

# Guidelines for Fine-grained Sentence-level Arabic Readability Annotation

Nizar Habash,<sup>†</sup> Hanada Taha-Thomure,<sup>‡</sup> Khalid N. Elmadani,<sup>†</sup>  
Zeina Zeino,<sup>‡</sup> Abdallah Abushmaes<sup>††</sup>

<sup>†</sup>Computational Approaches to Modeling Language Lab, New York University Abu Dhabi

<sup>‡</sup>Zai Arabic Language Research Centre, Zayed University

<sup>††</sup>Abu Dhabi Arabic Language Centre

nizar.habash@nyu.edu, Hanada.Thomure@zu.ac.ae

## Abstract

This paper presents the foundational framework and initial findings of the Balanced Arabic Readability Evaluation Corpus (**BAREC**) project,<sup>1</sup> designed to address the need for comprehensive Arabic language resources aligned with diverse readability levels. Inspired by the **Taha/Arabi21** readability reference (Taha-Thomure, 2017), **BAREC** aims to provide a standardized reference for assessing sentence-level Arabic text readability across 19 distinct levels, ranging in targets from kindergarten to postgraduate comprehension. Our ultimate goal with **BAREC** is to create a comprehensive and balanced corpus that represents a wide range of genres, topics, and regional variations through a multifaceted approach combining manual annotation with AI-driven tools. This paper focuses on our meticulous annotation guidelines, demonstrated through the analysis of 10,631 sentences/phrases (113,651 words). The average pairwise inter-annotator agreement, measured by Quadratic Weighted Kappa, is 79.9%, reflecting a high level of substantial agreement. We also report competitive results for benchmarking automatic readability assessment. We will make the **BAREC** corpus and guidelines openly accessible to support Arabic language research and education.

## 1 Introduction

Readability, the measure of how easily a reader can understand a written text, is essential for effective communication across diverse audiences. It is closely associated with text leveling, which categorizes texts into readability levels based on factors like orthography, morphology, syntax, and vocabulary complexity. Developing readability models is vital for improving literacy rates, aiding language learning, and enhancing academic achievement. However, in Arabic language education and

research, there is a significant lack of standardized resources for assessing text readability across various proficiency levels. This challenge is compounded by Arabic’s intricate linguistic features, such as rich morphology and lexicon, and its highly ambiguous orthography.

The work presented in this paper is part of a larger project – the Balanced Arabic Readability Evaluation Corpus (**BAREC**) – whose goal is to develop resources and tools for fine-grained readability assessment 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<sup>2</sup> Arabic text readability across 19 distinct levels, ranging from kindergarten to postgraduate comprehension.

Our contributions are as follows: (a) we define **detailed guidelines** for fine-grained sentence-level readability annotation across 19 levels; (b) we **curate and annotate a unique corpus** with a diverse mix of genres comprising 10,631 segments (113,651 words); and (c) we use the corpus to build **automatic readability assessment** models and benchmark them.

## 2 Related Work

### 2.1 Readability and Leveling

**Definitions** Readability correlates with understanding, retention, reading speed, and engagement (DuBay, 2004). Students given texts above their readability level may become unmotivated and disengaged. Klare (1963) defined readability as the ease of understanding a text, while Nassiri et al. (2023) noted that readability and legibility depend

<sup>1</sup>بارق *bAriq* is Arabic for ‘very bright and glittering’.

<sup>2</sup>We segment paragraphs down to syntactic sentences. However, we use the term *sentence* even for small standalone text segments such as phrases and single words (e.g. book titles).

on both external features (e.g., production, fonts, look and feel) and content-related features. In classrooms, text leveling helps match books to students' reading levels, fostering independent reading and comprehension (Allington et al., 2015).

**Readability Granularity** We distinguish two orthogonal aspects of readability *granularity*: **text granularity** and **level granularity**. Text granularity refers to the text unit size: a book, a chapter, a document, a paragraph, a sentence, a phrase or a word. Level granularity refers to the readability leveling scheme's degree of detail, e.g. Al-Khalifa and Al-Ajlan (2010) used a 3-level scale, the Common European Framework of Reference for Languages (CEFR) has 6 (Council of Europe, 2001), Fountas and Pinnell (2006)'s system has 27 levels from A to Z+ (Kindergarten to Highschool/Adult), while Taha-Thomure (2017)'s system has 19.

## 2.2 Arabic Readability Efforts

**Taha/Arabi21** Taha-Thomure (2017) presented an Arabic text leveling system that is inspired by Fountas and Pinnell (2006) and framed for the field of Arabic education. Her target text granularity is a book, and her level granularity is 19 levels, with special focus on the introductory levels (e.g., 11 of the 19 are up to around 4<sup>th</sup> grade). Taha-Thomure (2017)'s procedural framework employs ten qualitative and quantitative criteria to help school teachers level children's literature they use and match the right book level with each student's readability level. The criteria are as follows: text genre, abstract ideas used in the text, choice of vocabulary and its distance from dialects, text authenticity, book production, content, sentence structure, illustrations, use of diacritics, and number of words. This was a departure from the earliest text-leveling efforts that looked at the number of words in a sentence and the number of syllables in each word. These leveling criteria have been adopted by the Arab Thought Foundation (ATF), under the project Arabi21 which funded the leveling of 9,000 children's literature titles.

**Arabic CEFR** A number of efforts targeted the use of CEFR leveling for Arabic texts at different text granularities. The KELLY project (Kilgariff et al., 2014) developed monolingual and bilingual word lists for language learning. This project aims to map the most common 9,000 words in nine languages (including Arabic) onto CEFR levels through corpus-based frequency analysis and com-

parisons between translated language pairs across the said nine languages. Habash and Palfreyman (2022) manually annotated short essays written in Arabic and in English in CEFR. Abo Amsha et al. (2022) presented a detailed reference on Arabic CEFR leveling for non-native speakers. Naous et al. (2023) created a manually annotated CEFR-leveled dataset in five languages, including Arabic. Soliman and Familiar (2024) created an Arabic vocabulary profile suitable for CEFR Levels A1 and A2. They constructed it by prioritizing words based on their prevalence across multiple dialects, frequency of use, and linguistic complexity.

**SAMER** As part of the *Simplification of Arabic Masterpieces for Extensive Reading* (SAMER) project, Al Khalil et al. (2020) developed a 26K-lemma lexicon with a five-level readability scale, later extended to 40K lemmas (Jiang et al., 2020). The levels range from **L1** (Low Difficulty/Easy Readability) to **L5** (High Difficulty/Hard Readability). They relied on three annotators from different Arab countries to provide levels for each entry in their lexicon. The project further led to the creation of the SAMER Corpus, the first manually annotated Arabic parallel corpus for text simplification targeting school-aged learners (Alhafni et al., 2024). The corpus comprised 159K words from Arabic novels (**L5**) and was mapped to two lower levels (**L4**, **L3**).

**Automatic Readability Measurement** While our focus is on manual annotation of readability, we are inspired by ideas, techniques, and insights from previous efforts on automatic methods for readability measurement. Al-Dawsari (2004) described an Arabic readability formula that includes five features: average word length, average sentence length, word frequency, percentage of nominal clauses, and percentage of definite nouns. Al-Khalifa and Al-Ajlan (2010) targeted three readability levels: easy, medium, and difficult on manually collected data from the reading books of the elementary, intermediate, and secondary Saudi curriculum. They selected a number of text features such as the average number of syllables per word, word frequencies, and n-gram language model perplexity scores. Forsyth (2014) used a machine learning approach to process the online curriculum of the Defense Language Institute Foreign Language Center and concluded that most (19 out of 20) of the best features are from the POS-based frequency feature set. Al Tamimi et al. (2014)

presented AARI, an automatic readability index for Arabic which extracted seven features to calculate readability, including the number of characters, words, sentences and difficult words. They evaluated their work on Arabic texts from different grades in the Jordanian curriculum. [El-Haj and Rayson \(2016\)](#)’s OSMAN readability metric makes use of script markers of MSA, and counts the number of syllables through automatic diacritization. [Saddiki et al. \(2018\)](#) use a rich set of raw, syntactic, and morphological readability features to build feature vectors that represent documents. They use these representations to train a classifier that accurately predicts the readability level of documents in a four-level scale. Most recently, [Liberto et al. \(2024\)](#) explored Arabic readability assessment using rule-based methods and pretrained models, achieving 87.9% macro F1 score at the fragment level (**L5-L4-L3**) on the SAMER Corpus ([Alhafni et al., 2024](#)).

**Our Approach** Inspired by [Taha-Thomure \(2017\)](#), we extend their approach to the sentence/phrase level to offer greater control over text content and a more objective measure of variance across larger texts. Our guidelines incorporate relevant ideas from other efforts, focusing solely on readability features, and excluding aspects like legibility or book design.

### 3 Readability Annotation Desiderata

We outline below the key principles for the **BAREC** project guidelines:

**Comprehensive Coverage** Annotation guidelines will span a wide range of readability levels, from kindergarten (Easy) to postgraduate (Hard), with finer distinctions at lower levels.

**Objective Standardization** Standardized guidelines will minimize subjectivity, covering 19 readability levels based on factors like dialect, syntax, morphology, semantics, and content, avoiding oversimplifications like word or sentence length.

**Bias Mitigation** Guidelines will reflect the diversity of the Arab world’s religions, ethnicities, and dialects, ensuring inclusivity and considering regional variations, especially in easier levels.

**Balanced Coverage** Data annotation will try to balance readability levels, genres, and topics, acknowledging the scarcity of certain texts, like children’s books, and their inherent shorter length.

**Enriching Annotations** Texts with existing annotations (e.g., part-of-speech tagging, named-entity recognition) will be prioritized to support exploring readability in relation to other linguistic features in the future.

**Quality Control** Trained annotators will ensure high inter-annotator agreement, with additional consistency checks for methodology robustness.

**Open Accessibility** The **BAREC** corpus and guidelines will be openly available to support Arabic language research and education.

**Ethical Considerations** Annotation will respect fair-use copyright, and annotators will be fairly compensated, with measures in place to reduce task-related fatigue.

## 4 BAREC Guidelines

### 4.1 Readability Levels

We are inspired by [Taha-Thomure \(2017\)](#)’s naming convention of readability levels which use the Abjad order of Arabic letters.<sup>3</sup> We will refer to the **BAREC** readability level as **c+letter number-letter name**, giving us the following 19 levels: **c1-alif, c2-ba, c3-jim, c4-dal, c5-ha, c6-waw, c7-zay, c8-ha, c9-ta, c10-ya, c20-kaf, c30-lam, c40-mim, c50-nun, c60-sin, c70-ayn, c80-fa, c90-sad, and c100-qaf**. The higher increments pay homage to this traditional way of *letter counting*, but also signify that the levels are not equally spaced, with a lot more finer distinction in the early easier readability levels. Figure 1 illustrates the scaffolding relationship across the levels and their approximate mapping to another readability resource (SAMER) and education school grade levels. The **BAREC Pyramid** also highlights the different levels of involvements of various linguistic dimensions we use in the guidelines. Table 1 presents representative examples for each level.

### 4.2 Readability Annotation Principles

**Reading & Comprehension** The readability level of a specific sentence or phrase, henceforth

<sup>3</sup>The Abjad order lists the Arabic letters typically as أبجد هوز حطي كلمن سعفص قرشت ثخذ ضظغ *Ābjd hwz HTy klmn scfS qršt θxð DḌγ* – HSB Romanization ([Habash et al., 2007](#)). The order is connected with numerical counts starting from 1 to 10, followed by increments of 10 up to 100, and further increments of 100 up to 1,000.





RL	Arabic Sentence/Phrase	Translation	Level Reasoning
c1-alif	أُرْنَبْ	<b>Rabbit</b>	One word - two syllables - familiar noun
c2-ba	ملعب واسع	<b>A large playground</b>	Noun-adjective
c3-jim	أنا أحب اللون الأحمر.	I love <b>the</b> color red.	Definite article
c4-dal	الشمس تشرق في الصباح الباكر.	The sun rises early <b>in the morning</b> .	Prepositional phrase
c5-ha	القط تستريح على السرير وتستمتع بأشعة الشمس الدافئة.	The cat rests on the bed <b>and enjoys the warm sunshine</b> .	A conjoined sentence
c6-waw	سلوكي مسؤوليتي	My behavior is <b>my responsibility</b>	Five syllable word
c7-zay	الأصدقاء يحتفلون بعيد ميلاد صديقهم بكعكة وهدايا رائعة.	<b>Friends</b> celebrate their friend's birthday with cake and amazing gifts.	Broken plural
c8-ha	أستمع إلى كل فقرة من الفقرتين الآتيتين، ثم أجيب:	I listen to each of the following two paragraphs, <b>then</b> I answer:	Then: in level c8-ha ح
c9-ta	وقال بكلام فصيح مزعج: يا سمك يا سمك هل أنت على العهد القديم مقيم	He said in annoying, eloquent words: <b>Oh fish, oh fish</b> , do you abide by the old promise	Noun in the vocative case
c10-ya	وسألتك هل كنت تتهمونه بالكذب قبل أن يقول ما قال فذكرت أن لا.	I asked you whether <b>you were</b> accusing him of lying before he said what he said, and you said no.	Auxiliary Kaana
c20-kaf	حسام سعيد قلبه بسبب فوز فريقه.	Hossam, his <b>heart is happy</b> because of his team's victory.	Acting derivative (happy is predicative)
c30-lam	لا أحد يجمع هذه الزهور معاً في باقة، فهي منتشرة جداً — حتى إنه كان من المعروف عنها أنها تنمو بين أحجار الرصف، وتنبثق في كل مكان مثل الحشائش الضارة — وتحمل اسماً قبيحاً جداً وهو «زهور الكلاب» أو «الهندباء البرية».	No one puts these flowers together in a bouquet, they are so common— <b>they have even been known to grow between paving stones, and spring up everywhere like weeds</b> —and they have the very unsightly name of “dog-flowers” or “dandelions.”	Parenthetical phrase
c40-mim	ومن يفعل المعروف مع غير أهله يجازى كما جازى مجير أم عامر	<b>And whoever offers good deeds to someone undeserving will be rewarded like he who gave shelter to a hyena</b>	Conditional phrase
c50-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 <b>charged particles</b> indicates the spacecraft's departure from the influence of the <b>solar wind</b> , which is called <b>the heliosphere</b> (which, according to some definitions, is the border of the <b>solar system</b> ).	General geography vocabulary
c60-sin	وكان من عاداتها أن تقارن بينها وبين بطلة الرواية إذا أحسنت منه إعجاباً بها أو ثناء عليها، وتساله في ذلك أسئلة ذكية خبيثة 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 <b>that did not allow answering deceptively</b> , except by joking and teasing.	Specialized vocabulary that requires understanding the concept comprehend its use
c70-ayn	ويذهب المؤرخون إلى أن النابغة الذبياني كان من المخمخين، تقام له في هذه الأسواق قبة يذهب إليها الشعراء ليعرضوا شعرهم، فمن أشاد به ذاع صيته، وانتقلت شعره الركيان.	Historians assert that <b>Al-Nabigha Al-Dhubyani</b> was one of the <b>arbiters</b> . In these markets, a dome is erected for him where poets go to present their poetry. Whomever he praised, <b>his fame spread</b> , and his poetry circulated among the <b>caravans</b> .	Specialized and uncommon vocabulary
c80-fa	بين طعن القنا وخفق النود	Between the thrusts of <b>lances</b> and the fluttering of <b>ensigns</b>	Heritage vocabulary familiar to a novice specialist
c90-sad	ألا الأورى لأنا ما أبينها والنوى كالحوض بالمظلومة الجند	<b>I wasn't able to see except with extreme effort and difficulty like a water basin in solid undrillable land</b>	Specialist vocabulary, symbolic poetic ideas that require prior knowledge
c100-qaf	كان دوح المالكية غداة خلايا سفين بالتواصف من دد	As if <b>the camel saddles of the Malikyya caravan leaving the Dadi valley were great ships</b>	Advanced specialist vocabulary, symbolic poetic ideas that require prior knowledge

Table 1: 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.

or ‘this is a governmental authority without choices’ (harder readability).

We note that the decision to disregard diacritics is in departure from [Taha-Thomure \(2017\)](#) who values the use diacritics as a strong design feature of books intended for young readers. In a way, we consider adding them as a bookmaking design choice that complements and supports the chosen readability level.

### 4.3 Dimensions of Textual Features

To determine the **BAREC** level, we identified six dimensions of textual features, each specifying the necessary features (keys) for each level. Appendix A includes a *summary cheat sheet* of these guidelines in Arabic (as used by the annotators), along with an English translation. The full guidelines will be made publicly available.

**1. Number of Words** We count unique words separated by white space and punctuation, ignoring diacritization and meaning differences for words with the same spelling in the same text. For example, in Table 1(c3-jim), the text has 4 words. The maximum number of words is only used as a determining features for levels **c1-alif** (1 word) to **c20-kaf** (20 words).

**2. Orthography & Phonology** This dimension focuses on the difficulty of transferring from written to spoken form, especially regarding word length (syllable count), and the presence of certain letters (such as Hamzas and weak letters). Final diacritics are ignored in syllable counting, treating words as if they end in *waqf* (silent ending). For example the word in Table 1(c1-alif), أُرْنَبْ *Ār-nabū* ‘rabbit’ has a syllable count of 2 (*ar-nab*).

**3. Morphology: Inflection and Derivation** Arabic is a morphologically rich and complex language with templatic and concatenative morphological operations in productive use. This dimension focuses on leveling the various word morphology features from derivation (the root and pattern that determine the basic meaning) to inflection (the prefixes and suffixes added to the word to specify its meaning), as well as the relationship between them and linguistic features such as gender, number, person, tense, voice, etc. Examples of ordering decisions include introducing simple present tense verbs (**c1-alif**) before past tense (**c6-waw**), the singular (**c1-alif**) before the plural (**c4-dal**), and that before the dual (**c7-zay**), and delaying the introduction of passive voice, diminutive and energetic mood to higher levels – **c10-ya**, **c30-lam**, **c40-mim**, respectively. This dimension is used to distinguish up to level **c40-mim**.

**4. Syntactic Structures** This dimension focuses on the structure of the sentence, i.e., the syntactic relationship between words. Examples of ordering decisions include starting with single words (**c1-alif**), then introducing simple pairs of nominal sentences, noun-adjective and nonun-noun idafas (**c2-ba**). Temporal modifiers are introduced in **c7-zay**, vocatives in **c9-ta**, and conditional sentences in **c40-mim**. This dimension is used to distinguish up to level **c60-sin**, where we relegate ambiguous highly infrequent constructions that need diacritization to resolve.

**5. Vocabulary** This dimension focuses on the choice of words used in the sentences/phrases under evaluation. It is used with all levels and is especially important in higher levels. This dimension intersects with other dimensions that filter some of its options, e.g., the part-of-speech, spelling, and inflection limit some of the possible words at lower levels. Given Arabic’s evolving nature, we consider linguistically Arabized foreign words as part of the language and assess their readability accordingly. Words in non-Arabic scripts are excluded from classification. Examples of ordering decisions include introducing MSA vocabulary items that exactly match dialectal vocabulary before those that are similar but have predictable phonological differences. The guidelines occasionally reference SAMER levels (Al Khalil et al., 2018) as a rough guide. The harder levels introduce increasingly technical vocabulary in arts and sciences.

**6. Ideas & Content** This dimension focuses on organizing the levels of text in terms of three inter-related aspects: (i) what **prior knowledge** is necessary for comprehension? (nothing  $\ll$  Reader’s life  $\ll$  General knowledge  $\ll$  Other cultures’ knowledge  $\ll$  Specialized knowledge); (ii) what minimal degree of **symbolic unpacking** is necessary for direct understanding of the text? (no symbolism  $\ll$  some symbolism (one or two ideas)  $\ll$  a lot of symbolism and abstraction); and finally (c) what degree of prior knowledge **linking** and additional **analysis** are needed for direct understanding? (no need  $\ll$  link without analysis  $\ll$  link with analysis). At higher levels, we differentiate between general knowledge terms (arts and sciences for the general public) and specialized knowledge terms (language of specialists). We recognize that evaluating these aspects can be complex and subject to interpretation, and may vary among readers even within the same age or education level group.

**Problems and Difficulties** The annotators are encouraged to indicate any text problems or difficulties they encounter. Reportable problems include spelling errors (e.g., in Hamza or Ta-Marbuta), colloquial language, ungrammatical constructions, and inappropriate topics (racism, bullying, pornography, etc.). Difficulty is reported in case where it is not possible to make a decision because of conflicting considerations or guideline gaps.

## 5 BAREC Corpus Annotation

### 5.1 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.

### 5.2 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. (2022). Subsequently, A1-5 were tasked with assigning readability labels to the individually 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). Shared annotation sets were included covertly to ensure quality and measure inter-annotator agreement.

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. Finally, we conducted a thorough second review of the corpus data, resulting in every sentence being checked twice.

### 5.3 BAREC Dataset

We curated the **BAREC** dataset to include diverse genres and topics, resulting in 274 documents, categorized into four intended readership groups: **Children**, **Young Adults**, **Adult Modern Arabic**, and **Adult Classical Arabic**. The distribution of data for each group is shown in Table 2. We aimed to balance the total word count across these groups. As a result, children’s documents have more sentences due to the typically shorter sentence length in that genre. On average the length of sentences in the **Children** group is 7.0 words, whereas it is 13.7 for **Adult Classical Arabic**. On average we selected 419 words/document, although there is a lot of variation among *documents*, which range from complete books to chapters, sections, or ad hoc groupings. All selected texts are either out of copyright, or are within fair-use representative sample sizes. We collected data from various sources, including educational curriculum, books, Wikipedia, manually verified ChatGPT texts, children’s poems, UN documents, movie subtitles, classical and religious texts, literary works, and news articles. All details are available in Appendix B.

## 6 Results

### 6.1 Inter-Annotator Agreement

We conducted four inter-annotator agreement (IAA) studies: three 100-sentence pilots during *training* to enhance agreement, and a final official study using 200 sentences, which we report on next. The average pairwise exact-match over 19 **BAREC** levels between any two annotators is only 49.2%, which reflects the task’s complexity. Allowing a fuzzy match distance of up to 1, 2, 3, or 4 levels raises the match to 64.6%, 77.1%, 87.2%, and 93.2%, respectively. The overall average pairwise level difference is 1.38 levels. The average pair-

Group	#Docs	#Sents	#Words
Children	30	4,363	30,502
Young Adults	42	2,307	29,465
Adult Modern Arabic	74	1,952	26,108
Adult Classical Arabic	128	2,009	27,576
Total	274	10,631	113,651

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

wise Quadratic Weighted Kappa 79.9% (substantial agreement) confirms most disagreements are minor (Cohen, 1968; Doewes et al., 2023).

**Second Round QC** After the above-mentioned IAA, we made some minor guideline clarifications and did some continued training. Then we conducted a **second round of full annotation quality check** where every example was checked by a different annotator from the first round. In total 40% of the labels changed with an average level distance of 0.97; the average pairwise Quadratic Weighted Kappa between the two rounds is 85.5%.

### 6.2 Analysis of Annotation Distributions

**Flagged Segments** The actual number of annotated segments is 10,896; but 2.3% were excluded for flagged problems, and 0.13% excluded for flagged difficulties.

#### Readership Groups and Readability Levels

Figure 2 visualizes the annotation distributions across the four readership groups identified based on educated guesses and self-declared target readers. Full details are in Appendix D. Children’s texts dominate the easier levels (**c1-alif** to **c8-ha**), while Classical texts dominate the harder levels (**c90-sad** and **c100-qaf**), as expected. The middle levels contain a mix of all groups. Interestingly, some Children texts include advanced materials, which may need revision, or can be arguably justified for educational purposes.

**Readability Level Patterns** In terms of total counts, Figure 2 exhibits a slightly skewed distribution, notably with lower counts for **c9-ta** and higher counts for **c50-nun**. This pattern could stem from the limited sample size or potential biases in text selections. Notably, the guidelines for **c9-ta** feature specific uncommon linguistic elements like the dual command form, vocative, emotional vocabulary, and the Hamza interrogative particle.

**Readability Level and Text Length** Figure 3 presents two charts comparing readability levels with segment lengths. The overall averages show

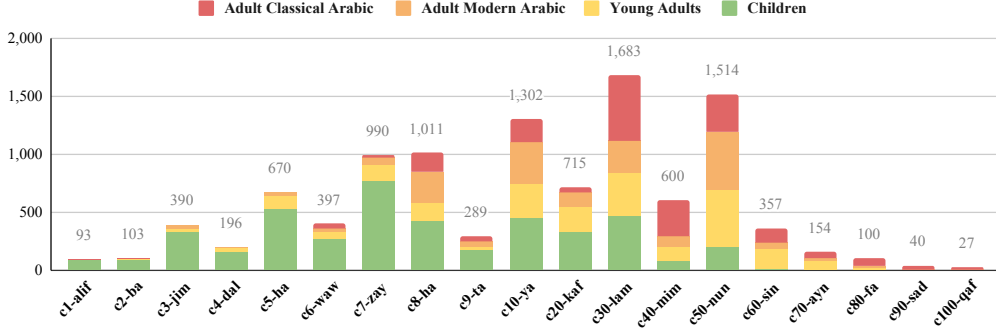


Figure 2: The distribution of annotated sentences among **BAREC** levels and Arabic readers groups

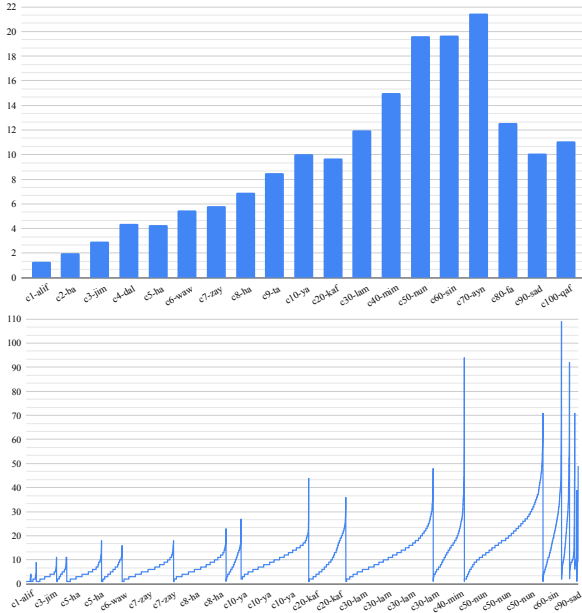


Figure 3: Charts comparing the average sentence length (left) and the distribution of lengths (right) per level

a generally expected linear pattern from **c1-alif** to **c10-ya/c20-kaf**, continuing to **c70-ayn** before dropping off, as higher-level texts, often poetry, are shorter than prose. The length distribution chart, in Figure 3(right), highlights variability within each readability level, confirming that annotators did not strictly use segment lengths for readability level annotation.

### 6.3 Automatic Readability Assessment

We train sentence-level classifiers by finetuning CAMELBERT-MIX (Inoue et al., 2021), MARBERT (Abdul-Mageed et al., 2021) and AraBERTv2 (Antoun et al., 2020) to benchmark the baseline performance given the dataset. We split the dataset into 90% for training and 10% for testing. We finetune the models using the Transformers library (Wolf et al., 2019) on a NVIDIA T4

Metric	CAMELBERT	MARBERT	AraBERT
Accuracy @1	58%	56%	57%
Accuracy @2	73%	72%	73%
Accuracy @3	83%	82%	82%
CL Rank	2.23	2.31	2.24
CL Distance	1.06	1.10	1.07
QWK	84%	84%	84%

Table 3: Results of automatic readability assessment comparing CAMELBERT-MIX, MARBERT, and AraBERTv2. CL Rank is the average rank of the correct label; CL Distance is the average distance from the correct label; and QWK is the Quadratic Weighted Kappa.

GPU for three epochs with a learning rate of  $5e-5$ , and a batch size of 16. Table 3 shows the results of finetuning the three models for readability prediction as a text classification task. We report with the following metrics: **Accuracy@n** (correct label is within the top  $n$  predictions), **Average Rank of the Correct Label**, **Average Distance from Correct Label**, and **Quadratic Weighted Kappa**. The performance of the compared systems is generally similar. Their results are comparable with the IAA numbers, showing a robust Quadratic Weighted Kappa score of 84%. We anticipate that performance will improve further with additional data.

## 7 Conclusions and Future Work

We introduced the **BAREC** project addressing the need for comprehensive Arabic language resources across various readability levels. We developed detailed guidelines, trained annotators, and labeled 10,000+ sentences. The guidelines and corpus will be publicly available. We also demonstrated the application of the corpus in automatic leveling, achieving promising results. Future work will expand the corpus’s size and diversity, refine the guidelines to address sources of disagreement, and enhance automatic readability models.



## 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.
- Khaled Hassan Abo Amsha, Raed Abdul Rahim, Hidayat Sheikh Ali, Al-Sayed Ezzat Abo Al-Wafa, Mohamed Ismail Aloui, Belkacem Al-Youbi, Mohamed Al-Jarrah, Saleh Al-Hajouri, Mohamed Mullaq, Bashir Al-Obaidi, Ibrahim Al-Shafie, Mohamed Haqqi Sawatchen, Riwaya Mohamed Jamous, and Suhad Nairat. 2022. *Applications of the Common European Framework of Reference (CEFR) in Teaching Arabic to Non-Native Speakers. Part One*. Dar Kunooz Al-Ma'arif for Publishing and Distribution, Amman, Jordan.
- 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.
- Muhammed 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.
- Muhammed 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).
- Abdel Karim Al Tamimi, Manar Jaradat, Nuha Al-Jarrah, and Sahar Ghanem. 2014. AARI: automatic Arabic readability index. *International Arab Journal of Information Technology*, 11(4):370–378.
- Bashar Alhafni, Reem Hazim, Juan David Pineros Liberato, Muhammed 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.
- Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213.
- Council of Europe Council of Europe. 2001. *Common European Framework of Reference for Languages: learning, teaching, assessment*. Cambridge University Press.

- Afrizal Doewes, Nughthoh Arfawi Kurdhi, and Akirati 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.
- 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).
- Salman Elgamal, Ossama Obeid, Mhd Kabbani, Go Inoue, and Nizar Habash. 2024. [Arabic diacritics in the wild: Exploiting opportunities for improved diacritization](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14815–14829, Bangkok, Thailand.
- 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. ZAEUBC: An annotated Arabic-English bilingual writer corpus. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Marseille, France.
- 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.
- 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.
- Zhengyang Jiang, Nizar Habash, and Muhamed Al Khalil. 2020. [An online readability leveled Arabic thesaurus](#). In *Proceedings of the 28th International Conference on Computational Linguistics: System Demonstrations*, pages 59–63, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).
- 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 Kilgariff, 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.
- 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.
- 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.
- Tarek Naous, Michael J. Ryan, Anton Lavrouk, Mohit Chandra, and Wei Xu. 2023. [ReadMe++: Benchmarking Multilingual Language Models for Multi-Domain Readability Assessment](#). *arXiv preprint. ArXiv:2305.14463 [cs]*.
- 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).
- 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).

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

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

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

# A BAREC Annotation Guidelines Cheat Sheet

## A.1 Arabic Original

مستوى بارقي	صف	ACTFL	عدد كلمات	تهجئة وإملاء	تصريف واشتقاق	تركييب 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 تركيبي) • ضمير منفصل • مفردات متطابقة مع العامية - سامر I • الأرقام (العربية أو الهندية) 1-10	• فكرة مباشرة وصريحة وحسية. • لا رمزية في النص.
ب		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	1	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 فضيحة شائعة - سامر I • اسم الإشارة المفرد • الأرقام (العربية أو الهندية) 10-100	
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	
هـ		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 استقها م أساسية: ماذا، متى، من، أين، ما، كيف • صيغة التعجب "ما أفعل"	• العدد الترتيبي • الأرقام (العربية أو الهندية) 101-1,000 • اسم إشارة مثنى، جمع	• المحتوى من حياة القارئ. • لا رمزية في النص.
و	2	مبتدئ أعلى	≤9	• كلمات من 5 مقاطع	• الفعل الماضي المفرد والجمع • جمع مذكر سالم	• جملة فيها فعلين (مثلا جملة فعلية مفعولها أن المصدرية)	• مفردات فضيحة - سامر I	
ز		متوسط أدنى	≤10	• كلمات من 6+ مقاطع • أفعال/أسماء معثلة الآخر • جمع التكسير • واو القسم (والله)	• 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 الاستقها م هل	• مفردات فضيحة 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، لكن، أما)	• فعل الأمر الجمع • نون النسوة في الأسماء والأفعال • سوابق أخرى: سين الاستقبال، واو الاستئناف، فاء العطف • أدوات ربط (ثم، حتى، أو، أم، لكن، أما)	• المفعول المطلق • المفعول لأجله • المفعول معه • جملة فعلية تتعدى إلى مفعولين	• مفردات فضيحة - سامر I و سامر II • أحرف التني • الأرقام (العربية أو الهندية) 1,001-1,000,000	• بعض الرمزية يحتاج معها القارئ إلى مساعدة من يدرج له المقصود من الفكرة
ط		متوسط متوسط	≤12	• فعل الأمر للمثنى • أداة الاستقها م: أ (أسمعت؟) • ياء القسم • القسم: أداة القسم والمقسم به وجواب القسم.	• فعل الأمر للمثنى • أداة الاستقها م: أ (أسمعت؟) • ياء 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 يتركها القارئ بنفسه أو من خلال معارفه السابقة
ي	4	متوسط متوسط	≤15	• المبنى للمجهول	• إن وأخواتها • كان وأخواتها • خبر مقدم / مبتدأ مؤخر • العطف/السند • زب (حرف جر شبه بالرائد) • جملة الصلة وجملة الصفة • جملة الحال وجملة المفعول به	• إن وأخواتها • كان وأخواتها • خبر مقدم / مبتدأ مؤخر • العطف/السند • زب (حرف جر شبه بالرائد) • جملة الصلة وجملة الصفة • جملة الحال وجملة المفعول به	• أسماء الوصل المفردة • (قد - لقد) • (مما - عما - عا - علام - فيم - لام - بم...)	
ك		متوسط أعلى	≤20	• المشتقات العاملة (مثلا اسم الفاعل)	• المشتقات العاملة (مثلا اسم الفاعل)	• جملة اسمية خبرها جملة اسمية • إضافة لفظية (طويل القامة)	• أسماء الوصل المثنى والجمع	• هناك درجة من الرمزية وحاجة للمعرفة السابقة كي يفهم المقصود من الجملة
ل	5	متقدم أدنى		• التصغير	• جعل اعراضية (تفسير، دعاء) • استثناء • محصر • بدل (بدل جملة بعض أو اشتمال) • ممييز	• جعل اعراضية (تفسير، دعاء) • استثناء • محصر • بدل (بدل جملة بعض أو اشتمال) • ممييز	• مفردات فضيحة - سامر III • اسم الفعل (مثلا أمين) • الأرقام (العربية أو الهندية) < 1,000,000 • ذو • (يل - بلى - أجل - قط)	
م	6-7	متقدم أوسط		• نون التوكيد • تاء القسم	• نون التوكيد • تاء القسم	• الجمل شرطية (مركية - عالية) • حرف الجزم لما	• كلمات تصف حالات نفسية صعبة مثل الاكتئاب، الضيق، الاستنفار النفسي • استخدام كلمات محدوتة غير متداولة (مثلا هجرع للخييف الأحمق مشتقة من هرع و هجع) • الرموز (ش.م.)	• أفكار رمزية ومعنى باطن خاصة على صعيد البعد النفسي للشخصيات أو الأحداث.
ن	8-9	متقدم أعلى				• التوكيد المعنوي • المدح والذم • جملة أن المصدرية في محل رفع مبتدأ • صيغة التعجب "أفعل به من"	• مفردات فضيحة - سامر IV • مفردات قانونية، علمية، دينية، سياسية... غير متخصصة/عامة • فو - حمو	• أفكار رمزية ومعنى باطن خاصة على صعيد البعد النفسي للشخصيات أو الأحداث.
س	10-11	متقن أدنى				• تراكييب غير متداولة فيها التنباس 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؟) • مفردات فضيحة - سامر V • مفردات متخصصة ومفردات عربية عالية غير شائعة كثيرا في الفضاء العام. • مفردات في الغالب بعيدة عن اللهجات العامية.	• أفكار رمزية، مجردة، علمية، أو شعرية وتحتاج إلى معارف لغوية ومعرفية سابقة للبناء عليها لأجل فهمها
ع	12	متقن أوسط					• مفردات علمية وترائية غير متداولة اليوم وغير مالوفة • غير المتخصص المبتدئ	
ف	جامعة 2-1	متقن أعلى					• مفردات علمية وترائية غير متداولة اليوم وغير مالوفة • غير المتخصص	
ص	جامعة 4-3	متقن					• مفردات علمية وترائية غير متداولة اليوم وغير مالوفة • غير المتخصص	
ق	متخصص	متميز					• غير المتخصص الباحث	
هناك صعوبة هذا الوسم يستخدم في حالة وجود صعوبة في تقييم المستوى، المفضل استخدام هذا الوسم حتى تتمكن ك فريق عمل أن تجد حلا (مثلا بتعديل المعايير أو إضافة تفاصيل شرحية لها) هناك مشكلة بصورة عامة، نستخدم هذا الوسم للجميل (الحوارية على: أخطاء إملائية (مثلا همزات، تاء مربوطة، ألف مقصورة/ياء) أخطاء في التشكيل رككة لغوية (أمية، عامية، ترجمة سينة من لغة أجنبية) مواضيع غير لائقة (عنصرية، حيادية، تنمرية، إباحية، إلخ) جمل وعبارات معظمها مكتوب بلغات غير العربية أو بغير الخط العربي								



## A.2 English Translation

BAREC Level	Grade	ACTFL	Word Count	Spelling/Pronunciation	Morphology	Syntax	Vocabulary	Idea / Content
X1-alif ا	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.
X2-ba ب		Novice Low	≤2	• Three-syllable words			• Verb • Adjective • Vocabulary similar to dialectal form - SAMER I • Spelled cardinal numbers • The five nouns: <i>Abu</i> (father), <i>Axw</i> (brother)	
X3-jim ج	1	Novice Mid	≤4		• Proclitic: Definite article <i>Al</i> + • Proclitic: Conjunction <i>wa</i> + • Enclitic: First Person Singular pronoun	• Apposition (full) • Demonstratives	• Common MSA vocabulary - SAMER I • Singular demonstrative pronoun • Numbers: 11-100	
X4-dal د		Novice Mid	≤6	• Words with an elongated Alif (e.g. / <i>ḏāṣif</i> /)	• Plural imperfective verb • Prepositional proclitics • Nunated adverbials	• Verbal sentence w/o direct object • Preposition and object	• Prepositions	
X5-ha ه		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.
X6-waw و	2	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 <i>Masdar 'an</i> [-to/that])	• MSA vocabulary - SAMER I	
X7-zay ز		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 <i>hal</i>	• High frequency MSA vocabulary - SAMER II	• Some symbolism, or not everything is stated directly in the sentence.
X8-ha ح		Intermediate Low	≤11		• Plural imperative verb • Feminine plural suffix ( <i>nun</i> ) in nouns and verbs • Other proclitics: future <i>sa</i> +, continuation <i>wa</i> +, conjunction <i>fa</i> + • 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.
X9-ta ط		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 oath	• 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.
X10-ya ي	4	Intermediate Mid	≤15		• Passive voice	• <i>fina</i> and its sisters (particles introducing a subject) • <i>Kana</i> and its sisters (past tense verbs) • Preposed predicate, postponed subject • Chain of narration • <i>rubba</i> preposition construction • Relative clauses • Circumstantial and object clauses	• Singular relative pronouns • Verbal particles <i>qad</i> and <i>laqad</i> • Preposition-Conjunctions: <i>mimma</i> , <i>fima</i> ...	
X20-kaf ك		Intermediate High	≤20		• Acting derivatives (e.g., the active participle) • False idafa (tall in stature)	• 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.
X30-lam ل	5	Advanced Low			• Diminutive form	• Parenthetical sentences (explanation, blessing) • Frozen Verbs (e.g., <i>Amiyn</i> Amen) • Numbers: > 1,000,000 • Five Nouns: Dhu (possession nominal) • Interjections: <i>hala</i> , <i>Ajal</i> , etc.	• MSA vocabulary - Samer III • Frozen Verbs (e.g., <i>Amiyn</i> Amen) • Numbers: > 1,000,000 • Five Nouns: Dhu (possession nominal) • Interjections: <i>hala</i> , <i>Ajal</i> , etc.	
X40-mim م	6-7	Advanced Mid			• Energetic mood (emphatic <i>nun</i> ) • Ta of oath	• Conditional sentences (compound - simple) • Jussive particle <i>lamma</i> (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.
X50-nun ن	8-9	Advanced High				• Semantic emphasis • Praise and dispraise • <i>Masdar 'an</i> clause as a subject • Exclamatory form: <comparative adjective> <i>biḥ min</i>	• MSA vocabulary - SAMER IV • General legal, scientific, religious, political vocabulary, etc. • Five Nouns: <i>fiv</i> , <i>Hmw</i>	• Local cultural expressions that may not be understood by those outside the culture.
X60-sin س	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., <i>fatim</i> for <i>fatima</i> )	• Symbolic, abstract, scientific, or poetic ideas that require prior linguistic and cognitive knowledge to understand.
X70-ayn ع	12	Superior Mid					• MSA vocabulary - SAMER V • Specialized and highly elevated Arabic vocabulary not commonly used in public discourse. • Vocabulary mostly distant from dialects.	
X80-fa ف	University Year 1-2	Superior High					• Scientific and heritage vocabulary not in use today, but familiar to a novice specialist	
X90-sad ص	University Year 3-4	Distinguished					• Scientific and heritage vocabulary not in use today, but familiar to a specialist	
X100-qaf ق	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 >> (ـ) • Dotted Yaa missing at the end of the word >> (ي)		

## B BAREC Dataset Details

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

**Hindawi** A subset of 8 books from Hindawi classified as children stories,<sup>4</sup> and Ahmed Shawqi’s collection of poems for Children.<sup>5</sup>

**Wikipedia** A subset of 20 Arabic wikipedia articles covering Culture, Figures, Geography, History, Mathematics, Sciences, Society, Philosophy, Religions and Technologies.<sup>6</sup>

**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.<sup>7</sup> Not all sentences generated by ChatGPT were correct. We discarded some sentences that were flagged by the annotators. Appendix C shows the prompts and the percentage of discarded sentences for each prompt.

**Collection of Children poems (Other)** Example of the included poems: My language sings (لغتي تغني), Poetry and news (أشعار وأخبار), and The cat and the Eid’s hat (القطعة وقبة العيد) (Al-Safadi, 2005; Taha-Thomure, 2007).

**UN** The Arabic translation of the Universal Declaration of Human Rights.<sup>8</sup>

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

**The Suspended Odes (Odes)** The first ten verses of the ten most celebrated poems from Pre-Islamic Arabia (المعلقات Mu’allaqat). All texts were extracted from Wikipedia.<sup>9</sup>

**Quran** The first Surah, the last 14 Surahs, the first 106 verses from the second Surah and the first 108 verses from the third Surah from the Holy

Quran. We selected the text from the Quran Corpus Project (Dukes et al., 2013).<sup>10</sup>

**Hadith** The first 47 Hadiths from Sahih Bukhari (al Bukhari, 846). We selected the text from the LK Hadith Corpus<sup>11</sup> (Altammami et al., 2019).

**One Thousand and One Nights (1001)** 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.<sup>12</sup>

**Hayy ibn Yaqdhan (Hayy)** A subset of the philosophical novel and allegorical tale written by Ibn Tufail (Tufail, 1150). We extracted the text from the Hindawi Foundation website.<sup>13</sup>

**Old Testament (OT)** The first 225 words from each of the first 20 chapters of the Book of Genesis (Smith and Van Dyck, 1865).<sup>14</sup>

**New Testament (NT)** The first 280 words from each of the first 16 chapters of the Book of Matthew (Smith and Van Dyck, 1860).<sup>14</sup>

**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.<sup>15</sup>

**WikiNews** 70 Arabic WikiNews articles covering politics, economics, health, science and technology, sports, arts, and culture (Abdelali et al., 2016).

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,** and **WikiNews** (Habash et al., 2022).

<sup>4</sup><https://www.hindawi.org/books/categories/children.stories/>

<sup>5</sup><https://www.hindawi.org/books/70706142/128/>

<sup>6</sup><https://ar.wikipedia.org/>

<sup>7</sup><https://chatgpt.com/>

<sup>8</sup><https://www.un.org/ar/about-us/universal-declaration-of-human-rights>

<sup>9</sup><https://ar.wikipedia.org/wiki/المعلقات>

<sup>10</sup><https://corpus.quran.com/>

<sup>11</sup><https://github.com/ShathaTm/LK-Hadith-Corpus>

<sup>12</sup><http://al-nada.eb2a.com/1000lela&lela/>

<sup>13</sup><https://www.hindawi.org/books/90463596/>

<sup>14</sup><https://www.arabicbible.com/>

<sup>15</sup><https://www.hindawi.org/books/72707304/>

Group	Source	Text	# Documents	# Sentences	# Words
Children	Other	أشعار وأخبار هنادا طه	1	364	1,163
		ماما تصنع خبزا	1	33	416
		أشعار سليمان العيسى	1	96	333
		القطعة وقبعة العيد	1	25	235
		لغتي تغني ببيان صفدي	1	359	1,879
	Hindawi	لَوْلِيَّةُ أَمِيرَةِ الْغَزَلَانِ	1	78	471
		الشَّاجِرُ مَرْمَرٌ	1	150	1,498
		أَخْلَامُ بَسْمِيَّةَ	1	104	750
		الوردة الشامية	1	13	247
	Emarati Curriculum	Grades 1 - 6	18	2,700	21,016
Young Adults	Hindawi	2-4 word sentences	1	195	849
		5-7 word sentences	1	152	766
		8-10 word sentences	1	94	879
	Hindawi	الكمبيوتر العربي	1	89	1,067
		ألوان من قصص الأطفال في الأدب العالمي	1	136	1,853
		قِصَصٌ صَبِيحِيَّةٌ لِلْأَطْفَالِ	1	148	1,812
		حكايات هانس أندرسن الخيالية	1	129	1,827
		الشوقيات - ديوان الاطفال	1	126	825
	Emarati Curriculum	Grades 7 - 12	17	1,026	9,805
	Wikipedia	الظهور في الثقافة	1	36	622
		إنسان رقمي	1	31	609
		عمر بن عبد العزيز	1	34	660
		الإسكندر الأكبر	1	32	656
		الإمارات العربية المتحدة	1	32	601
		القارة القطبية الجنوبية	1	24	651
		تاريخ فلسطين	1	29	638
		طريق الحرير	1	26	632
		الجبر	1	35	604
		خوارزمية	1	22	397
		علم الفلك	1	34	664
		فلسفة	1	41	691
		تجارة	1	33	679
		سيكولوجية التعلم	1	15	377
		المنطق	1	38	682
		تفكير	1	56	607
		اليهودية	1	33	635
		تاريخ الأديان	1	38	640
		ذكاء اصطناعي	1	37	664
		هندسة	1	27	567
Adult Modern Arabic	WikiNews	Wikinews	70	986	18,204
	Other	الكثكول	1	329	2,300
	UN	Universal Declaration of Human Rights	1	86	1,270
	Subtitles	Subtitles	1	498	3,169
	Sara	سارة (العقاد)	1	53	1,165
Adult Classical Arabic	Hayy	حي بن يقظان	1	65	1,038
	1001	ألف ليلة وليلة	17	426	4,559
	Hanging Odes	المعلقات	10	166	1,547
	Quran	Selected Surahs	17	294	4,825
	Old Testament	Selected Chapters	20	333	5,546
	New Testament	Selected Chapters	16	332	5,581
	Hadith	Selected Hadiths	47	393	4,480
Totals			274	10,631	113,651

Table 4: **BAREC** Dataset Details: the texts used to build the dataset, their groups and sources, and the number of documents, sentences, and words extracted from each text.

## C ChatGPT Prompts

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. البنات 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. الفرد يتسلق الأشجار بسرعة ويقفز ببراعة من فرع إلى فرع.	

Table 5: ChatGPT Prompts. % Discarded is the percentage of discarded sentences due to grammatical errors.

## D Detailed Annotation Stats

RL	Children	Young Adults	Adult Modern Arabic	Adult Classical Arabic	Total	%
c1-alif	86	4	3	0	93	0.9%
c2-ba	90	11	2	0	103	1.0%
c3-jim	322	31	35	2	390	3.7%
c4-dal	160	26	8	2	196	1.8%
c5-ha	526	109	29	6	670	6.3%
c6-waw	270	52	35	40	397	3.7%
c7-zay	772	135	64	19	990	9.3%
c8-ha	427	159	264	161	1,011	9.5%
c9-ta	167	31	47	44	289	2.7%
c10-ya	451	291	364	196	1,302	12.2%
c20-kaf	324	224	124	43	715	6.7%
c30-lam	469	362	286	566	1,683	15.8%
c40-mim	81	117	96	306	600	5.6%
c50-nun	198	489	509	318	1,514	14.2%
c60-sin	18	158	58	123	357	3.4%
c70-ayn	2	82	21	49	154	1.4%
c80-fa	0	23	7	70	100	0.9%
c90-sad	0	2	0	38	40	0.4%
c100-qaf	0	1	0	26	27	0.3%
Total	4,363	2,307	1,952	2,009	10,631	100.0%

Table 6: Detailed Annotation Statistics across Readability Levels and Reading Groups.