International Journal of Engineering Trends and Technology
ISSN: 2231–5381 / https://doi.org/10.14445/22315381/IJETT-V72I10P125

Volume 72 Issue 10, 257-266, October 2024

Original Article

DARIJA-C: A Crowdsourced Corpus for Moroccan DARIJA Speech-to-Text Translation

Maria Labied[1], Abdessamad Belangour[2], Mouad Banane[3]

[1,2]*Laboratory of Information Technology and Modeling LTIM Hassan II University, Ben M'sik Faculty of Sciences Casablanca, Morocco.*
[3]*Department, Laboratory of Artificial Intelligence & Complex Systems Engineering Hassan II University, Faculty of Legal, Economic, and Social Sciences Casablanca, Morocco.*

[1]*Corresponding Author : mr.labied@gmail.com*

Received: 25 May 2024 Revised: 08 October 2024 Accepted: 10 October 2024 Published: 25 October 2024

Abstract *- This paper outlines the development of a Moroccan Darija speech corpus named "DARIJA-C". The primary goal of this corpus is to facilitate the automatic translation of spoken Moroccan Darija into Modern Standard Arabic (MSA) text, offering potential applications across various sectors, including communication, education, and technology. To support ongoing and scalable data collection, we established a web platform that allows for the recording of speech and its corresponding text translation into MSA by anonymous contributors. Future iterations of this project aim to include translations into multiple international languages. The overarching aim of this initiative is to compile the largest and most diverse corpus of Moroccan Darija speech paired with textual translations in various languages. This will create a pioneering resource for the translation of Moroccan Darija speech into multiple languages, thus significantly contributing to the field of speech recognition and translation.*

Keywords *- Moroccan Darija, Speech corpus, Automatic speech recognition, Speech-to-text translation, Crowdsourcing, Modern standard arabic, Multilingual translation, Speech dataset, Language resources, DARIJA-C.*

1. Introduction

Building Automatic Speech Recognition (ASR) and Speech-to-Text (S2T) translation systems for under-resourced dialects such as Moroccan Darija is challenging due to the scarcity of essential resources like speech corpora, lexicons, and language models [1], [2]. These systems require large amounts of data for effective training, making the data collection process difficult and time-consuming. The diverse nature of Moroccan Darija, with significant variations in pronunciation, vocabulary, and grammar across regions and among speakers [3], further complicates this task. Additionally, the absence of standardized orthography leads to wide variations in transcriptions, affecting the creation of reliable datasets and, consequently, the performance of ASR and S2T systems. Despite these challenges, developing ASR and S2T systems for Moroccan Darija holds substantial promise. These technologies can enhance communication, educational opportunities, and access to information for Moroccan Darija speakers, fostering greater inclusion in the digital age. Efforts to create comprehensive and representative speech corpora for Moroccan Darija are crucial. Such corpora serve as foundational resources for training and evaluating ASR and S2T models. To address these needs, this project leverages crowdsourcing techniques, enabling the collection and validation of large volumes of speech data from a diverse group of speakers. This approach ensures that the dataset reflects the rich linguistic diversity of Moroccan Darija, including different accents, speaking styles, and regional variations. The construction of the DARIJA-C speech corpus significantly contributes to developing ASR and S2T systems for Moroccan Darija. A standardized dataset that accurately represents the phonology, syntax, and lexicon of Moroccan Darija helps overcome the challenges associated with developing systems for under-resourced dialects. This dataset serves as a critical resource for training and evaluating these systems, ensuring robustness and effectiveness in handling the linguistic nuances of Moroccan Darija. Furthermore, the corpus enables the development of effective language and acoustic models that capture the unique features of Moroccan Darija speech, supporting accurate recognition and translation of dialectal speech. The DARIJA-C corpus is designed to facilitate the development of high-performance ASR and S2T systems by offering a large and diverse collection of annotated speech data. This corpus can serve as a benchmark for evaluating model performance, allowing for meaningful comparisons and advancements in the field. By making the corpus freely available under a Creative Commons CC0 license, and aim to promote research and development efforts

on speech recognition and translation for Moroccan Darija. Open access to this resource can stimulate innovation and collaboration, leading to significant improvements in communication, information access, and technological inclusion for speakers of this dialect.

Building a high-quality speech corpus is challenging and time-consuming, requiring careful consideration of factors such as sampling rate, recording quality, speaker diversity, and transcription accuracy. The success of ASR and S2T systems heavily relies on the quality of the data they are trained on. High sampling rates and clear recording quality are essential to capture speech nuances accurately. At the same time, a diverse set of speakers ensures that the corpus represents the wide range of variations in Moroccan Darija. Ethical considerations, including privacy and consent of contributors, are paramount in the data collection process. Transparent and comprehensive consent processes ensure that all participants are fully informed about the use of their data and have willingly agreed to contribute. To achieve these goals, A community platform for collecting and validating DARIJA speech and translations is created. This platform enables the collection of large volumes of speech data from a diverse group of speakers, enhancing the corpus's representativeness and quality. By leveraging the power of crowdsourcing, a wide variety of speech samples are gathered, crucial for developing robust and generalizable ASR and S2T systems.

The validation process is critical for ensuring the dataset's quality and accuracy. This platform allows community members to verify the transcription and translation of recorded speech, helping to reduce errors and increase accuracy. Community-driven validation ensures the dataset remains reliable and high-quality, with multiple individuals cross-checking each entry, providing a robust mechanism for error correction and quality assurance. By making this corpus available for free download, it can serve as a valuable resource for researchers and developers working on Natural Language Processing (NLP) tasks for Moroccan Darija. Open access to this corpus can facilitate comparative studies and benchmarking, contributing to the advancement of speech recognition technologies for under-resourced dialects. Through its comprehensive and high-quality data, the DARIJA-C corpus aims to be a cornerstone resource for the continued growth and development of NLP applications for Moroccan Darija.

The following sections provide a comprehensive benchmark of related works focused on creating speech recognition datasets and compare the various methods used to collect the corpus in each study (Section 2). This comparison will highlight the strengths and limitations of existing approaches, offering valuable insights into the best practices and common challenges in the field. The aim is to contextualize the work within the broader landscape of speech recognition research, demonstrating the need for and contribution of the DARIJA-C corpus. The process of collecting and verifying our DARIJA-C corpus, including the creation of a community platform, is also detailed (Section 3). This section will delve into the specifics of data collection methodology, explaining how the issues addressed, such as speaker recruitment, recording procedures, and data validation. By leveraging crowdsourcing, The aim is to gather a diverse and representative dataset, ensuring high-quality transcriptions and translations through community verification. Section 4 analyzes the characteristics of the dataset, considering factors such as speaker diversity, recording quality, transcription accuracy, and overall size. This analysis will provide a thorough overview of the dataset's attributes, emphasizing its comprehensiveness and potential utility for various ASR and S2T applications. The measures taken to ensure the dataset's robustness and reliability are discussed, such as quality control protocols and data normalization techniques. Finally, the paper concluded by summarizing the contributions of the work and suggesting potential avenues for future research. The impact of the DARIJA-C corpus is reflected in the field of speech recognition, particularly for under-resourced dialects like Moroccan Darija. Additionally, possible directions for expanding and enhancing the dataset, including the incorporation of more diverse linguistic features, the addition of new languages for translation, and the application of advanced machine learning techniques, are outlined to improve ASR and S2T systems further. By highlighting these future research opportunities, we. This research aims to inspire continued innovation and collaboration in the development of speech technologies for Moroccan Darija and other under-resourced languages.

2. Related Work

The development of spoken language recognition and translation systems relies heavily on large-scale annotated corpora. The availability and quality of such corpora are crucial for training and evaluating deep learning models for these tasks. High-quality corpora provide the diverse and extensive data necessary for models to learn the complexities of spoken language, including phonetic variations, syntactic structures, and contextual nuances. However, building such corpora is a time and resource-intensive process, and the resulting datasets often have limitations. Some corpora may be limited in terms of language coverage, focusing on only a few languages or dialects. This narrow focus restricts the applicability of the resulting models to a wider array of languages and dialects, especially those that are under-resourced. Additionally, speaker diversity within these datasets can be limited, impacting the generalizability of the models. A dataset lacking sufficient variation in speaker attributes such as age, gender, accent, and socio-economic background may result in models that perform well on certain demographics but poorly on others. The morphology and syntax of some languages may be complex, which complicates

annotation and modeling. Languages with intricate grammatical rules, rich inflectional morphology, or significant regional dialectal variations pose additional challenges for corpus development. Annotators must possess a deep understanding of these linguistic features to accurately transcribe and label the data, which adds to the complexity and cost of the process. Moreover, the availability of corpora poses problems due to licensing restrictions. Some corpora may be proprietary and not accessible for research purposes, limiting the ability of researchers to build on previous work. Licensing issues can hinder the free exchange of data and methodologies, slowing down progress in the field. Open access to datasets is crucial for fostering innovation and collaboration among researchers.

Despite these challenges, many efforts have been made to build and disseminate large-scale annotated corpora for different languages and accents. For instance, the LibriSpeech[4] corpus provides a significant resource with around 1,000 hours of English speech data derived from audiobooks in the LibriVox project. LibriSpeech is freely available for academic and commercial use under the Creative Commons BY 4.0 license, making it a valuable resource for researchers and developers alike. Another important dataset is the MuST-C (Multilingual Speech Translation Corpus) [5], which includes around 385 hours of speech from English TED Talks, manually transcribed and translated into eight languages: German, Spanish, French, Italian, Dutch, Portuguese, Romanian, and Russian. The MuST-C corpus is available for research purposes under a Creative Commons BY-NC-ND 4.0 license, allowing for non-commercial use and dissemination with attribution.

The CoVoST (Common Voice Speech Translation) [6] corpus is another diverse multilingual dataset developed to support speech-to-text translation tasks. The initial release, CoVoST 1, includes over 60 hours of speech data for 15 languages translated into English. The extended version, CoVoST 2, offers translations from English into 15 languages and from 21 languages into English, with over 2,900 hours of speech data in total. CoVoST is part of the Common Voice project by Mozilla, and the data is available under the Creative Commons Zero (CC0) license, which allows for unrestricted use. The TED-LIUM corpus, based on TED Talks, is widely used for both speech recognition and translation tasks. The latest release, TED-LIUM 3, includes approximately 452 hours of English speech, with corresponding transcripts and speaker information. This corpus provides a rich source of data for training models in varied and natural-speaking conditions. It is freely available for research purposes under a permissive license that allows redistribution and modification.

The AMI Meeting Corpus[7] focuses on multi-party meeting recordings and is valuable for research on conversational speech recognition and diacritization. It includes around 100 hours of meeting recordings in English, with detailed annotations of speech, speaker turns, and transcriptions. The AMI corpus is available for research purposes under a non-commercial license, promoting its use in academic settings. Additionally, the FLEUR (Few-shot Learning Evaluation for Under-Resourced Languages) corpus[8] is a significant resource for evaluating the performance of models on under-resourced languages. Covering 102 languages, FLEUR provides few-shot learning benchmarks.

It includes a diverse array of languages with varying amounts of annotated data, making it an important resource for developing and testing models intended for languages with limited available resources. The corpus is freely available for research purposes, facilitating advancements in natural language processing for under-resourced languages. These corpora, although not without limitations, represent substantial steps forward in enhancing the accuracy and robustness of spoken language recognition and translation systems.

Compared to major international languages, under-resourced dialects suffer from a lack of available datasets for speech recognition and speech-to-text translation. This scarcity is primarily due to limited funding and resources for collecting and annotating such data. When datasets for these dialects do exist, they tend to be significantly smaller than those available for more widely spoken languages, which hampers the development of robust ASR and S2T systems. Major corpora like MuST-C and CoVoST offer extensive data for numerous languages, but equivalent resources for under-resourced dialects are often insufficient in size and scope. Moroccan dialectal Arabic, commonly known as Darija, exemplifies this issue. Spoken by over 40 million people in Morocco and used in daily conversations and activities [9-11], Moroccan Darija remains under-resourced[12].

The development of speech recognition and S2T systems for Darija is challenging due to the limited number and size of available datasets. One of the earliest efforts to create a Moroccan Darija speech dataset was by Bezoui et al. [13], who recorded words spoken by 20 speakers. This initial attempt provided a foundational dataset but was limited in scope. Hassine et al. [14] expanded on this by developing a corpus with 40 pronunciations of digits from 0 to 9, recorded from four speakers. These efforts, while significant, highlighted the need for larger, more diverse datasets.

Moroccan Darija has also been included in multi-dialect corpora. Belgacem [15] developed a corpus covering multiple Arabic dialects, including Moroccan Darija. Similarly, Amazouz et al. [16] introduced a dataset that includes about 15 hours of Moroccan Darija speech, adding valuable data to the resource pool. The inclusion of Moroccan Darija in the fifth edition of the Multi-Genre Broadcast Challenge (MGB-5) [17] further expanded the available data. This dataset

features 13 hours of speech transcribed using Arabic alphabets, extracted from 93 YouTube videos spanning seven genres, including TEDx talks.

Despite these contributions, the availability of Moroccan Darija-specific datasets remains limited. The Dvoice dataset, created by SI2M labs in collaboration with AIOX LAB[18], is currently the only freely available speech dataset dedicated to Moroccan Darija. It comprises 2992 audio files and their transcriptions, split into 2392 training files and 600 testing files, primarily intended for speech recognition tasks. This dataset represents a critical resource, yet the demand for more comprehensive and varied data persists. The scarcity of extensive datasets for Moroccan Darija underscores the urgent need for additional resources. Efforts to develop such datasets must address the linguistic diversity and specific characteristics of Darija to build effective ASR and S2T systems. By expanding the size and variety of available datasets, researchers can enhance the performance and generalizability of these systems, ultimately improving communication and information access for speakers of Moroccan Darija.

While several Moroccan Darija speech datasets have been created for dialect identification or automatic speech recognition tasks, to our knowledge, no dedicated speech-to-text translation dataset for Darija has been developed yet. This gap presents a significant challenge and an opportunity for advancing research in this area. Existing datasets, such as those developed by Bezoui et al. [13], Hassine et al. [14] and included in multi-dialect corpora like those by Belgacem [15] and Amazouz et al. [16], have primarily focused on speech recognition or dialect identification rather than translation tasks. These datasets, while valuable, do not address the need for comprehensive resources that facilitate the development of speech-to-text translation systems. To bridge this gap, this paper introduces the creation of a new dataset specifically designed for both speech recognition and speech-to-text translation of Moroccan Darija. This dataset, named DARIJA-C, will be freely available for researchers, providing an essential resource for advancing natural language processing technologies for Moroccan Darija. By making this dataset accessible, which support and stimulate further research and development in this under-resourced dialect.

The work of Kocabiyikoglu et al. inspires this methodology for collecting the DARIJA-C corpus [4], which utilized automated alignment techniques to match transcripts with text translations and subsequently with their corresponding audio segments. This approach ensures accurate and efficient pairing of speech data with textual translations, enhancing the quality and usability of the dataset. This work leverages a similar technique to align Moroccan Darija speech recordings with their textual translations into Modern Standard Arabic (MSA), creating a robust and reliable corpus for both ASR and S2T applications.

3. Data Collection Platform

To create the DARIJA-C corpus, the DARIJA-C Web platform was established, a collaborative and community-driven initiative aimed at collecting data for research on Moroccan Darija speech recognition technologies. This innovative platform empowers users to actively participate by contributing audio recordings of spoken Moroccan Darija. These recordings are essential for training and refining speech recognition systems, ensuring they accurately understand and process the nuances of this unique dialect. By harnessing the collective efforts of the community, the platform not only fosters a rich repository of linguistic data but also advances the field of speech recognition technology, making it more inclusive and representative of Moroccan Darija speakers. Through continuous user engagement and data collection, the DARIJA-C Web platform plays a pivotal role in enhancing the accuracy and effectiveness of speech recognition systems tailored to this specific language variant.

The dataset collection process on the DARIJA-C Web platform follows a structured approach to ensure the quality and reliability of the contributions. The process includes the following steps:

- User Registration: Individuals interested in contributing to the dataset must first create an account on the platform. This step helps to verify that genuine users rather than automated systems are making the recordings, thus maintaining the integrity of the data.
- Recording Submission: Once registered, users can submit audio recordings of themselves speaking Moroccan Darija. The platform offers an integrated recording tool that operates directly within the browser, simplifying the contribution process and encouraging user participation.
- Recording Review: Submitted recordings undergo a review process where evaluators assess their quality. Recordings that do not meet the established standards, whether due to poor audio quality or failure to meet specific criteria, are rejected. This ensures that only high-quality recordings are included in the dataset.
- Translation Submission: Users are also encouraged to submit textual translations of displayed Moroccan Darija utterances. This additional step enriches the dataset by providing paired audio and textual data.
- Translation Review: Similar to the recording review process, submitted translations are evaluated by reviewers to ensure accuracy and adherence to translation standards. Incorrect or substandard translations are rejected to maintain the dataset's reliability.
- Corpus Creation: Once a recording and its corresponding translation are approved, they are added to the corpus. The resulting corpus comprises a substantial collection of translated recordings, each annotated with metadata. This comprehensive dataset aids researchers in analyzing and understanding the nuances of Moroccan Darija, thereby advancing speech recognition technologies tailored to this dialect.

The DARIJA-C Web platform incorporates several quality control measures to ensure the accuracy and consistency of the dataset. These measures include:

- Random Sampling: The platform periodically displays random samples for manual review by users. This ongoing process helps maintain the consistency and accuracy of the dataset over time, as it ensures continuous oversight and verification.
- Consistency Checks: To detect inconsistencies in recordings, the platform allows users to listen to the recordings and identify any discrepancies between the audio, transcriptions, or associated metadata. This feature ensures that all elements of the dataset are in alignment, thereby enhancing its reliability.
- Regular Updates: The platform is consistently updated with new recordings and metadata. These regular updates ensure that the dataset remains current and relevant, providing researchers with up-to-date data for their work on Moroccan Darija speech recognition technologies.

The DARIJA-C Web platform stands as a valuable resource for researchers focused on developing Moroccan Darija speech recognition technologies. By harnessing community contributions and enforcing strict quality control measures, the platform successfully creates a high-quality corpus. This corpus is crucial for training and improving speech recognition systems, making them more effective and accurate in understanding Moroccan Darija.

4. Methodology of Collecting Darija-C

This section explores in detail the procedures employed for collecting the Moroccan Darija Speech Corpus through the DARIJA-C Web platform. Each step of the data collection process is examined, from user registration and recording submission to the evaluation and approval of contributions. Additionally, the stringent quality control measures implemented to ensure the accuracy and consistency of the collected data are discussed. These measures include random sampling for manual review, consistency checks for detecting discrepancies, and regular updates to keep the dataset current. Highlighting these procedures and quality control practices provides a comprehensive understanding of how the DARIJA-C Web platform maintains the integrity and reliability of the Moroccan Darija Speech Corpus.

4.1. Corpus Collection Method

To collect inputs for the dataset, a variety of sources were leveraged, including Darija learning websites and transcriptions of YouTube videos in Darija. The data collection strategy encompassed three main options: Arabic Darija text, Arabic Darija text paired with its Classical Arabic translation, and Darija audio accompanied by an Arabic Darija transcript. These diverse resources formed the foundation for the DARIJA-C Web platform, designed to facilitate the continuous expansion of the corpus. The DARIJA-C Web platform enables users to actively contribute to the dataset by recording their speech corresponding to the displayed Darija texts. Each user-submitted recording is stored as a single-channel audio file with a sampling rate of 16kHz, ensuring high-quality audio data. Furthermore, the platform offers the option to provide an Arabic translation of the displayed Darija text, which enhances the dataset by supporting speech-to-text translation research.

4.2. Data Validation

Each item in the dataset is composed of a meticulously annotated Darija sentence, its corresponding Arabic translation, and a variety of metadata. This metadata includes the audio file path, sampling rate, duration, and both the utterance recording and translation up and down votes. These elements provide comprehensive information about each entry, facilitating detailed analysis and research. To ensure the high quality of the recordings, a robust validation process is implemented on the DARIJA-C platform. This process involves peer evaluation, where other users review contributors' entries. Specifically, the platform displays the written utterance and requests reviewers to play the associated audio. Reviewers then vote up if the audio matches the displayed text or vote down if it does not. This community-driven approach helps maintain the integrity of the dataset by ensuring that only accurate recordings are included. Similarly, to verify the accuracy of translations, the platform displays the Darija text alongside its Arabic translation. Users are asked to confirm or disprove the provided translation, casting their votes accordingly. This method ensures that translations are accurate and consistent with the original Darija text.

By incorporating these validation steps, the DARIJA-C platform not only enhances the quality of the dataset but also builds a more reliable resource for advancing speech recognition technology. The collective effort of the community in vetting both audio recordings and translations significantly contributes to the overall improvement of speech recognition systems tailored for Moroccan Darija. This rigorous validation process ensures that the DARIJA-C corpus remains a valuable and trustworthy resource for researchers and developers working on speech-to-text translation and other related technologies.

4.3. Implementation

It is crucial for web platforms to be transparent about their data collection practices and to obtain explicit consent from users before gathering any data. In developing the DARIJA-C platform, user transparency and ethical data collection are prioritized. This ensures that users are fully informed about how their data will be used and obtain their explicit consent before any data collection begins. Given that the collected data will be freely available on the DARIJA-C platform as a corpus, it is paramount to collect and distribute this data ethically while safeguarding user privacy. To this end, the DARIJA-C platform has implemented a robust consent mechanism. Before any data can be collected, users must

consent to granting the platform access to their microphones. Upon first entering the web application (Figure 1), users are presented with a comprehensive page that explains the purpose of the platform and details the data collection and usage practices.

This page also includes user agreements, which users must accept if they wish to contribute to the DARIJA-C dataset. Access to the main features of the web application is restricted to those who have agreed to the terms outlined in the agreement statements. By implementing these measures, the DARIJA-C platform not only adheres to ethical standards but also ensures user trust and compliance with data protection regulations. This approach fosters a secure and transparent environment for data collection, encouraging user participation while maintaining the integrity and privacy of their contributions. It is exciting that the DARIJA-C platform is now available to the community at http://darija-c.com, allowing users to actively contribute to and benefit from this valuable resource.

When contributors access the "Record" page on the DARIJA-C web platform, they are presented with a randomly displayed utterance and a button labeled "Start Record." This setup is designed to be intuitive and user-friendly, ensuring that contributors can easily understand and engage with the recording process. Contributors are prompted to speak aloud the text corresponding to the displayed utterance. Upon finishing their recording, they press the "Stop Record" button, which saves the recorded speech as an audio file sampled at 16KHz. To maintain user engagement and streamline the data collection process, once the recording is saved, the platform will randomly select another utterance for the contributor to record (Figure 2).

This allows contributors to seamlessly continue contributing without needing to navigate away from the page, thereby maximizing the efficiency and volume of data collection. This process, which is clearly explained to users when they first access the platform and agree to the user terms, ensures that the DARIJA-C platform gathers a diverse range of high-quality recordings. By maintaining transparency and obtaining explicit user consent, the data collection is both ethical and efficient, fostering a secure environment for contributors while continuously expanding the corpus size.

The DARIJA-C platform features a comprehensive "Listen" page dedicated to the validation of recordings. On this page, users are invited to listen to the recorded speech of other contributors and determine if it matches the original displayed utterance. To indicate a match, users click the "Votes Up" button; if the recorded speech does not correspond to the displayed text, they click the "Votes Down" button. This peer review process is crucial for maintaining the accuracy and quality of the dataset, ensuring that only correctly matched recordings are retained (as illustrated in Figure 3).

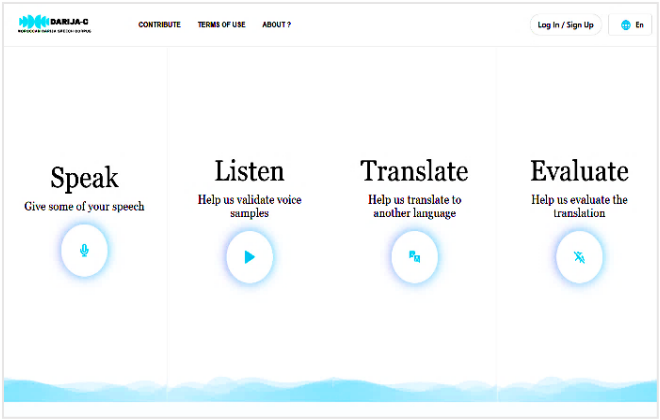

Fig. 1 DARIJA-C web plateform

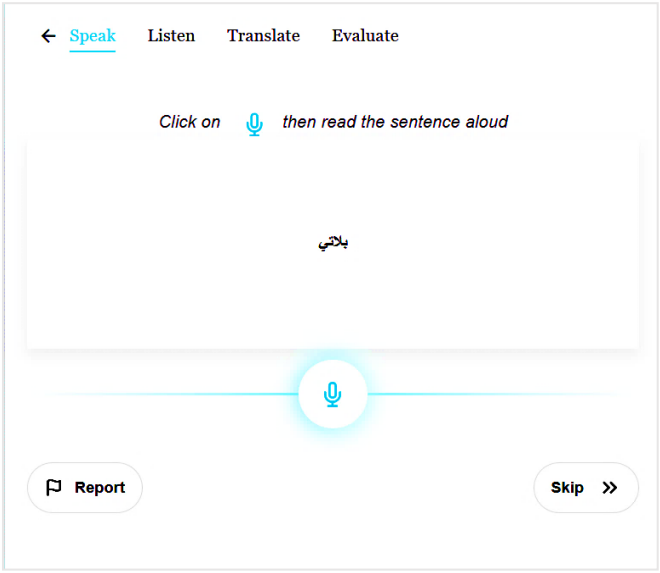

Fig. 2 DARIJA-C Speech recording interface

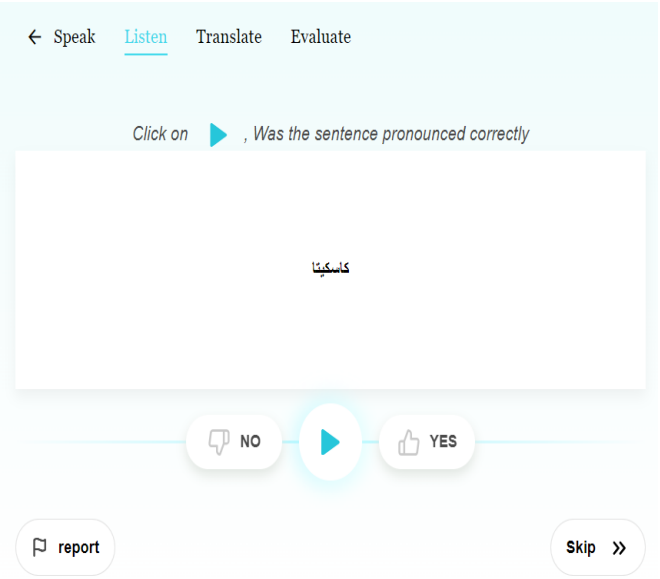

Fig. 3 DARIJA-C speech validation interface

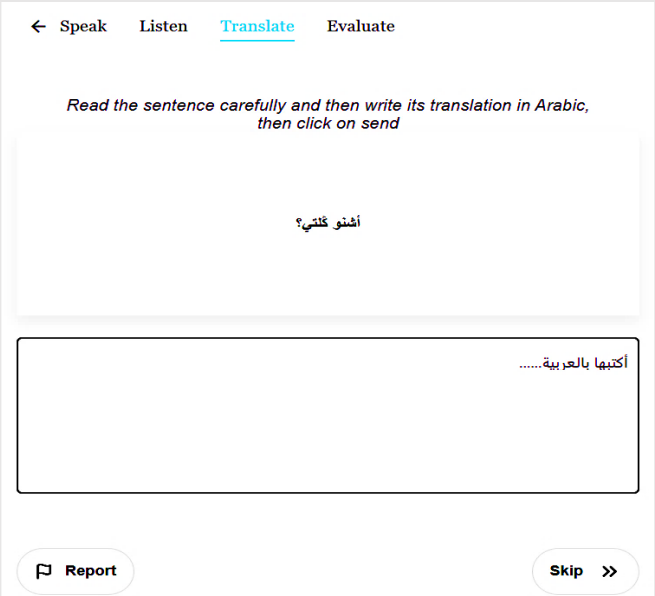

Fig. 4 DARIJA-C utterance translation interface

Fig. 5 DARIJA-C translation validation interface

The translation section of the DARIJA-C platform involves a structured process for contributing textual translations. Contributors access the "Translation" page, where they are presented with a randomly selected Darija utterance. Using only the Arabic alphabet, they write the translation of the given utterance. Once the translation is complete, contributors click the "Confirm Translation" button, which prompts the system to update and store the translation data. This methodical approach ensures that each Darija utterance is accurately translated, enriching the corpus with high-quality text data (as shown in Figure 4).The final component of the DARIJA-C platform is the "Evaluate Translation" page. Here, contributors review the translations submitted by others. They are required to compare the Darija utterance with its provided translation and then vote on its accuracy. If the translation is correct, they click the "Votes Up" button; if it is incorrect, they select the "Votes Down" button. This validation process is essential for maintaining the reliability and precision of the translation data within the corpus (as shown in Figure 5).

4.4. Corpus Quality

Given the variability in the quality of the audio clips collected, it establishes specific criteria to determine which data would be transferred to the official corpus. The informal guideline was to reject any audio files if listeners could not clearly distinguish the word spoken or if the word was incorrect based on the number of upvotes and downvotes received. Additionally, audio files that were either too quiet or too short had to be filtered out. Compressed audio files with minimal speech content would be small in size. Therefore, any file with a size less than 6 KB is most likely inaccurate and subsequently rejected. This size threshold helped us eliminate low-quality recordings and maintain a high standard for the dataset. In parallel with audio quality control, stringent criteria for managing translation entries on the DARIJA-C platform are also established, which assumes that a single-character translation was unlikely to be accurate and would, therefore, be rejected. To ensure consistency and accuracy, contributors must use only the Arabic alphabet when translating into Arabic. This requirement helped maintain uniformity in translation quality and prevented errors arising from the use of non-Arabic characters. By implementing these quality control measures, the DARIJA-C corpus remains a reliable and high-quality resource. These efforts not only enhance the accuracy of the speech and translation data but also support the development of robust speech recognition technologies tailored to Moroccan Darija. This meticulous approach underscores the researcher's commitment to creating a comprehensive and dependable dataset for researchers and developers in the field.

5. Collected Corpus Statistics and Characteristics

The following section provides an in-depth look at the size, structure, and quality of the corpus, which was collected over one year following the launch of the DARIJA-C platform.

5.1. Corpus Size and Scope

The DARIJA-C corpus consists of 50 hours (Figure 6) of recorded speech, amounting to a substantial resource for training and evaluating ASR and S2T models. This corpus is large enough to provide the necessary data for building machine learning models that require vast and varied linguistic input. With 18,000 individual recordings, the dataset ensures comprehensive coverage of common speech patterns, vocabulary, and sentence structures used in everyday conversations.

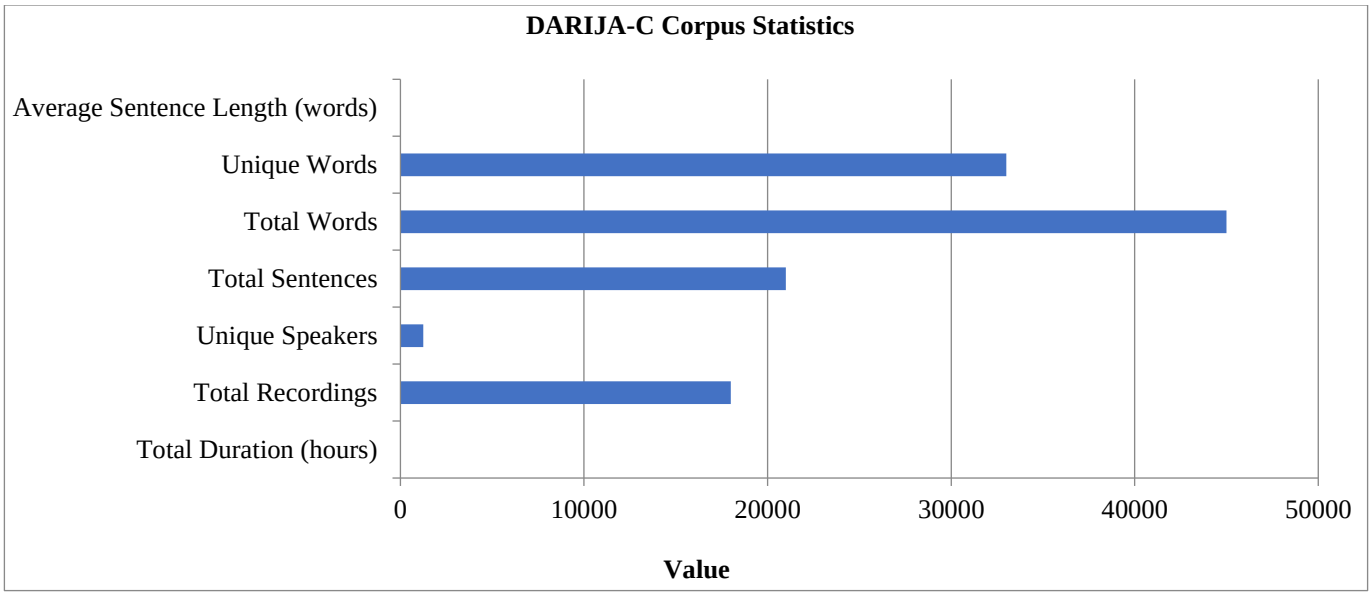

Fig. 6 Darija-C statistics

5.2. Speaker Demographics

A critical component of any speech corpus is the diversity of the speakers involved. The DARIJA-C corpus boasts contributions from 1,250 unique speakers, encompassing a wide range of accents, regional dialects, and individual speech variations. This diversity is essential for creating models that generalize well across different speakers and are not biased toward specific accents or demographics. Future analyses could explore the geographical distribution of speakers, gender balance, and socio-economic background to assess the corpus's representativeness further.

5.3. Sentence and Word Counts

The dataset includes 21,000 sentences, providing a rich source of structured linguistic data. These sentences span various contexts, offering both formal and informal speech patterns that reflect the natural flow of conversation in Moroccan Darija. With a total word count of 45,000, the corpus supports the training of models that can accurately recognize and translate everyday speech into Modern Standard Arabic (MSA) or other languages. The average sentence length is 4 words, which is typical for conversational speech, where short, fragmented sentences are common. Despite this, the dataset still captures a wide range of linguistic constructs, from simple phrases to more complex, multi-clause sentences. Additionally, the corpus exhibits a high degree of lexical diversity, with 33,000 unique words. This richness in vocabulary ensures that models trained on the corpus can handle a wide array of linguistic phenomena, from regional slang to formal expressions.

5.4. Speech Quality

Maintaining high-quality audio recordings is crucial for accurate ASR and S2T model development. The recordings in the DARIJA-C corpus were collected at a 16 kHz sampling rate, which is the standard for most modern speech recognition tasks. This ensures that the dataset captures sufficient detail in the audio signals to allow models to distinguish between subtle variations in pronunciation, tone, and intonation. The high-quality audio helps minimise errors in transcription and translation, especially for a dialect like Moroccan Darija, which is characterized by regional and social variations. The quality of the recordings also supports future applications in areas such as speaker identification and dialect classification, where fine-grained differences in speech need to be detected.

5.5. Applications and Potential Impact

The richness and diversity of the DARIJA-C corpus make it an invaluable resource for a variety of NLP applications. In addition to ASR and S2T, the dataset can be used for tasks such as sentiment analysis, topic modeling, and dialogue systems that understand and generate natural Darija speech. Furthermore, the corpus can serve as a benchmark for future research, enabling researchers to compare the performance of different models and approaches in handling Moroccan Darija.

6. Discussion

Building a speech corpus for Moroccan dialectal Arabic, known as Darija, is a challenging task due to the distinct characteristics and variations within the dialect. However, the aim is to overcome these challenges through the use of a collaborative and automated platform, the DARIJA-C Web platform. This innovative platform allows us to gather a large collection of Moroccan Darija speech along with corresponding translations, facilitating the development of advanced linguistic technologies for this dialect. These technologies can be applied to various fields, including speech recognition, machine translation, and natural language processing.

The DARIJA-C platform stores speech recordings as single-channel WAV files with a sampling rate of 16 kHz, ensuring high-quality audio data. To maintain the accuracy and reliability of the dataset, a robust validation system based on community feedback is implemented. Each audio file submitted to the platform undergoes a voting process where contributors rate the recordings. An audio file requires a minimum of three "Up Votes" to be marked as valid and included in the dataset. Conversely, if a recording receives three "Down Votes," it is considered invalid and is excluded from further validation. This same voting mechanism is applied to the translation validation process to ensure the accuracy of the textual data. To enhance the integrity of the validation process, the same contributor cannot validate the same input multiple times and cannot validate their submissions.

This prevents potential biases and maintains the quality and objectivity of the dataset. Each submission to the DARIJA-C platform must receive at least three "Up Votes" for both the audio file and its corresponding transcription and translation to be considered for inclusion in the official DARIJA-C corpus. If a submission receives three "Down Votes," it is reset, allowing other contributors to provide new data. This iterative process ensures that only high-quality data is included in the corpus, making it a valuable resource for researchers and developers. The DARIJA-C platform has successfully collected a significant amount of speech and text data, which will be freely available under a Creative Commons CC0 license. This makes the DARIJA-C corpus the largest public domain resource for Moroccan Darija, supporting research and development in automatic speech-to-text translation and automatic speech recognition. Once the first version of the DARIJA-C corpus is ready for use, it will be available for free download on the DARIJA-C platform, providing an invaluable tool for the advancement of linguistic technologies in the Moroccan Darija dialect.

By leveraging the collaborative efforts of contributors and implementing stringent quality control measures, the DARIJA-C platform aims to build a comprehensive and high-quality corpus. This resource will not only support the development of advanced linguistic technologies for Moroccan Darija but also contribute to the broader field of computational linguistics and natural language processing. The DARIJA-C platform is paving the way for more accurate and effective speech recognition systems that can understand and process the unique characteristics of Moroccan Darija, thereby enhancing communication and accessibility for Darija speakers worldwide. The large scale of the DARIJA-C corpus collected through the DARIJA-C platform, with over 50 hours of speech and contributions from 1,250 speakers, ensures that the trained models will be capable of handling a wide variety of accents and speech patterns in Moroccan Darija. The inclusion of speakers from diverse backgrounds helps the model generalize across different dialectical and social variations, increasing the model's applicability in real-world settings. The corpus's lexical diversity, with over 33,000 unique words, reflects the rich vocabulary present in Moroccan Darija. This variability presents challenges for automatic speech recognition, particularly when dealing with informal speech and regional slang. However, the richness of the vocabulary also provides an opportunity for models trained on this dataset to understand better and process the nuances of everyday Darija speech.

The average sentence length of 4 words is typical of spoken language, where utterances tend to be brief and conversational, adding further complexity to the task of speech recognition. The recordings in the DARIJA-C corpus are captured at a 16 kHz sampling rate, providing sufficient detail for ASR systems to distinguish phonetic and prosodic features accurately. The rigorous validation process ensures that only high-quality recordings are included, further minimizing errors caused by background noise or poor audio quality. This level of quality control is crucial for ensuring that the trained models can perform well even in challenging environments.

The combination of speaker diversity and lexical richness in the DARIJA-C corpus makes it a valuable resource not only for ASR but also for other applications such as speaker identification and dialect classification. The dataset's breadth ensures that it can be applied in a wide range of contexts, from conversational AI to educational tools, and it sets a strong foundation for future research into speech technologies for under-resourced dialects. While the DARIJA-C corpus provides a comprehensive foundation, future iterations could focus on expanding the range of sentence structures and increasing the representation of more complex speech patterns. Additionally, further diversification of speakers across different regions and socio-economic backgrounds could enhance the corpus's generalizability, allowing models to perform more effectively across various demographic groups.

7. Conclusion and Future Work

The development of the DARIJA-C corpus marks a significant milestone in addressing the challenges of Moroccan Darija speech recognition and translation. With its diverse dataset, including 50 hours of speech from 1,250 speakers, the corpus supports robust ASR and S2T models capable of handling the linguistic complexity of this dialect.

The community-driven approach ensures data accuracy and quality, fostering a reliable resource for future research. By making this corpus publicly available, it opens up new possibilities for technological advancements in under-resourced languages and contributes to the wider field of NLP. Further expansion of the corpus will enhance its utility, driving innovation in speech technologies for Moroccan Darija speakers.

References

[1] Ahmed Ali, Stephan Vogel, and Steve Renals, "Speech Recognition Challenge in the Wild: Arabic MGB-3," *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Okinawa, Japan, pp. 316-322, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[2] Mohamed G. Elfeky, Pedro Moreno, and Victor Soto, "Multi-Dialectical Languages Effect on Speech Recognition: Too Much Choice Can Hurt," *Procedia Computer Science*, vol. 128, pp. 1-8, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[3] Maria Labied, and Abdessamad Belangour, "Moroccan Dialect "Darija" Automatic Speech Recognition: A Survey," *2021 IEEE 2nd International Conference on Pattern Recognition and Machine Learning (PRML)*, Chengdu, China, pp. 208-213, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[4] Ali Can Kocabiyikoglu, Laurent Besacier, and Olivier Kraif, "Augmenting Librispeech with French Translations: A Multimodal Corpus for Direct Speech Translation Evaluation," *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, pp. 1-5, 2018. [Google Scholar] [Publisher Link]

[5] Roldano Cattoni et al., "MuST-C: A Multilingual Corpus for End-to-End Speech Translation," *Computer Speech & Language*, vol. 66, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[6] Changhan Wang et al., "CoVoST: A Diverse Multilingual Speech-To-Text Translation Corpus," *Proceedings of the Twelfth Language Resources and Evaluation Conference*, Marseille, France, pp. 4197-4203, 2020. [Google Scholar] [Publisher Link]

[7] Jean Carletta et al., "The AMI Meeting Corpus: A Pre-Announcement," *Second International Workshop: Machine Learning for Multimodal Interaction*, Edinburgh, UK, pp. 28-39, 2005. [CrossRef] [Google Scholar] [Publisher Link]

[8] Alexis Conneau et al., "FLEURS: Few-Shot Learning Evaluation of Universal Representations of Speech," *2022 IEEE Spoken Language Technology Workshop (SLT)*, Doha, Qatar, pp. 798-805, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[9] Nizar Y. Habash, *Introduction to Arabic Natural Language Processing*, 1st ed., Synthesis Lectures on Human Language Technologies, Springer Cham, pp. 1-187, 2010. [CrossRef] [Google Scholar] [Publisher Link]

[10] M. Amour, A. Bouhjar, and F. Boukhris, "Introduction to Amazigh Language," *Paris: IRCAM*, 2004. [Google Scholar]

[11] Fatima Sadiqi, *Women, Gender, and Language in Morocco*, Brill, pp. 1-336, 2003. [Google Scholar] [Publisher Link]

[12] Rabih Zbib et al., "Machine Translation of Arabic Dialects," *2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Montreal, Canada, pp. 49-59, 2012. [Google Scholar] [Publisher Link]

[13] Bezoui Mouaz, Beni Hssane Abderrahim, and Elmoutaouakkil Abdelmajid, "Speech Recognition of Moroccan Dialect Using Hidden Markov Models," *Procedia Computer Science*, vol. 151, pp. 985-991, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[14] Mohamed Hassine, Lotfi Boussaid, and Hassani Messaoud, "Maghrebian Dialect Recognition Based on Support Vector Machines and Neural Network Classifiers," *International Journal of Speech Technology*, vol. 19, pp. 687-695, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[15] Mohamed Belgacem, "Construction of a Robust Corpus of Different Arabic Dialects," *Proceedings of the 8th Young Researchers in Speech Meeting*, vol. 33, 2009. [Google Scholar]

[16] Djegdjiga Amazouz, Martine Adda-Decker, and Lori Lamel, "Addressing Code-Switching in French/Algerian Arabic Speech," *Interspeech 2017*, pp. 62-66, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[17] Ahmed Ali et al., "The MGB-5 Challenge: Recognition and Dialect Identification of Dialectal Arabic Speech," *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Singapore, pp. 1026-1033, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[18] Anass Allak et al., "Dialectal Voice : An Open-Source Voice Dataset and Automatic Speech Recognition Model for Moroccan Arabic Dialect," *NeurIPS Data-Centric AI Workshop*, 2021. [Google Scholar] [Publisher Link]