Interpretable Arabic Readability Assessment Using Linguistic Rules and Expert-Guided Annotations

Anonymous ACL submission

Abstract

This work focuses on explaining Arabic readability decisions by explicitly modeling the linguistic phenomena underlying text complexity. Building on the Balanced Arabic Readability Evaluation Corpus (BAREC), we introduce BAREC-X, a new dataset of 1,793 sentences annotated with expert-provided linguistic justifications aligned with the BAREC guidelines, enabling direct evaluation of explanation quality. We further propose a fully interpretable, rule-based readability model grounded in linguistically motivated features spanning morphology, syntax, vocabulary, syllabic structure, and content complexity. The model mirrors the BAREC annotation process, produces structured human-readable explanations, and supports both readability level prediction and linguistic reasoning generation. As a final contribution, we present the first reasoning-annotated Arabic readability dataset, achieving an average inter-annotator agreement of 93.3% in terms of at least one shared justification. We also report the first results for an automated Arabic readability reasoning system, which attains 65.8% agreement with human annotators under the same criterion.

1 Introduction

Readability assessment plays a key role in education, content creation, and language learning. While extensively studied in English (Fountas and Pinnell, 2006; Heilman et al., 2008; Vajjala and Lučić, 2018), Arabic readability at fine-grained and interpretable levels remains underexplored. Most existing approaches focus on predicting readability levels (Al-Khalifa and Al-Ajlan, 2010; Nassiri et al., 2018; Elmadani et al., 2025b), often at the expense of transparency. In educational settings, however, understanding why a text is assigned a particular level is as important as the prediction itself. This challenge is especially pronounced for Arabic due to its rich morphology, syntactic variation, and the limited availability of explanation-oriented resources.

Sentence	عالم غريب المعالم world$_{ms}$ strange$_{ms}$ features$_{mp}$ 'A world of strange features'
Readability	Level 11-kaf (Grade 4)
Linguistic Reason	إضافة لفظية False Idafa

Table 1: Readability Assessment Example based on Linguistic Features

This work addresses this gap by introducing an interpretable, feature-based approach to Arabic readability assessment grounded in the Balanced Arabic Readability Evaluation Corpus (BAREC) (Elmadani et al., 2025a). Linguistically motivated features are designed in accordance with the BAREC annotation framework (Habash et al., 2025). It covers morphology, syntax, vocabulary, syllabic structure, and content complexity. The approach explicitly models the linguistic phenomena underlying readability decisions and generates human-readable explanations alongside its predictions, as shown in table 1.

The contributions of this work are as follows:

- We introduce the first reasoning-annotated Arabic readability dataset (BAREC-X) aligned with the BAREC framework, comprising 1,793 sentences with expert-provided linguistic justifications alongside readability labels.
- We propose a fully interpretable, feature-based readability classifier grounded in rule-based linguistic features aligned with the BAREC framework, enabling transparent and linguistically grounded explanations.

The paper is organized as follows. After reviewing related work (§2), we present the linguistic

background and BAREC guidelines (§3), describe the proposed dataset (BAREC-X) and model (§4), and detail the experimental setup (§5). We then report and discuss the results (§6, §7, §8) and conclude in §9.

2 Related Work

Automatic readability assessment has been widely studied in NLP, resulting in a variety of datasets, models, and evaluation frameworks across languages. Recent advances in readability assessment have been driven by deep learning. Neural architectures, including recurrent models and hierarchical attention networks, have shown strong performance by learning representations directly from text (Sun et al., 2020). Transformer-based pretrained language models, such as BERT (Devlin et al., 2019), further improved accuracy by capturing contextualized linguistic information (Deutsch et al., 2020; Liu et al., 2025). Despite their effectiveness, these models are typically black-box systems that provide limited insight into the linguistic factors underlying their predictions.

As readability models became more complex, a critical question emerged: how to explain why a text is assigned a particular readability level. This has motivated research on computational methods to improve the interpretability of readability classifiers. One line of work applies post-hoc explanation methods that treat the underlying model as a black box. Model-agnostic approaches such as LIME (Ribeiro et al., 2016) and SHAP (Lundberg and Lee, 2017) have been used to explain individual predictions by estimating feature importance scores. While these methods can provide local explanations for complex models, their outputs are often difficult for non-expert users to interpret and may not align cleanly with linguistically meaningful categories (Gilpin et al., 2018). Moreover, such explanations are indirect and depend on the behavior of the underlying model rather than reflecting explicit linguistic reasoning.

An alternative direction focuses on incorporating linguistically motivated features directly into readability models. Traditional readability formulas are inherently interpretable but limited in expressiveness, while hybrid approaches combine neural representations with handcrafted linguistic features to balance performance and transparency (Deutsch et al., 2020). Although these methods demonstrate that linguistic features can be beneficial, particularly in low-resource settings, their explanatory capacity is often implicit and rarely evaluated against human-provided reasoning.

Arabic readability research has additional challenges due to the language's rich morphology, flexible word order, and orthographic variation. Machine learning approaches extended Arabic readability modeling by incorporating morphological, lexical, and syntactic features. For example, Al-Khalifa and Al-Ajlan (2010) demonstrated the effectiveness of SVM-based classifiers using handcrafted features, while more recent work showed that pretrained models such as AraBERT (Antoun et al., 2020) achieve higher predictive accuracy (Berrichi et al., 2024; Liberato et al., 2024). Nevertheless, these neural approaches largely prioritize performance and offer limited transparency into the linguistic drivers of readability decisions.

The development of annotated datasets has played a crucial role in advancing Arabic readability research. Existing resources include CEFR-based datasets for Arabic as a foreign language (Khallaf and Sharoff, 2021; Naous et al., 2023) and school-oriented corpora such as DARES (El-Haj et al., 2024). BAREC (Elmadani et al., 2025a) represents the largest and most fine-grained Arabic readability dataset to date, comprising 19 levels spanning early childhood to postgraduate comprehension, with high inter-annotator agreement. BAREC provides detailed linguistic guidelines for annotation (Habash et al., 2025), but its labels do not explicitly encode the reasoning behind each decision.

In contrast to prior work that emphasizes prediction accuracy or relies on post-hoc explanations, this study centers interpretability as a primary objective. We build on BAREC by introducing a reasoning-annotated subset in which expert annotators provide explicit linguistic justifications aligned with the BAREC guidelines. We further propose a fully interpretable, rule-based readability model that mirrors the human annotation process and produces structured, linguistically grounded explanations alongside predictions. By directly evaluating agreement between system-generated explanations and human reasoning, our work establishes a transparent baseline for interpretable Arabic readability assessment.

3 Linguistic Background

This section outlines the linguistic foundations of Arabic readability assessment as defined in BAREC guidelines (Habash et al., 2025). BAREC adopts a fine-grained, pedagogically motivated readability framework inspired by Taha-Thomure (2017), in which sentences are assigned to one of 19 ordered readability levels corresponding to progressive educational stages, from early literacy to postgraduate comprehension (Habash et al., 2025). These levels follow the Abjad order of Arabic letters (1-alif, 2-ba, 3-jim, through to 19-qaf) and are designed to provide fine distinctions at lower levels, where readability variation is greatest.

Readability in BAREC is modeled as the interaction of multiple linguistic dimensions rather than a single surface-level measure. Each sentence is assigned a readability level based on the most complex linguistic phenomenon it contains, reflecting a 'most complex feature' principle. As a result, the presence of a single high-level feature is often sufficient to determine the final readability level, regardless of simpler features present elsewhere in the sentence.

Annotators evaluate sentences across six core linguistic dimensions: spelling and phonology, word count, morphology, syntax, vocabulary, and content. These dimensions jointly define the progression of difficulty across the 19 levels, with different dimensions becoming dominant at different stages of readability development. The same dimensions guide both the annotation of our newly introduced dataset and the design of the interpretable models proposed in this work.

- **Spelling and Phonology:** Word length and syllabic structure influence difficulty, with more complex syllable patterns associated with higher levels. Diacritics are not considered. This dimension primarily constrains readability up to Level 7.
- **Word Count:** The number of unique words serves as an upper bound on readability at early levels and is mainly relevant up to Level 11.
- **Morphology:** Morphological complexity reflects inflectional and derivational patterns, including clitics, verb forms, plurality, and less frequent constructions. Morphology influences readability up to approximately Level 13.
- **Syntax:** Sentence structure and syntactic relations contribute to difficulty, with non-canonical and complex constructions appearing at higher levels. Syntactic complexity becomes a key determinant up to Level 15.
- **Vocabulary:** Lexical difficulty is driven by word familiarity and usage, with higher levels introducing rarer, technical, or classical vocabulary. Vocabulary spans the full readability range up to Level 19.
- **Content:** Conceptual difficulty reflects the level of abstraction and required prior knowledge, ranging from concrete everyday concepts to abstract or domain-specific content. Content applies across all levels and extends to Level 19.

The complete BAREC annotation guidelines are provided in Appendix A.

4 Methodology

Our methodology combines data-centric annotation with interpretable modeling approaches for Arabic readability assessment. Building on the Balanced Arabic Readability Evaluation Corpus (BAREC), we introduce a newly annotated reasoning-aware subset, BAREC-X, in which sentences are labeled with both readability levels and explicit linguistic justifications aligned with the BAREC guidelines. This dataset enables the direct study and evaluation of interpretability in readability modeling.

4.1 Reasoning-Annotated Readability Dataset

While BAREC provides a strong foundation for sentence-level readability classification, its annotations are limited to readability labels and do not explicitly capture the linguistic reasoning behind each decision. To support interpretable readability modeling, we introduce BAREC-X, a newly annotated subset of 1,793 sentences derived from BAREC. These sentences were selected from the inter-annotator agreement (IAA) portion of the corpus to ensure high-quality readability labels.

Three annotators, native Arabic-speaking educators with prior experience in Arabic readability annotation from a relevant project, were assigned approximately 800 sentences in total, with around 300 sentences shared across all annotators to enable the measurement of inter-annotator agreement on interpretive reasoning.The data was organized into 8 batches of approximately 100 sentences to

facilitate the annotation process. The exact sentence counts per batch and their distribution across readability levels are provided in Appendix A.3. Each annotator was given 5 unique batches and 3 batches were shared among all 3 annotators. They were asked to provide explicit linguistic justifications for each sentence underlying the given level from the BAREC corpus. Selecting from a fixed set of categories defined by the BAREC guidelines, each sentence may be associated with multiple linguistic reasons, reflecting the presence of different factors contributing to readability. An example of the annotation process sheet used by annotators, along with the structured output format, is provided in Appendix A.3. This annotation process results in aligned sentence-level readability labels and human-provided explanations, enabling direct evaluation of model interpretability.

4.2 Feature-Based Readability Model

We propose a fully interpretable, feature-based readability model designed to mirror the decision process followed by human annotators under the BAREC framework. Rather than learning implicit patterns from data, the model relies exclusively on handcrafted linguistic features and deterministic rules to assign readability levels and generate explicit explanations. The model supports two modes of operation: (i) as a readability classifier, where a sentence is provided as input and the model predicts a readability level together with its underlying linguistic justifications, and (ii) as an explanatory system, where a sentence and a specified readability level are provided, and the model outputs the linguistic reasons associated with that level.

Extraction of Linguistic Features Linguistic features corresponding to the dimensions described in Section 3 are extracted using a combination of morphological analysis with CAMeL Tools[1] (Obeid et al., 2020), syntactic parsing with CamelParser[2] (Elshabrawy et al., 2023) , surface-level measurements, and external lexical resources. In total, the model operates over 66 linguistically motivated features. These include 25 morphological features, 18 syntactic features, and 19 vocabulary-based features, in addition to surface-level features such as word count and syllable count, as well as a vocabulary and content-level feature.

[1] https://github.com/CAMeL-Lab/camel_tools, version 1.2.0

[2] https://github.com/CAMeL-Lab/camel_parser, version 2.0

Morphological features capture inflectional and derivational properties such as affixes, clitics, verb tense, voice, and plural forms. Syntactic features encode the presence of specific dependency structures associated with increased sentence complexity. Surface-level features include word count and syllable count, computed in accordance with the BAREC guidelines. Vocabulary difficulty is modeled through a level-based lexicon constructed from the BAREC training data, where lemma-POS pairs are associated with the earliest readability level at which they appear. To mitigate annotation noise, relaxed frequency thresholds are applied when assigning vocabulary levels. This lexicon is further augmented using the SAMER readability lexicon (Al Khalil et al., 2020) and a supplementary BAREC dialectal lexicon, enabling the model to account for curriculum-aligned vocabulary difficulty and lexical variation between Modern Standard Arabic and dialects. Detailed feature definitions and extraction rules are provided in Appendix B.

Readability Modeling In the first mode of operation, where the model functions as a readability classifier, extracted reasons for each sentence are mapped to their associated readability levels, with each activated reason contributing a candidate level. The final readability prediction is determined using a max-based aggregation strategy, whereby the sentence is assigned the highest level triggered by any reason. Table 2 illustrates this reason competition mechanism by showing an example sentence with its activated reasons and corresponding readability levels; only triggered reasons are considered, and the highest assigned level defines the final prediction. In addition to predicting a readability label, the model generates a structured set of linguistic reasons corresponding to the activated features. These reasons constitute the model's explanation output, ensuring full transparency and traceability between predictions and linguistic evidence.

In the second mode of operation, where the model is used as an explanatory system, the model restricts its output to the subset of extracted reasons whose associated levels match the given input level, returning only the linguistic features that justify that level for the sentence. When no explicit linguistic features are activated at the specified level, the model falls back to vocabulary and content-based features to provide a minimal yet informative explanation.

Category	Activated Reason	Assigned Level
WC	unique word count = 3	2
SP	syllable count = 3	2
M	definite article prefix (Al الـ)	3
	broken plural	7
V	adjective	2
	vocabulary difficulty	5
S	noun + adjective	2
	false idafa	**11**

Table 2: Example of feature competition for 'عالم غريب المعالم'. Only activated features are shown. WC=Word Count, SP=Spelling, M=Morphology, V=Vocabulary, S=Syntax.

Dataset	Train	Dev	Test
BAREC	54,845	7,310	7,286
BAREC-X	1,409	196	188

Table 3: Dataset statistics for BAREC and the reasoning-annotated subset BAREC-X across train, development, and test splits.

5 Experimental setup

5.1 Data Splits

The readability prediction experiment uses the official data split provided with the BAREC dataset. To evaluate the explanatory system, we augment the IAA portion of BAREC with reasoning annotations, resulting in BAREC-X (see section 4.1). Since the IAA portion spans across `Train`, `Dev` and `Test` splits, BAREC-X follows the same splits although we did not train the model for this task explicitly. In this paper, we report on the three splits separately. However, for future work that would use this data for training, we recommend following the splits shown in table 3.

5.2 Evaluation Metrics

We evaluate interpretability from two complementary perspectives: **Reason+Category (RC)** and **Category-only (C)**. The RC setting assesses whether a model correctly identifies the *exact linguistic reasons* within each category (e.g., specific morphological or syntactic phenomena) provided by human annotators. In contrast, the C setting evaluates performance at a coarser level by considering only the broader linguistic category (e.g., morphology, syntax, vocabulary), regardless of the specific reason selected within that category.

For both evaluation settings, we report three metrics to capture different aspects of alignment with human-provided explanations. First, **Exact Match** measures whether the model's predicted set of reasons exactly matches the annotator-provided set. Second, **Jaccard Similarity** is used to quantify the degree of overlap between predicted and gold sets, providing a graded measure of interpretability rather than a binary outcome. Finally, **At Least One** measures whether the model correctly predicts at least one gold reason, capturing partial but meaningful interpretive overlap.

We adopt the official evaluation metrics defined by BAREC to assess readability classification performance:

- **Quadratic Weighted Kappa (QWK)** - the primary evaluation metric, which penalizes larger misclassifications more heavily (Cohen, 1968; Doewes et al., 2023).
- **Accuracy ($\mathbf{Acc}^{19}$ / $\mathbf{Acc}^{7}$ / $\mathbf{Acc}^{5}$ / $\mathbf{Acc}^{3}$)** - classification accuracy computed at different levels of granularity by collapsing the original 19 readability labels into 7, 5, or 3 categories.
- **Adjacent Accuracy ($\pm\mathbf{1\ Acc}^{19}$)** - an off-by-one tolerance measure that counts predictions within one readability level of the gold label as correct.
- **Average Distance** - the average absolute difference between two sets of labels.

5.3 Model Architecture and Hyperparameters

The feature-based model does not require traditional training for most of its components, particularly for morphological and syntactic features. Each linguistic pattern is explicitly defined by a unique code and linked to one or more readability levels according to the BAREC guidelines; these patterns are treated as discrete codes within the system. During inference, all relevant features are extracted from a sentence, and each activated code contributes its associated readability level. A complete list of linguistic codes and their corresponding readability levels is provided in appendix B.1.

In its default configuration, the model assigns a final readability level using a max-based aggregation strategy, selecting the level associated with the most complex activated feature. In addition, the same feature extraction pipeline can be used in an explanatory mode, where the model filters the activated features to return only those corresponding

to a specified readability level. This design enables both forward prediction and targeted explanation using a shared, fully transparent mechanism.

Figure 1(a) shows the pipeline used for readability level prediction, while Figure 1(b) depicts the explanatory mode, in which a readability level is provided as input and the model returns the corresponding linguistic reasons.

For the content-based feature, which captures thematic or conceptual complexity, we rely on a learned approach. We fine-tune a transformer-based model to classify sentences into one of the eight content levels defined by the BAREC guidelines. The model is trained on the training split of the BAREC dataset. We fine-tune AraBERTv02 (Antoun et al., 2020),[3] configured using AdamW (β_1=0.9, β_2=0.999, ϵ=$1e{-}08$) with a learning rate of $5{\times}10^{-5}$, linear learning rate decay, batch sizes of 64. All experiments are evaluated one a single run.

The output of the content-level classifier is treated like any other feature and contributes to the final readability level using the same max-based logic.

6 Results

6.1 Reasoning Dataset IAA Evaluation

To assess the quality and consistency of the reasoning annotations in the newly introduced interpretable subset, we compute inter-annotator agreement (IAA) on the 300 selected sentences. Since each sentence may be associated with multiple reasoning labels, agreement is evaluated using the same interpretability metrics described in Section 5.2. Agreement is computed pairwise across annotators.

We report IAA results under both the **RC** and **C** evaluation settings, capturing agreement at fine-grained and coarse-grained levels of interpretation. The resulting scores provide an empirical measure of annotation reliability and validate the quality of the reasoning labels. Detailed agreement statistics are presented in Table 4.

In addition, we apply a unification procedure to derive a single consolidated set of reasoning labels per sentence. For each sentence, unified reasons are defined as those agreed upon by all three annotators; in cases where no such consensus exists, reasons agreed upon by at least two annotators are retained.

Annotator Pair	Exact		Jaccard		At Least 1	
	RC	C	RC	C	RC	C
A1-A2	68.4	79.5	68.0	78.0	91.3	94.3
A2-A3	82.8	85.5	80.2	84.4	95.3	97.0
A1-A3	70.4	80.8	70.9	79.5	93.3	95.0
Average	73.9	81.9	73.0	80.6	93.3	95.4

Table 4: Pairwise inter-annotator agreement (IAA). All values are reported as percentages(%)

The inter-annotator agreement results demonstrate strong consistency in the reasoning annotations, supporting the reliability of the proposed dataset. Agreement is consistently higher under the coarse-grained C setting than under the finer-grained RC setting, as expected. Substantial Exact Match and Jaccard scores are observed across annotator pairs, while the At Least One metric exceeds 90% in all cases, indicating near-universal agreement on at least one key linguistic factor. These results confirm that the annotation guidelines enable consistent interpretive judgments and that the dataset is suitable for evaluating interpretable readability models.

6.2 Feature-based System Evaluation

When the proposed feature-based model is used for readability level prediction, its performance varies systematically across the readability spectrum. In particular, evaluating the model separately on early-to-intermediate levels (1-11) yields substantially stronger results than treating all 19 levels as a single prediction task. This pattern reflects fundamental differences in how readability is defined within the BAREC framework. Levels 1-11 are governed by fine-grained, linguistically explicit guidelines that rely primarily on surface, morphological, and syntactic features, which are well suited to rule-based modeling. In contrast, higher levels (12-19) depend increasingly on vocabulary rarity and abstract content complexity, which are inherently more subjective and less amenable to deterministic approaches. Motivated by this distinction, we report results both for the full 19-level setting and for a focused evaluation on levels 1-11. As shown in Table 5, performance on the full level range remains moderate, reflecting the challenges posed by higher-level semantic abstraction. However, restricting evaluation to levels 1-11 leads to consistent and substantial im-

[3] https://huggingface.co/aubmindlab/bert-base-arabertv02

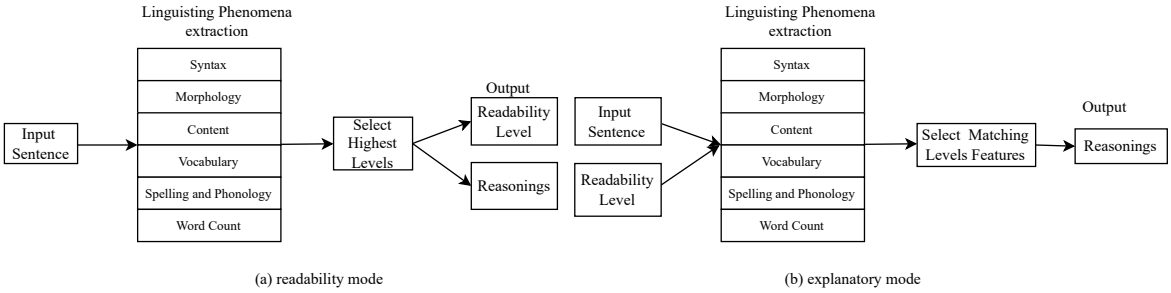

Figure 1: Feature-Based Custom Model Pipeline

	1–19		1–11	
	FBM	**BL**	**FBM**	**BL**
Count	7,310	7,310	3,949	3,949
Acc^{19}	34.9	56.6	58.2	64.1
±1 Acc^{19}	47.9	69.9	73.9	77.0
Avg. Distance	2.2	1.1	1.0	0.9
QWK	48.2	80.0	69.6	73.6
Acc^{7}	42.3	65.9	69.9	73.8
Acc^{5}	49.9	70.3	84.0	82.3
Acc^{3}	58.3	76.5	100.0	91.4

Table 5: BAREC dev set readability prediction performance of the feature-based model (FBM) compared with the BAREC baseline (BL). all metrics are in %

Split	**Exact**		**Jaccard**		**At Least 1**	
	RC	**C**	**RC**	**C**	**RC**	**C**
Train	43.0	59.3	40.7	50.3	65.4	73.0
Dev	36.2	51.5	35.4	49.1	65.8	77.0
Test	37.8	59.6	39.0	53.3	70.7	80.3

Table 6: Agreement between the feature-based system and human annotations across data splits. All values are percentages (%).

provements across all metrics. While the BAREC baseline achieves higher raw accuracy and QWK on levels 1-11, the performance gap is relatively modest. Crucially, the proposed model produces fully transparent, linguistically grounded explanations alongside its predictions, making it particularly well suited for educational and diagnostic applications where interpretability and actionable feedback are primary requirements.

The primary objective of the feature-based system is interpretability rather than maximizing predictive performance. Accordingly, we evaluate the system by comparing its generated explanations against the human-provided reasoning annotations in the BAREC-X subset.

Table 6 reports agreement between the system and human annotators across data splits inherited from BAREC. The development split is not used for model training or tuning in this work and is reported solely for completeness. The model operates in the same inference-only manner across the training, development, and test splits, and exploration of split-specific training or evaluation strategies is left for future work.

Within this evaluation setting, while Exact Match performance is lower in the fine-grained RC condition, agreement improves substantially at the category level, indicating that the system generally identifies the correct linguistic dimension influencing readability. Moreover, the relatively high At Least One scores suggest that the system frequently captures at least one core feature sufficient to justify a given readability level, even when it does not fully align with the annotators' complete set of reasons.

7 Error Analysis

Table 7 presents a category-level analysis of the system's reasoning prediction performance. Substantial variation is observed across linguistic dimensions, reflecting differences in both feature explicitness and model coverage. Categories grounded in surface-level and structurally explicit cues-most notably Word Count and Morphology, achieve the strongest performance, with F1 scores of 70.4% and 68.2%, respectively. Syntactic features achieve high precision (78.8%) but lower recall (40.7%), suggesting that identified constructions are typically correct, though many valid instances remain undetected. Vocabulary features show higher recall (74.3%) than precision (58.5%), indicating broader lexical coverage at the cost of occasional overgeneration, particularly in context-dependent cases. Performance is weakest for Orthography and Con-

Category	Precision (%)	Recall (%)	F1 (%)
Spelling & Phon.	55.8	25.3	34.8
Word Count	61.3	82.6	70.4
Morphology	87.4	55.9	68.2
Syntax	78.8	40.7	53.7
Vocabulary	58.5	74.3	65.4
Content	4.5	27.4	7.7

Table 7: Precision, recall, and F1 scores for reasoning prediction across linguistic categories.

tent. Orthographic features are relatively sparse and contribute less consistently to readability decisions, resulting in low recall (25.3%). Content-level reasoning performs poorest overall (F1 of 7.7%), which is expected given its reliance on semantic abstraction, cultural knowledge, and interpretive judgment-factors that are difficult to capture through deterministic, rule-based features. Overall, the results highlight a clear trade-off between interpretability and semantic coverage. The proposed feature-based system performs best in linguistically explicit categories, while performance degrades for phenomena requiring deeper semantic understanding. This analysis reinforces the model's suitability as an interpretable diagnostic tool for readability assessment and motivates future extensions that incorporate semantic modeling for content-level reasoning.

8 Discussion

This work demonstrates that Arabic readability can be modeled in an interpretable and linguistically faithful manner when annotation, feature design, and evaluation are tightly aligned. The proposed feature-based model performs strongest at early and intermediate readability levels, where BAREC guidelines rely on explicit morphological, syntactic, and surface-level cues. In contrast, performance declines at higher levels that depend more heavily on abstract vocabulary and content interpretation, highlighting the inherent limitations of deterministic rules for modeling semantic complexity. The interpretability evaluation shows that, while exact matching of fine-grained reasons remains challenging, the system reliably identifies the correct linguistic dimension influencing readability. High At Least One agreement indicates that the model frequently captures a core justification aligned with human reasoning, which is particularly relevant for educational and diagnostic use cases where partial but faithful explanations are sufficient.

Finally, the introduction of BAREC-X establishes the first reasoning-annotated Arabic readability dataset, with strong inter-annotator agreement, demonstrating that expert linguistic reasoning can be consistently captured and evaluated. Together, these results position interpretable, feature-driven models as a robust and transparent baseline for readability assessment, and motivate future work on extending semantic coverage through hybrid or learning-based approaches.

9 Conclusions and Future Work

This work presents an interpretable approach to Arabic readability assessment grounded in the BAREC annotation framework. We introduced BAREC-X, the first reasoning-annotated Arabic readability dataset, in which expert annotators provide explicit linguistic justifications, achieving an average inter-annotator agreement of 93.3% under the At Least One criterion. Building on this resource, we developed a fully interpretable rule-based model that mirrors the BAREC annotation process and produces structured, human-readable explanations for its predictions, constituting the first reported results for automated Arabic readability reasoning, with 65.8% agreement with human annotations under the same criterion.

Experimental results demonstrate that the proposed model performs competitively at early and intermediate readability levels (1-11), where linguistic guidelines are most explicit, while also providing strong alignment with human reasoning. These findings underscore the continued relevance of linguistically informed models for educational NLP tasks and establish a solid baseline for interpretable Arabic readability assessment.

Future work includes expanding the feature set to cover additional aspects of the BAREC guidelines that are not fully captured in the current implementation. Further effort will also be devoted to refining and debugging the linguistic tools used in feature extraction to improve accuracy and robustness. Finally, enlarging the reasoning-annotated dataset would enable more comprehensive evaluation and support the development of hybrid or learning-based models that retain interpretability while improving performance at higher readability levels.

Limitations

Despite its strengths, this work has several limitations. First, while the proposed feature-based model is grounded in the BAREC annotation guidelines, it does not yet cover the full spectrum of linguistic phenomena described in the guidelines. Certain fine-grained constructions and rare linguistic patterns are not explicitly modeled, which may affect performance at higher readability levels. Second, the accuracy of the extracted features depends on the reliability of the underlying NLP tools, including morphological disambiguation and dependency parsing. Errors or ambiguities in these tools may propagate to the feature representations and, consequently, to both predictions and explanations. Although we mitigate this by relying on well established resources, further debugging and validation of the feature extraction pipeline are necessary. Third, vocabulary- and content-based features are derived from data-driven resources trained on the BAREC training split. While this design choice is consistent with standard evaluation practices, it limits the interpretability evaluation to development and test subsets and may introduce bias toward the training distribution. Finally, the reasoning-annotated dataset remains relatively modest in size. Although it is sufficient for interpretability analysis and qualitative evaluation, scaling the dataset would enable more robust statistical analysis and support future learning-based explanation models.

Ethical considerations

The annotation process is conducted with transparency and fairness, with multiple annotators involved to mitigate biases and ensure reliability. All annotators are paid fair wages for their contribution.

The annotated dataset will be openly accessible to promote transparency, reproducibility, and collaboration in Arabic language research.

We used AI writing assistance within the scope of "Assistance purely with the language of the paper" described in the ACL Policy on Publication Ethics.

References

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.

Wissam Antoun, Fady Baly, and Hazem Hajj. Arabert: Transformer-based model for arabic language understanding. In *LREC 2020 Workshop Language Resources and Evaluation Conference 11–16 May 2020*, page 9.

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.

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

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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

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.

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

Khalid N Elmadani, Bashar Alhafni, Hanada Taha, and Nizar Habash. 2025b. Barec shared task 2025 on arabic readability assessment. In *Proceedings of The Third Arabic Natural Language Processing Conference: Shared Tasks*, pages 239–252.

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)*.

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

Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pages 80–89. IEEE.

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.

Michael Heilman, Kevyn Collins-Thompson, and Maxine Eskenazi. 2008. An analysis of statistical models and features for reading difficulty prediction. In *Proceedings of the third workshop on innovative use of NLP for building educational applications*, pages 71–79.

Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2021. The interplay of variant, size, and task type in Arabic pre-trained language models. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 92–104, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Nouran Khallaf and Serge Sharoff. 2021. Automatic difficulty classification of arabic sentences. *arXiv preprint arXiv:2103.04386*.

Juan Piñeros Liberato, Bashar Alhafni, Muhamed Al Khalil, and Nizar Habash. 2024. Strategies for arabic readability modeling. *arXiv preprint arXiv:2407.03032*.

Fengkai Liu, Tan Jin, and John SY Lee. 2025. Automatic readability assessment for sentences: neural, hybrid and large language models. *Language Resources and Evaluation*, pages 1–32.

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *CoRR*, abs/1705.07874.

Tarek Naous, Michael J Ryan, Anton Lavrouk, Mohit Chandra, and Wei Xu. 2023. Readme++: Benchmarking multilingual language models for multi-domain readability assessment. *arXiv preprint arXiv:2305.14463*.

Naoual Nassiri, Abdelhak Lakhouaja, and Violetta Cavalli-Sforza. 2018. Modern standard arabic readability prediction. In *Arabic Language Processing: From Theory to Practice*, pages 120–133, Cham. Springer International Publishing.

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.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA. Association for Computing Machinery.

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.

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

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

A BAREC Annotation Guidelines and Annotation Interface

A.1 BAREC Guidelines

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

A.2 English Translation

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

A.3 Annotation Interface

Figure 2 illustrates the Google Sheets interface used for annotation. The first two columns contain the sentence and its assigned readability level. The remaining columns correspond to the possible linguistic reasons underlying the assigned level, as defined in the BAREC annotation guidelines (Figure A.2). Annotators select one or more applicable reasons for each sentence. An additional notes column is provided for flagging ambiguous or problematic cases and for recording supplementary comments.

To facilitate the annotation process, the data was organized into batches of approximately 100 sentences, as summarized in Table 9, which reports the distribution of readability levels across all batches. This batching strategy ensured balanced coverage of levels while keeping annotation units manageable for annotators. Batches were then assigned to three expert annotators following a partially overlapping design, as shown in Table 8. Several batches were intentionally annotated by more than one annotator to enable the measurement of inter-annotator agreement (IAA), while the remaining batches were distributed to ensure full coverage of the dataset. Together, these tables illustrate both the level-wise composition of the annotation batches and the annotation assignment strategy used to support reliable and reproducible reasoning annotations.

B Modeling linguistic Phenomena

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 10 shows the specific rules for all morphological features.

- **Syntactic Features**
  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 11 shows the specific rules for all Syntactic features.

- **Content-Based Features**

  To automatically estimate this content complexity, a sentence-level classifier was developed by fine-tuning an AraBERT model (Antoun et al.). 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.

- **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.
* 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. 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 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

A1	A2	A3
1	7	13
2	2	2
3	8	14
9	9	9
4	10	15
16	16	16
5	11	17
6	12	18

Table 8: Assignment of annotation batches to annotators A1, A2, and A3.

Cameltools disambiguator, these are shown in table 12.

B.1 Linguistic codes and their associated readability levels

Table 13 presents the complete set of linguistic features defined in the BAREC framework and used by the proposed feature-based model. Each feature is represented by a unique code and is associated with a specific readability level, reflecting the guideline followed by human annotators. The table serves as a reference for the deterministic mapping between linguistic phenomena and readability levels employed by the system.

C License

We list below the licenses of the data and tools used in this work, all of which are employed in accordance with their intended use.

- SAMER Lexicon (Al Khalil et al., 2020): Non Commercial Use - NYUAD License.[4]
- BAREC Corpus (Elmadani et al., 2025a): Creative Commons Attribution Share Alike 4.0
- CAMeL Tools (Obeid et al., 2020) and CAMeLBERT (Inoue et al., 2021): MIT License

[4] https://camel.abudhabi.nyu.edu/samer-readability-lexicon/

Figure 2: Annotation interface

Batch/Level	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	Total
1	1	0	1	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	4
2	1	0	1	0	1	1	1	1	1	1	1	1	1	0	0	1	0	0	12
3	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	1	1	34
4	1	1	1	1	0	0	1	1	1	1	0	0	0	0	0	0	1	1	10
5	2	3	2	3	2	3	2	3	2	3	3	3	3	3	3	3	3	3	49
6	3	3	4	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	55
7	7	7	7	6	8	7	7	7	8	7	8	8	8	8	8	8	8	7	134
8	7	7	7	7	7	7	7	7	7	7	7	7	8	8	8	8	7	7	130
9	3	3	3	3	4	4	3	3	4	3	3	3	3	3	3	3	5	5	61
10	11	12	11	12	11	12	12	12	11	12	12	12	11	12	11	12	11	12	209
11	5	5	5	5	5	4	5	5	5	5	5	5	5	5	5	5	5	5	89
12	24	24	24	24	23	24	23	23	23	23	23	23	24	24	24	23	24	23	423
13	5	5	6	5	5	5	6	5	5	5	6	5	5	5	5	5	5	5	93
14	19	19	19	18	19	18	18	19	19	19	18	19	19	18	19	19	19	19	337
15	4	4	4	5	5	4	5	4	4	4	4	4	4	4	4	4	4	4	75
16	3	2	2	3	2	3	2	3	3	3	3	2	3	2	2	3	3	3	47
17	1	2	1	2	1	1	1	1	1	1	1	2	1	2	1	1	1	2	23
18	1	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	3
19	0	0	0	1	0	1	1	1	0	0	0	0	0	1	0	0	0	0	5
Total	100	99	100	100	98	99	100	100	99	99	100	99	100	101	99	100	100	100	**1793**

Table 9: Distribution of sentences per annotation batch and readability level.

Feature	Feature (Arabic)	Rule
Singular imperfective verb	الفعل المضارع المفرد	num=s, asp=i, pos=verb
Proclitic: 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=i, 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 (nouns and adjectives)	المثنى في الأسماء والصفات	num=d, pos=noun / adj
Sound feminine plural	جمع المؤنث السالم	form_num=p, form_gen=f, pos=noun / adj
Broken plurals	جمع التكسير	pos=noun / adj, form_num=s, num=p
Passive voice	المبني للمجهول	vox=p, pos=verb
Singular and plural perfective verb	الفعل الماضي المفرد والجمع	pos=verb, asp=p, num=s / p
Sound masculine plural	جمع مذكر سالم	form_gen=m, form_num=p, pos=noun / adj
Dual perfective verb	الفعل الماضي المثنى	pos=verb, asp=p, num=d
Dual imperfective verb	الفعل المضارع المثنى	pos=verb, asp=i, num=d
Singular imperative verb	فعل الأمر المفرد	pos=verb, asp=c, num=s
Enclitic: dual pronoun	لواحق: ضمير المثنى المتصل	enc0=[2d_dobj, . . . , 3d_pron]
Waw of oath	واو القسم (والله)	prc2=wa_prep and lex ∈ qassam_lex
Plural imperative verb	فعل الأمر الجمع	pos=verb, asp=c, num=p
Nun of feminine plural (nūn al-niswā)	نون النسوة في الأسماء والأفعال	ends with noon and 3FP is present
Conjunctions / connectives	(ثم ، حتى، أو، أم، لكن، أمّا)	lex ∈ connective_lex
Dual imperative verb	فعل الأمر للمثنى	pos=verb, asp=c, num=d
Interrogative Alif	أداة الاستفهام: أ (أسمعتَ؟)	prc3 in [>a_ques, A_ques]
Ba of oath	باء القسم	prc1=bi_prep and lex ∈ qassam_lex
Ta of oath	تاء القسم	prc1=ta_prep and lex ∈ qassam_lex

Table 10: Morphological features and extraction rules derived from the BAREC annotation guidelines.

Feature	Feature (Arabic)	Rule
Nominal sentence	الجملة الاسمية	parent ≠ inna and sisters; has a child with dependency relation SBJ; POS ≠ VRB
Verbal sentence without direct object	جملة فعلية بدون مفعول به	parent = verb; no dependency relation OBJ
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 = verb; has exactly one child with dependency relation OBJ
Sentence with two verbs	جملة فيها فعلين	verb count ≥ 2
Verbal sentence with clausal object (Masdar *an*)	جملة فعلية مفعولها أن المصدرية	token lemma = أن; POS = PRT; has a child with POS = VRB, deprel = OBJ, asp=i
Verbal sentence with two direct objects	جملة فعلية تتعدى إلى مفعولين	parent POS = VRB; two children with dependency relation OBJ
Vocative	المنادى	parent POS = PRT; FEATS include pos=part_voc; has child with deprel OBJ
Inna and its sisters	إن وأخواتها	parent lemma in inna set; has a child with dependency relation PRD
Kana and its sisters	كان وأخواته	parent POS = VRB; lemma in kana set; has a child with dependency relation PRD
Preposed predicate, postponed subject	الخبر المقدم والمبتدأ المؤخر	dependency relation = SBJ; child POS ≠ VRB; parent index < child index
Nominal sentence with nominal predicate	جملة اسمية خبرها جملة اسمية	sentence does not start with a verb; contains child with deprel TPC
False idafa (attributive)	إضافة خيالية (لفظية)	parent POS = NOM; pos=adj in FEATS; has child with deprel IDF
Exception	استثناء	POS = part_restrict
True idafa	إضافة حقيقية	parent POS = NOM; pos=adj not in FEATS; has child with deprel IDF
Noun + adjective	صفة وموصوف	adjacent NOM followed by ADJ (or noun with pos=adj); agreement in gender/number
Basic interrogatives	أدوات استفهام أساسية	lemma in {كيف, ما, أين, من, متى, ماذا}; POS indicates interrogative
Interrogative particle *hal*	أداة الاستفهام هل	lemma = هل; POS = part_interrog

Table 11: Syntactic features and extraction rules derived from the BAREC annotation guidelines.

Vocabulary	Vocabulary (Arabic)
Proper noun	اسم علم
Personal pronouns (non-clitics)	ضمير منفصل
Singular demonstrative pronoun	اسم الإشارة المفرد
Dual and plural demonstrative pronoun	اسم إشارة مثنى، جمع
Adjective	صفة
Verb	فعل
Prepositions	حروف الجر
Negation particles	أحرف النفي
Singular relative pronouns	أسماء الوصل المفردة
Dual and plural relative pronouns	أسماء الوصل المثنى والجمع
Father / Brother nouns	أبو ـ أخو
Ordinal numbers (written in words)	العدد الترتيبي
Cardinal numbers (written in words)	العدد الأصلي بالأحرف
Arabic or Indic digits (1–10)	الأرقام (العربية أو الهندية) من ١ إلى ١٠
Arabic or Indic digits (11–100)	الأرقام (العربية أو الهندية) من ١١ إلى ١٠٠
Arabic or Indic digits (101–1,000)	الأرقام (العربية أو الهندية) من ١٠١ إلى ١٠٠٠
Arabic or Indic digits (1,001–1,000,000)	الأرقام (العربية أو الهندية) من ١٠٠١ إلى ١٠٠٠٠٠٠

Table 12: Vocabulary feature levels in English and Arabic derived from the BAREC guidelines.

Table 13: Linguistic phenomena and their associated code and readability levels used for our feature-based model

Arabic Description	Code	Level
عدد المقاطع	O	–
كلمات تستخدم مد الألف (مثل آكل، آسف)	O4-1	4
أفعال معتلة الآخر	O7-1	7
عدد الكلمات	WC	–
مفرد مضارع (مذكر، مؤنث) (متكلم ، مخاطب، غائب)	M1-1	1
سوابق: ال التعريف	M3-1	3
سوابق: واو العطف	M3-2	3
لواحق: ضمير المتكلم المفرد المتصل	M3-3	3
الفعل المضارع الجمع	M4-1	4
سوابق: حروف جر متصلة (ب+ ل+ ك+)	M4-2	4
ظرف منون	M4-3	4
لواحق: ضمير متصل مفرد أو جمع	M5-1	5
المثنى (في الأسماء والصفات)	M5-2	5
جمع المؤنث السالم في الأسماء والصفات	M5-3	5
الفعل الماضي المفرد والجمع	M6-1	6
جمع مذكر سالم	M6-2	6
الفعل الماضي المثنى	M7-1	7
الفعل المضارع المثنى	M7-1	7
فعل الأمر المفرد	M7-3	7
لواحق: ضمير المثنى المتصل	M7-4	7
جمع التكسير	M7-5	7
واو القسم (والله)	M7-6	7
فعل الأمر الجمع	M8-1	8
نون النسوة في الأسماء والأفعال (انتظرنَ دورهنّ)	M8-2	8
سوابق أخرى: سين الاستقبال، واو الاستئناف، فاء العطف	M8-3	8
(ثم ، حتى، أو، أم، لكن، أمّا)	M8-4	8
فعل الأمر للمثنى	M9-1	9
أداة الاستفهام: أ (أسمعتَ؟)	M9-2	9
باء القسم	M9-3	9
القسم : أداة القسم والمقسم به وجواب القسم	M9-4	9
المبني للمجهول	M10-1	10
المشتقات على أنواعها	M11-1	11
التصغير	M12-1	12
نون التوكيد	M13-1	13

Continued on next page

Arabic Description	Code	Level
تاء القسم	M13-2	13
آخر ـ اشتقاق	M-O	–
كلمة واحدة (لا تراكيب نحوية ممكنة)	S1-1	1
إضافة حقيقية (باب بيت / بيت مريم)	S2-1	2
جملة اسمية	S2-2	2
صفة وموصوف	S2-3	2
بدل كل (صديقي أحمد)	S3-1	3
بدل إشارة (هذا البيت)	S3-2	3
جملة فعلية بدون مفعول به	S4-1	4
جار ومجرور	S4-2	4
جملة فعلية مع مفعول به واحد اسم	S5-1	5
جمل معطوفة	S5-2	5
أدوات استفهام أساسية: ماذا، متى، من، أين، ما، كيف	S5-3	5
صيغة التعجب ما أفعل	S5-4	5
جملة فيها فعلين	S6-1	6
مفعول فيه (ظروف زمان ومكان)	S7-1	7
حال	S7-2	7
أداة الاستفهام هل	S7-3	7
المفعول المطلق	S8-1	8
المفعول لأجله	S8-2	8
المفعول معه	S8-3	8
جملة فعلية تتعدى إلى مفعولين	S8-4	8
المنادى	S9-1	9
إن وأخواتها	S10-1	10
كان وأخواتها	S10-2	10
خبر مقدم / مبتدأ مؤخر	S10-3	10
العنعنة (حدثني ... قال)	S10-4	10
رُبّ (حرف جر شبيه بالزائد)	S10-5	10
جملة الصلة، جملة الصفة، جملة الحال، جملة المفعول به	S10-6	10
جملة اسمية خبرها جملة اسمية	S11-1	11
إضافة خيالية (لفظية)	S11-2	11
جمل اعتراضية (تفسير، دعاء...)	S12-1	12
استثناء	S12-2	12
حصر	S12-3	12
بدل غير بدل الكل	S12-4	12
تمييز	S12-5	12
الجمل الشرطية	S13-1	13

Continued on next page

Arabic Description	Code	Level
لما حرف الجزم	S13-2	13
التوكيد المعنوي	S14-1	14
المدح والذم	S14-2	14
جملة أن المصدرية في محل رفع مبتدأ	S14-3	14
صيغة التعجب أفعل به	S14-4	14
تراكيب غير متداولة تحتاج إلى التشكيل الإعرابي لفكها	S15	15
آخر ـ تراكيب نحوية	S-O	–
اسم علم (متداول بسيط تركيبيا)	V1-1	1
ضمير منفصل	V1-2	1
اسم جنس	V1-3	1
مفردات متطابقة مع العامية ـ سامر ١	V1-4	1
الأرقام (العربية أو الهندية) ١٠-١	V1-5	1
أبو ـ أخو	V2-1	2
مفردات متشابهة مع العامية ـ سامر ١	V2-2	2
صفة	V2-3	2
فعل	V2-4	2
العدد الأصلي بالأحرف	V2-5	2
اسم الإشارة المفرد	V3-1	3
مفردات فصيحة شائعة ـ سامر ١	V3-2	3
الأرقام (العربية أو الهندية) ١٠٠-١١	V3-3	3
حروف الجر	V4-1	4
العدد الترتيبي	V5-1	5
الأرقام (العربية أو الهندية) ١٠٠٠-١٠١	V5-2	5
اسم إشارة مثنى، جمع	V5-3	5
مفردات فصيحة ـ سامر ١	V6-1	6
مفردات فصيحة شائعة ـ سامر ٢	V7-1	7
مفردات فصيحة ـ سامر ١ و سامر ٢	V8-1	8
أحرف النفي	V8-2	8
الأرقام (العربية أو الهندية) ١٠٠٠٠٠٠-١٠٠١	V8-3	8
مفردات تصف حالات مزاجية وشعورية إيجابية وسلبية	V9-1	9
أسماء الوصل المفردة	V10-1	10
(قد ـ لقد)	V10-2	10
(ممّا ـ عمّا ـ عمَ ـ علامَ ـ فيمَ ـ إلامَ ـ بمَ)	V10-3	10
أسماء الوصل المثنى والجمع	V11-1	11
مفردات فصيحة ـ سامر ٣	V12-1	12
اسم الفعل: إيهِ، صَهْ، رويدَك ...	V12-2	12

Continued on next page

Arabic Description	Code	Level
الأرقام (العربية أو الهندية) أكبر من ١٠٠٠٠٠٠٠	V12-3	12
ذو	V12-4	12
(بل - بلى - أجل - قط)	V12-5	12
كلمات تصف حالات نفسية عميقة مثل الاكتئاب، الضياع	V13-1	13
استخدام كلمات منحوتة غير متداول...)	V13-2	13
الرموز (شم.)	V13-3	13
مفردات فصيحة - سامر ٤	V14-1	14
مفردات قانونية، علمية، دينية، سياسية، غير متخصصة	V14-2	14
فو - حمو	V14-3	14
المفردات المتخصصة التي لا تكفي معرفة الكلمة وحدها لفهمها...	V15-1	15
الترخيم في أسماء العلم (مثلا أفاطم؟)	V15-2	15
مفردات فصيحة - سامر ٥	V16-1	16
مفردات متخصصة ومفردات عربية عالية غير شائعة كثير...	V16-2	16
مفردات علمية وتراثية غير متداولة اليوم	V17	17
مفردات علمية وتراثية غير متداولة...	V18	18
مفردات علمية وتراثية غير متداولة اليوم...	V19	19
آخر - مفردات	V-O	–
لا رمزية في النص	C1-1	1
فكرة مباشرة وصريحة وحسية	C1-1	1
المحتوى من حياة القارئ	C5-1	5
لا رمزية في النص	C5-1	5
بعض الرمزية أو عدم التصريح المباشر...	C7-1	7
بعض الرمزية يحتاج معها القارئ إلى مساعدة...	C8-1	8
هناك شيء من الرمزية على مستوى الحدث ..	C9-1	9
هناك درجة من الرمزية وحاجة للمعرفة السابقة...	C11-1	11
أفكار رمزية ومعنى باطن خاصة على صعيد البعد ...	C13-1	13
تعابير ثقافية محلية قد لا يفهمها من لا يشترك في نفس الثقافة	C13-1	13
أفكار رمزية، مجردة، علمية، أو شعرية...	C15-1	15
آخر - محتوى	C-O	–