PHONETIC TOOL FOR THE TUNISIAN ARABIC

Abir Masmoudi[1,2], Yannick Estève[1], Mariem Ellouze Khmekhem[2], Fethi Bougares[1], Lamia Hadrich Belguith[2]

(1) LIUM, University of Maine, France
(2) *ANLP Research group, MIRACL Lab., University of Sfax, Tunisia*
masoudiabir@gmail.com, yannick.esteve@lium.univ-lemans.fr, Mariem.Ellouze@planet.tn, fethi.bougares@lium.univ-lemans.fr, l.belguith@fsegs.rnu.tn

ABSTRACT

A phonetic dictionary is an essential component of a speech recognition system or a speech synthesis system. Our work targets the generation of an automatic pronunciation dictionary for the Tunisian Arabic, in particular in the field of rail transport. To do this, we created two tools of phonetic vowelized and unvowelized words in the Tunisian Arabic. The proposed method to automatically generate phonetic dictionaries is based on rules and is presented in this article. This paper outlines the steps to create our own study corpus **TARIC**: **T**unisian **A**rabic **R**ailway **I**nteraction **C**orpus[1]. Then it details the phonetic and phonological exceptions of the Tunisian Arabic and illustrates some rules used for the construction of phonetic dictionary.

Index Terms— *Tunisian Arabic, Phonologic, Phonetic Dictionary, vowelized and unvowelized words.*

1. INTRODUCTION

This work is part of the implementation of a system of automatic speech recognition of the Tunisian Arabic used in the field of railway transport. In this article, we focus on one of the key system components of speech recognition, namely the phonetic dictionary. The lack of spoken and written resources is one of the main problems for the treatment of the Tunisian Arabic. Thus, we present in this paper the efforts we deployed to create our own resources, as well as our method for automatically generating a phonetic dictionary for vowelized words and unvowelized .

This article consists of 8 sections. First, we present an overview of the Tunisian Arabic. Secondly, we expose the related work. Then, in the section 4 we present some detail about our corpus **TARIC** . In the next section we review the writing system of the Tunisian Arabic. Then. We also show the phonological and phonetic exceptions of the Tunisian Arabic. In the seventh section of the article, we outline the principles and steps of the automatic generation of phonetic dictionaries. Finally, we will give the results of the evaluation of our tools of phonetization.

2. THE TUNISIAN ARABIC

Arabic is currently the sixth most widely spoken language in the world. It has a special status as an official standard language in the Arab world. It is important to realize that what we typically refer to as “Arabic” is not a single linguistic variety; rather, it is a collection of different dialects. These Dialects can be classified according to geographical areas as they can be classified according to sociological and regional differences. We can also divide Arabic Dialects into two major groups namely the Western group (the Arab Maghreb or the North African group) and the Eastern group (Levantine Arabic, Gulf Arabic and Egyptian Arabic).

The Tunisian Arabic, generally known as the "**Darija**" or "**Tounsi**", is a subset of the Arabic Dialects associated with the Arabic of the Maghreb (the west of the Arab world). It is used in the daily life of Tunisian people for spoken communication. As for the other dialects, we can usually find dialects used by urban residents, farmers and Bedouins (residents of the desert). These various Tunisian Arabic differ considerably from each other. Differences affect all levels of the language, i.e. pronunciation, phonology, vocabulary, morphology, and syntax. Despite these differences, the Tunisian Arabic always remains understood by all Tunisian people. Among the phonological differences, we can note that Bedouins pronounce the letter “ق” /q/ as 'ڤ' /g/, but generally, urban people pronounce it “ق” /q/.
At the vocabulary level, to mean “stop”, urban residents use the word "يوقف" / i:u:qif /, but farmers use the word "يحبس" /i:aħbis/. These two words have the same semantic denotation. Amongst the remarkable specificities of the Tunisian Arabic, we note the presence of borrowed words from French, Berber, Italian, Turkish, English and Spanish. Also, we note the presence of words which the origin is

Modern Standard Arabic (MSA). The presence of these words is the result of [1]many factors and historical events that occurred throughout the centuries such as: the Islamic invasions, the French colonization, migration, commercial exchanges, etc [3].

Today, the Tunisian Arabic is becoming more and more often used in interviews, news, telephone conversations, public services, etc., and it has a strong presence today in blogs, forums, and user/reader commentaries on the internet. So, it is so important to consider the Tunisian Arabic in the new technologies like speech recognition systems, systems Human-Machine Dialogue, etc. However, the dialect suffers from the absence of tools and linguistic resources. Tools to treat the MSA are very difficult to use, given the large difference between the MSA and the Tunisian Arabic. Thus, it should be noted that our work studying the Tunisian while we lack of linguistic tools for this dialect.

3. RELATED WORKS

In the context of automatic speech recognition, to make the link between the lexical level and the acoustic level, it is necessary to associate each word in the vocabulary with one or more sequences of acoustic base units. In the related literature, several approaches are used to obtain the sequences of phonemes. We can distinguish the data-driven approach that typically uses learning data and the rule-based approach that requires linguistic expertise [5], [6], [9], [10], [7], [12], [13]. In what follows, we will further explain each of these approaches.

3.1 The Data-driven Approach

The general idea of this approach is to use a dictionary of manually "phonetized" words. Probabilistic approaches based on joint models [11] or on approaches based on machine translation [2] can capture machine learning links between graphemes used to write the words and phonemes, units of sound used to represent the pronunciation of a word. These statistical techniques require training data. When these data are sufficient in number, they allow obtaining interesting results while minimizing human expertise. In the absence of data, these approaches are not applicable.

3.2 The Rule-based Approach

The construction of an automatic rule-based phonetic system requires a good knowledge of the language and its phonetic rules which, moreover, must not contain too many exceptions: the automatic phonetic transcription using linguistic knowledge does not require innovative techniques, but rather a significant expertise in the tongue [5] [6].
The advantage of this approach is that it allows better control of the quality of the construction of pronunciation dictionaries: if there is an error, it is possible to add a new rule. Further, the development cost can be compared to the cost necessary for the construction of the training data manual used by the statistical approach.

4. CORPUS STUDY: THE TUNISIAN ARABIC RAILWAY INTERACTION CORPUS

In addition to the lack of tools to deal with the Tunisian Arabic, lack of linguistic resources is also evident. To build an automatic speech recognition system, we need a corpus of audio recordings and their transcription. So, face this situation, we are obliged to create our own study corpus **TARIC**.
The creation of a corpus is done in three phases: *1)* the production of audio recordings, *2)* the manual transcription of these recordings and *3)* the standardization of these transcripts. For ample information about the steps for creating of **TARIC**, you can read this article [1].
Table 1 presents some statistics of the **TARIC** corpus.

Number of hours	Number of dialogues	Number of statements	Number of words
20 h	4,662	18,657	71,684

TABLE 1 - Statistics of **TARIC**

During the analysis of **TARIC**, we noticed that foreign words represent 20% of the vocabulary. Here are some examples of foreign words in the Tunisian Arabic: the word "تكاي" "Ticket" /tika:i:/ is of French origin, or the word "ترينو" "Trinou" "/tri:nu:/ is of Italian origin. One can also note the presence of words of English origin like the word "أوكاي" "Ok" /ʔɔ:kɑ:i:/.
Also, we noticed that some foreign words are used with modifications in relation to their origins. Thus, a foreign word undergoes an enclitic addition from Arabic. Here are some examples of words:
• The word "سيپرميها" /sIprimIha:/ "delete" is a word borrowed from the French; it underwent the addition of the enclitic Arabic "ها" / ha :/ "her",which is attached to the word.
• The word "ريزرڤيلي" / rIzirvIlI /, which means " make a reservation for me ", is a word borrowed from French, it underwent the addition of the enclitic Arabic "لي" / ll / "me" that is attached to the word.

[1] To represent phonemes, we use the symbols the International Phonetic Alphabet (http://fr.wiktionary.org/wiki/Annexe:Prononciation/arabe)

5. THE WRITING SYSTEM OF THE TUNISIAN ARABIC

The Tunisian Arabic is written in script from right to left. The alphabet consists of thirty-one letters: *i)* twenty-five of these are consonants taken from MSA, *ii)* three consonants are the result of the presence of foreign words : 'پ' /V/, 'ڤ' /G/ and 'پ' /P/ and *iii)* three letters represent the long vowels.

Each letter can appear in up to four different shapes, depending on whether it occurs at the beginning, in the middle, or at the end of a word, or even in isolation. The letters are most of the time connected to each other graphically and there is no capitalization.

A distinguishing feature of the MSA and its dialects writing system is the presence of short vowels called diacritics. These diacritics are not represented by the letters of the alphabet but they are marked by short strokes placed either above or below the consonant. We can distinguish nine diacritics: (a) three short-vowel diacritics; (b) three "nunation" diacritics representing a combination of a short vowel and the marker /n/ -This type of diacritic is rarely used in the Tunisian Arabic-, (c) one consonant lengthening diacritic (called Shadda) which repeats the previous consonant; (d) one diacritic for marking when there is no diacritic (called Sukun).

6. PHONOLOGIC AND PHONETIC EXCEPTIONS OF THE TUNISIAN ARABIC

In this section we will present the phonetic and phonological exceptions of the Tunisian Arabic.
There are several specific phonologic and phonetic features in the Tunisian Arabic. We cite below a few of these phonetic features:

- Short vowels are frequently omitted. For example, the verb "write" is pronounced in MSA «كَتَبَ » /kataba/ but in the Tunisian Arabic, we say "كْتِبْ" /ktib/.
- Most words in the Tunisian Arabic end in a "sukun". So, the short vowel is neglected when it is located at the end of a syllable [4].
- The intervention of the consonant "ش" /ʃ/ in the Tunisian Arabic occurs all over the country in rural and in urban areas. This appears especially in the interrogative voice of the dialect.
- Strong sound of the consonant "ت" /t/.
- In the Tunisian Arabic, the consonant "ق" /q/ has a double pronunciation. In the rural dialects, it is pronounced 'ڤ' /g/. In the urban dialects, the consonant "ق" is pronounced /q/, but there are some exceptions.
- There are many consonants in the Tunisian Arabic that can be pronounced in many different ways. Below we represent some of them:
  - The consonant "س" /s/ can be pronounced as "س" /s/ or "ص" /sˤ/.
  - The consonant "ظ" /ðˤ/ is realized as /ðˤ/ or "ض" /dˤ/
  - The consonant "ث" /θ/ can be pronounced in two ways: "ث" /θ/ or "ف" /f/.
  - The consonant "غ" /ɣ/ can be pronounced in two ways: "غ" /ɣ/ or "خ" /x/.
- In the Tunisian Arabic, the "Hamza" "أ " /ʔ/,, at the beginning of a word, is pronounced in different ways.
- In the Tunisian Arabic, the "Hamza" "أ " /ʔ/, is omitted when it is located at the end of a word.
- Generally, the "Hamza" "أ " /ʔ/, is omitted when it is located in the middle of a word.
- "Ta-Marbouta" is usually silent, but there are some exceptions.
- There are letters that are not taken into account. The "alif" in the word "خرجوا" /xardʒʊ:/ " they exit" does not correspond to a sound (silence).
- Long vowels become short vowels. For example, " التران في" / fItran / "on the train" "ف التران" / fi tran /.
- We noted the elimination of a consonant in certain words. For example, "مانعرفش" /ma:naʕˤrafʃ/ "I don't know" can be pronounced "مانعرش" /ma:naʕˤraʃ/. In this example, the consonant "ف" / f / is eliminated.
- The Tunisian Arabic has a new phoneme /EY/ which does not exist in MSA as in the word, "حْرِام " /HH R EY M/ "cover".
- The numbers have a specific characteristic:
  - the numbers between "three" and "nine" accept a double pronunciations which undergoes elimination of some consonants and a change of vowels;
  - the numbers from "Eleven" also accept two pronunciations, one of which requires the addition of the phoneme "N" at the end of the number.
- We noted the addition of the phoneme / E IH / to support the pronunciation of the first silent consonant of a word.

7. BUILDING A PHONETIC DICTIONARY

We created two phonetic tools: one for vowelized words and the other for unvowelized words. In the following sections, we explain the general principle of automatic phonetics then we show the specificities of the two phonetic tools.

7.1. General principle of phonetics

To automatically generate a pronunciation dictionary, we have shown, from our corpus **TARIC**, a set of phonetic rules and a lexicon containing exceptions.

The Tunisian Arabic is characterized by the presence of numerous phonetic exceptions. Indeed, one can find a word that can be pronounced in two or more ways.
The process for phonetic vowelized and unvowelized words occurs in two phases: the consultation of the base lexicon of exceptions and applying the phonetics rules.

7.1.1. The lexicon of exceptions

There are words that cannot follow our set of phonetic rules; it is necessary to define a lexicon of exceptions. This lexicon is consulted before the rules are applied. If the word is among the exceptions, it is encoded directly in phonetic form. Otherwise, we must apply the rules to generate its phonetic form. In our lexicon, we have more than 30 exceptions. Our basic lexicon of exceptions has been validated by three experts (native speakers).

Examples of exceptions:
• The word "نصف" /nisˤf/ /Half/ can be pronounced in three different ways: "نصف" /nisˤf/ "نص" / nisˤ / or "نفص" /nifsˤ/.

7.1.2. Phonetic rules

We have developed a set of phonetic rules that must be provided for each letter. Each rule attempts to match certain conditions relating to the context of the letter and provide a phoneme or phoneme sequences and sometimes silence. Our rules have also been validated by three experts. The total number of rules is about eighty [1].
Each rule is read from left to right and follows this format [1]:

{Right-Condition} + {Graph} + {Left-Condition} => Replacement

7.2. Phonetization of vowelized words

During transcription of our corpus, we chose vowelized words according to the pronunciation of the speaker. The presence of these vowels permits a substantial decrease in the degree of phonetic ambiguity.

We noticed that the words which are vowelized in the Tunisian Arabic generally end with Sukun (silent consonant) or a long vowel.

To automatically generate the phonetic form of a word, the application of phonetic rules is done in the direction of reading of the word, ie it starts with the first letter of the word and the order of the letters are respected. Below is an example of a phonetic vowelized word:
The word "حداش" / hda:ʃ/ "Eleven" belongs to the basic numbers, therefore it accepts a double word pronunciation with one ending with the phoneme "N".

The rules used are:

R1: {C = ح / ħ /} + { Sukun} => {HH} = Phoneme
When a consonant is followed by "Sukun" then we always obtain phoneme of the consonant.

R2: { / d / د} + { VL = Fatha alif} => {Phoneme = D AE:}

When a consonant is followed by a long vowel 'alif and Fatha ", so we obtains the two phonemes: the corresponding one of the consonant "D" and the other of the long vowel "AE".

R3: {ش + / ʃ /} + { Sukun} => {Phoneme = SH}
When a consonant is followed by "Sukun" then we always obtain phoneme of the consonant.
Phonetics forms as follows: "حداش" / hda: ʃ / "Eleven"
1. HH D AE: SH
2. HH D AE: N SH

7.3 Phonetization of unvowelized words

Starting from three principles: (a) a word in the Tunisian Arabic ends with either a silent consonant (with Sukun) or a long vowel, (b) each long vowel is always followed and preceded by a silent consonant (c) a word cannot have two successive consonants that carry a vowel (long or short) with the exception of words with "Shadda" (Doubling of consonants). Phonetic rules for unvowelized words can be divided into two groups: support and secondary rules.

The support rules are applied in a first step. The majority of these rules are at the origin of the production of phoneme of the long vowels in the word, which facilitates the application of secondary rules in a second time. The secondary rules are at the origin of the production of phoneme of the short vowels and "Sukun".

Unlike vowelized words, the application of phonetic rules does not comply with an order well defined and therefore does not follow the reading direction of the word.

In fact, the phonetic starts by locating in the word at least one of these four consonants ("ا" / a :/ "Alif", "ى" / a :/ "Alif Maksoura", "ي" / i :/ "Ya" and "وا" / u :/ "Alif and Waw") to apply the rules support. These rules permit to give a phonétisation of long vowels. Then and based on the principle that "every long vowel is always followed and preceded by a silent consonant", secondary rules are applied in order to add the phonemes short vowels or "Sukun".

In the absence of one of these four consonants, the phonetic transcription is based on two principles: the word ends with a silent consonant (with Sukun) and second, one cannot have two successive vowels consonants bearing.

8. EVALUATION

The phonetic system was tested on a corpus consisting of 400 words collected from Tunisian blogs of various field (politics, sports, culture, scientific ...). The corpus was in the first instance standardized by CODA : conventional orthographic for the Arabic Dialect" [8].
The evaluation of these two phonetic tools (with and without vowels) is done as follows:
(*i*) we did the manual phonetization of our test corpus then (*ii*) we did the automatic phonetization of the same corpus using our tools, (*iii*) next we used the tool Sclite to test our tools in terms of phonemes and words. This evaluation procedure is the same for the vowelized and unvowelized corpus.
The following table shows results obtained for the phonetic corpus with and without vowels.

Test corpus	Error rates in phoneme
vowelized	0 %
unvowelized	12.9 %

TABLE 2 - Results of the evaluation

As presented in Table 2, the system of phonetic of a Tunisian Arabic is 100% performer for vowelized texts.
Concerning non vowelized text, phonetisation is a more complex task where our system can provide several variants of phonetic. The error rate of 12.9% phoneme presented in Table 2 is the first proposal of the system.

9. CONCLUSION

This paper describes our effort to create two tools for the automatic phonetization of vowelized and unvowelized words used in the Tunisian Arabic in the field of railways. The tools generate a pronunciation dictionary based on phonetic rules of Tunisia. Each rule tries to match certain conditions relating to the context of the letter and to provide a replacement. The total number of rules is about 80.
To cope with the lack of linguistic resources in the Tunisian Arabic, we were compelled to create our own corpus **TARIC**. The creation of **TARIC** is done in three phases, namely, production of audio recordings, manual transcription of these records and standardization of these transcripts.
Also, this article gives a description of the Tunisian Arabic's writing system and the phonetic and phonological exceptions.
The phonetic system was tested on a corpus consisting of 400 words (vowelized and unvowelized), the error rate of vowelized words is 0 % and 12.9% for unvowelized words.

10. REFERENCES

[1] A. Masmoudi, M. Ellouze Khmekhem, Y. Estève, L. Hadrich Belguith, and N. Habash, " A corpus and a phonetic dictionary for Tunisian Arabic speech recognition", *In 19th edition of the Language Resources and Evaluation Conference (LREC)*, Reykjavik, Iceland, 2014.

[2]A. Laurent, P. Deléglise and S. Meignier, "Grapheme to phoneme conversion using an SMT system", *INTERSPEECH' 2009*, pp. 708-711, 2009.

[3] C. Pereira, 2011. "Arabic in the North African Region", In W. Stefan (Ed.), The Semitic Languages, pp. 954–969, 2011.

[4] D. Lajmi, "Spécificités du dialecte Sfaxien", Synergies Tunisie n° 1, Tunisie, 2009.

[5] F. Béchet, " LIA_PHON : un système complet de phonétisation de textes", *Traitement Automatique des Langues,* pp. 47-67, 2001.

[6] F. Biadsy, N. Habash, and J. Hirschberg, "Improving the Arabic Pronunciation Dictionary for Phone and Word Recognition with Linguistically-Based Pronunciation Rules", The 2009 Annual Conference of the North American Chapter of the ACL, pages 397–405, 2009.

[7] F. Diehl, M.J. Gales, M. Tomalin, and P.C. Woodland, "Phonetic pronunciations for Arabic speech-to-text systems", *IEEE International Conference on Acoustics*, Speech and Signal Processing, pp. 1573-1576, 2008.

[8] I. Zribi, R. Boujelben, A. Masmoudi, M. Ellouze Khmekhem, L. Hadrich Belguith and N. Habash, "A Conventional Orthography for Tunisian Arabic" , *In 19th edition of the Language Resources and Evaluation Conference (LREC)*, Reykjavik, Iceland, 2014.

[9] M. Algamdi, M. Elshafei, and H. Almuhtasib, "Speech Units for Arabic Text-to-speech", *Fourth Workshop on Computer and Inforamtion Sciences*, pp. 199-212, 2002.

[10] M. Algamdi, "KACST Arabic Phonetics Database", *Fifteenth International Congress of Phonetics Science*, Barcelona, pp. 3109-3112, 2003.

[11] M. Bisani and H. Ney. "Joint-Sequence Models for Grapheme-to-Phoneme Conversion", *Speech Communication*, Volume 50, Issue 5, pp. 434-451, 2008.

[12] M. J. F. Gales, F. Diehl, C. K. Raut, M. Tomalin, P. C. Woodland, and K. Yu, " Development of a phonetic system for large vocabulary Arabic speech recognition", *IEEE Workshop on Automatic Speech Recognition & Understanding*, pp. 24-29, 2007.

[13] Y. El-Imam, "Phonetization of Arabic: rules and algorithms", In Computer Speech and Language 18, pp. 339–373, 2004.