# A Corpus and Phonetic Dictionary for Tunisian Arabic Speech Recognition

**Abir Masmoudi[1,2], Mariem Ellouze Khemakhem[1], Yannick Estève[2], Lamia Hadrich Belguith[1] and Nizar Habash[3]**

(1) ANLP Research group, MIRACL Lab., University of Sfax, Tunisia
(2) LIUM, University of Maine, France
(3) Center for Computational Learning Systems, Columbia University, USA

masmoudiabir@gmail.com, mariem.ellouze@planet.tn, yannick.esteve@lium.univ-lemans.fr, l.belguith@fsegs.rnu.tn, habash@ccls.columbia.edu

**Abstract**

In this paper we describe an effort to create a corpus and phonetic dictionary for Tunisian Arabic Automatic Speech Recognition (ASR). The corpus, named TARIC (Tunisian Arabic Railway Interaction Corpus) has a collection of audio recordings and transcriptions from dialogues in the Tunisian Railway Transport Network. The phonetic (or pronunciation) dictionary is an important ASR component that serves as an intermediary between acoustic models and language models in ASR systems. The method proposed in this paper, to automatically generate a phonetic dictionary, is rule based. For that reason, we define a set of pronunciation rules and a lexicon of exceptions. To determine the performance of our phonetic rules, we chose to evaluate our pronunciation dictionary on two types of corpora. The word error rate of word grapheme-to-phoneme mapping is around 9%.

**Keywords:** Tunisian Arabic, speech recognition, phonetic dictionary, grapheme-to-phoneme

## 1. Introduction

Automatic Speech Recognition (ASR) is playing an increasingly important role in a variety of applications such as automatic query answering, telephone communication with information systems, speech-to-text transcription, etc. In this paper we describe an effort to create a corpus and phonetic dictionary for Tunisian Arabic ASR. The corpus, named **TARIC** (**T**unisian **A**rabic **R**ailway **I**nteraction **C**orpus) has a collection of audio recordings and transcriptions from dialogues in the Tunisian Railway Transport Network. The phonetic (or pronunciation) dictionary is an important ASR component that serves as an intermediary between acoustic models and language models in ASR systems. It contains a subset of the words available in the language and the pronunciation variants of each word in terms of sequences of the phonemes available in the acoustic models.

In the next section, we give a historical overview of Tunisian Arabic. Then, in Section 3, we present the steps of creating the corpus for our study and provide an analysis of this corpus in Section 4. Section 5 details the phonological variations of Tunisian Arabic. Sections 6 and 7 present the method we propose to build the Tunisian Arabic phonetic dictionary and its evaluation, respectively.

## 2. Historical Overview of Tunisian Arabic

Modern Standard Arabic (MSA) has a special status as an official standard language in the Arab world. It is in particular the language of the written press and official venues. Furthermore, there is a large variety of dialects that constitute the mother tongues of Arabic speakers. Arabic Dialects are divided into two major groups namely the Western group or North African group and the Eastern group. The North African Arabic is the variety of Arabic spoken in the Maghreb countries (Tunisia, Algeria, Morocco, Libya and Mauritania) while the Eastern group includes the varieties spoken in Egypt, the Levant, Iraq, the Gulf states, Yemen, Oman, etc.

Tunisian Arabic is the main variety used in the daily life of Tunisian people for spoken communication. It is becoming more widely used in interviews, news, debate programs, and public service announcements; and it has a strong online presence today in blogs, forums, and user/reader commentaries.

Historically, Berber was the original mother tongue of the inhabitants of North Africa. The spread of Islamin North Africa brought Arabic, the language of the Islam's Holy Book. Other historical facts occurred which influenced the language spoken in Tunisia such as the Ottoman empire, European colonialism and peaceful trade-based interactions between civilizations. So, Tunisian Arabic is an outcome of the interactions between Berber, Classical Arabic and many other languages. The trace of this interaction in the language is manifested in the introduction of borrowed words from French, Italian, Turkish and Spanish in Tunisian Arabic. These borrowings are used in the daily life of Tunisians with some phonological changes. However, many borrowed words are used in the discourse of the Tunisians without being adapted to the Tunisian phonology.

Table 1 below shows some examples of foreign words commonly used in Tunisian Arabic with or without phonological modification.

| Words | Transliteration | Origin | Sense |
|---|---|---|---|
| شكبَة | škub~aħ | Italian | card game |
| كاغث | kaAγiθ | Turkish | paper |

Table 1: Some examples of foreign words used in Tunisian Arabic .[1]

## 3. The Tunisian Arabic Railway Interaction Corpus

The building of an ASR system requires at least two types of corpora: audio recordings and the corresponding written text. Since we aim to build an ASR system, and due to the lack of such resources especially concerning Tunisian Arabic, we decided to create our own corpus, which we named TARIC: Tunisian Arabic Railway Interaction Corpus. The creation of the corpus was done in three steps. First is the production of audio recordings; second is the transcription of these recordings; and third is the normalization of these transcriptions. In the following three sub-sections we will detail the process of creation of TARIC**.**

### 3.1 The Recordings

The first step consisted in making audio recordings. We did that in the ticket offices of the Tunis railway station. We recorded conversations in which there was a request of information about such things as the train schedules, fares, bookings, etc. The equipment we used includes two portable PCs using the Audacity software and two microphones, one for the ticket office clerk and another one for the client. We chose to record in different periods, particularly holidays, weekends, festival days, and sometimes during the week. We obtained 20 hours of audio recordings.

### 3.2 The Transcription

Once our recordings were ready, we manually transcribed them because we did not have the tools for automatic transcription for Tunisian Arabic. This transcription was done by three university students. Our corpus consists of several dialogues; each dialogue is a complete interaction between a clerk and a client. All the words are written using the Arabic alphabet with diacritics. The diacritics indicate how the word is pronounced. The same word can have more than one pronunciation.
Table 2 presents some statistics of the TARIC corpus.

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

Table 2: Statistics of the TARIC corpus

[1] Transliteration of Arabic will be presented in the Habash-Soudi-Buckwalter scheme (Habash et al, 2007).

### 3.3 Normalization

To obtain coherent data and consistent corpora, we had to use standard orthographies. But until now, Tunisian Arabic has no standard orthographies since there are no Arabic dialect academies. In our laboratory, we developed our own orthographic guidelines to transcribe the spoken Tunisian Arabic following previous work by Habash et al. (2012) on developing a conventional orthography for dialectal Arabic – or CODA. Our guidelines are described in (Zribi et al.,2014).

## 4. Analysis of TARIC

In this section, we present an analysis of the collected corpus. The analysis consists of determining dialogue acts, foreign words, lexical variations and speech disfluencies.

### 4.1 Dialogue Acts

Dialogue acts are the actions caused by the speaker. The corpus had a variety of dialogue acts that pertain to requests and answers about scheduling and reservations. Table 3 shows an example of segmentation in dialogue act of a set of conversations between a client and an agent.

| Dialogue Act | Dialect Lexicon | Translation |
|---|---|---|
| Departure time requests | وقتاش التران للتونس | When is the train to Tunis? |
| Answer | ثمة في العشرة و في الماضي ساعة | there is at 10 hours and at 13 hours |
| Reservation requests | ريزرفيلي في التران متاع العشرة | Reserve me for the train at10. |
| Confirmation | أوكاي | OK |

Table 3: Analysis in dialogue act of a conversation between an agent and a Client

### 4.2 Lexical Variation

As indicated in Section 2, the use of foreign words is a common feature in Tunisian Arabic due to historical reasons. In TARIC, foreign words represent 20% of the corpus. Table 4 gives some examples of these words.

| Dialect words | Translation | Origin | Sense |
|---|---|---|---|
| تران | triAn | French | Train |
| كلاس | klaAs | French | Class |
| بلاصة | blaASaħ | French | Space |

Table 4: Examples of foreign words

Also, we noticed the presence of several different words from different backgrounds but with the same meaning. For example, the word "ticket" can be expressed in three different ways: تكاي tikaAy,تسكرة tiskraħor تذكرة tiðkraħ.

Table 5 illustrates other frequently used examples.

| Lexicon | | | Translation |
|---|---|---|---|
| ترين<br>triAn | ترينو<br>triynuw | أوتوراي<br>ÂuwtuwraAy | Train |
| بقايع<br>bqaAyiE | بلايص<br>bliAyiS | پلاس<br>plaAs | Places |

Table 5: Example of lexical variation in Tunisian Arabic

### 4.3 Speech Disfluencies

Disfluency is a frequently occurring phenomenon in spontaneous oral production resulting in new lexical classes that need to be properly handled. The principal phenomena of disfluency are: repetitions, self-corrections, hesitations and incomplete words. Next, we present an analysis of our corpus TARIC in terms of disfluencies to extract these new lexical classes.

- **Repetitions:** these consist of repeating a word or series of words. The majority of repetitions in TARIC are used by a speaker to affirm or to reformulate his request. Below are two examples of repetitions.

(a) **زوز** للتونس **ألاى رتور زوز ألاى رتور**

**two** to Tunis **go back two go back**

Example (a) represents a repetition in the speaker utterance to affirm the request.

(b) **تكاي بليصة** للصفاقس

**Ticket place** to Sfax

In the second example, the repetition is used by the speaker to press his claim. He used two different words that have the same meaning.

- **Self-corrections**: the speaker can make one or more mistakes and correct them in the same utterance. This phenomenon is similar to a repetition but the repeated portion is a reconstruction of a bad portion in the utterance. Below are two examples of self-corrections.

(a) **تونس** لا **سوسة**

**Tunis** no **Sousse**

(b) تكاي **ألاي** لا سامحني **ألاي رتور**

Ticket **go** no sorry **go back**

- **Hesitations:** these are phenomena which appear in spontaneous oral production. They can be manifested in various ways: either by using a specific morpheme (e.g., uh, um, etc.) or in the form of an elongation of syllable. These are lexical classes belonging only to spontaneous oral production. There are lexical classes that are similar to foreign languages such as French and others are specific to Tunisian Arabic. The following example shows hesitation markers present in our corpus.

(a) تران للتونس **آه** دراكت

Train to Tunis **ah** direct

- **Incomplete words:** these are the cases of the stopping the production of a word before the normal end of it. In his terminology, an incomplete word is always a word fragment that can be identified through knowledge of the phraseology.

(a)باللاهي **ترا تران** دوزيام كلاس

Please **tra train** second class

In this example, the speaker begins to pronounce the word "train" but he stops before the normal end of the word and then says the full word again.

## 5. Phonological Variations in Tunisian Arabic

Before creating a phonetic dictionary for Tunisian Arabic, it is necessary to study the phonological variations of this language. There are several specific phonological variations in Tunisian Arabic. We can find a variation in the pronunciation of some consonants. We cite below a few of these phonetic features:

- The presence of foreign words in Tunisian Arabic resulted in the introduction of three new phonemes: پ /V/, ڤ /G/ and پ /P/.
- In 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.
- The consonant ض /DD/ can have several possible pronunciations such asض /DD/ or ذ "ð" /DH/ or د "d" /D/. For example, the word مَاضِي /M AE: DD IY/ in the expression مَاضِي سَاعَة /M AE: DD IY S AE: AI AE/ '13 hours' is pronounced مَاضِي /M AE: DD IY/ or مَاذي /M AE: DH IY/ or مَادي /M AE: D IY/.
- The consonant س "s"/S/ can be pronounced as /S/ or ص "S" /SS/. For example, the word رسول /R AE S UW L / 'Prophet' is pronounced رسول /R AE S UW L/ or رصول /R AE SS UW L/.
- The consonant ظ /DH2/ is realized as /DH2/ or ض /DD/.
- In a few words such as ثَمَّةْ /TH AE M M AE/ 'exist', the consonant ث "v"/TH/ can be pronounced in two ways: ث /TH/ or ف "f"/F/
- The consonant ط "T" /TT/ is sometimes pronounced /TT/ and at other times ت "t"/T/. For example, أعطيني /E AE AI T IY N IY/ 'give-me' is pronounced أعطيني /E AE AI TT IY N IY/ or أعتيني /E AE AI T IY N IY/
- Tunisian Arabic Hamza (or glottal stop) at the beginning of the word, is sometimes pronounced with different ways:
  - If the word is at the beginning of the statement, the glottal stop is pronounced.
  - If the word is in the middle of the statement, the glottal stop is omitted.
- The consonant ع "E" /AI/ is sometimes

pronounced /AI/ and at other times ح "H" /HH/. For example, متَّاعهَا /M T AE: AI H AE:/ 'hers' is pronounced متَّاعهَا /M T AE: AI H AE: / or متَّاحهَا /M T AE: HH H AE:/.

- We noticed the elimination of a consonant in some word. For example, قلتلْك /Q UH L T L IH K / I told you can be pronounced قتلْك /Q UH T L IH K/, we noticed that the consonant ل "l" /L/ is eliminated.
- In Tunisian Arabic, starting from eleven, the phoneme (n) is added to numbers followed by a noun, for example, حدَاشنْ ألف /HH D AE: SH N E AE L F/.

## 6. The Tunisian Arabic Phonetic Dictionary

Pronunciation dictionaries map words to one or more pronunciation variants and take into account pronunciation variability. Our approach consists in using a set of phonetic rules and a lexicon of exceptions to automatically generate a pronunciation dictionary.

### 6.1 The Lexicon of Exceptions

There are some words that cannot follow our set of phonetic rules. So, it is necessary to define a lexicon of exceptions. This lexicon is consulted before the rules are used. If the word is among the exceptions, it is encoded directly in phonetic form. Otherwise, we must apply the rules to the word to generate its phonetic form. In our lexicon, we have more than 30 exceptions. Our lexicon of exceptions is evaluated by three judges (native speaker).Table 6 shows some examples of lexical exceptions.

| Exceptions | Transliteration | Phonetization |
|---|---|---|
| هذاhaðaA | this[masc. sg.] | H AE: DH AE: |
| هذيhaðiy | this[fem. sg.] | H AE: DH IY |
| الهAilaAh | god | E IH L AE: H |

Table 6: Lexicon of exceptions

This operation is called transcription by phonetic lexicon for each word as it directly generates a lexical entity that represents the pronunciation that matches it.

### 6.2 Phonetic Rules

We developed a set of phonetic rules to map written Tunisian Arabic. Rules are provided for each letter in Tunisian Arabic. Each rule tries to match certain conditions relative to the context of the letter and to provide a replacement. Our rules are evaluated by three judges (native speaker).These rules are stored in a rule base. The total number of rules is 80. Each rule is read from right to left and follows this format:

**Replacement**<={**Left-Cond**}+{**Graph**}+{**Right-Cond**}

- **Graph**: is the current letter in the word.
- **Right-Condition** has one of the following formats:

<? <= Pattern>: context before the current position "Graph" is to be considered.
<? <! Pattern>: context before the current position "Graph" is not to be considered.

- **Left-Condition:** can take one of these two formats:

<? = Pattern>: context after the current position "Graph" is to be considered.
<! Pattern>: context after the current position "Graph" is not to be considered.

- **Replacement:** is either a phoneme or more of a phoneme or a vacuum (*) if the graph is omitted in pronunciation.

The application of phonetic rules is done in the direction of reading of the word, that is to say it starts with the first letter of the word and respects the order of letters. The following are three examples of rules of Tunisian Arabic:

**1. Shadda rule**: shadda diacritic "ّ" is written on a consonant and never on a vowel. Its effect is to double the consonant on which it is placed.

**2. The rules of the "ا" (Alef):** at the end of a word and preceded by "w", the combination signifies a plural word. In this situation, the final "Alef" does not have any pronunciation. For example, in the plural word "خَلْصُوا" (they have paid) the final "ا " is deleted.

**3. Sun letter rule:** When a word starts with the definite article ال Al+ followed by a so-called "Sun" consonant letter, the /l/ of the definite article is assimilated to the consonant (Habash, 2010; Biadsy et al., 2009). For example, the word السما Al+samiA 'the sky' is pronounced /E IH S S M AE:/.

## 7. Evaluation

We evaluate the performance of our phonetic rules on two corpora: TARIC and another corpus downloaded from the website of Tunisian bloggers. This corpus is selected on several themes: political, sporting, cultural, social, etc. Since the web corpus does not follow our writing standard, we standardized the corpus according to Tunisian Arabic CODA (Zribi et al., 2014)and manually diacritized it.
The evaluation set contained around 3K unique words from TARIC and 3K unique words from the web.
Our pronunciation dictionary is evaluated by three experts (native speaker).
Table 8 shows the evaluation size of each type of corpus.

| TARIC corpus | Web corpus |
|---|---|
| 8% | 10% |

Table 8: Results of the evaluation (word error rate)

As presented in Table 8, the system of phonetic of a Tunisian Arabic has 8% word error rate for vowelized words of our corpus **TARIC** and 10% word error rate for diacritized words from the web corpus. These errors are due to the order of rules, for example it is necessary to

make the rules of long vowels before rules of short vowels. Also, you can find errors due to the contradiction of two rules.

## 8. Conclusion and Future Work

To deal with the lack of linguistic resources in Tunisian Arabic for ASR, we create our own corpus **TARIC**. We described TARIC creation and highlighted some of its features. We also presented a tool for rule-based grapheme to phoneme mapping that converts graphemes of Tunisian Arabic into their corresponding phonemes. The process of implementation is based on the list of graphemes, phonemes, the lexicon of exceptions and phonetic rules. Each rule attempts to match certain conditions relating to the context of the letter and provides a replacement. The total number of rules is about 80.The resulting software is tested on a word list in Tunisian Arabic using two independent test sets and reached an error rate of ~9%. The data that has been prepared: **TARIC** and phonetic dictionary and tool will be used to build ASR systems in the Tunisian Railway Transport Network.

In future work, we plan to extend our research to improving of the phonetization of diacritized and undiacritized words in Tunisian Arabic. We will consider methods for data driven grapheme-to-phoneme mapping.

## 9. References


Algamdi, M. (2003). KACST Arabic Phonetics Database.Fifteenth International Congress of Phonetics Science, Barcelona. pages. 3109-3112.

Algamdi, M., Elshafei, M., & Almuhtasib, H. (2002).,Speech Units for Arabic Text-to-speech. Fourth Workshop on Computer and Inforamtion Sciences. pages. 199-212.

Biadsy, F., Habash, N., and Hirschberg, J., (2009), 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, Boulder, Colorado.

Bisani, M., Ney, H., (2008).Joint-sequence models for grapheme-to-phoneme conversion. Speech Communication 50, 434–451.

Diehl, F., Gales, M. J. F., Tomalin, M., & Woodland, P. C. (2008).Phonetic pronunciations for Arabic speech-to-text systems. IEEE International Conference on Acoustics, Speech and Signal Processing .pages. 1573-1576.

El-Imam. Y., (2004). Phonetization of Arabic: rules and algorithms. In Computer Speech and Language 18, pages 339–373.

Gales, M. J. F., Diehl, F., Raut, C. K., Tomalin, M., Woodland, P. C., & Yu, K. (2007).Development of a phonetic system for large vocabulary Arabic speech recognition. IEEE Workshop on Automatic Speech Recognition & Understanding. pages. 24-29.

Habash, Nizar. (2010) Introduction to Arabic Natural Language Processing, Synthesis Lectures on Human Language Technologies, Graeme Hirst, editor. Morgan & Claypool Publishers.

Habash, N., Soudi, A., and Buckwalter T. (2007). On Arabic Transliteration. Book Chapter. In Arabic Computational Morphology: Knowledge-based and Empirical Methods. Editors Antal van den Bosch and Abdelhadi Soudi.

Habash, N., Diab, M., Rambow, O. (2012).Conventional Orthography for Dialectal Arabic. In: Proceedings of the Language Resources and Evaluation Conference (LREC), Istanbul.

Hiyassat, H. A. R. (2007). Automatic Pronunciation Dictionary Toolkit for Arabic Speech Recognition Using SPHINX Engine. Ph.D. thesis, Arab Academy for Banking and Financial Sciences, Amman, Jordan.

Maamouri, M., Buckwalter, T., Cieri, C. (2004). Dialectal Arabic Telephone Speech Corpus: Principles, Tool Design, and Transcription Conventions. In: NEMLAR International Conference on Arabic Language Resources and Tools, Cairo, September, pages. 22-23. Paris-sud, Centre d'Orsay.

Masmoudi, A., Estève, Y., Ellouze Khmekhem, M., Hadrich Belguith, L., (2014), Phonetic tools for the Tunisian Dialect, The 4th International Workshop on spoken Language Technologies for Under-resourced Languages, Russia.

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