Noor at BAREC Shared Task 2025: A Hybrid Transformer-Feature Architecture for Sentence-level Readability Assessment

Nour Rabih
Mohamed bin Zayed University of Artificial Intelligence
Nour.rabih@mbzuai.ac.ae

Abstract

This paper presents my participation in the Sentence-level Readability Assessment, Strict track of the BAREC Shared Task 2025 (Elmadani et al., 2025a). Building upon prior work that fine-tuned pre-trained transformer models (Elmadani et al., 2025b), this work explores the impact of incorporating a rich set of handcrafted features on readability prediction performance. A total of 51 features were extracted from the BAREC corpus (Elmadani et al., 2025b), including morphological, lexical, and syntactic indicators, leveraging established computational linguistics tools. These features were integrated into a hybrid architecture that combines transformer-based contextual embeddings with dense layers for feature processing. To optimize performance, experiments included freezing strategies and gradual unfreezing, alongside architectural variations with additional classification layers. Among the tested models, the best performance was achieved with MARBERT, reaching a Quadratic Weighted Kappa (QWK) of 80.95% on the test set, and 83.1% on the blind test set.

1 Introduction

Readability assessment aims to determine the ease or difficulty with which a reader can comprehend a given text. In educational contexts, accurate readability prediction supports tasks such as tailoring learning materials to students' proficiency levels, selecting appropriate reading passages, and developing adaptive learning systems. While research in English readability assessment has been extensive, Arabic remains comparatively underexplored, even though it has a rich morphology, diglossic nature, and complex orthographic and syntactic structures, all of which present unique challenges for computational modeling. The Sentence-level Readability Assessment task introduced in the BAREC Shared Task 2025 (Elmadani et al., 2025a) addresses these challenges by focusing on predicting 19 readability levels, based on the Taha/Arabi21 readability framework (Taha, 2017), for isolated Arabic sentences. Sentence-level assessment is inherently more challenging than document-level assessment, as the absence of broader discourse and contextual cues limits the available linguistic signals for prediction. Previous work (Elmadani et al., 2025b), has demonstrated competitive performance using fine-tuned transformer models without incorporating additional features. In this work, we present a hybrid approach that integrates 51 handcrafted linguistic and structural features with transformer-based contextual embeddings. These include counts of specific morphological forms (e.g., dual and plural noun/adjective inflections, broken plurals, verb tense and voice distinctions), syntactic constructions (e.g., nominal and verbal sentence types, complex clausal structures, object presence), functional particles (e.g., negation, prepositions, demonstratives, vocatives), and broader lexical indicators (e.g., unique word count, content word proportion, vocabulary richness measures. We further explore strategies such as layer freezing, gradual unfreezing, and the addition of extra classification layers to enhance performance. The results on both the test and blind test sets demonstrate that the inclusion of complementary features alongside transformer representations can yield improvements over purely transformer-based baselines, though the degree of improvement is model-dependent.

2 Background

The BAREC Shared Task 2025 (Elmadani et al., 2025a) introduced the Sentence-level Readability Assessment challenge for Arabic, designed to promote the development of models capable of fine-grained readability prediction. The task is framed as a multi-class classification problem with 19 dis-

Proceedings of The Third Arabic Natural Language Processing Conference, pages 331–342

crete readability levels. These levels are assigned to individual sentences based on a combination of linguistic, lexical, and pedagogical criteria, enabling precise targeting of reading materials to learner proficiency levels. I participated in the Strict Track where participants are restricted to using only the provided training data without incorporating any external corpora or embeddings. The dataset is split into training, development, test, and blind test sets. Each instance comprises an Arabic sentence and its readability label. The challenge lies in the granularity of the classification (19 levels), the diglossic and morphologically rich nature of Arabic, and the limited contextual cues available at the sentence level.

While readability assessment for English has benefited from decades of research using both handcrafted features and neural models (Sun et al., 2020; Deutsch et al., 2020; Heilman et al., 2007; Petersen and Ostendorf, 2009), Arabic-specific efforts remain comparatively limited. Early efforts relied on textbook corpora and statistical machine learning models (Al-Khalifa and Al-Ajlan, 2010). More recent work has explored both handcrafted linguistic features and modern pretrained language models (PLMs) such as AraBERT (Berrichi et al., 2024). The latest research trends emphasize hybrid approaches that combine traditional rule-based methods with PLMs, leveraging their complementary strengths for improved Arabic readability prediction (Liberato et al., 2024).

The SAMER project has further advanced Arabic readability resources. The SAMER Lexicon (Al Khalil et al., 2020) provides a five-level readability-annotated lexicon of approximately 26K lemmas, covering multiple dialects and achieving high inter-annotator agreement. Building on this, the SAMER Corpus (Alhafni et al., 2024) constitutes the first manually annotated Arabic parallel corpus for text simplification, consisting of around 159K words from 15 Arabic novels, each accompanied by two simplified parallel versions at different readability levels. These resources provide an important foundation for readability and simplification research in Arabic.

In addition, (Hazim et al., 2022) introduced a Google Docs add-on for Arabic word-level readability visualization. The tool integrates the SAMER Lexicon with morphological analysis and Arabic WordNet to highlight difficult words in context and suggest simpler alternatives. This practical interface enables annotators and educators to assess, simplify, and edit text directly within a familiar document editor, thereby making readability resources more accessible and actionable for corpus creation and pedagogical tasks.

3 System Overview

In this paper, the system adopts a hybrid architecture that integrates transformer-based contextual embeddings with handcrafted linguistic features for sentence-level readability prediction in Arabic. The design was motivated by the need to capture both deep semantic representations and explicit linguistic signals grounded in the BAREC annotation framework (Habash et al., 2025).

Five pre-trained models were experimented with: MARBERT , MARBERTv2 (Abdul-Mageed et al., 2021), AraBERTv2, AraBERTv02 (Antoun et al., 2020), and CamelBERT-MSA (Inoue et al., 2021). These models were selected for their strong performance on Arabic NLP tasks and their coverage of Modern Standard Arabic (MSA) and dialectal variants. For each model, the final hidden state of the [CLS] token, was extracted as the sentence representation.

3.1 Features

To complement the transformer embeddings, we engineered 51 handcrafted features inspired by the BAREC guidelines (Habash et al., 2025). These include:

- **Morphological features:** counts of prefixes, suffixes, verb tenses, plural forms, passive voice, etc.
- **Syntactic features:** dependency-based indicators such as presence of nominal sentences, verbal sentences with/without objects, vocatives, preposed predicates, and coordination structures.
- **Word/syllable counts:** normalized counts of unique words and syllables, leveraging diacritized forms for accuracy.
- **Vocabulary-based features:** sentence-level lexical difficulty scores derived from a lemma–POS vocabulary dictionary, augmented with the SAMER (Al Khalil et al., 2020) and dialect-sensitive markers.
- **Content-based features:** estimated idea/conceptual difficulty levels (ranging

from concrete to symbolic/abstract), obtained by fine-tuning a sentence-level AraBERT (Antoun et al., 2020) classifier.

Full details of the features and their extraction process are provided in the appendix A, but a summary of the methodology is given here. Feature extraction combined a range of resources, including CAMeL Tools (Obeid et al., 2020), regular expressions, external lexicons, and custom Python scripts. Morphological features were obtained using the CAMeL Tools morphological disambiguator, which decomposed tokens into base forms and affixes. This enabled systematic counting of prefixes, suffixes, and clitics at the sentence level, as well as identifying verb tense and voice distinctions such as active versus passive forms. Broken plurals and feminine plurals were similarly detected through combinations of morphological tags, following the rules outlined in Table 4 in the appendix. Syntactic features were extracted from dependency parses generated by CamelParser (Elshabrawy et al., 2023). Each sentence was transformed into a syntactic tree, from which binary indicators were derived to mark the presence of grammatical phenomena such as nominal versus verbal sentences, vocatives, and coordination structures. This rule-based approach ensured that subtle markers of syntactic complexity were systematically encoded, as summarized in Table 5 in the appendix.

Content-based features followed the BAREC framework, which defines eight levels of conceptual difficulty from concrete ideas to abstract or symbolic knowledge. A sentence-level AraBERT (Antoun et al., 2020) classifier was fine-tuned to predict these levels, which were then used as categorical features. Vocabulary-based features were derived through a multi-step pipeline aimed at quantifying lexical difficulty. First, a lemma–POS dictionary was constructed from the BAREC training set by tracking the distribution of each pair across all 19 readability levels. To account for noise and rare outliers, three thresholding strategies were evaluated (strict, relaxed-1%, and relaxed-2%), where the relaxed-1% variant provided the best balance between robustness and sensitivity. This dictionary allowed each sentence to be assigned a vocabulary score based on the most advanced lemma–POS pair it contained. To further strengthen coverage and align with curriculum-based readability scales, the dictionary was enriched with entries from the SAMER lexicon (Al Khalil et al., 2020), which provided additional structured mappings between words and difficulty levels. Together, these methods ensured that the handcrafted features captured complementary dimensions of linguistic complexity-morphological, syntactic, semantic, and lexical-beyond transformer embeddings.

3.2 Hybrid Architecture

The system combines transformer embeddings with feature representations through a dual-branch architecture.

Transformer branch. A BERT encoder produces contextual embeddings for the input sentence.

Feature branch. Handcrafted features $\mathbf{f} \in R^d$ (where $d = 51$) are processed through a Multi-Layer Perceptron (MLP) with batch normalization and ReLU activations.

Fusion. The feature representation is concatenated with the transformer [CLS] embedding.

Classification. A linear layer (softmax) maps the fused representation to 19 readability levels.

4 Experimental Setup

4.1 Dataset and Splits

We conduct experiments on the BAREC Shared Task 2025 dataset (Elmadani et al., 2025b) , which provides labeled sentences for sentence-level readability assessment. Following the official setup, we use the train, development (dev), and test splits provided. Additionally, we evaluate the blind test set, which contains hidden labels released only for the final submission phase.

4.2 Preprocessing

We integrated both raw text and handcrafted features into our pipeline. For each split, we merged the sentence text with 51 extracted linguistic features. One-hot encoding was done on categorical features such as content and vocabulary related features. Finally, labels were shifted to a 0–18 range for compatibility with PyTorch's classification layer.

4.3 Training Setup

We experiment with five transformer models: **MARBERT, MARBERTv2, CamelBERT-MSA,**

Model	Development Set							Test Set						
	Acc19	±1	Dist	QWK	Acc7	Acc5	Acc3	Acc19	±1	Dist	QWK	Acc7	Acc5	Acc3
CamelBERT-MSA	46.36	61.60	1.47	72.97	56.81	62.80	70.18	48.39	64.27	1.35	75.37	58.81	63.71	71.21
MARBERTv2	52.64	67.91	1.23	78.40	62.38	67.10	73.80	53.23	68.49	1.16	80.06	62.65	66.98	73.24
AraBERTv02	45.31	60.82	1.58	67.22	55.75	62.04	68.91	46.73	63.01	1.48	68.80	56.82	62.15	69.13
AraBERTv2	43.42	59.86	1.53	72.37	53.98	60.94	68.00	44.99	62.65	1.41	74.29	55.56	61.05	68.76
MARBERT	**54.69**	**69.28**	**1.19**	**79.38**	**63.93**	**68.44**	**75.08**	**54.45**	**69.75**	**1.11**	**80.95**	**63.93**	**67.99**	**74.11**

Table 1: Sentence-level readability results on BAREC (Dev/Test). Best per column in **bold**.

	Acc19	±1	Dist	QWK	Acc7	Acc5	Acc3
Blind Set (submitted system)	56.10	72.50	1.00	83.10	67.00	70.50	75.80

Table 2: Official hidden-set results of our submission. Acc19 = exact 19-class accuracy; ±1 = adjacent accuracy; Dist = mean absolute distance.

AraBERTv2, and AraBERTv02. We use the Hugging Face Transformers library for model initialization and PyTorch for training. Tokenization is performed with the respective model's pretrained tokenizer, truncating or padding sequences to a fixed maximum length. Models were trained using AdamW (lr=2e-5), batch size 16, for 6 epochs with Cross-Entropy loss, linear warmup/decay scheduling, and 0.3 dropout on an NVIDIA CUDA-enabled GPU. To improve generalization, we adopt a gradual unfreezing strategy: BERT embeddings are frozen at the start, with the last 4 layers unfrozen after epoch 1 and the full encoder unfrozen after epoch 2. Early stopping with patience 3 is applied based on validation QWK.

4.4 Evaluation Metrics

Readability assessment is treated as an ordinal classification task. We adopt the official metrics of the shared task:

- **Quadratic Weighted Kappa (QWK)** – primary metric, penalizing larger misclassifications more heavily.
- **Accuracy (Acc19/Acc7/Acc5/Acc3)** – classification accuracy at different granularities (collapsing 19 labels into 7, 5, or 3 bins), as show in table 3.
- **Adjacent Accuracy (±1 Acc19)** – off-by-1 tolerance measure.

Granularity	Group	BAREC Levels (1-19)
Acc3	1	1–11
	2	12–13
	3	14–19
Acc5	1	1–7
	2	8–11
	3	12–13
	4	14–15
	5	16–19
Acc7	1	1
	2	2–5
	3	6–8
	4	9–10
	5	11–13
	6	14–15
	7	16–19

Table 3: Coarse-grained groupings of the 19 BAREC readability levels used to compute Acc3, Acc5, and Acc7.

5 Results

Table 1 reports the performance of the hybrid architecture on all five pretrained models on the development and test splits of the BAREC Shared Task 2025. Overall, MARBERT achieved the strongest performance, reaching a QWK of 79.38% on the dev set and 80.95% on the test set. It also achieved the lowest average distance (1.11) and the highest exact accuracy (54.45%), confirming its robustness for fine-grained sentence-level readability assessment. MARBERTv2 followed closely, with a test QWK of 80.06%, suggesting that both MARBERT variants are particularly well-suited for the task.

We compared the hybrid models to text-only baselines for each pretrained model from (**?**) . The feature branch produced consistent improvements only with MARBERT (best QWK and lowest dis-

tance), whereas CamelBERT-MSA and AraBERT (v2/v02) showed very similar scores with and without features across Acc[19], ±1 Acc, Dist, and QWK. This indicates that the benefit of feature–text fusion is model-dependent rather than universal, and that strong PLM representations can already capture much of the signal for some encoders.

Blind set (official leaderboard). On the blind set used for the leaderboard, our submitted system achieved the results in Table 2. This placed us **9th on the Strict Path**.

6 Conclusion

This work presented a hybrid transformer–feature architecture for sentence-level Arabic readability assessment in the context of the BAREC 2025 Shared Task. By integrating 51 handcrafted linguistic, syntactic, morphological, and lexical features with contextual embeddings from pretrained Arabic language models, the system sought to capture complementary signals for fine-grained readability classification across 19 levels. Experimental results highlighted that MARBERT delivered the strongest performance, achieving a QWK of 80.95% on the test set and 83.1% on the hidden leaderboard, underscoring its robustness for handling sentence-level complexity in Arabic. The findings demonstrate that while transformer-based models alone provide strong baselines, combining them with structured linguistic indicators can further enhance performance, though the degree of improvement is model-dependent. This work contributes valuable insights into how feature engineering and representation learning can be jointly leveraged for readability modeling in morphologically rich and diglossic languages like Arabic. Future research may focus on incorporating dialectal diversity, enriching the dataset with larger and more varied corpora, and further engineering linguistic features to capture nuanced aspects of Arabic sentence complexity.

7 Limitations

This work is constrained by several limitations that restrict the scope and generalizability of its findings. First, the observed benefits of combining handcrafted linguistic features with transformer-based embeddings appear to be model-dependent. Improvements were most notable with MARBERT, while other pretrained models showed less consistent gains. This raises questions about the robustness and generalizability of the hybrid approach, suggesting the need for broader experimentation across architectures and domains. Second, the current setup focuses on sentence-level readability assessment, which inherently overlooks discourse-level context. Cohesion, coherence, and pragmatic cues that extend beyond individual sentences are often crucial for determining text difficulty, and their absence limits the granularity of prediction. Third, although the handcrafted features were carefully designed to reflect BAREC annotation guidelines, they rely on rule-based extraction pipelines that may introduce errors or fail to capture more nuanced aspects of Arabic syntax, morphology, and dialectal variation. These constraints highlight the need for richer and more diverse datasets and the development of adaptive, data-driven feature engineering techniques. Addressing these challenges will be essential for advancing the accuracy, and real-world applicability of Arabic readability assessment systems.

Ethics Statement

While the author is affiliated with institutions linked to the shared-task organizers, no organizer was involved in the design, development, or evaluation of the systems presented in this paper. The work was conducted exclusively using the resources and information publicly released as part of the shared task, without any privileged access or special guidance. This statement is provided to clarify that no conflict of interest has influenced the reported results.

References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021. ARBERT & MARBERT: Deep bidirectional transformers for Arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7088–7105, Online. Association for Computational Linguistics.

Hend S Al-Khalifa and Amani A Al-Ajlan. 2010. Automatic readability measurements of the arabic text: An exploratory study. *Arabian Journal for Science and Engineering*, 35(2 C):103–124.

Muhamed Al Khalil, Nizar Habash, and Zhengyang Jiang. 2020. A large-scale leveled readability lexicon for Standard Arabic. In *Proceedings of the*

Twelfth Language Resources and Evaluation Conference, pages 3053–3062, Marseille, France. European Language Resources Association.

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.

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.

Safae Berrichi, Naoual Nassiri, Azzeddine Mazroui, and Abdelhak Lakhouaja. 2024. Exploring the impact of deep learning techniques on evaluating arabic l1 readability. In *Artificial Intelligence, Data Science and Applications*, pages 1–7, Cham. Springer Nature Switzerland.

Tovly Deutsch, Masoud Jasbi, and Stuart Shieber. 2020. Linguistic features for readability assessment. *arXiv preprint arXiv:2006.00377*.

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*, 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 The First Arabic Natural Language Processing Conference (ArabicNLP 2023)*.

Nizar Habash, Hanada Taha-Thomure, Khalid 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.

Reem Hazim, Hind Saddiki, Bashar Alhafni, Muhamed Al Khalil, and Nizar Habash. 2022. Arabic word-level readability visualization for assisted text simplification. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 242–249, Abu Dhabi, UAE. Association for Computational Linguistics.

Michael Heilman, Kevyn Collins-Thompson, Jamie Callan, and Maxine Eskenazi. 2007. Combining lexical and grammatical features to improve readability measures for first and second language texts. In *Human language technologies 2007: the conference of the North American chapter of the association for computational linguistics; proceedings of the main conference*, pages 460–467.

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*, Kyiv, Ukraine (Online). Association for Computational Linguistics.

Juan Liberato, Bashar Alhafni, Muhamed Khalil, and Nizar Habash. 2024. Strategies for Arabic readability modeling. In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 55–66, Bangkok, Thailand. Association for Computational Linguistics.

Ossama Obeid, Nasser Zalmout, Salam Khalifa, Dima Taji, Mai Oudah, Bashar Alhafni, Go Inoue, Fadhl Eryani, Alexander Erdmann, and Nizar Habash. 2020. CAMeL tools: An open source python toolkit for Arabic natural language processing. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 7022–7032, Marseille, France. European Language Resources Association.

Sarah E Petersen and Mari Ostendorf. 2009. A machine learning approach to reading level assessment. *Computer speech & language*, 23(1):89–106.

Yuxuan Sun, Keying Chen, Lin Sun, and Chenlu Hu. 2020. Attention-based deep learning model for text readability evaluation. *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

H. Taha. 2017. معايير هنادا طه لتصنيف مستويات النصوص العربية. دار الكتاب التربوي للنشر والتوزيع.

A Appendix

In this work, the Balanced Arabic Readability Evaluation Corpus (BAREC) is used as the primary dataset. The BAREC Annotation Guidelines (Habash et al., 2025) offer a detailed framework for readability annotation, considering six major linguistic dimensions: Spelling/Pronunciation, morphology, syntax, vocabulary, Idea/content, and

word count. To align with this framework, a comprehensive set of linguistic features was engineered, rooted in these dimensions. These features fall under five main categories: morphological, syntactic, word/syllable counts, vocabulary-based, and content-based. Feature selection was guided by the criteria outlined in the BAREC guidelines to reflect the linguistic signals that influence sentence complexity.

Feature extraction was conducted using a combination of CAMeL Tools (Obeid et al., 2020), regular expressions, external lexicons, and custom Python scripts. Below is a breakdown of each feature group and the extraction methods used:

- **Morphological Features**
  - **Number of prefixes, suffixes, and clitics**: Extracted using CAMeL Tools' morphological disambiguator. Each token was decomposed into its base form and affixes, and counts were aggregated per sentence.
  - **Verb tense and voice (e.g., passive, active)**: Identified using the POS tags and morphological features provided by CAMeL Tools.
  - **Use of different forms (broken plurals, feminine plurals)**: Detected using morphological patterns and specific tag combinations (e.g.,singular form and plural num) from CAMeL analysis.

  Table 4 shows the specific rules for all morphological features.

- **Syntactic Features**
  Syntactic complexity plays a key role in determining sentence-level readability in Arabic. To capture this, a set of rule-based syntactic features was developed using dependency parsing outputs using the CamelParser(Elshabrawy et al., 2023).

  A dependency parse was first used to construct syntactic trees for each sentence, allowing for the identification of grammatical relations between words. From these structures, a set of binary features was extracted to reflect the presence or absence of key syntactic phenomena.

  Table 5 shows the specific rules for all Syntactic features.

- **Content-Based Features**
  The BAREC annotation guidelines include a dedicated dimension for evaluating the conceptual and semantic difficulty of a sentence, referred to as the content level. This dimension considers the type of knowledge required for comprehension, the presence of abstract or symbolic ideas, and the cognitive demands placed on the reader. The guidelines define eight content levels, ranging from direct and concrete ideas (e.g., daily life topics requiring no prior knowledge) to highly abstract, symbolic, or culturally nuanced content that assumes specialized background knowledge. To automatically estimate this content complexity, a sentence-level classifier was developed by fine-tuning an AraBERT model. The model was trained to predict one of the eight content levels defined in the guidelines, treating this as a multi-class classification task. These predicted levels were then included as features in the broader feature set used for readability prediction. Table 6 provides a summary of the eight content levels defined in the BAREC framework, along with example indicators used during annotation.

- **word/syllable counts**
  - **Word count**: Computed as the number of unique words in a sentence, ignoring repetitions, or punctuation.
  - **syllable count**: The number of syllables in each word is computed by incorporating morphological and phonetic information. The CAPHI (consonant–vowel pattern) representation, the diacritized form of the word, and morphological prefix annotations are used for a more accurate count of syllables. The CAPHI string is tokenized and scanned for vowel segments, each indicating a potential syllable. Specific linguistic rules are applied to refine the syllable count:
    * The final vowels are excluded if it is a diacritic (حركات الإعراب).
    * Morphological prefixes such as the definite article (ال التعريف) and conjunction (واو عاطفة) are excluded, as they do not contribute to the core syllabic structure of the main word.

ccc

Feature	Feature (Arabic)	Rule
Singular imperfective verb	الفعل المضارع المفرد	num=s, asp=i, pos=verb
Prtoclitic: Definite article Al+	سوابق: ال التعريف	prc0=Al_det
Proclitic: Conjunction wa+	سوابق: واو العطف	prc2=wa_conj
Enclitic: First Person Singular pronoun	لواحق: ضمير المتكلم المفرد المتصل	enc0=1s_pron / 1s_poss / 1s_dobj
Plural imperfective verb	الفعل المضارع الجمع	pos=verb, asp=imp, num=p
Prepositional proclitics	سوابق: حروف جر متصلة	prc1=bi_prep / li_prep / ka_prep
Enclitic: Singular and Plural pronouns	لواحق: ضمير متصل مفرد أو جمع	enc0 in [1p_dobj, ..., 3p_pron]
Dual (in nouns and adjectives)	المثنى في الأسماء والصفات	num=d, pos=noun / adj / noun_quant / adj_comp
Sound feminine plural	جمع المؤنث السالم	form_num=p, form_gen=f, pos=noun / adj
Singular and plural perfective verb	الفعل الماضي المفرد والجمع	pos=verb, asp=p, num=s / p
Sound masculine plural	جمع المذكر السالم	form_gen=m, form_num=p, pos=noun / adj
Dual perfective verb	الفعل الماضي المثنى	asp=p, num=d, pos=verb
Dual imperfective verb	الفعل المضارع المثنى	asp=i, num=d, pos=verb
Singular imperative verb	فعل الأمر المفرد	pos=verb, asp=c, num=s
Enclitics: dual pronoun	لواحق: ضمير المثنى المتصل	enc0=[2d_dobj, ..., 3d_pron]
Broken plurals	جمع التكسير	pos=noun / adj, form_num=s, num=p
Waw of oath	واو القسم	prc2=wa_prep and followed by qassam_lex
Plural imperative verb	فعل الأمر الجمع	asp=c, num=p, pos=v
Conjunctions (e.g., then, until, or...)	أدوات ربط	match of lex
Dual imperative verb	فعل الأمر للمثنى	asp=c, num=d, pos=verb
Ba of oath	باء القسم	prc1=bi_prep, lex in qassam_lex
Passive voice	المبني للمجهول	vox=p, pos=verb
Ta of oath	تاء القسم	prc1=ta_prep, lex in qassam_lex

Table 4: Morphological Features and Rules from BAREC Guidelines

Table 5: Syntactic Features and Rules from BAREC Guidelines

Feature	**Feature (Arabic)**	**Rule**
Nominal sentence	الجملة الاسمية	parent != inna and sisters, has a child with Dependency relation: SBJ (subject), POS tag not equal to VRB
Verbal sentence w/o direct object	جملة فعلية بدون مفعول به	parent = verb, no OBJ Dependency relation
Preposition and object	جار+مجرور	parent pos: PRT, pos=prep in FEATS, has a child with Dependency relation: OBJ
Verbal sentence with one nominal direct object	جملة فعلية مع مفعول به واحد اسم	parent = PRT with pos=prep, has a child with Dependency relation: OBJ
Sentence with two verbs	جملة فيها فعلان	verb count
Verbal sentence with a clausal direct object introduced with Masdar 'an [~to/that]	جملة فعلية مفعولها أن المصدرية	Token is , pos = PRT Has a child: pos = VRB deprel = OBJ asp=i(imperfective)
Verbal sentence with two direct objects	جملة فعلية تتعدى إلى مفعولين	parent pos = VRB Two children with Dependency relation = OBJ
Vocative	المنادى	parent has pos:PRT, FEATS include pos=part_voc Has a child with Dependency relation = OBJ
Inna and its sisters	إن وأخواتها	parent matches the lemma set, has a child with Dependency relation=PRD
Kana and its sisters	كان وأخواته	parent pos=verb, lemma in kana set, has a child with Dependency relation = PRD
Preposed predicate, postponed subject	الخبر المقدم والمبتدأ المؤخر	Dependency relation= SBJ, has a child: pos != VRB, index of parent < index of child
Nominal sentence with a nominal predicate	جملة أسمية خبرها جملة أسمية (فيها مبتدآن)	The sentence does not start with a verb It contains a child node with deprel == "TPC" (topic)
False idafa (tall in stature)	إضافة خيالية (لفظية)	parent pos: NOM, pos=adj in FEATS, Has a child with Dependency relation = IDF
Exception	استثناء	pos '='part$_{r}estrict$

Table 6: Idea / Content Levels in English and Arabic

Idea / Content	فكرة ومحتوى
Direct, explicit, and concrete idea. No symbolism in the text.	فكرة مباشرة وصريحة وحسية. لا رمزية في النص.
Content is from the reader's life. No symbolism in the text.	المحتوى من حياة القارئ. لا رمزية في النص.
Some symbolism, or not everything is stated directly in the sentence.	بعض الرمزية أو عدم التصريح المباشر بكل المقصود في الجملة
Some symbolism that requires the reader to seek help to understand the idea.	بعض الرمزية يحتاج معها القارئ إلى مساعدة من يشرح له المقصود من الفكرة
Some symbolism at the event level in the sentence that the reader understands through prior knowledge.	هناك شيء من الرمزية على مستوى الحدث في الجملة يدركها القارئ بنفسه أو من خلال معارفه السابقة
A degree of symbolism and a need for prior knowledge to understand the meaning of the sentence.	هناك درجة من الرمزية وحاجة للمعرفة السابقة كي يُفهم المقصود من الجملة.
Symbolic ideas and deeper meanings, especially in terms of the psychological dimension of characters/events. Local cultural expressions that may not be understood by those outside the culture.	أفكار رمزية ومعنى باطن خاصة على صعيد البعد النفسي للشخصيات أو الأحداث. تعابير ثقافية محلية قد لا يفهمها من لا يشترك في نفس الثقافة.
Symbolic, abstract, scientific, or poetic ideas that require prior linguistic and cognitive knowledge to understand.	أفكار رمزية، مجردة، عِلمية، أو شعرية وتحتاج إلى معارف لغوية ومعرفية سابقة للبناء عليها لأجل فهمها.

    - In the absence of CAPHI information, syllables are counted by identifying diacritic characters corresponding to short vowels within the diacritized form of the word.

- **Vocabulary-based Features**

  Vocabulary was handled in three different ways to estimate the lexical difficulty of sentences. Firstly, to estimate the vocabulary difficulty of a sentence, a level-based vocabulary scoring system was constructed using the training set of the BAREC dataset. The process began by extracting all lemma–Part of Speech (POS) pairs from the training data using the cameltools disambiguator. For each pair, the number of occurrences was counted across all 19 BAREC readability levels. This allowed for identifying the earliest level at which each lemma–POS pair appeared in the corpus.

  To account for annotation noise or occasional use of advanced vocabulary in lower levels, three variants of vocabulary level assignment were considered:

  - **Strict**: The lowest level at which the lemma–POS pair appeared.
  - **Relaxed (1%)**: The lowest level where the pair appeared, allowing for a 1% error margin of frequency across levels.
  - **Relaxed (2%)**: Similar to the above but with a 2% margin.

  These thresholds introduced flexibility, ensuring that a few early occurrences of complex vocabulary in lower-level sentences did not skew the overall difficulty estimation.

  Once each lemma–POS pair was associated with a level, a vocabulary-level dictionary was constructed containing all lemma–POS pairs from the training data along with their assigned difficulty levels. This dictionary was then used to map vocabulary in the development and test sets. New input sentences were transformed into lists of lemma–POS pairs, and for each sentence, the vocabulary level was defined as the highest (i.e., most difficult) level among all matched pairs. Experiments were conducted using all three threshold variants, and the version yielding - 1% error margin- the best performance was selected for use in the final model.

  In addition to the data-driven vocabulary extracted from the training set, it was observed that expanding the lexical coverage further improved performance. The BAREC annotation guidelines specifically reference certain levels from the SAMER readability lexicon (Al Khalil et al., 2020) as indicative of vocabulary difficulty. To incorporate this, the SAMER lexicon was used to augment the existing vocabulary-level dictionary. Lemma–POS pairs from SAMER were assigned levels in accordance with the BAREC guidelines, thereby enriching the vocabulary feature set with structured, curriculum-aligned information.

  To introduce dialectal sensitivity—also highlighted in the BAREC guidelines—a supplementary lexicon from the BAREC project was utilized. This lexicon consists of approximately 5,000 annotated words, each marked with a dialectal match indicator. Although this represents a relatively small subset of the overall vocabulary, it introduces an important dimension of variation and adds a foundational layer of dialectal awareness to the feature set.

  This layer is particularly valuable because it enables the model to distinguish between vocabulary that overlaps across Modern Standard Arabic and dialects versus vocabulary that exists only in dialectal usage. Words that are common across both MSA and dialects—such as “chair" (كرسي), which appears consistently in both—are typically introduced at earlier reading levels and thus ranked lower in complexity. In contrast, words like “window," which differ in MSA and dialectal forms (e.g., “نافذة" vs. “شباك"), are treated as more complex and are ranked at higher readability levels. Incorporating this information allows the model to better reflect the lexical difficulty that dialectal divergence introduces, especially for learners who are trained primarily on MSA vocabulary.

  In addition to the above features, the barec dataset specifes certain closed groups of vocabs that can be identified using the Cameltools disambiguator, these are shown in table 7.

Table 7: Vocabulary Feature Levels (English and Arabic)

Vocabulary	المفردات
Proper noun Personal pronouns (non-clitics)	اسم علم ضمير منفصل
Singular demonstrative pronoun	اسم الإشارة المفرد
Prepositions	حروف الجر
Dual and plural demonstrative pronoun	اسم اشارة مثنى، جمع
Negation particles	أحرف النفي
Singular relative pronouns	أسماء الوصل المفردة
Dual and plural relative pronouns.	أسماء الوصل المثنى والجمع