WhiteHouse: Translation of the Casablanca Corpus for Multi-dialectal Arabic Speech Translation

Fethi Bougares[1,2], **Salima Mdhaffar**[2], **Yannick Estève**[2]
[1]ELYADATA, [2]Laboratoire Informatique d'Avignon, Avignon, France
Correspondence: fethi.bougares@elyadata.com

Abstract

Remarkable progress has been made recently in the speech processing of Arabic dialects. This is primarily due to the availability of large multilingual pre-trained models as well as the development of multiple well-annotated datasets that support training, fine-tuning, and evaluation of various speech models. However, most existing research on Arabic speech processing did not consider Automatic Speech Translation (AST) and focused mainly on Dialect Identification (DI) and Automatic Speech Recognition (ASR) tasks. To address this gap, we introduce WhiteHouse, the first multi-dialectal Arabic-English Speech Translation Corpus. WhiteHouse supplements the recently created Casablanca dataset with English translation for each utterance in the transcripts. This results in a three-way parallel speech-transcription-translation multi-dialectal Arabic dataset. WhiteHouse dataset is used to evaluate various State-of-The-Art (SoTA) speech translation models. Our experiments show that SoTA speech translation models performs poorly when evaluated on Arabic dialectal conditions. All the data used during training and testing are released for public use and further improvements at : https://huggingface.co/datasets/fbougares/WhiteHouse.

Keywords: Arabic dialects, speech translation corpus, speech translation benchmark

1. Introduction

The Speech-to-Text Translation (STT) task aims to convert a speech in one source language into text in another target language. Historically, the STT problem has been solved by cascading an ASR module, which generates the transcript in the source language, and an Machine Translation (MT) module, which translates the transcript into the target language. This pipeline solution suffers from error propagation, hight latency and and high training costs. To address these shortcomings, researchers have shifted the focus towards end-to-end (E2E) models that unify these components into a single trainable model that reduces latency and prevents error propagation (Duong et al.; Berard et al., 2016). This approach has gained increasing popularity and achieved great success (Sung et al., 2019; Salesky et al., 2019; Zhang et al., 2019).

Since end-to-end models are known to be severely data-hungry, the performance of the E2E STT model is highly correlated with the size and quality of the training corpus. Therefore, substantial effort has been put into the annotation of large speech datasets. However, such efforts were largely concentrated on a small subset of languages such as English (Di Gangi et al., 2019), Chinese (Zhang et al., 2021) and French (Kocabiyikoglu et al., 2018). Other previous work introduced multi-lingual speech translation datasets such as CoVoST (Wang et al., 2020a,b), mTEDx (Salesky et al., 2021), and FLEURS (Conneau et al., 2022). Although they made a substantial contribution to advance speech translation for many languages, the vast majority of low-resource languages and spoken dialects are left behind with very limited or not available training and evaluation data sets. The availability of large multilingual speech datasets, on the other hand, allows researchers to train and investigate the effectiveness of these all-in-one multilingual systems for low-resource scenarios. For instance, it has been shown (Ma et al., 2025) that such a system can benefit the performance of both high- and low-resourced languages.

It is also important to emphasize that the nature of E2E STT systems, which removes the need for transcription in the source language, is particularly convenient for spoken languages. In fact, spoken languages are characterized by the lack of a writing system and orthographic conventions. This makes the cascading approach challenging since it relies on the lexical form of the source language (output of the ASR system) as an intermediate and input to the MT module.

The lack of a writing system and orthographic conventions are exactly the characteristics of Arabic dialects. In reality, Arabic sets out a wide range of linguistic varieties called Arabic dialects. These dialects are the spoken informal versions of the Modern Standard Arabic. They are significantly different and nearly mutually unintelligible. This situation poses challenges in adapting technologies from the Modern Standard Arabic and from one variety to another. Although significant effort has been made to build orthographic conventions for multiple Arabic dialects (Habash and al., 2018; Zribi et al., 2014; Habash and al., 2015), these conventions are still not widely embraced because of their lack of details and the absence of automatic

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 5849–5855
11–16 May 2026.

processing tools, such as adapted grammar and spell checker.

Despite these challenges, there exist previous studies that have introduced multiple datasets for Arabic dialects and developed various dedicated speech models. The vast majority of these previous works have been allocated to few tasks, including Automatic Speech Recognition (Masmoudi et al., 2018; Hussein et al., 2021; Abdallah et al., 2024; Mdhaffar et al., 2024a,b; Djanibekov et al., 2025; Talafha and al., 2025) and Dialect Identification (Ali et al., 2017, 2020; Shon et al., 2020; Kulkarni and Aldarmaki, 2023; Elleuch et al., 2025).

Unlike the aforementioned studies, the primary objective of this work is to address the lack of resources that enable the development and evaluation of E2E speech translation systems for multiple Arabic dialects. To that end, we supplement the recently published Casablanca multi-dialectal ASR dataset with English translation for each utterance in the transcripts. This resulted in a three-way parallel multi-dialectal Arabic dataset named **Whitehouse** and used to evaluate multiple multilingual speech translation models under zero-shot and fine-tuned conditions. In summary, our contributions are as follows.

1. We introduce **Whitehouse**, the first speech translation dataset that covers eight different Arabic dialects;

2. We evaluate SoTA multilingual Speech translation models and report obtained results;

3. We distribute all the annotated data sets to foster research on dialect speech processing.

2. Related work

Early efforts to develop Arabic dialect speech translation systems began in 2006 with the DARPA Spoken Language Communication and Translation System for Tactical Use (TRANSTAC) programs (Sanders et al., 2008). The goal of the TRANSTAC program was to demonstrate capabilities to rapidly develop two-way speech-to-speech translation systems that enable speakers of different languages to communicate with each other in real-world tactical situations. TRANSAC program focused on English to/from multiple languages including Iraqi Arabic. Work has since continued and several studies and initiatives have been carried out in order to create new dialectal Arabic speech translation resources. The following is a list of datasets that we were able to identify and that might be used for Arabic dialect to English speech translation task.

2.1. Callhome Egyptian Arabic Speech Translation dataset

This data set was introduced in (Kumar et al., 2014) to support research in Egyptian-Arabic to English speech translation. This dataset supplements three existing ASR oriented LDC corpus (LDC97T19, LDC2002T39 and LDC2002T38) with four reference translations for each utterance in the transcripts. The speech part of the corpus consists of telephone conversations between native speakers of Egyptian Colloquial Arabic. The translations were obtained using crowd-sourcing techniques. In total, 838 translators participated in this process, producing 143,568 translations in English.

2.2. ArzEnST

ArzEnST (Hamed et al., 2022) is a code-switched Egyptian Arabic - English Speech Translation Corpus. This corpus was introduced as an extension of the ArzEn (Hamed et al., 2020) speech corpus, which was collected through informal interviews with bilingual speakers. These interviews were initially transcribed by Egyptian Arabic-English bilingual speakers. The transcriptions are later translated into monolingual English and monolingual Egyptian Arabic sentences by human translators. This result on a three-way Egyptian Arabic - English speech translation corpus of 12 hours of speech, containing 6,216 sentences. This data set was used to build a cascaded speech translation system, where an ASR system is used to transcribe the speech, followed by MT system that translates the transcripts.

2.3. UFAL Speech Corpus of North Levantine Arabic

The corpus contains recordings by the native speakers of the North Levantine Arabic (apc) acquired during 2020, 2021, and 2023 in Prague, Paris, Kabardia, and St. Petersburg (Zemánek et al., 2023). Recordings contain both monologues and dialogues on the topics of everyday life (health, education, family life, sports, culture) as well as information on both host countries (living abroad) and country of origin (Syria traditions, education system, etc.). Audio recordings are transcribed and translated into English by students of Arabic at Charles University. An additional quality check is performed by the native speakers of the dialect. This data set was used to run Dialectal Speech Translation Shared Task organized for 2024 (Ahmad and al., 2024) and 2025 (Abdulmumin and al., 2025) IWSLT editions.

2.4. Tunisian Arabic conversational telephone speech

This data set was introduced during the International Conference on Spoken Language Translation (IWSLT). It contains 210 hours of transcribed Tunisian Arabic conversational telephone speech developed by the Linguistic Data Consortium[1] (LDC). A subset of 175 hours of that speech is translated into English. Speech data are conversational telephone recordings FLAC-compressed files in 16-bit 8 kHz PCM format. This corpus was used to run the Dialectal Speech Translation Shared Task organized for 2022 (Anastasopoulos and al., 2022), 2023 (Agarwal and al., 2023) and 2025 (Abdulmumin and al., 2025) IWSLT editions.

2.5. TEDxTN

TEDxTN (Bougares et al., 2025) is a three-way code-switched Tunisian Arabic-English Speech Translation Corpus. This data set is a collection of 108 TEDx talks that represents 25 hours of speech with code-switching that cover speakers with various accents from over 11 different regions of Tunisia. TEDxTN talks are first collected, segmented, and transcribed using a predefined annotation guidelines. Manual transcription is performed following a two-stage process. The first stage takes as input the audio files and produces a segmented output with an initial transcription systematically reviewed during a second validation stage to ensure compliance with the guidelines and correct inattention errors. Each utterance in the transcripts was also manually translated into English with possible access to the corresponding audio recording in case of need.

3. WhiteHouse

In this section, we present the WhiteHouse Arabic dialects-English Speech translation Corpus. This supplements the existing Casablanca (Talafha et al., 2024) corpus with manual translation for each utterance in the Casablanca transcripts.

3.1. Data source

Casablanca was recently introduced in (Talafha et al., 2024), as the largest supervised dataset for Arabic dialects. Casablanca data set was built to advance Arabic speech processing, especially ASR, gender identification, and dialect identification. It includes a diverse representation of eight Arabic dialects that contain some dialects that have not been featured in any previous NLP research. The data set was created by 15 native speakers who collected TV series episodes that represent the dialects of their countries. The collected episodes are segmented into shorter utterances, and annotated with orthographic transcription and gender information. Overall, the annotation process led to the creation of a 48-hours data set covering eight Arabic dialects, namely Algerian, Egyptian, Emirati, Jordanian, Mauritanian (Hassaniya), Moroccan, Palestinian, and Yemeni. Unfortunately, only the Casablanca development and test subsets were made vailable[2], with the aim of supporting further research and innovation in speech processing and linguistic research targeting Arabic dialects. These subsets were also used to run the NADI 2025 shared task (Talafha and al., 2025) for Automatic Speech Recognition and Spoken Dialect Identification.

Dialect	**Duration** Valid / Test	**#Segments** Valid / Test
Algeria	0:59:19 / 0:55:03	834 / 832
Egypt	0:58:50 / 0:59:12	835 / 824
Jordan	1:00:01 / 0:58:54	848 / 848
Mauritania	0:58:43 / 0:56:15	943 / 948
Morocco	1:00:04 / 1:00:23	1,045 / 1,045
Palestine	1:00:00 / 0:58:22	667 / 667
UAE	1:00:08 / 0:55:59	813 / 813
Yemen	1:00:04 / 1:02:30	803 / 803

Table 1: Duration and number of segments of the distributed Casablanca valid and test sets.

Table 1 provides the duration details and number of segments in the validation and test sets for each dialect. As shown in the table, the validation and test sets are around 1 hour per country.

3.2. Translation guidelines

Translations for the multi-dialectal Casablanca corpus were performed following the LDC Arabic-to-English Translation Guidelines (LDC, 2013). In order to achieve a high-quality dataset, we design a two-stage translation process: The first stage takes as input an audio file and its transcription in the source dialect and produces an initial translation that may contain transcription errors or may also not be fully compliant with the translation guideline. Translators are asked to use audio recordings only if the transcription in the source dialect is unclear. The output of the first stage is systematically reviewed during a second validation stage, in which non-compliance with the guidelines and inattention errors are corrected.

[1] https://catalog.ldc.upenn.edu/LDC2025S05

[2] https://huggingface.co/datasets/UBC-NLP/Casablanca

3.3. WhiteHouse corpus statistics

Based on the Casablanca dialectal transcriptions described in Table 1, we were able to create a multi-dialectal Arabic speech translation data set. Table 2 shows the number of words in the transcription and the English translation of each dialectal dataset (Valid and Test). All counts are reported after removing special tags such as "[music]", "[crying]" and "[noise]".

Dialect	Valid AR/EN	Test AR/EN
Algeria	8,548/11,286	8,032/10,559
Egypt	10,046/13,019	10,116/12,228
Jordan	9,065/12,333	8,804/12,188
Mauritania	10,149/12,531	10,116/12,228
Morocco	11,820/15,218	11,959/15,796
Palestine	8,958/11,809	8,833/11,729
UAE	9,336/12,579	8,766/12,163
Yemen	9,175/12,955	9,359/13,296

Table 2: Distribution of data in WhiteHouse. For each set, counts represent the number of words per dialect in the original dialectal transcription (AR) and the corresponding English translation (EN).

As reported in Table 2, for all dialects we observed a text expansion that is common when translating Arabic into English due to the structural differences between these languages. In fact, Arabic is a highly concise, root-based language, while English often requires auxiliary verbs, articles, and prepositions to convey the same meaning, causing the volume of text to increase.

4. Speech translation models

We perform a number of experiments on the Valid and Test splits presented earlier in Table 1. First, we evaluate various Speech-to-text translation models under a zero-shot condition. We also evaluated under few-shot training condition by fine-tuning using a subset of the validation data of each country. We have opted for the latter solution because of the non-availability of speech translation data sets that cover the considered dialects.

4.1. Pre-trained models

We evaluated the following Speech-to-text translation models in order to assess their adaptability and performance across the eight Arabic dialects:
Whisper: OpenAI Whisper is a family of robust Transformer-based encoder-decoder model trained for several tasks including speech transcription and translation into English (Radford et al., 2023). Its modular design and flexibility make it a great choice for low-resource scenarios with limited and specialized datasets. In this work, we compare the performance of three Whisper model versions (small, medium, and large-v3) in zero-shot and few-shot scenarios.
SeamlessM4T: SeamlessM4T is a pre-trained multilingual model that integrates speech and text translation into a unified framework (Communication and al., 2023). The model is able to perform end-to-end translations in both spoken and written languages without requiring intermediate transcriptions. SeamlessM4T is training using a very large linguistically and acoustically diverse dataset that includes both speech and text modalities. This enables the model to generalize effectively to low-resource languages and dialects. In this work, we used it in zero-shot and few-shot scenarios.
Qwen2-Audio: Qwen2-Audio is a pre-trained large audio-language model capable of understanding information from different modalities (Chu et al., 2024). This model is a combination of the Whisperlarge-v3 model audio encoder and the Qwen-7B large language model (Bai and al., 2023) as its foundational component. Overall, Qwen2-Audio is an 8.2B parameter audio model that has been (1) pre-trained using language prompts, (2) fine-tuned in a supervised fashion using set of high-quality dataset, and (3) optimized to follow human preferences. In this work, we used Qwen2-Audio-7B-Instruct under zero-shot scenario. We decided to keep Qwen2-Audio for future research because of the computational resource needed to fine-tune this large model.

4.2. Experimental Setup

For all experiments, we utilize speechbrain[3] (Ravanelli et al., 2024) and transformers libraries to perform zero-shot decoding and few-shot training. We resample all audio segments to a 16kHz rate. Zero-shot whisper experiments are performed using a single node with Quadro RTX 6000-24GB. For few-shot fine-tuning, we used a single-node Tesla A100-80GB GPU. All experimented models are evaluated with BLEU and chrF++ scores and obtained using the Sacrebleu toolkit (Post, 2018). Additionally, scores are calculated using case-sensitive after removing punctuation from the reference and system outputs. As regards the few-shot experiments, we split the validation set of each country to keep 15 minutes for validation and use the remaining samples (around 45 minutes per dialect) as few-shot samples during fine-tuning experiments. All fine-tuning experiments are performed for 5 training epochs and with the defaults parameters of each used SoTA model.

[3] https://github.com/speechbrain

Dialect	Whisper Small	Whisper Medium	Whisper Large-v3	seamless-m4t-v2-large	Qwen2-7B-Instruct
Algeria	2.5 / 18.2	5.6 / 23.3	**7.6 / 26.3**	6.2/ 23.7	4.0 / 23.0
Egypt	5.2 / 23.9	14.2 / 35.4	**16.5 / 38.2**	11.4 / 32.1	11.8 / 36.1
Jordan	9.8 / 30.6	19.0 / 40.1	19.5 / 41.6	13.9 / 35.8	**21.2 / 43.8**
Mauritania	1.5 / 14.7	3.4 / 18.9	**3.9 / 19.8**	3.0 / 19.7	2.1 / 18.2
Morocco	1.5 / 14.8	3.2 / 20.1	4.0 / 21.4	**5.2 / 22.3**	2.2 / 19.1
Palestine	5.8 / 27.4	12.2 / 35.8	**13.2 / 37.0**	8.7 / 28.7	11.7 / 37.0
UAE	4.6 / 21.4	11.3 / 30.3	12.3 / 32.7	6.9 / 25.4	**14.3 / 36.0**
Yemen	4.1 / 20.8	9.5 / 28.6	**9.5 / 30.0**	7.7 / 26.0	6.5 / 29.5

Table 3: **Zero-Shot** Speech Translation performances on the Whitehouse test set. Results are reported in case-sensitive BLEU (↑) and chrF++ (↑) scores separated by /.

Dialect	Whisper-small	Whisper-medium	Whisper-large-v3	seamless-m4t-v2-large
Algeria	7.1 / 27.3	12.6 / 34.1	**13.2 / 35.8**	9.8 / 33.3
Egypt	5.2 / 27.5	**11.8 / 38.5**	7.5 / 33.6	**12.1 / 36.6**
Jordan	15.4 / 38.5	24.1 / 48.0	**25.5 / 47.2**	21.7 / 44.6
Mauritania	3.8 / 23.5	7.6 / 28.7	**7.7 / 29.3**	6.5 / 28.2
Morocco	3.1 / 21.8	5.2 / 26.4	**8.6 / 28.6**	8.4 / 29.1
Palestine	11.9 / 38.0	**21.0 / 47.0**	19.6 / 46.1	17.2 / 41.5
UAE	10.4 / 31.1	**16.9 / 40.1**	15.4 / 38.1	14.3 / 36.3
Yemen	4.2 / 25.2	**11.0 / 36.1**	9.8 / 34.8	10.9 / 33.0

Table 4: **Few-Shot** Speech Translation performances on the Whitehouse test set. Results are reported in case-sensitive BLEU (↑) and chrF++ (↑) scores separated by /.

5. Results

Table 3 provides a comparison between different SoTA speech translation models under the zero-shot condition. As we can see, all models struggle to achieve good results, showing their inability to effectively translate from Arabic dialectal input speech. That being said, we also noted that all the evaluated SoTA models scores are particularly low for Algerian, Moroccan, Mauritanian, and Yemeni dialects. As for Algerian and Moroccan dialects, we can explain this by the high usage of code-switching in Algerian and Moroccan as reported in (Talafha et al., 2024). Results on Yemeni, and Mauritanian dialects, on the other hands, may be related to the fact that they are nearly zero resourced dialects.

Compared to the above dialects, the SoTA models performed significantly better in the Egyptian, Jordanian, Palestinian, and UAE dialects. We believe that this is due to their closeness to Modern Standard Arabic and the availability of previously developed resources for these dialects.

Across models, results show that Whisper large-v3 performs better when compared to medium and small versions. Compared to seamless-m4t-v2-large and Qwen-7B-Instruct, whisper large-v3 generally scores also better except for UAE and Jordan where it is outperformed by Qwen2-7B-Instruct (21.2 compared to 19.5 and 14.3 compared to 12.3 for Jordan and UAE respectively).

As shown in 4, almost all the models benefit from few-shot learning, although fine-tuning was performed using only about 45 minutes of training data for each dialect. An exception to this trend is observed for Whisper-large-v3 model with Egyptian dialect. Egyptian zero-shot model achieves 16.5 Bleu points, while the same model performs worse after fine-tuning and obtains 7.5 Bleu points. Another interesting observation, is the scores obtained with whisper-medium models with and without fine-tuning. In fact, whisper medium was systematically worse than Whisper large-v3 under zero-shot conditions. Under few-shot learning conditions, whisper medium performs better than the large-v3 model with half of the considered dialects.

6. conclusion

This paper introduces WhiteHouse, a new resource for Arabic dialects speech translation task. WhiteHouse supplements Casablanca dataset with English translation for eight Arabic dialects. We evaluated several pre-trained SoTA encoder-decoder and audio-language models. We experimented under zero-shot and few-shot learning conditions and found that these models are unable to process Arabic dialectal speech effectively. All of our data are released for public use to enable the reproducibility of our results. We hope that this data set will foster research on spoken Arabic dialects, particularly Arabic dialect speech translation.

7. Limitations

Whitehouse is an augmentation of the Casablanca data set created for speech transcription of multiple Arabic dialects. Therefore, it presents inherently the same limitations as Casablanca. In addition, it must be mentioned that Whitehouse augmentation is limited to the only subsets made available by the creators of the Casablanca corpus. However, this limitation could be overcome when the training set of Casablanca becomes available.

8. Bibliographical References

Ahmed Amine Ben Abdallah, Ata Kabboudi, Amir Kanoun, and Salah Zaiem. 2024. Leveraging data collection and unsupervised learning for code-switched tunisian arabic automatic speech recognition. In *ICASSP 2024*.

Idris Abdulmumin and al. 2025. Findings of the IWSLT 2025 evaluation campaign. In *IWSLT 2025*, pages 412–481, Vienna, Austria. ACL.

Milind Agarwal and al. 2023. FINDINGS OF THE IWSLT 2023 EVALUATION CAMPAIGN. In *IWSLT 2023*, pages 1–61, Toronto. ACL.

Ibrahim Said Ahmad and al. 2024. FINDINGS OF THE IWSLT 2024 EVALUATION CAMPAIGN. In *IWSLT 2024*, pages 1–11, Bangkok.

Ahmed Ali, Suwon Shon, Younes Samih, Hamdy Mubarak, Ahmed Abdelali, James Glass, Steve Renals, and Khalid Choukri. 2020. The mgb-5 challenge: Recognition and dialect identification of dialectal arabic speech. In *ASRU 2019*.

Ahmed Ali, Stephan Vogel, and Steve Renals. 2017. Speech recognition challenge in the wild: Arabic mgb-3.

Antonios Anastasopoulos and al. 2022. Findings of the IWSLT 2022 evaluation campaign. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 98–157, Dublin, Ireland. ACL.

Jinze Bai and al. 2023. Qwen technical report.

Alexandre Berard, Olivier Pietquin, Christophe Servan, and Laurent Besacier. 2016. Listen and translate: A proof of concept for end-to-end speech-to-text translation.

Fethi Bougares, Salima Mdhaffar, Haroun Elleuch, and Yannick Estève. 2025. TEDxTN: TEDx speech translation corpus for code-switched tunisian arabic - english. In *WANLP 2025*.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. 2024. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.

Seamless Communication and al. 2023. Seamlessm4t: Massively multilingual multimodal machine translation.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2022. Fleurs: Few-shot learning evaluation of universal representations of speech.

Mattia A. Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019. MuST-C: a Multilingual Speech Translation Corpus. In *NAACL*.

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

Long Duong, Antonios Anastasopoulos, David Chiang, Steven Bird, and Trevor Cohn. An attentional model for speech translation without transcription. In *NAACL 2016*, San Diego.

Haroun Elleuch, Salima Mdhaffar, Yannick Estève, and Fethi Bougares. 2025. Adi-20: Arabic dialect identification dataset and models. In *Interspeech*, The Netherlands.

Nizar Habash and al. 2015. Palestinian arabic conventional orthography guidelines - technical report. In *Birzeit University and New York Univesity Abu Dhabi*.

Nizar Habash and al. 2018. Unified guidelines and resources for Arabic dialect orthography. In *LREC 2018*, Miyazaki, Japan.

Injy Hamed, Nizar Habash, Slim Abdennadher, and Ngoc Thang Vu. 2022. Arzen-st: A three-way speech translation corpus for code-switched egyptian arabic - english.

Injy Hamed, Ngoc Thang Vu, and Slim Abdennadher. 2020. ArzEn: A speech corpus for code-switched Egyptian Arabic-English. In *LREC 2020*, pages 4237–4246, Marseille, France. ELRA.

Amir Hussein, Shinji Watanabe, and Ahmed Ali. 2021. Arabic speech recognition by end-to-end, modular systems and human.

Ali Can Kocabiyikoglu, Laurent Besacier, and Olivier Kraif. 2018. Augmenting librispeech with french translations: A multimodal corpus for direct speech translation evaluation.

Ajinkya Kulkarni and Hanan Aldarmaki. 2023. Yet another model for arabic dialect identification.

Gaurav Kumar, Yuan Cao, Ryan Cotterell, Chris Callison-Burch, Daniel Povey, and Sanjeev Khudanpur. 2014. Translations of the callhome Egyptian Arabic corpus for conversational speech translation. In *SLT 2014*, pages 244–248, Lake Tahoe, California.

Linguistic Data Consortium LDC. 2013. Bolt program: Arabic to english translation guidelines.

Rao Ma, Mengjie Qian, Yassir Fathullah, Siyuan Tang, Mark Gales, and Kate Knill. 2025. Cross-lingual transfer learning for speech translation.

Abir Masmoudi, Fethi Bougares, Mariem Ellouze, Yannick Estève, and Lamia Belguith. 2018. Automatic speech recognition system for tunisian dialect. *Lang. Resour. Eval.*, page 249–267.

Salima Mdhaffar, Fethi Bougares, Renato De Mori, Salah Zaiem, Mirco Ravanelli, and Yannick Estève. 2024a. Taric-slu: A tunisian benchmark dataset for spoken language understanding. In *LREC-COLING 2024*.

Salima Mdhaffar, Haroun Elleuch, Fethi Bougares, and Yannick Estève. 2024b. Performance analysis of speech encoders for low-resource slu and asr in tunisian dialect. In *WANLP*, pages 130–139.

Matt Post. 2018. A call for clarity in reporting bleu scores.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Mirco Ravanelli, Titouan Parcollet, Adel Moumen, Sylvain De Langen, Cem Subakan, Peter Plantinga, Yingzhi Wang, Pooneh Mousavi, Luca Della Libera, Artem Ploujnikov, et al. 2024. Open-source conversational ai with speechbrain 1.0. *Journal of Machine Learning Research*, 25(333):1–11.

Elizabeth Salesky, Matthias Sperber, and Alan W Black. 2019. Exploring phoneme-level speech representations for end-to-end speech translation.

Elizabeth Salesky, Matthew Wiesner, Jacob Bremerman, Roldano Cattoni, Matteo Negri, Marco Turchi, Douglas W. Oard, and Matt Post. 2021. The multilingual tedx corpus for speech recognition and translation.

Gregory Sanders, Sébastien Bronsart, Sherri Condon, and Craig Schlenoff. 2008. Odds of successful transfer of low-level concepts: a key metric for bidirectional speech-to-speech machine translation in DARPA's TRANSTAC program. In *LREC 2008*, Marrakech, Morocco.

Suwon Shon, Ahmed M. Ali, Younes Samih, Hamdy Mubarak, and James R. Glass. 2020. Adi17: A fine-grained arabic dialect identification dataset. *ICASSP 2020*, pages 8244–8248.

Tzu-Wei Sung, Jun-You Liu, Hung yi Lee, and Lin-Shan Lee. 2019. Towards end-to-end speech-to-text translation with two-pass decoding. *ICASSP 2019*, pages 7175–7179.

Bashar Talafha and al. 2025. Nadi 2025: The first multidialectal arabic speech processing shared task.

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

Changhan Wang, Juan Pino, Anne Wu, and Jiatao Gu. 2020a. CoVoST: A diverse multilingual speech-to-text translation corpus. In *LREC*, pages 4197–4203, Marseille, France.

Changhan Wang, Anne Wu, and Juan Pino. 2020b. Covost 2 and massively multilingual speech-to-text translation.

Petr Zemánek, Adam Pospíšil, Hashem Sellat, Mateusz Krubiński, and Pavel Pecina. 2023. UFAL speech corpus of north levantine arabic 1.0 - part 1. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL).

Biao Zhang, Ivan Titov, and Rico Sennrich. 2019. Improving deep transformer with depth-scaled initialization and merged attention.

Ruiqing Zhang, Xiyang Wang, Chuanqiang Zhang, Zhongjun He, Hua Wu, Zhi Li, Haifeng Wang, Ying Chen, and Qinfei Li. 2021. Bstc: A large-scale chinese-english speech translation dataset.

Inès Zribi, Rahma Boujelbane, Abir Masmoudi, Mariem Ellouze, Lamia Belguith, and Nizar Habash. 2014. A conventional orthography for Tunisian Arabic. In *LREC 2014*, pages 2355–2361, Reykjavik, Iceland.