A Large and Balanced Multi-Domain Arabic Corpus Annotated for Morphology, Syntax, and Readability

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

Abstract

We present BAREC-10M, an expanded version of the Balanced Arabic Readability Evaluation Corpus (BAREC). This new release extends the original 1M-word corpus to 10 million words and broadens its scope to include balanced multi-domain coverage annotated for morphology, syntax, and readability. The corpus integrates 45 sub-corpora drawn from diverse sources, including news, educational materials, literature, children's texts, and religious discourse. Each text is labeled for domain, readership level, and genre, and automatically analyzed using state-of-the-art morphological and syntactic tools. To enhance coverage of underrepresented varieties, we manually digitized and included children's materials, magazines, and curriculum-based content. The resulting dataset provides a balanced resource for studying Arabic linguistic variation across styles, audiences, and levels of complexity.

Keywords: Arabic, Balanced Corpus, Readability

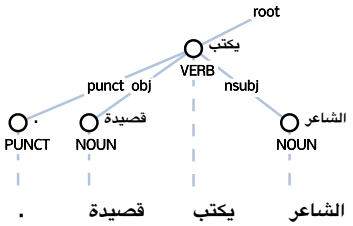
1. Introduction

Over the past two decades, Arabic has witnessed extensive corpus development for linguistic and computational research (Zaghouani, 2014; Al-Thubaity, 2015; Al-Sulaiti and Atwell, 2006). Existing corpora, however, often focus on a single dimension such as morphology/syntax (Maamouri and Bies, 2004), domain coverage (Arts et al., 2014), or readability (Elmadani et al., 2025b), and rarely integrate multiple linguistic layers with balanced sampling across genres and readership levels.

This paper introduces BAREC-10M, the 10-million-word expansion of the Balanced Arabic Readability Evaluation Corpus (BAREC) (Elmadani et al., 2025b), designed to unify multiple linguistic annotations in a single resource. BAREC-10M integrates automatic morphological and syntactic annotation with readability labeling, organized across three domains (Arts & Humanities, Social Sciences, STEM) and three readership groups (Foundational, Advanced, Specialized). The data selection process prioritized diversity and proportional representation across educational levels, domains, and genres. Table 1 presents an example sentence with associated annotations.

To address the scarcity of child-directed and educational content, we manually digitized a range of children's books and magazines (e.g., Arabic comics *Batman*, *Superman*, *Lulu*). We also incorporated existing corpora whose annotations strengthen the dataset's linguistic richness: the DARES corpus (El-Haj et al., 2024) and national curricula (Tunisian, Emirati) provide grade-aligned readability levels, while ReadMe++ (Naous et al., 2024) offers CEFR-based annotation.

Syntax (UD)



Morphology				
Diacritization	.	قَصِيدَةً	يَكْتُبُ	الشّاعِرُ
Lemma	.	قَصِيدَة	كَتَب	شاعِر
Gloss	.	poem	write	poet
POS	punc	noun	verb	noun
PRC	na	0	0	Al_det
Aspect	na	na	i	na
Voice	na	na	a	na
Mood	na	na	i	na
Gender	na	f	m	m
Number	na	s	s	s
State	na	i	na	d
Case	na	a	na	n
Person	na	na	3	na
Rationality	na	i	n	r
Lexical Readability		3	1	3
Sentence Readability			5	
Domain			Arts and Humanities	
Level			Foundational Text	
Category			Educational Materials	

Table 1: Annotations for the sentence: الشاعر يكتب قصيدة 'the poet writes a poem'. Readability in BAREC (1-19) (Habash et al., 2025).

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 11761–11775
11–16 May 2026.

Authors	Corpus	Size	Morphology	Syntax	Readability	Domains	Level	Access
Maamouri and Bies (2004)	LDC ATB	1M	Manual	Manual	-	News	A	L
Al-Sulaiti and Atwell (2006)	CCA	0.8M	-	-	-	Mix	A	B
Arts et al. (2014)	arTenTen	5.8B	Auto	-	-	Mix	AS	B
Al-Thubaity (2015)	KACST	730M	-	-	-	Mix	AS	B
Al-Sulaiti et al. (2016)	ACC	1.8M	-	-	-	Child	FA	B
Abu El-Khair (2016)	Abu El-Khair	1.5B	-	-	-	News	AS	O
Belinkov et al. (2016)	Shamela	1B	Auto	-	-	Religion	S	O
Al Khalil et al. (2018)	Al Khalil's	7M	Auto	-	Given	Curr, Fict	FAS	NR
Sawalha et al. (2019)	JCCA	100M	Portions	-	-	Mix	FAS	O
Habash et al. (2022)	CamelTB	188K	Manual	Manual	-	Mix	AS	O
Habash and Palfreyman (2022)	ZAEBUC	33K	Manual	-	Manual	Essays	A	O
Alhafni et al. (2024)	SAMER	159K	-	-	Manual	Fict	S	O
Naous et al. (2024)	ReadMe++	47K	-	-	Manual	News	A	O
Almujaiwel et al. (2025)	DARES2.0	7M	-	-	Given	Curr	FAS	O
Elmadani et al. (2025b)	BAREC	1M	-	-	Manual	Mix	FAS	O
Ours	**BAREC-10M**	10M	Auto	Auto	Auto	Mix	FAS	O

Table 2: Comparison of major Arabic corpora annotated for Morphology, Syntax, and Readability. **Domains**: **Curr**iculum; **Fict**ion. Readership **Level**: **F**oundational, **A**dvanced, **S**pecialized. **Access**: **O**pen, **B**rowsable, **L**icensed, Not Released (**NR**).

BAREC-10M includes 45 sub-corpora spanning news, literature, education, religion, and benchmarks, sampled for balanced coverage of Arabic textual variation. Only brief excerpts were used where required to comply with fair-use and licensing terms. The corpus will be released as an open-access resource for research and educational use.[1]

The paper is organized as follows. Section 2 reviews related work. Section 3 describes data selection and compilation. Section 4 details the annotation pipeline. Section 5 presents corpus analysis and key findings.

2. Related Work

Over the past two decades, there has been a sustained and extensive effort to develop Arabic corpora for a wide range of linguistic and computational purposes. In this work, we discuss fifteen representative datasets, though many others exist. For broader overviews of Arabic corpus construction, see (Zaghouani, 2014; Al-Thubaity, 2015; Al-Sulaiti and Atwell, 2006).

The selected corpora vary across multiple dimensions, including their time of creation, size, linguistic annotation (morphology, syntax, and readability), domain diversity, readership level, and accessibility. Early efforts often prioritized certain aspects while trading off others. The LDC's Arabic Treebank (ATB) (Maamouri and Bies, 2004) set the standard for syntactic annotation, pairing manual morphological and syntactic layers over approximately one million tokens. However, it was limited in domain coverage (news only) and restricted in access. Al-Sulaiti and Atwell (2006) introduced the Corpus of Contemporary Arabic (CCA), which expanded genre diversity but lacked linguistic annotation and remained small in scale. The KACST corpus (Al-Thubaity, 2015) further increased data size to 730M words, and *arTenTen* (Arts et al., 2014) scaled this up to 5.8B words of web text, automatically annotated for morphology.

Subsequent corpora incorporated manual or semi-automatic annotations to support more specialized linguistic analysis. Examples include JCCA (Sawalha et al., 2019), ACC for child-directed language (Al-Sulaiti et al., 2016), and Shamela (Belinkov et al., 2016), a large religious and philosophical corpus. Al Khalil's corpus (Al Khalil et al., 2018) contributed automatic morphological annotation alongside readability leveling derived from curriculum data.

Recent years have seen a renewed emphasis on manual annotation and balanced design. Habash et al. (2022) introduced *CamelTB*, the first multi-genre Arabic treebank with manual morphological and syntactic annotation. Habash and Palfreyman (2022) released *ZAEBUC*, which includes manual morphological annotation and CEFR-based readability labels (Council of Europe, 2001). Similarly, Naous et al. (2024) presented *ReadMe++*, a news-focused corpus with manual CEFR readability levels, and Alhafni et al. (2024) introduced *SAMER*, the first Arabic parallel corpus aligned across readability levels. Almujaiwel et al. (2025) developed *DARES2.0*, a large educational corpus with predefined curricular readability levels. Most recently, Elmadani et al. (2025b) introduced the *Balanced Arabic Readability Corpus Evaluation (BAREC)*, comprising one million words from diverse domains and readership levels, manually annotated for fine-gained readability (Habash et al., 2025).

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

This paper is part of the broader BAREC project, whose goal is to develop a comprehensive reference resource to facilitate the study and evaluation of Arabic readability across the Arab world.[1] Prior to this work, we released several key resources and initiatives. First, we introduced the BAREC Annotation Framework (Habash et al., 2025), a set of guidelines for fine-grained (1-19), sentence-level Arabic readability annotation based on Taha-Thomure (2017). We then released the BAREC Corpus (Elmadani et al., 2025b), a balanced 1M-word dataset manually annotated for sentence-level readability assessment. To support practical use, we developed BAREC Demo (Altarbouch et al., 2025), a web-based system for fine-grained sentence-level Arabic readability assessment. In addition, we organized the BAREC Shared Task 2025 on Arabic readability assessment to encourage community engagement and benchmarking (Elmadani et al., 2025a). Finally, in this paper, we introduce BAREC-10M, a 10M-word dataset that integrates automatic morphological, syntactic, and readability annotations. BAREC-10M is carefully curated to include balanced, fair-use samples spanning multiple genres and readership levels.

As shown in Table 2, BAREC-10M stands out for its size, multi-layered annotation, and balanced representation of domains and readership levels. In the following sections, we describe the collection and annotation process, followed by an analysis and evaluation of the resulting corpus.

3. Corpus Selection

The data selection process aimed to ensure broad coverage across educational levels, domains, and genres, with attention to balanced distributions. To address the scarcity of child-directed materials in various genres and domains, we manually typed a number of children's Arabic books and magazines (e.g. *Batman*, *Superman* and *Lulu*). Several sub-corpora were also included because existing annotations made them particularly valuable. For instance, the **DARES** corpus (El-Haj et al., 2024) and the Tunisian and Emirati curricula naturally provide readability levels aligned with grade levels, while the **ReadMe++** dataset (Naous et al., 2024) includes CEFR-based manual annotations. In addition, our corpus incorporates the **CamelTB** (Habash et al., 2022) (partial) and **BAREC** (Elmadani et al., 2025b) (full) corpora, which offer sentence-level syntactic and readability annotations, respectively. We include **45 sub-corpora** spanning news, educational, literary, children's, religious, and benchmark domains. For each, we sample small, balanced portions to ensure diversity of genre, style, and readability. We thematically group the sources as follows:

- **News and Media** We draw from three SANAD (Einea et al., 2019) news sources: Akhbarona (6 categories, 300 files each),[2] AlArabiya (6 categories, 350 files each),[3] and AlKhaleej (7 categories, 350 files each).[4] We further include 62 articles from WikiNews (Abdelali et al., 2016) and 175 from Wikipedia, covering diverse topics.[5] Majarra[6] contributed 50 curated online articles across five genres.
- **Educational and Curriculum Materials** This group includes full or partial Arabic textbooks from multiple national curricula: DARES1.0 (El-Haj et al., 2024), the Emirati Curriculum (Arabic, Islamic, and Social Studies books across 12 grades), and the Tunisian Curriculum (grades 2–5). We also incorporate DARES2.0 (Almujaiwel et al., 2025) for comparison with curricular readability levels.
- **Children's Literature and Media** We sample extensively from digitized and manually typed children's materials, including Arab Youth Publisher (74 books), Green Library (61 books), Reading Corner (37 books), and the Arab Thought Foundation's *Ofoq* series (827 articles, 2018–2024)[7]. We further include periodicals and magazines such as *Majed* (10 editions, 1983–2019), *Lulu* (10 editions), *Batman* and *Superman* comics (10 editions each), and *The Five Adventurers* (**5-Adventurers**, المغامرون الخمسة, 187 editions). We supplement these with 56 Spacetoon theme songs, selected children's poems (Al-Safadi, 2005; Taha-Thomure, 2007), and 450 synthetic children's sentences generated by ChatGPT.[8]
- **Literary and Philosophical Texts** Our literary subset includes samples from the Hindawi Foundation (2,758 books),[9] Kalima (596 books)[10], and the Library of Arabic Literature (**Lib. Arab Lit.**, 36 volumes)[11]. We also include classical and modern works: *Hayy ibn Yaqdhan* (Tufail, 1150), *Sara* by Al-Akkad (Al-Akkad, 1938), and *The Arabian Nights* (**1001 Nights**) (Unknown, 12th century). Poetic heritage is represented through the Arabic Poem Comprehensive Dataset (**Arab Poem CD**) (Yousef et al., 2019) (1,041 poems) and the pre-Islamic *Hanging Odes* (المعلقات).

[2] www.akhbarona.com
[3] www.alarabiya.net
[4] www.alkhaleej.ae
[5] ar.wikipedia.org
[6] www.majarra.com
[7] arabthought.org/ar/researchcenter/ofoqelectronic
[8] www.chatgpt.com
[9] www.hindawi.org
[10] www.alc.ae/publications/kalima
[11] www.libraryofarabicliterature.org

Sub-corpus	Status	Level	Domain	Text Category	#Docs	#Sents	#Words
ChatGPT	Existing	F	A&H	Educational Materials	3	449	2,544
Green Library	Existing	F	A&H	Literature; Art & Music	61	2,972	47,551
Spacetoon	Existing	F	A&H	Literature; Art & Music	56	958	4,219
Majed	Existing	F	A&H	Media & Culture	318	12,770	130,897
Other	Existing	FA	A&H	Literature; Art & Music	37	1,296	6,716
DARES	Existing	FAS	Mixed	Educational Materials	1,832	13,870	1,092,399
ArabicMMLU	Existing	FAS	Mixed	Educational Materials	358	7,086	217,084
BTEC	Existing	A	A&H	Educational Materials	20	2,000	15,935
ZAEBUC	Existing	A	A&H	Educational Materials	100	1,109	15,778
Arabic Learner	Existing	A	A&H	Educational Materials	16	727	9,228
Sara	Existing	A	A&H	Literature; Art & Music	1	59	1,175
ANERCorp	Existing	A	A&H	Media & Culture	244	4,898	151,163
ReadMe++	Existing	A	A&H	Media & Culture	97	1,945	49,646
Subtitles	Existing	A	A&H	Media & Culture	11	567	3,594
Wikipedia	Existing	A	Mixed	Encyclopedic	175	5,749	124,726
AlKhaleej	Existing	A	Mixed	Media & Culture	2,450	31,376	1,038,825
AlArabiya	Existing	A	Mixed	Media & Culture	2,099	17,736	567,218
Akhbarona	Existing	A	Mixed	Media & Culture	1,791	19,105	551,997
WikiNews	Existing	A	Mixed	Media & Culture	62	879	16,010
UN	Existing	A	SS	Academic	1	93	1,426
Arab Poem CD	Existing	S	A&H	Literature; Art & Music	1,041	33,075	299,945
1001 Nights	Existing	S	A&H	Literature; Art & Music	35	1,145	11,831
Hanging Odes	Existing	S	A&H	Literature; Art & Music	10	784	7,465
Hayy	Existing	S	A&H	Religion & Philosophy	20	1,198	19,674
Hadith	Existing	S	A&H	Religion & Philosophy	134	1,190	12,467
Quran	Existing	S	A&H	Religion & Philosophy	54	572	11,699
New Testament	Existing	S	A&H	Religion & Philosophy	16	573	9,593
Old Testament	Existing	S	A&H	Religion & Philosophy	20	535	9,097
Constitutions	Existing	S	SS	Academic	17	1,643	34,023
MedArabiQ	Existing	S	STEM	Academic	2,205	34,323	1,321,794
5-Adventurers	New	F	A&H	Literature; Art & Music	187	10,328	130,365
Superman	New	F	A&H	Literature; Art & Music	127	8,477	72,350
Batman	New	F	A&H	Literature; Art & Music	69	4,688	37,639
Arab Youth Pub.	New	F	A&H	Literature; Art & Music	74	2,633	34,267
Lulu	New	F	A&H	Literature; Art & Music	56	5,237	30,220
Ladybird	New	F	A&H	Literature; Art & Music	35	1,846	22,031
Reading Corner	New	F	A&H	Literature; Art & Music	37	1,094	8,082
Tunisian Curr.	New	FA	A&H	Educational Materials	152	8,783	89,778
Emirati Curr.	New	FAS	A&H, SS	Educational Materials	2,228	141,979	1,336,002
Hindawi	New	FAS	Mixed	Mixed	2,758	81,489	1,528,793
Majarra	New	A	SS, STEM	Media & Culture	50	995	28,515
Lib. Arab Lit.	New	S	A&H	Literature; Art & Music	36	1,731	22,557
Anthems	New	S	A&H	Literature; Art & Music	19	360	1,595
Kalima	New	S	Mixed	Mixed	596	25,627	436,858
Ofoq	New	S	SS	Academic	827	18,930	512,398
Total					**20,535**	**514,879**	**10,077,169**

Table 3: Overview of the sub-corpora in BAREC-10M Corpus. **Domain**: **A**rts **&** **H**umanities, **S**ocial **S**ciences, **STEM**. Readership **Level**: **F**oundational, **A**dvanced, **S**pecialized. See Appendix A for more information about the sub-corpora.

- **Religious and Foundational Texts** We include major religious texts for linguistic and stylistic contrast: 134 Hadiths from *Sahih Bukhari* (al Bukhari, 846; Altammami et al., 2019), the Quran (first three and last fourteen Surahs) (Dukes et al., 2013), and Arabic translations of the *Old Testament* and *New Testament*.[12] We also include the Arabic version of the *Universal Declaration of Human Rights* (**UN**).[13]
- **Educational Benchmarks and Learner Data** This category covers datasets designed for language learning and assessment: the **Arabic Learner** Corpus (Alfaifi, 2015), ReadMe++ (Naous et al., 2024), ZAEBUC (Habash and Palfreyman, 2022), and SAMER (Alhafni et al., 2024). We also include ANERCorp (Benajiba et al., 2007) and the subsets of ArabicMMLU (Koto et al., 2024) and MedArabiQ (Daoud et al., 2025).
- **Other Sources** Additional general-language data were drawn from the Basic Travel Expressions Corpus (BTEC) (Eck and Hori, 2005; Takezawa et al., 2007), Arabic subtitles from OpenSubtitles (Lison and Tiedemann, 2016), and national anthems from 19 Arabic-speaking countries.

Table 3 provides an overview of all sub-corpora included in our dataset, detailing the number of documents, sentences, and words, as well as their domains, readership groups, and text categories. The table also indicates whether each sub-corpus was pre-existing and previously used in NLP research, or newly added in this work. The new sub-corpora include additions to existing sources such as Hindawi, Kalima, and the Emirati Curriculum; newly identified born-digital corpora such as Majarra, Library of Arabic Literature, Anthems, and Ofoq; and manually transcribed corpora created from scanned sources such as Superman, Batman, Lulu, 5-Adventurers, and Ladybird. Overall, 38% of the words in the BAREC-10M corpus come from these newly collected resources. More information about different sub-corpora is available in Appendices A and B.

4. Corpus Annotation

Our corpus is enriched with multiple layers of annotation to support a wide range of NLP tasks. These include annotations at the document, sentence, and word levels. Document-level annotations were performed manually, while sentence- and word-level annotations were generated automatically.

[12] www.arabicbible.com

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

4.1. Manual Annotation

Each document was manually labeled for domain, readership, and text category.

Domains Each document was assigned to one of three broad **domains**, *Arts & Humanities*, *Social Sciences*, or *STEM*, based on its primary subject matter and communicative purpose.

- **Arts & Humanities:** Includes literature and fiction (novels, poetry, short stories), religious and philosophical texts, as well as educational, reference, and news materials related to cultural and artistic topics.
- **Social Sciences:** Covers business, law, and social studies (sociology, anthropology, history), along with associated educational materials, encyclopedic entries, and news or commentary on social and political issues.
- **STEM:** Encompasses scientific and technical texts across science, technology, engineering, and mathematics, including research papers, educational materials, reference works, and science or technology news.

Readership Groups Each document was labeled according to its intended **readership group**, reflecting expected reading proficiency and linguistic complexity.

- **Foundational:** Texts for early learners (typically up to grade 4 or age 10) focusing on basic literacy, word decoding, and simple sentence comprehension.
- **Advanced:** Materials for general adult readers, featuring moderate linguistic and conceptual complexity suited to everyday reading and learning contexts.
- **Specialized:** Domain-specific or technical texts for expert readers (roughly grade 9 and above) requiring advanced vocabulary and conceptual understanding.

Text Categories Texts were classified into six broad **categories** based on three complementary dimensions: (1) observable linguistic and stylistic features, (2) communicative purpose, and (3) social or conventional expectations associated with the text type. Together, these dimensions define the dominant functional and stylistic profile of each text.

- **Media & Culture:** Journalistic and popular media texts that blend information and entertainment, use accessible language, and address topical issues, that are typically short-lived and produced for rapid consumption, e.g., AlKhaleej News.
- **Religion & Philosophy:** Texts conveying moral, doctrinal, or spiritual content, often using figurative and didactic language drawn from religious or philosophical traditions, e.g., the Quran.

- **Literature, Art & Music:** Creative and expressive texts emphasizing aesthetic form, stylistic richness, and rhetorical devices such as metaphor and rhythm, e.g., the Green Library.
- **Academic:** Scholarly or scientific writing organized through argumentation, evidence, and reasoning. Characterized by explicit structure, technical vocabulary, and analytical tone, e.g. Ofoq.
- **Educational Materials:** Pedagogical texts intended for instruction or language learning, including textbooks and curricular materials with explanations, examples, and exercises, e.g., Emirati Curriculum.
- **Encyclopedic:** Expository reference texts presenting established knowledge in a factual, neutral, and definitional manner, emphasizing clarity and stability of information, e.g., Wikipedia.

4.2. Automatic Annotation

Automatic annotations cover sentence-level readability assessment, syntactic parsing, lemmatization, and morphosyntactic tagging.

Morphosyntactic Tagging and Lemmatization The entire corpus is processed using the morphological disambiguation component of *CAMeL Tools* to obtain the top-ranked morphological analysis for each token in context (Obeid et al., 2020a; Inoue et al., 2022).[14] The resulting analysis provides the lemma, part-of-speech tag, and 15 other morphosyntactic features, including gender, number, person, state, case, mood, rationality, and clitics. In cases where the disambiguator outputs multiple possible lemmas, we resolve the ambiguity using lemma clusters[15] proposed by Saeed and Habash (2025). These clusters group semantically related lemmas and enable the selection of a single lemma.[16]

Syntactic Parsing We generate syntactic trees for all sentences in the corpus using *Camel-Parser 2.0* (Elshabrawy et al., 2023). Each sentence is parsed in both the Columbia Arabic Treebank (*CATiB*) (Habash and Roth, 2009; Habash et al., 2009) and Universal Dependencies (*UD*) (Taji et al., 2017; Nivre et al., 2017) styles.

Readability Assessment For non-BAREC corpus sentences, we assigned a readability level following the BAREC 19-level framework (Habash et al., 2025). Before prediction, the text is cleaned and diacritics are removed using *CAMeL Tools*. We then apply the Elmadani et al. (2025b) `AraBERTv02+Word+CE` model[17] to predict the readability level for each sentence. Finally, lemma-level readability annotations are derived from the BAREC-10M corpus by combining sentence-level readability scores with lemmatization. For each lemma, we assign the readability level of the easiest sentence in which it appears in BAREC-10M. This approach assigns the highest possible readability level to a lemma (upper boundary) assuming ideal lemmatization and readability prediction.

5. Corpus Analysis

This section analyzes the BAREC-10M corpus across domains, readership groups, and text categories, linking readability to lexical and syntactic complexity.

5.1. Domain, Readership Groups, and Text Categories

Domain × Text Category Table 4 shows the distribution of words across different domains and text categories in the BAREC-10M corpus. The Arts & Humanities domain covers a wide range of genres, especially Literature, Art & Music, Media & Culture, and Educational Materials. In contrast, Social Sciences and STEM are mainly composed of Academic texts, showing their focus on scholarly and scientific writing. Overall, Academic writing dominates technical fields, while Cultural and Artistic genres are concentrated in the Arts & Humanities.

Domain × Readership Group Table 5 shows the distribution of words across different domains and readership groups in the BAREC-10M corpus. Across domains, there is a clear progression from Foundational to Specialized readerships. Arts & Humanities exhibits the most balanced distribution across all three readership groups, indicating a diversity of materials for emerging and expert readers. In contrast, Social Sciences and STEM are heavily skewed toward Specialized Texts, reflecting their focus on academic content. Overall, Arts & Humanities spans a broad educational range, while STEM and Social Sciences mainly target competent and expert readers.

Readership Group × Text Category Table 6 shows the distribution of words across different readership groups and text categories in the BAREC-10M corpus. When examining readership

[14] *CAMeL Tools v1.5.5:* `BERT-Disambig + CamelMorph-MSA v1.1 db (Khairallah et al., 2024).`

[15] `https://huggingface.co/CAMeL-Lab/camelbert-msa-pos-msa-lemma-clustering`

[16] `https://github.com/CAMeL-Lab/lemmatization-as-classification`

[17] `https://huggingface.co/CAMeL-Lab/readability-arabertv02-word-CE`

Text Category	Arts & Humanities		Social Sciences		STEM		All	
	Words	%	Words	%	Words	%	Words	%
Educational Materials	1,545,083	15.3	672,058	6.7	561,607	5.6	2,778,748	27.6
Literature, Art & Music	1,937,864	19.2	—	—	—	—	1,937,864	19.2
Media & Culture	1,447,963	14.4	492,229	4.9	597,673	5.9	2,537,865	25.2
Academic	45,885	0.5	1,109,545	11.0	1,474,543	14.6	2,629,973	26.1
Encyclopedic	35,493	0.4	63,864	0.6	25,369	0.3	124,726	1.2
Religion & Philosophy	67,993	0.7	—	—	—	—	67,993	0.7
All	**5,080,281**	**50.4**	**2,337,696**	**23.2**	**2,659,192**	**26.4**	**10,077,169**	**100.0**

Table 4: Word counts and proportions across domains and text categories in the BAREC-10M corpus.

Readership Group	Arts & Humanities		Social Sciences		STEM		All	
	Words	%	Words	%	Words	%	Words	%
Foundational Text	1,164,194	11.6	119,228	1.2	49,460	0.5	1,332,882	13.2
Advanced Text	2,195,598	21.8	827,090	8.2	746,789	7.4	3,769,477	37.4
Specialized Text	1,720,489	17.1	1,391,378	13.8	1,862,943	18.5	4,974,810	49.4
All	**5,080,281**	**50.4**	**2,337,696**	**23.2**	**2,659,192**	**26.4**	**10,077,169**	**100.0**

Table 5: Word counts and proportions across domains and readership groups in the BAREC-10M corpus.

Text Category	Foundational Text		Advanced Text		Specialized Text		All	
	Words	%	Words	%	Words	%	Words	%
Educational Materials	499,954	5.0	970,844	9.6	1,307,950	13.0	2,778,748	27.6
Literature, Art & Music	697,408	6.9	221,183	2.2	1,019,273	10.1	1,937,864	19.2
Media & Culture	130,897	1.3	2,406,968	23.9	—	—	2,537,865	25.2
Academic	4,623	0.0	45,756	0.5	2,579,594	25.6	2,629,973	26.1
Encyclopedic	—	—	124,726	1.2	—	—	124,726	1.2
Religion & Philosophy	—	—	—	—	67,993	0.7	67,993	0.7
All	**1,332,882**	**13.2**	**3,769,477**	**37.4**	**4,974,810**	**49.4**	**10,077,169**	**100.0**

Table 6: Word counts and proportions across readership groups and text categories in the BAREC-10M corpus.

by genre, Foundational Texts cluster strongly in Educational Materials and Literature, indicating their pedagogical focus. Advanced Texts are common in Media & Culture, consistent with intermediate or general-knowledge audiences. Specialized Texts, however, are concentrated almost entirely in Academic writing, with minor representation in educational and literature materials. This reflects a clear gradient of specialization across readership levels, from instructional and narrative genres to scholarly and technical writing.

5.2. Readership Groups and Readability

Figure 1 illustrates the distribution of readability levels across different readership groups. We observe a gradual increase in the proportion of higher readability levels as we move from Foundational to Specialized texts. This trend reflects the alignment between the manual annotations at the document level and the automatic readability annotation at the sentence level.

5.3. Readability across Sub-corpora

Figure 2 presents the sentence-level readability distribution for all sub-corpora, ordered from the easiest to the most difficult based on their average readability levels. As expected, materials intended for children appear on the left side of the figure, reflecting lower readability levels. In contrast, specialized corpora such as *MedArabiQ*, *APCD*, and *Odes* show much higher readability levels.

Surprisingly, the *DARES* corpus ranks among the most difficult, even though it is extracted from school textbooks. This can be explained by the nature of its content: it only includes sentences that describe concepts, which tend to be longer and more complex.

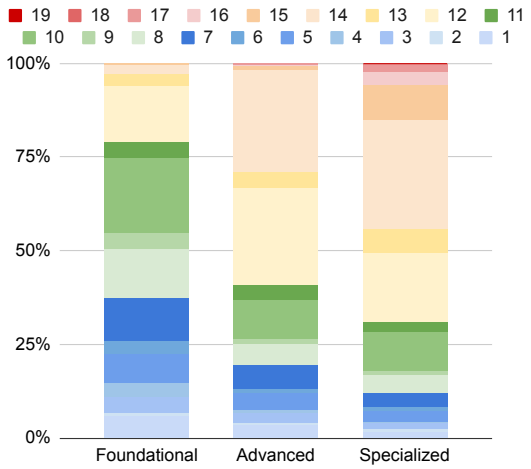


Figure 1: The distribution of readability levels across readership groups.

5.4. Readability: Sentences vs. Lemmas

Figure 3 illustrates how lemma-level readability is distributed across different sentence-level readability levels. The x-axis represents sentence-level readability, while the y-axis shows the distribution of lemma-level readability.

By definition (see Section 4.2), all lemmas in a sentence must have readability levels equal to or lower than the sentence's level. This analysis aims to understand the nature of difficult sentences — whether they are composed mainly of difficult lemmas, or if they consist mostly of easy lemmas with only a few hard ones raising the overall difficulty.

The figure reveals that most high-level sentences are dominated by lemmas from lower readability levels. In fact, over 50% of the lemmas in the hardest sentences (Level-19) belong to Level-5 or below. Additionally, Level-1 lemmas form the majority across all sentence-level readability groups.

5.5. Readability and Syntax

We analyzed how several syntactic tree features relate to sentence readability. Using the `conllx_df` tool,[18] we extracted the following features from the CATiB-style syntactic trees: **Depth**, **Breadth** (the maximum number of nodes at any tree level), **Maximum Branching Factor** (the largest number of child nodes from a single parent), and **Sentence Length** (number of words).

Figure 4 shows the average readability level for each value of these features. The x-axis represents the feature values, and the y-axis shows the corresponding average readability level. For example, at $x = 5$, the blue line represents the average readability level for sentences whose syntactic tree has a depth of 5, while the red line represents sentences of length 5.

Overall, sentence depth shows the strongest relationship with readability, spanning the widest range of levels (4–14), while Breadth varies within a narrower range (6–14). The maximum branching factor becomes unstable at higher values due to limited data (only 0.1% of trees exceed a branching factor of 15). Sentence length is a good indicator of readability for short sentences (up to 10 words), but its influence decreases for longer ones. Numerically, the correlation with readability is highest for depth (0.60), followed by maximum branching factor (0.53), breadth (0.50), and length (0.50).

6. Conclusion and Future Work

We introduced **BAREC-10M**, a large-scale, multi-domain Arabic corpus annotated for morphology, syntax, and readability. By combining 45 heterogeneous sources into a single harmonized framework, BAREC-10M bridges existing efforts and offers balanced coverage across domains and readership levels. The corpus provides a foundation for future research on linguistic variation, readability modeling, and cross-genre NLP in Arabic.

Future work will expand BAREC-10M with dialectal data, deeper semantic annotation, and manual validation of the automatic layers. We also plan to incorporate additional educational and social media sources to improve domain diversity and temporal coverage.

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. Special thanks go to Abdallah Abushmaes, Karin Aghadjanian, and Omar Al Ayyoubi of the ALC for their continued support. We would also like to thank the Zayed University ZAI Arabic Language Research Center team, in particular Hamda Al-Hadhrami, Maha Fatha, and Metha Talhak, for their valuable contributions to typing materials for the project. We also acknowledge Ali Gomaa and his team for their additional support in this area. We are also grateful to Mostafa Saeed for running the corpus through the morphosyntactic tagging and lemmatization systems. Finally, we thank members of the Arabic NLP community who have made datasets publicly available. We also acknowledge our publishing partners: Arab Thought Foundation, Kalima, Library of Arabic Literature, Majarra, and NYUAD's Clinical AI Lab.

[18] www.github.com/CAMeL-Lab/conllx_df

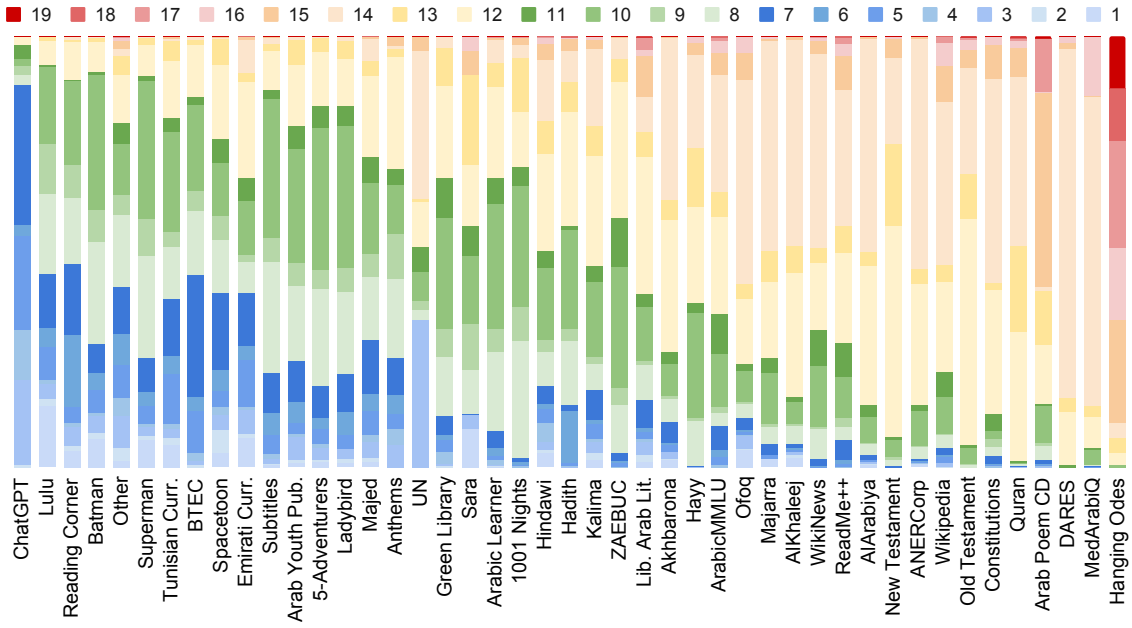


Figure 2: The distribution of readability levels of sentences across all sub-corpora.

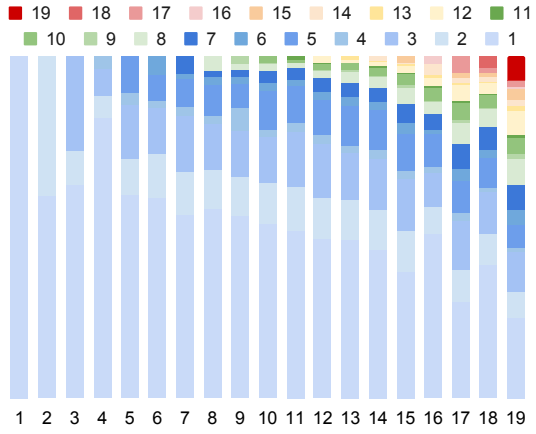


Figure 3: The distribution of lemma levels at each sentence-level readability.

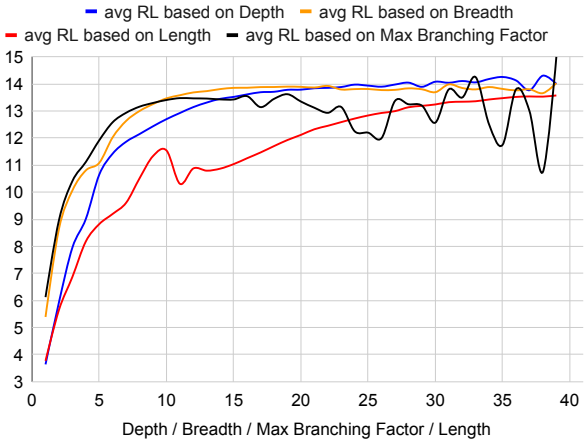


Figure 4: Average readability level (RL) as a function of Syntactic Tree Depth, Breadth, Maximum Branching Factor, and Sentence Length.

Limitations

Although BAREC-10M covers a wide range of text types, it remains limited in dialectal and spoken Arabic coverage, as most data are in Modern Standard Arabic. Automatic morphological and syntactic annotation may introduce errors, particularly for ambiguous forms or informal text. Readability labels are automatically assigned for most texts, which, while consistent, may not fully reflect pedagogical difficulty. Furthermore, despite our efforts toward balance, the dataset may still overrepresent certain domains (e.g., news) due to data availability.

Ethics Statement

All included materials were either open-access, publicly available under permissive terms, or excerpted in small portions consistent with fair use for research. Manually typed materials from educational and children's sources were reproduced solely for research purposes without redistributing the full works. The dataset does not contain personal, private, or sensitive information. All annotators and transcribers were paid fair wages. Researchers are encouraged to use BAREC-10M responsibly and to credit original sources when reusing or extending the data.

7. Bibliographical 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.

Ibrahim Abu El-Khair. 2016. Abu el-khair corpus: A Modern Standard Arabic Corpus. *International Journal of Recent Trends in Engineering & Research*, 2:5–13.

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

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

Muhamed Al Khalil, Hind Saddiki, Nizar Habash, and Latifa Alfalasi. 2018. A leveled reading corpus of Modern Standard Arabic. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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).

Latifa Al-Sulaiti, Noorhan Abbas, Claire Brierley, Eric Atwell, and Ayman Alghamdi. 2016. Compilation of an Arabic children's corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1808–1812, Portorož, Slovenia. European Language Resources Association (ELRA).

Latifa Al-Sulaiti and Eric Steven Atwell. 2006. The design of a corpus of contemporary Arabic. *International journal of corpus linguistics*, 11(2):135–171.

Abdulmohsen O Al-Thubaity. 2015. A 700m+ Arabic corpus: Kacst Arabic corpus design and construction. *Language Resources and Evaluation*, 49(3):721–751.

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

Bashar Alhafni, Nizar Habash, and Houda Bouamor. 2022. The Arabic parallel gender corpus 2.0: Extensions and analyses. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1870–1884, Marseille, France. European Language Resources Association.

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.

Sultan Almujaiwel, Damith Premasiri, Tharindu Ranasinghe, Mo El-Haj, and Ruslan Mitkov. 2025. Complex Concept-Based Readability Estimation from Arabic Curriculum. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* Just Accepted.

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*.

Kinda Altarbouch, Khalid N. Elmadani, Ossama Obeid, Hanada Taha, and Nizar Habash. 2025. BAREC demo: Resources and tools for sentence-level Arabic readability assessment. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 950–959, Suzhou, China. Association for Computational Linguistics.

Tressy Arts, Yonatan Belinkov, Nizar Habash, Adam Kilgarriff, and Vit Suchomel. 2014. arten-ten: Arabic corpus and word sketches. *Journal of King Saud University - Computer and Information Sciences*, 26(4):357–371. Special Issue on Arabic NLP.

Yonatan Belinkov, Alexander Magidow, Maxim Romanov, Avi Shmidman, and Moshe Koppel. 2016. Shamela: A large-scale historical Arabic corpus. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, pages 45–53, Osaka, Japan. The COLING 2016 Organizing Committee.

Yassine Benajiba, Paolo Rosso, and José Miguel BenedíRuiz. 2007. ANERsys: An Arabic Named Entity Recognition System Based on Maximum Entropy. In *Computational Linguistics and Intelligent Text Processing*, pages 143–153, Berlin, Heidelberg. Springer Berlin Heidelberg.

Council of Europe Council of Europe. 2001. *Common European Framework of Reference for Languages: learning, teaching, assessment*. Cambridge University Press.

Mouath Abu Daoud, Chaimae Abouzahir, Leen Kharouf, Walid Al-Eisawi, Nizar Habash, and Farah E. Shamout. 2025. MedArabiQ: Benchmarking Large Language Models on Arabic Medical Tasks.

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.

Omar Einea, Ashraf Elnagar, and Ridhwan Al Debsi. 2019. SANAD: Single-label Arabic News Articles Dataset for automatic text categorization. *Data in Brief*, 25:104076.

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, Bashar Alhafni, Hanada Taha, and Nizar Habash. 2025a. BAREC shared task 2025 on Arabic readability assessment. In *Proceedings of The Third Arabic Natural Language Processing Conference: Shared Tasks*, pages 239–252, Suzhou, China. Association for Computational Linguistics.

Khalid N. Elmadani, Nizar Habash, and Hanada Taha-Thomure. 2025b. A large and balanced corpus for fine-grained Arabic readability assessment. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 16376–16400, Vienna, Austria. Association for Computational Linguistics.

Ahmed Elshabrawy, Muhammed AbuOdeh, Go Inoue, and Nizar Habash. 2023. CamelParser2.0: A state-of-the-art dependency parser for Arabic. In *Proceedings of ArabicNLP 2023*, pages 170–180, Singapore (Hybrid). Association for Computational Linguistics.

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, Reem Faraj, and Ryan Roth. 2009. Syntactic Annotation in the Columbia Arabic Treebank. In *Proceedings of the International Conference on Arabic Language Resources and Tools*, Cairo, Egypt.

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 and Ryan Roth. 2009. CATiB: The Columbia Arabic treebank. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 221–224, Suntec, Singapore. Association for Computational Linguistics.

Nizar Habash, Hanada Taha-Thomure, Khalid N. Elmadani, Zeina Zeino, and Abdallah Abushmaes. 2025. Guidelines for fine-grained sentence-level Arabic readability annotation. In *Proceedings of the 19th Linguistic Annotation Workshop (LAW-XIX-2025)*, pages 359–376, Vienna, Austria. Association for Computational Linguistics.

Go Inoue, Salam Khalifa, and Nizar Habash. 2022. Morphosyntactic tagging with pre-trained language models for Arabic and its dialects. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1708–1719, Dublin, Ireland. Association for Computational Linguistics.

Christian Khairallah, Salam Khalifa, Reham Marzouk, Mayar Nassar, and Nizar Habash. 2024. Camel morph MSA: A large-scale open-source morphological analyzer for Modern Standard Arabic. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2683–2691, Torino, Italia. ELRA and ICCL.

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.

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.

Mohamed Maamouri and Ann Bies. 2004. Developing an Arabic Treebank: Methods, Guidelines, Procedures, and Tools. In *Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages (CAASL)*, pages 2–9, Geneva, Switzerland.

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.

Joakim Nivre, Daniel Zeman, Filip Ginter, and Francis Tyers. 2017. Universal Dependencies. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, Valencia, Spain. Association for Computational Linguistics.

Ossama Obeid, Nasser Zalmout, Salam Khalifa, Dima Taji, Mai Oudah, Bashar Alhafni, Go Inoue, Fadhl Eryani, Alexander Erdmann, and Nizar Habash. 2020a. CAMeL tools: An open source python toolkit for Arabic natural language processing. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 7022–7032, Marseille, France. European Language Resources Association.

Ossama Obeid, Nasser Zalmout, Salam Khalifa, Dima Taji, Mai Oudah, Bashar Alhafni, Go Inoue, Fadhl Eryani, Alexander Erdmann, and Nizar Habash. 2020b. CAMeL tools: An open source python toolkit for Arabic natural language processing. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 7022–7032, Marseille, France.

Mostafa Saeed and Nizar Habash. 2025. Lemmatization as a classification task: Results from Arabic across multiple genres. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 30014–30029, Suzhou, China. Association for Computational Linguistics.

Majdi Sawalha, Faisal Alshargi, Abdallah AlShdaifat, Sane Yagi, and Mohammad A. Qudah. 2019. Construction and annotation of the Jordan comprehensive contemporary Arabic corpus (JCCA). In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 148–157, Florence, Italy. Association for Computational Linguistics.

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

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

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).

Dima Taji, Nizar Habash, and Daniel Zeman. 2017. Universal Dependencies for Arabic. In *Proceedings of the Third Arabic Natural Language Processing Workshop*, pages 166–176, Valencia, Spain. Association for Computational Linguistics.

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*.

Waleed A. Yousef, Omar M. Ibrahime, Taha M. Madbouly, and Moustafa A. Mahmoud. 2019. Learning meters of arabic and english poems with recurrent neural networks: a step forward for language understanding and synthesis. *arXiv preprint arXiv:1905.05700*.

Wajdi Zaghouani. 2014. Critical survey of the freely available arabic corpora. In *Workshop on Free/Open-Source Arabic Corpora and Corpora Processing Tools Workshop Programme*, page 1.

A. BAREC-10M Corpus Details

This section provides details of the sub-corpora used in BAREC-10M. They follow the same order as table 3.

ChatGPT A total of 450 sentences generated by ChatGPT[19] as synthetic children's materials.

Green Library 61 manually typed books from the Green Library for children.[20]

Spacetoon Theme songs from 56 animated children's series on Spacetoon channel.

Majed 10 manually typed editions of the *Majed* magazine for children (1983–2019).[21]

Collection of Children's Poems (Other) Includes poems such as *My Language Sings* (لغتي تغني), *Poetry and News* (أشعار وأخبار), and *The Cat and the Eid's Hat* (القطة وقبعة العيد) (Al-Safadi, 2005; Taha-Thomure, 2007).

DARES The full DARESv1 corpus (El-Haj et al., 2024), comprising Saudi school materials spanning 38 subjects across 12 grades.

ArabicMMLU A subset of 6,093 MCQs from the ArabicMMLU benchmark (Koto et al., 2024), selected across different levels, subjects, and countries.

Basic Travel Expressions Corpus (BTEC) The Modern Standard Arabic (MSA) translation of the Basic Travel Expressions Corpus (Eck and Hori, 2005; Takezawa et al., 2007). This subset is selected from the Camel Treebank (Habash et al., 2022).

Zayed Arabic-English Bilingual Undergraduate Corpus (ZAEBUC) 100 student-written essays from the ZAEBUC corpus (Habash and Palfreyman, 2022). This subset is selected from the Camel Treebank (Habash et al., 2022).

Arabic Learner Corpus 20 L2 articles from the Arabic Learner Corpus (Alfaifi, 2015).

Sara The first 1,000 words of *Sara*, a novel by Al-Akkad (1938) (Al-Akkad, 1938), extracted from Hindawi.[22] This subset is selected from the Camel Treebank (Habash et al., 2022).

ANERCorp The full ANERCorp (Benajiba et al., 2007) corpus using the data splits provided by Obeid et al. (2020b).

ReadMe++ The Arabic split of the ReadMe++ dataset (Naous et al., 2024), containing sentences annotated with CEFR levels.

Subtitles A subset of the Arabic side of the OpenSubtitles dataset (Lison and Tiedemann, 2016). This subset is selected from the Arabic Parallel Gender Corpus v2 (Alhafni et al., 2022).

Wikipedia A subset of 175 Arabic Wikipedia articles covering diverse fields such as culture, history, science, and philosophy.[23]

AlKhaleej The first 350 files from each of seven news categories in the AlKhaleej News[24] subset of the SANAD dataset (Einea et al., 2019).

AlArabiya The first 350 files from each of six news categories in the AlArabiya News[25] subset of the SANAD dataset (Einea et al., 2019).

Akhbarona The first 300 files from each of six news categories in the Akhbarona News[26] subset of the SANAD dataset (Einea et al., 2019).

Wikinews 62 Arabic WikiNews articles across multiple categories, including politics, economics, health, science and technology, sports, arts, and culture (Abdelali et al., 2016). This subset is selected from the Camel Treebank (Habash et al., 2022).

UN The Arabic version of the Universal Declaration of Human Rights.[27]

Arabic Poem Comprehensive Dataset (Arab Poem CD) The first 500 words of 1,041 poems from the APCD dataset (Yousef et al., 2019).

[19] https://chatgpt.com/

[20] https://archive.org/details/201409_201409

[21] https://archive.org/details/majid_magazine

[22] https://www.hindawi.org/books/72707304/

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

[24] https://www.alkhaleej.ae

[25] https://www.alarabiya.net

[26] https://www.akhbarona.com

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

Arabian Nights (1001 Nights) The opening narrative and the first eight nights from *The Arabian Nights* (Unknown, 12th century), extracted from an online forum.[28] This subset is selected from the Camel Treebank (Habash et al., 2022).

Hanging Odes The complete text of the ten celebrated pre-Islamic poems (المعلقات), obtained from Wikipedia.[29] This subset is selected from the Camel Treebank (Habash et al., 2022).

Hayy ibn Yaqdhan (Hayy) The full philosophical novel by Ibn Tufail (Tufail, 1150), extracted from the Hindawi Foundation.[30] This subset is selected from the Camel Treebank (Habash et al., 2022).

Hadith The first 134 Hadiths from *Sahih Bukhari* (al Bukhari, 846), sourced from the LK Hadith Corpus[31] (Altammami et al., 2019). This subset is selected from the Camel Treebank (Habash et al., 2022).

Quran The first three Surahs and the last fourteen Surahs from the Holy Quran, obtained from the Quran Corpus Project (Dukes et al., 2013).[32] This subset is selected from the Camel Treebank (Habash et al., 2022).

New Testament The first 16 chapters of the Book of Matthew (Smith and Van Dyck, 1860).[33] This subset is selected from the Camel Treebank (Habash et al., 2022).

Old Testament The first 20 chapters of the Book of Genesis (Smith and Van Dyck, 1865).[33] This subset is selected from the Camel Treebank (Habash et al., 2022).

Constitutions The first 2,000 words of Arabic constitutions from 17 Arab countries, collected from the MCWC dataset (El-Haj and Ezzini, 2024).

MedArabiQ The full MedArabiQ benchmark (Daoud et al., 2025), containing 34K medical MCQs across 33 domains and five academic years.

The Five Adventurers (5-Adventurers) The first 500 words of 187 editions of the *Five Adventurers* magazine (المغامرون الخمسة), manually typed from scanned copies.[34]

Superman 10 manually typed editions of the *Superman* comic series for children (10th to 100th editions).[35]

Batman 10 manually typed editions of the *Batman* comic series for children (10th to 100th editions).[36]

Arab Youth Publisher 74 manually typed books from the Arab Youth Publisher (دار الفتى العربي), digitized from scanned copies available on the Internet Archive.

Lulu 10 manually typed editions of the *Lulu* magazine for children (10th to 100th editions).[37]

Ladybird The first 500 words from 35 manually typed editions of Ladybird children's books.[38]

Reading Corner 37 manually typed books from the *Reading Corner* series (سلسلة زاوية القرائة).

Tunisian Curriculum Full Arabic language textbooks from grades 2 to 5 of the Tunisian curriculum, manually typed.

Emirati Curriculum Subsets of the Arabic Language, Islamic Studies, and Social Studies books across 12 grades. We included the first 500 words from each unit.

Hindawi The first 500 words from 2,758 Hindawi books spanning diverse genres.[39]

Majarra The first 500 words of 50 articles from Majarra.[40] Articles were selected to represent five distinct genres.

[28] http://al-nada.eb2a.com/1000lela&lela/

[29] https://ar.wikipedia.org/wiki/المعلقات

[30] https://www.hindawi.org/books/90463596/

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

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

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

[34] https://archive.org/details/Ar_Mistery_01

[35] https://archive.org/details/004_20201219/001/

[36] https://archive.org/details/004_20201219/001/

[37] https://archive.org/details/42_20201222

[38] https://archive.org/details/LadyBird_201808

[39] https://www.hindawi.org/books/categories/

[40] https://majarra.com

Library of Arabic Literature (Lib. Arab Lit.) The first 500 words of 36 books from the Library of Arabic Literature.[41]

National Anthems (Anthems) The full text of national anthems from 19 Arabic-speaking countries.

Kalima The first 500 words of 596 books from the Kalima project.[42]

Ofoq The first 600 words of 827 articles from the *Ofoq* books (2018 - 2024) of the Arab Thought Foundation.[43]

B. Licenses

All resources used in this work were collected and used in accordance with their respective licensing conditions. A portion of the data was obtained from publicly available materials used under fair-use, including sources such as Green Library, Majed, Wikipedia, The Five Adventurers, Superman, Batman, Arab Youth Publisher, Lulu, Ladybird, Reading Corner, Tunis and Emirati curricula, Hindawi, and Spacetoon songs. Additional resources were drawn from openly licensed datasets, including materials distributed through SANAD (e.g., AlKhaleej, AlArabiya, and Akhbarona news), Camel Treebank collections (e.g., religious texts, classical literature, and WikiNews), and other publicly available datasets such as DARES, ArabicMMLU, BTEC, ZAEBUC, ALC, ANERCorp, ReadMe++, APCD, MCWC (Constitutions), MedArabiQ, and Arabic Parallel Gender corpus. We also used very limited content generated with ChatGPT and open resources from the United Nations and National Anthems (Wikipedia and other sources). Finally, several datasets were incorporated through direct agreements with publishers, including Majarra, the Library of Arabic Literature, Kalima, and Ofoq. All resources were used solely for research purposes and in compliance with their respective usage terms. See Table 3 and Appendix A for more details.

[41] https://www.libraryofarabicliterature.org/

[42] https://alc.ae/publications/kalima/

[43] https://arabthought.org/ar/researchcenter/ofoqelectronic