

TEDxTN: A Three-way Speech Translation Corpus for Code-Switched Tunisian Arabic - English

Fethi Bougares^{1,2}, Salima Mdhaffar², Haroun Elleuch^{1,2}, Yannick Estève²

¹ELYADATA, Paris, France, ²Laboratoire Informatique d'Avignon, Avignon, France

Correspondence: fethi.bougares@elyadata.com

Abstract

In this paper, we introduce TEDxTN, the first publicly available Tunisian Arabic to English speech translation dataset. This work is in line with the ongoing effort to mitigate the data scarcity obstacle for a number of Arabic dialects. We collected, segmented, transcribed and translated 108 TEDx talks following our internally developed annotations guidelines. The collected talks represent 25 hours of speech with code-switching that cover speakers with various accents from over 11 different regions of Tunisia. We make the annotation guidelines and corpus publicly available. This will enable the extension of TEDxTN to new talks as they become available. We also report results for strong baseline systems of Speech Recognition and Speech Translation using multiple pre-trained and fine-tuned end-to-end models. This corpus is the first open source and publicly available speech translation corpus of Code-Switching Tunisian dialect. We believe that this is a valuable resource that can motivate and facilitate further research on the natural language processing of Tunisian Dialect.

1 Introduction

Speech translation is the task of translating speech in a given source language into text in another target language. This task is traditionally accomplished through a cascading approach, where a first Automatic Speech Recognition system (ASR) recognizes spoken words, followed by a Machine Translation (MT) system that translates the transcribed text into the target language. This approach is generally criticized because it suffers from cascaded error propagation and high resource and training costs (Sethiya and Maurya, 2025). To overcome these weaknesses, researchers proposed end-to-end (E2E) models (Cho et al., 2014) (Bahdanau et al., 2016) (Vaswani et al., 2023) that generate translation directly from speech in the source language without relying on its transcription

as an intermediate representation. It turns out that this approach is well suited for speech translation from spoken languages characterized by the lack of a standardized orthographic convention, which is the case for multiple low-resourced languages across the world including all Arabic dialects. In addition to being a way to get around the need of source language transcription, E2E models enables a simple and effective framework for transfer learning from pre-trained models on high-resource language pairs. In this work, we report our efforts to collect, annotate, and release the first open-source annotated Tunisian Arabic to English speech translation dataset. We also release a set of ready-to-use Speech Recognition and Speech Translation models alongside with a SpeechBrain recipe and the instructions needed to reproduce our results.

Our contributions are fourfold:

1. **Data:** Release of TEDxTn, the first open source code-switching Tunisian to English speech translation corpus.
2. **Annotation quality:** Consistent and high-quality annotated corpus transcribed by professional transcribers.
3. **ASR and AST:** Development and evaluation of ASR and AST systems using multiple pre-trained Self-Supervised and multilingual models.
4. **Open-Sourcing:** Data¹, annotation guidelines and models are released together with their code and training recipe².

¹Annotations released under a CC BY-NC-ND 4.0 license.

²<https://huggingface.co/datasets/fbougares/TedxTn>

2 Related work

Deep neural network approaches have revolutionized modern Natural Language Processing (NLP) tasks. However, these methods require large amounts of training data, which remain very limited for a large number of languages, including all Arabic dialects. Indeed, despite the considerable effort made to build datasets for multiple Arabic dialects, none of them could be considered today richly resourced. This is the case of all Arabic dialects where available speech datasets are, in general, scarce and even much scarcer when it comes to Code-Switching (CS) speech. CS speech processing has been gaining attention in recent years. This is particularly true for some languages, such as English-Mandarin (Li and Fung, 2013) (Li, 2013) (Chiou et al., 2022) or English-Hindi CS (Dey and Fung, 2014) (Sreeram et al., 2018). Previous works have studied Arabic speakers CS from linguistic and sociolinguistic perspectives (Alowidha, 2024) (Abuhakema, 2013). Arabic speakers often switch from their dialectal Arabic to French or English³. In fact, Arabic speakers generally code-switch to these two languages due to historical factors, since the Arab countries were mainly French and British colonies. Several studies investigated the reasons behind code-switching and pointed out that speakers generally switch for different reasons. People can alternate languages in order to fill a lexical gap, when using more technical terms than native equivalents, to reflect modernity and sophistication, or when using foreign names without translation (Takashi, 1990). In Eldin (2014), the author studied the main drivers of Arabic-to-English switching among Facebook users and highlighted that incompetence, lack of facility, habitual expressions, and the speaker’s mood are the main motivations behind CS.

Although limited, there exist some previous works addressing CS in the domain of Arabic dialect Speech Recognition and Translation. In Elfahal et al. (2019), one hour of mixed Sudanese Arabic and English speech corpus was collected and recorded. Afterward, they used this corpus to train and evaluate a speech recognition system that achieved a 33% word error rate (WER) on a test set of 25 sentences. A much larger amount of work was done to build ArzEn, a larger Egyptian Arabic and English CS corpus (Hamed et al.,

2020). ArzEn is a 12-hour corpus of mixed Egyptian Arabic-English speech. It is a collection of 38 recorded and transcribed interviews on broad topics, including education, hobbies, work, and life experiences. They achieved 57.9% of WER (Hamed et al., 2022) using a CTC/attention-based end-to-end ASR system trained with the ESPnet toolkit (Watanabe et al., 2018). This corpus was also extended to create ArzEn-ST (Hamed et al., 2022), with translations into monolingual Egyptian Arabic and monolingual English. This a three-way speech translation corpus was used to train and evaluate various ASR, MT and AST systems. A multilingual strategy was proposed to model CS in Arabic speech recognition in Chowdhury et al. (2021). They trained an E2E model using Arabic, English, and French data sets. Results are reported for Egyptian and Moroccan dialects. Although a low word error rate (WER) was reported for the Egyptian dialect CS ASR, a higher WER was observed for the Moroccan CS test set. Recently, a 48-hour Multi-dialectal Arabic Speech data set called Casablanca was collected and published in Talafha et al.. This data set aims to mitigate the data scarcity obstacle for a number of Arabic dialects. Casablanca covers eight Arabic dialects. It was used to evaluate various pre-trained SoTA multilingual speech models and fine-tuned Whisper-large-v2 models. We emphasize that only a subset of Casablanca is publicly available and does not include the Tunisian dialect. More details of the current literature on code-switched Arabic NLP are recently presented in Hamed et al. (2025).

With regard to the Tunisian dialect, the number of previous works related to ASR is still limited, and developed data sets are generally not available. Currently, there are only three publicly available ASR Tunisian dialect corpora, namely TARIC (Mdhaftar et al., 2024), TunSwitch (Abdallah et al., 2023) and LinTo (Naouara et al., 2025). TARIC is an 8-hour dataset that target the domain of human-to-human dialogues for train reservation tasks. Therefore, it was transcribed using only Arabic script. TunSwitch, on the other hand, was collected from radio broadcasts that intentionally targeted the Tunisian Code-Switched ASR task (Abdallah et al., 2023). Overall, 8h15m of spontaneous Tunisian speech corpus has been collected as part of TunSwitch data set. This data set was used to train an end-to-end ASR system by fine-tuning the pre-trained speech encoder WavLM (Graves et al., 2013) followed by three dense trainable lay-

³There is one notable exception in Morocco, where some people use Spanish as CS language.

ers trained with CTC loss (Graves et al., 2013). Using a test set of about 25 minutes, the authors reported a WER of 29.47% using an end-to-end ASR system and a 4-gram language model trained with an additional textual corpus of ten thousand monolingual English and French sentences. Recently, (Naouara et al., 2025) extended TunSwitch to create LinTo that contains an annotated data set of 81h38m. A kaldi (Povey et al., 2011) based ASR system was trained using LinTo dataset, and a WER of 20.51% was reported on the TunSwitch test set. In addition to the Tunisian dialect Speech resources mentioned above, there exists a data set used during the IWSLT (Anastasopoulos et al., 2022) evaluation campaign and published recently in the Linguistic Data Consortium (LDC) catalogue. This data set represents 383 hours of manually transcribed conversational speech. A subpart of 160 hours of it is augmented with English translations. This data set was used by several teams within the context of IWSLT to develop multiple ASR and ST systems (Yan et al., 2022) (Yang et al., 2022) (Boito et al., 2022). Although, the latter data set is adapted for Tunisian Arabic to English speech translation, we should point out that, unlike our data set, it is not publicly available and the input speech is conversational telephone recording sampled at 8Khz.

3 Code-switching in Tunisian Arabic

Tunisia is the northernmost country in Africa. Its language is generally referred to as "Tunisian Dialect" or "Tunisian Arabic" or "Tounsi". Tunisia is an ethnically and linguistically homogeneous country, where 98% of Tunisians identify as Arabs and speak Tunisian Dialect (Youssef and Gries., 2023). Today's linguistic situation of Tunisia is strongly shaped by its history, trade, and today's world. That's why Tunisian Arabic co-exists with Modern Standard Arabic (MSA) and French, in a 'triglossic' relationship⁴. As a result of this situation, Tunisian daily communication is characterized by an alternation between multiple languages within a single conversation. This alternation between languages is commonly known as CS. It is defined as "the alternating use of two languages in the same stretch of discourse by a bilingual speaker." (Bul). According to Myers-Scotton (2013) CS is at the same time a mechanism and an outcome of language contact. It is a significant

⁴We would like to highlight an increasing trend towards code-switching with English, compared to French, among Tunisian youth.

and common linguistic phenomenon in Tunisia. It has been shown in Sayahi (2011) that the direction of the switch is almost always from Arabic to French, the most frequently switched categories are single nouns and noun phrases. With regard to CS frequency, the latter shows that education is the most important criterion. People with a higher education code-switch more compared to people with only a high school education. People with a university degree show a much higher frequency of CS, which reflects a higher degree of competence in the French language. However, gender does not affect the frequency of CS. Generally speaking, CS is studied at the sentence boundaries (Myers-Scotton, 1989; Poplack, 1980) and classified into three types: inter-sentential, intra-sentential and extra-sentential switching. The following are descriptions of each type.

- **Inter-sentential switching** defines the situation in which the alternation between languages occurs at sentence boundaries.
- **Intra-sentential switching**, on the other hand, refers to the alternation that occurs within the sentence without any indication of the shift.
- **Extra-sentential switching** also known as **tag-switching** is transplanting a tag from one language to another.

In addition to the above, there exists also the **intra-word switching**, where people change language within a single word occurs where Tunisian speakers attach Arabic clitics and affixes to foreign French or English words. Table 2 provides a concrete example for each CS in the Tunisian dialect type extracted from the TEDxTN corpus.

4 Corpus Creation

4.1 Data Collection

The source for this corpus is a collection of TEDx talks⁵. TEDx events are planned and coordinated independently. TEDx talks share the same format as TED talks. However, while TED talks are all in English, TEDx talks can be in a variety of languages, including local spoken languages and dialects. TEDx events aim to help communities, organizations and individuals produce TED-style events at the local level. They are planned and coordinated independently, on a community-by-community basis, under a free license from TED. TEDx talks are

⁵<https://www.ted.com/about/programs-initiatives/tedx-program>

a valuable source for multiple speech processing tasks. They have been used to create many data sets for many languages and multiple tasks. Some examples of this are: (1) TED-LIUM (Hernandez et al., 2018) created for English speech recognition; (2) MTEDx (Salesky et al., 2021) built to support speech recognition and speech translation research across many languages and (3) TED-EL (Li et al., 2024) created for Speech Entity Linking. TEDx talks are particularly a valuable source for speech processing of low-resource languages. However, they are usually the fruit of local non-funding initiatives. Therefore, the available recordings may be difficult to find on the Internet⁶ or of poor audio quality for speech processing tasks. Another notable difference between TEDx and TED events is the lack of volunteers who subtitle TEDx talks. In this version of the TEDxTN corpus, we were able to collect 108 talks with an acceptable audio quality ranging from 2 to 23 minutes. The audio quality of each TEDx talk was manually verified before saving it in WAV format sampled at 16Khz. Table 1 shows some key statistics about the collected data.

TEDxTN Corpus	
#Tedx Talks	108
#Tedx Events	38
#Different cities	11
Languages	TN/FR/EN
Date range	2010 - 2023
#Speakers	130
Total audio duration	28h39min

Table 1: Overview of the TEDxTN corpus.

4.2 Corpus Annotation

All TEDxTN talks were manually transcribed by professional tri-lingual (Arabic, English, and French) transcribers. Like all Arabic dialects, Tunisian Arabic does not have a standardized orthography. Therefore, words can have multiple correct spellings, and several letters can be used interchangeably. Moreover, in the context of CS speech, some loanwords are adapted and transformed by changing their pronunciation or integrating number, gender, or case agreement. All of that makes the definition and application of a unified transcription

⁶Unlike TED talks, TEDx talks are not gathered in a common website and could be sometimes shared only on personal social network accounts

guideline particularly challenging. We have chosen to follow the CODA* (Habash et al., 2018) design principles to develop our annotation guidelines. Although CODA* included a seed lexicon, it remains limited and covers only five dialects (GLF, MOR, EGY, TUN, and LEV) with a very small lexicon for each dialect. In the context of this work, we derived a set of rules and patterns used to unify as much as possible the spelling for each annotator and between annotators. Below are some annotation rules extracted from our transcription and translation guidelines:

1. Use Arabic script for Arabic words and Roman script for foreign words.
2. Use Arabic script for foreign words when they are adapted to Tunisian dialect.
3. Arabic clitics and affixes are written in Arabic script, and French or English words are written in Roman script. For example "هذا point عال", "*About this point*" in English.
4. Use a predefined fixed spelling for common words like days of the week, numbers, quantities, percentages, etc.
5. Negative pronouns are written attached such as مانيش, "*I am not*" in English.
6. Translate to provide natural translations with the intended meaning rather than literal translations.
7. Translate foreign words (i.e French) into fluent English while preserving the meaning present in the original code-switched text.
8. Disfluencies such as partial words and repetitions should also be included in translations.

In order to ensure a high-quality dataset, we followed a two-stage transcription process. The first stage takes as input the audio files and produces a segmented output with an initial transcription that may contain transcription errors or may also not be fully compliant with the transcription guidelines. 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. The English translation is performed using the Tunisian transcription with possible access to the corresponding audio recording in case of need. In most cases, we followed

the LDC Arabic-to-English Translation Guidelines (LDC, 2013).

4.3 Corpus Statistics

In this section, we present an overview of the annotated TEDxTN corpus. Table 3 includes the number of transcribed segments, words, and speakers. It also includes total speech duration, average segment length (words) and duration (seconds), as well as gender distribution.

Category	Value
# Segments	17,278
# different speakers	130
Speech duration	25h01min
Avg segment Duration (seconds)	5.20 sec
Gender dist (M/F) - Count	86/44
Gender dist (M/F) - Duration	18h/07h
#Total source words	321,220
#Src TUN words	177,079
#Src Intra CS words	4,176
#Src foreign words	43,932
#Seg. full Tun	7,979
#Seg. full foreign	459
#Seg. mixed	9,299
#Src Vocab size	31,064
#Total target Words (Translation)	280,353
#Target Vocab size	20,982

Table 3: Detailed statistics of TEDxTN corpus.

As reported in Table 3, we were able to transcribe around 25 hours of speech out of 28 hours and 39 minutes of audio signal (87.3%). This represents about 17.2k segments containing more than 321k words from 133 different speakers and a vocabulary size of around 31k.

4.4 Code switching statistics

Only 7,963 segments out of 17,200 total segments are fully in Tunisian dialect. That means that 53.70% of the TEDxTN-ST corpus segments contain at least one foreign word. In order to better quantify the amount of code-switching present in TEDxTN-ST data, we calculate the Code-Mixing Index (CMI). CMI was introduced by Das and Gambäck (2014) as a method to compare different code-mixed corpora to each other. CMI is defined as:

$$CMI = \frac{\sum_{i=1}^N w_i - \max\{w_i\}}{n - u} \quad (1)$$

where $\sum_{i=1}^N w_i$ is the total number of words from N languages, w_i is the number of words in language i , n is the total number of words regardless of language, and u is the number of tokens given language-independent tags. CMI is equal to 0 for utterances that contain only tokens from one language. A high CMI score is an indicator of the high degree of code-mixing in the text. The CMI for the entire TEDxTN corpus is 21.50%. This indicates a high rate of CS in this corpus. As shown in Table 3, we also include statistics on the number of Tunisian words, written in Arabic script, (**Src. w TUN**), the number of foreign words fully spelled in Latin script (**Src w. foreign**) and the number of words written using a mix of Arabic and Latin script (a.k.a intra-word switching). On the word level, among code-mixed sentences (the 9,299 sentences reported in **#Seg. mixed** row), 67.78% of the words are Arabic, 26.38% are foreign, and 5.84% are intra-words code-switch.

4.5 Trigger Words

As defined in Hamed et al. (2018), code-switching trigger words are words that can prompt a bilingual speaker to switch languages during a conversation. TEDxTN includes 2729 unique Arabic code-switching trigger words.

Word	English	Frequency
ال	The	2,688
أ - آ	Hesitation	988 / 225
و	And	436
في	In	399
لل	To	247
متاع	Belongs to	200
بال	With	173
معتهها	This means	129

Table 4: TEDxTN most frequent trigger words.

Table 4 shows the most common trigger words that precede a code switching point in TEDxTN. The most common switches occur after the definite article ال (The). This is reasonable because ال is placed before a foreign noun or adjective that the speaker wants to specify. The و and في trigger words are aligned with the observations reported for the Egyptian dialect in Hamed et al. (2018). As for متاع and معتهها, they are very common transition

CS type	TEDxTN samples
Extra-sent	<u>Mais bon it happens.</u> أذاكا لي خلاياي إنزید إندافع أكثر. Anyways, it happens. This is what made me defend them even more.
	<u>حبتي.</u> c’était un déclenchement d’amour بالجواب هذاكا. She loved me. Love was triggered through this letter.
Intra-sent	<u>إذا كان نعطيكم تويكا نص ساعة ال كلمم تكتبوا لي</u> five statements على رواحكم. So, when I saw it my heart started beating fast and I said to myself "Isn't this it?"
	<u>عند</u> comme je suis هبطت لل bouquinistes. Stubborn as I used to be. I went to the booksellers.
Inter-word	<u>الشباب أذاكا وال l'énergie</u> اللي عنده وال passion اللي عنده أذیکا ثروة. These young people and their energy and passion are wealth.
	<u>وقت لي مشيت لل</u> préparatoire كان عندي حلمة. When I started studying at the preparatory institute, I had a dream.

Table 2: Examples of different CS types in TEDxTN corpus followed by their English translation. The underlining marks the non-Tunisian phrases and their corresponding translation in English.

words used in Tunisian dialect.

5 Experiments and results

5.1 Data split

We created standardized data splits for training, validation, and evaluation. We have chosen to put full TEDx talks in dev and test sets in order to avoid contamination between the training and evaluation. The number of talks, segments, and words are reported in Table 5. We also report the total duration, the number of unique speakers, the gender distribution, and the CMI score per dataset. As shown in Table 5, speakers belonging to the validation and test subsets are not seen in the training set. In addition, validation and test sets have higher CMI scores compared to training data. Finally, we also ensured that both male and female speakers are kept within the validation and test set.

	Train	Valid.	Test
#Talks	97	05	06
#Segments	15,626	731	842
#Words	205,753	11,250	11,834
Duration	22h40m	01h07m	01h14m
#Speakers	117	10	07
CMI score	20.66	24.37	33.09
Gender: M/F	77/40	5/5	5/2

Table 5: TEDxTN corpus split to train, valid and test.

5.2 Automatic Speech Recognition

Given the relatively small size of our datasets, we opt for a fine-tuning approach rather than training a Tunisian dialect ASR system from scratch. As regards the choice of the pre-trained models to use, there are various options available to us, ranging from small models with a few hundred million parameters to bigger models with around 1 billion parameters. In addition to the model size, we also have the choice between multiple model architectures. In this work, we experimented with fine-tuning 5 different pre-trained models. Namely, we use the TEDxTN training set to adapt Whisper (Radford et al., 2022), Massively Multilingual Speech (MMS) (Pratap et al., 2023), XLSR (Babu et al., 2021) and w2v-Bert-2.0T (Communication et al., 2023) models. All of our experiments were performed using the SpeechBrain toolkit (Ravanelli et al., 2024) without using a language model. All our models were trained for 80 epochs. For whisper based models, we used the original encoder-decoder architecture without any parameters freezing. MMS and XLSR models are trained using an additional linear layer of size 1024 followed by a Connectionist Temporal Classification layer (CTC) for transcribing the labels. Finally, for W2v-Bert-2.0T model we added two transformer layers of size 1024 each, followed by a CTC layer for transcribing the labels. We used Adam optimizer for all our ASR models. Our results are reported in Ta-

Model	Model Size (#Params)	Valid.		Test	
		WER (↓)	CER (↓)	WER (↓)	CER (↓)
Whisper-small (zero-shot)	244M	133.24	100.00	183.81	142.00
Whisper-meduim (zero-shot)	769M	130.61	103.00	150.71	122.00
Whisper-Lg-v3 (zero-shot)	1550M	92.50	63.20	94.00	67.90
Whisper-Small	244 M	26.66	11.81	27.78	13.38
Whisper-Medium	769 M	23.10	10.46	25.37	13.00
Whisper-Lg-v3	1550 M	22.72	09.77	25.19	12.33
MMS Large	316.6 M	35.43	13.02	37.29	14.67
MMS 1B	964.3 M	26.78	09.90	27.91	11.27
XLSR Large	316.6 M	35.74	13.79	37.11	15.26
XLSR 1B	964.3 M	28.12	10.82	29.98	12.12
w2v-Bert-2.0T	590.1 M	19.92	07.10	21.37	08.34

Table 6: ASR results of TEDxTN Tunisian Arabic speech. Lower WER and CER indicate better quality.

ble 6. We evaluated Whisper (large-v3), one of the best multilingual open source speech recognition models, in a zero shot setting (line 1 in Table 6) and we observed, as shown in previous work for other Arabic dialects (Talafha et al., 2023), that Whisper did not reach reasonable performance with 92.50% and 94.00% WER on TEDxTN dev and test sets, respectively.

Fine-tuning Whisper models using domain-specific data results in a significant reduction in WER. We started by fine-tuning Whisper-small, which already gives a significant WER reduction compared to a much bigger model (large-v3) in a zero-shot setting. Using larger Whisper models (large-v3) incrementally decreases the WER to achieve **25.19%** WER on the test set. We also report obtained results when fine-tuning MMS Large and MMS 1B models. As shown in the table, MMS 1B obtained better results compared to MMS Large. However, MMS 1B results are comparable to Whisper-small, although the latter has about 4 times fewer parameters. An interesting observation is that fine-tuning the w2v-Bert-2.0T model gives much better results compared to Whisper-large-v3 while having 3 times fewer parameters. w2v-Bert-2.0T model achieves **19.92%** and **21.37%** WER on TEDxTN dev and test sets, respectively.

5.3 Automatic Speech Translation

For the same reasons set out above, we decided to opt for a fine-tuning approach of pretrained models. We started by using pre-trained Speech translation models. Particularly speaking, we began by translating the dev and test set in a zero shot fashion using different pre-trained Whisper models. Next,

we fine-tuned these models using TEDxTN dataset. To be consistent with the ASR experiments, we kept the same data split used reported in section 5.1. All our translation outputs are evaluated with TrueCased BLEU score without punctuation using sacrebleu (Post, 2018). Table 7 shows the speech translation performance of each trained model. All our models are fine-tuned for 80 epochs following the default Whisper recipe of SpeechBrain toolkit.

Model	Valid. (↑)	Test (↑)
Whisper-small (zero-shot)	3.98	5.70
Whisper-med. (zero-shot)	10.23	12.85
Whisper-lg-v3 (zero-shot)	10.96	13.95
Whisper small	17.31	18.53
Whisper med.	23.02	24.50
Whisper-lg-v3	25.19	25.68

Table 7: BLEU scores of TEDxTN Speech Translation.

As expected, we obtained better BLEU score using larger Whisper models for both zero-shot and fine-tuning settings. For instance, the best zero-shot BLEU scores are obtained using whisper-Large-v3 (row whisper-Lg-v3) with 10.96 and 13.95 for valid and test respectively. Likewise, whisper-Large-v3 achieves the best performing model after fine-tuning on TEDxTN training set with **25.19** and **25.68** BLEU scores for validation and test sets respectively. Note that we trained also speech translation models by feeding WLSR and w2v-Bert-2.0T encoders outputs to the NLLB decoder, in its 1.3B parameters configuration. Contrary to what we expected, the model did not in exceed a BLEU score of 5. More investigation of this model is left to be done in future work.

Output	TEDxTN Samples
Reference	ما يحبش بيدل Ooredoo في متاعه Il il est près à passer le reste de jours
Prediction	ما يحبش بيدل في أوريدو ما يحبش بيدل Il il est près à passer le reste de jours
Reference	أنا شمة شكون ما يتكيفش ويقعد في ال+fumoir+إنحب نقول له حاجة
Prediction	أنا شمة شكون ما يتكيفش ويقعد في الفيومار إنحب نقول له حاجة
Reference	تي حتى مالمسيركيا سيون ولات أمورها واضحة
Prediction	تي حتى مال circulation ولات أمورها واضحة
Reference	à toi وحدك تنجم تبدل ال mentalité متاع تونس عال chef
Prediction	à toi وحدك تنجم تبدل ال mentalité متاع تونس عالشاف

Table 8: Examples of ASR (w2v-Bert-2.0T model) errors from TEDxTN test set.

5.4 Qualitative Analysis

Speech transcription: To understand the quality of ASR transcription per segment type, we divided the test set into the following 3 subsets: (1) **TUN** subset with segments uttered only in the Tunisian dialect, (2) **MIXED** subset includes code-switching segments and (3) **FOR** subset with segments fully in foreign language. Using our best ASR system (w2v-Bert-2.0T from Table 6), we calculated the WER for each subset.

	TUN	MIXED	FOR	ALL
#Seg.	215	551	76	842
#Words	1,912	9,282	639	11,833
WER (↓)	24.16	21.31	13.93	21.37

Table 9: ASR Error analysis per segment type.

As shown in Table 9, most ASR errors are made for Tunisian-only and code-switched segments. Manual inspection of the code-switched segments shows that the system outputs the correct transcription but using different script from the one used in the reference. Some examples of this are provided in Table 8. In the first example, our ASR system transcribed the word الفيموار in Arabic while this word is written using the prefix ال plus the same word in Latin script "fumoir". Same for the word "circulation", but in the opposite direction: The reference is written in Arabic script (مالمسيركيا سيون) while the human transcription is in Latin script. As regards, speech translation system, we analyzed the output of the fine-tuned Whisper large-v3 model but no particular error pattern was identified.

6 Conclusion and future works

In this study, we propose TEDxTN, the first Tunisian Arabic to English Code-switching Speech Translation annotated corpus. **TEDxTn** is carefully annotated by linguistic experts following a detailed annotation guideline. This corpus was used to train and evaluate multiple strng Tunisian dialect speech transcription and translation baselines. Our best models achieves **21.37%** WER and **25.68** BLEU scores on the transcription and translation tasks of TEDxTn test set respectively. We believe that this corpus fills an important resource gap in Code-switching research for Tunisian dialect. For future work, we plan to extend **TEDxTn** as new talks are available and use it for other NLP tasks.

Ethical considerations and limitations

Like any other dataset, the collected speech corpus is not representative of all the spoken forms of Tunisian Dialect. This corpus is likely unbalanced in terms of any demographic aspect since it includes talks from only 11 different cities in Tunisia. Nevertheless, we think that the lack of previous code-switching speech Tunisian Arabic to English translation data set, would make it valuable resource for training and evaluating code-switching speech models of Tunisian Dialect. TEDx talks are governed by the CC BY-NC-ND 4.0 license. Under this license, "NoDerivatives" implies that any modifications, remixes, or transformations cannot be distributed. In compliance with this we distribute only transcriptions and translations. For the audio recordings, we provide the YouTube URL of each video for users to download.

Acknowledgments

This work was performed using HPC resources from GENCI-IDRIS (grant AD011015051R1) and received funding from the ESPERANTO research and innovation programme under the Marie Skłodowska-Curie (grant No 101007666).

References

- The Cambridge Handbook of Linguistic Code-switching*. Cambridge Handbooks in Language and Linguistics. Cambridge University Press, 2009.
- 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.
- Ghazi Abuhakema. 2013. Code switching and code mixing in arabic written advertisements: Patterns, aspects, and the question of prestige and standardisation.
- Kais Sultan Mousa Alowidha. 2024. English-arabic code switching and identity in bilingual saudis living in saudi arabia: A comparative study between large and small cities. *Educational Administration: Theory and Practice*, 30(5).
- Antonios Anastasopoulos, Loïc Barrault, Luisa Bentivogli, Marceley Zanon Boito, Ondřej Bojar, Roldano Cattoni, Anna Currey, Georgiana Dinu, Kevin Duh, Maha Elbayad, Clara Emmanuel, Yannick Estève, Marcello Federico, Christian Federmann, Souhir Gahbiche, Hongyu Gong, Roman Grundkiewicz, Barry Haddow, Benjamin Hsu, and 24 others. 2022. [Findings of the IWSLT 2022 evaluation campaign](#). In *IWSLT*, pages 98–157, Dublin, Ireland.
- Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2021. [Xls-r: Self-supervised cross-lingual speech representation learning at scale](#). *Preprint*, arXiv:2111.09296.
- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2016. [Neural machine translation by jointly learning to align and translate](#). *Preprint*, arXiv:1409.0473.
- Marceley Zanon Boito, John Ortega, Hugo Riguidel, Antoine Laurent, Loïc Barrault, Fethi Bougares, Firas Chaabani, Ha Nguyen, Florentin Barbier, Souhir Gahbiche, and 1 others. 2022. On-trac consortium systems for the iwslt 2022 dialect and low-resource speech translation tasks.
- Chung-Pu Chiou, Hou-An Lin, and Chia-Ping Chen. 2022. Mandarin-English code-switching speech recognition system for specific domain. In *ROCLING*.
- Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. [Learning phrase representations using rnn encoder-decoder for statistical machine translation](#). *Preprint*, arXiv:1406.1078.
- S. A. Chowdhury, A. Hussein, Ahmed Abdelali, and Ahmed Ali. 2021. Towards one model to rule all: Multilingual strategy for dialectal code-switching arabic asr. *ArXiv*, abs/2105.14779.
- Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenhaler, Paul-Ambroise Duquenne, Brian Ellis, Hady Elsahar, Justin Haaheim, and all. 2023. [Seamless: Multilingual expressive and streaming speech translation](#). *Preprint*, arXiv:2312.05187.
- Amitava Das and Björn Gambäck. 2014. Identifying languages at the word level in code-mixed Indian social media text. In *ICON*, Goa, India.
- Anik Dey and Pascale Fung. 2014. A Hindi-English code-switching corpus. In *LREC*.
- Ahmad Abdel Tawwab Sharaf Eldin. 2014. Socio linguistic study of code switching of the arabic language speakers on social networking. *International Journal of English Linguistics*, 4:78.
- Mohammed O. Elfahal, Mohammed Elhafiz Mustafa Supervisor, and Rashid A. Saeed Co-Supervisor. 2019. Automatic recognition and identification for mixed sudanese arabic – english languages speech.
- Alex Graves, Abdel rahman Mohamed, and Geoffrey E. Hinton. 2013. Speech recognition with deep recurrent neural networks. *ICASSP*.
- Nizar Habash, Fadhil Eryani, Salam Khalifa, Owen Rambow, Dana Abdulrahim, Alexander Erdmann, Reem Faraj, Wajdi Zaghouni, Houda Bouamor, Nasser Zalmout, Sara Hassan, Faisal Al-Shargi, Sakhar B. Alkhereyfy, Basma Abdulkareem, Ramy Eskander, Mohammad Salameh, and Hind Saddiki. 2018. Unified guidelines and resources for arabic dialect orthography. In *LREC*.
- Injy Hamed, Mohamed Elmahdy, and Slim Abdennadher. 2018. Collection and analysis of code-switch Egyptian Arabic-English speech corpus. In *LREC*.
- Injy Hamed, Nizar Habash, Slim Abdennadher, and Ngoc Thang Vu. 2022. ArEn-ST: A three-way speech translation corpus for code-switched Egyptian Arabic-English. In *WANLP*.
- Injy Hamed, Caroline Sabty, Slim Abdennadher, Ngoc Thang Vu, Tamar Solorio, and Nizar Habash. 2025. A survey of code-switched Arabic NLP: Progress, challenges, and future directions. In *COLING*.
- Injy Hamed, Ngoc Thang Vu, and Slim Abdennadher. 2020. ArEn: A speech corpus for code-switched Egyptian Arabic-English. In *LREC*.

- François Hernandez, Vincent Nguyen, Sahar Ghannay, Natalia Tomashenko, and Yannick Estève. 2018. *TED-LIUM 3: Twice as Much Data and Corpus Repartition for Experiments on Speaker Adaptation*, page 198–208. Springer International Publishing.
- Linguistic Data Consortium LDC. 2013. Bolt program: Arabic to english translation guidelines.
- Fung P. Li, Y. 2013. Language modeling for mixed language speech recognition using weighted phrase extraction. In *INTERSPEECH 2013*, Lyon, France.
- Silin Li, Ruoyu Song, Tianwei Lan, Zeming Liu, and Yuhang Guo. 2024. *TED-EL: A corpus for speech entity linking*. In *LREC-COLING 2024*, pages 15721–15731, Torino, Italia. ELRA and ICCL.
- Ying Li and Pascale Fung. 2013. Improved mixed language speech recognition using asymmetric acoustic model and language model with code-switch inversion constraints. In *ICASSP*.
- 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*.
- Carol Myers-Scotton. 1989. Codeswitching with english: types of switching, types of communities. *World Englishes*, 8:333–346.
- Carol. Myers-Scotton. 2013. Contact linguistics: Bilingual encounters and grammatical outcomes.
- 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.
- Shana Poplack. 1980. *Sometimes i’ll start a sentence in spanish y termino en español: toward a typology of code-switching 1*. *Linguistics*, 18:581–618.
- Matt Post. 2018. *A call for clarity in reporting BLEU scores*. In *Conference on Machine Translation*, pages 186–191, Brussels, Belgium. ACL.
- Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Luká Burget, Ondrej Glembek, Nagendra Kumar Goel, Mirko Hannemann, Petr Motlíček, Yanmin Qian, Petr Schwarz, Jan Silovský, Georg Stemmer, and Karel Veselý. 2011. The kaldi speech recognition toolkit.
- Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2023. *Scaling speech technology to 1,000+ languages*. *Preprint*, arXiv:2305.13516.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. *Robust speech recognition via large-scale weak supervision*. *Preprint*, arXiv:2212.04356.
- Mirco Ravanelli, Titouan Parcollet, Adel Moumen, Sylvain de Langen, Cem Subakan, Peter Plantinga, Yingzhi Wang, Pooneh Mousavi, Luca Della Libera, Artem Ploujnikov, and 1 others. 2024. Open-source conversational ai with speechbrain 1.0. *Journal of Machine Learning Research*, 25(333):1–11.
- 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*. *Preprint*, arXiv:2102.01757.
- Lotfi Sayahi. 2011. *Code-switching and language change in tunisia*.
- Nivedita Sethiya and Chandresh Kumar Maurya. 2025. *End-to-end speech-to-text translation: A survey*. *Comput. Speech Lang.*, 90(C).
- Ganji Sreeram, Kunal Dhawan, and Rohit Sinha. 2018. *Hindi-english code-switching speech corpus*. *CoRR*, abs/1810.00662.
- Kyoko Takashi. 1990. A sociolinguistic analysis of english borrowings in japanese advertising texts. *World Englishes*.
- Bashar Talafha, Karima Kadaoui, Samar Mohamed Magdy, Mariem Habiboullah, Chafei Mohamed Chafei, Ahmed Oumar El-Shangiti, Hiba Zayed, Mohamedou Cheikh Tourad, Rahaf Alhamouri, Rwa Assi, Aisha Alraeesi, Hour Mohamed, Fakhreddin Alwajih, Abdelrahman Mohamed, and booktitle = El Mekki, Abdellah et al. Casablanca: Data and models for multidialectal Arabic speech recognition.
- Bashar Talafha, Abdul Waheed, and Muhammad Abdul-Mageed. 2023. *N-shot benchmarking of whisper on diverse arabic speech recognition*. *Preprint*, arXiv:2306.02902.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. *Attention is all you need*. *Preprint*, arXiv:1706.03762.
- Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplín, Jahn Heymann, Matthew Wiesner, Nanxin Chen, Adithya Renduchintala, and Tsubasa Ochiai. 2018. *EspNet: End-to-end speech processing toolkit*. In *Interspeech*.
- Brian Yan, Patrick Fernandes, Siddharth Dalmia, Jiantong Shi, Yifan Peng, Dan Berrebbi, Xinyi Wang, Graham Neubig, and Shinji Watanabe. 2022. *Cmu’s iwslt 2022 dialect speech translation system*. In *IWSLT 2022*, pages 298–307.
- Jinyi Yang, Amir Hussein, Matthew Wiesner, and Sanjeev Khudanpur. 2022. *Jhu iwslt 2022 dialect speech translation system description*. In *IWSLT*, pages 319–326.
- Chadi Ben Youssef and Stefan Th. Gries. 2023. Code-switching in tunisian arabic: A multifactorial random forest analysis. In *Corpora, Volume 18 Issue 3*.