ABSTRACT

We aim to develop a Speech-to-Speech translation system between Modern Standard Arabic and Algiers dialect. Such a system must include a Text-to-Speech module which itself must include a Grapheme-to-Phoneme converter. Algiers dialect is an Arabic dialect concerned by the most problems of Modern Standard Arabic in NLP area. Furthermore, it could be considered as an under-resourced language because it is a vernacular language for which no substantial corpus exists. In this paper we present a grapheme-to-phoneme converter for this language. We used a rule based approach and a statistical approach, we got an accuracy of 92% VS 85% despite the lack of resource for this language.

Index Terms— Modern Standard Arabic (MSA), Algiers Dialect, Grapheme-to-phoneme conversion, Statistical Machine Translation.

1. INTRODUCTION

Grapheme-to-Phoneme (G2P) conversion or phonetic transcription is the process which converts a written form of a word to its pronunciation form. G2P conversion is not a simple deal, especially for non-transparent languages like English where a phoneme may be represented by a letter or a group of letters and vice-versa [1].

Unlike English, Arabic is considered as a transparent language; in fact, the relationship between grapheme and phoneme is one to one, but note that this feature is conditioned by the presence of diacritics. The lack of vocalization generates ambiguity at the phonetic level and consequently at the lexical, syntactic and semantic levels. For example the word "كتب" /kth/, can have different phonetic transcriptions\(^1\) such as /kataba/, /kutiba/, /kutubun/, /kutubi/, /katbin/... Algiers dialect obeys to the same rule: without diacritics, G2P conversion will be a difficult issue to resolve. That is why we first dedicated our efforts to develop an automatic diacritizer system for this language.

Most of works on G2P conversion have used two approaches: the first one is a dictionary-based approach, where a phonetized dictionary contains for each word of the language its correct pronunciation. The G2P conversion is reduced to a lookup of this dictionary. The second approach is rule-based[2][3][4], in which the conversion is done by applying phonetic rules, which are deduced from phonological and phonetic studies of the considered language or learned on a phonetized corpus using a statistical approach based on significant quantities of data[5],[6].

For Algiers dialect which is a non-resourced language, a dictionary based solution for a G2P converter is not feasible since a phonetized dictionary with a large amount of data is not available. The first intuitive approach (regards to the lack of resource) is a rule-based one, but the specificity of Algiers dialect had led us to a statistical approach in order to consider all features related to this language.

2. ALGIERS DIALECT

Algiers dialect is one of the most important dialects of Algeria; it represents the dialectal Arabic spoken in Algiers and its periphery. Algiers dialect simplifies the morphological and syntactic rules of the written Arabic. It uses the Arabic alphabet which includes 28 letters (14 solar consonants which assimilate the ٧ of a preceding definite article ٧, and 14 lunar consonants which do not assimilate it) and three non Arabic letters ـ /G/, ـ /V/ and ـ /P/. It uses all Arabic diacritics except the Tanween doubled case endings see [7] for more details.

2.1. Writing System

Algiers dialect is a spoken language without conventional writing rules. No writing system is adopted for it by scientific community of the country. In fact, the absence of

\(^1\) We use SAMPA Speech Assessment Methods Phonetic Alphabet for phoneme representation.
writing system for Algiers dialect is due to the lack of resources such as corpora for this language. Working on this issue (lack of resources) began with a corpus construction which has taken us to adopt a writing system. The main idea for elaborating writing rules is based on the fact that Algiers language is an under-resourced Arabic language, so we adopted the Arabic writing system when it is possible. That means when writing a word in Algiers dialect we look if there is an Arabic word close to this dialect word: if it exists, we adopt the Arabic writing for the dialect word; otherwise, the word is written as it is spoken. It should be noted that Algiers dialect is written from right to left (like Arabic language).

2.2. Issues of G2P Conversion for Algiers Dialect

Algiers dialect G2P conversion obeys to the same rules as MSA. Indeed, Algiers dialect could be considered as a transparent language since alignment between grapheme and phoneme is one to one when the input text is vocalized. This transparency conducts us to adopt a rule-based approach to build a G2P converter. Furthermore, Algiers dialect contains borrowed words from foreign languages (mainly French). Its vocabulary contains many French words used in everyday conversation. French borrowed words could be divided into two categories: the first one includes French words phonologically altered such as the word “فامليه” (famille, family) and the second one includes words which are pronounced as in French like the word “سور” (sûr, sure).

This last category constitutes a serious deal for G2P conversion since these words do not obey to Arabic pronunciation rules.

<table>
<thead>
<tr>
<th>Dialect word</th>
<th>Dialect phonetic transcription</th>
<th>French word</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>كوزينة</td>
<td>/ku:zina/</td>
<td>Cuisine</td>
<td>Kitchen</td>
</tr>
<tr>
<td>طابلة</td>
<td>/ t'a:bla/</td>
<td>Table</td>
<td>Table</td>
</tr>
<tr>
<td>كونكورس</td>
<td>/kûnn/kksj3/</td>
<td>Connexion</td>
<td>Connection</td>
</tr>
<tr>
<td>دوفيز</td>
<td>/doviz/</td>
<td>Devis</td>
<td>Currency</td>
</tr>
</tbody>
</table>

Table 1. Examples of French words used in Algiers dialect

In the examples of Table1, although the first two words are French, they are phonetized as Arabic words. The French word “table” is phonologically altered and written in Algiers dialect (with Arabic script) “طابلة”. On the other side, the last two words are phonetized as French words since they are pronounced as in French by Algiers dialect speakers. In order to take account of this word category, the French phonemes like /ε/, /ɛ/ and /s/ must be included in Algiers dialect phonemes. Note that, depending on the speaker, we can find different words related to the same meaning falling into both categories. An example of this, the word “فامليه” (famille), which is a French word (famille) phonologically altered is widely used in Algiers dialect, but even the word “فاملي” (famille) with French pronunciation is used.

3. RULE BASED APPROACH

As mentioned above the rule-based approach for G2P conversion applied to Algiers dialect requires a diacritized text. Hence, we implemented ADAD, the Automatic Diacritizer of Algerian Dialect texts which then can be converted into their phonetic forms by applying the established rules mentioned in Section 3.2.

3.1. Automatic Diacritic Restoration

ADAD[7] is based on a statistical approach; we used available tools for machine translation where source and target were respectively undiacritized and diacitized texts. We considered diacritization as a Machine Translation problem; we built a Statistical Machine Translation system based on parallel corpora of diacritized and undiacritized texts. We used a trigram language model trained on diacritized texts, a phrase table with undiacritized and diacritized entries. This choice is due to the fact that this approach does not require any linguistic knowledge or resources except vocalized corpora. First, we experimented our solution on MSA diacritized corpora Tashkeela2 and LDC Arabic Treebank (Part3, V1.0) [8], we got motivating results: a DER (Diacritization Error Rate) of respectively 4.1% and 5.7%, and precision rate of 93.1% and 96%.

Figure 1. Iterative Diacritization Process

For dialectal side, we vocalized by hand a part of Algiers dialect corpus and we used it for training our

2 http://sourceforge.net/project/tashkeela
3.2. Rule Based G2P Conversion

After the diacritization step described above, the text is converted into phonemes by applying a set of rules detailed in the following. It should be noted that most of these rules are those adopted also for Arabic [2] [3] and are applicable only for Arabic words and foreign words phonologically altered in our corpus.

Let us consider: BS is a mark of the beginning of a sentence, ES is a mark of the end of a sentence, BL is a blank character, C is a consonant, V is a vowel, LC is a lunar consonant, SC is a solar consonant, and LV is a long vowel. A conversion rule can be written as follows:

\[ \text{LFT} + \text{GR} + \text{RGT} \Rightarrow /\text{PH}/ \]

The rule is read as follows: a grapheme GR having respectively as left and right context LFT and RGT is converted to the phoneme PH. Left and right contexts could be a grapheme, a word separator, the beginning or the end of a sentence or empty.

3.2.1. \( \ddot{\text{A}}, \text{\( \ddot{\text{A}} \) and} \( \ddot{\text{E}} \) rules

In Algiers dialect, the letters \( \ddot{\text{A}} \), \( \ddot{\text{E}} \), \( \ddot{\text{D}} \), \( \ddot{\text{D}} \) and \( \ddot{\text{T}} \) are not used, they are in most cases pronounced respectively as the graphemes \( \ddot{\text{a}} \), \( \ddot{\text{e}} \), \( \ddot{\text{d}} \), \( \ddot{\text{d}} \) and \( \ddot{\text{t}} \).

\[ \{\text{C, V}\} + \ddot{\text{a}} + \{\text{C, V}\} \Rightarrow /\text{d}/ \]
\[ \{\text{C, V}\} + \ddot{\text{e}} + \{\text{C, V}\} \Rightarrow /\text{d}/ \]
\[ \{\text{C, V}\} + \ddot{\text{d}} + \{\text{C, V}\} \Rightarrow /\text{t}/ \]

3.2.2. Foreign letters rule

Algiers dialect alphabet corresponds to Arabic alphabet extended to three foreign letters g, v and p.

\[ \{\text{C, V}\} + \ddot{\text{g}} + \{\text{C, V}\} \Rightarrow /\text{g}/ \]
\[ \{\text{C, V}\} + \ddot{\text{v}} + \{\text{C, V}\} \Rightarrow /\text{v}/ \]
\[ \{\text{C, V}\} + \ddot{\text{p}} + \{\text{C, V}\} \Rightarrow /\text{p}/ \]

3.2.3. Definite article rule

When the definite article “\( \ddot{\text{A}} \)” is followed by a lunar consonant (which does not assimilate the "\( \ddot{\text{J}} \)"") the "\( \ddot{\text{A}} \)" is not pronounced.

\[ \{\text{LC}\} + \ddot{\text{A}} + \{\text{BL, BS}\} \Rightarrow /\text{a}/ + /\text{LC}/ \]

Example: "الشعب" (the nation) \( \Rightarrow /\text{al-sha\text{b}}/ \)

This rule is the same as in MSA with the difference that in MSA the "\( \ddot{\text{A}} \)" is pronounced if the definite article is in the beginning of the sentence.

When the definite article “\( \ddot{\text{A}} \)” is followed by a solar consonant the "\( \ddot{\text{A}} \)" is not pronounced and the consonant following it is doubled (gemination).

\[ \{\text{SC}\} + \ddot{\text{A}} + \{\text{BL, BS}\} \Rightarrow /\text{a}/ + /\text{SC}/ + /\text{SC}/ \]

Example: "النافورة" (the moon) \( \Rightarrow /\text{laqmar}/ \)

3.2.4. Words Case-ending

Words case ending in Algiers dialect is the Sukun (Absence of diacritics), so the last consonant of a word should be pronounced without any diacritic.

\[ \{\text{BL, ES}\} + \ddot{\text{A}} + \{\text{C, V}\} \Rightarrow /\text{C}/ \]
\[ \{\text{BL, ES}\} + \ddot{\text{A}} + \sigma + \{\text{C, V}\} \Rightarrow /\text{C}/ \]

Example: "كل" (before) \( \Rightarrow /\text{qbal}/ \)

3.2.5. Long vowel rules

When \( \ddot{\text{A}}, \ddot{\text{V}}, \) and \( \ddot{\text{E}} \) occur in a word preceded respectively by the short vowels \( \ddot{\text{a}}, \ddot{\text{e}}, \) and \( \ddot{\text{e}} \), their relative long vowels are generated.

\[ \{\text{C} + \ddot{\text{a}}\} + \ddot{\text{a}} + \{\text{C}\} \Rightarrow /\text{a}/ \]

Example: "الكأس" (a cup) \( \Rightarrow /\text{ka\text{s}}/ \)

\[ \{\text{C} + \ddot{\text{e}}\} + \ddot{\text{e}} + \{\text{C}\} \Rightarrow /\text{e}/ \]

Example: "الألوى" (a well) \( \Rightarrow /\text{bi\text{r}}/ \)

3.2.6. Glottal stop rule

In Algiers dialect, when a word begins with a Hamza, its phonetic representation begins with a glottal stop. In the end of a word the Hamza is not pronounced.

\[ \{\text{C, V}\} + \ddot{\text{a}} + \{\text{BS, BL}\} \Rightarrow /\text{a}/ \]

Example: "سكته" (stop talking) \( \Rightarrow /\text{askut}/ \)

\[ \{\text{BL, ES}\} + \ddot{\text{e}} + \{\text{C, V}\} \Rightarrow /\text{Null}/ \]

Example: "سم" (sky) \( \Rightarrow /\text{sm}/ \)

It should be noted that the hamza in the middle of the word is replaced by the long vowels \( \ddot{\text{a}}, \ddot{\text{e}} \) or \( \ddot{\text{e}} \) in Algiers dialect. For example the Arabic words "\( \ddot{\text{a}}\)" (a well) and "\( \ddot{\text{e}}\)" (a poleax) correspond respectively to /\text{bi\text{r}}/ and /\text{fa\text{s}}/.

3.2.7. Alif Maqsura \( \ddot{\text{E}} \) rule

Alif Maqsura \( \ddot{\text{Arabic}} \) (which is always preceded by a fatha /\text{a}/) at the end of a word is realized as the short vowel /\text{a}/.

\[ \{\text{BL, ES}\} + \ddot{\text{a}} + \{\text{C}\} \Rightarrow /\text{a}/ \]
5. EXPERIMENTS

5.1. Experimental Material

We used an Algiers dialect corpus constructed by hand. It consists of more than 6K sentences including 10.7k different words. This corpus is a collection of movies and TV shows and some everyday recorded conversations which we transcribed into text (in Arabic script) by hand by adopting writing rules. Initially, this corpus was not vocalized, we proceeded to vocalize it as explained in section 3.1.

The corpus is consisting of three categories of words:

1. Arabic words.
2. French words phonologically altered and their pronunciation is realized with Arabic phonemes.
3. French words for which the pronunciation is realized with French phonemes.

5.2. Results for Rule Based Approach

We applied phonetization rules exposed above on a corpus of Algiers dialect containing only Arabic words (8993 words) and French words phonologically altered (801 words), the system accuracy was 100%. In addition to Arabic words, French words in the second category are correctly phonetized because their phonetic realization is close to Algiers dialect. For example the word “کوزینه” (kitchen, original French word is "cuisine") which is a borrowed French word phonologically altered is correctly converted to /ku:zina/ while a word in the third category as "کونکسیون" (connection, original French word connexion) is incorrectly converted to /ku:niksju:n/ since it is realized /konneksij/ with French phonemes. Considering these words(906 words), the system accuracy decreases to 92%

The issue of these words is that we cannot introduce rules for French words written in Arabic script, since the relationship between Arabic graphemes and French phonemes is not a one-to-one. For example, the graphemes “ـ“, a French word written in Arabic script could correspond to the French phonemes /y/, /u/, /s/ or /O/ (see some examples in Table 2).
5.3. Results for Statistical approach

In order to evaluate the statistical approach, we needed a parallel corpus (text and its phonetic representation) which was not available. To build this resource we proceeded as follows: we used the rule-based system described below to convert Arabic words and French words phonologically altered (category 1 and 2) into Arabic phonemes. Whereas French words (of category 3) were identified and transliterated into Latin script by hand and then converted to French phonemes by using a free French converter. For example the word “شمس” is transliterated to “connexion” then converted to /kmeŋksjɔ̃/.

For test, we split randomly the parallel corpus into three datasets: training data (80%, 8560 words), tuning data (10%, 1070 words) and testing data (10%, 1070 words). Note that the test set includes 913 words of category 1, 77 words of category 2 and 80 words of category 3.

First we tested the statistical approach on a test set containing only Arabic words and French words phonologically altered (category 1 and 2). We got an accuracy of 93%. Then we proceeded to a test on a corpus including the three words categories, system accuracy decreases to 85%. This result is due to the increase of hypothesis number of each grapheme because of introducing French phonemes in the training data. The graphemes “شمس” for example in some Arabic words (category 1) are phonetized as the French phonemes /y/ or /ɛ/ instead of the Arabic long vowel /u:/, the phoneme /ɛ/ instead of /u:/.

Whereas, some words in category 3 are phonetized with Arabic phonemes by substituting for example the phonemes /y/, /u:/, /ɛ/ or /O/ by the /u/; /ɛ/ or /æ/ by /æ/.

5.4. Discussion

In Table 3 we summarize the achieved results. At first glance, and regarding the accuracy rate, we could deduce that rule-based approach is more efficient than statistical approach.

6. CONCLUSION

We presented two approaches to build a G2P converter for Algiers dialect which is an under-resourced language. The first approach is rule based; it gives perfect results for Arabic words and French words phonologically altered. Unfortunately, Algiers dialect includes a big proportion of French words pronounced as in the original language. This category of words could not be phonetized with a rule-based approach since the relationship between Arabic grapheme and French phoneme is not a one-to-one process. We used the statistical approach to overcome this problem. Regarding the small amount of training data, we can consider that the results are acceptable.

To improve the statistical G2P converter, French words pronounced as in the original language must be processed in a separate corpus. That is why we have to develop a module for identifying these words. We can use the rule-based approach for Arabic and French words phonologically altered. This hybrid approach could carry out better results.

In terms of resources, this work allowed us to build a phonetized dictionary for Algiers dialect; at our knowledge no such resource is available at this time for this language.

7. REFERENCES


