NADI 2025: The First Multidialectal Arabic Speech Processing Shared Task

Bashar Talafha[λ] **Hawau Olamide Toyin**[ξ] **Peter Sullivan**[λ] **AbdelRahim Elmadany**[λ]
Abdurrahman Juma[γ] **Amirbek Djanibekov**[ξ] **Chiyu Zhang**[λ] **Hamad Alshehhi**[ξ]
Hanan Aldarmaki[ξ] **Mustafa Jarrar**[αγ] **Nizar Habash**[ηξ] **Muhammad Abdul-Mageed**[λ]
[λ]The University of British Columbia [α]Hamad Bin Khalifa University
[γ]Birzeit University [ξ]MBZUAI [η]NYU Abu Dhabi
{btalafha@mail.,a.elmadany@,muhammad.mageed@}ubc.ca

Abstract

We present the findings of the sixth Nuanced Arabic Dialect Identification (NADI 2025) Shared Task, which focused on Arabic speech dialect processing across three subtasks: spoken dialect identification (Subtask 1), speech recognition (Subtask 2), and diacritic restoration for spoken dialects (Subtask 3). A total of 44 teams registered, and during the testing phase, 100 valid submissions were received from eight unique teams. The distribution was as follows: 34 submissions for Subtask 1 "*five teams*", 47 submissions for Subtask 2 "*six teams*", and 19 submissions for Subtask 3 "*two teams*". The best-performing systems achieved 79.8% accuracy on Subtask 1, 35.68/12.20 WER/CER (overall average) on Subtask 2, and 55/13 WER/CER on Subtask 3. These results highlight the ongoing challenges of Arabic dialect speech processing, particularly in dialect identification, recognition, and diacritic restoration. We also summarize the methods adopted by participating teams and briefly outline directions for future editions of NADI.[1]

1 Introduction

Spoken Arabic exhibits a remarkable degree of linguistic diversity. Beyond Modern Standard Arabic (MSA) and Classical Arabic (CA), which have historically dominated computational work, Arabic encompasses numerous regional and national dialects that differ across all linguistic levels (phonology, morphology, lexicon, and syntax) and in discourse/pragmatics (Talafha et al., 2024; Jarrar et al., 2023). These varieties also frequently exhibit intra- and inter-sentential code-switching with other languages (Abdul-Mageed et al., 2024). These varieties dominate everyday communication across the Arab world yet remain under-represented in annotated datasets and resources (Bouamor et al., 2018; Darwish et al., 2021; Abdul-Mageed et al., 2020, 2023). At the same time, many downstream applications—from automatic transcription and virtual assistants to text-to-speech and educational tools—depend on accurate handling of dialectal speech and the diacritics that indicate short vowels and phonological features. Existing systems trained on CA/MSA (Elmadany et al., 2023a; Toyin et al., 2023) often ignore these diacritics or assume text forms, leaving a large gap between technology and real-world usage.

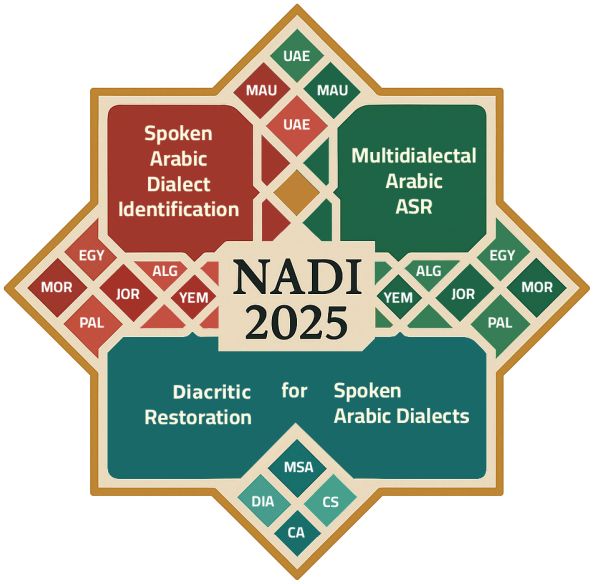

Figure 1: Overview of the **NADI 2025** shared tasks.

NADI shared task series, hosted at the Arabic-NLP conference[2] since 2020, was created to alleviate this bottleneck by providing curated datasets and standardized evaluation settings for dialect identification, translation and related tasks (Abdul-Mageed et al., 2020, 2021; Abdul-Mageed et al., 2022, 2023, 2024). These earlier, text-focused editions—together with the general observation that

[1]The official leaderboards and datasets for NADI 2025 are available at https://nadi.dlnlp.ai/2025.

[2]Formerly the Workshop on Arabic Natural Language Processing, WANLP

Arabic dialects remain under-studied due to limited resources—motivate a shift in NADI 2025 toward speech and diacritization.

NADI 2025 marks the *sixth* edition of the NADI shared task series, hosted by the Third Arabic Natural Language Processing Conference (ArabicNLP 2025[3]). In the following, we introduce several key new features that set it apart from previous versions, focusing on the challenges of real-world, spoken Arabic dialects:

A unified speech processing benchmark. This edition brings together three distinct but complementary tasks, "*dialect identification*", "*automatic speech recognition*", and "*diacritic restoration*", under one umbrella. This creates a comprehensive benchmark for evaluating system performance across the full spectrum of challenges in Arabic speech processing.

New evaluation datasets and unified benchmarking framework. We introduce a comprehensive suite of newly-curated datasets across all three subtasks. This includes a high-quality blind test sets *eight-hours* speech corpus for spoken dialect identification, a large-scale $10,807$ *utterances* for ASR, and a $1,332$ utterances for diacritic restoration, all covering diverse Arabic varieties. Beyond the data itself, NADI 2025 establishes a robust and unified evaluation framework featuring large-scale blind test sets to ensure fair comparison. This framework introduces novel paradigms, such as benchmarking model *adaptation* in the ADI task and offering distinct *open* and *closed* tracks for Diacritic Restoration.

A novel diacritic restoration task. We introduce the first shared task for diacritic restoration that moves beyond formal written Arabic (CA and MSA) to target *spoken dialects* and *code-switched language*. The task is uniquely designed to encourage multimodal solutions that leverage both speech and text as input.

Figure 1 provides a schematic overview of the NADI 2025 shared task, illustrating its three main subtasks including Spoken Arabic Dialect Identification, which covers *eight* regional dialects as "*Algerian*" (ALG), "*Egyptian*" (EGY), "*Emirati*" (UAE), "*Jordanian*" (JOR), "*Mauritanian*" (MAU), "*Moroccan*" (MOR), "*Palestinian*" (PAL), and "*Yemeni*" (YEM); Multidialectal Arabic ASR, which targets the exact same set of dialects; and Diacritic Restoration for Spoken Arabic Dialects, which encompasses MSA, mixed dialects, code-switched varieties, and CA.

The rest of the paper is organized as follows: Section 2 provides a review of related work on spoken Arabic processing and the history of the NADI shared task. In Section 3, we describe the NADI 2025 shared task in detail, including the three subtasks, their datasets, and evaluation metrics. Section 4 presents the results for all participating teams and baselines, followed by an overview of the submitted systems in Section 5. We conclude the paper in Section 6.

2 Literature Review

Unlike previous NADI tasks that relied on text, NADI 2025 concentrates on spoken Arabic dialects. Accordingly, this section covers related work on the subtasks of spoken language identification, ASR, and diacritic restoration. Before delving into the related work, it is useful to explore the history of NADI and its growth since its inception.

2.1 NADI Shared Task: Origins and Growth

`NADI-2020`, the first NADI shared task (Abdul-Mageed et al., 2020) involved two subtasks, one targeting country level (21 countries) and another focusing on province level (100 provinces), both exploiting X, *formerly Twitter*, data. NADI 2020 was the first shared task to exploit naturally occurring fine-grained dialectal text at the sub-country level.

`NADI-2021`, the second version (Abdul-Mageed et al., 2021) targeted the same 21 Arab countries and 100 corresponding provinces as NADI 2020, also using X data. However, it improved upon the previous version by removing non-Arabic data and distinguishing between MSA and dialectical Arabic (DA). It involved four subtasks: MSA-country, DA-country, MSA-province, and DA-province.

`NADI-2022` (Abdul-Mageed et al., 2022) continued the focus on studying Arabic dialects at the country level, but also included dialectal sentiment analysis with an objective to explore variation in socio-geographical regions that had not been extensively studied before.

`NADI-2023`, the fourth edition (Abdul-Mageed et al., 2023), proposed new machine translation subtasks from four dialectal Arabic varieties to MSA, in two themes (open-track and closed-track) as well as a dialect identification subtask at the country level.

[3] https://arabicnlp2025.sigarab.org/

Finally, NADI-2024, the fifth edition (Abdul-Mageed et al., 2024), targeted both dialect ID cast as a multi-label task, identification of the Arabic level of dialectness, and dialect-to-MSA machine translation.

2.2 Spoken Dialect Identification

Although CA and MSA have been extensively examined (Harrell, 1962; Badawi, 1973; Brustad, 2000; Holes, 2004), dialectal Arabic (DA) became the center of attention only relatively recently. A significant challenge in studying DA has been the scarcity of resources, prompting researchers to create new DA datasets targeting limited regions (Gadalla et al., 1997; Diab et al., 2010; Al-Sabbagh and Girju, 2012; Sadat et al., 2014; Harrat et al., 2014; Jarrar et al., 2016; Khalifa et al., 2016; Al-Twairesh et al., 2018; Alsarsour et al., 2018; Abu Kwaik et al., 2018; El-Haj, 2020; Haff et al., 2022; Nayouf et al., 2023; Jarrar et al., 2023). Several works introducing multi-dialectal datasets and models for region-level dialect identification (Zaidan and Callison-Burch, 2011; Elfardy et al., 2014; Bouamor et al., 2014; Meftouh et al., 2015) and the VarDial workshop series employing transcriptions of speech broadcasts (Malmasi et al., 2016) also followed. Other work developed relatively small-sized commissioned data (Bouamor et al., 2018; Salameh et al., 2018; Obeid et al., 2019).

Subsequently, larger datasets that cover between 10 to 21 countries were introduced (Mubarak and Darwish, 2014; Abdul-Mageed et al., 2018; Zaghouani and Charfi, 2018; Abdul-Mageed et al., 2020; Abdelali et al., 2021; Issa et al., 2021; Baimukan et al., 2022; Althobaiti, 2022; Elleuch et al., 2025; Hamad et al., 2025). The majority of these datasets are compiled from social media posts, especially X (formerly Twitter). More recently, benchmarks such as ORCA (Elmadany et al., 2023b) and DOLPHIN (Nagoudi et al., 2023) boast dialectal coverage.

Spoken dialect ID shares with text-based dialect ID a scarcity of labeled data. Important efforts to counter this include the introduction of the multi-genre and multidialectal ADI-5 (Ali et al., 2017) and ADI-17 (Ali et al., 2019; Shon et al., 2020) datasets (covering coarse regional and fine-grain country-level dialects, respectively). Moving from text to speech as a modality, however, introduces additional complexities such as potential channel mismatch between train and test sets due to differences in recording conditions, as is in the case with ADI-5 (Ali et al., 2017). Furthermore, dialect ID models may capture non-linguistic information such as gender and channel features (Chowdhury et al., 2020), and may experience major performance degradation in cross-domain and cross-dialect settings (Sullivan et al., 2023; Hamad et al., 2025).

2.3 Automatic Speech Recognition

Arabic ASR systems often struggle with dialectal speech, primarily due to lack of (or limited) dialectal data (Waheed et al., 2023). Mozilla Common Voice (Ardila et al., 2020) and MASC (MSA and Dialectal Speech) (Al-Fetyani et al., 2022) were introduced to alleviate this issue. However, both of these corpora label the data under a single label (Arabic) instead of different dialect names. Some of the audio and text samples in these datasets are also misaligned (Lau et al., 2025). The Casablanca project (Talafha et al., 2024) compiled high quality multidialectal speech for eight countries, providing a significant boost towards research in multidialectal ASR. Djanibekov et al. (2025) have also recently presented strong results for dialectal Arabic ASR as well as training strategies that work best based on data availability for each dialect.

2.4 Diacritic Restoration

Several text-based approaches (Alasmary et al., 2024; Elgamal et al., 2024; Fadel et al., 2019; Harrat et al., 2013) and resources (Toyin et al., 2025; Zerrouki and Balla, 2017) have been proposed for Arabic diacritic / vowel restoration. Aldarmaki and Ghannam (2023) found the speech based approach to outperform text only approaches. More recently, speech based or multi-modal approaches have also been proposed -albeit at a slow rate, mainly due to lack of parallel speech-text resources (Shatnawi et al., 2024). Elmadany et al. (2023a) report strong diacritization models as part of the Octopus toolkit, based on simple finetuning of AraT5 (Elmadany et al., 2022). Shatnawi et al. (2024) also propose an ASR-based diacritic restoration framework, where a pretrained ASR model generates vowelized transcripts refined by a secondary diacritization model. While their approach achieved high accuracy for CA, it fails to generalize to dialectal Arabic due to dataset limitations.

3 NADI 2025

NADI 2025 is the sixth edition of NADI shared task series. Since we extend the scope of the shared task to address broader challenges in multidialectal Arabic speech processing, we refer to NADI 2025 as "*the first multidialectal Arabic speech processing shared task*". This edition comprises three complementary subtasks: spoken Arabic dialect identification, multidialectal Arabic ASR, and diacritization restoration. Collectively, these subtasks target critical components of the Arabic speech technology pipeline, each addressing long-standing challenges arising from the language's rich dialectal variation, frequent code-switching, and the absence of diacritics in most written Arabic. By curating diverse, high quality datasets and establishing standardized evaluation protocols, NADI 2025 aims to catalyze the development of robust, generalizable systems that advance state of the art in Arabic speech and language processing.

3.1 Subtask 1 - Spoken Dialect Identification

Task Description. This subtask is an 8-way classification task to identify which of country-level dialect is being spoken in an utterance, with our set of countries being *Algeria, Egypt, Jordan, Mauritania, Morocco, Palestine, United Arab Emirates (UAE), and Yemen*.

Data. In this subtask, we follow similar procedure in selecting utterances to the data collection procedure of Casablanca (Talafha et al., 2024). For each dialect, different series were identified and the dialect spoken was verified by fluent speakers. For the Adaptation set, we utilize the same series as in Casablanca, but ensure there is no overlap with the series used for the Test set. By doing so, we aim to minimize the influence of potentially overlapping speakers, and to try to disentangle the dialect ID task from simple domain classification.

Evaluation Metric. We use both accuracy as well as the Language Recognition Evaluation 2022 average Cost metric (C_{avg}) (Lee et al., 2022). Because Cost is based on the probability of missed detections as well as false alarms for a given system it provides a complementary way to characterize model performance. At a high level, for two models that have similar accuracy but different Cost, the lower Cost model will providing a larger positive margin between the probability of the correct classes in comparison to incorrect classes, while the higher Cost model would have a smaller margin between correct and incorrect class probabilities.

3.2 Subtask 2 - Multidialectal Arabic ASR

Task Description. The ASR subtask2 in NADI-2025 focuses on building speech recognition systems that can handle spoken Arabic across a range of regional dialects: *Algerian, Egyptian, Jordanian, Mauritanian, Moroccan, Palestinian, Emirati, and Yemeni*. The task includes both monolingual and code-switched speech, which captures the variation speakers naturally use in different settings.

Data. The dataset used in this subtask is a subset of the Casablanca corpus (Talafha et al., 2024). In this subtask, we select balanced samples from each dialect. The training set is intended primarily for adaptation rather than full model training, encouraging participants to leverage transfer learning, domain adaptation, and other data-efficient strategies. We provide a total of $47,027$ utterances, evenly distributed across the eight dialects for the training, validation, and test sets ($1,600$ utterances per dialect per split). The only exceptions are Algeria, Palestine, and Yemen, which have 727, 900, and $1,180$ utterances, respectively, in the test set. These lower counts are due to the limited availability of samples for these dialects in the original Casablanca dataset.

Evaluation Metric. System performance is evaluated using the word error rate (WER) as the primary metric, reported both overall and per dialect. We also report character error rate (CER)[4] for additional insight into system performance, particularly for short utterances and morphologically rich forms. During evaluation, in line with Talafha et al., 2024, we apply a consistent text normalization pipeline to both system outputs and reference transcripts. Specifically, we: (a) retain only the % symbol, removing other special characters, (b) eliminate diacritics, (c) normalize Hamzas and Maddas to bare alif (ا), (d) convert Eastern Arabic numerals to Western Arabic numerals (e.g., ٢٩ becomes 29), and (e) preserve all Latin characters, as Casablanca contains code-switching segments in other languages.

Subtask 3 - Diacritic Restoration for Spoken Arabic Varieties. This subtask aims to advance

[4] In the case of a tie, we use the average CER as the tiebreaker.

Dataset	Type	Diacritized	Train	Dev	Test (Ours)
MDASPC	Multi-dialectal	True	60,677	—	5,164
TunSwitch	Dialectal, CS	True	5,212	165	110 (**110**)
ArzEn	Dialectal, CS	False	3,344	1,402	1,470 (**104**)
Mixat	Dialectal, CS	False	3,721	—	1,583 (**100**)
ClArTTS	CA	True	9,500	—	204
ArVoice	MSA	True	2,507	258	(**11**)
MGB2	MSA	False	—	—	5,365 (**40**)

Table 1: Number of sentences in datasets provided for the diacritic restoration sub-task. **Ours.** refers to the held-out test set for this shared task which we manually diaritize. **CA.** refers to Classical Arabic. **CS.** refers to code-switching.

research on automatic diacritic restoration for spoken Arabic varieties. As the vast majority of existing vowelization or diacritic restoration efforts focus on CA or MSA, we aim to raise attention to more challenging spoken varieties, such as dialects and code-switching, with a focus on generalization across different varieties. The objective of this subtask is to restore the diacritics of a given text. The text can be in a variety of forms, including MSA and Arabic dialects and may even include code-switched instances. In addition to text, all inputs have an associated speech utterance to encourage multi-modal approaches.

Data. This subtask encourages the development of multi-modal (speech + text) diacritic restoration models that generalize across Arabic variants. To enable the development of such models, we identified several high-quality data sets (Almeman et al., 2013; Abdallah et al., 2023; Al Ali and Aldarmaki, 2024; Hamed et al., 2020; Kulkarni et al., 2023; Toyin et al., 2025) of Arabic variants (CA, MSA, dialectal, CS) that include parallel speech and text. Table 1 shows a summary of the data sets provided to the participants for this subtask. The MDASPC dataset contains multi-dialectal speech with diacritized transcriptions and we include it for training. For the *TunSwitch* (Abdallah et al., 2023) training data, we used GPT-4o with a chain-of-thought prompt to initially diacritize the transcriptions. The diacritized output of GPT-4o was subsequently manually corrected with the corresponding audio as a reference by a native Arabic speaker. For code-switching, we provide undiacritized resources for training; *ArzEn* (Hamed et al., 2020), *Mixat* (Al Ali and Aldarmaki, 2024) and *MGB2* (Ali et al., 2016); for each dataset, we provide diacritized test sets by manually annotating random subsets of their test set.

Evaluation Metric. Similar to subtask 2, we use WER and CER as performance metrics for this subtask, which are chosen to enable the evaluation of diacritic restoration performance even for models that may change the underlying text, such as ASR-based or sequence-to-sequence models.

4 Shared Task Teams & Results

4.1 Participating Teams

A total of 44 teams registered for the NADI 2025. At the testing phase, a total of 100 valid entries were submitted by *eight* unique teams. The breakdown across the subtasks as follow: 34 submissions for subtask 1 by *five* teams, 47 submissions for subtask 2 by *six* teams and 19 submissions by *two* teams for subtask 3. Table 2 list NADI 2025 participated teams which completed the testing phase.

4.2 Baselines

We developed baseline (BL) models for each subtask to serve as reference points for evaluating the teams' systems. These models were not shared with participants during the competition.

Subtask 1. We finetune SpeechBrain's VoxLingua107 (Valk and Alumäe, 2021) ECAPA-TDNN (Desplanques et al., 2020) system[5] on the adaptation split of the dataset. We replace the classification layers of the pretrained system with new randomized layers corresponding to the smaller number of output classes (8); and train these new layers with the rest of the model frozen for 5K steps, and then unfreeze the model and train for an additional 25K steps. We use AdamW with base learning rate of $1e-4$, and apply a linear ramp up from $1/3$ base LR over 3K steps followed by constant LR until unfreezing, and then repeat the linear ramp up and plateau. Finally, Starting at 20K steps we applying an exponential decay.

Subtask 2. A zero-shot baseline is built on the pre-trained Whisper-Large-v3 model (Radford et al., 2022). Dialect-wise inference is performed on the official NADI 2025 subtask 2 ASR release available on Hugging Face[6], which provides validation splits for eight country-level dialects; official evaluation is conducted on a private Codabench test set. During inference, audio inputs are transcribed

[5] https://huggingface.co/speechbrain/spkrec-ecapa-voxceleb

[6] https://huggingface.co/datasets/UBC-NLP/NADI2025_subtask2_ASR

Team Name	Affiliation	Subtask
Abjad AI (Ghannam et al., 2025)	Abjad AI, Jordan & Saudi Arabia	1,3
BYZÖ (Abdullah et al., 2025)	Saarland University, Germany	2
Elyadata (Elleuch et al., 2025)	Elyadata, Tunisia	1,2
Hamsa	Hamsa	2
Lahjati (ALBawwab and Qawasmeh, 2025)	Princess Sumaya University, Jordan	1
MarsadLab (Attia et al., 2025)	Hamad Bin Khalifa University, Qatar	1,2
Munsit (Salhab et al., 2025b)	Lebanese American University, Lebanon	2
Unicorn (Elrefai, 2025)	Ain shams University, Egypt	3

Table 2: List of teams that participated in NADI 2025 shared task. Teams with accepted papers are cited.

using Whisper's default decoding parameters with language explicitly set to Arabic.

	Accuracy ↑	C_{avg} ↓
ELYDATA-LIA (Elleuch et al., 2025)	**79.8**	**17.88**
BYZÖ-ADI (Abdullah et al., 2025)	76.4	22.65
MarsadLab (Attia et al., 2025)	61.6	30.68
Abjad AI	61.2	34.77
Baseline	61.1	34.22
Lahjati (ALBawwab and Qawasmeh, 2025)	50.8	48.99

Table 3: Performance of the systems on the test set for Subtask 1. Results are sorted by Accuracy, while the average cost (C_{avg}) score is also reported, with lower values indicating better performance. The best performance is highlighted in bold.

Subtask 3. In this subtask, we provide three baselines: (I) A *text only* baseline based on the publicly available CATT model (Alasmary et al., 2024), which we use without further fine-tuning (II) an *ASR based* baseline where we use the ArTST v3 checkpoint (Djanibekov et al., 2025), which is pretrained on dialectal and code-switched Arabic, and finetune it for Arabic ASR with diacritics using the provided training data, and (III) a *multi-modal* diacritc restoration model designed as follows:

The raw waveform and corresponding transcriptions are passed in parallel to a speech and text encoder, respectively. The speech encoder is derived from ArTST ASR (Djanibekov et al., 2025), and the text encoder is derived from ArTST TTS model (Toyin et al., 2023). We then align the resulting text and speech embeddings using multi-head attention with 8 heads, followed by a trainable prediction component comprising 2 bi-directional LSTM layers, a 30% dropout layer, and a final linear prediction head to predict the corresponding diacritics. Simple ad-hoc post-processing is applied to add the predicted diacritics to the input text to produce the fully diacritized text output. This approach is inspired by the multi-modal diacritization model described in Shatnawi et al. (2024).

4.3 Results

Tables 3, 4, and 5, present the preformernce of the submitted systems on the test set for subtask 1, subtask 2, and subtask 3 respectively.

Subtask1. The `ELYDATA-LIA` team (Elleuch et al., 2025) achieved the best performance in terms of both accuracy and average cost C_{avg} (79.8 / 17.88), followed closely by `BYZÖ-ADI` (Abdullah et al., 2025) (76.4 / 22.65). Both top teams addressed the limited size of the Adaptation set in novel ways: ELYDATA-LIA leveraged the much larger ADI-20 dataset (Elleuch et al., 2025), while BYZÖ-ADI employed kNN voice conversion (Baas et al., 2023) to augment the training data with synthetic samples. In third place, `MarsadLab` (Attia et al., 2025) improved upon the baseline system through additional data augmentation and the introduction of an attention mechanism prior to the classification layer. In fourth place, `Abjad AI` fine-tuned a Whisper Small encoder with further data augmentation. While the third- and fourth-place systems were close in terms of accuracy (61.6 vs. 61.2), the approach by MarsadLa achieved a notably better C_{avg}, reducing it by approximately 4 points. Finally, we note that one team (Lahjati (ALBawwab and Qawasmeh, 2025)) perform below the baseline. Overall, these results highlight the effectiveness and diversity of data augmentation strategies.

Subtask 2. The `Munist` team (Salhab et al., 2025b) obtain the lowest overall average WER/CER scores (35.68/12.10) among all participating systems, achieving the best performance across all dialects except Moroccan, where it ranked second in both WER and CER, and Mauritanian, where it ranked first in CER and second in WER. The `ELYADATA-LIA` team (Elleuch et al., 2025) ranked second with scores of 38.52/14.52. They achieved the best performance on the

	Average	JOR	EGY	MOR	ALG	YEM	MAU	UAE	PAL
Munsit (Salhab et al., 2025b)	**35.68/12.20**	**20.68/5.64**	**20.88/7.33**	41.71/14.04	**53.62/18.44**	**44.62/14.30**	59.03/**23.28**	**22.66/6.55**	**22.27/8.05**
ELYADATA-LIA (Elleuch et al., 2025)	38.53/14.52	28.03/9.36	26.83/11.43	**38.26/13.66**	53.73/20.43	46.63/16.66	**58.10**/24.53	29.35/9.91	27.36/10.20
BYZÖ-Whisper (Abdullah et al., 2025)	39.78/14.75	28.84/9.47	29.50/11.91	43.06/15.52	55.04/20.59	46.41/16.05	59.36/24.84	28.38/9.04	27.65/10.59
Hamsa	42.04/16.18	32.24/9.90	24.72/10.21	48.21/18.11	60.32/23.33	51.76/20.41	66.23/29.11	28.00/8.98	24.87/9.41
BYZÖ-CTC (Abdullah et al., 2025)	44.14/15.58	31.74/9.94	37.23/12.57	43.31/15.07	56.12/21.38	46.14/15.68	63.32/26.70	38.65/11.14	36.62/12.18
Baseline	93.89/72.79	46.09/19.28	100.06/81.37	100.38/80.42	101.03/79.58	101.09/80.58	100.59/82.89	101.15/80.27	100.76/77.92
MarsadLab (Attia et al., 2025)	104.89/84.69	44.97/19.19	113.97/97.65	104.07/87.58	116.59/94.26	113.54/94.56	111.59/92.84	116.79/97.00	117.60/94.42

Table 4: Performance of the systems on the test set for Subtask 2. Results are sorted by the overall average WER/CER score across all dialects, with lower values indicating better performance. The best performance is highlighted in bold.

	WER ↓	CER ↓
Abjad AI (Ghannam et al., 2025)	**55**	**13**
Unicorn (Elrefai, 2025)	64	15
Baseline-I (***ASR based***)	88	45
Baseline-II (***text-only***)	65	16
Baseline-III (***multi-modal***)	66	16

Table 5: Performance of the systems on the test set for Subtask 3. Results are sorted by the overall average WER/CER score across all dialects, with lower values indicating better performance. The best performance is highlighted in bold.

Moroccan dialect (WER/CER of 38.26/13.66) and obtain the lowest CER for the Mauritanian dialect. Their performance on the Algerian dialect was only marginally lower than that of the first-ranked team, suggesting that their system demonstrates strong capabilities for North African dialects in general. The `BYZÖ-Whisper` team (Abdullah et al., 2025) ranked third, with average WER/CER scores of 39.78/14.75. The `Hamsa` team follow in fourth place, scoring 42.04/16.18, while the `BYZÖ-CTC` team (Abdullah et al., 2025) ranked fifth with 44.14/15.58. Only one team, `MarsadLab` (Attia et al., 2025), perform below the baseline, with notably higher average WER/CER scores of 104.89/84.69. The winning team `Munist` (Salhab et al., 2025b) surpassed the baseline by **58.21** WER points (93.89 $\rightarrow$ 35.68; $\approx$ 62% reduction). Furthermore, the variation in WER scores among the teams that surpassed the baseline is relatively low ($\sigma \approx 3.25$), corresponding to about 8.1% of the mean WER for these systems.

Subtask 3. The `Abjad AI` (Ghannam et al., 2025) perform the best with the lowest WER of 55% and CER of 13%. The `Unicorn` team (Elrefai, 2025) follow closely behind with WER of 64% and CER of 15%. Both teams improve over the provided baselines, the best of which achieve WER and CER of 65% and 16%, respectively.

5 Overview of Submitted Systems

In this section, we present an overview of the submitted systems for each subtask and summarize the methodological approaches adopted by the participating teams.

5.1 Subtask 1

ELYDATA-LIA (Elleuch et al., 2025). Using Whisper Large-v3 encoder as their base model, they adopt a two stage finetuning procedure to first finetune on the forthcoming ADI-20 dataset (Elleuch et al., 2025), and then use the NADI ADI Adaptation set for a second finetuning. Features of this approach include freezing the first 16 layers of the encoder and using plenty of data augmentation methods including speed perturbation, added noise, and frequency and chunk dropping.

BYZÖ-ADI (Abdullah et al., 2025) The authors choose a straightforward finetuning approach using w2v-BERT-2.0 (Barrault et al., 2023) model finetuned on the NADI ADI split (69% accuracy). However, in order to improve the robustness of the model, they add a data augmentation approach by using a voice conversion model (Baas et al., 2023) to re-synthesizing the training utterances using voice samples from the 4 Arabic speakers from the LibriVox project, and training on the mixed natural and synthetic audio, leading to their final model.

MarsadLab (Attia et al., 2025) Adopts a starting point similar to the baseline with a VoxLingua107 ECAPA-TDNN system that was finetuned on the ADI task. They introduce a number of features in the process including feature reweighting of the hidden representation just prior to the classification layer through the use of a lightweight attention mechanism, discriminative learning rate of the classification head, progressive unfreezing, as well as data augmentation using SpecAugment and injected noise.

Abjad AI Like the ELYDATA-LIA approach, this team used Whisper model, Whisper Small, and finetuned the encoder for dialect ID. They use only the NADI Adaptation set for finetuning, using SpecAugment (time and frequency masking) for data augmentation, and unfreezing the model partway through training.

Lahjati (ALBawwab and Qawasmeh, 2025) Using both the VoxLingua107 ECAPA-TDNN system as well as WavLM encoder, this fusion approach concattenates the outputs from the two models (WavLM pooled to match the ECAPA 256 dimension), and passes this combined representation through a layer normalization layer and then a two layers feedforward network to perform classification. Similar to other approaches the underlying models start frozen, with unfreezing at 8000 steps, followed by a ramp up, plateau, and then cosine annealing learning rate schedule.

5.2 Subtask 2

Munsit (Salhab et al., 2025b) This system follows a two-stage training pipeline combining large-scale weakly supervised pretraining and continual supervised fine-tuning, inspired by Salhab et al., 2025a. In the first stage, a Conformer-large model (Gulati et al., 2020) (121M parameters) was pretrained on 15K hours of weakly labeled Arabic speech, covering MSA and various dialects, with automatic labeling and no manual verification. In the second stage, the model was fine-tuned using a high-quality dataset composed of 3,000 hours of rigorously filtered weakly labeled data, excluding news content, and the official Casablanca Challenge training set, expanded via data augmentation. Training used the CTC (Graves et al., 2006) objective with a SentencePiece (Kudo and Richardson, 2018) vocabulary of 128 tokens, AdamW optimizer, Noam learning rate schedule, and dropout of 0.1, in a distributed setup across 8 NVIDIA A100 GPUs with bfloat16 precision. This approach enabled robust performance across all dialects, achieving the lowest average WER and CER in the shared task.

ELYADATA & LIA (Elleuch et al., 2025) For the ASR subtask, this team fine-tuned the SeamlessM4T-v2 (Barrault et al., 2023) Large Egyptian model separately for each of the eight dialects in the Casablanca dataset, producing eight distinct models. Training was performed for 6 epochs on NVIDIA A100 GPUs with a label-smoothed NLL loss (smoothing 0.2), AdamW optimizer, and a learning rate schedule with 100 warm-up steps ramping from 1e-9 to 5e-5. A batch size of 2 was used for all runs. This per-dialect fine-tuning approach yielded second overall in the shared task.

BYZÖ (Abdullah et al., 2025) The team submitted two independent systems. The first, `BYZÖ-Whisper`, fine-tuned the Whisper-Large-v3 model (Radford et al., 2023) (1.54B parameters) for Arabic dialect ASR using only the NADI shared task data, without external datasets or data augmentation. Text labels were preprocessed by removing bracketed content and normalizing spacing. Training followed a two-stage process: (1) domain adaptation on combined data from all dialects for 9000 steps (learning rate 1e-5, batch size 32), and (2) dialect-specific adaptation for 2000 steps per dialect using CER as the metric. The second, `BYZÖ-CTC`, fine-tuned the w2v-BERT-2.0 model (Barrault et al., 2023) (580M parameters) using a mix of public Arabic ASR datasets, then further fine-tuned per dialect on the shared task data (learning rate 1e-5, batch size 16). A multi-dialectal 3-gram Kneser-Ney smoothed language model, trained on collected dialect-specific text data, was integrated to reduce WER. This encoder-only CTC-based system was noted for efficiency and competitive zero-shot performance compared to Whisper large.

MarsadLab (Attia et al., 2025) For the ASR subtask, this team adopted Whisper-Large model (Radford et al., 2023) in a zero-shot setting, without any fine-tuning, preprocessing, or post-processing. Leveraging Whisper's multilingual capabilities, the system directly transcribed Arabic speech from multiple dialects in the test set. While the ECAPA-TDNN architecture was central to their ADI submission, it was not applied to ASR.

Hamsa Submissions were received from the `Hamsa` team; however, a system description was not made available.

5.3 Subtask 3

Unicorn (Elrefai, 2025) This team addressed the diacritic restoration task by fine-tuning the GEMM3N[7] multimodal model on both audio and text inputs. They formed diacritic restoration as a structured generation task where the model receives an undiacritized sentence and its corresponding audio and generates a fully diacritized ver-

[7] `https://unsloth.ai/`

sion. They fine-tuned with LoRA adaptation to efficiently adapt the model with the provided data for the sub-task only. They applied *nlpaug* for speech augmentation to simulate more diverse audio inputs. They perform inference by prompting GEMM3N with the raw audio and the corresponding undiacritized text.

Abjad AI (Ghannam et al., 2025) This team presented CATT-Whisper, which is a multimodal approach that combines both textual and speech information. Their model represents the text modality using an encoder extracted from their pre-trained model named CATT (Alasmary et al., 2024). The speech component is handled by the encoder module of the OpenAI Whisper base model (Radford et al., 2022). Their approach uses two integration strategies. The former consists of fusing the speech tokens with the input at an early stage, where the 1500 frames of the audio segment are averaged on the basis of 10 consecutive frames, resulting in 150 speech tokens only. To ensure embedding compatibility, these averaged tokens are processed through a linear projection layer prior to merging them with the text tokens. Contextual encoding is guaranteed by the CATT encoder module. The latter strategy relies on cross-attention, where text and speech embeddings are fused. Then, finally, the cross-attention output is fed to the CATT classification head for token-level diacritic prediction. They randomly deactivate the speech input during training for robustness, which allows the model to perform well with or without speech.

6 Conclusion

The *sixth* NADI shared task extends the scope of the series beyond text-based processing to encompass speech and diacritization, introducing three new subtasks: spoken dialect identification, Arabic ASR, and diacritic restoration. By releasing high-quality resources and providing clear evaluation protocols, our goal is to foster progress in inclusive Arabic speech processing. This edition, we received 44 registrations, with *eight* teams submitting system outputs and *seven* system description papers accepted. Results across the three subtasks highlight substantial headroom for improvement: even strong pretrained models continue to face challenges with multidialectal variability, code-switching, and diacritic restoration. We hope that this edition not only advances the state of the art on each individual subtask but also inspires future research toward unified, dialect-aware speech technologies for Arabic.

Limitations & Ethical Considerations

Despite the contributions of this year's shared task, several limitations remain across the three subtasks:

Coverage of dialects: Not all Arabic dialects are represented in the test sets, which limits the generalizability of results across the full dialect continuum.

Country-level labeling: We acknowledge that the use of country-level labels may be problematic. The continuum of Arabic dialects is complex, and using country affiliation as a stand-in for well-defined linguistic boundaries is not without limitations. This choice was made to ensure a reasonable degree of diversity in dialect coverage, while avoiding assumptions about the generalizability of models trained on a subset of dialects to unseen but related varieties.

Code-switching: The datasets capture only a limited subset of code-switching phenomena, whereas real-world Arabic speech often involves more diverse language mixing.

Real-world conditions: Background noise, spontaneous disfluencies, and accented speech are underrepresented in the datasets, limiting ecological validity.

Evaluation metrics: Metrics such as WER and CER may be misleading in the ASR task, since a dialectal utterance can often have multiple valid references. As the data provides only one reference per utterance, evaluation scores may underestimate system performance by penalizing alternative but correct transcriptions.

Acknowledgments

Muhammad Abdul-Mageed acknowledges support from Canada Research Chairs (CRC), the Natural Sciences and Engineering Research Council of Canada (NSERC; RGPIN-2018-04267), the Social Sciences and Humanities Research Council of Canada (SSHRC; 435-2018-0576; 895-2020-1004; 895-2021-1008), Canadian Foundation for Innovation (CFI; 37771), Digital Research Alliance of Canada,[8] and UBC ARC-Sockeye.

[8] https://alliancecan.ca

References

Ahmed Amine Ben Abdallah, Ata Kabboudi, Amir Kanoun, and Salah Zaiem. 2023. Leveraging data collection and unsupervised learning for code-switched tunisian arabic automatic speech recognition. *Preprint*, arXiv:2309.11327.

Ahmed Abdelali, Hamdy Mubarak, Younes Samih, Sabit Hassan, and Kareem Darwish. 2021. Qadi: Arabic dialect identification in the wild. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 1–10, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Muhammad Abdul-Mageed, Hassan Alhuzali, and Mohamed Elaraby. 2018. You tweet what you speak: A city-level dataset of arabic dialects. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Muhammad Abdul-Mageed, AbdelRahim Elmadany, Chiyu Zhang, El Moatez Billah Nagoudi, Houda Bouamor, and Nizar Habash. 2023. Nadi 2023: The fourth nuanced arabic dialect identification shared task. *arXiv preprint arXiv:2310.16117*.

Muhammad Abdul-Mageed, Amr Keleg, AbdelRahim Elmadany, Chiyu Zhang, Injy Hamed, Walid Magdy, Houda Bouamor, and Nizar Habash. 2024. Nadi 2024: The fifth nuanced arabic dialect identification shared task. *arXiv preprint arXiv:2407.04910*.

Muhammad Abdul-Mageed, Chiyu Zhang, Houda Bouamor, and Nizar Habash. 2020. NADI 2020: The first nuanced arabic dialect identification shared task. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop, WANLP@COLING 2020, Barcelona, Spain (Online), December 12, 2020*, pages 97–110. Association for Computational Linguistics.

Muhammad Abdul-Mageed, Chiyu Zhang, AbdelRahim Elmadany, Houda Bouamor, and Nizar Habash. 2022. Nadi 2022: The third nuanced arabic dialect identification shared task. *arXiv preprint arXiv:2210.09582*.

Muhammad Abdul-Mageed, Chiyu Zhang, AbdelRahim Elmadany, and Lyle Ungar. 2020. Toward micro-dialect identification in diaglossic and code-switched environments. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5855–5876, Online. Association for Computational Linguistics.

Muhammad Abdul-Mageed, Chiyu Zhang, Abdel-Rahim A. Elmadany, Houda Bouamor, and Nizar Habash. 2021. NADI 2021: The second nuanced arabic dialect identification shared task. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop, WANLP 2021, Kyiv, Ukraine (Virtual), April 9, 2021*, pages 244–259. Association for Computational Linguistics.

Badr Abdullah, Yusser Al-Ghussein, Zena Al-Khalili, Ömer Özyilmaz, Matias Valdenegro-Toro, Simon Ostermann, and Dietrich Klakow. 2025. Saarland-groningen at nadi 2025 shared task: Effective dialectal arabic speech processing under data constraints. In *The Third Arabic Natural Language Processing Conference (ArabicNLP 2025)*, Suzhou. Association for Computational Linguistics.

Kathrein Abu Kwaik, Motaz Saad, Stergios Chatzikyriakidis, and Simon Dobnik. 2018. Shami: A corpus of levantine arabic dialects. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Maryam Khalifa Al Ali and Hanan Aldarmaki. 2024. Mixat: A data set of bilingual emirati-English speech. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*, pages 222–226, Torino, Italia. ELRA and ICCL.

Mohammad Al-Fetyani, Muhammad Al-Barham, Gheith A. Abandah, Adham Alsharkawi, and Maha Dawas. 2022. MASC: Massive arabic speech corpus. In *Proceedings of the 2022 IEEE Spoken Language Technology Workshop (SLT)*, page 1002206.

Rania Al-Sabbagh and Roxana Girju. 2012. Yadac: Yet another dialectal arabic corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2882–2889, Istanbul, Turkey. European Language Resources Association (ELRA).

Nora Al-Twairesh, Rawan N. Al-Matham, Nora Madi, Nada Almugren, Al-Hanouf Al-Aljmi, Shahad Alshalan, Raghad Alshalan, Nafla Alrumayyan, Shams Al-Manea, Sumayah Bawazeer, Nourah Al-Mutlaq, Nada Almanea, Waad Bin Huwaymil, Dalal Alqusair, Reem Alotaibi, Suha Al-Senaydi, and Abeer Alfutamani. 2018. Suar: Towards building a corpus for the saudi dialect. In *Fourth International Conference on Arabic Computational Linguistics, ACLING 2018*, volume 142 of *Procedia Computer Science*, pages 72–82, Dubai, United Arab Emirates. Elsevier.

Faris Alasmary, Orjuwan Zaafarani, and Ahmad Ghannam. 2024. Catt: Character-based arabic tashkeel transformer. *Preprint*, arXiv:2407.03236.

Sanad ALBawwab and Omar Qawasmeh. 2025. Lahjati at nadi 2025 a ecapa-wavlm fusion with multi-stage optimization. In *The Third Arabic Natural Language Processing Conference (ArabicNLP 2025)*, Suzhou. Association for Computational Linguistics.

Hanan Aldarmaki and Ahmad Ghannam. 2023. Diacritic recognition performance in arabic asr. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, volume 2023, pages 361–365.

Ahmed Ali, Peter Bell, James Glass, Yacine Messaoui, Hamdy Mubarak, Steve Renals, and Yifan Zhang. 2016. The mgb-2 challenge: Arabic multi-dialect

broadcast media recognition. In *2016 IEEE Spoken Language Technology Workshop (SLT)*, pages 279–284.

Ahmed Ali, Suwon Shon, Younes Samih, Hamdy Mubarak, Ahmed Abdelali, James Glass, Steve Renals, and Khalid Choukri. 2019. The mgb-5 challenge: Recognition and dialect identification of dialectal arabic speech. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1026–1033. IEEE.

Ahmed Ali, Stephan Vogel, and Steve Renals. 2017. Speech recognition challenge in the wild: Arabic mgb-3. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 316–322. IEEE.

Khalid Almeman, Mark Lee, and Ali Abdulrahman Almiman. 2013. Multi dialect arabic speech parallel corpora. In *2013 1st International Conference on Communications, Signal Processing, and their Applications (ICCSPA)*, pages 1–6.

Israa Alsarsour, Esraa Mohamed, Reem Suwaileh, and Tamer Elsayed. 2018. Dart: A large dataset of dialectal arabic tweets. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Maha J. Althobaiti. 2022. Creation of annotated country-level dialectal arabic resources: An unsupervised approach. *Natural Language Engineering*, 28(5):607–648.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC)*, pages 4218–4222.

Kais Attia, Md. Rafiul Biswas, Shimaa Ibrahim, Mabrouka Bessghaier, Firoj Alam, and Wajdi Zaghouani. 2025. Marsadlab at nadi: Arabic dialect identification and speech recognition using ecapa-tdnn and whisper. In *The Third Arabic Natural Language Processing Conference (ArabicNLP 2025)*, Suzhou. Association for Computational Linguistics.

Matthew Baas, Benjamin van Niekerk, and Herman Kamper. 2023. Voice conversion with just nearest neighbors. In *Interspeech 2023*, pages 2053–2057.

As-Said Muhámmad Badawi. 1973. *Mustawayat al-arabiyya al-muasira fi Misr*. Dar al-maarif.

Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2022. Hierarchical aggregation of dialectal data for arabic dialect identification. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC 2022)*, pages 4586–4596, Marseille, France. European Language Resources Association (ELRA).

Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady Elsahar, Justin Haaheim, and 1 others. 2023. Seamless: Multilingual expressive and streaming speech translation. *arXiv preprint arXiv:2312.05187*.

Houda Bouamor, Nizar Habash, and Kemal Oflazer. 2014. A multidialectal parallel corpus of arabic. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1240–1245, Reykjavik, Iceland. European Language Resources Association (ELRA).

Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, and Kemal Oflazer. 2018. The MADAR arabic dialect corpus and lexicon. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018*. European Language Resources Association (ELRA).

Kristen Brustad. 2000. *The Syntax of Spoken Arabic: A Comparative Study of Moroccan, Egyptian, Syrian, and Kuwaiti Dialects*. Georgetown University Press.

Shammur A Chowdhury, Ahmed Ali, Suwon Shon, and James R Glass. 2020. What does an end-to-end dialect identification model learn about non-dialectal information? In *INTERSPEECH*, pages 462–466.

Kareem Darwish, Nizar Habash, Mourad Abbas, Hend Al-Khalifa, Huseein T. Al-Natsheh, Houda Bouamor, Karim Bouzoubaa, Violetta Cavalli-Sforza, Samhaa R. El-Beltagy, Wassim El-Hajj, Mustafa Jarrar, and Hamdy Mubarak. 2021. A Panoramic survey of Natural Language Processing in the Arab Worlds. *Commun. ACM*, 64(4):72–81.

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. 2020. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. In *Interspeech 2020*, pages 3830–3834.

Mona Diab, Nizar Habash, Owen Rambow, Mohamed Altantawy, and Yassine Benajiba. 2010. Colaba: Arabic dialect annotation and processing. In *Proceedings of the LREC Workshop on Semitic Language Processing*, pages 66–74.

Amirbek Djanibekov, Hawau Olamide Toyin, Raghad Alshalan, Abdullah Alitr, and Hanan Aldarmaki. 2025. Dialectal coverage and generalization in arabic speech recognition. *Preprint*, arXiv:2411.05872.

Mahmoud El-Haj. 2020. Habibi – a multi dialect multi national arabic song lyrics corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1318–1326, Marseille, France. European Language Resources Association (ELRA).

Heba Elfardy, Mohamed Al-Badrashiny, and Mona Diab. 2014. Aida: Identifying code switching in informal

arabic text. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 94–101, Doha, Qatar. Association for Computational Linguistics.

Salman Elgamal, Ossama Obeid, Mhd Kabbani, Go Inoue, and Nizar Habash. 2024. Arabic diacritics in the wild: Exploiting opportunities for improved diacritization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14815–14829, Bangkok, Thailand. Association for Computational Linguistics.

Haroun Elleuch, Salima Mdhaffar, Yannick Estève, and Fethi Bougares. 2025. ADI-20: Arabic Dialect Identification dataset and models. In *Interspeech 2025*, pages 2775–2779.

Haroun Elleuch, Youssef Saidi, Salima Mdhaffar, Yannick Estève, and Fethi Bougares. 2025. Elyadata lia at nadi 2025: Asr and adi subtasks. In *The Third Arabic Natural Language Processing Conference (ArabicNLP 2025)*, Suzhou. Association for Computational Linguistics.

AbdelRahim Elmadany, Muhammad Abdul-Mageed, and 1 others. 2022. Arat5: Text-to-text transformers for arabic language generation. In *Proceedings of the 60th annual meeting of the association for computational linguistics (Volume 1: Long papers)*, pages 628–647.

Abdelrahim Elmadany, Muhammad Abdul-Mageed, and 1 others. 2023a. Octopus: A multitask model and toolkit for arabic natural language generation. In *Proceedings of ArabicNLP 2023*, pages 232–243.

AbdelRahim Elmadany, ElMoatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023b. Orca: A challenging benchmark for arabic language understanding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9559–9586, Toronto, Canada. Association for Computational Linguistics.

Mohamed Lotfy Elrefai. 2025. Unicorn at nadi 2025 subtask 3: Gemm3n-dr: Audio-text diacritic restoration via fine-tuned multimodal arabic llm. In *The Third Arabic Natural Language Processing Conference (ArabicNLP 2025)*, Suzhou. Association for Computational Linguistics.

Ali Fadel, Ibraheem Tuffaha, Bara' Al-Jawarneh, and Mahmoud Al-Ayyoub. 2019. Neural Arabic text diacritization: State of the art results and a novel approach for machine translation. In *Proceedings of the 6th Workshop on Asian Translation*, pages 215–225, Hong Kong, China. Association for Computational Linguistics.

Hassan Gadalla, Hanaa Kilany, Howaida Arram, Ashraf Yacoub, Alaa El-Habashi, Amr Shalaby, Krisjanis Karins, Everett Rowson, Robert MacIntyre, Paul Kingsbury, David Graff, and Cynthia McLemore. 1997. Callhome egyptian arabic transcripts ldc97t19. Web Download.

Ahmad Ghannam, Naif Alharthi, Faris Alasmary, Kholood Al Tabash, Shouq Sadah, and Lahouari Ghouti. 2025. Abjad ai at nadi 2025: Catt-whisper: Multimodal diacritic restoration using text and speech representations. In *The Third Arabic Natural Language Processing Conference (ArabicNLP 2025)*, Suzhou. Association for Computational Linguistics.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and 1 others. 2020. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*.

Karim El Haff, Mustafa Jarrar, Tymaa Hammouda, and Fadi Zaraket. 2022. Curras + Baladi: Towards a Levantine Corpus. In *Proceedings of the International Conference on Language Resources and Evaluation(LREC 2022)*, Marseille, France.

Nagham Hamad, Mohammed Khalilia, and Mustafa Jarrar. 2025. Konooz: Multi-domain Multi-dialect Corpus for Named Entity Recognition. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pages 0–0, Vienna, Austria. Association for Computational Linguistics.

Injy Hamed, Ngoc Thang Vu, and Slim Abdennadher. 2020. ArzEn: A speech corpus for code-switched Egyptian Arabic-English. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4237–4246, Marseille, France. European Language Resources Association.

S. Harrat, M. Abbas, K. Meftouh, and K. Smaili. 2013. Diacritics restoration for arabic dialect texts. In *Interspeech 2013*, pages 1429–1433.

Salima Harrat, Karima Meftouh, Mourad Abbas, and Kamel Smaïli. 2014. Building resources for algerian arabic dialects. In *INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association*, pages 2123–2127, Singapore. ISCA.

Richard S. Harrell. 1962. *A Short Reference Grammar of Moroccan Arabic: With Audio CD*. Georgetown Classics in Arabic Language and Linguistics. Georgetown University Press.

Clive Holes. 2004. *Modern Arabic: Structures, Functions, and Varieties*. Georgetown Classics in Arabic Language and Linguistics. Georgetown University Press.

Elsayed Issa, Mohammed AlShakhori, Reda AlBahrani, and Gus Hahn-Powell. 2021. Country-level arabic dialect identification using rnns with and without

linguistic features. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 276–281, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Mustafa Jarrar, Nizar Habash, Faeq Alrimawi, Diyam Akra, and Nasser Zalmout. 2016. Curras: An annotated corpus for the palestinian arabic dialect. *Language Resources and Evaluation*, pages 1–31.

Mustafa Jarrar, Fadi Zaraket, Tymaa Hammouda, Daanish Masood Alavi, and Martin Waahlisch. 2023. Lisan: Yemeni, Irqi, Libyan, and Sudanese Arabic Dialect Copora with Morphological Annotations. In *The 20th IEEE/ACS International Conference on Computer Systems and Applications (AICCSA)*, pages 1–7. IEEE.

Salam Khalifa, Nizar Habash, Dana Abdulrahim, and Sara Hassan. 2016. A large scale corpus of gulf arabic. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4282–4289, Portorož, Slovenia. European Language Resources Association (ELRA).

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*.

Ajinkya Kulkarni, Atharva Kulkarni, Sara Abedalmon'em Mohammad Shatnawi, and Hanan Aldarmaki. 2023. Clartts: An open-source classical arabic text-to-speech corpus. In *2023 INTERSPEECH*, pages 5511–5515.

Mingfei Lau, Qian Chen, Yeming Fang, Tingting Xu, Tongzhou Chen, and Pavel Golik. 2025. Data quality issues in multilingual speech datasets: The need for sociolinguistic awareness and proactive language planning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7466–7492, Vienna, Austria. Association for Computational Linguistics.

Yooyoung Lee, Craig Greenberg, Lisa Mason, and Elliot Singer. 2022. Nist 2022 language recognition evaluation plan.

Shervin Malmasi, Marcos Zampieri, Nikola Ljubešić, Preslav Nakov, Ahmed Ali, and Jörg Tiedemann. 2016. Discriminating between similar languages and Arabic dialect identification: A report on the third DSL shared task. In *Proceedings of the third workshop on NLP for similar languages, varieties and dialects (VarDial3)*, pages 1–14.

Karima Meftouh, Salima Harrat, Salma Jamoussi, Mourad Abbas, and Kamel Smaili. 2015. Machine translation experiments on padic: A parallel arabic dialect corpus. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*, pages 26–34, Shanghai, China.

Hamdy Mubarak and Kareem Darwish. 2014. Using twitter to collect a multi-dialectal corpus of arabic. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, pages 1–7, Doha, Qatar. Association for Computational Linguistics.

El Moatez Billah Nagoudi, Ahmed El-Shangiti, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2023. Dolphin: A challenging and diverse benchmark for arabic nlg. *arXiv preprint arXiv:2305.14989*.

Amal Nayouf, Mustafa Jarrar, Fadi zaraket, Tymaa Hammouda, and Mohamad-Bassam Kurdy. 2023. Nâbra: Syrian Arabic Dialects with Morphological Annotations. In *Proceedings of the 1st Arabic Natural Language Processing Conference (ArabicNLP), Part of the EMNLP 2023*, pages 12–23. ACL.

Ossama Obeid, Mohammad Salameh, Houda Bouamor, and Nizar Habash. 2019. Adida: Automatic dialect identification for arabic. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 6–11, Minneapolis, Minnesota. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *arXiv preprint*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Fatiha Sadat, Farzindar Kazemi, and Atefeh Farzindar. 2014. Automatic identification of arabic language varieties and dialects in social media. In *Proceedings of the Second Workshop on Natural Language Processing for Social Media (SocialNLP)*, pages 22–27, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Mohammad Salameh, Houda Bouamor, and Nizar Habash. 2018. Fine-grained arabic dialect identification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1332–1344, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Mahmoud Salhab, Marwan Elghitany, Shameed Sait, Syed Sibghat Ullah, Mohammad Abusheikh, and Hasan Abusheikh. 2025a. Advancing arabic speech recognition through large-scale weakly supervised learning. *arXiv preprint arXiv:2504.12254*.

Mahmoud Salhab, Shameed Sait, Mohammad Abusheikh, and Hasan Abusheikh. 2025b. Munsit at nadi 2025 shared task 2: Pushing the boundaries of multidialectal arabic asr with weakly supervised pretraining and continual supervised fine-tuning. In *The Third Arabic Natural Language Processing Conference (ArabicNLP 2025)*, Suzhou. Association for Computational Linguistics.

Sara Shatnawi, Sawsan Alqahtani, and Hanan Aldarmaki. 2024. Automatic restoration of diacritics for speech data sets. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4166–4176.

Suwon Shon, Ahmed Ali, Younes Samih, Hamdy Mubarak, and James Glass. 2020. Adi17: A fine-grained arabic dialect identification dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8244–8248. IEEE.

Peter Sullivan, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2023. On the robustness of arabic speech dialect identification. In *Interspeech 2023*, pages 5326–5330.

Bashar Talafha, Karima Kadaoui, Samar Mohamed Magdy, Mariem Habiboullah, Chafei Mohamed Chafei, Ahmed Oumar El-Shangiti, Hiba Zayed, Rahaf Alhamouri, Rwaa Assi, Aisha Alraeesi, and 1 others. 2024. Casablanca: Data and models for multidialectal arabic speech recognition. *arXiv preprint arXiv:2410.04527*.

Hawau Toyin, Amirbek Djanibekov, Ajinkya Kulkarni, and Hanan Aldarmaki. 2023. Artst: Arabic text and speech transformer. In *Proceedings of ArabicNLP 2023*, pages 41–51.

Hawau Toyin, Rufael Marew, Humaid Alblooshi, Samar M. Magdy, and Hanan Aldarmaki. 2025. ArVoice: A Multi-Speaker Dataset for Arabic Speech Synthesis. In *Interspeech 2025*, pages 4808–4812.

Jörgen Valk and Tanel Alumäe. 2021. Voxlingua107: a dataset for spoken language recognition. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 652–658. IEEE.

Abdul Waheed, Bashar Talafha, Peter Sullivan, Abdelrahim Elmadany, and Muhammad Abdul-Mageed. 2023. Voxarabica: A robust dialect-aware arabic speech recognition system. In *Proceedings of ArabicNLP 2023*, pages 441–449.

Wajdi Zaghouani and Anis Charfi. 2018. Arap-tweet: A large multi-dialect twitter corpus for gender, age and language variety identification. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Omar F. Zaidan and Chris Callison-Burch. 2011. The arabic online commentary dataset: An annotated dataset of informal arabic with high dialectal content. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 37–41, Portland, Oregon, USA. Association for Computational Linguistics.

Taha Zerrouki and Amar Balla. 2017. Tashkeela: Novel corpus of arabic vocalized texts, data for auto-diacritization systems. *Data in Brief*, 11.