Revue d'Intelligence Artificielle
Vol. 38, No. 3, June, 2024, pp. 857-866
Journal homepage: http://iieta.org/journals/ria

Building a Corpus for the Underexplored Moroccan Dialect (CFMD) Through Audio Segmentations

Hajar Zaidani[1*], Abderrahim Maizate[1], Mohammed Ouzzif[1], Rim Koulali[2]

[1] RITM Laboratory, Higher School of Technology Casablanca, CED ENSEM, Hassan II University Casablanca, Casablanca 20430, Morocco
[2] LIS Labs, Faculty of Sciences Ain Chock, Hassan II University, Casablanca 20100, Morocco

Corresponding Author Email: hajar.zaidani-etu@etu.univh2c.ma

https://doi.org/10.18280/ria.380312

Received: 30 October 2023
Revised: 1 February 2024
Accepted: 12 March 2024
Available online: 21 June 2024

Keywords:
artificial intelligence, natural language processing, data collection, Moroccan dialect

ABSTRACT

The advancement of artificial intelligence has deeply influenced numerous domains. One particular area that has experienced remarkable progress is natural language processing. This progress can be largely attributed to the widespread use and popularity of social media platforms. With the increasing use of social media, dialects have taken on a new importance, as the diversity of dialects has an important role to consider in the relevance of Natural Language Processing, as it allows a greater number of people to communicate using a pertinent and appropriate local context. As evidenced by the rise of Chatbots that allow people to interact with machines using their own native dialects. The significance of dialects, especially in the Arabic-speaking world, cannot be understated. Many Arabic dialects have been under-researched and not adequately addressed in natural language processing applications. Among these, the Moroccan dialect stands out, prompting researchers to focus their efforts on understanding and incorporating it into artificial intelligence technologies. To facilitate the development of Chatbots that can effectively understand and respond in Moroccan dialect, the availability of suitable datasets becomes vital. For this reason, we adopt a targeted strategy for creating datasets by exploiting the extensive resources offered by platforms such as YouTube, where audio content is highly diverse in terms of language. This involves classifying each audio according to its theme and dividing it into 30 second segments to simplify manual transcription into text. This meticulous process enabled us to accumulate and annotate a large volume of data. As a result, NLP models built on these extensive and comprehensive datasets can efficiently and accurately understand Moroccan dialect speech and text. With the aim to employ this dataset as training data for the future development of a Moroccan-dialect conversational Chatbot. The methodologies and techniques can be adapted and applied to other underexplored dialects, creating opportunities for further advancements in natural language processing in a global context.

1. INTRODUCTION

Artificial intelligence is becoming increasingly powerful these days. It is being integrated into our daily lives in many disciplines, including engineering, medicine, education and economics. There are many instances of conversational artificial intelligence in our daily life. Dialogue systems, conversational agents and personal assistants are systems designed to converse with humans using speech, gestures, graphics and other methods of interaction [1]. Furthermore, a chatbot is a computer programme that reacts like an intelligent entity when spoken to through text or voice, and which understands one or more human languages thanks to natural language processing (NLP) [2]. The chatbot is one of the most elementary and widespread forms of intelligent human-machine interaction. it plays a critical function in many real-life applications, like smart speakers, customer service systems, etc [3]. In general, Conversational AI is a constantly emerging discipline that has attracted the research efforts of natural language processors and companies such as Google, Amazon, Facebook, etc., who have developed speech and language technologies and are currently developing text and voice dialogue systems [4]. They have become more powerful nowadays thanks to developments in Natural Language Processing (NLP) as well as other fields of artificial intelligence. Not only are these developments gaining ground in Latin languages, but they have also extended to Arabic. However, Arabic dialects are not exploited as much due to the lack of corpora in these languages. This is why Arabic dialects Chatbots are not yet available.

Clearly, dealing with dialects is more complicated than Modern Standard Arabic (MSA) in Natural Language Processing for many reasons. First, Arabic dialects are spoken languages, so we do not need to write them down. Second, they differ from country to country. Third, each dialect is further divided into sub-dialects. Therefore, we need to pay careful attention when choosing a specific dialect to work on

[5]. On the other hand, Arabic dialects have been little published, and suffer from a lack of NLP datasets. However, the increased use of the Internet and social media. The Moroccan dialect, for example, faces challenges due to the unavailability of corpora, which poses a great challenge in the use of this dialect [6].

Research into understudied dialects in developing countries is crucial to maintaining cultural and linguistic diversity and preventing the extinction of these languages. Whereas most research concentrates on the more formal, wealthier languages. The rarity of research into understudied dialects and the challenges they present are often due to the predominance of mainstream languages and the limited means to collect resources, even when they are available. Language is the main medium for preserving a country's heritage; it is the main reference for traditional knowledge, values, practices, and history over the centuries.

Indeed, there is current work on the Moroccan dialect. Firstly, AIOX LAB worked on an Open-Source Voice Dataset but only on single words [7]. Secondly, Issam and Mrini [8] used the model developed by AIOX Labs Darijabert and other models to summarize a long text. Then, the creation of corpora for sentiment analysis by exploiting social media comments [9]. Unfortunately, all this research is not suited to our research objective.

With a view to harnessing the power of the neural networks, we aim to generate firstly a Moroccan dialect dataset that could help a ChatBot learn Moroccan communication. To strengthen this content, we seek to create an open dataset, as a solution to increase the innovation and application of NLP for the Moroccan dialect. Written content in Moroccan dialect has grown significantly on the internet in recent years thanks to by social media. Therefore, social media has become the appropriate source for collecting data [10, 11]. While we were unable to collect audio data exclusively from the YouTube platform, we will have the opportunity to use this data in the future to train a chatbot capable of understanding speech.

To accomplish this objective, our study involved a series of methodical steps to transcribe the Moroccan dialect videos from the YouTube platform into text. Initially, we chose the appropriate videos with clear audio and general Moroccan dialect. Subsequently, we divided each recording into small recordings (30 seconds). Then, we wrote the content manually to make sure that our dataset included correct words in Moroccan dialect. our study distinguishes itself as the pioneering effort in creating a dataset that combines transcripts of text with corresponding audio recordings, using the YouTube platform to extract recordings and transcribe them manually with the assistance of different editors for quality assurance.

The rest of this article is organized as follows: Section 2 describes related works. Section 3 includes an overview of the Moroccan Dialect at the phonological, morphological, syntactic and lexical levels. Section 4 explains the main challenges concerning the Moroccan Dialect, such as morphological, semantic and syntactic ambiguities Section 5 outlines in detail the methodology of creating the corpus. Results and discussion are presented in Section 6. Finally, we provide a conclusion and lay the ground for future research in Section 7.

2. RELATED WORK

Over the past several years, interest in Arabic NLP resources has grown, significantly, methods to collect corpora are also diversified. The MADAR Arabic Dialect Corpus, has adopted the Machine translation approach for the corpus creation, whose domain is Travel and Tourism. This method focuses on converting text from a source language into a target language, while retaining the same meaning, by choosing items from the Basic Travel Expression Corpus (BTEC), the English phrases are translated into French, Modern Standard Arabic and dialectical Arabic of five regional area. MADAR could reach over 62K sentences at size, with a public accessibility to corpus and private reuse [11]. The authors of Darija Open Dataset (DODA) adopted the same method of MADAR, unfortunately, they use arabizi form to build the corpus, with multiple categories, such as: Food, Animals, Health...etc. Corpus could achieve more than +10K entries, which have a public access and reuse [12]. MDED, a bilingual dictionary, created using the same method as the previous ones, but without a specific theme, with private rights of access and re-use, it contains 18,000 entries mainly constructed by manual translation of an MSA dictionary and a Moroccan dialect dictionary [13].

Nowadays, Moroccan dialect is attracting more and more attention due to the massive use and popularity of social media, blogs, … etc. Even though there are not many research studies that have covered Darija [14]. GOUD.MA is a dataset of news articles, gathered using manual data collection, based on the extraction of data from a news website, it should be noted that this corpus includes 158,000 news articles of various categories, which are accessible to the public and available for re-use [8]. In particular, Dvoice adopts the same method as GOUD.MA but adds voice as an extra option. It is a web application that allows users to contribute their own voice by submitting recordings that correspond to texts written in Darija or to approve the recordings of other users. Furthermore, Dvoice gives subscribers public access to data and its reuse, with a file size of 2,990, in various categories [7].

Our work is also based on the same approach as that adopted by GOUD.MA and Dvoice. We have chosen to collect data manually from the web, but on specific themes, which the other datasets do not contain. In addition, we have been trying a new method of converting audio to text that will enable us to collect both voice and text data. This corpus therefore contains sentences from everyday communication, whose meaning is comprehensible, as well as textual data and voice recordings on different themes.

3. MOROCCAN DIALECT

3.1 Overview

Arabic is one of the oldest Semitic languages in the world. The development of this language has made it possible to distinguish four different varieties. First, Old Arabic, which is no longer used today, is only present in earlier literature, more specifically in poetry. Secondly, Classical Arabic, also known as Literary Arabic, refers to the official language of Islam's Holy Book. Third, Modern Standard Arabic (MSA), a more modern version of Classical Arabic and finally dialects. MSA is the official language of all countries in the Arab world. It is used in a range of contexts; namely, media, education, business, literature and in official or juridical written documents [15, 16].

Furthermore, we could categorize Arabic dialects regionally, such as: North African, Egyptian, Levantine, Yemeni, Gulf. Or, Sub-regionally: Moroccan, Tunisian, Lebanese, Jordanian, Kuwaiti, and Qatari [17].

Moroccan dialect poses many challenges, for several reasons:

(1) The Moroccan constitution recognizes Arabic and Tamazight as the two official languages of Morocco. The Moroccan dialect is used among Arabic speakers, among Arabic and Amazigh speakers, and among Amazigh speakers of different Amazigh dialects. It is used in informal social settings.

(2) Moroccan dialect is the most used language in Morocco according to the official population census of 2014 (90% of Moroccans use the Moroccan dialect).

(3) Only 27% of them could use at least one of the oral forms of Tamazight [14, 18, 19].

(4) Moroccan Arabic diverges from MSA at the lexical and phonological levels due to various factors, including Morocco geographic location, colonial history and other considerations. Moroccan dialect has had some linguistic influences from the French and Spanish protectorates through loanwords. Alternatively, Arabic, especially at the lexical level, heavily influences Moroccan dialect as 81% of Moroccan words come from the Arabic language. Table 1 shows some Moroccan dialect loanwords from Arabic language.

(5) Moroccan dialect is challenging to understand because of the great distance between Morocco and the Middle East. Furthermore, Moroccan culture is not as widespread as Egyptian or Gulf culture through television broadcasts.

(6) Morocco has been colonized by different countries in the past, which has affected its dialect. The Moroccan dialect is influenced by foreign words from the French or Spanish language [5]. To a greater or a lesser degree 11% from French, and less than 1% from both Spanish and Tamazight languages [18, 12, 16].

Table 1. Moroccan dialect loanwords from Arabic language

Moroccan Dialect	Arabic Language	English Translation
بزاف	بجزاف	a lot
كيفاش	كيف هو الشيء	how
كيخربق	خربق الشيء أفسده	mess
بغى	بغى الشيء طلبه	want
فالطة	فلتة	fault

Nowadays, Morocco has reached approximately 37 million people divided into 12 regions of the kingdom. Each region uses a specific dialect. Moroccan dialect is considered the most diverse dialect, containing urban, Amazigh and Sahraouian dialects. Along with each dialect, there are sub-dialects, for instance Tamazight contains Rifia (of the Rif region), Soussia (of the Middle and High Atlas region), among others [10].

Linguists provide different classifications of Moroccan dialect types. For one, Fatima Sadiqi (2002) distinguishes five varieties of Darija:

(1) The Shamali variety in the north of Morocco.
(2) The Fassi variety in the center.
(3) The Rabat/Casablanca variety around these two cities.
(4) The Marrakshi/Agadiri variety in the south.
(5) The hassaniya variety in the Sahara.

Other researchers, such as Boukous and Amour, have divided Darija into four varieties only: The Mdini, which designates people living in the city, the Jebli, referring to people in the mountains, the Arubi, meaning the Bedouin population, and the Aribi for Hassani in southern Morocco [16].

As is the case with most Arabic dialects, Darija remains highly under-resourced from a computational point of view (NLP, machine translation, etc.) even though the written content of Moroccan dialect has developed rapidly on the Internet in recent years. Notably, through social media platforms, such as Facebook, Twitter, YouTube, etc [10]. This occurs through diverse forms such as written texts, audio and video documents, by using either the Arabic alphabet or a combination of the Latin alphabet and numbers [12]. Adoption of new modes of communication (SMS, Facebook, Twitter...), which are widespread in Arab countries, have strengthened dialect writing, particularly in Latin characters, the written content based on Latin Script is called " Arabizi " [19]. The next section reviews some of the features of Moroccan Darija.

3.2 Characteristics

3.2.1 Phonological level

The Moroccan Arabic (MA) consonantal system contains 28 consonant phonemes and four vowel phonemes [20] less vowels compared to Classical Arabic, but share the same features. MA uses non-Arabic phonemes /g/, /p/ and /v/, which are mainly used in words borrowed from foreign languages as French [21], for instance, the word Virage meaning a bend or turn, is used in Moroccan dialect; pronounced as /Virage/ but written as (ڤيراج). The influence of Berber on MA is mainly seen at the phonological level. This influence is shown in many areas through the deletion or addition of a segment [21]. For example deleting Hamza (Arabic ء : همزة). The word الماء (Water) is pronounced in Classical Arabic /Almaa/, but in MA it is pronounced as /Lma/.

3.2.2 Morphological level

Moroccan Arabic is similar to MSA in a few ways. First, the order subject-verb object [22] dominates the composition of a sentence in Darija. Also, the original MSA words keep their morphology, or undergo some changes in their patterns or affixes [23]. For instance, the prefix غا or the word غادي in Darija, indicates the intention of speaker to do the action in the future. Moreover, writing the verb between the prefix ما and the letter ش designate negation (ما بغيتش) which means (I won't). Furthermore, adding (كـ) to the verb refers to doing something at the same time of speaking (indicative present tense) like (كنرسم) which means I am drawing. Finally, we use ن as a prefix for present tense first person singular (ناكل) which means (I eat) [24].

3.2.3 Syntactic level

Moroccan dialect, like other Arabic dialects, has some syntactic specificities. Unlike English which uses Subject Verb Object (SVO) [25]. Moroccan dialect offers many possibilities for word order. The speaker will usually start the sentence with the element he or she wants to emphasize. In this sense, SVO and VSO can be used interchangeably. However, OSV and OVS are the most rarely applied [21]. Moreover, the disappearance of the feminine plural of the second person أنتن and the feminine plural of the third person هن. Beyond that, regional variations [25] in syntax play a key role in changes of Moroccan dialect. For example, the gender marker in the second person singular in Casablanca is different from Fes, Table 2 gives some Gender marker difference between cities.

(1) Word order flexibility and Agreement: a sentence in Moroccan dialect has a subject and a predicate. It could have a simple SVO form, where the subject leads the verb and the complement of the verb. It can also have a VSO Structure, where the verb comes before the subject and the direct object [26]. Moroccan dialect demands complete agreement in person, number, and gender between the verb and the subject, independently of the word order.

(2) Pro-drop nature: SA permits missing subjects in the subject position … of clauses [27], on the other hand, Moroccan dialect also drops the pronouns in the subject position, each the sentence below lacks subject yet remains grammatically correct. Table 3 gives an example of Moroccan dialect sentences without Personal Pronouns.

Table 2. Gender marker difference between cities

Personal Pronouns	Casablanca	Fes
The Second Person Singular (You) masculine	أَنْتَ	نْتِينَا
The Second Person Singular (You) feminine	أَنْتِ	نْتِينَا

Table 3. Moroccan dialect sentences without personal pronouns

Without Personal Pronouns	Moroccan Dialect Sentence
He made tea	صاوب آتاي
They arrived last night	وصلوا اليوم

3.2.4 Lexical level

Moroccan dialect consists of Arabic with a number of borrowed words from many different languages. This was a direct influence of the colonization of Morocco in the past by different countries such as France, Spain [21] and Portugal, which affected daily communication. Moroccan dialect vocabulary is rich due to historical events. Moroccan Arabic speakers use it and so do Amazigh, who represent about 50% of population in Morocco [19], as well as Amazigh speakers who do not understand each other; when the Amazigh variety of each person is not identical. Moroccan dialect includes verbs, nouns, and pronouns, like Arabic language.

3.3 Challenges

3.3.1 Phonological variety

Many phonological ambiguities make dealing with Moroccan dialect challenging due to the different varieties of this dialect:

(1) The /q/ variant differ according to geographical regions, for example in Fes, Tetouan and Tangier the community use /q/. In the other cities most people use /g/ for example: /gal/ becomes /qal/, which translates in English as 'he says' in Fes variety.

(2) Metathesis: the same word keeps meaning with a transposition of letters, for instance (زعما) /z3ma/ becomes (زمعا) /zm3a/

(3) The double process of assimilation-deletion: words undergo some small transformation, but the meaning does not change. For purposes of fast connected speech, the word (بنتي) /bnty/ 'my daughter' becomes (بتي) /bty/ deleting "ن" but keeping the same meaning.

(4) The use of /d/ could be different from one region to another. Some communities use it as a simple /d/ but others use it as a /t/ [25].

(5) Intonation: this key point varies from language to language [20]. This is why we are under the obligation of keeping punctuation marks to know the difference between a declarative sentences, interrogative sentences, and imperative sentences.

3.3.2 Morphological richness

Diacritical marks in Arabic give different meanings to the same word, which is the case for Moroccan dialect. Deleting diacritical marks give rise to serious problems while determining the location of the vowels. Moreover, decomposing words occurs in two cases:

(1) Decomposition at the level of words: verb, name …

(2) Decomposition at the level of grammatical information: noun (singular), verb (1st person) [6].

Thanks to the richness of this dialect, it is common to create a sentence from one word that translates into a five-word English sentence "غيبيعوها" and 'they will sell it' [28]. The fact that there is a huge number of vocabulary items in Moroccan dialect compared to English represents a challenge for models based on Machine Learning [29].

3.3.3 Syntactic ambiguity

A sentence has only one meaning when we use diacritical marks, but without them the sentence can be interpreted in several ways, all of which are syntactically correct [6], which we could observe on Table 4.

Table 4. Meanings of sentence

Sentence	First Meaning	Translation	Second Meaning	Translation
بغيت هاد الورقة	بْغِيتْ هَادْ الوَرْقَة	I want this document	بْغِيتِ هَادْ الوَرْقَة	Do you want this document?
عمر السيارة	عَمَّرْ الطُّونُوبِيلْ	Fill the car	عْمَرْ الطُّونُبِيلْ	How old the car is?

This syntactic ambiguity exists because of the difficulty of adding diacritical marks in conversation.

3.3.4 Lexical ambiguity

Darija is a spoken dialect without rules, so everyone writes the same word in different ways. Many users on the web write Moroccan dialect with the Latin letters, and numbers to illustrate letters, which does not exist in Latin. To deal with this ambiguity, we need to transform all illustrated letters to Arabic letters. Table 5 gives examples of these:

Table 5. Illustration of Arab letters

Arab Letters	Illustration Letters
ح	7
خ	5
ع	3

Moroccan dialects suffer from many problems, including phonological variations, which depend on the geographical region, as well as metathesis, adding that we sometimes have words with the same meaning but with a small transformation, ultimately pronouncing the same letter differently, and intonation can change the meaning. In fact, the large number of vocabulary and diacritical marks in the Moroccan dialect represent a morphological richness and, at the same time, a challenge for models based on machine learning. To mention that the absence of diacritical marks represents a syntactic ambiguity This syntactic ambiguity exists because of the difficulty of adding diacritical marks in conversation. Then there are Arabizi letters, which require us to transform each of them into Arabic letters. Table 5 shows some illustration of Arab letters.

4. CORPUS BUILDING

4.1 Data collection

The purpose of the corpus was to create a new dataset for Moroccan dialect because of a lack of data in this language and with the aim of training a Chatbot on how to speak Moroccan dialect easily. YouTube proved to be the most suitable platform for many reasons:

(1) The popularity of YouTube and its use by many different Moroccan speakers.
(2) The capability to collect audio and text data at the same time.
(3) The possibility to download records using a playlist and python scripts.

The choice of an appropriate playlist was not a straightforward process. We started by selecting content that represented the Moroccan population, using informal language with no swearing. Each video had to contain different forms of greeting, which we planned to use later to train the chatbot. In parallel, we were looking for clear voices with good pronunciation. We were interested in various themes relevant to the Moroccan society, including cooking videos where everyone could discover our delicious and traditional meals, social activities that raise awareness of the importance of helping others and showing solidarity. Stories relating to Moroccan history and important events in our country give a glimpse of modern Morocco. Finally, we have included proverbs inherited from our predecessors and commonly used in everyday conversation. Each theme contains a different number of videos, depending on their length. Table 6 below gives the relevant numbers.

After the preparation of appropriate playlists on the YouTube platform, with different themes, we used YouTube-dl, which is a command-line program to download videos from YouTube website. This requires a Python interpreter and is suitable for any platform.

Each audio file was filed in the suitable folder according to its category. All files had the same "MP3" format. Each recording was further divided into small 30 seconds recordings, using "mp3splt-gtk" to facilitate the next step (Transcription). This is an application which allows for cutting MP3 tracks into as many pieces as you want. Finally, we obtained the number of divisions mentioned in the Table 6 below:

Table 6. Dataset summary

#	Category	Records	Number of Splits	Mean of Records	Sum of Minutes
1	Social	7	98	6.61min	46.33min
2	Cooking	8	172	10.29min	82.32min
3	History	11	281	12.32min	135.53min
4	Proverbs	37	159	1.68min	62.47min
Total	4	63	710	-	326.65min

Figure 1 below summarises the various phases of data collection.

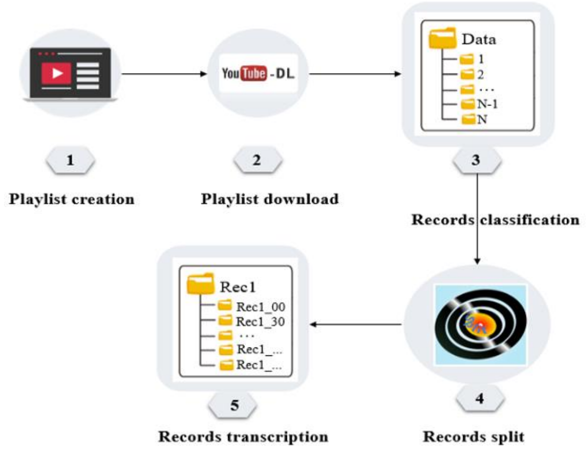


Figure 1. The steps of data collection

4.2 Data transcription

Chatbots are able to understand text messages. In recent years, there is a great deal of research to develop the ability to understand voice and play both roles at the same time: chatbot and voice assistance. Therefore, we chose to create a voice and text dataset at the same time in order to obtain the same text from the voice. Therefore, to transcribe these recordings, we adopted two test methods to decide which method we would use for the whole transcription.

4.2.1 Automatic transcription

Many websites offer transcription services, but we were looking for a good result, while we had no time to waste on correction. After a long search, we chose four websites where we uploaded the same recording. After a short time, we obtained the result. Automatic transcription seemed like a promising idea at first, but after testing several websites, we made several observations about the quality of the transcription. First of all, some websites transcribe in a dialect other than Moroccan, which gave us extra work. In addition, we tried to ensure that the text corresponded precisely to the words in the audio file, but we sometimes came across additional words that did not exist in the original audio file. In addition, some websites changed the meaning of the sentence, which could pose a problem when testing the dataset further with machine learning models. This Table 7 below shows the observations on each transcription.

Table 7. List of transcription services websites

Website	URL	Observation
Kateb!	https://kateb.ai/home	Egyptian transcription
Speeh to text	https://spechtotext.com/VoiceTyping/SpeechToTextArabic	Words do not exist in the recording
Speech texter	https://www.speechtexter.com/	Correct words but different from the original meaning
Talk typer	https://talktyper.com/	Correct words but we need to correct the script manually

4.2.2 Manual transcription

This method was achieved using a stopwatch and several recordings of different speakers. We noticed that all recordings have the same duration of 30 secs, we concluded that speech pace influences the writing time parameter. Listening and writing at the same time was a bit difficult, but we were able to obtain the results presented in Figure 2 below which represent time of writing/min:

In order to make the right decisions, we compared all the parameters of each method.

We chose manual transcription for many reasons:

(1) We need a correct text spelling to train the model.
(2) Manual transcription is time consuming but gives the best result in the end. However, automatic transcription requires a double effort: transcription and correction.
(3) We are obliged to have the correct spellings of the same word.

We adopted the manual method for all the reasons cited above. However, we cannot deny the challenges we encountered during this process and the efforts we made to overcome them. Firstly, we divided the tasks between several transcribers in order to rationalize the process. The pace of the speech influenced the time taken to transcribe; some videos had speakers who transmitted information rapidly, which increased the transcription time compared to videos of the same length but with slower speech. We asked the transcribers to take the time to listen and write the content accordingly. Furthermore, we recognize that manual transcription is time-consuming and requires a high level of commitment to achieve the desired result. In addition, to ensure that the generated text matches the content of the extract, transcribers need to listen carefully to the extract several times at the end of the transcription. Knowing that each word can have different forms of spelling, we were flexible in accepting different forms of a word in order to enrich the dataset. Given that women are often used to cooking and can easily understand each split, we entrusted this task to a woman. Transcription was done with the assistance of young native speakers, who are proficient in Moroccan Dialect, and who had experience in technological fields. Four people, one woman and three men transcribed the recordings. We chose different people for this mission to make sure we get different spellings of the same word. The transcriber must be careful when he or she is writing the content, and they should read the transcription several times to be sure it does not contain mistakes. After transcribing 710 recordings, we were able to obtain 816 rows in Excel, each row containing several lines.

Apart from the rationale mentioned earlier, we opted for the manual transcription for several other compelling reasons. One of the main considerations was the uniqueness and complexity of the Moroccan Dialect, which demanded the expertise of young native speakers who were not only well-versed in the language but also had a background in technological fields. This combined knowledge ensured that the transcriptions would accurately capture the nuances and intricacies of the dialect, making it an invaluable resource for further research and applications in natural language processing. To ensure diversity and comprehensive coverage of the language, a team of four skilled transcribers was assembled, comprising one woman and three men. This deliberate selection aimed to obtain a wide range of perspectives and insights into the different variations and spellings of words in the Moroccan Dialect. By having multiple transcribers, the dataset could capture the rich tapestry of linguistic expressions found within the language, leading to a more robust and adaptable model for future language-based AI technologies. The transcribers were well aware of the responsibility they carried in this endeavor. They understood the significance of their meticulous work in building a high-quality dataset. Precision and accuracy were paramount, as even a minor mistake in the transcription could have far-reaching consequences on the AI model's performance. Therefore, the transcribers approached their task with great care and attention to detail.

Given the complexity of the Moroccan Dialect, the transcribers were aware of the potential challenges they might encounter. They made it a point to read the content they transcribed multiple times to ensure its faithfulness to the original recordings. This rigorous reviewing process not only eliminated errors but also served as an opportunity to gain a deeper understanding of the dialect's unique features and linguistic patterns. After dedicating substantial effort and time, the team successfully transcribed a total of 710 records. The resulting dataset proved to be a valuable repository of language data, containing 816 rows in Excel, Figure 3 represent a screenshot of Dataset, with each row containing several lines of transcribed text. This substantial dataset laid the foundation for the development of a powerful Chatbot and other AI applications, capable of interacting seamlessly with users in the Moroccan Dialect. The insights gained from this project can be applied to other dialects and languages, creating a ripple effect of progress in the field of natural language processing and AI.

The decision to adopt the manual method for transcription, the careful selection of transcribers, and the meticulous approach taken in creating the dataset have proven to be vital steps in the evolution of language-based AI technologies.

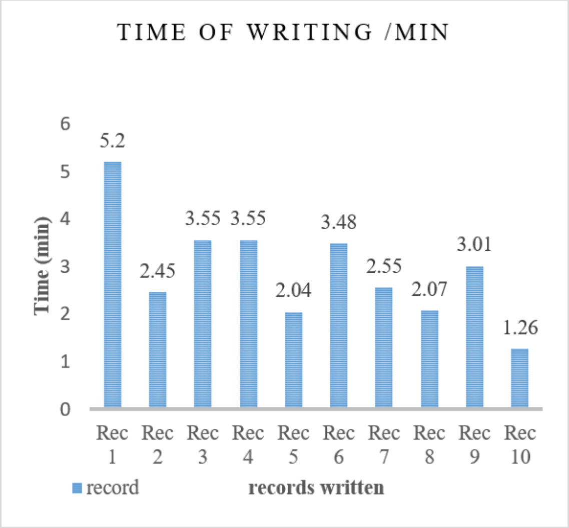
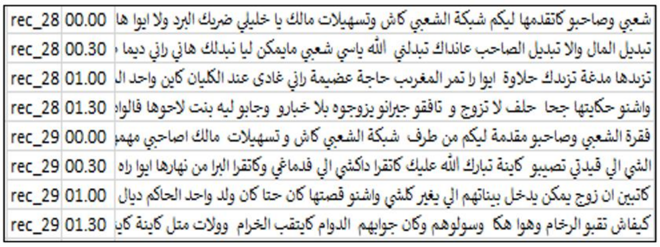


Figure 2. Time of writing/min



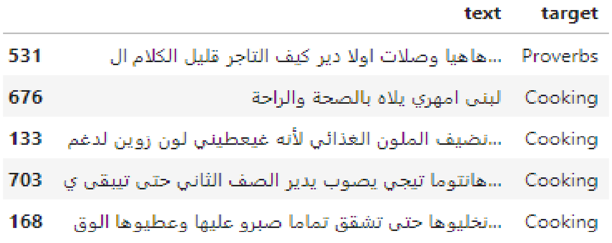
rec_28	00.00	شعبي وصاحبو كاتقدمها ليكم شبكة الشعبي كاش وتسهيلات مالك يا خليلي ضريك البرد ولا ايوا ها
rec_28	00.30	تبديل المال والا تبديل الصاحب عانداك تبدلني الله ياسي شعبي مايمكن ليا نبدلك هاني راني ديما
rec_28	01.00	تزيدها مدغة تزيدك حلاوة ايوا را تمر المغريب حاجة عضيمة راني غادى عند الكليان كاين واحد ال
rec_28	01.30	واشنو حكايتها جحا حلف لا تزوج و تافقو جيرانو يزوجوه بلا خبارو وجابو ليه بنت لاحوها فالوا
rec_29	00.00	فقرة الشعبي وصاحبو مقدمة ليكم من طرف شبكة الشعبي كاش و تسهيلات مالك اصاحبي مهم
rec_29	00.30	الشي الي قيدتي تصيبو كاينة تبارك الله عليك كاتقرا داكشي الي فدماغي وكاتقرا البرا من نهارها ايوا راه
rec_29	01.00	كاتبين ان زوج يمكن يدخل بيناتهم الي يغير كلشي واشنو قصتها كان حتا كان ولد واحد الحاكم ديال
rec_29	01.30	كيفاش تقبو الرخام وهوا هكا وسولوهم وكان جوابهم الدوام كايتقب الخرام وولات متل كاينة كاي

Figure 3. Screenshot of dataset

5. CORPUS STATISTICS

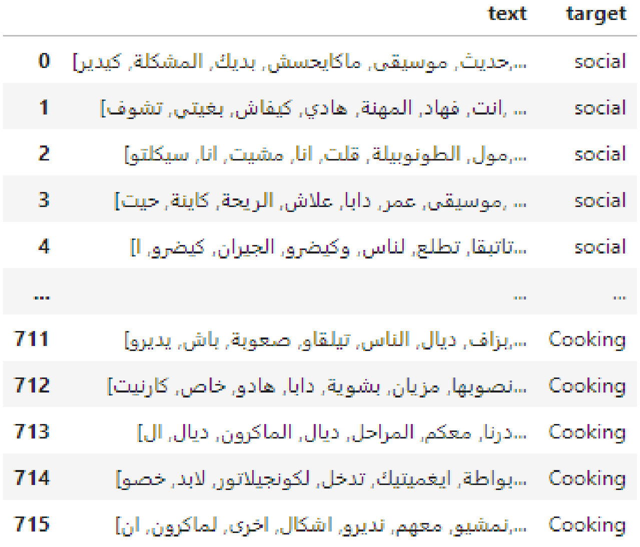
It should be noted that before creating a machine learning or deep learning model, we had to clean the dataset. The cleaning process is a crucial step in preparing our dataset to be understandable by a computer. Cleaning tools are quite poor when we are dealing with Moroccan dialect. Hence, we chose to do it manually. Figure 4 shows a screenshot after a manual cleaning of our dataset. At first, by not only deleting columns without useful information, but also empty rows, and assigning a suitable target for each row.

Data processing is time-consuming, but using the appropriate libraries to do it automatically helps us to reduce time. Although we do not have libraries devoted to the Moroccan dialect, we were able to use Natural Language Toolkit (NLTK), which is a platform for building Python programs to work with human language data. Text pre-processing techniques involves converting the raw data into an understandable structure, focusing on keywords that emphasise the context of the sentence or paragraph. The order in which the NLP pipeline is built has a considerable influence on the result. Tokenization, empty word removal, punctuation removal and lemmatisation are widely recognised and effective text pre-processing techniques [30]. We consider that the pre-processing steps have a significant impact on the accuracy of machine learning algorithms [31].



	text	target
531	هاهيا وصلات اولا دير كيف التاجر قليل الكلام ال...	Proverbs
676	لبنى امهري يلاه بالصحة والراحة	Cooking
133	تضيف الملون الغذائي لأنه غيعطيني لون زوين لدغم...	Cooking
703	هاتتوما تيجي يصوب يدير الصف الثاني حتى تيبقى ي...	Cooking
168	تخليوها حتى تشقق تماما صبرو عليها وعطيوها الوق...	Cooking

Figure 4. Screenshot of dataset after manual cleaning



	text	target
0	[حديث, موسيقى, ماكايحسش, بديك, المشكلة, كيدير...	social
1	[انت, فهاد, المهنة, هادي, كيفاش, بغيتي, تشوف, ...	social
2	[مول, الطونوبيلة, قلت, انا, مشيت, انا, سيكلتو...	social
3	[موسيقى, عمر, دابا, علاش, الريحة, كاينة, حيت, ...	social
4	[تاتيقا, تطلع, لناس, وكيضرو, الجيران, كيضرو, ا...	social
...	...	...
711	[بزاف, ديال, الناس, تيلقاو, صعوبة, باش, يديرو...	Cooking
712	[نصوبها, مزيان, بشوية, دابا, هادو, خاص, كارنيت...	Cooking
713	[درنا, معكم, المراحل, ديال, الماكرون, ديال, ال...	Cooking
714	[بواطة, ايغميتيك, تدخل, لكونجيلاتور, لابد, خصو...	Cooking
715	[نمشيو, معهم, تديرو, اشكال, اخرى, لماكرون, ان...	Cooking

Figure 5. Screenshot of dataset after automatic cleaning

After removal of punctuation and repetitive characters, we moved to the normalization of the dataset to make it standard and to reduce inflectional forms through stemming. Additionally, the segmentation of the dataset into sentences and words was done through tokenization, which helped us to get each word as a token, thus allowing us to remove stop words using the NLTK library. Figure 5 exposes the result of our automatic cleaning.

The analysis of our dataset allows us to obtain the number of words per theme, the frequency of words used, and some other information that we are going to show in Figure 6 and Table 8 below.

Table 8. Dataset in numbers

	Words Total	Vocabulary Size	Max Sentence Length
Social	1909	974	68
Cooking	10902	3701	62
History	17816	6951	170
Proverbs	8318	3587	71

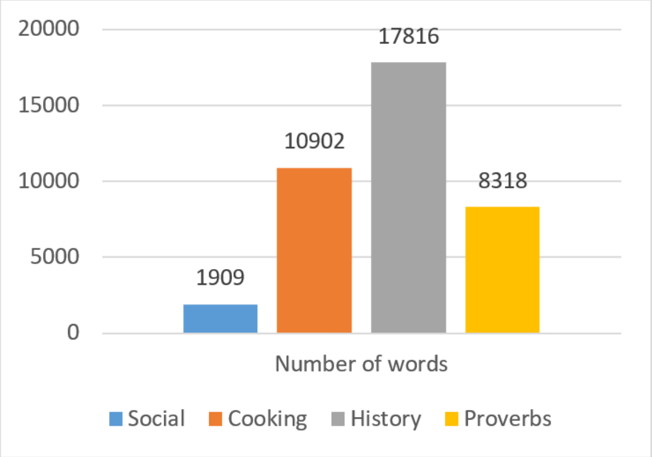


Figure 6. Number of words on each theme

The Figure 7 shows the word cloud of social context content, we observe word frequencies such as " هاد، دابا، الناس، الله، هادي، علاش". These words provide contextual understanding and meaning of the social theme. In another way, words like " شكرا، كامل، الخدمة، كاين" emphasize the importance of the general context, indicating its relevance to the social theme.

The Figure 8 represents the word cloud of cooking content, we see word frequencies such as "شويا، بزاف، ديال، باش، يعني". These words help to make sense of the context and contribute to the meaning of the cooking theme. On the other hand, such words as "الكمية، السكر، الضروري، كنزيدو، الطريقة، البصلة" highlight the importance of the general context, indicating its relevance to the cooking theme through the presence of words linked to this topic.

The Figure 9 exposes the word cloud of historical content, we see word occurrences such as 'ديال، هاد، المغرب، غادي، الناس'. These words contribute to understanding the context and help to convey the meaning of the story's theme. In contrast, words such as "الجيش، بزاف، المغاربة، كلشي، البلاد" emphasise the importance of the overall context, indicating its pertinence to the theme of the story through the presence of words related to this topic.

Overall, as we can see on the figures, words are different for each theme, which could give us a global view about the diversity of our dataset. This dataset contains written and vocal data with over 650 recordings. Each recording has a playback time of 30 seconds and is divided into four classes. In comparison, many of the older corpora do not have specific classes. The CFMD is accessible to the public, whereas the other corpora do not contain voice recordings. For all these reasons, we can see that the CFMD is created with a unique combination that does not exist in other corpora, which gives it added value. Table 9 below shows a comparison of all the corpora mentioned.

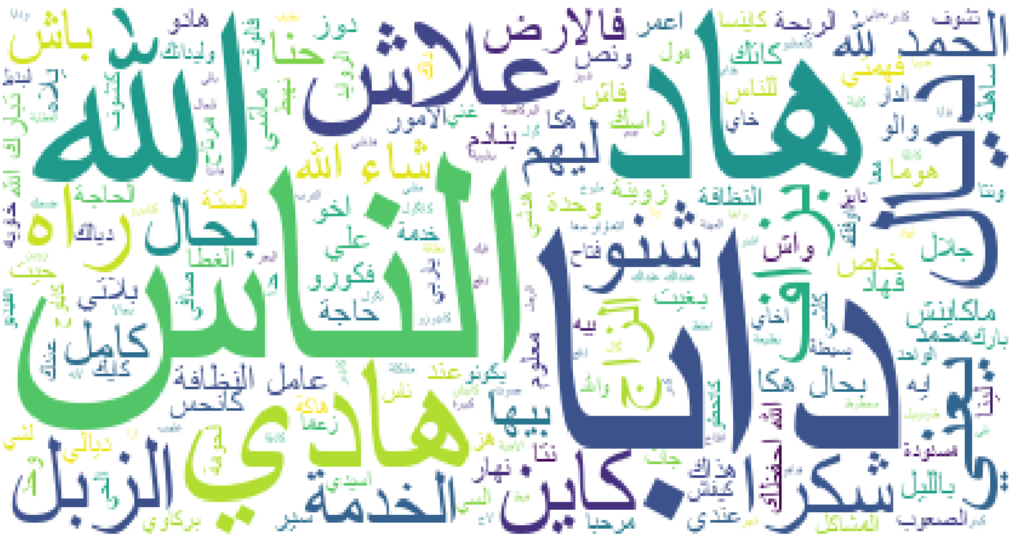


Figure 7. Word cloud of top words in social

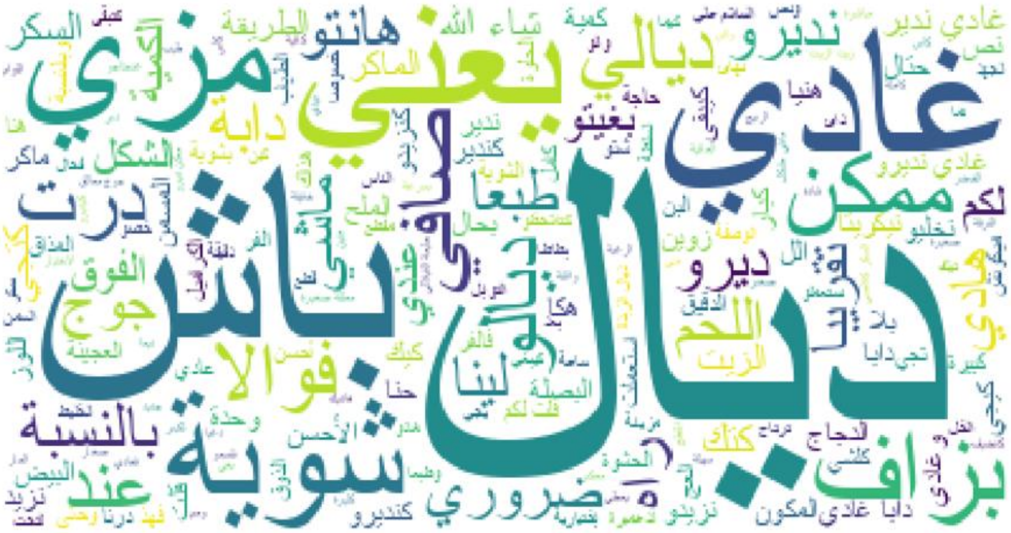


Figure 8. Word cloud of top words in cooking

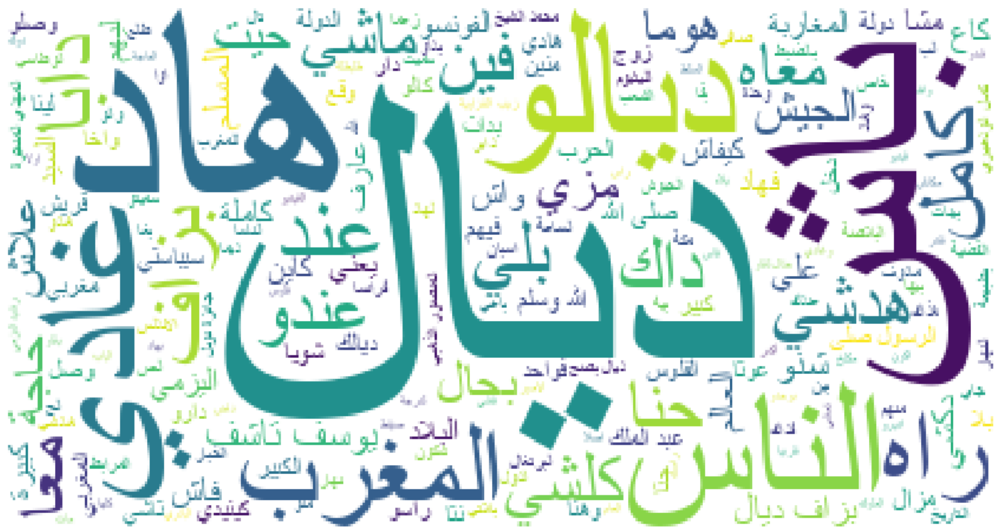


Figure 9. Word cloud of top words in history

Table 9. Existing Moroccan dialect corpora

#	Corpus	Language	Source	Size	Year	Accessibility	Reuse	No. of Text	No. of Classes	Medium	Purpose
1	The MADAR Arabic Dialect Corpus	25 Arabic city dialects+MSA	translating selected sentences from the Basic Traveling Expression Corpus (BTEC)	+372K words	2018	Public	Private	+62000 sentences	25 cities from the Arab World	Written	Translation
2	GOUD.MA : a news article dataset	Moroccan Darija MSA	News articles for automatic summarization in code-switched Moroccan Darija	+1,229,993 unique tokens	2022	Public	Public	+158k news articles	10 categories over articles of the GOUD.MA dataset	Written	to establish baselines for Moroccan Darija summarization
3	Moroccan Dialect-Darija-Open Dataset (DODA)	Moroccan dialect English	design with NLP-applications	+ 10K entries	2021	Public	Public	+10K entries	8 classes	Written	Understand and study Moroccan dialect
4	Building a Moroccan dialect electronic Dictionary (MDED)	Moroccan dialect MSA	translated manually to Moroccan dialect content of an MSA dictionary	+18K entries	2014	Private	Private	+18K entries	2 classes	Written	will be used as a lexical resource to build a machine translation for the Moroccan dialect
5	Dialectal Voice: An Open-Source Voice Dataset	Moroccan dialect MSA	-Dvoice platform -Web scrape -Label a set of collected recordings via the Speech Recognition library	+2990 Files	2021	Public	Public	+2990 Files	——	Written & speech	Automatic voice recognition system

6. CONCLUSIONS

In this paper, our primary objective was to address the significant lack of resources for the Moroccan dialect, which is severely under-resourced. To achieve our goal, we devised a unique approach that involved harnessing the vast resources available on the YouTube platform by collecting diverse audio data related to various topics, we further categorized each audio clip into smaller records, then engaging native Moroccan speakers as transcribers. Their inherent familiarity with the dialect and cultural context was invaluable in ensuring accurate and authentic transcriptions. However, we acknowledged that this manual transcription, process was time-consuming, however, because we recognized the significance of producing a high-quality corpus that accurately represented the nuances and intricacies of this unique dialect, we persevered through our quest.

Unfortunately, time did not allow us to expand our database to include topics such as tourism and travel, as well as other relevant topics related to Moroccan culture. In addition, with fewer spelling variations associated with each region of Morocco, our database has remained relatively small compared to databases in other languages including English and French. However, our future efforts will focus on expanding this dataset.

This corpus, which contains both textual and vocal data, represents an important asset for training chatbots and voice assistants using the Moroccan dialect. The inclusion of both modalities enhances its usefulness in various applications, including natural language processing (NLP) tasks, speech recognition systems and conversational agents.

REFERENCES

[1] Wołk, K., Wołk, A., Wnuk, D., Grześ, T., Skubis, I. (2022). Survey on dialogue systems including slavic languages. Neurocomputing, 477: 62-84. https://doi.org/10.1016/j.neucom.2021.11.076

[2] Adamopoulou, E., Moussiades, L. (2020). An overview of chatbot technology. In IFIP International Conference on Artificial Intelligence Applications and Innovations. Springer, Cham, pp. 373-383. https://doi.org/10.1007/978-3-030-49186-4_31

[3] Zhang, L., Yang, Y., Zhou, J., Chen, C., He, L. (2020). Retrieval-polished response generation for chatbot. IEEE Access, 8: 123882-123890. http://doi.org/10.1109/ACCESS.2020.3004152

[4] McTear, M. (2021). Challenges and future directions. In: Conversational AI. Synthesis Lectures on Human

Language Technologies. Springer, Cham. https://doi.org/10.1007/978-3-031-02176-3_6

[5] Harrat, S., Meftouh, K., Smaïli, K. (2017). Creating parallel Arabic dialect corpus: Pitfalls to avoid. In 18th International Conference on Computational Linguistics and Intelligent Text Processing (CICLING).

[6] Masmoudi, A., Mdhaffar, S., Sellami, R., Belguith, L.H. (2019). Automatic diacritics restoration for Tunisian dialect. ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), 18(3): 1-18. http://doi.org/10.1145/3297278

[7] Allak, A., Naira, A.M., Benelallam, I., Gaanoun, K. (2021). Dialectal voice: An open-source voice dataset and automatic speech recognition model for Moroccan Arabic dialect. In: NeurIPS Data-Centric AI Workshop. https://scholar.google.com/citations?view_op=view_citation&hl=fr&user=ED8WE9QAAAAJ&citation_for_view=ED8WE9QAAAAJ:u-x6o8ySG0sC.

[8] Issam, A., Mrini, K. (2022). Goud. ma: A news article dataset for summarization in Moroccan Darija. In 3rd Workshop on African Natural Language Processing.

[9] Gaanoun, K., Naira, A.M., Allak, A., Benelallam, I. (2024). Darijabert: A step forward in NLP for the written Moroccan dialect. International Journal of Data Science and Analytics, 1-13. http://doi.org/10.1007/s41060-023-00498-2

[10] Jarrar, M., Habash, N., Alrimawi, F., Akra, D., Zalmout, N. (2017). Curras: An annotated corpus for the Palestinian Arabic dialect. Language Resources and Evaluation, 51: 745-775. https://doi.org/10.1007/s10579-016-9370-7

[11] Boujou, E., Chataoui, H., Mekki, A.E., Benjelloun, S., Chairi, I., Berrada, I. (2021). An open access NLP dataset for Arabic dialects: Data collection, labeling, and model construction. arXiv Preprint arXiv: 2102.11000. https://doi.org/10.48550/arXiv.2102.11000

[12] Outchakoucht, A., Es-Samaali, H. (2021). Moroccan dialect-darija-open dataset. arXiv Preprint arXiv: 2103.09687. https://doi.org/10.48550/arXiv.2103.09687

[13] Tachicart, R., Bouzoubaa, K., Jaafar, H. (2014). Building a Moroccan dialect electronic dictionary (MDED). In 5th International Conference on Arabic Language Processing, pp. 216-221.

[14] Tachicart, R., Bouzoubaa, K. (2022). Moroccan Arabic vocabulary generation using a rule-based approach. Journal of King Saud University-Computer and Information Sciences, 34(10): 8538-8548. https://doi.org/10.1016/j.jksuci.2021.02.013

[15] Masmoudi, A., Bougares, F., Ellouze, M., Estève, Y., Belguith, L. (2018). Automatic speech recognition system for Tunisian dialect. Language Resources and Evaluation, 52: 249-267. https://doi.org/10.1007/s10579-017-9402-y

[16] Labied, M., Belangour, A. (2021). Moroccan dialect "Darija" automatic speech recognition: A survey. In 2021 IEEE 2nd International Conference on Pattern Recognition and Machine Learning (PRML), Chengdu, China, pp. 208-213. https://doi.org/10.1109/PRML52754.2021.9520690

[17] Bouamor, H., Habash, N., Salameh, M., Zaghouani, W., Rambow, O., Abdulrahim, D., Obeid, O., Khalifa, S., Eryani, F., Erdmann, A., Oflazer, K. (2018). The madar arabic dialect corpus and lexicon. In Proceedings of The Eleventh International Conference on Language Resources and Evaluation (LREC 2018).

[18] Tachicart, R., Bouzoubaa, K. (2021). Moroccan data-driven spelling normalization using character neural embedding. Vietnam Journal of Computer Science, 8(01): 113-131. https://doi.org/10.1142/S2196888821500044

[19] Younes, J., Souissi, E., Achour, H., Ferchichi, A. (2020). Language resources for Maghrebi Arabic dialects' NLP: A survey. Language Resources and Evaluation, 54: 1079-1142. https://doi.org/10.1007/s10579-020-09490-9

[20] Ennaji, M., Makhoukh, A., Es-saiydi, H., Moubtassime, M., Slaoui, S. (2004). A grammar of Moroccan Arabic. Faculté des Lettres et des Sciences Humaines, Université Sidi Mohamed Ben Abdellah.

[21] Harrat, S., Meftouh, K., Smaïli, K. (2018). Maghrebi Arabic dialect processing: An overview. Journal of International Science and General Applications, 1.

[22] Alshammari, A.R. (2023). Analyzing word order variation and agreement asymmetry in SVO and VSO structures of Standard Arabic: Towards a unified account.

[23] Tachicart, R., Bouzoubaa, K. (2014). A hybrid approach to translate Moroccan Arabic dialect. In 2014 9th International Conference on Intelligent Systems: Theories and Applications (SITA-14), Rabat, Morocco, IEEE, pp. 1-5. https://doi.org/10.1109/SITA.2014.6847293

[24] Al-Shargi, F., Kaplan, A., Eskander, R., Habash, N., Rambow, O. (2016). Morphologically annotated corpora and morphological analyzers for Moroccan and Sanaani Yemeni Arabic. In 10th Language Resources and Evaluation Conference (LREC 2016). Portoroz, Slovenia.

[25] Announi, I. (2021). Moroccan linguistic variation: An overview. Lingua. Language and Culture, 20(1): 153-174.

[26] Announi, I. (2021). The problem of word order and verbal movement in Moroccan Arabic. International Journal of Linguistics, Literature and Translation, 4(4): 34-54. http://doi.org/10.32996/ijllt.2021.4.4.6

[27] Altamimi, M.I. (2015). Arabic pro-drop. Master Thesis, Eastern Michigan University, USA.

[28] Moukafih, Y., Sbihi, N., Ghogho, M., Smaïli, K. (2021). Improving machine translation of Arabic dialects through multi-task learning. In International Conference of the Italian Association for Artificial Intelligence. Cham: Springer International Publishing, pp. 580-590. https://doi.org/10.1007/978-3-031-08421-8_40

[29] Darwish, K., Habash, N., Abbas, M., Al-Khalifa, H., Al-Natsheh, H.T., Bouamor, H., Bouzoubaa, K., Cavalli-Sforza, V. El-Beltagy, S.R., El-Hajj, W., Jarrar, M., Mubarak, H. (2021). A panoramic survey of natural language processing in the Arab world. Communications of The ACM, 64(4): 72-81. https://doi.org/10.1145/3447735

[30] Tabassum, A., Patil, R.R. (2020). A survey on text pre-processing & feature extraction techniques in natural language processing. International Research Journal of Engineering and Technology (IRJET), 7(06): 4864-4867.

[31] Alam, S., Yao, N. (2019). The impact of preprocessing steps on the accuracy of machine learning algorithms in sentiment analysis. Computational and Mathematical Organization Theory, 25: 319-335. http://doi.org/10.1007/s10588-018-9266-8