An Innovative Methodology to Transcribe Local Spoken Arabic Languages

Ahcene Farah, American University in the Emirates, UAE
Sameh Ghwanmeh, American University in the Emirates, UAE

The IAFOR International Conference on Language Learning – Dubai 2016
Official Conference Proceedings

Abstract
In this research we propose an innovative methodology, named Saoutharf, for standardizing writing and reading spoken local or regional Arabic languages. The transcription of spoken Arabic languages or Arabic dialects uses the standard Arabic letters themselves with four modified letters to write the letters sounds which do not exist in the standard Arabic. The proposed methodology targets speeches, cinema and drama, TV-talk-shows, advertising and Arabic in the street, where spoken Arabic is prevalent. Likewise, it targets learning spoken Arabic to foreigners. The objective is to make the rules of this methodology a standard which permits spoken Arabic languages to be transcribed the same way making texts readable by everybody, and transportable between people and locations, which is not the case today. Furthermore, UNESCO recommends that “Mother Tongue Matters and local language is a key to effective learning”.

When writing Arabic texts in a classic way, the vowels are not written explicitly but they are kept implicit. Without knowing the word, we cannot pronounce it correctly; Saoutharf vocalize the word making it easy to read. The traditional letters (ا و ي) are no longer used to sustain the sounds (double vowel) of implicit vowels, but they are used as explicit vowels like in Latin languages (English, French,...etc.). Further the transcription orthography provides also a clear and easy way to read and pronounce correctly standard Arabic. The proposed methodology will be implemented as a real-time computer based algorithm and would be valuable for information retrieval and web search algorithms for Arabic dialect content.

Keywords: Saoutharf, Arabic Dialects, Writing spoken Arabic, Arabic morphology
Introduction

After the speaking languages, the human being did a major step by the invention of writing, and thereafter making paper as a recording medium. Writing was and still a powerful instrument of communication, transmission and storage of information and learning. Fast learning a native or nonnative language depends primarily on the word formation or the morphology of the language, then its syntax (grammar rules) and semantics (word meanings). Colloquial Arabic or spoken Arabic has nor uniform morphology, neither same semantics due to its expansion through large geographical regions (Fig. 1), and the influence of non-Arabic local languages.

A document published by the UNESCO [1] in 2008 states that using mother tongue is highly recommended to effective learning. Standard Arabic is considered morphologically a complex language [2]. In a long article, Zaidan and Callison-Burch from Pennsylvania University state [3] that “The written form of the Arabic language, differs in a nontrivial manner from the various spoken regional dialects of Arabic speakers”. Further, it is well known that there is morphological ambiguities in Standards Arabic [4]; many words in Arabic have the same orthographic form, with different pronunciations. If we do not know Arabic, we cannot read Arabic texts. This problem militates for a new orthographic morphology to transcribe Arabic dialects even though we use the same alphabet as the Standard Arabic. In Saoutharf we use Arabic letters to overcome any psychological or cultural barriers that might be raised if Latin alphabet has been proposed (the term ‘dialect’ in the Arabic tradition of Arabic countries automatically carries a pejorative connotation due to the modern opposition between ‘dialect’ vs. Standard Arabic [5]).

In this Methodology we propose an innovative approach, named Saoutharf (literally letter sound), for standardizing writing and reading spoken regional Arabic languages [6]. We target the Morphology of these languages, that is to say the structure of words, with a little incursion in the syntax of the end of verbs and nouns in sentences. The transcription of spoken local Arabic languages or Arabic dialects uses the standard Arabic letters themselves plus four modified letters to write the letters sounds which do not exist in the standard Arabic. The transcription orthography provides a clear and easy way, without morphological ambiguities, for reading and stating the correct pronunciation for Arabic dialects as well as standard Arabic if this method is used, because it’s an effective phonetic writing. The proposed methodology will be implemented as a real-time computer based algorithm and would be valuable for information retrieval and web search algorithms for Arabic dialect content.
Methodology
As we said previously, spoken Arabic dialects are largely used in communication, cinema scripts and theater plays. The fact that they are used in the two last areas supposes that these languages are written even they are not widely published. When they are written they do not use a standard orthography and different orthographies may be used even within a same text [6]. We thus need a standard rules that permit to spoken Arabic to be written in the same way in all Arabic countries.

Rule one (writing syllables): In the methodology Saoutharf, the traditional letters (ا و ي) are no longer used to sustain the sounds (double vowel) of implicit vowels, but they are used as explicit vowels like in Latin languages (English, French etc.) like in (يكتب yaktab). This rule applies almost to all letters of the Arabic alphabet (three exceptions only reported in rule two).

Rule Two: a letter without added-vowel is pronounced like in English, and it corresponds to a letter with Sukun in Arabic like in (تكلم) [Ikamar]. However, we have three exceptions: ا is pronounced (a) like in [adam], د is pronounced like (i) in [islam] and ء at the end of the word is pronounced (a) like [ra a].

Rule Three: the letter for female nouns (ة) is no longer used like in [madrassa].

Rules Four: a letter with ashedda is doubled; the first letter is read with a Soukoun, the second is read as syllable with the next vowel like in [a shams], حاتم [hat tam], وريح [a rrih] or صورة [as ssoura].
Rule Five: In Arabic dialects, many sounds or syllables are coming from local languages (non-arabic languages or from Arabic but with different pronunciation or from English, etc.). For this purpose we have introduced four modifiers: Hamza, Ramza, Tarza and Elmed. The first three called Hamza, Ramza and Tarza) permit to transcribe the letters sounds which do not exist in the standard Arabic and the fourth one, Elmed which is generally not used in dialects speaking, may be used to write Sustained vowels (double-vowels) in standard Arabic or specific sound in spoken languages. These modifiers and some applications are listed in Table 1.

Table 1: Modifiers in Action.

<table>
<thead>
<tr>
<th>Modifier</th>
<th>Name</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>ء</td>
<td>Hamza</td>
<td>موبايل, Fatma</td>
</tr>
<tr>
<td>ﺍ</td>
<td>Ramza</td>
<td>أبي, Papa</td>
</tr>
<tr>
<td>ﺍ</td>
<td>Tarza</td>
<td>مات، salon</td>
</tr>
<tr>
<td>ﺍ</td>
<td>Elmed</td>
<td>مواء، حي</td>
</tr>
</tbody>
</table>

Rules Six (writing and reading sentences): The standard Arabic language grammar is well codified; but the formalism used tends to make reading and speaking more complex due the great variations of the words ends depending on the position, the type (subject, verb or complement), the number of people considered (singular, double, and plural-more-than-two) and the gender. Thus why the pronunciation of Arabic syntax is not used in dialects, even some time it is simplified by standard Arabic formal speakers on TV, Radio and conferences. We propose in Saoutharf methodology the followings:

a. Use only singular and plural for verbs and nouns (no double);
b. Use Soukoun at the end of nouns and singular verbs, when standard Arabic uses variations, like in [Hmad aktab wa msha].
c. Use gender only in singular verb like in [hakmat Fatima].
d. Use same and simplified plural for men and women like in [lbnat wafkou].

Writing Standard Arabic using Saoutharf Rules.
In the classic or modern Standard Arabic texts, the vowels are not written explicitly but they are kept implicit. Even when the word is included in a sentence, if we do not know the word, we cannot pronounce it correctly. Figure 2 shows an example cited from [4] which illustrate the case the word (يعد).

Figure 2: Legal Ambiguity in Arabic.

The methodology Saoutharf overcomes this ambiguity. If we rewrite the words, the result is: [yu’id (‘a’aada), yu’ud (‘aada), ya’idd (‘adda), ya’idd (‘adda)].
Southarf on the Web
We dedicated a Facebook page to the methodology. This page uses French language as explicative media and target essentially Algerian people; we have published three different texts chosen from Algeria: An extract from a popular song (Fig. 3), an extract from a film dialog (Fig. 4) and an extract from the national hymn (Fig. 5) transcribed in Southarf from a text written in standard Arabic (https://www.facebook.com/southarf/).

Figure 3: An extract from an Algerian popular song

Figure 4: An extract from a film dialog

Figure 5: An extract from the Algerian national hymn

Some statistics, obtained between 1/10/2015 and 29/02/2016, are pictured into Figure 6 and Figure 7. The number of like mentions in 4 months is 7359 (around 11000 total number of visitors), of which 57% are between 18 and 24 old. These statistics can be considered as indicators not as measures.
Figure 6: Total like mentions (7359)

Figure 7: Distribution of like mentions by age and gender (women 54%, men 46%)

Conclusion and Perspective
We propose a methodology for standardizing writing and reading spoken local or regional Arabic languages. The objective is to provide spoken Arabic languages or dialects with a methodology for writing and reading the same way everywhere, and making texts transportable between people and locations, which is not the case today. Despite that the methodology has been tested only in a set of few people, it showed fast learning capability. However, it has to be tested in real field in large scale. We plan to collaborate with specialists in communication and education to confirm the efficiency of the methodology. Further, the proposed system will be implemented as a real-time computer based algorithm and would be valuable for information retrieval and web search algorithms for Arabic content.
References


Contact email: ahcene.farah@aue.ae
Contact email: sameh.ghawanmeh@aue.ae