# AraDiaWER: An Explainable Metric For Dialectical Arabic ASR

**Abdulwahab Sahyoun** and **Shady Shehata***
Mohamed bin Zayed University of Artificial Intelligence
Abu Dhabi, United Arab Emirates
abdulwahab.sahyoun@mbzuai.ac.ae, shady.shehata@mbzuai.ac.ae

## Abstract

Linguistic variability poses a challenge to many modern ASR systems, particularly Dialectical Arabic (DA) ASR systems dealing with low-resource dialects and resulting morphological and orthographic variations in text and speech. Traditional evaluation metrics such as the word error rate (WER) inadequately capture these complexities, leading to an incomplete assessment of DA ASR performance. We propose AraDiaWER, an ASR evaluation metric for Dialectical Arabic (DA) speech recognition systems, focused on the Egyptian dialect. AraDiaWER uses language model embeddings for the syntactic and semantic aspects of ASR errors to identify their root cause, not captured by traditional WER. MiniLM generates the semantic score, capturing contextual differences between reference and predicted transcripts. CAMeLBERT-Mix assigns morphological and lexical tags using a fuzzy matching algorithm to calculate the syntactic score. Our experiments validate the effectiveness of AraDiaWER. By incorporating language model embeddings, AraDiaWER enables a more interpretable evaluation, allowing us to improve DA ASR systems. We position the proposed metric as a complementary tool to WER, capturing syntactic and semantic features not represented by WER. Additionally, we use UMAP analysis to observe the quality of ASR embeddings in the proposed evaluation framework.

## 1 Introduction

State-of-the-art (SoTA) ASR systems such as Wav2Vec2 XLSR-53 (Baevski et al., 2020), HuBERT (Hsu et al., 2021), and Whisper (Radford et al., 2022) are designed to perform on a wide range of languages, including Arabic speech. To benchmark these models, WER and character error rate (CER) metrics are used to calculate the number of words inserted, substituted, and deleted in transcribed speech. WER then calculates the error percentage by dividing by the total number of words in the predicted transcript. This yields an error rate that quantifies a basic comparison without any language-specific analysis. WER is not designed to consider any form of syntactical or semantic differences in the reference transcript and the predicted transcript, but rather to compare them word by word. This form of calculation poses a gap in the evaluation methodology used for benchmarking ASR systems that deal with multiple languages and dialects of the same language, particularly Dialectical Arabic (DA), which imposes a multitude of morphological and orthographic variations. In the evaluation landscape, metrics present themselves as the source of truth for the quantities they report. However, most metrics used in research today do not give researchers enough insight into the reasoning behind the results and the methodologies used within the metric. This poses a critical issue for explaining results when a system deals with a multitude of morphological variations in speech. To improve the explainability of the results, our research work focuses on proposing a transparent method that provides a new metric named AraDiaWER that is based on WER with a new explainable identity, allowing the metric to report additional semantic and syntactic scores.

It is well established that WER could be used as a benchmark metric for most speech recognition tasks, and in most languages, it works fairly well. However, the challenges imposed by synthetic languages and the lack of syntactic and semantic context of WER, as shown in (Kim et al., 2021), have required researchers to explore methods designed around the language itself. (Ali et al., 2015, 2017; Ali and Renals, 2018; Ali et al., 2019; Ali and Renals, 2020) are five SoTA systems that propose supervised, unsupervised, and objective-based evaluation of Arabic ASR systems. SemDist (Kim et al., 2021), FLORES (Goyal et al., 2021), and the study of lexical distance (Kwaik et al., 2018) provide a

*Corresponding author



*Proceedings of the Second Workshop on NLP Applications to Field Linguistics*, pages 64–73

semantic distance component combined with NER tags and intent recognition to improve the evaluation of ASR systems. Our proposed AraDiaWER metric incorporates semantic and syntactic scoring, fluency scoring, and a UMAP analysis to better explain the performance of DA ASR, while also keeping the traditional metrics (*i.e.,* WER) intact and available for benchmarking purposes.

## 2 AraDiaWER Methodology

To introduce additional syntactic and semantic variances to the existing WER metric and enhance its explainability, we proposed the AraDiaWER metric. We used a weighted sum approach to capture more differences in utterances while maintaining the integrity and distribution of WER.

We designed AraDiaWER to depend on the semantic and syntactic weight generated by other models through a factor we call the error weight or $W_{err}$. The portable dependency on other LMs for the semantic and syntactic scores provides flexibility for other researchers to improve AraDiaWER or adjust the SoTA models used for the syntactic and semantic components to their specific use cases.

To assess the performance of the proposed metric, we fine-tuned a Wav2Vec2-based model with a Connectionist Temporal Classification scorer on a large Arabic speech dataset with more than ten dialects. We compared the performance of the fine-tuned model with five other state-of-the-art ASR systems.

### 2.1 Datasets

This work focuses on evaluating the performance of Dialectical Arabic (DA) ASR systems, which must deal with low-resource dialects and resulting morphological and orthographic variations in text and speech. To evaluate the AraDiaWER metric, datasets that represent the dialectical variations of Arabic, particularly Egyptian dialects are required. We evaluate various ASR systems, including those developed by AALTO (Smit et al., 2017), MIT (Najafian et al., 2017), JHU (Manohar et al., 2017), BUT (Vesely et al., 2017), Mo, and NDSC, and the TDNN-based ASR system in (Ali et al., 2014). All evaluated models vary in their specific implementations but are primarily based on TDNN and LSTM architectures, which are considered hybrid ASR systems. These systems rely on an acoustic model, language model, and lexicons or phonemes to effectively process and recognize dialectal variations

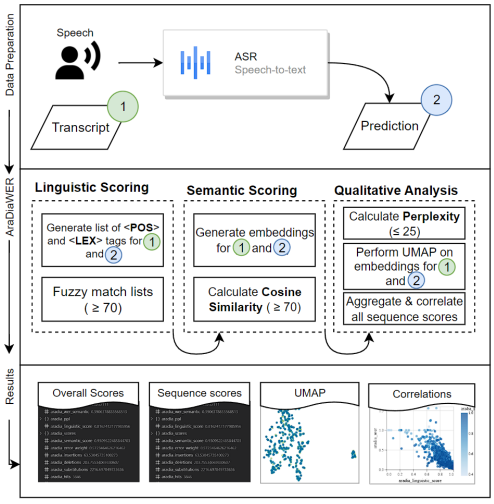



Figure 1: Illustration of AraDiaWER end-to-end approach. The main inputs are the references and predictions.

within Arabic speech.

Our experiments use several datasets that are specifically designed for Arabic speech recognition, including the MGB-3 dataset, which contains 1,000 Egyptian speech samples in the adaptation set and 2,058 samples in the development set. We also use the Arabic subset in FLEURS, which contains 428 samples of Egyptian speech from 180 unique speakers. These datasets were chosen because they represent the specific dialectical variations of Arabic that we aim to evaluate with our metric.

Moreover, we evaluate several SoTA ASR systems, including our fine-tuned AraDia-CTC model, Whisper, Wav2Vec2 XLSR-53, and HuBERT, on the MGB3 test set with 297 samples. The chosen SoTA ASR systems represent the current state of the art in Arabic speech recognition and provide a benchmark for the AraDiaWER metric.

### 2.2 Metric End-to-End Approach

AraDiaWER metric is designed to add explainability to the existing WER metric by incorporating additional syntactic and semantic variances. To achieve this, our framework includes three pipelines: data loading, prediction, and evaluation. The data loading pipeline converts speech audio data into feature tensors, which are then used by the prediction pipeline to transcribe the speech data into text using any ASR system (e.g., Whisper, Wav2Vec2). Once the prediction transcripts are generated, the evaluation pipeline uses two language models to determine the syntactic match and semantic similarity between the reference and

Table 1: Ablation of all syntactic (syntax) tags for the AALTO ASR system using the MGB3 evaluation set.

| Configuration | AraDiaWER | Syntactic Score | Semantic Score | Error Weight |
|---|---|---|---|---|
| syntax7 (pos,lex,prc0,prc1,prc2,prc3,enc0) | 0.268 | 0.863 | 0.925 | 0.559 |
| **syntax6 (pos,lex,prc0,prc1,prc2,prc3)** | **0.268** | **0.863** | **0.925** | **0.559** |
| syntax5 (pos,lex,prc0,prc1,prc2) | 0.268 | 0.855 | 0.925 | 0.561 |
| syntax4 (pos,lex,prc0,prc1) | 0.270 | 0.853 | 0.925 | 0.562 |
| syntax3 (pos,lex,prc0) | 0.269 | 0.848 | 0.925 | 0.564 |
| syntax2 (pos,lex) | 0.271 | 0.837 | 0.925 | 0.567 |
| syntax1 (pos) | 0.270 | 0.844 | 0.925 | 0.565 |

Table 2: Ablation of all syntactic (syntax) tags for the TDNN ASR system using the MGB3 development set.

| Configuration | AraDiaWER | Syntactic Score | Semantic Score | Error Weight |
|---|---|---|---|---|
| syntax7 (pos,lex,prc0,prc1,prc2,prc3,enc0) | 0.604 | 0.708 | 0.833 | 0.648 |
| **syntax6 (pos,lex,prc0,prc1,prc2,prc3)** | **0.604** | **0.708** | **0.833** | **0.648** |
| syntax5 (pos,lex,prc0,prc1,prc2) | 0.605 | 0.698 | 0.833 | 0.653 |
| syntax4 (pos,lex,prc0,prc1) | 0.606 | 0.694 | 0.833 | 0.654 |
| syntax3 (pos,lex,prc0) | 0.607 | 0.685 | 0.833 | 0.658 |
| syntax2 (pos,lex) | 0.611 | 0.665 | 0.833 | 0.670 |
| syntax1 (pos) | 0.607 | 0.690 | 0.833 | 0.656 |

prediction transcripts. These two language models act as the basis for the semantic and syntactic components of the AraDiaWER metric. Figure 1 illustrates the end-to-end AraDiaWER process, highlighting the significance of the syntactic and semantic components in improving the evaluation of DA ASR systems.

The explainability of AraDiaWER with respect to the correlation between substitution/insertion/deletion and semantic and syntactic scores allows researchers to evaluate ASR systems using any chosen language model configuration (see Tables 1 and 2). The AraDiaWER metric consists of two components, the syntactic component, and the semantic component, which capture different aspects of the accuracy and fluency of the predicted transcript. The syntactic component captures changes in parts of speech and lemmas, which are crucial for capturing the grammatical structure of the predicted transcript. On the other hand, the semantic component aims to capture variances in meaning and context, providing a more comprehensive and interpretable evaluation of the DA ASR system. Additionally, the use of embeddings from each LM is essential for extracting an explainable correlation between errors made by the ASR and AraDiaWER's two scores (semantic and syntactic). The transparent and interpretable nature of the AraDiaWER metric facilitates a comprehensive evaluation of the performance of DA ASR systems by accounting for the linguistic, semantic, and fluency features of dialectical Arabic speech, which are not fully captured by the traditional WER metric.

#### 2.2.1 Syntactic Component

The syntactic component of the AraDiaWER assigns morphological and lexical tags to the reference and predicted transcripts. This has been achieved by utilizing a BERT-based disambiguation model (Inoue et al., 2021) out of the box, which uses a pre-trained CAMeLBERT-Mix language model to classify the morphological and lexical features of an input sequence. Firstly, we use CAMeLBERT-Mix LM to determine the syntactic tag. Secondly, we use a unigram-based morpho-syntactic analyzer (Inoue et al., 2022)) to refine the unfactored parent tag (e.g., POS) to an individual subtag (e.g., noun).

For the purpose of our study, the output of the syntactic model was limited to the following tags: parts-of-speech (POS) tags, lemmas (lex), and five clitic features: article proclitic (prc0), preposition proclitic (prc1), conjunction proclitic (prc2), question proclitic (prc3), and pronominal enclitic (enc0). A fuzzy matching algorithm runs on the set of tags assigned for each word and calculates the syntactic score. Tables 1 and 2 show how different syntactic tag configurations affect the final weight $W_{err}$.

The syntactic score in Eq. 3 aims to capture the syntactic variances in the reference and predicted transcripts. It is calculated using the Levenshtein distance (LD) (Eq. 1) between the syntactic characteristics (list of POS, lexicons, and clitics) of the

reference ($L_{\text{ref}}$) and prediction ($L_{\text{hyp}}$). Using the LD, the fuzzy ratio (FR) Eq. (2) is calculated for each pair of words, and the total ratio for the entire sequence is calculated by dividing the sum of the fuzzy ratios by the total number of words (N). This scoring process is repeated for each syntactic tag and the total syntactic score is the sum of the fuzzy ratios of all unfactored tags (POS, lexicons, and clitics) over the total number of sequences, a value bound between 0 and 1. The formulas are as follows.

$$\text{LD}\left(str1, str2\right) = \text{LevDist}\left(str1, str2\right) \quad (1)$$

$$\text{FR}\left(str1, str2\right) = \frac{(\text{len}(str1) + \text{len}(str2) - \text{LD})}{(\text{len}(str1) + \text{len}(str2))} \quad (2)$$

$$\text{ScoreSyn}\left(L_{\text{ref}}, L_{\text{hyp}}\right) = \frac{\sum_i^N (\text{FR}_i\left(L_{\text{ref,i}}, L_{\text{hyp,i}}\right))}{N} \quad (3)$$

#### 2.2.2 Semantic Component

The semantic score of AraDiaWER aims to capture contextual differences between the reference and predicted transcripts by using the pre-trained MiniLM sentence transformer (Wang et al., 2020) out of the box. This language model is designed to perform various NLP tasks, such as feature extraction, question answering, natural language generation, question generation, abstractive summarization, and more. Our semantic scoring component uses the 6-layer all-MiniLM-L6-v2 variant to vectorize the input sequences and perform cosine similarity calculations on the resulting high-dimensional vectors.

Our semantic component focuses specifically on the contextual differences between the reference and predicted transcripts, which are not captured by syntactic information alone. By encoding the prediction and reference transcript pairs ($e_{\text{pre}}$,$e_{\text{ref}}$), using the MiniLM language model, the cosine similarity is calculated to obtain the semantic score, as shown in Eq. 4. This allows the capture of the contextual differences between the reference and predicted transcripts, which are indicative of the semantic differences.

$$\text{ScoreSem}\left(e_{\text{ref}}, e_{\text{hyp}}\right) = \frac{\left(e_{\text{ref}}\right)^T \cdot e_{\text{hyp}}}{\|e_{\text{ref}}\| \cdot \|e_{\text{hyp}}\|} \quad (4)$$

#### 2.2.3 Error Weight & AraDiaWER

In our AraDiaWER metric, we introduced an error weight ($W_{err}$) to determine the influence of semantic and syntactic changes that occur in the language on the estimation of the errors made by ASR systems. Our error weight is based on the theory of weighted sums and weighted averages. In statistics, it is important to account for biases in the data by looking at possible variances within the sample. For example, using variance $\sigma_i^2$, we can compose a weight $\frac{1}{\sigma_i^2}$ that can be used to calculate the weighted average of all measurements to obtain an estimate of a signal. Using the weighted sum approach, we take the syntactic and semantic variances of a sample and build a weight function using the following formula:

$$W_{\text{err}} = \frac{1}{\text{ScoreSem} + \text{ScoreSyn}} \quad (5)$$

By incorporating our error weight into the AraDiaWER metric, we obtain a more comprehensive and interpretable evaluation of DA ASR systems, which takes into account syntactic and semantic variances. The estimated errors are calculated using the new AraDiaWER function, which is a weighted sum of the errors based on their corresponding error weight. WER is computed by summing up all substitutions, insertions, and deletions and dividing them by the total number of words in the reference transcript. AraDiaWER computes WER in terms of a weighted sum of errors, as shown in Eq. 7

WER is the sum of all substitutions, insertions, and deletions (SUB, INS, and DEL) on the total number of words in the reference ($N_{\text{ref}}$), which includes correct words (HIT). The formulas are as follows.

$$\text{WER} = \frac{\text{SUB} + \text{INS} + \text{DEL}}{\text{SUB} + \text{DEL} + \text{HIT}} \quad (6)$$

$$\text{AraDiaWER} = \frac{WER}{ScoreSem + ScoreSyn} \quad (7)$$

The relationship between WER and AraDiaWER in terms of ranking or score correlation can be interpreted as follows: AraDiaWER refines the standard WER by incorporating the error weight, which considers both semantic and syntactic variances. As a result, the AraDiaWER values will generally be correlated with WER but provide a more nuanced ranking of ASR systems, as it accounts for these variances in the Egyptian Arabic dialects.

In order to ensure the interpretability and practical relevance of the AraDiaWER metric, it is necessary to impose a constraint on the syntactic and semantic scores. Specifically, both ScoreSem and ScoreSyn must exceed a threshold of 0.5. This requirement guarantees that the error weight remains within a reasonable range, avoiding excessively large or small values that could undermine the metric's interpretability. By stipulating that ScoreSem and ScoreSyn surpass 0.5, we preserve a balanced representation of the semantic and syntactic variances within the AraDiaWER metric, thereby facilitating more accurate and reliable evaluations of DA ASR systems.

The use of the error weight in our AraDiaWER metric is crucial in assessing the performance of DA ASR systems. The weight determines the importance of semantic and syntactic variances, and it ensures that the evaluation is not biased toward a particular component. This approach allows a better understanding of the performance of ASR systems in dialectical Arabic speech and provides more accurate and reliable evaluations.

### 2.3 Quantitative Analysis of Syntactic and Semantic Errors

The main objective of AraDiaWER is to explain the performance of ASR systems in terms of syntactic and semantic errors. We calculate the Pearson correlation between the WER errors (SUB, INS, DEL) made by the ASR system and the semantic and syntactic scores. The correlation analysis helps to understand which type of errors the ASR system is making and how those errors are reflected in the semantic and syntactic scores. We also utilized p-values to determine the statistical significance of the correlations. By analyzing the correlation and p-values, we can determine the strengths and weaknesses of the ASR system and identify areas for improvement. This information can be used to optimize the ASR system and improve its overall performance. Additionally, the use of AraDiaWER allows for a more interpretable and transparent assessment of the ASR system's performance, making it easier to communicate the results to stakeholders and end-users. The AraDiaWER metric provides a more comprehensive and interpretable assessment of ASR system performance in order to make recommendations based on the traceable assessment. For instance, we can identify the areas where the ASR system is underperforming and recommend improvements to the language models or training data.

### 2.4 Qualitative Analysis using UMAP & Language Models

To analyze the fluency of the predicted transcript, we measure perplexity and combine it with quantitative results to provide a clear assessment of ASR performance. We measure the perplexity score using a dedicated language model. In our implementation, we use GPT-2 base model (Radford et al., 2019) to measure perplexity, and the inverse of perplexity is reported as fluency. Another key component in the quality analysis is the comparison of the reference and prediction embeddings. Our objective is to visualize the semantic embeddings of references and predictions in a low-dimensional space using UMAP to determine overlaps between reference and prediction samples; more dispersed overlaps can indicate better performance.

The UMAP projections for the Whisper ASR model, as shown in Figure 3, provide a way to visualize the quality of the ASR output in a 2D space. By looking at the 2-component UMAP projections for references and hypotheses in different datasets, we can assess the ability of the ASR system to generalize and capture the unique linguistic features of the target dialect. For instance, the UMAP projection for the MGB3 test set, as shown in Figure 3b, shows a low-quality projection, indicating a poor performance of the Whisper model on this dataset. Conversely, the UMAP projection for the FLEURS test set, as shown in Figure 3d, shows an excellent projection, indicating that the Whisper model was able to capture the unique features of this dataset well. The UMAP projections provide an additional tool for evaluating the performance of ASR models, beyond just quantitative metrics. It enables a visual representation of the quality of the ASR output that can aid in identifying areas for improvement and optimizing ASR systems.

The semantic and syntactic scores are used in conjunction with other evaluation metrics, such as the ASR model fluency and the quality of UMAP projections, to provide a more comprehensive and interpretable assessment of the performance of DA ASR systems. Figure 4 shows the comparison between the Whisper scores and the transcript fluency and overall quality extracted from UMAP projections. The figure highlights the negative correlation between the semantic scores and transcript fluency,

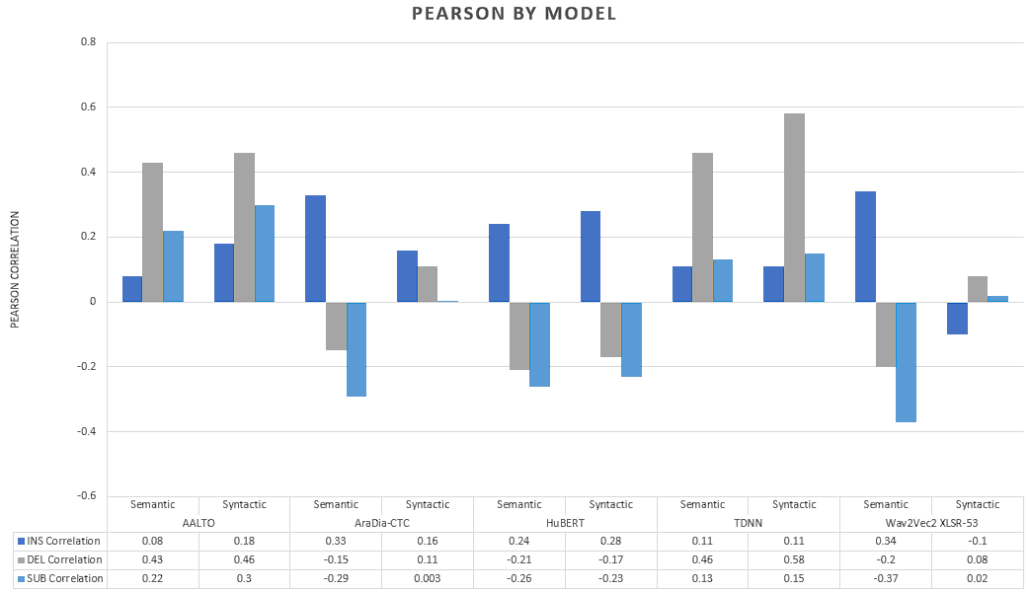



Figure 2: A grouped bar chart showing the semantic and syntactic correlations of different ASR models with AraDiaWER. The bars show the Pearson correlation coefficients of the WER SUB, DEL, and INS. The results indicate that the correlations between semantic errors and WER are generally negative, while the correlations between syntactic errors and WER are generally positive. AALTO and TDNN show strong correlations with both semantic and syntactic errors, while Wav2Vec2 XLSR-53 and HuBERT show negative correlations with semantic errors and weaker positive correlations with syntactic errors. AraDia-CTC, on the other hand, shows strong negative correlations with semantic errors and positive correlations with syntactic errors. In the context of the paper, positive correlation means that when the WER errors increase, the corresponding semantic or syntactic errors also increase, while negative correlation means that when the WER errors decrease, the corresponding semantic or syntactic errors decrease.

| | Avg | Semantic Correlation with WER | | | Syntactic Correlation with WER | | |
|---|---|---|---|---|---|---|---|
| Model | Sem/Syn Error | SUB Pearson / pVal | DEL Pearson / pVal | INS Pearson / pVal | SUB Pearson / pVal | DEL Pearson / pVal | INS Pearson / pVal |
| AALTO | 0.07 / 0.16 | 0.22 / 1.08E-11 | 0.43 / 2.21E-47 | 0.08 / 1.41E-02 | 0.30 / 2.22E-21 | 0.46 / 1.86E-54 | 0.18/ 5.43E-09 |
| TDNN | 0.16 / 0.33 | 0.13 / 8.76E-09 | 0.46 / 4.28E-109 | 0.11 / 3.86E-07 | 0.15 / 5.83E-12 | 0.58 / 7.42E-184 | 0.11 / 7.97E-07 |
| Wav2Vec2 XLSR-53 | 0.19 / 0.47 | -0.37 / 2.97E-11 | -0.20 / 6.20E-04 | 0.34 / 1.96E-09 | 0.02 / 7.61E-01 | 0.08 / 1.66E-01 | -0.10 / 8.08E-02 |
| HuBERT | 0.15 / 0.30 | -0.26 / 4.47E-06 | -0.21 / 2.34E-04 | 0.24 / 2.18E-05 | -0.23 / 8.56E-05 | -0.17 / 3.67E-03 | 0.28 / 9.42E-07 |
| AraDia-CTC | 0.15 / 0.33 | -0.29 / 5.83E-07 | -0.15 / 1.19E-02 | 0.33 / 5.17E-09 | 0.003 / 9.51E-01 | 0.11 / 5.82E-02 | 0.16 / 4.75E-03 |

Table 3: AraDiaWER Correlations with Semantic and Syntactic Errors

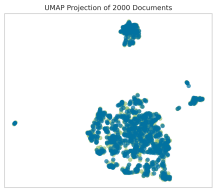
| System | Dataset | WER | AraDia WER | RMSE |
|---|---|---|---|---|
| AALTO | MGB3(A) | 0.400 | 0.268 | 0.11 |
| TDNN | MGB3(D) | 0.710 | 0.604 | **0.09** |
| Wav2Vec2 XLSR-53 | FLEURS | 0.600 | 0.470 | 0.12 |
| HuBERT | FLEURS | 0.480 | 0.330 | 0.14 |
| AraDia-CTC | FLEURS | 0.540 | 0.400 | 0.13 |
| Whisper | FLEURS | **0.210** | **0.120** | 0.10 |

Table 4: Results on the MGB-3 and FLEURS datasets extracted from the study. (D) is the development set and (A) is the adaptation set.

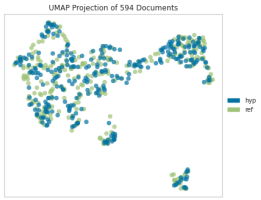
| System | WER | AraDia WER | RMSE |
|---|---|---|---|
| AALTO | **0.400** | **0.268** | 0.11 |
| TDNN | 0.710 | 0.604 | **0.09** |
| Wav2Vec2 XLSR-53 | 0.753 | 0.660 | **0.09** |
| HuBERT | 0.733 | 0.580 | 0.18 |
| AraDia-CTC | 0.695 | 0.575 | 0.12 |
| Whisper | 0.565 | 0.446 | 0.15 |

Table 5: Average results across all tests. AALTO captures the Egyptian dialect well, while Whisper is capable of generalizing to any dataset.

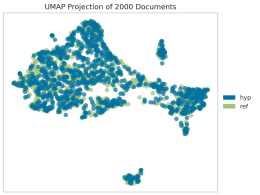
indicating that the higher the semantic score, the lower the fluency of the transcript. On the other hand, the transcript quality is impacted the most when the ASR model commits more syntactic errors. This suggests that the syntactic score is more sensitive to the variations in the ASR output across different dialects, making it a useful tool for identifying areas for improvement in DA ASR systems.



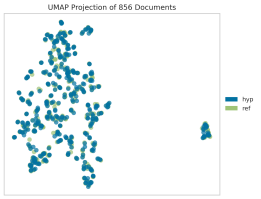

(a) 2-component UMAP for references and hypotheses in CommonVoice 6.1 (Good)




(b) 2-component UMAP for references and hypotheses in MGB3 test set (Poor)




(c) 2-component UMAP for references and hypotheses in MGB2 train set (Good)




(d) 2-component UMAP for references and hypotheses in FLEURS test set (Excellent)

Figure 3: UMAP projections for Whisper ASR model in different datasets. The scatter plots show 2-component UMAP projections of references and hypotheses for (a) CommonVoice 6.1, (b) MGB3 test set, (c) MGB2 train set, and (d) FLEURS test set. The UMAP projections help assess the quality of the ASR output and the similarity between reference and hypotheses.

The use of multiple evaluation metrics, including the semantic and syntactic scores, transcript fluency, and UMAP projections, enables us to obtain a more complete picture of the performance of DA ASR systems and make more informed recommendations for improving their performance.

## 3 Results and Analysis

Table 3 illustrates the results of our experiments, which aim to evaluate the AraDiaWER metric's effectiveness in assessing ASR systems in dialectal Arabic. The evaluated models include AALTO, TDNN, Wav2Vec2 XLSR-53, HuBERT, and AraDia-CTC. The average semantic and syntactic errors are presented in the table. We observe that the TDNN model has the highest average semantic and syntactic errors, followed by Wav2Vec2 XLSR-53, AraDia-CTC, AALTO, and HuBERT. The results show that the AraDiaWER metric is effective in capturing the syntactic and semantic errors of ASR systems and that different LM models have varying degrees of performance in capturing these errors. Our approach relies on the semantic and syntactic components of AraDiaWER. The semantic component measures the variances in meaning and context between the reference and predicted transcripts, while the syntactic component captures the syntactic variations in dialectal utterances. The results show that the semantic correlation with WER is generally negative, while the syntactic correlation with WER is positive. The AALTO and TDNN models have high syntactic correlations with WER, indicating that these models have significant syntactic errors. On the other hand, Wav2Vec2 XLSR-53, HuBERT, and AraDia-CTC have low syntactic correlations with WER, indicating that these models have low syntactic errors. Furthermore, the AraDia-CTC model has the highest semantic correlation with WER, indicating that it has the highest semantic errors among the models evaluated. Conversely, the TDNN model has the lowest semantic correlation with WER, indicating that it has the lowest semantic errors among the models. The experimental results show that the p-values of the correlations are all statistically significant ($p < 0.05$) which provides insight into the underlying factors that contribute to the ASR system's performance.

Tables 4 and 5 summarize the results for each system. The averaged results of AALTO showed that the best performance is observed when a system is trained and tested in a fully supervised approach on the same distribution and language. The results of TDNN show that even legacy systems can perform well when it comes to capturing syntactic and semantic patterns. This is further proven in the ablation studies for AALTO and TDNN, where the number of syntactic tags captured is negatively correlated with the penalty-reducing error weight $W_{err}$ (see Tables 1 and 2). Linking this to the correlation analysis in AALTO, it is possible to deduce that capturing additional syntactic tags can lead to improved syntactic scores and better overall capture of dialectical variations in utterances, decreasing the error weight and AraDiaWER value. Higher semantic scores indicate a better contextual understanding of the utterance, allowing for more accurate prediction of words that are similar and reducing the RMSE between WER and AraDiaWER. In addition, less complex utterances observe higher syntactic and semantic scores. Lastly, certain outliers in the dataset still achieve high semantic and syntactic scores but fail at fluency; this shows the metric's ability to pinpoint low-quality utterances that are not intelligible. The inclusion of RMSE in the calculation of the results serves as a means of

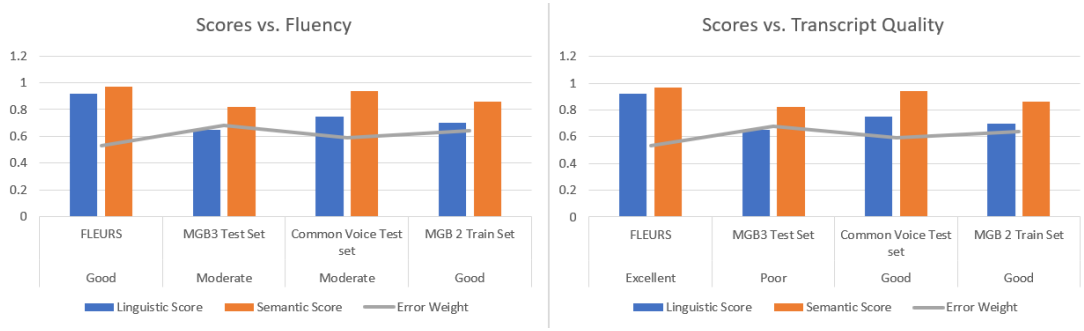



Figure 4: Comparison of Whisper ASR scores with transcript fluency and overall quality extracted from UMAP projections. The labels 'Excellent', 'Good', 'Moderate', and 'Poor' indicate the visual quality of the UMAP projections and the range of Perplexity values for the ASR model output. The scatter plots show the correlation between the semantic, syntactic, and AraDiaWER scores with the transcript fluency and overall quality. The results indicate that AraDiaWER is positively correlated with the overall quality of the ASR output, while the semantic and syntactic scores show a stronger correlation with transcript fluency. These findings highlight the usefulness of AraDiaWER as a more comprehensive and interpretable metric for evaluating DA ASR systems.

quantifying the differences between WER and AraDiaWER. By measuring the RMSE, we can assess the degree of agreement between the two metrics and determine the extent to which AraDiaWER captures variations in dialectal Arabic speech that are not fully represented by the traditional WER. This additional analysis provides further insight into the strengths and limitations of AraDiaWER, enabling researchers and practitioners to better understand the implications of adopting this new metric in the context of DA ASR systems evaluation.

The experimental study reveals that the use of AraDiaWER brings an average improvement of 18.65% in error rate compared to WER. This improvement does not necessarily suggest that our metric is a direct replacement for WER or that it outperforms it in all aspects. Rather, our approach offers a transparent and traceable method that utilizes language models to evaluate DA ASR systems in a more comprehensive and interpretable manner.

## 4 Conclusion

The focus of this paper is to propose an explainable evaluation metric, AraDiaWER, that complements WER and is designed to assess the performance of Automatic Speech Recognition (ASR) systems for dialectal Arabic speech. This metric combines three different scoring systems, namely syntactic, semantic, and fluency, by utilizing state-of-the-art models. The main objective of AraDiaWER is to provide a more detailed and inclusive assessment of the performance of ASR systems in the context of dialectal Arabic speech, which is a significant improvement compared to the conventional word error rate (WER) metric alone.

This work can be considered a resource tool to capture the dialectal variations in speech, where the addition of syntactic features, such as parts of speech tags and lemmas, is helpful for improving the overall performance of the metric. Moreover, the incorporation of semantic features allows the ASR to be evaluated based on meaning, thus ensuring a more holistic assessment of the ASR system. Therefore, we do not seek to undermine the importance of WER but to offer a complementary tool that enables a more extensible evaluation of DA ASR systems

The AraDiaWER framework relies on language models (LMs) to extract both syntactic and semantic features from the text. While LMs are primarily trained for syntactic features, they also contain information about semantic features. The syntactic component assigns morphological and lexical tags to the text using the embeddings of CAMeLBERT-Mix LM. The semantic component uses the embeddings of MiniLM and cosine similarity to calculate the semantic similarity between the reference and predicted transcripts. The embeddings of each LM are used to extract explainable correlations between errors made by the ASR and AraDiaWER's two scores (semantic, and syntactic). This allows us to capture both semantic and syntactic features and make more comprehensive and interpretable assessments of the ASR systems. Additionally, the proposed evaluation framework uses a UMAP analysis to evaluate the semantic

patterns in a low-dimensional space and the GPT-2 generated perplexity score to determine the fluency of an utterance.

In conclusion, while there is still room for improvement, our proposed AraDiaWER metric represents a step forward in the comprehensive evaluation of ASR systems, especially in the context of dialectal variations. In future work, we plan to further improve the metric by incorporating multilingual language models to capture additional morphological and orthographic patterns in the transcripts, target a wider range of diverse datasets, and use modern LMs like GPT-4, LaMDA, and LLaMA to interpret perplexity and AraDiaWER results even further for a more detailed analogy.

## Limitations

One of the limitations is the reliance on the available language models for calculating the semantic and syntactic scores. The quality of these scores may depend on the training data and domain specificity, which may have an impact on the generalizability of our findings. Additionally, the scope of our experiments is limited to one set of Arabic dialects, namely Egyptian, which may not be representative of all dialectical variations in the language. Further work is needed to evaluate the effectiveness of the AraDiaWER metric on a wider range of dialects and to improve the quality of the language models used in our study.

## Ethics Statement

In compliance with the ACL Ethics Policy, we acknowledge the potential ethical considerations associated with this research on automatic speech recognition for dialectical Arabic. The proposed AraDiaWER metric is intended to provide a more comprehensive and explainable evaluation of DA ASR systems that can better account for dialectical variations. However, we acknowledge the potential impact of any inaccuracies in the system, particularly regarding sociocultural implications. As such, we urge caution in the use and application of this metric and encourage future research to further explore the impact of such technology on diverse groups and communities. We are committed to ethical research practices and will prioritize transparency and accountability in all future studies.

## References


Ahmed Ali, Salam Khalifa, and Nizar Habash. 2019. Towards variability resistant dialectal speech evaluation: 20th Annual Conference of the International Speech Communication Association: Crossroads of Speech and Language, INTERSPEECH 2019. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2019-September:336–340.

Ahmed Ali, Walid Magdy, and Steve Renals. 2015. Multi-Reference Evaluation for Dialectal Speech Recognition System: A Study for Egyptian ASR. In *Proceedings of the Second Workshop on Arabic Natural Language Processing*, pages 118–126, Beijing, China. Association for Computational Linguistics.

Ahmed Ali, Preslav Nakov, Peter Bell, and Steve Renals. 2017. WERd: Using Social Text Spelling Variants for Evaluating Dialectal Speech Recognition. Technical Report arXiv:1709.07484, arXiv. ArXiv:1709.07484 [cs] type: article.

Ahmed Ali and Steve Renals. 2020. Word Error Rate Estimation Without ASR Output: e-WER2. Technical Report arXiv:2008.03403, arXiv. ArXiv:2008.03403 [cs, eess] type: article.

Ahmed Ali, Yifan Zhang, Patrick Cardinal, Najim Dahak, Stephan Vogel, and James Glass. 2014. A complete KALDI recipe for building Arabic speech recognition systems. In *2014 IEEE Spoken Language Technology Workshop (SLT)*, pages 525–529.

Ahmed M. Ali and S. Renals. 2018. Word Error Rate Estimation for Speech Recognition: e-WER. In *ACL*.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. *arXiv:2006.11477 [cs, eess]*. ArXiv: 2006.11477.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzman, and Angela Fan. 2021. The FLORES-101 Evaluation Benchmark for Low-Resource and Multilingual Machine Translation. ArXiv:2106.03193 [cs].

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. *arXiv:2106.07447 [cs, eess]*. ArXiv: 2106.07447.

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. ArXiv:2110.06852 [cs].

Suyoun Kim, Abhinav Arora, Duc Le, Ching-Feng Yeh, Christian Fuegen, Ozlem Kalinli, and Michael L. Seltzer. 2021. Semantic Distance: A New Metric for ASR Performance Analysis Towards Spoken Language Understanding. ArXiv:2104.02138 [cs].

Kathrein Abu Kwaik, Motaz Saad, Stergios Chatzikyriakidis, and Simon Dobnika. 2018. A Lexical Distance Study of Arabic Dialects. *Procedia Computer Science*, 142:2–13.

Vimal Manohar, Daniel Povey, and Sanjeev Khudanpur. 2017. JHU Kaldi system for Arabic MGB-3 ASR challenge using diarization, audio-transcript alignment and transfer learning. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 346–352.

Maryam Najafian, Wei-Ning Hsu, Ahmed Ali, and James Glass. 2017. Automatic speech recognition of Arabic multi-genre broadcast media. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 353–359.

Alec Radford, John Kim, and Xu. 2022. Robust Speech Recognition via Large-Scale Weak Supervision.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Peter Smit, Siva Reddy Gangireddy, Seppo Enarvi, Sami Virpioja, and Mikko Kurimo. 2017. Aalto system for the 2017 Arabic multi-genre broadcast challenge. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 338–345.

Karel Vesely, Murali Karthick Baskar, Mireia Diez, and Karel Benes. 2017. MGB-3 but system: Low-resource ASR on Egyptian YouTube data. pages 368–373.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers. ArXiv:2002.10957 [cs].