TUNIFRA: A Tunisian Arabic Speech Corpus with Orthographic Transcriptions and French Translations

Alex Choux[1,2], Marko Avila[2], Fethi Bougares[3,4], Hugo Riguidel[1,2], Josep Crego[2], Antoine Laurent[1]

[1]LIUM, [2]SYSTRAN by ChapsVision, [3]Elyadata, [4]Laboratoire Informatique d'Avignon

{achoux,mavila,jcrego}@chapsvision.com fethi.bougares@elyadata.com antoine.laurent@univ-lemans.fr

Abstract

We introduce TUNIFRA, a novel and comprehensive corpus developed to advance research in Automatic Speech Recognition (ASR) and Speech-to-Text Translation (STT) for Tunisian Arabic, a notably low-resourced language variety. The TUNIFRA corpus comprises 15 hours of native Tunisian Arabic speech, carefully transcribed and manually translated into French. While the development of ASR and STT systems for major languages is supported by extensive datasets, low-resource languages such as Tunisian Arabic face significant challenges due to limited training data, particularly for speech technologies. TUNIFRA addresses this gap by offering a valuable resource tailored for both ASR and STT tasks in the Tunisian dialect. We describe our methodology for data collection, transcription, and annotation, and present initial baseline results for both Tunisian Arabic speech recognition and Tunisian Arabic–French speech translation.

1 Introduction

In recent years, AI has become increasingly integrated into daily life, largely due to the rise of powerful foundation models that support a wide array of downstream applications, like Radford et al. (2023) for ASR and Communication et al. (2023) for STT. Nevertheless, a significant portion of the population remains unable to benefit from these technological advances, as there are few models specifically adapted to their languages. This limitation is especially pronounced for under-resourced languages and dialects, despite increasing research efforts aimed at overcoming these barriers (Xu et al., 2024; Bhogale et al., 2024).

When it comes to Arabic dialects, only a handful are represented in widely used corpora such as Common Voice (Ardila et al., 2020), MGB (Ali et al., 2016, 2017, 2019), and FLEURS (Conneau et al., 2023). This has led to an imbalanced representation in the coverage and representation of the diverse range of Arabic dialects.

As a result, many Arabic dialects remain under-resourced, which reduces the effectiveness of Arabic language models due to significant differences in pronunciation and orthographic rules. Furthermore, Talafha et al. (2023) shows that even within the limited set of dialects represented in available corpora, data sparsity is a persistent issue, with the Egyptian dialect (EGY) dominating over other varieties. The Tunisian dialect, in particular, is among the many under-resourced dialects, a challenge that is common to most Arabic dialects, as also highlighted by Talafha et al. (2023). Table 1 summarizes previous efforts to collect and annotate datasets for the Tunisian dialect (Abdallah et al., 2023; Mdhaffar et al., 2024; Naouara et al., 2025).

Corpus	**Hours**	**Languages**
TunSwitchCS	8.15	Tunisian with CS
TunSwitchTO	2.29	Tunisian
TARIC	8.0	Tunisian
Linto	81.34	Tunisian

Table 1: Available Tunisian Arabic speech corpora. CS refers to code-switching.

We present TUNIFRA, the first publicly available three-way corpus specifically designed for Tunisian Arabic to French speech translation. In this work, we offer a comprehensive account of the data collection methodology and detail the annotation process, encompassing both transcription and translation steps, to ensure high-quality and reliable data. Furthermore, we report baseline experimental results for both ASR and STT tasks, demonstrating the utility and impact of the TUNIFRA corpus for advancing research in under-resourced language technologies.

Proceedings of The Third Arabic Natural Language Processing Conference, pages 64–68

2 Corpus

2.1 Data Collection

All recordings were sourced from Tunisian YouTube podcasts. The dataset consists of 19 raw audio recordings, each ranging from 20 to 80 minutes in duration, resulting in a total of around 15 hours of Tunisian speech. All speakers involved are native Tunisian speakers; however, some frequently code-switch, primarily between Tunisian and French, with occasional use of English.

The 19 recordings encompass a broad spectrum of topics, such as ecology, the education system, and economy. The formats also vary: some recordings are structured as interviews, while others are debates featuring between four and six participants. This results in a corpus that is diverse in both subject matter and speaker composition, providing substantial coverage of different speaking situations. We anticipate that this variety will help enhance the robustness of speech processing systems.

2.2 Data Annotation

The annotation of the raw audio recordings was performed using Transcriber, a specialized audio annotation tool. Human annotators, all native Tunisian linguists with degrees in French linguistics, ensured accurate alignment between the audio and its corresponding transcriptions. Two linguists were responsible for the ASR (automatic speech recognition) annotations, while four others handled the Tunisian-to-French translation annotations. It should be noted that neither inter-annotator nor intra-annotator agreement was assessed during this process. All recordings were fully annotated for both ASR and speech-to-text translation tasks. For the annotation of the raw transcription files, we adhere to the following specific rules:

- Foreign words are transcribed using the Roman script.
- When foreign words have been adapted to Tunisian dialect pronunciation, they are written in Arabic script.
- Arabic clitics/affixes are written in Arabic script and attached directly to foreign words.
- A predefined, fixed spelling is used for frequently occurring terms such as days of the week, numbers, quantities, percentages, and similar expressions.

2.3 Data Analysis

We present the results of our analysis in Table 2. Our analysis is conducted at a global level, without providing detailed statistics for each individual file.

Category	Value
Speech duration (hours)	15
# Segments	9,189
Avg segment Duration (seconds)	5.90
# Different speakers	41
Gender distribution (M/F/?)	29/8/4
# Src w. Tunisian	130,815
# Src w. foreign	16,889
# Seg. full Tunisian	5,353
# Seg. full foreign	132
# Seg. mixed	3,704
Avg transcription length (words)	16.07
# Src Words (Transcription)	147,704
# Src Vocab size	22,386
# Tgt Words (Translation)	190,640
# Tgt Vocab size	11,977
# Overlap. Speech (hours)	6
# Overlap. Speech segments	2,710

Table 2: Statistics of the TunFra corpus. The first section provides general speech corpus statistics (? indicates unknown gender). The second section presents code-switching statistics. The third section analyzes vocabulary diversity in both the source and target texts. The final section highlights the prevalence and significance of overlapping speech in the dataset.

3 Experiments and Results

3.1 Data Splitting and Preprocessing

We partitioned our dataset into three distinct sets: training (Train), development (Dev), and testing (Test). The split is performed at the file level, which means that each file is assigned exclusively to a single set. Consequently, no speaker appears in more than one set.

Table 3 provides a breakdown of the speech duration and the number of utterances for each set. Due to the distribution of annotated files, we were limited to including two male speakers in both the development and test sets. Each female speaker participated in at least two audio files, so assigning female speakers to both the Dev and Test sets would have further reduced the available training.

To prepare the data for developing ASR and STT systems, we applied several filtering steps based on the reference transcriptions and translations:

- We excluded samples with empty transcrip-

	Train	Dev	Test
#Segments	7,795	693	701
Duration	13h	01h	50m
#Speakers	37	2	2
Gender: M/F	25/8*	2/0	2/0

Table 3: TUNIFRA corpus split to training, development and testing sets. h and m stand for hours and minutes. *4 speakers are not annotated with gender information.

tions or translations to prevent silent audio segments from being included in the corpus.

- Specific tokens were removed in accordance with our annotation guidelines to maintain clarity in the transcriptions and translations.
- No normalization processing was performed on the Tunisian transcriptions.

3.2 Automatic Speech Recognition

Given that the training set contains only 13 hours of data, this amount is insufficient to train a transformer model from scratch. Therefore, we opt to fine-tune a pre-trained model. Specifically, we select the Whisper model (Radford et al., 2023) for fine-tuning on the Tunisian dialect, as its effectiveness for similar tasks has been demonstrated in previous studies (Talafha et al., 2023; Waheed et al., 2023). We fine-tune the small, medium and large versions of the Whisper model. In addition to our primary approach, we also fine-tune a self-supervised learning (SSL) speech encoder, specifically Wav2Vec 2.0 (Baevski et al., 2020). This encoder is combined with a linear layer, which acts as the decoder to produce transcriptions using CTC loss. By leveraging the knowledge captured during pretraining, this SSL-based pipeline is expected to enhance performance, as reflected in lower word error rates (WER) and character error rates (CER). The results obtained using this method are summarized in Table 4.

Model	Zero-shot WER / CER	TUNIFRA WER / CER
$\text{Whisper}_{\text{Small}}$	104.97 / 72.84	46.78 / 19.04
$\text{Whisper}_{\text{Medium}}$	86.94 / 64.29	37.48 / 14.87
$\text{Whisper}_{\text{Large}}$	90.84 / 62.41	34.22 / 13.72
$\text{Whisper}_{\text{Large-v3}}$	76.46 / 48.50	**29.94 / 11.57**
W2v-Bert + CTC	-	**28.03 / 9.81**

Table 4: ASR results on TUNIFRA test set.

As shown, error rates decrease as the size of the Whisper model increases, both for the original (Zero-shot) models and those fine-tuned on our TUNIFRA dataset. As expected, fine-tuning with TUNIFRA leads to a significant reduction in error rates. Utilizing an SSL model further enhances performance, as demonstrated by Wav2Vec-Bert outperforming the best Whisper model by nearly 2 WER points on the TUNIFRA test set.

3.3 Speech-to-Text Translation

To assess the suitability of our dataset for the STT task, we utilize several systems based on two main approaches: a cascade method (**ASR** $\rightarrow$ **NMT**), where transcriptions generated by the ASR model are subsequently translated by the NMT model; and a direct method (**ASR** + **NMT**), where the speech encoder is coupled with the NMT model to produce translations directly from the audio signal. For the ASR component, we use both models described in the ASR section (Whisper and wav2vec-bert+ctc), while for the NMT component, we employ several sizes of the NLLB (Team et al., 2022) model. Results for the cascade approach are presented in Table 5.

ASR$\rightarrow$ NMT	Dev ($\uparrow$)	Test ($\uparrow$)
$\text{Whisper}_{\text{Small}} \rightarrow \text{NLLB}_{\text{600M}}$	20.59	12.22
$\text{Whisper}_{\text{Small}} \rightarrow \text{NLLB}_{\text{1.3B}}$	21.76	14.10
$\text{Whisper}_{\text{Large-v3}} \rightarrow \text{NLLB}_{\text{1.3B}}$	26.71	17.77
$\text{Whisper}_{\text{Large-v3}} \rightarrow \text{NLLB}_{\text{3.3B}}$	30.74	18.34
W2V-bert $\rightarrow \text{NLLB}_{\text{1.3B}}$	26.62	18.31
W2V-bert $\rightarrow \text{NLLB}_{\text{3.3B}}$	30.06	18.28

Table 5: BLEU score for cascade STT systems using ASR and NLLB models.

Our direct (end-to-end) approach is based on the methodology proposed by (Avila and Crego, 2025). We use the Whisper encoder as the speech encoder and retain NLLB as the NMT model. To bridge the two models, we introduce a CNN layer, applying a transposition before and after the CNN. This process adjusts the speech embeddings by modifying their dimensionality (based on the number of channels) and then restores the original orientation. The input sequence length of NLLB is limited to a maximum of 512 vectors. To achieve this with Whisper, we set the stride value to 3, which reduces the sequence length from 1500 to 500 vectors.

Given the low-resource setting, we are unable to fully fine-tune the entire network. Instead, we restrict updates to the CNN layer and the two adjacent layers on each side, specifically, the last two

Task	Reference	Prediction
ASR	هل آنا تجرأت على أشياء ولا هوما تجرأتوا على الله؟	هل آنا تجرأت على أشياء ولا هوما تجرؤوا على الله؟
STT	*Est-ce que j'ai osé dire des choses ou bien ce sont eux qui ont osé parler de Dieu?*	*Est-ce que maintenant, les personnes qui se débattent sur quelque chose, ou bien les personnes qui se débattent sur Dieu.*
ASR	شيخ مرحبا بيك أهلا وسهلا نورتنا وينورك قلبك	شيخ مرحبا بيك أهلا وسهلا نورتنا إنور قلدك
STT	*Bienvenue Cheikh . . . bienvenue . . . tu nous as honorés . . . que ton cœur soit illuminé.*	*Bienvenue à toi! . . . Bienvenue! . . . Nourra . . . le monstre.*

Figure 1: ASR and STT French hypotheses of two Tunisian Arabic audio segments.

layers of the speech encoder and the first two layers of the NLLB encoder. This strategy is based on the assumption that layers nearest to the CNN are most critical for effective embedding adaptation.

We experiment with two initialization strategies for our end-to-end (E2E) network: using the original pretrained models, and using models that have already been fully fine-tuned on the TUNIFRA dataset. We anticipate that pre-adapting the models to TUNIFRA will enhance performance, as the models will have prior exposure to the dialect, thereby facilitating more effective E2E training. Results for the E2E approach are presented in Table 6.

ASR + NMT	**Dev (↑)**	**Test (↑)**
Whisper/NLLB original pre-trained models		
$\text{Whisper}_{\text{Small}} + \text{NLLB}_{\text{600M}}$	7.40	5.10
$\text{Whisper}_{\text{Small}} + \text{NLLB}_{\text{1.3B}}$	10.85	7.45
$\text{Whisper}_{\text{Large-v3}} + \text{NLLB}_{\text{1.3B}}$	17.86	11.91
Whisper/NLLB adapted to TUNIFRA		
$\text{Whisper}_{\text{Small}} + \text{NLLB}_{\text{600M}}$	11.60	9.35
$\text{Whisper}_{\text{Small}} + \text{NLLB}_{\text{1.3B}}$	16.71	11.62
$\text{Whisper}_{\text{Large-v3}} + \text{NLLB}_{\text{1.3B}}$	22.50	**15.68**

Table 6: BLEU score for STT using our E2E approach. The top section uses original pre-trained models, while the bottom section employs Whisper and NLLB models that were each fine-tuned on the TUNIFRA dataset before being coupled together.

For the STT task, the cascade pipeline clearly outperforms both end-to-end (E2E) approaches, with the best cascade models achieving nearly 3 BLEU points higher than their E2E counterparts. Across all experiments, increasing model size consistently leads to improved performance, as shown in Tables 5 and 6. Additionally, fine-tuning the models on TUNIFRA before jointly training them in the E2E approach with a reshape module yields better results than using the pretrained models directly. This approach results in an improvement of approximately 4 BLEU points for each model pairing. Figure 1 shows two examples of ASR and STT hypotheses. These were generated using our best-performing models: $\text{Whisper}_{\text{Large-v3}}$ and $\text{Whisper}_{\text{Large-v3}}+\text{NLLB}_{\text{1.3B}}$ respectively.

4 Conclusions

By making the data presented in this paper publicly available, we aim to support research in Tunisian and Arabic speech processing, with particular focus in STT. This corpus provides a valuable resource for building more robust models for under-resourced Arabic dialects, advancing both ASR and machine translation. Alongside releasing the corpus, we present baseline results for ASR and STT tasks to support future research and facilitate meaningful comparisons. We encourage further exploration of new architectures, training methods, and data augmentation to improve Tunisian speech processing. We also plan to apply this corpus to code-switching and dialectal speech tasks, aiming to help bridge the digital language divide and improve language technology accessibility.

Limitation

Our end-to-end (E2E) approach has demonstrated efficiency in high-resource settings, matching cascade performance as reported in (Avila and Crego, 2025). However, with limited data, the E2E approach falls short of cascade results. Data scarcity also restricted us to modifying only a few layers during E2E training; with more data, greater model adaptation would be possible. We did not explore data augmentation or incorporate Tunisian data from other corpora (see Table 1) in this work. Future research should investigate additional training pipelines, such as using wav2vec-bert + NLLB in the E2E setup. Given wav2vec's strong results in ASR and cascade S2T, it may offer the best E2E performance as a speech encoder.

5 Acknowledgments

This work has been funded by the French Ministry of Defense through the DGA-RAPID 2022190955, COMMUTE project. This project has received funding from the European Union's Horizon 2020 research and innovation programme ESPERANTO under the Marie Skłodowska-Curie grant agreement No. 101007666. This work was granted access to the HPC resources of IDRIS under the allocations A0161014876 and A0181012527 made by GENCI.

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 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. 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.

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.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

Marko Avila and Josep Crego. 2025. Leveraging large pre-trained multilingual models for high-quality speech-to-text translation on industry scenarios. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7624–7633, Abu Dhabi, UAE. Association for Computational Linguistics.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc.

Kaushal Bhogale, Deovrat Mehendale, Niharika Parasa, Sathish G, Tahir Javed, Pratyush Kumar, and Mitesh Khapra. 2024. Empowering low-resource language asr via large-scale pseudo labeling. pages 2519–2523.

Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, Christopher Klaiber, Pengwei Li, Daniel Licht, Jean Maillard, Alice Rakotoarison, Kaushik Ram Sadagopan, Guillaume Wenzek, Ethan Ye, and 49 others. 2023. Seamlessm4t: Massively multilingual multimodal machine translation. *Preprint*, arXiv:2308.11596.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2023. Fleurs: Few-shot learning evaluation of universal representations of speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 798–805.

Salima Mdhaffar, Fethi Bougares, Renato de Mori, Salah Zaiem, Mirco Ravanelli, and Yannick Estève. 2024. TARIC-SLU: A Tunisian benchmark dataset for spoken language understanding. In *LREC-COLING 2024*, pages 15606–15616, Torino, Italia. ELRA and ICCL.

Hedi Naouara, Jérôme Louradour, and Jean-Pierre Lorré. 2025. Linto audio and textual datasets to train and evaluate automatic speech recognition in tunisian arabic dialect. Good Data Workshop, AAAI 2025.

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.

Bashar Talafha, Abdul Waheed, and Muhammad Abdul-Mageed. 2023. N-shot benchmarking of whisper on diverse arabic speech recognition. pages 5092–5096.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

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

Tianyi Xu, Kaixun Huang, Pengcheng Guo, Yu Zhou, Longtao Huang, Hui Xue, and Lei Xie. 2024. Towards rehearsal-free multilingual asr: A lora-based case study on whisper. pages 2534–2538.