Automatic speech recognition system for Tunisian dialect

Abir Masmoudi, Fethi Bougares, Mariem Ellouze, Yannick Estève, Lamia Belguith

Automatic Speech Recognition System for Tunisian Dialect

Abir Masmoudi, Fethi Bougares, Mariem Ellouze, Yannick Estève, Lamia Belguith

Noname manuscript No.
(will be inserted by the editor)

Automatic Speech Recognition System for Tunisian Dialect

Abir Masmoudi · Fethi Bougares · Mariem Ellouze · Yannick Esteve · Lamia Belguith

the date of receipt and acceptance should be inserted later

Abstract Although Modern Standard Arabic is taught in schools and used in written communication and TV/radio broadcasts, all informal communication is typically carried out in dialectal Arabic. In this work, we focus on the design of speech tools and resources required for the development of an Automatic Speech Recognition system for the Tunisian dialect. The development of such a system faces the challenges of the lack of annotated resources and tools, apart from the lack of standardization at all linguistic levels (phonological, morphological, syntactic and lexical) together with the mispronunciation dictionary needed for ASR development. In this paper, we present a historical overview of the Tunisian dialect and its linguistic characteristics. We also describe and evaluate our rule-based phonetic tool. Next, we go deeper into the details of Tunisian dialect corpus creation. This corpus is finally approved and used to build the first ASR system for Tunisian dialect with a Word Error Rate of 22.6%.

Keywords Under-resourced Language · Rule-based Grapheme-to-phoneme conversion · Automatic speech recognition · Tunisian dialect

A. Masmoudi
LIUM, Le Mans University, France - ANLP Research group, MIRACL Lab., University of Sfax, Tunisia
E-mail: masmoudiabir@gmail.com

F. Bougares
LIUM, Le Mans University, France
E-mail: fethi.bougares@univ-lemans.fr

M. Ellouze
ANLP Research group, MIRACL Lab., University of Sfax, Tunisia
E-mail: Mariem.Ellouze@planet.tn

Y. Estève
LIUM, Le Mans University, France
E-mail: yannick.esteve@univ-lemans.fr

L. Belguith
ANLP Research group, MIRACL Lab., University of Sfax, Tunisia
E-mail: l.belguith@fsegs.rnu.tn

1 Introduction

Arabic is a Semitic language among the oldest in the world. The evolution of this language from antiquity to the present day has given birth to several linguistic registers; some forms have disappeared and others are still spoken. In accordance with the great periods of the Arabic language history, four linguistic registers can be cited: **(1)** Old Arabic, which is not used currently; it is found in ancient literary works (mainly poems); **(2)** Classical Arabic or literary Arabic, which is the language of Islams Holy Book; spreaded through Islamic conquests; **(3)** Modern Standard Arabic (MSA) is the official language of all the Arab countries and it has extensive presence in various settings: education, business, arts and literature, and official and legal written documents; **(4)** dialectal Arabic, which is the mother tongue of every Arabic speaker and the main variety used in daily life for spoken communication. There are many Arabic dialects across the Arab World.

The dialects are not standardized, they are not taught, and they do not have an official status. Indeed, most native speakers of Arabic are unable to produce a sustained spontaneous discourse in MSA; in unscripted situations where spoken MSA would normally be required (such as talk shows on TV); speakers usually resort to repeated code-switching between their dialect and MSA [1].

In the Natural Language Processing (NLP) context, MSA has received the lions share of attention. There are lots of parallel and monolingual data collections, richly annotated collections (e.g., treebanks[1]), sophisticated tools for morphological analysis and disambiguation, syntactic parsing, etc.[22]. Unfortunately, a large number of these MSA tools cannot effectively model Arabic dialects because the latter diverge considerably from MSA and differ widely among themselves. Indeed, these differences make the direct use of MSA NLP tools and applications for handling dialects impractical. Therefore, it is necessary to construct dedicated tools and resources to process dialectal Arabic.

Tunisian dialect is a subset of the Arabic dialects usually associated with the Arabic of the Maghreb. In this work, we aim to develop speech processing tools and resources for the purpose of creating an ASR system in the framework of human/human conversation between customers and agents of the Tunisian Railway Transport Network. Despite the significant progress and the considerable effort recently made to study this dialect, much work remains to be done. On the other hand, the development of such tools faces many challenges: **(1)** The limited number and the non-availability of dedicated resources. **(2)** The lack of standard orthographies for this dialect. Therefore, the task of resources creation is more complicated because there is no consensus about the

[1] Treebanks are language resources that provide annotations of natural languages at various levels of structure: at the word level and the sentence level.

transcription conventions. **(3)** The impossibility to apply phonetic transcriptions techniques dedicated to MSA, since the Tunisian dialect is characterized by the addition of other supplementary phonemes. **(4)** The irregularities relative to the lack of correspondence between the spelling of the lexical items and their sound contents.

The main contributions of this paper are as follows:

- Presenting the characteristics of the Tunisian dialect at different levels *i.e.*, lexical, morphological, phonological and syntactic.
- Designing an accurate approach for G2P conversion of the Tunisian dialect in order to develop a phonetic dictionary. The latter is an important ASR component that serves as an intermediary between acoustic models and language models in ASR systems [34].
- Describing the creation process of our Tunisian dialect corpus and giving the most important characteristics of this corpus.
- Developing and analyzing a dedicated ASR system for the Tunisian Railway Transport Network.

The remainder of this paper is organized as follows. Section **2**, briefly overviews earlier research in automatic speech recognition for under-resourced languages. Section **3** sets a historical review of the Tunisian dialect and presents its linguistic characteristics. Section **4** describes the variety of published G2P conversion approaches and presents our methodology to develop G2P rule-based approach for Tunisian dialect. We present, discuss and analyse the experimental results of the proposed method as well. In section **5**, we explain how we proceeded to create our manually transcribed Tunisian dialect corpus. In this section we describe the steps taken in order to implement and analyse the results of our ASR system in the Tunisian Railway Transport Network domain. Finally, we conclude and present future work in section **6**.

2 Background and Related Works

In order to train automatic speech recognition (ASR) systems very large amounts of speech and textual data are generally required. Such spoken language resources are available for a large variety of languages such as English and some European languages. However, these resources are still unavailable or very limited for other languages. The latter are generally defined as under-resourced languages. It should be noted that under-resourced languages include languages spoken by a minority of the population such as Breton, Castilian, Catalan... or spoken languages in regions of underdeveloped countries. They include also Arabic dialects that are considered as the first language of several Arab countries and used for everyday communication.

Recently we have seen an increasing interest in Automatic Speech Recognition (ASR) for under-resourced languages. Several systems were presented and discussed over the last couple of years. These systems are different in terms of their origins and in the techniques used to solve the problem of lack of resources.

For instance, the earliest works on ASR for Vietnamese and Khmer low-resourced languages were proposed by [7]. Afterward [50] used a method for a rapid adaptation of an automatic speech recognition system for Vietnamese. Another interesting work was carried out for the Khmer language [46]. In this work, the problem of the lack of data was addressed. Authors crawled the web to extract a large number of khmer documents. These documents were pre-processed and used for language model training and a part of them was uttered by several speakers in order to constitute training data for acoustic models.

Other researchers, proposed different works for African languages: Somali [37], Amharic [39], South African languages [6] and West African languages [44]. For more detailed information, readers may refer to the survey on ASR for under-resourced languages presented in [8].

With respect to the Arabic dialects, a large gap is still present between them and the MSA. Indeed, MSA is characterized by a large amount of resources to develop an ASR system in contrast to the Arabic dialects, which suffer from lack of resources. In the Arabic dialect context, we point to the earlier work on the Qatari dialect [14]. Although the resources for the Qatari dialect are derisory, it is possible to benefit from the MSA resources: spoken and textual corpora. In this respect, authors proposed to add MSA corpora to the small specific dialectal data. Adding MSA data enhance the acoustic modeling of the Qatari dialect [14].

3 The Tunisian dialect

3.1 Historical Overview of the Tunisian dialect

The Tunisian dialect is considered as the current mother tongue shared by all Tunisians independently of their origin and their social belonging. Therefore, this dialect occupies an important linguistic situation in Tunisia. Historically, Berber was the original mother tongue of the North Africa inhabitants. Later on, the Arab invasions and the Islamic conquests introduced the Arabic language which is the language of Islams Holy Book. The prolonged Ottoman Turkish political domination of North Africa, roughly from the mid-fifteenth to the late nineteenth century, and the French colonization, from 1881 to 1956, had an impact on the absorption of foreign vocabulary into the lexicon of the local Arabic dialect. In addition to Turkish and French, we find plentiful examples of lexical elements from the European languages in the Tunisian dialect. We can identify a significant number of expressions and words of Spanish and

Italian origin, and even of Maltese origins.

In this regard, the interactions between the Tunisian dialect and these foreign languages affect the different levels of the spoken language, *i.e* phonology, morphology, lexical and syntax, which have rendered the linguistic situation in Tunisia rather complex. The linguistic situation in Tunisia is described in [30] as poly-glossic where multiple languages and language varieties coexist. In the following sub-sections, we present some characteristics of the Tunisian Dialect at different levels.

3.2 Linguistic characteristics

3.2.1 Phonological level

There are many substantial phonological differences between Tunisian dialect and MSA [36,52]. This includes the introduction of some foreign phonemes like /P/, /V/ and /G/ and the neglection of short vowels when they are located at the end of a syllable. Furthermore, we can also quote the pronunciation variation of some consonants. For example, the consonant س /s/ [s] can be pronounced as س /s/ [s] or ص /S/ [S], as in the example of the word رَسول /raswl/ [Prophet] is pronounced رَسول /raswl/ or رَصول /raSwl/.

3.2.2 Morphologic level

Compared to MSA, the Tunisian Dialect shows many divergences at the morphological level. The overarching themes are those of simplifying inflections and introducing new clitics [52]. The Tunisian Dialect has lost the feminine singular and the feminine plural in the conjugation of verbs. Similarly, there is no difference between the feminine and the masculine plural in the Tunisian Dialect. In addition, there is no longer any difference between the conjugation of the first and second person singular. In order to indicate duality, the Tunisian uses a numeral word زوز /zwz/ [two] before or after the plural noun. New clitics are also introduced such as negation clitic شِي /Siy/ + verb + مَا /mA/ which are manifested in MSA with different particles ie مَا /mA/, لَا /lA/, لَن /lan/ and لَم /lam/.

3.2.3 Syntactic level

Like other Arabic dialects, the Tunisian dialect respects in most cases the regular grammar of MSA or that of Classical Arabic [20]. Yet it has some syntactic particularities. Indeed, the order of syntactic constituents in the sentence seems to relatively be more important. The canonical order of words in

Tunisian dialect verbal sentence is SVO [Subject Verb Object] [5]. In MSA, the word order can have the following three forms (SVO, VSO and VOS)[5].

Among the Tunisian dialect syntactic particularities, we can also cite the reduction of the number and the form of MSA pronouns [36] from twelve to seven personal pronouns. Moreover, we have reported the disappearance of the duel of the second person, the feminine plural of the second person and the feminine plural of the third person.

3.2.4 Lexical level

As mentioned above, the Tunisian dialect is an outcome of the interactions between Berber, Arabic and many other languages. The outcome of this interaction is manifested in the introduction of borrowed words in the Tunisian daily communication. Examples include, but are not limited to: سَبَاط /sabAT/ [shoes] from Spanish and فيشطَة /fySTap/ [holiday] from Italian. Some foreign words undergo an addition of the Arabic enclitic or proclitic to express either an action or an order or possession of a thing etc. An example of this, the French verb /supprimer/ [delete] which undergoes generally with the addition of the Arabic enclitic : the Arabic enclitic هَا /hA/ [her] is attached at the end of the verb to indicate an order سيبرِميهَا /syprpyhA/ [delete it].

3.3 CODA writing convention

CODA is a conventionalized orthography for dialectal Arabic [21]. In CODA, each word has a single orthographic representation. CODA has five key properties. First, CODA is an internally consistent and coherent convention for writing dialects. Second, it is primarily created for computational purposes, but it is easy to be learnt and recognized by educated Arabic speakers. Third, it uses the Arabic script as used for MSA, with no extra symbols from (i.e Persian or Urdu). Fourth, it is intended as a unified framework for writing all dialects. CODA has been defined for Egyptian [21] and was extended to the Tunisian dialect [52] (some authors of our paper were involved in this work). Finally, it aims to maintain a level of dialectal uniqueness while using conventions based on similarities between MSA and the dialects. For a full presentation of CODA and a justification and explanation of its choices, see [21], [42] and [52].

4 G2P conversion of the Tunisian dialect

Grapheme-to-Phoneme (G2P) conversion is the task of converting a sequence of graphemes (written letters) to a sequence of phonetic symbols (representation of speech sounds). This conversion is usually generated using an automatic G2P conversion tool [10]. It has an application in many areas such as speech

recognition and speech synthesis systems. A fairly large number of researches have been conducted on phonetic conversion. Proposed techniques could be classified into three categories briefly described in the following sub-sections.

4.1 G2P approaches

4.1.1 Look-up approach

This approach involves a manually created phonetic dictionary. A good example of this approach is the CMU dictionary [2] for North American English, which has been built by human experts (linguists) and contains over 125,000 words and their transcriptions. Such a resource is particularly useful for speech recognition and synthesis, as it has mappings from words to their pronunciations into a given phoneme set. Despite its good quality and usefulness for different NLP applications, creating phonetic dictionaries manually is a very expensive and tedious task. To overcome the limitations of this approach, the rule-based and data-driven approaches were proposed as alternatives.

4.1.2 Rule-based approach

This type of approach uses a comprehensive set of G2P rules and a lexicon of exceptions. The development of the G2P rules requires a highly qualified phonetician. For morphologically rich languages, such as MSA, the manual creation of pronunciation dictionaries is a painful task because of the large number of word forms, each of which has a large number of possible pronunciations.

Fortunately for MSA-language, the relationship between orthography and pronunciation is relatively regular and well understood [9]. More recently, most MSA speech recognition systems are trained using pronunciation dictionary constructed by mapping every undiacritized [3] word to all its diacritized versions. Morphological disambiguation obtained using tools such as MADA[4] are used to determine the most likely pronunciation of the word given its context. A typical example of this approach is presented in [49] where morphological information is used to predict word pronunciation.

In the context of Arabic dialect G2P conversion, [25] used the rule-based approach in order to generate the pronunciation of Algerian dialect[5]. To avoid ambiguity caused by the absence of diacritics, they first added diacritic marks to the texts using ADAD (Automatic Diacritizer of Algerian dialect), a system for automatic diacritization. Then, they applied several manually defined

[2] http://www.speech.cs.cmu.edu/cgi-bin/cmudict

[3] Undiacritized or unvowelized word refers to a word without short vowels.

[4] MADA is a POS tagger for Arabic languages

[5] The Algerian dialect is the language used in the daily spoken communication of Algerian.

phonetic rules to perform the G2P transformation. The techniques described above show that the rule-based approach provides decent quality pronunciation dictionaries without requiring a great amount of data. However, this approach requires specific linguistic skills to design the rules.

4.1.3 The data-driven approach

In contrast to the rule-based approach outlined above, the G2P data-driven approach is example based. Hence, this approach make uses of a manually annotated corpus rather than a linguistic one. These examples are used to train a machine learning classifier to predict the pronunciation of unseen words [10]. In the literature, several machine learning algorithms have been used to train data-driven G2P conversion including pronunciation by analogy, local classification: decision trees [24], Hidden Markov Models (HMMs)[47], Conditional Random Fields (CRFs) [26] and Joint Sequence Model (JMM)[10].

4.2 Methodology of G2P conversion of the Tunisian dialect

Pronunciation dictionaries map words to one or more pronunciation variants, taking account of pronunciation variability. In our study, we have chosen to use the rule-based approach. The reasons for this choice were threefold: first, the lack of hand-crafted corpus for Tunisian Dialect needed to train a G2P classifier; second the results from previous studies show the effectiveness of rule based approach and give a pronunciation dictionary of good quality; and third the similarity between Tunisian dialect and MSA, gives us the possibility to benefit from the well defined phonetic rules of MSA [13] and build on it.

4.2.1 Format of phonetic rules

We have identified a set of pronunciation rules. The development and the conception of these rules were carried out collaboratively with linguistic and phonetic experts. Rules are provided for each letter in the Tunisian dialect. Each rule tries to match certain conditions relative to the right and left context of the letter and provide subsequently a replacement. Each rule is read from right to left by following this format:

$$Phonetisation \leftarrow \{Left\,Condition\} + \{Grapheme\} + \{Right\,Condition\}$$

- **Right-Condition** context before the current position "*Grapheme*".
- **Grapheme**: is the current letter in the word.
- **Left-Condition** context after the current position "*Grapheme*".
- **Phonetisation**: is either a phoneme or more of a phoneme or a vacuum if the "*Grapheme*" is omitted in pronunciation.

4.2.2 Lexicon of exceptions

The Tunisian dialect spelling is fairly regular, except in certain situations where some words do not comply with the regular pronunciation rules. It is therefore necessary to define a lexicon of exceptions (93 exceptions of unvowelized words and 90 exceptions of vowelized words). The phonetic form of each exceptional word is entered with its graphemic form. The exceptions dictionary is scanned first. When the word does not exist in that dictionary, the pronunciation rules are applied in order to generate the corresponding phonetic form. The exceptions lexicon is subdivided into four predominant categories. The first category includes 10 dialecatel demonstrative pronouns of the Tunisian dialect such as هذَا /h*A/ [This male] or هذي /h*y/ [this female]. The second one contains some exceptional names like the names of ALLAH الّله /Allah/ [GOD]. we recorded 6 exceptional names. The third category contains 9 elements representing the personal pronouns of the Tunisian dialect like أَنَا /<nA/ [I]. The last category gathers about 65 various exceptions which do not belong to any previous category. By doing this, we could solve some problems related to the phonetic and phonological variations in the Tunisian dialect [34].

4.2.3 Application of phonetic rules

This section highlights the phonetic rules application strategy in order to generate the phonetic form of vowelized and no-vowelized words. It is worth noting that no-vowelized category also includes the partially vowelized words.

Vowelized words

In this case the application of the phonetic rules straightforward. Rules are applied in the word reading direction (from right to left). This begins with the first letter of each word whilst taking into account the order and the context of each letter.

No-vowelized words

As it is the case for MSA, vowels are generally omitted in Tunisian dialect written texts. Therefore, the determination of the phonetic form of a given word is ambiguous. It is thus quite natural that G2P conversion becomes more difficult compared to vowelized words. This ambiguity was targeted for MSA and several solutions have been proposed in the NLP community to over-come the lack of vowels. For example, [3] proposed an automatic restoration of arabic diacritics. However, the proposed method is specific for undiacritized MSA

text and cannot handle the dialects given the morphological, syntactic, lexical and phonetic differences.

In order to get the exact pronunciation of words unvowelized many challenges must be addressed. Firstly, the primary challenge consists in the determination of the phonemes corresponding to the short vowels location. In this regard, we have agreed to put short vowels during the manual transcription. Then, according to the phonetic study that we have performed in this corpus, we have identified these three phonetisation principles:

1. Tunisian dialect words end with either a silent consonant (with Sukun) or a long vowel.
2. Each long vowel is always followed and preceded by a silent consonant (without short vowels).
3. Tunisian dialect words cannot have two successive consonants that carry a vowel (long or short) with the exception of words with "Shadda" (Doubling of consonants).

Based on these previous principles and the pronunciation rules that we presented above, we were able to distinguish four classes of G2P conversion of no-vowelized words:

1. First class contains words with long vowels.
2. Second class contains words with Shadda.
3. Third one includes words with long vowels and Shadda together.
4. The last class integrates simple words without Shadda and long vowels.

Unlike the case of vowelized words, the application of phonetic rules to no-vowelized words does not follow the reading direction of the word. Indeed, the phonetic rules for no-vowelized words are divided into two groups: support rules and secondary rules.

As a first step, the support rules are applied. This produces the graphemes and the long vowels phonemes and facilitates the application of secondary rules. Secondary rules generate the production of the short vowels, Tanween and "Sukun" phonemes [34]. Support rules are applied in the following order:

1. The grapheme rules and foreign-letter rules;
2. Long-vowel rules;
3. Rules of the Ya-Maqsoura;
4. Gemination rule;
5. Ligature rules;
6. Elision rules;
7. Shamsi and Ghamari rules.

Once the support rules is applied and based on the three phonetisation principles identified above, the secondary phonetic rules are applied in the following sequence: first short-vowel and sukun rules, then tanween rules.

Table 1 shows an example of the phonetisation process of no-vowelized word. First, the support rules are applied (phase 1 in table 1) and produce the phonemes: "KH R JH UW". Afterwards, the secondary rules are applied (phase 2 in table 1) in order to produce the phonetisation results of the last column of the table.

Table 1 The G2P conversion process of the word خرجوا. /xrjwA/ [they go out]

Word	Phonetisation Rules & comments	Phonetisation
خرجوا	**Phase 1: Application of support rules**	KH AE R JH UW KH IH R J UW KH UH R J UW
	- By applying graphemes rules we obtain the phoneme "KH" which corresponds to the grapheme خ, the phoneme "R" which corresponds to the grapheme ر, and the phoneme "JH" which corresponds to the grapheme ج. -The grapheme و /w/ is a part of وَاو الجمَاعة. Thus, the long vowel rule must be applied to obtain the "UW" phoneme directly.	
	Phase 2: Application of secondary rules	
	-Based on the principle that each long vowel is always followed and preceded by a mute consonant, in this case, the insertion of phoneme of the short vowels "AE" or "IH" or "UH" must be only between the first phoneme "KH" and the second phoneme "R".	

4.2.4 Evaluation and discussion

In order to evaluate our G2P conversion tool, an expert-annotated Tunisian dialect corpus is needed. This kind of manually annotated corpus does not exist for the Tunisian dialect. To cope with this situation, we create our evaluation set from the three corpus described in table 2.

Table 2 Corpus details used in G2P experiments. 4k words are randomly selected to create the evaluation set

Corpus	Description	Size (#words)
(1) Blogs corpus	texts collected from Tunisian blogs of various fields (politics, sports, culture, science ...)	21917
(2) Arabizi conv corpus	automatic converted texts originally written in Latin script (called Arabizi)	8057
(3)TARIC	Tunisian Arabic Railway Interaction Corpus	4648
(4) Evaluation set	random selection from (1), (2) and (3)	4083

As a first step, these corpora are standardized using the CODA convention [52], then about 4,000 different words are randomly selected in order to

build the G2P evaluation set. This evaluation subset is then automatically phonetized using our rule based G2P system. The error rate is calculated as a function of phonemes, by estimating the following errors in the hypothesis compared to the reference:

- **Substitutions** (Sub): misrecognized phonemes.
- **Insertions** (Del): represent extra phonemes.
- **Deletions** (Ins): represent unrecognized phonemes.

The evaluation was performed based on two distinct levels using Sclite[6], according to word and phoneme levels. At the word level the generated phonetic form of a given word is considered correct when all its phonemes are correctly generated. The Final score is calculated according to the following formula :

$$\text{WER/PER} = (\text{Sub} + \text{Del} + \text{Ins}) \,/\, \text{N} \text{ (where N is the reference size)}$$

Table 3 Word/Phone Error Rate of the rule based G2P on the evaluation set (row (4) table 2)

	Words (WER)		**Phonemes (PER)**	
	Vowelized	**Unvowelized**	**Vowelized**	**Unvowelized**
Substitutions	0%	0.3%	0%	0.1%
Insertions	0%	0.3%	0%	0.2%
Deletions	0%	0.2%	0%	0.1%
Overall Error Rate	0%	0.8%	0%	0.4%

Table 3 reports the obtained scores using our G2P conversion system on the evaluation set. As reported in the table, the G2P conversion system always predicts the correct phonemes when the test words are vowelized. Otherwise, when words are non-vowelized the G2P conversion system predicts the phonetic form with an accuracy of 99.6% and 99.2% at the phoneme and word levels respectively.

A list of erroneous words was compiled, in order to determine the source of errors made with the no-vowelized G2P conversion. It turned out that the errors are caused by the presence of foreign words, irregular words and some proper names. The rule-based approach adopted in this work is unable to generate the correct G2P conversion of the foreign words used in the Tunisian dialect and pronounced the same way as in the original language. One possible solution to deal with foreign words would be statistical approaches. Indeed, with statistical approaches we can train the G2P including foreign phonemes.

[6] http://www1.icsi.berkeley.edu/Speech/docs/sctk-1.2/sclite.htm

5 Tunisian dialect speech recognition system

5.1 Tunisian Arabic Railway Interaction Corpus (TARIC)

The quality and number of NLP tools available for a given language are usually correlated with the amount of annotated resources (transcribed speech corpora, phonetic dictionaries, vocabularies, text corpora) for that language. Although these resources are limited for the under-resourced languages, they are still needed to develop tools such as ASR systems. Tunisian dialect is an example of this so called under-resourced languages with limited amount of available annotated resources.

In this context, we took the initiative to create our own spoken dialogues corpus. This corpus is gathered in collaboration with the Tunisian National Railway Company (SNCFT[7]) and corresponds to recordings of information request in railway stations. This corpus, called TARIC (**T**unisian **A**rabic **R**ailway **I**nteraction **C**orpus), contains information requests in the Tunisian dialect about the railway services such as train type, train schedule, train destination, train path, ticket price and ticket booking. TARIC corpus is created in three stages : first, audio recording; second the orthographic transcription, and finally the transcription normalization. The following sub-sections give more details about these stages.

5.1.1 Speech data collection

This first stage consists in speech audio recordings. This task was performed in the ticket offices at the railway station in Tunis. We recorded conversations in which there was a request for information about train schedules, fares, bookings, etc. The equipment we used included two portable PCs using the Audacity software[8] and two microphones, one for the ticket office clerk and the other for the client. The micro of the client is installed invisibly in order to collect spontaneous speech using the daily form of the Tunisian dialect. The recording was carried out in full respect of the laws and regulations in force and the internal procedures of the SNCFT.

In order to increase the vocabulary size, we recorded at various periods, particularly holidays, weekends, festival days, and sometimes during the week. This strategy leads us to create a heterogeneous corpus with a rich vocabulary, given that fares and schedules vary in holiday periods compared to normal days requests. Finally, we recorded also in the promotion days and on festival days. At the end of this stage, we came up with 20 hours of audio recording. We note, however, that we opt for the re-recording solution when the original session was very noisy (i.e. some sentences were re-recording by another person for a better acoustic conditions).

[7] http://www.sncft.com.tn/
[8] https://sourceforge.net/projects/audacity/

5.1.2 Private life policies

Since the speech corpus collection is based on real conversations between the ticket office clerk and the client, personal data of the client may be present in these recordings. These data can be references to places or human characteristics such as the national identity cards numbers, the work card numbers or the fidelity card numbers. Consequently, the presence of such personal data in a corpus engenders a lack of respect for people's privacy. To prevent this, we have manually identified all passages of the corpus including personal information about the speaker "client". Thereafter, these passages are deleted. Doing so ensures that we fulfill our commitment with the SNCFT in addition to compliance with local applicable laws.

5.2 Orthographic transcriptions

The transcription is the first level of a symbolic representation of an oral communication. It is an essential step in the constitution of the TARIC corpus. The goodness of the transcription is crucial and it has repercussions for all posterior treatments. Yet, the transcription in itself is not a trivial task. This task has been even more difficult since no transcription conventions are available for the Tunisian dialect. For this reason, we have developed first our own orthographic guidelines called CODA [52] (see section 3.3 for more details), which was later adopted during the annotation of our speech corpus.

In practical terms, we used *Transcriber*[9] to perform the alignment of transcription and audio data. Using the 20 hours described in previous sections, we ended up creating a corpus of roughly 10 hours of transcribed speech. This corpus is then divided into three parts: training, development and testing sets. The main characteristics of the three sets are shown in Table 3.

Table 4 Main characteristics of the TARIC training, development and test sets.

	# of hours	**# of statements**	**Vocabulary size (different words)**
Train	8 Hours and 57 Minutes	18027	3027
Dev	33 Minutes and 40 Seconds	1052	612
Test	43 Minutes and 14 Seconds	2023	1009

The TARIC corpus contains more than 21k statements. The corpus contains 82k running words with a vocabulary of 3207 words. As presented in Table 4, about 9.5 hours of the transcribed data are devoted to both training and development while the 43 remaining minutes composed the test set.

Table 5 reports the speakers distribution over train, dev and test set. Overall, we recorded 108 different speakers 60 males and 37 females. Among these

[9] Transcriber is distributed as free software and is available at http://trans.sourceforge.net

Table 5 Speakers distribution over train, dev and test set

	Number of speaker		#unseen speaker (duration)
	Male	Female	
Train	60	37	-
Dev	3	8	7 (15min 19Sec)
Test	1	5	4 (24 min 15 Sec)

speakers, 11 are on dev and 6 on test. We wish to point out that development and test sets include 7 and 4 unseen speakers with roughly 15 minutes and 24 minutes respectively.

5.3 Corpus distribution

TARIC will be the first freely distributed transcribed speech corpora for Tunisian dialect. This will be distributed as a package available for research purposes[10]. This package will be constituted of the audio recordings along with their corresponding aligned transcripts and the pronunciation dictionary. We will provide also the language model and the monolingual data used to train it. Finally we plan to continuously enhance the corpus with data acquired from other domains (mainly from broadcast news) with more speakers having different accents and different speaking styles.

5.4 Tunisian dialect ASR system

In this section we describe our ASR system for the Tunisian dialect built using the TARIC corpus. This ASR system is developed using the KALDI speech recognition toolkit [40]. KALDI toolkit is very actively developed over the past few years and is widely used by the research community. Moreover, several recipes for training ASR systems with various speech corpora have been made available and frequently updated to include the newest techniques developed in the community. In the following subsections, we describe the main design characteristics of our Tunisian dialect ASR system.

5.4.1 Acoustic Modeling

Acoustic model is trained using PLP (Perceptual linear predictive) features. For each frame, we also include its neighboring ± 4 frames and apply Linear Discriminative Analysis (LDA) transformation to project the concatenated frames to 40 dimensions, followed by Maximum Likelihood Linear Transform (MLLT). We also applied a speaker adaptation through feature-space Maximum Likelihood Linear Regression (fMLLR) technique. These models are all standard 3-states context-dependent triphone models. The GMM-HMM model has about 15K Gaussians for 2.5K tied states.

[10] `http://www-lium.univ-lemans.fr/~bougares/ressources.php`

5.4.2 Language Model and lexicon

The language model (LM) was built using TARIC training data transcripts and corpora described in table 2 (section 4.2.4). A total corpus of 100K words is obtained and normalized following the CODA recommendations. We trained different 3-gram language models with modified Kneser-Ney discounting. These models were interpolated to create the final LM by minimizing the perplexity on the development corpus. Given the limited size of our training data, we did not apply neither cut-offs nor pruning to the final language model. Moreover, we would point out we deliberately chose to avoid out-of-vocabulary (OOV) words. This was done by including all dev and test words in the lexicon.

5.4.3 Results

The KALDI system produces lattices as recognition results. The language model scaling factor is optimized on the development set and used to compute the best path through the lattice.

Table 6 Results of the first large vocabulary ASR system for the Tunisian dialect.

	LM PPL	**WER (%)**	**Substitutions**	**Deletion**	**Insertions**
Dev	41.69	21.5	15.2	4.1	2.3
Test	53.71	22.6	16.0	4.0	2.5

The results of the first large vocabulary ASR system for the Tunisian dialect are presented in table 6. A WER of 22.6% is obtained on the test set. In order to have a clear analysis about the type of errors made by our ASR system, we looked at the frequently confused pairs on the development set. Given the ASR results presented in previous table, we especially focused on the substitute and we found that the top 6 frequent substitutions are words with *Shadda*, which represent the doubling of consonants. These errors will be deeply studied in future work in order to improve this baseline ASR system.

Table 7 Examples of output our Tunisian dialect ASR system.

	Arabic script	Buckwalter transliteration
Reference	اَلتِسعة غير ربع متَاع قْفصة هذَاكة	Altsep gyr rbe mtAe qfsp h*Akp
ASR output	اَلتِسعة و ربع متَاع قْفصة هذَاك	tsep w rbe mtAe qfsp h*Akp
Reference	بقدَّاش تطلعلي التكَاي	bqd∼A$ ttlely tkAy
ASR output	قد تطلعلي التكَاي	qd ttlely tkAy
Reference	أي زوز لبيرويص فمّة	>y zwz lbrwys fm∼p
ASR output	أيه زوز لبيرويص أمم	>yh zwz lbrwys >mm

We showed examples of the system output in Table 7. For each example we presented the ASR output and the reference. We presented also the Arabic script and the Buckwalter transliteration. The two last examples showed the case in which our system recognize correctly all the words except these with *Shadda* (فمّة and بقدّاش).

6 Conclusion and future work

In this paper, we presented some historical features of the Tunisian dialect. We also reported the linguistic characteristics of this dialect at multiple levels including phonological, morphological, syntactic as well as the lexical level. We then exposed an in-depth analysis of the problems of converting G2P in the Tunisian dialect and the challenges related to spelling irregularities. An outcome of this study was the development of the G2P tools based on a comprehensive set of G2P rules and a dictionary of exceptions. This G2P tools was evaluated and validated using a manually annotated evaluation set.

On a further step, we created a spoken dialogues corpus in the Tunisian dialect in the Tunisian Railway Transport Network domain called TARIC. This corpus was manually transcribed and used together with our G2P tool to run our experience in order to develop an automatic speech recognition system of the Tunisian dialect. This ASR reaches a word error rate of 22.6% on a held-out test set. G2P tool and corpus are freely distributed for the research community.

For the future research, we are planning to improve the G2P conversion for unvowelized words. We also plan to expand our ASR system of the Tunisian dialect in the Tunisian Railway Transport Network to more generic field like broadcast news. Last, starting from the baseline ASR system presented in this paper, we will train deep neural network acoustic models in order to apply some transfert learning techniques that have been showed as to be very effective [43] to take benefit from resources available in closely-related language. In the case of Tunisian dialect, resources related to MSA should be really helpful to improve the acoustic models.

References

1. Abdel-Rahman A. (1991). Code-switching and Linguistic Accommodation in Arabic, Perspectives on Arabic Linguistics III: Papers from the Third Annual Symposium on Arabic Linguistics, 80, 231250, John Benjamins Publishing.
2. Alghamdi, M., Elshafei, M. & and Al-Muhtaseb, H. (2002). Speech Units for Arabic Text-to-speech, Fourth Workshop on Computer and Information Sciences, 199-212.
3. Alghamdi, M., Muzaffar, Z. & Alhakami, H. (2010). Automatic restoration of Arabic diacritics : a simple, purely statical approach, The Arabian Journal for Science and Engineering, Volume 35, Number 2.
4. Andersen, O., Kuhn, R., Lazaridès, A., Dalsgaard, P., Haas, J. & Nth, E. (1996). Comparison of two tree-structured approaches for grapheme-to-phoneme conversion, Spoken Language Processing, Philadelphia (PA), USA, 3, 1700–1703.

5. Baccouche, T. (2003). Larabe, dune koin dialectale une langue de culture, Mmoires de la socit linguistique de Paris, TomeXI, (les langues de Communication...), 87-93.
6. Barnard, E., Davel, M.H., Van Huyssteen G.B. (2010). Speech technology for information access: a south african case study, AAAI Spring Symposium: Artificial Intelligence for Development.
7. Besacier, L., Le, V.B., Castelli, E., Sethserey, S. & Protin, L. (2005). Reconnaissance automatique de la parole pour des langues peu dotees: Application au vietnamien et au khmer, TALN'2005.
8. Besacier, L., Barnard, E., Karpov, A. & Schultz, T. (2014). Automatic speech recognition for under-resourced languages: A survey. Speech Communication.
9. Biadsy, F., Habash, N. & Julia Hirschberg, J. (2009). Improving the Arabic Pronunciation Dictionary for Phone and Word Recognition with Linguistically-Based Pronunciation Rules, Annual Conference of the North American, Boulder, Colorado, 397405.
10. Bisani; M., Ney, H. (2008). Joint-Sequence Models for Grapheme-to-Phoneme Conversion, Speech Communication, 50, 434-451.
11. Blachona, D., Gauthiera, E., Besacier, L., Kouarata, G., Adda-Deckerb, M. & Rialland, A. (2016). Parallel Speech Collection for Under-resourced Language Studies using the Lig-Aikuma Mobile Device App, 5th Workshop on Spoken Language Technology for Under-resourced Languages, SLTU'2016.
12. Cucu, H., Buzo, A., Besacier, L., & Burileanu, C. (2014). SMT-based ASR domain adaptation methods for under-resourced languages: Application to Romanian, Speech Communication.
13. El-Imam, Y. (2003). Phonetization of Arabic: rules and algorithms, Computer Speech and Language.
14. Elmahdy, M., Hasegawa-Johnson, M. & Mustafawi, E. (2014). Development of a tv broadcasts speech recognition system for qatari arabic, The 9th edition of the Language Resources and Evaluation Conference : LREC'2014.
15. Elshafei, M., Al-Muhtaseb, H. & Alghamdi. M. (2006). Statistical methods for automatic diacritization of Arabic text, The Saudi 18th National Computer Conference, 18, 301-306.
16. Gauthier, E., Besacier, L.,Voisin, S., Melese, M. & Elingui, U.P. (2016). Collecting resources in sub-saharan african languages for automatic speech recognition: a case study of wolof, LREC'2016.
17. Gauthiera, E., Besacier, L. & Voisinb, S. (2016). Automatic Speech Recognition for African Languages with Vowel Length Contrast, 5th Workshop on Spoken Language Technology for Under-resourced Languages, SLTU'2016.
18. Gelas, H., and Solomon Teferra, S., Besacier, L. & Pellegrino, F. (2012). Analyse des performances de modles de langage sub-lexicale pour des langues peu-dotees morphologie riche, JEP-TALN-RECITAL 2012, Atelier TALAf 2012 : Traitement Automatique des Langues Africaines.
19. Graja, M., Jaoua, M. & Belguith, L. (2010). Lexical Study of A Spoken Dialogue Corpus in Tunisian Dialect, ACIT2010: the International Arab Conference on Information Technology, Benghazi-Libya, December 1416.
20. Graja, M., Jaoua, M. & Belguith, L. (2015). Statistical Framework with Knowledge Base Integration for Robust Speech Understanding of the Tunisian Dialect, IEEE/ACM Transactions on Audio, Speech & Language Processing, 23, 2311–2321.
21. Habash, N., Diab, D. & and Rambow, O. (2012). Conventional Orthography for dialectal Arabic,Proceedings of the Eighth International Conference on Language Resources and Evaluation LREC'2012.
22. Habash, N. (2010). Introduction to Arabic Natural Language Processing, Synthesis Lectures on Human Language Technologies, Graeme Hirst, Morgan & Claypool Publishers/
23. Habash, N. (2006). On Arabic and its Dialects, Multilingual Magazine, 17.
24. Häkkinen, J., Suontausta, J., Riis, S. & Jensen, K. (2003). Assessing text-to-phoneme mapping strategies in speaker independent isolated word recognition, Speech Communication, 41, 455-467.
25. Harrat, S., Meftouh, K., Abbas, M. & and Smaïli, K. (2014). Grapheme To Phoneme Conversion-An Arabic Dialect Case, In Spoken Language Technologies for Under-resourced Languages (SLTU'2014).

26. Illina, I., Fohr, D. & Jouvet, D. (2011). Grapheme-to-Phoneme Conversion using Conditional Random Fields, Interspeech' 2011.
27. Jensen, J., Riis, S. (2000). Self-organizing letter code-book for text-to-phoneme neural network model, Spoken Language Processing, 3, 318 321.
28. Juan, S.& Besacier, L. (2013). Fast bootstrapping of grapheme to phoneme system for under-resourced languages-application to the iban language, WSSANLP-2013.
29. Kheang, S., Katsurada, K., Iribe, Y. & Nitta, T. (2014). Solving the Phoneme Conflict in Grapheme-to-Phoneme Conversion Using a Two-Stage Neural Network-Based Approach, IEICE Transactions on Information and Systems, 97.
30. Lawson, S. & Itesh, S. (1997). Accommodation communicative en Tunisie: une tude empirique, Plurilinguisme et identits au Maghreb, Publications de lUniversite de Rouen, 101-114.
31. Lileikyta, R., Gorinaa, A., Lamela, L., Gauvaina, J. & Fraga-Silva, Th. (2016). Lithuanian Broadcast Speech Transcription using Semi-supervised Acoustic Model Training, 5th Workshop on Spoken Language Technology for Under-resourced Languages, SLTU'2016.
32. Loots, L., Niesler, T. (2011). Automatic conversion between pronunciations of different English accents, Speech Communication, 53, 7584.
33. Marchand, Y. & Damper, R. (2000). A multistrategy approach to improving pronunciation by analogy, Computational Linguistics, 26, 195-219.
34. Masmoudi, A., Khmekhem, M., Estève, Y., Belguith, L., & Habash, N. (2014). A Corpus and Phonetic Dictionary for Tunisian Arabic Speech Recognition, Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014), Reykjavik, Iceland, 306-310.
35. Masmoudi, A., Habash, N., Khmekhem, M., Estève, Y. & Belguith, L. (2015). Arabic Transliteration of Romanized Tunisian Dialect Text: A Preliminary Investigation, Computational Linguistics and Intelligent Text Processing - 16th International Conference, CICLing 2015, Cairo, Egypt, 608-619.
36. Mejri, S., Said, S. & Sfar, I. (2009). Pluringuisme et diglossie en Tunisie, Synergies Tunisie, 1, 53-74.
37. Nimaan, A., Nocera, P. & Torres-Moreno, J.M. (2006). Boites a outils tal pour les langues peu informatisees: Le cas du somali. JADT06: actes des 8es Journees internationales danalyse statistique des donnees textuelles: Besancon.
38. Pagel. V., Lenzo, K. & W. Black, A. (1998). Letter-to-sound rules for accented lexicon compression,Spoken Language Processing, Sydney, Australia, 2015 2018.
39. Pellegrini, T. (2008). Transcription automatique de langues peu dotees, Ph.D. thesis; Universite Paris Sud-Paris XI.
40. Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, K., Stemmer, G. & Vesely, K. (2011). The Kaldi Speech Recognition Toolkit, IEEE 2011 Workshop on Automatic Speech Recognition and Understanding.
41. Rasipuram, R. & Doss, M. (2012). Acoustic data-driven grapheme-to-phoneme conversion using KL-HMM, Acoustics, Speech and Signal Processing (ICASSP'2012), 4841-4844.
42. Saadane, H. & Habash, N. (2015). A Conventional Orthography for Algerian Arabic, Proceedings of the Second Workshop on Arabic Natural Language Processing, 69-79.
43. Samson, S., Besacier, L., Lecouteux, B. & Dyab, M. (2015). Using Resources from a Closely-related Language to Develop ASR for a Very Under-resourced Language: A Case Study for Iban, Interspeech'2015, Dresden, Germany.
44. Schlippe, T., Djomgang, E., Vu, N., Ochs, S. & Schultz, T. (2012). hause large vocabulary continuous speech recognition,The third International Workshop on Spoken Languages Technologies for Under-resourced Languages, Cape Town, South Africa, SLTU'2012.
45. Sejnowski, T. & Rosenberg, CH. (1987). Parallel networks that learn to pronounce English text, Complex Systems Publications, 145-168.
46. Seng. K., Iribe, Y., Nitta, T. (2011). Letter-to-Phoneme Conversion Based on Two-Stage Neural Network Focusing on Letter and Phoneme Contexts, INTERSPEECH'2011, 12th Annual Conference of the International Speech Communication Association, ISCA, 1885–1888.

47. Taylor, P. (2005). Hidden Markov models for grapheme to phoneme conversion, INTERSPEECH' 2005 - Eurospeech, 9th European Conference on Speech Communication and Technology, ISCA, 1973–1976.
48. Tebbi, H. (2007). Transcription orthographique phonétique en vue de la synthèse de la parole partir du texte de lArabe, Univrersité de Blida, Algérie.
49. Vergyri, D., Mandal, A., Wang., W., Stolcke, A., Zheng, J., Graciarena, M., Rybach, D., Gollan, CH., Schlter, R., Kirchhoff, K., Faria, A. & Morgan, N. (2008). Development of the SRI/Nightingale Arabic ASR system, Interspeech'2008, 14371440.
50. Vu, N.T, Kraus, F. & Schultz, T. (2011). Rapid building of an asr system for underresourced languages based on multilingual unsupervised training, Interspeech, Citeseer (2011)
51. Wang, X. & Sim, K. (2013). Integrating Conditional Random Fields and Joint Multi-gram Model with Syllabic features for Grapheme-to-Phone Conversion,INTERSPEECH'2013.
52. Zribi, I., Boujelbane, R., Masmoudi, A., Ellouze, M., Belguith, L. & Habash, N. (2014). A Conventional Orthography for Tunisian Arabic, Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014) Reykjavik, Iceland, 2355–2361.