Strategies for Arabic Readability Modeling

Juan Piñeros Liberato, Bashar Alhafni, Muhamed Al Khalil, Nizar Habash
Computational Approaches to Modeling Language Lab
New York University Abu Dhabi
{juanpl,alhafni,muhamed.alkhalil,nizar.habash}@nyu.edu

Abstract

Automatic readability assessment is relevant to building NLP applications for education, content analysis, and accessibility. However, Arabic readability assessment is a challenging task due to Arabic's morphological richness and limited readability resources. In this paper, we present a set of experimental results on Arabic readability assessment using a diverse range of approaches, from rule-based methods to Arabic pretrained language models. We report our results on a newly created corpus at different textual granularity levels (words and sentence fragments). Our results show that combining different techniques yields the best results, achieving an overall macro F_1 score of 86.7 at the word level and 87.9 at the fragment level on a blind test set. We make our code, data, and pretrained models publicly available.[1]

1 Introduction

The task of automatic readability assessment aims at modeling the reading and comprehension difficulty of a given piece of text for a particular target audience. This is relevant to building and enhancing pedagogical natural language processing (NLP) applications, which aid students in language learning (Xia et al., 2016; Vajjala and Meurers, 2012), help teachers with designing curricula and writing assessments (Collins-Thompson and Callan, 2004b), and enable the personalization of NLP systems' output to target users with different readability levels (Marchisio et al., 2019; Agrawal and Carpuat, 2019). Research on English automatic readability assessment have garnered substantial interest in terms of dataset creation (Heilman et al., 2007; Vajjala and Meurers, 2013; Xia et al., 2016; Vajjala and Lučić, 2018) and modeling advancements (Deutsch et al., 2020; Martinc et al., 2021; Lee and Vajjala, 2022). In contrast, other languages such as Arabic have not received as much attention.

[1] https://github.com/CAMeL-Lab/samer-arabic-readability

Arabic is a morphologically rich and orthographically ambiguous language. Words have many inflected forms varying in terms of gender, number, person, case, aspect, mood, voice, as well as a large number of attachable clitics, such as pronominal objects and prepositions (Habash, 2010). Arabic's high level of complexity poses a significant challenge for new learners. Furthermore, while Modern Standard Arabic (MSA) is used in education and the media, modern-day Arabs natively speak a variety of Arabic dialects that differ from MSA, making MSA readability a relevant issue for them too. There are growing research efforts on Arabic readability assessment (Al-Khalifa and Al-Ajlan, 2010; Al Tamimi et al., 2014; El-Haj and Rayson, 2016; Saddiki et al., 2018). However, we are not aware of any work that systematically explores modeling approaches for Arabic readability at different textual granularity levels. In this paper, we present Arabic readability assessment results using diverse approaches relying on frequency and rule-based models as well as pretrained language models (PLMs). We use the newly created SAMER Arabic Text Simplification Corpus (Alhafni et al., 2024) and report on word-level and fragment-level readability. Our contributions are as follows:

- We systematically explore different modeling approaches to report on the task of Arabic readability assessment, ranging from rule-based methods to Arabic PLMs.
- We benchmark our models on a new corpus with different readability levels.
- We show that combining different modeling techniques yields optimal results: 86.7 word-level macro F_1 and 87.9 fragment-level macro F_1 on a blind test set.

We discuss related work in §2, provide an overview of our dataset in §3, describe our models for Arabic readability assessment in §4, and discuss results in §5.

Proceedings of The Second Arabic Natural Language Processing Conference, pages 55–66
August 16, 2024

2 Related Work

2.1 Readability Assessment Datasets

Automatic readability assessment has received considerable attention, leading to the development of many resources (Collins-Thompson and Callan, 2004a; Pitler and Nenkova, 2008; Feng et al., 2010; Vajjala and Meurers, 2012; Xu et al., 2015; Xia et al., 2016; Nadeem and Ostendorf, 2018; Vajjala and Lučić, 2018; Deutsch et al., 2020; Lee et al., 2021). Most of the English datasets were initially derived from textbooks as they are considered to be naturally suited for readability assessment research, given that the linguistic characteristics of texts become more complex as school grade increases (Vajjala, 2022). However, many textbooks are under copyright restrictions and may not be accessible in a digitized form. This led to relying on crowd sourcing to annotate data collected from the web (Vajjala and Meurers, 2012; Vajjala and Lučić, 2018) or from English assessment exams targeting second-language (L2) learners (Xia et al., 2016), where the Common European Framework of Reference (CEFR) (Council of Europe, 2001) is used.

When it comes to Arabic, specifically Modern Standard Arabic (MSA), early work on readability assessment relied mainly on academic curricula (Al-Khalifa and Al-Ajlan, 2010; Al Tamimi et al., 2014; Forsyth, 2014; Khalil et al., 2018). More recently, there have been more efforts to create Arabic readability assessment resources. Khallaf and Sharoff (2021) consolidated multiple annotated L2 datasets and mapped their readability levels to CEFR. Habash and Palfreyman (2022) created the ZAEBUC dataset that contains essays written by native Arabic speakers, which were manually corrected and annotated for writing proficiency using the CEFR levels. Naous et al. (2023) introduced a manually annotated multi-domain multilingual dataset for readability assessment. In our work, we use the newly introduced publicly available SAMER Arabic Text Simplification Corpus (Alhafni et al., 2024), which was manually annotated for readability leveling. We discuss the corpus in more detail in §3. It is noteworthy that this corpus is one of the publicly available resources created by the Simplification of Arabic Masterpieces for Extensive Reading (SAMER) project which includes a readability leveled lexicon (Al Khalil et al., 2020a; Jiang et al., 2020), and a Google Doc add-on (Hazim et al., 2022).

2.2 Approaches to Readability Assessment

Early approaches for automatic readability assessment relied on surface-level features that could be extracted from raw text such as the average number words per sentence and the average number of characters per word. Such approaches include commonly used readability measures such as the Dale-Chall Readability Score (Dale and Chall, 1948) and the Flesch-Kincaid Grade Level (FKGL) (Flesch, 1948). With the emergence of machine learning and data driven methods, approaches were extended to leverage statisical language models (Si and Callan, 2001) and linguistic features (Heilman et al., 2007; Petersen and Ostendorf, 2009; Ambati et al., 2016). More recently, deep learning approaches were explored (Cha et al., 2017; Jiang et al., 2018; Azpiazu and Pera, 2019), including the use of Transfomer-based PLMs (Deutsch et al., 2020; Lee and Vajjala, 2022; Naous et al., 2023; Imperial and Kochmar, 2023).

Although a lot of this research evolved on English, approaches to modeling Arabic readability assessment witnessed a similar trend. Inspired by English readability formulas, Al-Tamimi et al. (2014) developed the Arabic Automatic Readability Index (AARI). Similarly, El-Haj and Rayson (2016) introduced OSMAN, an adaptation of conventional readability formulas such as FKGL to Arabic. When it comes to machine learning models, the majority were based on linguistic features such as perplexity scores from statistical language models (Al-Khalifa and Al-Ajlan, 2010), morphological information (e.g., lemmas, morphemes, part-of-speech tags) (Cavalli-Sforza et al., 2014; Forsyth, 2014; Saddiki et al., 2015; Nassiri et al., 2017), and syntactic features (Saddiki et al., 2018). Despite the various efforts on modeling Arabic readability assessment, only few attempts were made to explore deep learning approaches. Khallaf and Sharoff (2021) and Naous et al. (2023) presented results on using BERT (Devlin et al., 2019; Antoun et al., 2020) for readability assessment. Moreover, it is worth noting that the majority of research on Arabic readability assessment report results at either the document or sentence levels.

In our work, we draw inspiration from previous efforts to explore various modeling approaches for Arabic readability assessment at both the word and fragment levels, encompassing a spectrum from rule-based models to PLMs and their combinations.

<table>
<tr><td>Word Level</td><td>3</td><td>3</td><td>3</td><td>3</td><td>5</td><td>3</td><td>4</td></tr>
<tr><td>Original</td><td>الفتى</td><td>يمين</td><td>إلى</td><td>واحد</td><td>فربض</td><td>إليها</td><td>فأشارت</td></tr>
<tr><td colspan="8">She signaled to them (the dogs) and one of them lodged itself to the right of the boy</td></tr>
<tr><td>Level 4</td><td>الفتى</td><td>يمين</td><td>إلى</td><td>واحد</td><td>فجلس</td><td>إليها</td><td>فأشارت</td></tr>
<tr><td colspan="8">She signaled to them and one of them sat to the right of the boy</td></tr>
<tr><td>Level 3</td><td>الفتى</td><td>يمين</td><td>إلى</td><td>واحد</td><td>فجلس</td><td>إليها</td><td>فلوحت</td></tr>
<tr><td colspan="8">She pointed to them and one of them sat to the right of the boy</td></tr>
</table>

Figure 1: An example illustrating the word-level labeling process. A word in the original text is labeled at the lowest level where it appears unchanged across the parallel versions of the text in the SAMER Corpus.

	Level 3	**Level 4**	**Level 5**	**All**
Train	5,947	4,543	3,766	14,256
Dev	1,256	926	766	2,948
Test	1,477	901	776	3,154
Total	**8,680**	**6,370**	**5,308**	**20,358**

Table 1: SAMER Corpus fragment readability level statistics per split.

	All Tokens		**Train Tokens**	
Level 3	136,805	86.6%	97,616	86.5%
Level 4	14,145	8.9%	10,151	9.0%
Level 5	7,104	4.5%	5,056	4.5%
Total	**158,054**	**100%**	**112,823**	**100%**

Table 2: SAMER Corpus word token readability level statistics after the word-level labeling process.

3 Data

3.1 SAMER Corpus

We extensively use the SAMER Arabic Text Simplification Corpus (Alhafni et al., 2024). The corpus consists of *original* texts selected from 15 publicly available Arabic fiction novels. It includes two simplified parallel versions for each text targeting learners at two readability levels (*Level 4* and *Level 3*). The levels are based on Al Khalil et al. (2020a)'s five-level lexical readability scale which ranges from Level 1 (Low Difficulty/Easy Readability) to Level 5 (High Difficulty/Hard Readability). The SAMER Corpus simplification guidelines consider the readability level of a *text* to be equal to the highest readability level found among the *words* in the *text*. So, a Level 4 *text* cannot have any Level 5 *words*, but must have at least one Level 4 word. As part of the manual simplification process, the human annotators simplified the original text to Level 4, and then to Level 3. This was done by first automatically obtaining the word-level readability of the original text using the SAMER Google Doc add-on (Hazim et al., 2022) and then manually performing minimal replacements, insertions, and deletions to simplify the text from a higher to lower readability levels. In some cases the annotators minimally modified some words to maintain grammatical agreement without changing their lexical readability levels. The add-on was used to confirm the target levels were reached; however, the annotators were allowed to overwrite incorrect automatically assigned word readability levels. The SAMER Corpus release includes the original paragraphs and their simplified counterparts segmented into smaller parallel sentence fragments using punctuation marks. Our readability assessment experiments only use the original text fragments. The release also includes the final word readability levels; however, we do not use them as we opted to employ a more generic solution for word-level readability assignment, which we discuss next.

3.2 Fragments, Words, and Readability Levels

We model readability assessment on three levels: Levels 5, 4, and 3, at the word level and fragment level. To assign a readability label to each word in the original fragments, we first obtain word-level alignments between the original fragments and the simplified parallels using an edit distance word alignment tool (Alhafni et al., 2023; Khalifa et al., 2021). We then derive the readability labels based on whether the words in the original fragments were changed in the simplified Levels 4 and 3 texts. See example in Figure 1. Similar to the SAMER Corpus, we consider the readability level of the

Model	Resources Used			
	SAMER Corpus	CAMeL BERT	CAMeL Tools	SAMER Lexicon
Lexicon		X	X	X
BERT	X	X		
Frequency	X	Counts		
MLE	X			
Default L3				

Table 3: Resources used in the word-level models.

fragment to be equal to the highest word readability level found among the words in the fragment. Tables 1 and 2 present the statistics of the corpus at the fragment and word levels, respectively.

While this alignment-based approach is applicable to any parallel original-simplified text, it struggles to distinguish between lexical readability changes and grammatical agreement changes. Nevertheless, this holistic approach is valuable for text simplification tasks that require altering both words and their grammatical dependents. Importantly, this limitation does not affect the readability level of the text fragments.

4 Approach

We present below the set of models we use for word-level and fragment-level readability labeling.

4.1 Word-Level Readability Labeling

We investigate four models to label words according to their readability levels: Maximum Likelihood Estimation (MLE), Lexicon lookup, Frequency-based labeling, and BERT-based token classification. Each model relies on different resources as summarized in Table 3. Moreover, each model has a different set of parameters, which were tuned to optimize the performance on the Dev set. We further investigate combining the models in a cascaded setup, leveraging their complementary strengths to address the limitations of each model individually.

4.1.1 Maximum Likelihood Estimation

The **MLE** model assigns the readability level R that maximizes the conditional probability $P(R|W)$, where R is the readability level of word W as estimated over the training data. Among in-vocabulary words, 97.2% appear with one readability level in the training, and 0.2% appear with all three. For out-of-vocabulary (OOV) words, we back-off to a default readability level or to one of other models we discuss below.

4.1.2 Lexicon

Our second model, **Lex**, leverages the SAMER Readability Lexicon (Al Khalil et al., 2020a), which consists of over 40K lemmas manually annotated with their readability levels (1 to 5). For the purposes of our task, we consider lemmas of Levels 1 and 2 to be included under Level 3. During inference, we use Inoue et al. (2022)'s morphological disambiguator as implemented in CAMel Tools (Obeid et al., 2020) to identify the lemma and part-of-speech tag for each word. We infer the readability level of the word using its lemma's readability level in the SAMER lexicon. In cases where the morphological disambiguator returns multiple top lemma analyses, we select the lowest readability associated with these lemmas. For OOV words, we back-off to a default readability level or to one of other models we discuss below.

4.1.3 Frequency-Based Models

Given the limited vocabulary seen the SAMER Corpus, we explore different approaches to derive readability levels from frequency data building on the known observation about the inverse correlation between frequency and readability levels (Al Khalil et al., 2020b): more frequent words have easier/lower readability levels. We leverage our SAMER Corpus training data and link it with type frequency data from a corpus of 12.6B word tokens (11.4M types) used to pretrain the CAMeLBERT models (Inoue et al., 2021).[2] We sort the 11.4M word types by frequency and divide them into adjacent bins for which we assign readability levels using one of two methods:

Distribution-based Labeling (Dist-Freq) In this method, we divide the word types into three bins that mirror the distribution of readability levels in our training data (as seen in Table 2). More concretely, the most frequent words that account for 86.5% of the total distribution mass are assigned to Level 3, followed by 9.0% assigned to Level 4, and the remaining tail assigned to Level 5.

Example-based Labeling (Ex-Freq) Based on the assumption that types of a certain readability level tend to have similar frequencies in a large

[2] https://github.com/CAMeL-Lab/Camel_Arabic_Frequency_Lists

corpus, we divide the frequency-sorted types into equally sized bins based on cumulative frequency. Utilizing our training data as training examples, we assign a readability level to all the types within each bin according to the majority readability level of training words found in that bin. During inference, if words are not observed in any bin, we default to assigning a readability level of 5, reflecting the expectation that rare words are typically harder to read. We empirically experiment with different numbers of bins and found that 10,000 bins yield the highest performance in terms of macro F_1 score.

4.1.4 BERT Token Classification

We build a word-level classifier by leveraging a Transformer-based PLM. There are many Arabic monolingual BERT (Devlin et al., 2019) models available such as AraBERT (Antoun et al., 2020), ARBERT (Abdul-Mageed et al., 2021), and JABER (Ghaddar et al., 2022). However, we chose to use CAMeLBERT MSA (Inoue et al., 2021) as it was pretrained on the largest MSA dataset to date, and following the recommendations of Inoue et al. (2021) to use it for tasks on MSA. We fine-tune CAMeLBERT MSA using Hugging Face's Transformers (Wolf et al., 2020) by adding a fully-connected linear layer with a softmax on top of its architecture. Given that BERT operates at the subword-level (i.e., wordpieces), we assign to each subword the readability level of the word it belongs to. During inference, we label each word according to the highest readability level among its subwords. We fine-tune our model on a single GPU for 10 epochs with a learning rate of 1e-5, a batch size of 32, and a maximum sequence length of 30.

4.1.5 System Combination

In addition to evaluating the various models discussed above, we consider their combinations to exploit their complementarities. Our approach to combining the systems runs the **Lex** and **MLE** models first (independently and together in different orders – four combinations) followed by one of the following *six* models: default level 3, 4, or 5, **Dist-Freq**, **Ex-Freq**, or **BERT**. The total is 24 combinations layered in two or three steps. See Table 9 in Appendix A. Since the early layers, Lex and MLE, do not handle unknown words, the later layers resolve these cases. We evaluate all these system combinations on both word and fragment leveling in terms of accuracy and macro F_1 score.

4.2 Fragment-Level Readability Labeling

We consider two approaches to fragment-level labeling: a direct BERT-based approach and an aggregation of the word-level predictions.

4.2.1 BERT Fragment Classification

We train a fragment-level classifier by fine-tuning CAMeLBERT MSA. We add a fully connected linear layer on top of the representation of the whole fragment. We experiment with different ways of obtaining the fragment representation from the BERT model: using the `[CLS]` token, mean-pooling, and max-pooling, and found mean-pooling to outperform the other representations in all evaluation metrics. We fine-tune our model using Hugging Face's Transformers (Wolf et al., 2019) on a single GPU for 10 epochs with a learning rate of 5e-5, a batch size of 32, and a maximum sequence length of 20.

4.2.2 Aggregating Word-Level Predictions

Finally, we aggregate the word-level labels produced by the various models discussed in §4.1 above to assign fragment-level labels: the fragment label equals the highest readability level found among its words.

5 Results

We present and discuss the results of our evaluation below. The complete set of results for word-level and fragment-level labeling across all experimental setups is available in Appendix A.

5.1 Word-Level Labeling Results

Table 4 presents the results on the Dev set. We start off with the results of the standalone models in Table 4**(a)**. The frequency-based approaches (Dist-Freq and Ex-Freq) improve over the majority class baseline (Default Level 3). However, they are outperformed by BERT. This improvement is attributed to the significant increase in the F_1 scores for Level 4 and Level 5 words. In Table 4**(b)** we show that the results improve further when combining the MLE model with BERT as a back-off system.

Results in Table 4(**c**) show that using the frequency-based and BERT models as back-off systems to Lex improve the results compared to defaulting to Level 3, with Lex $\rightarrow$ BERT being the best performer. However, the improvements when using a back-off model to Lex are not as large as the ones observed when using the MLE model (Table 4**(b)**). This is due to the larger coverage the

	Model	$F_1(3)$	$F_1(4)$	$F_1(5)$	F_1	Acc.
(a)	Default Level 3	92.8	0.0	0.0	30.9	86.5
	Dist-Freq	84.2	20.8	28.6	44.5	71.1
	Ex-Freq	93.0	21.5	14.7	43.1	86.4
	BERT	96.5	67.9	59.3	74.6	92.4
(b)	MLE → Level 3	95.1	57.5	41.2	64.6	91.0
	MLE → Dist-Freq	91.6	51.0	39.8	60.8	83.1
	MLE → Ex-Freq	95.0	56.9	42.4	64.7	90.3
	MLE → BERT	96.7	69.9	61.0	75.9	92.8
(c)	Lex → Level 3	97.8	85.2	74.1	85.7	95.7
	Lex → Dist-Freq	97.8	84.5	75.2	85.8	95.5
	Lex → Ex-Freq	97.8	85.1	74.6	85.8	95.7
	Lex → BERT	98.0	85.1	76.5	86.5	95.9
(d)	Lex → MLE → Level 3	97.8	85.2	74.5	85.8	95.8
	Lex → MLE → BERT	98.0	85.1	76.5	86.5	95.9
	MLE → Lex → Level 3	98.0	85.7	76.9	86.8	96.0
	MLE → Lex → BERT	98.1	85.5	78.8	87.5	96.2
	Tuned-MLE → Lex → BERT	**98.2**	**86.1**	**79.4**	**87.9**	**96.3**

Table 4: Word-level results on the Dev set. $F_1(3)$, $F_1(4)$, and $F_1(5)$ are the macro F_1 scores for levels 3, 4, and 5, respectively. F_1 is the overall macro F_1 score. Underlined numbers represent the best results in each subcategory of experiments. Best overall results are in bold.

Model	$F_1(3)$	$F_1(4)$	$F_1(5)$	F_1	Acc.
BERT (Fragment-Level)	88.7	68.2	58.8	71.9	79.4
BERT (Word-Level)	85.1	70.2	69.4	74.9	76.4
Lex → MLE → Level 3	90.9	85.0	81.1	85.7	86.6
Lex → MLE → BERT	91.2	85.3	82.9	86.4	87.2
MLE → Lex → Level 3	91.2	85.9	83.0	86.7	87.5
MLE → Lex → BERT	91.6	86.1	84.6	87.5	88.1
Tuned-MLE → Lex → BERT	**92.1**	**86.7**	**84.9**	**87.9**	**88.6**

Table 5: Fragment-level results on the Dev set.

Lexicon has on the Dev set (96.4% of all tokens) compared to the MLE system (79.0%).

Finally, in Table 4**(d)** we present the maximal combination results. We find that using the MLE model, followed by Lex and then BERT yields the best results. We further tune this combination by considering different probability thresholds at which to back-off from MLE. We found 85% MLE minimum probability to give the best results on the Dev set. Our best model combination is thus **Tuned-MLE → Lex → BERT** with 87.9 F_1.

5.2 Fragment-Level Labeling Results

Table 5 presents the fragment-level results on the Dev set. We find that, although the fragment-level BERT classifier does better than its word-level counterpart, the aggregated word-level models perform better on the fragment-level. We obtain the best results using the (**Tuned-MLE → Lex → BERT**) model, achieving an F_1 score of 87.9. It is interesting to note that the best system coincidentally achieves the same overall F_1 macro at the word and fragment levels. Our best system is better at predicting Level 3 words compared to Level 3 fragments (98.2 v.s. 92.1). Conversely, the system is better at predicting Level 4 and Level 5 fragments compared to the words. This makes sense given that Level 3 fragments are exclusively composed of Level 3 words, any word-level error on a Level 3 fragment leads to fragment error. In sum, the 3.7%

Fragment Label	Word Errors	# of Fragments		
Correct	0	2,300	78.%	78.0%
Correct	1	239	8.1%	
Correct	2	56	1.9%	10.6%
Correct	3+	16	0.5%	
Incorrect	1	283	9.6%	
Incorrect	2	38	1.3%	11.4%
Incorrect	3+	16	0.5%	

Table 6: Summary of fragment-word error combinations on the Dev set. We identify three groups: correct fragments with no word errors, correct fragments with some word errors, and incorrect fragments with word errors.

accuracy errors at the word level lead to 11.4% accuracy errors at the fragment level. Table 6 presents a detailed breakdown of the combinations of word and fragment errors.

Finally, we revisit our best model combination **Tuned-MLE → Lex → BERT** in Table 7, where we give a summary of the decisions and mistakes made by each of its three components and their effect on word-level and fragment-level performance. We notice that most of the decisions were taken by the MLE model, which had the lowest error rate, and the lowest rate of error propagation to the fragment level. However, when errors at the word level happen, there is a large chance a fragment error will follow suit in all three models. Moreover, we note that the performance is highly degraded by the last model (BERT) decisions, with 35.6% word-level and 67.7% fragment-level errors.

5.3 Blind Test Results

Table 8 presents the results on the Test set. We observe consistent conclusions to the Dev results. Our best system (**Tuned-MLE → Lex → BERT**) achieves an overall F_1 score of 86.7 at the word level and 87.9 at the fragment level.

5.4 Manual Error Analysis

We manually classified 100 cases of word readability errors from the Dev set (out of 814 or 3.7% of all words) into seven distinct error types. We provide a brief description of each error type below, with its percentage of occurrence. The errors are presented in order of precedence, so if there is an Input error, we do not consider any other error below it, and so on.

	MLE	LEX	BERT
Decisions	17,058	4,843	174
Mistakes	377	375	62
Applied	77.3%	21.9%	0.8%
Word Error	2.2%	7.7%	35.6%
Fragment Error	41.4%	56.3%	67.7%

Table 7: The word-level decisions taken by each of the layers of the best-performing system on the Dev set's 22,075 tokens, and their error rates in terms of word-level and fragment-level labeling.

Input Error: 3% The word is malformed in terms of spelling; e.g., معارفة *mςArfħ* instead of معارفه *mςArfh* 'his features'.

Gold Reference Annotation Error: 18% The human annotator made a mistake of under-simplification or over-simplification, e.g., rewriting فقطعا أكثر الطريق 'they *crossed* most of the road' as فمشيا أكثر الطريق 'they *walked* most of the road' (L4) is unnecessary since the *original* is not L5.

Gold Reference Determination Error: 8% As discussed in §3.2, our process to determine the word-level readability confused grammatical agreement changes with lexical simplification changes, e.g., the phrase ملامحه المتجعدة *mlAmHh Almtjςdħ* 'his wrinkled features' is simplified correctly to وجهه المتجعد *wjhh Almtjςd* 'his wrinkled face' by changing the first word's lemma and only changing the gender agreement of the second word; however both are considered changed and thus assigned a higher level.

MLE Error: 22% The MLE model misclassified a word, e.g., confusing مهد *mahd* 'cradle' (L5) with the verb *mah∼ad* 'he paved' (L4).

Disambiguation Error: 31% The Lex model misclassified a word whose lemma is in the lexicon, because of morphosyntactic or lemmatization choice errors, e.g., منفذ *manfað* 'outlet' (L4) is incorrectly identified as *munaf∼ið* 'executor' (L3).

Lexicon Error: 11% The correct lemma is not in the lexicon, and an incorrect lemma is chosen, e.g., for the word لامسا *lAmsA* the system chose the verbal analysis *laAmas* 'touched' instead of the nominal active participle *laAmis* 'touching'.

Model	Word-Level					Fragment-Level				
	$F_1(3)$	**$F_1(4)$**	**$F_1(5)$**	**F_1**	**Acc.**	**$F_1(3)$**	**$F_1(4)$**	**$F_1(5)$**	**F_1**	**Acc.**
BERT	96.8	70.4	56.9	74.7	92.9	87.9	71.8	66.8	75.5	78.3
MLE → BERT	96.9	71.9	58.0	75.6	93.2	88.7	73.4	67.5	76.5	79.3
Lexicon → BERT	97.8	84.8	72.5	85.0	95.6	91.1	85.6	79.4	85.4	86.7
Tuned-MLE → Lex → BERT	**98.1**	**86.2**	**75.9**	**86.7**	**96.2**	**93.3**	**87.3**	**82.9**	**87.9**	**89.1**

Table 8: Results on the Test set at both the word and fragment levels.

BERT Error: 7% The word is OOV in the lexicon, and BERT misclassified it, e.g., the lemma حائم *HAŷm* 'hovering' (annotators assigned L5) is not in the lexicon, and BERT misclassified it as L4.

The lexicon and disambiguation errors take a significant share of all errors and direct us towards working on improving these resources in the future; as better generalizing models are developed, we would rely less on the MLE model. The rate of gold errors is low and within reason given the complexity of the task.

6 Conclusions and Future Work

We explored the problem of Arabic readability assessment using a diverse set of approaches relying on frequency and rule-based models as well as Arabic pretrained language models (PLMs). We reported results using a newly manually created corpus at both the word and fragment levels. We further highlighted the strengths and weaknesses of each approach and underscored the importance of employing different strategies to address Arabic readability assessment effectively. Our findings demonstrate that combining different modeling techniques yields the best results, achieving an overall macro F_1 score of 86.7 at the word level and 87.9 at the fragment level.

In future work, we plan to explore the effect of various linguistic features in enhancing machine learning models for Arabic readability assessment. We plan to continue to improve basic enabling technologies such as morphological disambiguation and lemmatization and study their effect on readability models. We further plan to employ our best results in the development of online tools to support pedagogical NLP applications.

Acknowledgements

We acknowledge the support of the High Performance Computing Center at New York University Abu Dhabi.

Limitations

We acknowledge the following limitations.

- By focusing on lexical readability, the approach used to create the SAMER corpus ignores many readability related phenomena such as phonological, morphological and syntactic complexity.
- The SAMER corpus does not cover all variations of Arabic text genres, which limits the robustness of the results.
- The assessment at three readability levels might not capture the full complexity of text readability at wider age and education level ranges.
- The study lacks human evaluation to corroborate the automatic readability assessments, which is crucial for validating the practical effectiveness of the models.

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.

Sweta Agrawal and Marine Carpuat. 2019. Controlling text complexity in neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1549–1564, Hong Kong, China. 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. 2020a. 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.

Muhamed Al Khalil, Nizar Habash, and Zhengyang Jiang. 2020b. 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.

Abdel Karim Al Tamimi, Manar Jaradat, Nuha Al-Jarrah, and Sahar Ghanem. 2014. AARI: automatic Arabic readability index. *International Arab Journal of Information Technology*, 11(4):370–378.

Abdel-Karim Al-Tamimi, Manar Jaradat, Nuha Aljarrah, and Sahar Ghanim. 2014. Aari: Automatic arabic readability index. *International Arab Journal of Information Technology*, 11:370–378.

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.

Bashar Alhafni, Go Inoue, Christian Khairallah, and Nizar Habash. 2023. Advancements in Arabic grammatical error detection and correction: An empirical investigation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6430–6448, Singapore. Association for Computational Linguistics.

Bharat Ram Ambati, Siva Reddy, and Mark Steedman. 2016. Assessing relative sentence complexity using an incremental CCG parser. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1051–1057, San Diego, California. Association for Computational Linguistics.

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.

Ion Madrazo Azpiazu and Maria Soledad Pera. 2019. Multiattentive recurrent neural network architecture for multilingual readability assessment. *Transactions of the Association for Computational Linguistics*, 7:421–436.

Violetta Cavalli-Sforza, Mariam El Mezouar, and Hind Saddiki. 2014. Matching an Arabic text to a learners' curriculum. In *Proceedings of the Conference on Arabic Language Processing (CITALA)*, pages 79–88, Oujda, Morocco.

Miriam Cha, Youngjune Gwon, and H. T. Kung. 2017. Language modeling by clustering with word embeddings for text readability assessment. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17, page 2003–2006, New York, NY, USA. Association for Computing Machinery.

Kevyn Collins-Thompson and James P. Callan. 2004a. A language modeling approach to predicting reading difficulty. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 193–200, Boston, Massachusetts, USA. Association for Computational Linguistics.

Kevyn Collins-Thompson and Jamie Callan. 2004b. Information retrieval for language tutoring: an overview of the reap project. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '04, page 544–545, New York, NY, USA. Association for Computing Machinery.

C. o. E. Council of Europe. 2001. Common european framework of reference for languages: learning, teaching, assessment. Cambridge University Press.

Edgar Dale and Jeanne S. Chall. 1948. A formula for predicting readability. *Educational Research Bulletin*, 27(1):11–28.

Tovly Deutsch, Masoud Jasbi, and Stuart Shieber. 2020. Linguistic features for readability assessment. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–17, Seattle, WA, USA → Online. Association for Computational Linguistics.

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.

Mahmoud El-Haj and Paul Rayson. 2016. OSMAN: A novel Arabic readability metric. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Portorož, Slovenia.

Lijun Feng, Martin Jansche, Matt Huenerfauth, and Noémie Elhadad. 2010. A comparison of features for automatic readability assessment. In *Coling 2010: Posters*, pages 276–284, Beijing, China. Coling 2010 Organizing Committee.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of Applied Psychology*, 32(3):p221 – 233.

Jonathan Forsyth. 2014. Automatic readability prediction for modern standard Arabic. In *Proceedings of the Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT)*.

Abbas Ghaddar, Yimeng Wu, Sunyam Bagga, Ahmad Rashid, Khalil Bibi, Mehdi Rezagholizadeh, Chao

Xing, Yasheng Wang, Xinyu Duan, Zhefeng Wang, Baoxing Huai, Xin Jiang, Qun Liu, and Phillippe Langlais. 2022. Revisiting pre-trained language models and their evaluation for Arabic natural language processing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3135–3151, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Nizar Habash and David Palfreyman. 2022. ZAEBUC: An annotated Arabic-English bilingual writer corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 79–88, Marseille, France. European Language Resources Association.

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

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, Rochester, New York. Association for Computational Linguistics.

Joseph Marvin Imperial and Ekaterina Kochmar. 2023. Automatic readability assessment for closely related languages. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5371–5386, Toronto, Canada. Association for Computational Linguistics.

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.

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

Zhengyang Jiang, Nizar Habash, and Muhamed Al Khalil. 2020. An online readability leveled Arabic thesaurus. In *Proceedings of the 28th International Conference on Computational Linguistics: System Demonstrations*, pages 59–63, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

Zhiwei Jiang, Qing Gu, Yafeng Yin, and Daoxu Chen. 2018. Enriching word embeddings with domain knowledge for readability assessment. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 366–378, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Salam Khalifa, Ossama Obeid, and Nizar Habash. 2021. Character Edit Distance Based Word Alignment. `https://github.com/CAMeL-Lab/ced_word_alignment`.

Muhamed Al Khalil, Hind Saddiki, Nizar Habash, and Latifa Alfalasi. 2018. A Leveled Reading Corpus of Modern Standard Arabic. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Miyazaki, Japan.

Nouran Khallaf and Serge Sharoff. 2021. Automatic difficulty classification of Arabic sentences. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 105–114, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Bruce W. Lee, Yoo Sung Jang, and Jason Lee. 2021. Pushing on text readability assessment: A transformer meets handcrafted linguistic features. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10669–10686, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Justin Lee and Sowmya Vajjala. 2022. A neural pairwise ranking model for readability assessment. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3802–3813, Dublin, Ireland. Association for Computational Linguistics.

Kelly Marchisio, Jialiang Guo, Cheng-I Lai, and Philipp Koehn. 2019. Controlling the reading level of machine translation output. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 193–203, Dublin, Ireland. European Association for Machine Translation.

Matej Martinc, Senja Pollak, and Marko Robnik-Šikonja. 2021. Supervised and unsupervised neural approaches to text readability. *Computational Linguistics*, 47(1):141–179.

Farah Nadeem and Mari Ostendorf. 2018. Estimating linguistic complexity for science texts. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 45–55, New Orleans, Louisiana. Association for Computational Linguistics.

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

Naoual Nassiri, Abdelhak Lakhouaja, and Violetta Cavalli-Sforza. 2017. Modern standard Arabic readability prediction. In *Proceedings of the International Conference on Arabic Language*, pages 120–133.

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.

Emily Pitler and Ani Nenkova. 2008. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 186–195, Honolulu, Hawaii. Association for Computational Linguistics.

Hind Saddiki, Karim Bouzoubaa, and Violetta Cavalli-Sforza. 2015. Text readability for Arabic as a foreign language. In *Proceedings of the International Conference of Computer Systems and Applications (AICCSA)*, pages 1–8, Marrakech, Morocco.

Hind Saddiki, Nizar Habash, Violetta Cavalli-Sforza, and Muhamed Al Khalil. 2018. Feature Optimization for Predicting Readability of Arabic L1 and L2. In *Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications*, pages 20–29, Melbourne, Australia. Association for Computational Linguistics.

Luo Si and Jamie Callan. 2001. A statistical model for scientific readability. In *Proceedings of the Tenth International Conference on Information and Knowledge Management*, CIKM '01, page 574–576, New York, NY, USA. Association for Computing Machinery.

Sowmya Vajjala. 2022. Trends, limitations and open challenges in automatic readability assessment research. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5366–5377, Marseille, France. European Language Resources Association.

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

Sowmya Vajjala and Detmar Meurers. 2012. On improving the accuracy of readability classification using insights from second language acquisition. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 163–173, Montréal, Canada. Association for Computational Linguistics.

Sowmya Vajjala and Detmar Meurers. 2013. On the applicability of readability models to web texts. In *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations*, pages 59–68, Sofia, Bulgaria. Association for Computational Linguistics.

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

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Menglin Xia, Ekaterina Kochmar, and Ted Briscoe. 2016. Text readability assessment for second language learners. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 12–22, San Diego, CA. Association for Computational Linguistics.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297.

A All Word-level Model Combination Results

Model	Word-Level					Fragment-Level				
	$F_1(3)$	$F_1(4)$	$F_1(5)$	F_1	Acc.	$F_1(3)$	$F_1(4)$	$F_1(5)$	F_1	Acc.
Default level 3	92.8	0	0	30.9	86.5	59.8	0	0	19.9	42.6
Default level 4	0	16.3	0	5.4	8.9	0	47.8	0	15.9	31.4
Default level 5	0	0	8.7	2.9	4.6	0	0	41.2	13.7	26.0
Dist-Freq	84.2	20.8	28.6	44.5	71.1	39.2	27.1	49.0	38.4	40.3
Ex-Freq	93.0	21.5	14.7	43.1	86.4	64.9	29.3	29.5	41.3	50.5
BERT	96.5	67.9	59.3	74.6	92.4	85.1	70.2	69.4	74.9	76.4
MLE → L3	95.1	57.5	41.2	64.6	91.0	75.1	62.1	50.2	62.5	67.1
MLE → L4	89.9	46.0	41.2	59.0	81.0	54.6	57.6	50.2	54.1	55.4
MLE → L5	89.9	57.5	29.8	59.1	79.7	54.6	26.6	49.4	43.5	46.7
MLE → Dist-Freq	91.6	51.0	39.8	60.8	83.1	60.5	44.0	55.9	53.5	54.2
MLE → Ex-Freq	95.0	56.9	42.4	64.7	90.3	75.7	60.8	56.0	64.2	67.3
MLE → BERT	96.7	69.9	61.0	75.9	92.8	85.4	71.0	70.7	75.7	77.1
MLE → Lex → L3	98.0	85.7	76.9	86.8	96.0	91.2	85.9	83.0	86.7	87.5
MLE → Lex → L4	98.0	82.9	76.9	85.9	95.8	91.0	83.3	83.0	85.8	86.5
MLE → Lex → L5	98.0	85.7	77.5	87.1	96.0	91.0	85.5	83.5	86.7	87.3
MLE → Lex → Dist-Freq	98.0	85.5	78.0	87.2	96.0	91.4	85.9	84.1	87.2	87.8
MLE → Lex → Ex-Freq	98.0	85.6	77.2	86.9	96.0	91.3	85.8	83.4	86.9	87.6
MLE → Lex → BERT	98.1	85.5	78.8	87.5	96.2	91.6	86.1	84.6	87.4	88.1
Lex → L3	97.8	85.2	74.1	85.7	95.7	90.7	85.0	80.9	85.5	86.5
Lex → L4	97.3	78.6	74.1	83.4	94.6	88.6	80.3	80.9	83.2	83.9
Lex → L5	97.3	85.2	68.8	83.8	94.9	88.6	82.3	77.5	82.8	83.5
Lex → Dist-Freq	97.8	84.5	75.2	85.8	95.5	90.6	84.4	82.1	85.7	86.3
Lex → Ex-Freq	97.8	85.1	74.6	85.8	95.7	90.9	84.9	81.5	85.8	86.6
Lex → BERT	98.0	85.1	76.5	86.5	95.9	91.3	85.4	83.0	86.6	87.3
Lex → MLE → L3	97.8	85.2	74.5	85.8	95.8	90.9	85.0	81.1	85.7	86.6
Lex → MLE → L4	97.9	82.5	74.5	84.9	95.5	90.8	82.4	81.1	84.8	85.6
Lex → MLE → L5	97.9	85.2	75.3	86.1	95.7	90.8	84.9	82.3	86.0	86.7
Lex → MLE → Dist-Freq	97.9	85.1	75.8	86.2	95.7	91.1	85.1	82.6	86.3	87.0
Lex → MLE → Ex-Freq	97.8	85.1	74.8	85.9	95.7	91.1	84.9	81.6	85.9	86.7
Lex → MLE → BERT	98.0	85.1	76.5	86.5	95.9	91.2	85.3	82.9	86.4	87.2
Tuned-MLE → Lex → BERT	**98.2**	**86.1**	**79.4**	**87.9**	**96.3**	**92.1**	**86.7**	**84.9**	**87.9**	**88.6**

Table 9: Word-level results on the Dev set for all the layered experiments.