Guidelines for Fine-grained Sentence-level Arabic Readability Annotation

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

Abstract

This paper presents the annotation guidelines of the Balanced Arabic Readability Evaluation Corpus (**BAREC**), a large-scale resource for fine-grained sentence-level readability assessment in Arabic. **BAREC** includes 69,441 sentences (1M+ words) labeled across 19 levels, from kindergarten to postgraduate. Based on the Taha/Arabi21 framework, the guidelines were refined through iterative training with native Arabic-speaking educators. We highlight key linguistic, pedagogical, and cognitive factors in determining readability and report high inter-annotator agreement: Quadratic Weighted Kappa 81.8% (substantial/excellent agreement) in the last annotation phase. We also benchmark automatic readability models across multiple classification granularities (19-, 7-, 5-, and 3-level). The corpus and guidelines are publicly available.[1]

1 Introduction

Text readability plays a crucial role in comprehension, retention, reading speed, and engagement (DuBay, 2004). When texts exceed a reader's ability, they can lead to frustration and disengagement (Klare, 1963). Readability is shaped by both the content and presentation (Nassiri et al., 2023). In educational settings, readability leveling is widely used to align texts with students' reading abilities, promoting independent and more effective learning (Allington et al., 2015; Barber and Klauda, 2020).

Fine-grained readability systems, like Fountas and Pinnell's 27-level scale in English (Fountas and Pinnell, 2006), and Taha's 19-level Arabic system (Taha-Thomure, 2017), guide progression from early readers to adult fluency. These levels support instructional goals and can be mapped to broader categories for practical use in NLP.

We present the Balanced Arabic Readability Evaluation Corpus (**BAREC**), a large-scale dataset of 69K+ sentences[2] (1M+ words) across a broad space of genres and 19 readability levels. Based on the Taha/Arabi21 framework (Taha-Thomure, 2017), which has been instrumental in tagging over 9,000 children's books, **BAREC** guidelines enable standardized, sentence-level readability evaluation across diverse genres and educational levels, ranging from kindergarten to postgraduate comprehension (see Table 1). Our contributions are as follows:

RL	Grade	Example
1	KG	كُرَة Ball
3	1st	غُرْفَةُ النَوم The bedroom
6	2nd	سُلوكي مَسْؤوليَتْي My behavior is my responsibility
10	4th	كانت الحديقة واسعة، تطل على شاطئ النيل، The garden was spacious, overlooking the Nile.
14	8th	تعريف أصول الفقه Definition of Islamic Jurisprudence Principles
17	Uni	بين طعن القَنا وخَفْق البُنودِ Between lance thrusts and ensign flutters

Table 1: Examples by Reading Level (RL) and grade.

- We **define detailed annotation guidelines** for Arabic sentence-level readability across a fine-grained 19-level scale.
- We **apply and refine these guidelines** through annotation of a diverse, large-scale corpus, analyzing annotator agreement and sources of difficulty in this nuanced task.
- We **build and evaluate readability models** across multiple granularities (19, 7, 5, and 3 levels) to provide baseline results for various research and application needs.

Next, §2 reviews related work, §3 outlines the annotation framework, §4 covers data selection, and §5 discusses evaluation results.

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

[2] We use *sentence* to refer to syntactic sentences as well as shorter standalone text segments (e.g., phrases or titles).

Authors	Project	Metric	Levels	Unit	Size	Content
Al-Khalifa and Al-Ajlan (2010)	Arability	Readability	3	Document	150	School Textbooks
Forsyth (2014)	DLI Corpus	ILR	5 (3)	Document	179	L2 Learner
Kilgarriff et al. (2014)	KELLY	CEFR	6	Word	9,000	Most Frequent
Taha-Thomure (2017)	Taha/Arabi21	Readability	19	Document	9,000	Children's Books
Al Khalil et al. (2020)	SAMER Lexicon	Readability	5	Word	40,000	General Vocab
Habash and Palfreyman (2022)	ZAEBUC	CEFR	6	Document	214	Prompted Essays
Naous et al. (2024)	ReadMe++	CEFR	6	Sentence	1,945	Multi-domain
Soliman and Familiar (2024)	Arabic Vocab Profile	CEFR	2	Word	1,200	L2 Learner (A1, A2)
El-Haj et al. (2024)	DARES	Grade Level	12	Sentence	13,335	School Textbooks
Alhafni et al. (2024)	SAMER Corpus	Readability	3	Word	159,265	Literature
Bashendy et al. (2024)	QAES	AES	7×5	Document	195	Argumentative Essays
Our Work	BAREC	Readability	19 (7–5–3)	Sentence	69,441	Multi-domain

Table 2: Overview of Arabic readability and proficiency-related corpora.

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 explored text leveling and assessment in multiple frameworks (Nassiri et al., 2023).

Taha-Thomure (2017) proposed a 19-level Arabic text leveling framework for educators, inspired by Fountas and Pinnell (2006) and focused on children's literature. Targeting full texts (books), particularly for early education, with 11 of the 19 levels covering up to 4th grade, the system supports teachers in matching books to students' reading abilities. Taha-Thomure (2017)'s procedural framework outlines ten qualitative and quantitative criteria: text genre, abstractness of ideas, vocabulary and its proximity to dialects, text authenticity, book production quality, content suitability, sentence structure, illustrations, use of diacritics, and word count. The Arab Thought Foundation adopted this framework under its Arabi21 initiative, which funded the leveling of over 9,000 children's books.

Other efforts applied CEFR leveling to Arabic, including the KELLY project's frequency-based word lists, manually annotated corpora such as ZAEBUC (Habash and Palfreyman, 2022) and ReadMe++ (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). Bashendy et al. (2024) presented a corpus of Arabic essays annotated across organization and style traits.

Automated readability assessment in Arabic has evolved from rule-based models using surface features (Al-Dawsari, 2004; Al-Khalifa and Al-Ajlan, 2010) to machine learning approaches with POS, morphology (Forsyth, 2014; Saddiki et al., 2018), and script features like OSMAN (El-Haj and Rayson, 2016). Recent work (Liberato et al., 2024) shows strong results with pretrained models on the SAMER corpus.

Our Approach Building on prior work, we curated the **BAREC** corpus across diverse genres and readability levels, manually annotating it at the sentence level using adapted Taha/Arabi21 guidelines (Taha-Thomure, 2017). Sentence-level annotation balances the coarse granularity of document-level labels and the limited context of word-level labels. This allows finer control and more objective assessment of textual variation. Table 2 compares **BAREC** with earlier efforts. To our knowledge, **BAREC** is the largest and most fine-grained manually annotated Arabic readability resource.

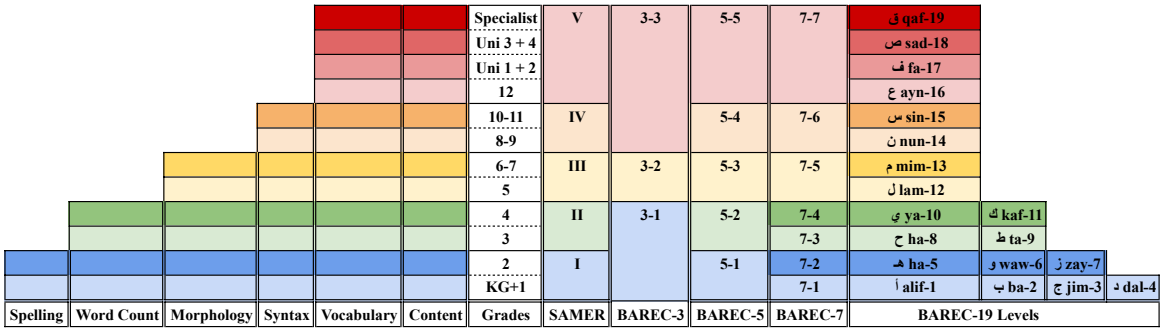

Figure 1: The **BAREC** *Pyramid* illustrates the relationship across **BAREC** levels and linguistic dimensions, three collapsed variants (3 levels, 5 levels and 7 levels), and educational grades.

3 BAREC Annotation Guidelines

3.1 Annotation Desiderata

Our guidelines and annotation decisions follow several key principles. **Comprehensive Coverage** ensures representation across all 19 levels, from kindergarten to postgraduate, with finer distinctions at early stages. **Objective Standardization** defines levels using consistent linguistic and content-based criteria, avoiding overreliance on surface features like word or sentence length. **Bias Mitigation** promotes inclusivity across Arab world regions and cultural content. **Balanced Coverage** supports diversity in levels, genres, and topics, especially addressing material scarcity in areas like children's literature. **Quality Control** is maintained through trained annotators and regular checks for inter-annotator agreement and consistency. Finally, **Ethical Considerations** include respecting copyrights and fairly compensating annotators.

3.2 Readability Levels

BAREC readability annotation assigns one of 19 levels to each sentence in the corpus. We retain Taha-Thomure (2017)'s 19-level naming system based on the Abjad order: **1-alif**, **2-ba**, **3-jim**, ... **19-qaf**, but extend and adjust the original guidelines, which were designed for book-level annotation to this task. 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 allow easy mapping from fine-grained to coarse-grained levels, but manual annotation only happened on 19 levels. For example, level **11-kaf** maps to level **4** (of 7), level **2** (of 5) and level **1** (of 3). See Table 3 for representative examples.

3.3 Readability Annotation Principles

Reading and Comprehension Readability reflects how easily independent readers can both read and comprehend a text without teacher or parent support. We focus on basic pronunciation (recovering lexical diacritics) and literal understanding, not on grammatical analysis or deep interpretation.

Sentence-level Focus We assess readability at the sentence level, independent of broader context, source, or author intent. This deliberate choice avoids genre-based assumptions and enables fair, objective comparison across diverse texts. Mapping sentence-level judgments to larger units is left for future work.

Target Audience While religious content is part of basic public education in the Arab world, we make no assumptions about readers' religious backgrounds or prior knowledge. Readability is judged purely on linguistic and cognitive grounds. Our guidelines reflect Modern Standard Arabic (MSA) as used in Egypt, the Gulf, and the Levant, leaving variations in other regions for future work.

Readability Level Keys Annotators start from the lowest (easiest) level and raise it based on key features: lexical, morphological, syntactic, or semantic. See Sections 3.4 and 3.5 below for details.

A Note on Arabic Diacritics While diacritics can aid comprehension, we assess readability without relying on them. This departs from Taha-Thomure (2017), who consider diacritics a key design feature in children's books. In ambiguous cases, we choose the simpler meaning, e.g., هذه سلطة بدون خيار *hðh slTħ bdwn xyAr*[3] is read as 'a salad without cucumbers' not 'an authority without choices'.

[3]HSB transliteration (Habash et al., 2007).

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 لا،	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 جدًا — **حتى إنه كان من المعروف عنها أنها تنمو بين أحجار الرصف، وتنبثق في كل مكان مثل الحشائش الضارة** — وتحمل اسمًا قبيحًا جدًا 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 الأواريَّ لأيًا ما أُبَيِّنُها والنُّؤيُ كالحَوْضِ بالمَظْلومَةِ الجَلَدِ**	**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*

Table 3: 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.

3.4 Dimensions of Textual Features

To determine the **BAREC** level, we define six textual dimensions that identify *key* features necessary to unlock each level:

1. Number of Words Counts unique printed words (ignoring punctuation and diacritics). Used only up to level **11-kaf** (max 20 words).

2. Orthography & Phonology Focuses on word length (syllables) and letters like Hamzas. Final diacritics are ignored (words read in *waqf*), e.g., أَرْنَبٌ *Âar.nabũ* 'rabbit' has 2 syllables: *ar-nab*.

3. Morphology Covers derivation and inflection (tense, voice, number, etc.). Simpler forms appear at lower levels (e.g., present tense before past, singular before plural). Used up to level **13-mim**.

4. Syntactic Structures Tracks sentence complexity, from single words (**1-alif**) to complex constructions. Used up to level **15-sin**.

5. Vocabulary Central at all levels. Overlapping dialect and MSA vocabulary appear at easier levels; technical terms are introduced at harder levels. Arabized foreign words are treated as part of the language, while non-Arabic script is excluded.

6. Ideas & Content Evaluates needed prior knowledge, symbolic unpacking, and conceptual linking. Levels progress from familiar to specialized knowledge and from literal to abstract ideas. We recognize that such evaluations are complex and may vary subjectively among readers within the same age or education group.

Problems and Difficulties Annotators are instructed to report issues such as spelling errors, colloquial language, or sensitive topics. Difficulty is noted when annotations cannot be made due to conflicting guidelines.

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**. A full set of examples with explanations of leveling choices is in Table 3. The ***Annotation Cheat Sheet*** used by the annotators in Arabic and its translation in English are included in Appendix A. The full guidelines are publicly available.[1] For more on Arabic linguistic features, see Habash (2010).

3.5 Annotation Process

Sentence Segmentation Since our starting point is a text excerpt, typically a paragraph or two (∼500±200 words) from each source, we begin with sentence-level segmentation and initial text flagging. We followed the Arabic sentence segmentation guidelines by Habash et al. (2022).

Sentence Readability Annotation Each annotator is presented with a batch of 100 randomly selected sentences to annotate. The annotation was done through a simple Google Sheet interface (see Appendix A.3), which provides details such as sentence word count, and the guidelines constraints for the selected level to provide feedback confirmation to the annotator. The annotators are instructed to follow this procedure: **First** they read the sentence and make sure it has no flaws that can lead to excluding it. **Second**, they think about the meaning of the sentence noting any ambiguities due to diacritic absence or limited context, and consciously decide on the simpler reading in case of multiple readings. **Third**, they make an initial assessment of the lowest possible level based on word count. **Fourth**, they look for specific phenomena that allow increasing the level to the highest possible. For example, the sixth sentence in Table 3, سلوكي مسؤوليتي *slwky msŵwlyty* 'my behavior is my responsibility' has two words, which automatically sets it as level **2-ba** or higher. The presence of the first person pronominal clitic ي+ *+y* elevates the level to **3-jim**; however, the fact that the second word has five syllables raises the level further to **6-waw**. No other keys can take it higher.

Annotation averaged 2.5 hours per 100-sentence batch (1.5 minutes per sentence), reflecting the careful and rigorous approach taken by annotators to ensure high-quality, consistent labeling across a diverse and challenging dataset.

3.6 Annotation Team

The **BAREC** annotation team included six native Arabic-speaking educators (A0-A5), most with advanced degrees in Arabic Literature or Linguistics. A0 had prior experience in computational linguistics annotation, while A1-A5 brought extensive expertise in readability assessment from the Taha/Arabi21 project. A0 handled sentence segmentation and initial text selection; and A5 led the annotation team in assigning readability labels. Annotator profiles, covering demographic, educational, linguistic, and teaching backgrounds, are listed in Appendix A.4.

3.7 Training and Quality Control

Annotators A1-A5 received thorough training, including three shared pilot rounds that enabled in-depth discussion and refinement of the guidelines.

To ensure consistency, the initial 10,658 sentences (Phase 1) were double-reviewed before annotating the full 69K (1M+ words). Inter-annotator agreement (IAA) was assessed on 19 blind batches (excluding pilots 1 and 2), followed by group unification to support quality control and prevent drift. Only unified labels appear in the official release. The multiple IAA annotations will be released separately to support research on readability annotations.[1] Details on IAA are in Section 5.3).

In total, the annotators labeled 92.6K sentences; 25% were excluded from the final corpus: 3.3% were problematic (typos and offensive topics), 11.5% from early double annotations, and 10.3% from IAA rounds (excluding unification).

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

Table 4: **BAREC** corpus statistics in documents, sentences, and words, across domain and readership levels.

4 BAREC Corpus

4.1 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**,[4] and three readership groups: **Foundational**, **Advanced**, and **Specialized**.[5] Table 4 shows the distribution of the documents, sentences and words across domains and groups. The corpus emphasizes educational coverage, with a higher-than-usual proportion of foundational-level texts. Domain variation reflects text availability and reader interest (more Arts & Humanities, less STEM). Texts were sourced from 30 resources, all either public domain, within fair use, or used with permission. Some were selected due to existing annotations. Notably, 25% of sentences came from new sources that were manually digitized. See Appendix C for resource details.

4.2 Readability Statistics

Figure 2 shows sentence distribution across **BAREC**-19 levels and their mappings to coarser levels (7, 5, and 3). The distribution is uneven, with 63% of sentences in the middle levels (**10-ya**∼fourth grade to **14-nun**∼ninth grade) reflecting natural text complexity and real-world usage.

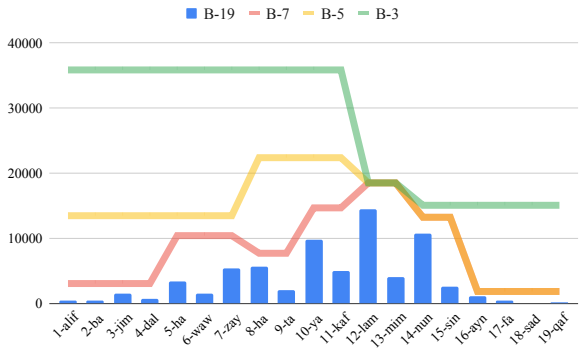

Figure 2: The distribution of sentences across **BAREC**-19 levels (blue), and their mapping to coarser levels.

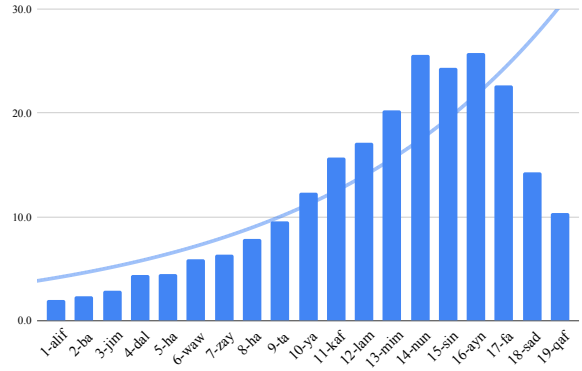

Figure 3: The average sentence word count across **BAREC**-19 levels, with trend line.

Figure 3 shows average sentence length by level, which correlates strongly with readability (Pearson r=81%). The drop at higher levels may result from shorter classical poetry lines.

Figure 4 shows *relative* distribution of readership groups and domains across readability levels. Foundational texts dominate lower levels and specialized texts higher ones. STEM and Social Science texts have a higher relative appearance in the upper mid levels.

[4]**Arts & Humanities:** literature, philosophy, religion, education, and related news. **Social Sciences:** business, law, social studies, education, and related news. **STEM:** science, technology, engineering, math, education, and related news.

[5]**Foundational:** Learners up to 4th grade (age 10), focused on basic literacy skills. **Advanced:** Adult readers with average abilities, handling moderate complexity texts. **Specialized:** Advanced readers (typically 9th grade+), engaging with domain-specific texts.

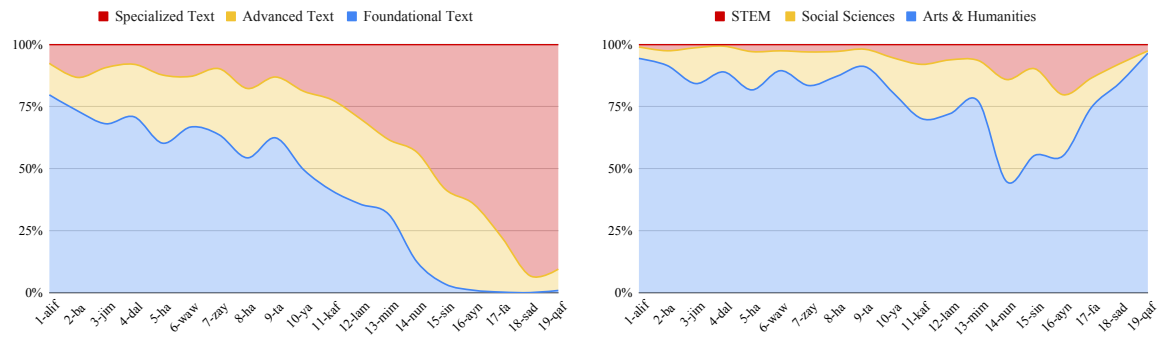

Figure 4: The relative distribution of readership groups and domains across **BAREC** levels.

5 Evaluation and Analysis

5.1 Metrics

We evaluate readability models and IAA using Accuracy, Adjacent Accuracy, Average Distance, and Quadratic Weighted Kappa (QWK), with QWK as our primary metric.

Accuracy (Acc) The percentage of cases where the predicted class matches the reference class in the 19-level scheme (**Acc**^{19}), as well as three variants, **Acc**^{7}, **Acc**^{5}, and **Acc**^{3}, which collapse the 19-level scheme into 7, 5, and 3 levels, respectively (Section 3.2).

Adjacent Accuracy (±1 Acc^{19}) The proportion of predictions that are either exactly correct or off by at most one level.

Average Distance (Dist) The average absolute difference between two sets of labels. For example, the distance between **2-ba** and **4-dal** is 2.

Quadratic Weighted Kappa (QWK) An extension of Cohen's Kappa (Cohen, 1968; Doewes et al., 2023), measuring agreement between predicted and true labels, with a quadratic penalty for larger misclassifications.

5.2 Corpus Splits

We split the corpus at the document level into **Train (∼80%)**, **Dev (∼10%)**, and **Test (∼10%)**. Sentences from IAA studies are distributed across splits. For resources with existing splits, such as CamelTB (Habash et al., 2022) and ReadMe++ (Naous et al., 2024), we adopted their original splits. Table 5 reports the splits by documents, sentences, and words. Due to IAA and external corpus constraints, final proportions slightly deviate from exact 80-10-10. See Appendix B for full and split readability level distributions.

Split	#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 5: **BAREC** corpus splits.

Stage	#Sets	Distance	Acc^{19}	±1Acc^{19}	QWK
Pilot 3	1	1.69	37.5%	58.5%	79.3%
Phase 1	2	1.38	48.4%	64.4%	80.2%
Phase 2A	6	1.21	49.4%	67.4%	72.4%
Phase 2B	10	0.80	67.6%	78.3%	78.8%
Overall / Macro	19	1.04	58.2%	72.3%	76.9%
Phase 2 / Macro	16	0.96	60.8%	74.2%	76.4%
Phase 2 / Micro	16	0.95	61.1%	74.4%	81.8%

Table 6: Average pairwise inter-annotator agreement (IAA) across different annotation stages. Macro/Micro indicate the form of averaging, over sets or sentences, respectively. Phase 2 = Phase 2A and 2B.

5.3 Inter-Annotator Agreement (IAA)

Pairwise Agreement Table 6 summarizes results for 19 IAA sets (excluding Pilots 1 and 2). We observe steady improvement from Pilot 3 to Phase 2B, with reduced distance and higher accuracy. The overall macro-average QWK is 76.9%, indicating substantial agreement and suggesting that most disagreements are minor (Cohen, 1968; Doewes et al., 2023). In Phase 2, the final and largest phase, the micro-average QWK rises to 81.8%.

Figure 5 presents a confusion matrix of sentence-level pairwise agreements for Phase 2 IAA sentences, using F-scores to account for the unbalanced level distribution. The strong diagonal (exact matches) reflects a high degree of agreement, consistent with the overall IAA results. However, accuracy varies across levels, with more disagree-

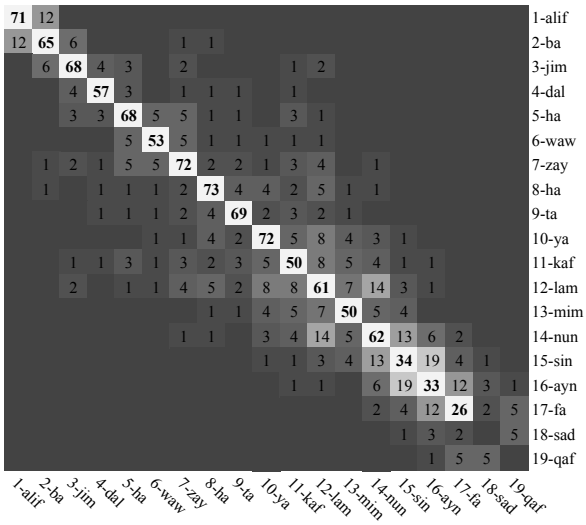

Figure 5: Confusion matrix for annotator pairwise agreement on Phase 2 IAA sentences normalized as F-scores.

ment at the harder higher levels. This may stem from the guidelines emphasizing vocabulary and content at the higher levels, features that are inherently more subjective than the textual feature cues used at lower levels.

Unification Agreement After each IAA study, annotators determined a unified readability level (UL) for each sentence. The UL falls within the Max-Min range of annotator labels 99.2% of the time and matches one of the annotators 86.8% of the time. Table 7 compares the micro-average performance of annotators in Phase 2, using both pairwise comparisons and the comparison between the UL and the rounded average level (AL) of annotators' choices. Table 7 also presents the results mapped to lower granularity levels (7, 5 and 3). We observe that overall, the AL-UL distance is smaller than the average pairwise distance among the annotators, and that its ±1 Acc is much higher, which suggests the average (AL) is more often than not closer to UL than any pair of annotators are to each other. The comparison across granularity levels shows that although the absolute Distance decreases, its relative magnitude (compared to the label range) increases. As expected, both Acc and ±1 Acc are higher with coarser level groupings. Appendix A.5 presents the results for each annotator against UL.

Error analysis To better understand annotator disagreement, we manually analyzed 100 randomly selected sentences with divergent readability labels. Table 8 presents representative examples

	19 Level	7 Level	5 Level	3 Level
Pairwise Distance	0.95	0.39	0.30	0.23
Relative to Range	5.0%	5.5%	6.0%	7.5%
Acc	61.1%	73.1%	75.2%	80.0%
±1 Acc	74.4%	92.0%	95.0%	97.3%
AL-UL Distance	0.52	0.26	0.22	0.18
Relative to Range	2.7%	3.7%	4.4%	5.9%
AL-UL Acc	61.2%	75.5%	78.9%	82.9%
AL-UL ±1 Acc	90.1%	98.5%	99.4%	99.5%

Table 7: Comparison of pairwise agreement micro averages across level granularities for all Phase 2 IAA sentences. UL = Unified Label; AL = Average Label.

with explanations. We found that 25% of disagreements were due to basic linguistic features (e.g., morphology, syntax, spelling), 12% involved emotional or symbolic content, 18% related to general advanced vocabulary, and 45% stemmed from domain-specific terminology in STEM, Humanities, or Social Sciences. This suggests that specialized vocabulary is the leading source of inconsistency, often due to differing expectations about what counts as general versus domain-specific language, and how specialization is defined. Some variation also stems from subjective views on what an *educated* Standard Arabic reader should know. In the future, we plan to develop readability lexicons to anchor our guidelines, building on efforts like the SAMER Lexicon (Al Khalil et al., 2020) and the Arabic Vocabulary Profile (Soliman and Familiar, 2024), but targeting 19 levels.

5.4 Automatic Readability Assessment

To establish a baseline for sentence-level readability classification, we fine-tune AraBERTv02 (Antoun et al., 2020) using the Transformers library (Wolf et al., 2019). Training is conducted on an NVIDIA V100 GPU for three epochs with a learning rate of 5×10^{-5}, a batch size of 64, and a cross-entropy loss function for multi-class classification across 19 levels. Table 9 presents the model's learning curve. We evaluate performance using varying proportions of the training data: $\frac{1}{8}$, $\frac{1}{4}$, $\frac{1}{2}$, and the full dataset. As shown in the table, model performance improves consistently with larger training data. Compared to the Phase 2 IAA micro averages (Table 6), the model's best Distance is 15.3% higher, and its best Accuracy is 5.3% absolute (8.7% relative) lower. However, the QWK is only marginally lower by just 0.8% absolute.

For a more extensive discussion of the automatic annotation results, see Elmadani et al. (2025).

Sentence (Arabic)	A1	A2	A3	A4	A5	UL	MM	Comments
أبي.. أبي.. *Dad .. Dad .. [lit. my father .. my father ..]*	2	2	2	3	3	**3**	*1*	First person singular pronoun is level 3.
احْتِضانُ الْأُمِ لَهُم. *The mother's embrace for them.*	9	12	5	5	5	**5**	*7*	Disagreement over احتضان 'embrace': standard or dialect aligned.
أشعر بالتعب والجوع.. *I feel tired and hungry..*	9	9	9	9	4	**9**	*5*	Vocabulary describing emotions (level 9).
يتم ضمان حيادية الإدارة بموجب القانون. *Administrative neutrality is guaranteed by law.*	12	12	12	14	12	**12**	*2*	Disagreement over حيادية 'neutrality': general advanced or specialized.

Table 8: Examples of Annotator Disagreements with Unified Levels (UL) and Max-Min Differences (MM)

Train	Distance	Acc^{19}	±1 Acc^{19}	QWK	Acc^{7}	Acc^{5}	Acc^{3}
12.5%	1.35	45.0%	61.3%	77.2%	56.8%	63.0%	71.3%
25.0%	1.33	46.9%	63.0%	77.6%	58.8%	64.3%	72.3%
50.0%	1.16	52.4%	68.1%	80.7%	62.9%	67.6%	74.0%
100.0%	1.09	55.8%	69.4%	81.0%	64.9%	69.1%	74.7%

Table 9: Performance at different training data sizes across multiple evaluation metrics.

6 Conclusions and Future Work

This paper presented the annotation guidelines of 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, to our knowledge, the largest Arabic readability corpus, covering diverse genres, topics, and audiences. We report high inter-annotator agreement (QWK 81.8% in Phase 2) that ensures reliable annotations. Benchmark results across multiple classification granularities (19, 7, 5, and 3 levels) demonstrate both the difficulty and feasibility of automated Arabic readability prediction.

Looking ahead, we plan to expand the corpus by increasing its size and diversity to include more genres and topics. We also aim to add annotations for vocabulary leveling and syntactic treebanks to study the effect of vocabulary and syntax on readability. Future work will analyze readability variations across genres and topics. Additionally, we intend to integrate our tools into a system that assists children's story writers in targeting specific reading levels.

The **BAREC** dataset, its annotation guidelines, and benchmark results, are publicly available to support future research and educational applications in Arabic readability assessment.[1]

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 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.

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.

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.

May Bashendy, Salam Albatarni, Sohaila Eltanbouly, Eman Zahran, Hamdo Elhuseyin, Tamer Elsayed, Walid Massoud, and Houda Bouamor. 2024. Qaes: First publicly-available trait-specific annotations for automated scoring of arabic essays. In *Proceedings of The Second Arabic Natural Language Processing Conference*, pages 337–351.

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).

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.

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.

Khalid N. Elmadani, Nizar Habash, and Hanada Taha-Thomure. 2025. A large and balanced corpus for fine-grained Arabic readability assessment. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL 2025)*, Vienna, Austria. Association for Computational Linguistics.

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. 2022. 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 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, Abdelhadi Soudi, and Tim Buckwalter. 2007. On Arabic Transliteration. In A. van den Bosch and A. Soudi, editors, *Arabic Computational Morphology: Knowledge-based and Empirical Methods*, pages 15–22. Springer, Netherlands.

Nizar Y Habash. 2010. *Introduction to Arabic natural language processing*, volume 3. Morgan & Claypool Publishers.

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.

Adam Kilgarriff, Frieda Charalabopoulou, Maria Gavrilidou, Janne Bondi Johannessen, Saussan Khalil, Sofie Johansson Kokkinakis, Robert Lew, Serge Sharoff, Ravikiran Vadlapudi, and Elena Volodina. 2014. Corpus-based vocabulary lists for language learners for nine languages. *Language Resources and Evaluation*, 48(1):121–163.

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.

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).

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.

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.

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 Annotation Interface

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، الاستفار 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 • التوخيم 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 Interface

Sentence/Phrase	Length	Level		Word Count	Spelling/Pronunciation	Morphology	Syntax	Vocabulary	Idea/Content	Notes
الجملة \ العبارة	عدد الكلمات	المستوى		عدد الكلمات	تهجئة/املاء	تصريف واشتقاق	تراكيب نحوية	مفردات	فكرة / محتوى	ملاحظات
خَبَّرَ	1	و (صف 2)	6-waw	٩ هو أعلى عدد كلمات مطبعية غير متكررة بدون علامات الترقيم	• كلمات من ٥ مقاطع (بدون حساب حركات الإعراب)	• الفعل الماضي المفرد والجمع • جمع مذكر سالم	• جملة فيها فعلين (مثلا جملة فعلية مفعولها أن المصدرية)	• مفردات فصيحة - سامر ١	• المحتوى من حياة القارئ. • لا رمزية في النص.	
جودي بقربي	2	ز (صف 2)	7-zay	١٠ هو أعلى عدد كلمات مطبعية غير متكررة بدون علامات الترقيم	• كلمات من ٦ مقاطع أو أكثر (بدون حساب حركات الإعراب) • أفعال/أسماء معتلة الآخر	• الفعل الماضي المثنى •الفعل المضارع المثنى • فعل الأمر المفرد • جمع التكسير • واو القسم (والله)	• مفعول فيه (ظروف زمان ومكان) • حال • أداة الاستفهام هل	• مفردات فصيحة شائعة - سامر ٢	• بعض الرمزية أو عدم التصريح المباشر بكل المقصود في الجملة	
بيروت في يوليو ١٩٦٦	4	ح (صف 3)	8-ha	١١ هو أعلى عدد كلمات مطبعية غير متكررة بدون علامات الترقيم		• فعل الأمر الجمع • نون النسوة في الأسماء والأفعال (انتظرنَ دورهنَ) • سوابق أخرى: سين الاستقبال ، واو الاستئناف، فاء العطف • (ثم ، حتى ، أو ، أم ، لكن ، أمّا)	• المفعول المطلق • المفعول لأجله • المفعول معه • جملة فعلية تتعدى إلى مفعولين	• مفردات فصيحة - سامر ١ و سامر ٢ • أحرف النفي • الأرقام (العربية أو الهندية) 1,000,000-1,001	• بعض الرمزية يحتاج معها القارئ إلى مساعدة من يشرح له المقصود من الفكرة	
كتابةُ خطَّةِ لمشروع الوحدة	4	ك (صف 4)	11-kaf	٢٠ هو أعلى عدد كلمات مطبعية غير متكررة بدون علامات الترقيم		• المشتقات على أنواعها (تركز على المشتقات العاملة لاسيما اسم الفاعل واسم المفعول)	• جملة أسمية خبرها جملة أسمية (فيها مبتدأ) • إضافة خيالية (لفظية) طويل القامة	• أسماء الوصل المثنى والجمع • متلازمات لفظية مثل شارد الذهن، وارف الظلال	• هناك درجة من الرمزية وحاجة للمعرفة السابقة كي يُفهم المقصود من الجملة	
اجْتَمَعَ الأَهْلُ في العِيدِ.	4	و (صف 2)	6-waw	٩ هو أعلى عدد كلمات مطبعية غير متكررة بدون علامات الترقيم	• كلمات من ٥ مقاطع (بدون حساب حركات الإعراب)	• الفعل الماضي المفرد والجمع • جمع مذكر سالم	• جملة فيها فعلين (مثلا جملة فعلية مفعولها أن المصدرية)	• مفردات فصيحة - سامر ١	• المحتوى من حياة القارئ. • لا رمزية في النص.	
وَلا يُخَالِطُنا عَجُزٌ وَلا خَوَرُ	4	ل (صف 5)	12-lam	لا حد لعدد الكلمات المطبعية		• التصغير	• جمل اعتراضية (تفسير - دعاء...) • استثناء • حصر • البدل (غير حالات ما بعد اسم الإشارة) • تمييز	• مفردات فصيحة - سامر ٣ • اسم الفعل : إيه - صَهْ - أمين - خيَّ - هاؤُم - هاك - هَيَّا - هيتَ - هَلُمَّ إلى - مَهْ - رويدَك - • الأرقام (العربية أو الهندية) < 1,000,000 • ذو • (بل - بلى - أجل)	• هناك درجة من الرمزية وحاجة للمعرفة السابقة كي يُفهم المقصود من الجملة	

This is a screenshot of the Google Sheet interface used for annotation. The first two columns on the left are the sentence and its word count. The third column is the readability level which is selected by drop down menus. The fourth yellow column and the first yellow row are not part of the interface, we added them for the purpose of explaining the structure to readers of this paper who do not know Arabic. The next 6 columns automatically display the text features from the annotation guidelines to help the annotators confirm their choices. The last column is for extra notes such as flagging problematic sentences.

A.4 Annotation Team

	A0^P	**A1**	**A2**	**A3**	**A4**	**A5**^{L}
Native Language	Arabic	Arabic	Arabic	Arabic	Arabic	Arabic
Other Language	En, Fr	En	En, Fr	En, Fr	En, Fr	En, Fr
Nationality	Syrian	Lebanese	Lebanese	Lebanese	Lebanese	Lebanese
Residence	USA	Lebanon	Lebanon	Lebanon	UAE	Lebanon
Gender	Female	Female	Female	Female	Female	Female
Background	Muslim	Muslim	Muslim	Muslim	Christian	Muslim
Degree	MA	BA	BA	MA	MA	B MA
Major	Applied Ling.	Arabic Lit.	Geography	Arabic Lit.	Arabic Lit.	Arabic Lit.
Experience	CT, LA, RA	PT, LA	PT, LA	CT, LA	CT, LA	CT, LA, RA
School	Private	-	-	Public&Private	Private	Public
Level	University	Elementary	Elementary	Secondary	Secondary	Secondary
Students	L2	L1	L1	L1	L1	L1
Years	16	16	22	22	8	25

Table 10: Annotator background information. All have extensive linguistic annotation experience. Certified Teacher (CT), Private Tutor (PT), Linguistic Annotator (LA), Research Assistant (RA). **A0**^{P} is the preprocessing and segmentation lead; and **A5**^{L} is the readability annotation lead.

A.5 Inter-Annotator Agreement between Annotator Labels and Unified Labels

	Acc^{19}	$\pm 1\ \text{Acc}^{19}$	Dist	QWK	Acc^{7}	Acc^{5}	Acc^{3}
A1	78.4%	89.0%	0.42	93.4%	85.3%	87.0%	89.7%
A2	65.1%	76.4%	0.87	82.2%	71.6%	73.6%	79.3%
A3	66.4%	78.4%	0.78	86.0%	73.7%	75.8%	79.0%
A4	63.7%	76.6%	0.86	83.8%	71.8%	74.2%	79.5%
A5	85.1%	91.2%	0.31	94.8%	89.2%	90.3%	92.9%
Avg	71.7%	82.3%	0.65	88.1%	78.4%	80.2%	84.1%

Table 11: Inter-Annotator Agreement (IAA) results comparing initial annotations by A1-A5 to unified labels (UL).

B BAREC Corpus Level Distributions Across Splits

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-ḥa	5,683	8%	4,529	8%	613	8%	541	7%
9-ṭa	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%

Table 12: Distribution of sentence counts and percentages across readability levels and data splits.

C BAREC Corpus Sources

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

Some datasets are chosen because they already have annotations available for other tasks. We list them independently of other collections they may be part of. For example, dependency treebank annotations exist (Habash et al., 2022) for the texts we included from the Arabian Nights, Quran and Hadith, Old and New Testament, Suspended Odes Odes, and Sara (which comes from Hindawi Foundation).

C.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.[6] Not all sentences generated by ChatGPT were correct. We discarded some sentences that were flagged by the annotators. Table 13 shows the prompts and the percentage of discarded sentences for each prompt.

[6] https://chatgpt.com/

C.2 Literature

Hindawi A subset of 264 books extracted from the Hindawi Foundation website across different different genres.[7]

Kalima The first 500 words of 62 books from Kalima project.[8]

Green Library 58 manually typed books from the Green Library.[9]

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.[10]

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.[11]

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.[12]

The Suspended Odes (Odes) The ten most celebrated poems from Pre-Islamic Arabia (المعلقات Mu'allaqat). All texts were extracted from Wikipedia.[13]

C.3 Media

Majed 10 manually typed editions of Majed magazine for children from 1983 to 2019.[14]

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).

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

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

C.4 References

Wikipedia A subset of 168 Arabic wikipedia articles covering Culture, Figures, Geography, History, Mathematics, Sciences, Society, Philosophy, Religions and Technologies.[15]

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.[16]

C.5 Religion

Old Testament The first 20 chapters of the Book of Genesis (Smith and Van Dyck, 1865).[17]

New Testament The first 16 chapters of the Book of Matthew (Smith and Van Dyck, 1860).[17]

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).[18]

Hadith The first 75 Hadiths from Sahih Bukhari (al Bukhari, 846). We selected the text from the LK Hadith Corpus[19] (Altammami et al., 2019).

[15] https://ar.wikipedia.org/

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

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

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

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