common MSA vocabulary - SAMER I • Singular demonstrative pronoun • Numbers: 11-100	
4-dal		Novice Mid	≤6	• Words with an elongated Alif (e.g. /ʔāsif/)	• Plural imperfective verb • Prepositional proclitics • Nunated adverbials	• Verbal sentence w/o direct object • Preposition and object	• Prepositions	
5-ha	2	Novice High	≤8	• Four-syllable words	• Enclitic: Singular and Plural pronouns • Dual (in nouns and adjectives) • Sound feminine plural	• Verbal sentence with one nominal direct object • Conjoined sentences • Basic interrogative particles: what, when, who, where, how • Exclamatory form: how <comparative adjective>	• Ordinal numbers • Numbers: 101-1,000 • Dual and plural demonstrative pronoun	• Content is from the reader's life. • No symbolism in the text.
6-waw		Novice High	≤9	• Five-syllable words	• Singular and plural perfective verb • Sound masculine plural	• Sentence with two verbs (e.g., a verbal sentence a clausal direct object introduced with *Masdar 'an [~to/that]*)	• MSA vocabulary - SAMER I	
7-zay		Intermediate Low	≤10	• Six-syllable or more words • Verbs/nouns with weak final letters	• Dual perfective verb • Dual imperfective verb • Singular imperative verb • Enclitics: dual pronoun • Broken plurals • Waw of oath	• Adverbial accusative (time and place adverbs) • Circumstantial accusative • Interrogative particle *hal*	• High frequency MSA vocabulary - SAMER II	• Some symbolism, or not everything is stated directly in the sentence.
8-ha	3	Intermediate Low	≤11		• Plural imperative verb • Feminine plural suffix (*nun*) in nouns and verbs • Other proclitics: future *sa+*, continuation *wa+*, conjunction *fa+* • Conjunctions (e.g., then, until, or, whether, but, as for)	• Absolute object (emphasizing the verb) • Object of purpose • Object of accompaniment • Verbal sentence with two direct objects	• MSA vocabulary - SAMER I and II • Negation particles • Numbers: 1,001-1,000,000	• Some symbolism that requires the reader to seek help to understand the idea.
9-ta		Intermediate Mid	≤12		• Dual imperative verb • Interrogative Hamza • Ba of oath • Oath: The particle of oath, the object of the oath, and the answer to the oat	• Vocative	• Vocabulary describing positive and negative emotional and mood states like joy, happiness, anger, regret, sorrow	• Some symbolism at the event level in the sentence that the reader understands through prior knowledge.
10-ya	4	Intermediate Mid	≤15		• Passive voice	• *Inna* and its sisters (particles introducing a subject) • *Kana* and its sisters (past tense verbs) • Preposed predicate, postponed subject • Chain of narration • *rubba* preposition construction • Relative clauses • Circumstantial and object clauses	• Singular relative pronouns • Verbal particles *qad* and *laqad* • Preposition-Conjunctions: *mimma*, *fima*...	
11-kaf		Intermediate High	≤20		• Acting derivatives (e.g., the active participle)	• Nominal sentence with a nominal predicate • False idafa (tall in stature)	• Dual and plural relative pronouns	• A degree of symbolism and a need for prior knowledge to understand the meaning of the sentence.
12-lam	5	Advanced Low			• Diminutive form	• Parentheticals (explanation, blessing) • Exception • Exclusivity • Apposition (e.g., partitive or containing) • Specification (*tamyiyz* construction)	• MSA vocabulary - Samer III • Frozen Verbs (e.g., *Āmiyn* Amen) • Numbers: > 1,000,000 • Five Nouns: Dhu (possession nominal) • Interjections: *bala*, *Ajal*, etc.	
13-mim	6-7	Advanced Mid			• Energetic mood (emphatic *nun*) • Ta of oath	• Conditional sentences • Jussive particle *lamma* (not yet)	• Words describing deep psychological states like depression, loss, psychological alertness • Use of coined, uncommon words • Abbreviations (e.g., LLC)	• Symbolic ideas and deeper meanings, especially in terms of the psychological dimension of characters/events. • Local cultural expressions that may not be understood by those outside the
14-nun	8-9	Advanced High				• Semantic emphasis • Praise and dispraise • *Masdar 'an* clause as a subject • Exclamatory form: <comparative adjective> *bih min*	• MSA vocabulary - SAMER IV • General legal, scientific, religious, political vocabulary, etc. • Five Nouns: *fw*, *Hmw*	
15-sin	10-11	Superior Low				• Uncommon constructions that are ambiguous and need diacritization for clarification	• Specialized vocabulary that requires understanding the concept/idea to comprehend it • Shortening in proper names (e.g., *fatim* for *fatima*)	• Symbolic, abstract, scientific, or poetic ideas that require prior linguistic and cognitive knowledge to understand.
16-ayn	12	Superior Mid					• MSA vocabulary - SAMER V • Specialized and highly elevated Arabic vocabulary. • Vocabulary mostly distant from dialects.	
17-fa	University Year 1-2	Superior High					• Scientific and heritage vocabulary not in use today, but familiar to a novice specialist	
18-sad	University Year 3-4	Distinguished					• Scientific and heritage vocabulary not in use today, but familiar to a specialist	
19-qaf	Specialist	Distinguished+					• Scientific and heritage vocabulary not in use today, but familiar to the advanced researcher specialist	
Difficulty	This tag is used when there is difficulty in assessing the level. It is preferred to use this tag so that the team can find a solution (for example, by adjusting the criteria or adding explanatory details).							
Problem	Generally, we use this tag for sentences containing:		• Spelling mistakes (e.g., Hamzas, Ta Marbuta, Alif maqsura/Ya) • Errors in diacritics • Linguistic awkwardness (illiteracy, colloquialism, poor translation from a foreign language) • Inappropriate topics (racism, bias, bullying, pornography, etc.) • Sentences and phrases mostly written in languages other than Arabic or in non-Arabic script	However, in the following cases, we provide the level and add a note in the comments column: • Error in Hamzat al-Wasl/Hamzat al-Qat' >> (ا) • Offensive words >> (ع) • Error in diacritics at the beginning of the sentence >> (ب) • Dotted Yaa missing at the end of the word >> (ي)		

A.3 Annotation Examples

Representative examples of the 19 **BAREC** readability levels, with English translations, and readability level reasoning. Underlining is used to highlight the main keys that determined the level.

RL	Arabic Sentence/Phrase	Translation	Reasoning
1-alif	**أَرْنَب**	**Rabbit**	*One bisyllabic familiar noun*
2-ba	**ملعبٌ واسعٌ**	**A large playground**	*Noun-adjective*
3-jim	أنا أحب **اللون** الأحمر.	I love **the** color red.	*Definite article*
4-dal	الشمس تشرق **في الصباح الباكر**.	The sun rises early **in the morning**.	*Prepositional phrase*
5-ha-	القطة تستريح على السرير **وتستمتع بأشعة الشمس الدافئة**.	The cat rests on the bed **and enjoys the warm sunshine**.	*A conjoined sentence*
6-waw	سُلوكي **مَسْؤولِيَّتي**	My behavior is **my responsibility**	*Five syllable word*
7-zay	**الأصدقاء** يحتفلون بعيد ميلاد صديقهم بكعكة وهدايا رائعة.	**Friends** celebrate their friend's birthday with cake and amazing gifts.	*Broken plural*
8-ha	أَسْتَمِعُ إِلى كُلِّ فِقْرةٍ مِنَ الفِقْرَتَيْنِ الآتِيَتَيْنِ، **ثُمَّ** أُجيبُ:	I listen to each of the following two paragraphs, **then** I answer:	*ثُمَّ (then) is in level 8-ha ح*
9-ta	وقال بكلام فصيح مزعج: **يا سمك يا سمك** هل أنت على العهد القديم مقيم	He said in annoying, eloquent words: **Oh fish, oh fish**, do you abide by the old promise	*Vocative construction*
10-ya	وَسَأَلْتُكَ هَلْ **كُنْتُمْ** تَتَّهِمُونَهُ بِالْكَذِبِ قَبْلَ أَنْ يَقُولَ مَا قَالَ فَذَكَرْتَ أَنْ لاَ،	I asked you whether **you were** accusing him of lying before he said what he said, and you said no.	*Auxiliary Kaana*
11-kaf	حسام **سعيدٌ قلبُه** بسبب فوز فريقه.	Hossam, his **heart is happy** because of his team's victory.	*Acting derivative (happy is predicative)*
12-lam	لا أحد يجمع هذه الزهور معًا في باقة، فهي منتشرة جدًّا — **حتى إنه كان من المعروف عنها أنها تنمو بين أحجار الرصف، وتنبثق في كل مكان مثل الحشائش الضارة** — وتحمل اسمًا قبيحًا جدًّا وهو «زهور الكلاب» أو «الهندباء البرية».	No one puts these flowers together in a bouquet, they are so common—**they have even been known to grow between paving stones, and spring up everywhere like weeds**—and they have the very unsightly name of "dog-flowers" or "dandelions."	*Parenthetical phrase*
13-mim	**ومن يفعل المعروف مع غير أهله يجازَ كما جوزي مجير أم عامر**	**And whoever offers good deeds to someone undeserving will be rewarded like he who gave shelter to a hyena**	*Conditional phrase*
14-nun	حيث إن هذه الزيادة في **الجسيمات المشحونة** تشير إلى خروج المركبة من نطاق تأثير **الرياح الشمسية** الذي يسمى **الغلاف الشمسي** (والذي يعتبر حسب بعض التعاريف حدود **المجموعة الشمسية**).	This increase in **charged particles** indicates the spacecraft's departure from the influence of the **solar wind**, which is called **the heliosphere** (which, according to some definitions, is the border of the **solar system**).	*General geography vocabulary*
15-sin	وكان من عادتها أن تقارن بينها وبين بطلة الرواية إذا أحسَّت منه إعجابًا بها أو ثناءً عليها، وتسأله في ذلك أسئلةً ذكيةً خبيثةً**لا تسهل المغالطة في جوابها**، إلا على سبيل المزاح والمداعبة.	It was her habit to compare herself with the heroine of the novel when she felt his admiration or praise for her, asking him smart and tricky questions **that did not allow answering deceptively**, except by joking and teasing.	*Specialized vocabulary that requires understanding the concept to comprehend its use*
16-ayn	ويذهب المؤرخون إلى أن **النابغة الذبياني** كان من **المُحَكَّمين**، تقام له في هذه الأسواق قبة يذهب إليها الشعراء ليعرضوا شعرهم، فمن أشاد به **ذاع صيته**، وتناقلت شعره **الركبان**.	Historians assert that **Al-Nabigha Al-Dhubyani** was one of the **arbiters**. In these markets, a dome is erected for him where poets go to present their poetry. Whomever he praised, **his fame spread**, and his poetry circulated among the **caravans**.	*Specialized and uncommon vocabulary*
17-fa	بين طعن **القَنا** وخَفْق **البُنود**	Between the thrusts of **lances** and the fluttering of **ensigns**	*Heritage vocabulary familiar to a novice specialist*
18-sad	**إلاّ الأوارِيَّ لأيًا ما أُبَيِّنُها والنُّؤيُ كالحَوْضِ بالمَظلومَةِ الجَلَد**	**I wasn't able to see except with extreme effort and difficulty like a water basin in solid undrillable land**	*Specialist vocabulary, symbolic poetic ideas requiring prior knowledge*
19-qaf	كأن **حدوج المالكية غدوةً خلايا سفينٍ بالنواصف من دد**	As if **the camel saddles of the Malikiyya caravan leaving the Dadi valley were great ships**	*Advanced specialist vocabulary, symbolic poetic ideas requiring prior knowledge*

B BAREC Corpus Splits

B.1 Sentence-level splits across readability levels

Level	All		Train		Dev		Test	
1-alif	409	1%	333	1%	44	1%	32	0%
2-ba	437	1%	333	1%	68	1%	36	0%
3-jim	1,462	2%	1,139	2%	182	2%	141	2%
4-dal	751	1%	587	1%	78	1%	86	1%
5-ha	3,443	5%	2,646	5%	417	6%	380	5%
6-waw	1,534	2%	1,206	2%	189	3%	139	2%
7-zay	5,438	8%	4,152	8%	701	10%	585	8%
8-Ha	5,683	8%	4,529	8%	613	8%	541	7%
9-ta	2,023	3%	1,597	3%	236	3%	190	3%
10-ya	9,763	14%	7,741	14%	1,012	14%	1,010	14%
11-kaf	4,914	7%	4,041	7%	409	6%	464	6%
12-lam	14,471	21%	11,318	21%	1,491	20%	1,662	23%
13-mim	4,039	6%	3,252	6%	349	5%	438	6%
14-nun	10,687	15%	8,573	16%	1,072	15%	1,042	14%
15-sin	2,547	4%	2,016	4%	258	4%	273	4%
16-ayn	1,141	2%	866	2%	114	2%	161	2%
17-fa	480	1%	364	1%	49	1%	67	1%
18-sad	103	0%	67	0%	13	0%	23	0%
19-qaf	116	0%	85	0%	15	0%	16	0%
Total	69,441	100%	54,845	100%	7,310	100%	7,286	100%

B.2 Sentence-level splits across domains and readership groups

Domain	Readership Group	All		Train		Dev		Test	
Arts & Humanities	**Foundational**	24,978	36%	20,161	37%	2,397	33%	2,420	33%
Arts & Humanities	**Advanced**	15,285	22%	11,982	22%	1,653	23%	1,650	23%
Arts & Humanities	**Specialized**	10,179	15%	7,755	14%	1,090	15%	1,334	18%
STEM	**Foundational**	533	1%	453	1%	80	1%	0	0%
STEM	**Advanced**	1,948	3%	1,741	3%	137	2%	70	1%
STEM	**Specialized**	2,199	3%	1,600	3%	258	4%	341	5%
Social Sciences	**Foundational**	2,270	3%	1,355	2%	600	8%	315	4%
Social Sciences	**Advanced**	5,463	8%	4,394	8%	514	7%	555	8%
Social Sciences	**Specialized**	6,586	9%	5,404	10%	581	8%	601	8%
Arts & Humanities		50,442	73%	39,898	73%	5,140	70%	5,404	74%
STEM		4,680	7%	3,794	7%	475	6%	411	6%
Social Sciences		14,319	21%	11,153	20%	1,695	23%	1,471	20%
	Foundational	27,781	40%	21,969	40%	3,077	42%	2,735	38%
	Advanced	22,696	33%	18,117	33%	2,304	32%	2,275	31%
	Specialized	18,964	27%	14,759	27%	1,929	26%	2,276	31%
		69,441	100%	54,845	100%	7,310	100%	7,286	100%

C BAREC Corpus Details

C.1 Resources

We present the corpus sources in groups of their general intended purpose.

C.1.1 Education

Emarati Curriculum The first five units of the UAE curriculum textbooks for the 12 grades in three subjects: Arabic language, social studies, Islamic studies (Khalil et al., 2018).

ArabicMMLU 6,205 question and answer pairs from the ArabicMMLU benchmark dataset (Koto et al., 2024).

Zayed Arabic-English Bilingual Undergraduate Corpus (ZAEBUC) 100 student-written articles from the Zayed University Arabic-English Bilingual Undergraduate Corpus (Habash and Palfreyman, 2022).

Arabic Learner Corpus (ALC) 16 L2 articles from the Arabic Learner Corpus (Alfaifi, 2015).

Basic Travel Expressions Corpus (BTEC) 20 documents from the MSA translation of the Basic Traveling Expression Corpus (Eck and Hori, 2005; Takezawa et al., 2007; Bouamor et al., 2018).

Collection of Children poems Example of the included poems: My language sings (لغتي تغني), and Poetry and news (أشعار وأخبار) (Al-Safadi, 2005; Taha-Thomure, 2007).

ChatGPT To add more children's materials, we ask Chatgpt to generate 200 sentences ranging from 2 to 4 words per sentence, 150 sentences ranging from 5 to 7 words per sentence and 100 sentences ranging from 8 to 10 words per sentence.[5] Not all sentences generated by ChatGPT were correct. We discarded some sentences that were flagged by the annotators. Table 9 shows the prompts and the percentage of discarded sentences for each prompt.

C.1.2 Literature

Hindawi A subset of 264 books extracted from the Hindawi Foundation website across different different genres.[6]

Kalima The first 500 words of 62 books from Kalima project.[7]

Green Library 58 manually typed books from the Green Library.[8]

Arabian Nights The openings and endings of the opening narrative and the first eight nights from the Arabian Nights (Unknown, 12th century). We extracted the text from an online forum.[9]

Hayy ibn Yaqdhan A subset of the philosophical novel and allegorical tale written by Ibn Tufail (Tufail, 1150). We extracted the text from the Hindawi Foundation website.[10]

Sara The first 1000 words of *Sara*, a novel by Al-Akkad first published in 1938 (Al-Akkad, 1938). We extracted the text from the Hindawi Foundation website.[11]

The Suspended Odes (Odes) The ten most celebrated poems from Pre-Islamic Arabia (المعلقات Mu'allaqat). All texts were extracted from Wikipedia.[12]

C.1.3 Media

Majed 10 manually typed editions of Majed magazine for children from 1983 to 2019.[13]

ReadMe++ The Arabic split of the ReadMe++ dataset (Naous et al., 2024).

Spacetoon Songs The opening songs of 53 animated children series from Spacetoon channel.

Subtitles A subset of the Arabic side of the OpenSubtitles dataset (Lison and Tiedemann, 2016).

WikiNews 62 Arabic WikiNews articles covering politics, economics, health, science and technology, sports, arts, and culture (Abdelali et al., 2016).

C.1.4 References

Wikipedia A subset of 168 Arabic wikipedia articles covering Culture, Figures, Geography, History, Mathematics, Sciences, Society, Philosophy, Religions and Technologies.[14]

[5] https://chatgpt.com/
[6] https://www.hindawi.org/books/categories/
[7] https://alc.ae/publications/kalima/
[8] https://archive.org/details/201409_201409
[9] http://al-nada.eb2a.com/1000lela&lela/
[10] https://www.hindawi.org/books/90463596/
[11] https://www.hindawi.org/books/72707304/
[12] https://ar.wikipedia.org/wiki/المعلقات
[13] https://archive.org/details/majid_magazine
[14] https://ar.wikipedia.org/

Prompt	Targeted #Words per Sentence	Prompt Text	% Discarded
Prompt 1	2-4	I am creating a children's textbook to practice reading in Arabic. I need short sentences containing 2 to 4 words that are limited to children's vocabulary. Give me 200 sentences in Standard Arabic -- no need to include English.	1.5%
	Examples	الشمس مشرقة. البنت تأكل الفاكهة.	
Prompt 2	5-7	I am creating a children's textbook to practice reading in Arabic. I need 5-word, 6-word, and 7-word sentences that are limited to children's vocabulary. Give me 150 sentences in Standard Arabic -- no need to include English.	1.3%
	Examples	الأسد ينام تحت شجرة كبيرة. الأطفال يلعبون في الملعب ويضحكون بسعادة كبيرة.	
Prompt 3	8-10	I am creating a children's textbook to practice reading in Arabic. I need long sentences (8-word, 9-word, and 10-word sentences) that are limited to children's vocabulary. Give me 100 sentences in Standard Arabic -- no need to include English.	1.0%
	Examples	الأرنب يقفز فوق العشب الأخضر في الصباح الباكر. القرد يتسلق الأشجار بسرعة ويقفز ببراعة من فرع إلى فرع.	

Table 9: ChatGPT Prompts. % Discarded is the percentage of discarded sentences due to grammatical errors.

Constitutions The first 2000 words of the Arabic constitutions from 16 Arabic speaking countries, collected from MCWC dataset (El-Haj and Ezzini, 2024).

UN The Arabic translation of the Universal Declaration of Human Rights.[15]

C.1.5 Religion

Old Testament The first 20 chapters of the Book of Genesis (Smith and Van Dyck, 1865).[16]

New Testament The first 16 chapters of the Book of Matthew (Smith and Van Dyck, 1860).[16]

Quran The first three Surahs and the last 14 Surahs from the Holy Quran. We selected the text from the Quran Corpus Project (Dukes et al., 2013).[17]

Hadith The first 75 Hadiths from Sahih Bukhari (al Bukhari, 846). We selected the text from the LK Hadith Corpus[18] (Altammami et al., 2019).

Some datasets are chosen because they already have annotations available for other tasks. For example, dependency treebank annotations exist for **Odes**, **Quran**, **Hadith**, **1001**, **Hayy**, **OT**, **NT**, **Sara**,**WikiNews**, **ALC**, **BTEC**, and **ZAEBUC** (Habash et al., 2022a).

[15] https://www.un.org/ar/about-us/universal-declaration-of-human-rights

[16] https://www.arabicbible.com/

[17] https://corpus.quran.com/

[18] https://github.com/ShathaTm/LK-Hadith-Corpus

C.2 Domains

Arts & Humanities The Arts and Humanities domain comprised the following subdomains.

- *Literature and Fiction:* Encompasses novels, short stories, poetry, and other creative writing forms that emphasize narrative and artistic expression.
- *Religion and Philosophy:* Contains religious texts, philosophical works, and related writings that explore spiritual beliefs, ethics, and metaphysical ideas.
- *Education and Academic Texts (on Arts and Humanities):* Includes textbooks, scholarly articles, and educational materials that are often structured for learning and academic purposes.
- *General Knowledge and Encyclopedic Content (on Arts and Humanities):* Covers reference materials such as encyclopedias, almanacs, and general knowledge articles that provide broad information on various topics.
- *News and Current Affairs (on Arts and Humanities):* Includes newspapers, magazines, and online news sources that report on current events and issues affecting society.

Social Sciences The Social Sciences domain comprised the following subdomains.

- *Business and Law:* Encompasses legal texts, business strategies, financial reports, and corporate documentation relevant to professional and legal contexts.

- *Social Sciences and Humanities:* Covers disciplines like sociology, anthropology, history, and cultural studies, which explore human society and culture.
- *Education and Academic Texts (on Social Sciences):* Includes textbooks, scholarly articles, and educational materials that are often structured for learning and academic purposes.
- *General Knowledge and Encyclopedic Content (on Social Sciences):* Covers reference materials such as encyclopedias, almanacs, and general knowledge articles that provide broad information on various topics.
- *News and Current Affairs (on Social Sciences):* Includes newspapers, magazines, and online news sources that report on current events and issues affecting society.

STEM The Science, Technology, Engineering and Mathematics domain comprised the following subdomains.

- *Science and Technology:* Includes scientific research papers, technology articles, and technical manuals that focus on advancements and knowledge in science and tech fields.
- *Education and Academic Texts (on STEM):* Includes textbooks, scholarly articles, and educational materials that are often structured for learning and academic purposes.
- *General Knowledge and Encyclopedic Content (on STEM):* Covers reference materials such as encyclopedias, almanacs, and general knowledge articles that provide broad information on various topics.
- *News and Current Affairs (on STEM):* Includes newspapers, magazines, and online news sources that report on current events and issues affecting society.

C.3 Readership Groups

Foundational This level includes learners, typically up to 4th grade or age 10, who are building basic literacy skills, such as decoding words and understanding simple sentences.

Advanced Refers to individuals with average adult reading abilities, capable of understanding a variety of texts with moderate complexity, handling everyday reading tasks with ease.

Specialized Represents readers with advanced skills, typically starting in 9th grade or above in specialized topics, who can comprehend and engage with complex, domain-specific texts in specialized fields.

Resource	#Documents	#Sentences	#Words
al-Kashkuul	17	330	2,306
Arabian Nights	24	669	6,835
ALC	16	676	8,395
ArabicMMLU	344	6,205	187,604
BTEC	20	1,865	14,663
chatGPT	3	443	2,502
Constitutions	16	1,490	30,370
Emarati Curriculum	126	13,365	113,952
Green Library	58	2,809	45,078
Hadith	75	672	7,057
Hanging Odes	10	764	7,269
Hayy ibn Yaqdhan	1	65	1,038
Hindawi	275	13,195	227,677
Kalima	62	2,767	43,423
Majed	294	11,490	121,126
Mama Makes Bread	1	39	468
My Language Sings	16	362	1,897
New Testament	16	566	9,471
Old Testament	20	525	8,874
Poems and News	1	391	1,239
Poems of Suleiman Al-Issa	1	97	336
Quran	42	405	7,744
ReadMe++	88	1,371	32,131
Sara	1	57	1,169
Spacetoon Songs	53	870	3,836
Subtitles	11	502	3,207
Universal Declaration of Human Rights	1	88	1,276
WikiNews	62	875	15,967
Wikipedia	168	5,402	117,100
ZAEBUC	100	1,086	15,361
Totals	**1,922**	**69,441**	**1,039,371**

Table 10: **BAREC** Corpus Details: the texts used to build the dataset, and the number of documents, sentences, and words extracted from each text.

D Additional Results

D.1 Confusion Matrix

Figure 3 shows the confusion matrix for the best-performing model from Stage 1: the AraBERTv2 model trained on D3Tok sentences with Cross-Entropy (CE) loss. The matrix uses F-scores to account for the unbalanced distribution of readability levels. The strong diagonal indicates a high rate of exact matches between predicted and gold labels. However, the model exhibits more disagreement at the higher, more difficult levels— likely due to the scarcity of training examples in those levels. Additionally, the model shows a tendency to under-estimate readability levels, favoring lower labels. This aligns with the patterns observed in the error analysis discussed in Section 6.5.

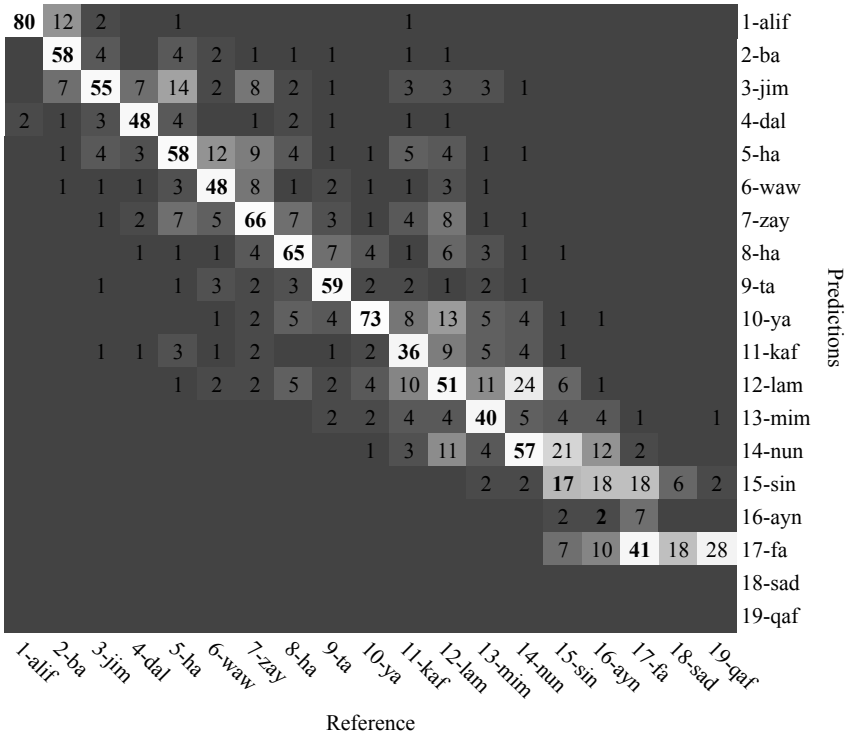

Figure 3: Confusion matrix of F-score across the different readability levels for the best model from stage 1.

D.2 All Loss Functions

Input	Model	Acc^{19}	±1 Acc^{19}	Dist	QWK
Word	SVM	36.2%	47.9%	2.03	53.4%
D3Tok	SVM	37.2%	49.3%	1.92	56.5%
Word	DecisionTree	27.2%	41.2%	2.50	44.2%
D3Tok	DecisionTree	29.9%	44.2%	2.33	52.5%
D3Tok	AraBERTv2				
	+CE	**56.6%**	69.9%	1.14	80.0%
	+EMD	55.3%	70.3%	**1.11**	81.2%
	+OLL2	35.2%	70.3%	1.25	82.0%
	+OLL15	47.3%	71.1%	1.13	82.8%
	+OLL1	50.8%	71.5%	1.12	81.7%
	+OLL05	53.1%	68.8%	1.18	79.7%
	+SOFT2	55.8%	69.8%	1.15	80.0%
	+SOFT3	56.4%	69.9%	1.14	80.1%
	+SOFT4	56.4%	69.9%	1.15	79.6%
	+SOFT5	56.2%	69.5%	1.17	79.3%
	+Reg	43.1%	**73.1%**	1.13	**84.0%**

Table 11: Loss functions comparisons on **BAREC** Dev set. For SVM and Decision Tree classifiers, we used count vectorizer.

D.3 Impact of Training Granularity on Readability Level Prediction

To analyze the effect of training granularity on readability level prediction, we compare two approaches: (1) training on all 19 levels and then mapping predictions to lower levels (7, 5, or 3), and (2) training directly on the target granularity.

Table 12 presents the results of this comparison. Overall, training on 19 levels and then mapping achieves slightly better performance across for 5-level and 3-level granularities compared to direct training. Moreover, the performance gap between the two approaches widens as the target granularity becomes coarser, suggesting that finer-grained supervision during training provides more informative learning signals, which translate into improved generalization when predictions are mapped into broader scales.

Train Gran	Dev Gran	Input	Model	Acc	±1 Acc	Dist	QWK
19	7	D3Tok	CE	**65.9%**	88.9%	0.51	79.9%
7	7	D3Tok	CE	65.2%	**89.5%**	**0.50**	**81.0%**
19	5	D3Tok	CE	**70.3%**	93.5%	**0.37**	**78.3%**
5	5	D3Tok	CE	67.8%	**93.7%**	0.39	77.3%
19	3	D3Tok	CE	**76.5%**	**97.6%**	**0.26**	**74.7%**
3	3	D3Tok	CE	74.4%	96.9%	0.29	74.0%

Table 12: Comparison between training on 19 levels then mapping to the target granularity vs. training directly on the target granularity.

D.4 Ensembles & Oracles

CE Word	CE Lex	CE D3Tok	CE D3Lex	EMD D3Tok	Reg D3Tok	Metrics			
						Acc[19]	±1 Acc[19]	Dist	QWK
✓						51.6%	65.9%	1.32	76.3%
	✓					50.1%	65.4%	1.29	77.7%
		✓				**56.6%**	69.9%	1.14	80.0%
			✓			53.2%	67.1%	1.24	78.6%
				✓		55.3%	70.3%	**1.11**	81.2%
					✓	43.1%	**73.1%**	1.13	**84.0%**
Average						46.9%	72.5%	**1.11**	83.4%
Most Common						56.3%	70.0%	1.13	80.4%
Oracle Combinations									
✓	✓					62.4%	76.6%	0.88	88.4%
✓		✓				63.5%	76.7%	0.89	87.7%
✓			✓			63.2%	76.6%	0.88	88.2%
✓				✓		63.3%	77.9%	0.83	89.2%
✓					✓	62.0%	80.7%	0.77	90.8%
✓	✓	✓	✓			69.5%	82.3%	0.67	91.4%
✓	✓	✓	✓	✓		72.0%	84.5%	0.59	92.6%
✓	✓	✓	✓		✓	73.6%	86.6%	0.53	93.4%
✓	✓	✓	✓	✓	✓	75.2%	87.4%	0.50	93.8%

Table 13: Comparison between individual models, ensembles and oracles on **BAREC** Dev set.