*Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 6405–6411
Marseille, 11–16 May 2020


# Text and Speech-based Tunisian Arabic Sub-Dialects Identification

**Najla Ben Abdallah**[1], **Saméh Kchaou**[2], **Fethi BOUGARES**[3]
[1]FLAH Mannouba , [2]Sfax Université, [3]Le Mans Université
Tunis-Tunisia, Sfax-Tunisia, Le Mans-France
benabdalahnajla@gmail.com, samehkchaou4@gmail.com, fethi.bougares@univ-lemans.fr

**Abstract**

Dialect IDentification (DID) is a challenging task, and it becomes more complicated when it is about the identification of dialects that belong to the same country. Indeed, dialects of the same country are closely related and exhibit a significant overlapping at the phonetic and lexical levels. In this paper, we present our first results on a dialect classification task covering four sub-dialects spoken in Tunisia. We use the term 'sub-dialect' to refer to the dialects belonging to the same country. We conducted our experiments aiming to discriminate between Tunisian sub-dialects belonging to four different cities: namely Tunis, Sfax, Sousse and Tataouine. A spoken corpus of 1673 utterances is collected, transcribed and freely distributed. We used this corpus to build several speech- and text-based DID systems. Our results confirm that, at this level of granularity, dialects are much better distinguishable using the speech modality. Indeed, we were able to reach an F-1 score of **93.75%** using our best speech-based identification system while the F-1 score is limited to **54.16%** using text-based DID on the same test set.

**Keywords:** Tunisian Dialects, Sub-Dialects Identification, Speech Corpus, Phonetic description

## 1. Introduction

Due to historical and sociological factors, Arabic-speaking countries have two main varieties of languages that they use in daily life. The first one is Modern Standrad Arabic (MSA), while the second is the dialect of that country. The former (i.e. MSA) is the official language that all Arab countries share and commonly use for their official communication in TV broadcasts and written documents, while the latter (i.e. dialect) is used for daily oral communication. Arabic dialects have no official status, they are not used for official writing, and therefore do not have an agreed-upon writing system. Nevertheless, it is necessary to study Arabic dialects (ADs) since they are the means of communication across all the Arab World. In fact, in Arab countries, MSA is considered to be the **H**igh **V**ariety (HV) as it is the official language of the news broadcasts, official documents, etc. The HV is the standardized written variety that is taught in schools and institutions. On the other hand, ADs are considered to be the **L**ow Variety. They are used for daily non-official communication, they have no conventional writing system, and they are considered as being chaotic at the linguistic level including syntax, phonetics and phonology, and morphology. ADs are commonly known as spoken or colloquial Arabic, acquired naturally as the mother tongue of all Arabs. They are, nowadays, emerging as the language of informal communication on the web, including emails, blogs, forums, chat rooms and social media. This new situation amplifies the need for consistent language resources and processing tools for these dialects.

Being able to identify the dialect is a fundamental step for various applications such as machine translation, speech recognition and multiple NLP-related services. For instance, the automatic identification of speaker's dialect could enable call centers to orient the call to human operators who understand the caller's regional dialect.

**D**ialect **ID**entification (DID) task has been the subject of several earlier research and exploration activities. It is generally perceived that the number of arabic dialects is equal to the number of Arab countries. This perception is falsified when researchers come across remarkable differences between the dialects spoken in different cities of the same country. An example of this phenomenon is the dialectal differences found in Egypt (Cairo vs Upper Egypt) (Zaidan and Callison-Burch, 2014). Generally speaking, dialects are classified into five main groups: *Maghrebi*, *Egyptian*, *Levantine*, *Gulf*, and *Iraqi* (El-Haj et al., 2018). Latterly, a trend of considering a finer granularity of DID is emerging (Sadat et al., 2014; Salameh et al., 2018; Abdul-Mageed et al., 2018).

This work follows this trend and addresses the DID of several dialects belonging to the same country. We focus on the identification of multiple **T**unisian sub-**D**ialects[1]. The **T**unisian **D**ialect (TD) is part of the Maghrebi dialects. This former is generally known as the "Darija" or "Tounsi". Tunisia is politically divided into 24 administrative areas. This political division has also a linguistic dimension. Native speakers claim that differences at the linguistic level between these areas do exist, and they are generally able to identify the geographical origin of a speaker based on their speech. In this work we considered four varieties of the Tunisian Dialect in order to check the validity of this claim. We are mainly interested in the automatic classification of four different sub-dialects belonging to different geographical zones in Tunisia. The main contributions of this paper are two-folds: (i) we create and distribute the first speech corpus of four Tunisian sub-dialects. This corpus is manually segmented and transcribed. (ii) Using this corpus, we build multiple automatic dialect identification systems and we experimentally show that speech-based DID

[1]also referred to as regional *accents*

systems outperform text-based ones when dealing with dialects spoken in very close geographical areas.
The remainder of this paper is structured as follows: In Section 2, we review related work on Arabic DID. In section 3, we describe the data that we collected and annotated. We present the learning features in section 4. Sections 5 and 6 are devoted to the DID experimental setup and the results. In section 7 we conclude upon this paper and outline our future work.

## 2. Related Work

Although Arabic dialects are rather under-resourced and under-investigated, there have been several dialect identification systems built using both speech and text modalities. To illustrate the developments achieved at this level, the past few years have consequently been marked by the organization of multiple shared tasks for identifying Arabic dialects on speech and text data. For instance Arabic Dialect Identification shared task at the VarDial Evaluation Campaign 2017 and 2018, and MADAR Shared Task 2019.

Overall, the development of DID systems has attracted multiple interests in the research community. Multiple previous works, such as (Malmasi and Zampieri, 2016; Shon et al., 2017), who have used both acoustic/phonetic and phonotactic features to perform the identification at the regional level using a speech corpus. (Djellab et al., 2016) designed a GMM-UBM and an i-vector framework for accents recognition. They conducted an experiment based on acoustic approaches using data spoken in 3 Algerian regions: East, Center and West of Algeria. Five Algerian sub-dialects were considered by (Bougrinea et al., 2017) in order to build a hierarchical identification system with Deep Neural Network methods using prosodic features.

As regards DID systems of written documents, there are also several prior works that targeted a number of Arabic dialects. For instance, (Abdul-Mageed et al., 2018) presented a large Twitter data set covering 29 Arabic dialects belonging to 10 different Arab countries. In the same trend, (Salameh et al., 2018) presented a fine-grained DID system covering the dialects of 25 cities from several countries, including cities in the same country in the Arab World. In an interesting study, (Barkat, 1999) had as objective the identification of the most important cues that are salient for Automatic Dialect Identification (ADI). To this end, she selected six ADs, with three from the Eastern Area (Syria, Lebanon and Jordan), and three from the Western part of the Arab world (Morocco, Algeria, Tunisia). The data is in the form of audio files that were recorded with native speakers of the corresponding areas, who were asked to describe a photo. After recording, the researcher ran a perception test. She invited eighteen other native speakers to listen to the audio files. Their tasks were to first identify the geographical areas to which the speaker belongs, and then provide the segmental, lexical and prosodic cues that they relied on to decide upon the geographical origin of the recorded speakers. Interestingly, they identified phonetic cues along with morpho-syntactic, lexical and prosodic cues. Barakat argued that there was identification errors, but they were systematic and logical. In fact, she remarked that errors occur for geographically close areas. This means that when the speaker is from Morocco for example, they may wrongly be perceived as being from Algeria or Tunisia. She adds that these errors mostly occur when the perceiver is from a different geographical area than the speaker. The best identification results prove that when the speaker and the identifier are from the same geographical area of the speaker, correct identification trials reached 100 percent. Another interesting conclusion is that most correct identifications were for the Maghrebi dialects. In fact, even native identifiers who belong to the Eastern Area were able to correctly differentiate between the Western three dialects under study with a rate that reaches up to 75percent for the Moroccan dialect. In a similar approach, (Biadsy et al., 2009) use phonotactic modelling for the identification of Gulf, Iraqi, Levantine, Egyptian dialects, and MSA. they define the phonotactic approach as being the one which uses the rules governing phonemes as well as phoneme sequences in a language, for Arabic dialects identification. They justify their use of phonetics, and more specifically phonotactic modelling, by the fact that the most successful language identification models are based upon phonotactic variation between languages. Based on these phonotactic features. (Biadsy et al., 2009) introduced the Parallel Phone Recognition followed by Language Modelling (PPRLM) model, which is solely based on the acoustic signal of the thirty-second utterances that researchers used as their testing data. The PPRLM makes use of multiple phone recognizers previously trained on one language for each, in order to be able to model the different ADs phonemes. The use of several phone recognizers at the same time enabled the research team to reduce error rates along with the modelling the various Arabic and non-Arabic phonemes that are used in ADs. This system achieved an accuracy rate of 81.60%, and they claim that they found no previous research on PPRLM effectiveness for ADI.

Regardless of whether DID is carried at the speech or text levels, there is a clear trend leading towards finer-grained level by covering an increasing number of Arabic dialects. This makes the identification task harder since dialects become increasingly similar when the geographical distance is reduced, which results in the increase of social contact. Thus, dialects of neighbouring cities become hardly distinguishable even for native speakers, as stated in (Kchaou et al., 2019).

## 3. Corpus description

In this section, we describe the current version of the collected corpus. The first version of our corpus contains 1673 utterances spoken by Tunisian native speakers. It covers 4 varieties of the Tunisian dialect. As previously mentioned, we use the term 'sub-dialect' to refer to these varieties. The corpus contains roughly 45.95 minutes of transcribed speech. Table 1, gives the statistics for each sub-dialect.
We collected our data from online sources. In fact, social media have become a preliminary means of communication. Different people from a variety of social and

| | Duration (mins) | #Utterances | #Sentences |
|---|---|---|---|
| Sfax | 8.37 | 298 | 298 |
| Tunis | 12.94 | 411 | 411 |
| Souse | 13 | 626 | 626 |
| Tataouine | 11.64 | 338 | 338 |
| Total | 45.95 | 1673 | 1673 |

Table 1: Training Data Details

geographical backgrounds are present to spontaneously speak up and express their opinions. For this reason, we chose online data collection method. This method saved us time and effort. We nevertheless encountered some difficulties at this stage, especially for the region of Tataouine and Sousse.

The difficulties in data collection that we had were especially with the region of Tataouine, and to a lesser degree the region of Sousse. This is due to the scarcity of online-available resources like interviews for the former (Tataouine).As for Sousse, the problem is caused by the phenomenon of dialects continuum that exists in the Tunisian coast. Concerning Tunis and Sfax, we did not find much difficulties while searching for primary sources. For instance, the sub-dialect of Tunis is the received dialect of TV and radio programs, especially news broadcasts, and formal interviews and debates. As for the sub-dialect of Sfax, we found a number of subjects originating from Sfax and speaking in their regional dialect.

Speakers of the different regions are from both sexes, with their ages ranging between eighteen and fifty at the time of recording. Raw data is in the form of long WAV files that we segmented, lexically-annotated, and saved into separate utterances using Tanscriber. As summarized in the table, the total duration for each region is the sum of all the preserved segments for each region.

## 3.1. Phonetic analysis

We divide this analysis into two main parts. First, we briefly introduce the vowel system of each region. We do this through a table that we included below. We then introduce some major vocalic patterns. To do this, we compare the four regions' vowels use. In the second part, we present the consonant system of TA in general, and how it is used in the four sub-dialects under study. We finally combine the two parts that make the phonemic system for the sake of presenting the possible syllable structures. We also include some statistics of the sub-dialects under study throughout the section. We also classify the vowels, consonants and syllabic systems according to (Maddieson, 2013) in (Dryer and Haspelmath, 2013).

### 3.1.1. Vowels:

Table 2 presents the different vowels that TA speakers use. We collected this inventory of vowels from the phonetic transcription of our collected data. As is clear from the table, the speakers of the regions under study use fourteen vowels in their speech according to the phonetic transcription that we established. Interestingly, some of the vowels are considered to be allophones at the inter-regions level. An example of this phenomemon is the vowel /i/, /ɛ/ and /æ/ to designate the "I" subject. In fact, "I" is uttered as /ɛni/ in Sousse, /ɛna/ or /ænæ/ in Tunis and Sfax, and /æne/ in Tataouine. At the intra-regional level, phonemes perception changes. Vowels and consonants are the same in all regions, i.e. all Tunisians use the same vowel and consonant systems.Yet, in the same phonetic environment, speakers from the four regions, use different vowels. By this we mean that vocalic differences come to surface in the process of syllable and word formation.

According to (Maddieson, 2013)'s vowel systems categorization of languages, TA is considered to have a large vowel inventory. This is because each of the regions under study has 14 vowels in average, and this is according to the phonetic transcription that we established. The classification of a given dialect or language according to this scale does not consider allophones of the same vowel as a separate phoneme. The nasalized form for example does not make a different vowel. The main distinctive feature in TA in general is the place of articulation. Nasalization, vowel length and aspiration are present in all the sub-dialects under study, but their presence or absence does not make different lexical entries for interlocutors. (Maddieson, 2013) claims that, in general, there is an inverse relationship between consonant and vowel inventories. However, he continues that, by reference to the languages that have large consonant and vowel inventories at the same time, the inverse relationship between consonant and vowel inventories "is not part of a general pattern in languages". Maddieson also highlights that the relationship between these two elementary components of human speech in general by insisting on the consonant-vowel ratio (C/VQ) which is calculated by dividing the number of consonants by the number of vowel qualities. The average vowel quality number for the regions under study is 14, while consonants are of 30. The resulting ratio is 2.14, which is, according to the consonant-vowel ratio scale, considered to be part of the 101 world languages having a moderately low consonant-vowel ratio.

Maddieson claims also that vowels typically occur at the center of syllables. This is not always the case for our TA sub-dialects. These latter share with MSA the possiblity of having vowels at the end of syllables as in مشينَا ( /mʃinæ/), كليتُو (/klitu/) and كرهبَة (/karəhba/).

Differences at the vocalic level arise at the level of phonemes choice across the Studied regions. This means that, for the same token, different vowels are used. An example for his phenomemon is the verb مشينَا which is uttered as /mʃenæ/ in Tataouine, /mʃinæ/ in Sousse, /mʃina/ in Tunis, and typically /mʃaɪna/ or /mʃeɪna/ in Sfax. In general, the main difference between the four sub-dialects is in long vowels choice. Speakers from the four regions use long vowels differently, especially word-medially or word-finally. Words containing long vowels are pronounced differently across sub-dialects. The example mentioned above of the pronoun أنَا in MSA

| Vowels | Region | | | |
|---|---|---|---|---|
| | Tunis | Sousse | Sfax | Tataouine |
| i | ✓ | ✓ | ✓ | ✓ |
| ɪ | ✓ | ✓ | ✓ | ✓ |
| ʌ | ✓ | ✓ | ✓ | ✓ |
| a | ✓ | ✓ | ✓ | ✓ |
| o | ✓ | ✓ | ✓ | ✓ |
| ʊ | ✓ | | ✓ | ✓ |
| u | ✓ | ✓ | ✓ | ✓ |
| ɔ | | ✓ | | ✓ |
| æ | ✓ | ✓ | ✓ | ✓ |
| ɛ | ✓ | ✓ | ✓ | ✓ |
| e | ✓ | ✓ | ✓ | ✓ |
| ə | ✓ | ✓ | ✓ | ✓ |
| eɪ | ✓ | ✓ | ✓ | ✓ |
| aɪ | ✓ | ✓ | ✓ | ✓ |
| aʊ | ✓ | ✓ | ✓ | |
| Proportion vowel/consonant | V: 33.97% | V : 31.50% | V : 31.57% | V : 33.33% |
| | C: 66.03% | C : 68.5% | C : 68.33% | C : 66.66% |

Table 2: Per region vowels distribution and percentages

(which means 'I'), is an illustration of this variation. Given the above and the collected data, we observed a tendency in the sub-dialect of Tataouine to use close-mid and open-mid long vowel forms more than Tunis and Sfax, while the sub-dialect of Sousse makes a mixture of both. The sub-dialect of Tunis combines the upper and lower parts of the vowel quadrilateral, while the sub-dialect of Sfax adds to it the use of diphthongs.

#### 3.1.2. Consonants and Possible Syllable Structures:

In order to determine to what language group the TA sub-dialects under study belong in terms of possible syllable structures, we ultimately have to briefly present the consonant system used in TA. As previously stated, TA's consonantal system encompasses 30 consonants, twenty-seven of which are the same as MSA except for two which are treated as one. The other three consonants are borrowed from other languages. These are /p/, /b/ and /g/. According to (Maddieson, 2013), this number of consonants places the four TA sub-dialects among 94 world languages with moderately-large consonant inventories.

Furthermore, we estimated the proportion of consonants compared to vowels in continuous speech based on the phonetic transcription. We found that consonants in our four sub-dialects are present more than two times as compared to vowels in connected speech. The classification of consonantal and vocalic systems, added to the former proportion helped us have a vision on the quantity of consonant combinations in TA in general, and in the four sub-dialects in our case. Since TA does not use diacritization in the way MSA does, consonant clusters are widely present in the four sub-dialects as compared to vowels. It is a general rule in TA that consonant clusters are permitted word-initially as in the word شبيك (/ʃbik/), word-medially as in تنقلب (/tɪŋglæb/), and word finally as in سيبتش (/sæjjæbtʃ/).

Added to the fact that consonant clusters are allowed in all word positions, TA allows for more than binary consonant clusters. The combination of three consonants is also allowed, like in the word مبلدك (/mæbldək/). The most prominent allowed syllable structures in TA in general are: C, CV, CCV, CCCV, CCCVC. The syllabification system that we opted for is the one that concatenates words uniquely according to the presence of vowels, without any account for the presence of meaning in each syllable. This is because the TA dialect in general is one of the understudied dialects. To our knowledge, the account for syllable patterns is still absent. Another common ground between the sub-dialects under study is the fact that ض (/dˤ/) and ظ (/ð/) are both pronounced as ظ. Words like أبيض (/ʌbjəðˤ/) and يظهرلي (/joðˤhorli/) are examples of the fusion of these two Arabic phonemes into one single phoneme.

Along with the common ground, TA sub-dialects under study have some differences. For the sub-dialect of Tataouine, the /g/ is used most of the times in place of the /q/. The other three sub-dialects use the /q/ in the same phonetic environment as the /g/ phoneme in the sub-dialect of Tataouine. For the latter sub-dialect, the /g/ phoneme is used mainly in verbs like تلڤى (/tælgæ/), and in nouns like ڤبل (/gæbəl/), in both of which which the g becomes q in the three other sub-dialects. Interestingly however, the /g/ phoneme is not always an allophone for /q/. Examples like ڤعود (/gʕuːd/) which means the camel's baby, and ڤرست (/grɪst/) which means 'I am cold', are examples of the treatment of /g/ as an independent phoneme. Substituting /g/ with /q/ will change the meaning for both words. The former's meaning becomes 'sitting' like in MSA (and this word is not used as far as we searched), while the latter loses its meaning, and one has to substitute س with ص in order to obtain قرصت (/qrʌsˤt/) which means 'I tweaked

(somebody)'. Statistically speaking, for 21 verbs uttered in the sub-dialect of Tataouine, the /q/ phoneme is uttered as /g/ in eighteen instances, leaving only three instances for the /q/. As for nouns and adjectives, the native speakers of Tataouine sub-dialect opt for the use of /q/ instead of /g/ in twenty five out of twenty eight instances. The /q/ phoneme is used in 57.14 percent of the instances, while /g/ is used 42.85 percent of the instances. The sub-dialects of Tunis, Sfax and Sousse use /g/ in approximately 5.5 percent.

### 3.2. Corpus distribution

Our corpus will be freely distributed for the research community[2]. This will enable researchers to reproduce our results and, hopefully, to push forward the research in the field of Arabic dialect identification. The corpus will be a package that consists of the audio files along with their corresponding aligned transcripts.

## 4. Learning Features

Our DID task is a multi-class classification task. It consists of 4 classes. We considered different classification algorithms in order to create various identification systems using either the speech signal or its transcripts. For both types of systems we derived a suite of acoustic and lexical features. In this section we describe our feature sets used to build various DID systems.

### 4.1. N-gram features

N-gram features have been used in the vast majority of previous works related to DID and text classification. We extracted Word unigrams as word level features. Character n-grams have also shown to be the most effective in language and dialect identification tasks (Zampieri et al., 2017). We extracted character n-grams ranging from 1- to 3-grams and from 1- to 5-grams. We used Term Frequency-Inverse Document Frequency (Tf-Idf) scores instead of count weights since they are known to perform better.

### 4.2. Spectral features Extraction

As regards the speech identification systems, we extracted the Mel-Frequency Cepstrum Coefficients (MFCC). MFCCs has proven to be effective in modeling the subjective pitch and frequency content of audio signals (Mubarak et al., 2006). In addition, they have been shown to be reliable for speech recognition as well as for speaker identification. MFCC are based on a double Fourier transform or discrete cosine transform of the signal energy. This transformation highlights the harmonic properties of an acoustic signal. We partitioned the speech into frames and computed the cepstral features for each frame. Given that MFCC features describes only the power spectral envelope of a single frame, for richer information about the frames, we also measureed changes in the speech spectrum over multiple frames of speech to model long-term language characteristics using first and second derivatives.

[2]https://github.com/fbougares/Tunisian-Sub-Dialects-Corpus

## 5. Experimental setup

In this section, we present our experimental setup to train speech and text-based DID systems. We also present the classification accuracy using several classification algorithms.

### 5.1. Data, Algorithms and Evaluation metrics

**Data splitting :** The collected corpus presented in section 3. is divided into Train ad Test sets. The splits are balanced for each sub-dialect and the distribution of each split is presented in Table 3. For each sub-dialect, 238 utterances are devoted to training while 60 utterances are kept for Testing. This corresponds to about 18 and 8 minutes of speech in the train and test set respectively.

| | Train set | Test set |
|---|---|---|
| # of minutes | 18 | 8 |
| # of Utterances | 238 | 60 |
| # of Sentences | 238 | 60 |

Table 3: Data splitting.

**ML algorithms:** In the recognition phase, we tested three traditional classifiers. To form these classifiers, we used machine learning algorithms that are suitable for classification tasks. We considered Support Vector Machines (SVM), Naïve Bayes (NB), and a MultiLayer Perceptron (MLP). All of them were applied without any pre-processing.

**Evaluation Metrics:** We report the results of classification using the F1 scores. F1 is calculated according to the weighted F1-score which provides a balance between precision and recall. We also provide the confusion matrix of all sub-dialects for both text- and speech-based DID systems.

### 5.2. Text-based DID system

The first set of experiments are performed using the transcribed data of our corpus without any pre-processing.
As a first step, three classifiers (NB, SVM, MLP) are trained using word uni-grams and character n-grams features separately. Thereafter, they are trained using a various combination of both word uni-grams and character n-grams. Table 4 reports the results of our various systems on the test set. The best F1-score (**54.16** in Table 4) is obtained using SVM classifier trained over word uni-gram and 1→3 character n-gram.

| | **N-Gram Features** | | **F1 score** | | |
|---|---|---|---|---|---|
| | Word | Char | SVM | NB | MLP |
| 1. | 1 | - | 51.66 | 53.75 | 43.33 |
| 2. | - | 1 | 32.06 | 29.16 | 26.66 |
| 3. | - | 1→5 | 52.91 | 50.0 | 31.25 |
| 4. | 1 | 1 | 49.16 | 51.66 | 43.75 |
| 5. | 1 | 1→3 | **54.16** | 52.08 | 36.66 |
| 6. | 1 | 1→5 | 52.5 | 50.41 | 36.66 |

Table 4: F1-scores of text-based DID systems on Test set.

### 5.3. Speech based DID system

In addition to the above text-based DID systems, we trained various speech-based systems using different feature sets. For speech-based systems we decided to use the classifier that performed best on text data: SVM. Indeed, Several SVM models were trained using feature vectors of different sizes extracted from the speech signal using a frame size of 20 seconds duration.

Table 5 reports the results of dialect identification using different dimensional feature vectors. The best result is obtained by a feature vector with a dimension composed of 13 MFCC coefficients and their corresponding delta coefficients.

| **Spectral Features** | **F-1 score** |
|---|---|
| MFCC (13) | 92.08 |
| MFCC (13) + Delta | 93.33 |
| MFCC (13) +Delta +DD | **93.75** |

Table 5: F1-scores of SVM speech-based DID systems on Test set.

As it can be seen from the table above, the speech-based system performed better with MFCC+Delta+DD features. This is due to the richer context information we get using Dela+DD. If we compare the speech-based DID system results against the text-based ones, we notice a significant F-1 score increase. This emphasizes the presupposition that the text-based DID systems reach their limits when they have to deal with dialects belonging to close geographical areas.

## 6. Analysis and Discussion

In this section, we present an analysis of the classification results of our best text- and speech-based DID systems.
The confusion matrix of the text-based DID system is presented in table 6. As we can noticed from this table, the most confused sub-dialects are from Sousse (SOUS) and Tunis (TUN). 16 segments from SOUS class are wrongly predicted as TUN and 13 from TUN wrongly predicted as SOUS. The most confused pair is Sfax (SFX) and Sousse (SOUS): 16 sentences from SFX incorrectly predicted as SOUS and 7 SOUS sentences incorrectly assigned to SFX. As expected, the identification confusions tend to be bigger for close geographic zones. In fact, sentences from close regions has a big vocabulary overlap and therefore they are harder to discriminate using textual features.

| | | Predicted label | | | |
|---|---|---|---|---|---|
| | | SFX | SOUS | TAT | TUN |
| Actual label | SFX | 37 | 16 | 9 | 8 |
| | SOUS | 7 | 27 | 7 | 16 |
| | TAT | 5 | 4 | 38 | 8 |
| | TUN | 11 | 13 | 6 | 28 |

Table 6: Confusion matrix of text-based DID system.

In order to investigate further the reasons behind the high confusability, we conducted a deeper analysis by manually evaluating the mis-classified sentences for each sub-dialect. This evaluation was performed as follows: each of the incorrectly predicted sentences was presented to a native speaker who had to decide whether or not to approve the system's decision. Table 7 presents examples of sentences from the test and their transliteration.

| Sentences | Prediction | Gold |
|---|---|---|
| أرفع شوية بش نّجم آنَا نستعمل سَاقَى<br>'rf' $^{s}wyTb^{s}nnjm'AnAnst'mlsAqY$ | TUN | SFX |
| خذيت اول مغَامرة و بعدت علَى دَارنَا<br>$x_{d}ytawlm.gAmrTwb'dt'lYdArnA$ | TUN | SFX |
| أني مشيت سَافرت لفرَانسَا و<br>قعدت غَاديكَا<br>'ny m$^{s}ytsAfrtlfrAnsAw$<br>q'dt .gAdykA | TUN | SFX |
| تو موزَالك يتلع فهمت<br>tw mwrAlk ytl' fhmt | SFX | TUN |
| سَاعَات نبدَا فرحَانة بروحي برشة<br>سَاعَات نبدَا فرحَانة شوية<br>sA'At nbdA fr.hAnT brw.hy br$^{s}T$<br>sA'At nbdA fr.hAnT $^{s}wyT$ | SFX | TUN |
| تغشت<br>t.g$^{s}t$ | SFX | TUN |
| تَتسبَب و توصل مَمكن<br>tatsbab w tw.sl mamkn | TUN | TAT |
| شمس رمضَان غَاديكَا كل شيء<br>$^{s}msrm.dAn.gAdykAkl^{s}y'$ | TUN | TAT |
| فمَا خيَام فَارغة وهمية<br>و لَا فمَا أعتصَام<br>fmA xyAm fAr.gT whmyT<br>w lA fmA ''t.sAm | TUN | TAT |
| تبدَا يَاسر قوية عليه<br>tbdA yAsr qwyT 'lyh | SFX | SOUS |
| جَا في مخي طول فَازة<br>gA fy mxy .twl fAzT | SFX | SOUS |
| من عند بوك و ألَا أمك<br>mn 'nd bwk w 'lA 'mk | SFX | SOUS |

Table 7: Examples of text-based classification errors from test set and their ground truth label.

The evaluation has shown that almost all the studied examples may belongs to both dialect and are hardly distinguishable even for native speakers. This exemplifies the increasing complexity the DID task in written text when we consider close dialects having a large lexical overlap. As with the text-based DID system, we also analyzed the confusion matrix of our speech-based DID system.

As is highlighted in table 8, the identification confusions are much lower even for close geographic zone. The speech-based system almost completely corrected the dialect prediction of utterances for which text-based system fails.

| | | Predicted | | | |
|---|---|---|---|---|---|
| | | SFX | SOUS | TAT | TUN |
| Actual | SFX | 51 | 3 | 1 | 0 |
| | SOUS | 9 | 57 | 1 | 1 |
| | TAT | 0 | 0 | 58 | 0 |
| | TUN | 0 | 0 | 0 | 59 |

Table 8: Confusion matrix of speech-based DID system.

## 7. Conclusion

In this paper we wanted to show whether conventional DID systems are able to distinguish Arabic sub-dialects at a narrow geographical scale or not. To achieve this goal, a corpus of 4 varieties of spoken sub-dialects from north to south of the Tunisian country was created. We used this spoken corpus and its manual transcription to carry out experiments dedicated to the build text- and speech-based dialect identification at this fine-grained level. Our experiences have shown that speech-based DID systems outperform text-based systems since the latter has to deal with close sub-dialects with highly overlapping vocabularies. We also presented a small phonetic description of the sub-dialects at the vocalic, consonantal and syllabic levels despite the challenge of having a limited corpus. This study confirms the native speaker statements regarding the higher suitability of speech to better discriminate closely related dialects. This is our first work towards drawing the dialectal map of Tunisia, and to establish a well-comprehensive linguistic description of the TA dialect as a whole, and highlight the variation between the different sub-dialects spoken across the different cities. Therefore, this work will be pursued by exploring the dialectal characteristics of additional Tunisian cities. We also plan to enlarge the data for the current cities in order to draw solid descriptions of the different Tunisian sub-dialects, and then use the results to build systems that rely more on linguistic rules including phonotactics and prosody.

## 8. Bibliographical References


Abdul-Mageed, M., Alhuzali, H., and Elaraby, M. (2018). You tweet what you speak: A city-level dataset of Arabic dialects. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Barkat, M. (1999). Identification of arabic dialects and experimental determination of distinctive cues.

Biadsy, F., Hirschberg, J., and Habash, N. (2009). Spoken Arabic dialect identification using phonotactic modeling. In *Proceedings of the EACL 2009 Workshop on Computational Approaches to Semitic Languages*, pages 53–61, Athens, Greece. Association for Computational Linguistics.

Bougrinea, S., Cherrouna, H., and Ziadib, D. (2017). Hierarchical classification for spoken arabic dialect identification using prosody: Case of algerian dialects. In *arXiv:1703.10065v1 [cs.CL]*, March.

Djellab, M., Amrouche, A., Bouridane, A., and Mehallegue, N. (2016). Algerian Modern Colloquial Arabic Speech Corpus (AMCASC): regional accents recognition within complex socio-linguistic environments. In *Language Resources and Evaluation 1â29.*

Matthew S. Dryer et al., editors. (2013). *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

El-Haj, M., Rayson, P., and Aboelezz, M. (2018). Arabic dialect identification in the context of bivalency and code-switching. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Kchaou, S., Bougares, F., and Hadrich-Belguith, L. (2019). LIUM-MIRACL participation in the MADAR Arabic dialect identification shared task. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 219–223, Florence, Italy, August. Association for Computational Linguistics.

Maddieson, I. (2013). Consonant inventories. In Matthew S. Dryer et al., editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Malmasi, S. and Zampieri, M. (2016). Arabic dialect identification in speech transcripts. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects*, page 106â113, Osaka, Japan, December.

Mubarak, O., rajah, E. A., and Epps, J. (2006). Novel features for effective speech and music discrimination. In *IEEE Engineering on Intelligent Systems*, pages 342–346.

Sadat, F., Kazemi, F., and Farzindar, A. (2014). Automatic identification of Arabic language varieties and dialects in social media. In *Proceedings of the Second Workshop on Natural Language Processing for Social Media (SocialNLP)*, pages 22–27, Dublin, Ireland, August. Association for Computational Linguistics and Dublin City University.

Salameh, M., Bouamor, H., and Habash, N. (2018). Fine-grained Arabic dialect identification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1332–1344, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Shon, S., Ali, A., and Glass, J. (2017). MIT-QCRI Arabic dialect identification system for the 2017 multi-genre broadcast challenge . In *Automatic Speech Recognition and Understanding Workshop (ASRU), 2017*.

Zaidan, O. F. and Callison-Burch, C. (2014). Arabic dialect identification. *Comput. Linguist.*, 40(1), March.

Zampieri, M., Malmasi, S., LjubeÅ¡iÄ, N., Nakov, P., Ali, A., Tiedemann, J., Scherrer, Y., and Aepli, N. (2017). Findings of the vardial evaluation campaign 2017. In *In Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, Valencia, Spain.