A Study of a Non-Resourced Language: The Case of one of the Algerian Dialects
Karima Meftouh, Najette Bouchemal, Kamel Smaïli

To cite this version:

HAL Id: hal-00727042
https://hal.archives-ouvertes.fr/hal-00727042
Submitted on 14 Sep 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
A STUDY OF A NON-RESOURCED LANGUAGE: THE CASE OF ONE OF THE ALGERIAN DIALECT

K. Meftouh, N. Bouchemal

UBMA
Badji Mokhtar University
Informatic Department
BP 12, 23000 Annaba, Algeria

K.Smaili
LORIA
Campus scientifique
BP 139, 54500 Vandoeuvre Lès Nancy Cedex, France

ABSTRACT
This paper presents a linguistic study of an algerian arabic dialect, namely the dialect of Annaba (AD). It also presents the methodology applied in the construction of a parallel corpus MSA-AD. This work is done in a future goal of developing a machine translation system of standard Arabic (MSA) to algerian arabic dialects.

Index Terms— Machine translation system, Standard Arabic, Algerian arabic dialect, parallel corpus, dialect of Annaba, cosine similarity measure

1. INTRODUCTION
Arabic is a Semitic language, it is used by around 250 million people, but is understood by up to four times more among Muslims around the world [1]. Arabic is a language divided into 3 separate groups: Classical written Arabic, written modern standard Arabic and spoken Arabic.

Classical written Arabic is principally defined as the Arabic used in the Qur’an and in the earliest literature from the arabian peninsula, but also forms the core of much literature up until our time. written modern standard Arabic (or MSA, also called Alfus’ha), is the variety of Arabic most widely used in print media, official documents, correspondence, education, and as a liturgical language. It is essentially a modern variant of classical Arabic. Standard Arabic is not acquired as a mother tongue, but rather it is learned as a second language at school and through exposure to formal broadcast programs (such as the daily news), religious practice, and print media. Spoken Arabic is often referred to as colloquial Arabic, dialects, or vernaculars. It’s a mixed form, which has many variations, and often a dominating influence from local languages (from before the introduction of arabic). Differences between the various variants of spoken Arabic can be large enough to make them incomprehensible to one another. Hence, regarding the large differences between such spoken languages, we can consider them as disparate languages or more exactly as different dialects depending on the geographical place in which they are practiced : Morocco, Algeria, Egypt, ...

In this paper, we will focus on algerian dialect. We have to understand that the concept of dialect here is different from what is admitted in west. In fact, people in their day life do not use standard Arabic but dialect, which is in most cases different from standard Arabic. Consequently, people who are not educated can not understand standard Arabic which is considered as a foreign language.

This work is part of a project TORJMAN1 which is dedicated to translating standard Arabic to algerian arabic dialect. Interest in such extremely complicated problem can be very surprising. In fact, it is difficult to understand this issue but when we analyze the spoken language in different places in Algeria for instance, we can notice that almost nobody speaks standard Arabic even if the official language of Algeria is standard Arabic. Furthermore, this spoken language is not written. The idea of this project is twofold, first understand the function and the underlying structure of algerian dialects and then provide the population and social-economic actors, a tool enabling the average user to understand the standard Arabic. We present in the following section (section 2) why should we be interested in algerian dialect.

2. WHY ARE WE INTERESTED IN COLLOQUIAL ARABIC?
We see at international conferences post September 11, 2001, a craze increasingly important for machine translation of standard Arabic to Indo-European languages. These studies are important when it comes to translating official documents, however if you want to develop applications for the average citizen, it is necessary to take into account his mother tongue, it means his dialect.

The main dialectal division is between the Maghreb dialects and those of the middle east, followed by that between sedentary dialects and bedouin ones.

Watson writes "Dialects of Arabic form a roughly continuous

1TORJMAN is a national research project which is totally financed by the algerian research ministry
3. ALGERIAN ARABIC

In Algeria, as elsewhere, spoken Arabic differs from written Arabic; Algerian Arabic has a vocabulary inspired from Arabic but the original words have been altered phonologically, with significant Berber substrates, and many new words and loanwords borrowed from French, Turkish, and Spanish. Like all Arabic dialects, Algerian Arabic has dropped the case endings of the written language. It is not used in schools, television or newspapers, which usually use standard Arabic or French, but is more likely, heard in music if not just heard in Algerian homes and on the street. Algerian Arabic is spoken daily by the vast majority of Algerians [5]. Algerian Arabic is part of the Maghreb Arabic dialect continuum, and fades into Moroccan Arabic and Tunisian Arabic along the respective borders. Algerian Arabic vocabulary is pretty much similar throughout Algeria, although the easterners sound closer to Tunisians while the westerners speak an Arabic closer to that of the Moroccans.

We focus, in this paper, on one of the easterners dialects of Algeria: Annaba’s dialect (AD). This choice is justified by the fact that this dialect is the one we know best. We present in section 4 its peculiarities.

4. SPECIFICITIES OF ANNABA’S DIALECT

To develop any application based on a language, at least a basic linguistic study is necessary even if we use a statistical model. In this section, we present the main features of the dialect of Annaba in which we are concerned.

Annaba’s dialect is spoken in the city of Annaba located east of Algeria. It is spoken by more than one million people. Like for Maghreb Arabic dialects, the most notable features of this dialect, is the collapse of short vowels in some positions. The word كتّاب (كتاب) in MSA correspond to كتّاب: the short vowel َ in kasra on the first consonant ﺔ in MSA is deleted in dialectal and replaced by the sukun.

In AD, the consonant ق q is generally pronounced ﻋ v. For example قال قال vāl. For some words both alternatives exist like the word قطع qata which can be also pronounced ﻋ vāta. We give in Table 2 a list of other consonants which pronunciation differs from standard Arabic, and their respective pronunciation.

<table>
<thead>
<tr>
<th>Consonant</th>
<th>Dialectal pronunciation</th>
</tr>
</thead>
<tbody>
<tr>
<td>ﺔ</td>
<td>ﻋ</td>
</tr>
<tr>
<td>ﺔ</td>
<td>ﻋ</td>
</tr>
<tr>
<td>ﺔ</td>
<td>ﺔ</td>
</tr>
</tbody>
</table>

The Hamza, which is very present in standard Arabic, is
avoided or bypassed by almost all the dialects including the one used in Annaba.
This is practically systematic in the middle of a word or at the end. Either it disappears altogether at the pronunciation, or it is replaced by ي like in مَرَداً which correspond respectively to mardah or جلطةā'ilah in MSA which correspond respectively to مَيْدَة maydah and جلطةā'ilah in dialect form. At the beginning of a word, the Hamza can be preserved as in the case of imperative form, it is replaced by ø end. Either it disappears altogether at the pronunciation, or this is practically systematic in the middle of a word or at the end. It’s intro-
duced after the noun to which the possessive refers, it then becomes necessary that that noun be defined by the addition of the defining article, as لكتَّاب ثَابَت (My book).

4.1. Personal pronouns
The personal pronoun appears in two forms:
a the separate form which is used in the nominative "I", "he", etc.
b the suffixed form which is used for the possessive "my", "his", etc., or for the objective "me", "him"

The first form stands alone, the second can only be used attached to a noun, verb, or certain particles.

• The Personal Pronoun: Separate Form.

Singular.
1. ي y is used for "My", for example كتَّابي كتَّابي (My book).
2. لَدَ k is used for "Your", as كتَّابك كتَّابك (Your book).
3. masc. هَوَّ k or هَوَّ "His" as كتَّابه كتَّابه (His book), هَوَّ (his brother); fem. هَا “Her", as كتَّبَهُ كتَّبَهُ (Her book).

Plural.
1. نَا nā is used for "Our", as داَرُنا داَرُنا (Our house).
2. دَارُكم dārūm (Your house).
3. هَمَّ hum is used for "Their", as داَرُهم dārūm (Their house).

In the case of feminine nouns ending with ظَ dābah as شَمَعَة shāmah, the suffixes are said to both plural masculine and feminine. The form تَعَدَ تَعَدَ combined with personal pronouns as suffixes is also used to denote property. It’s introduced after the noun to which the possessive refers, it then becomes necessary that that noun be defined by the addition of the defining article, as لكتَّاب ثَابَت (My book), داَرُكَمَ داَرُكَم (Your house)

4.2. Interrogatives
We list in table 3 the commonest forms of interrogative particles and pronouns used in the dialect of Annaba.

4.3. The interrogative sentence
Any dialectal sentence can be turned into a question in any one of two ways.
Table 3. Interrogative particles and pronouns in AD and their equivalents in MSA.

<table>
<thead>
<tr>
<th>English</th>
<th>Annaba dial.</th>
<th>MSA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Who</td>
<td>ſkán</td>
<td>من man</td>
</tr>
<tr>
<td>Which</td>
<td>wana</td>
<td>أي yu</td>
</tr>
<tr>
<td>Where</td>
<td>wajn</td>
<td>أين yna</td>
</tr>
<tr>
<td>What</td>
<td>wšiyā</td>
<td>ماذā madā</td>
</tr>
<tr>
<td>When</td>
<td>waqtāš</td>
<td>متي matā</td>
</tr>
</tbody>
</table>
| Why           | waltāš       | كيف kayfa
| How           | kifjāš        |             |

4.4. The negative sentence

The form مشر maš (Not) is in general use as a negative particle, and may be found with all the persons. It can also be combined with the personal pronouns to get negatives: مشنī mašnī (I am not); مشدا maškūm (You are not); مشها mašnā (We are not); مشهو mašā (He is not); مشه maššī (She is not) and مشهم mašhum (They are not). The negative sentence can also be obtained by adding affixes مā mā (as a prefix) and ِش ِش s (as a suffix) to verbs. Table 4 gives examples of negative sentences.

Table 4. Negative sentences

<table>
<thead>
<tr>
<th>English</th>
<th>Annaba Dialect</th>
</tr>
</thead>
<tbody>
<tr>
<td>I do not go</td>
<td>مشر راية maš rāyāh</td>
</tr>
<tr>
<td>I do not remember</td>
<td>مشنی متفکر mašnī matfakar</td>
</tr>
<tr>
<td>You did not eat</td>
<td>ما كنتي什 måklišt</td>
</tr>
</tbody>
</table>

4.5. Pluralization

Algerian Arabic uses broken and regular plural. Like all other Arabic dialects, suffix ُن wn used for the nominative in classical Arabic is no longer in use in regular plural. Suffix ُن bn used in classical Arabic for the accusative and the genitive is used for all cases. For example the plural of ميمن ميمن mūman (believer) is ميمن ميمن måmmûn.

For feminine nouns, the plural is mostly regular (obtained by postfixing -at): the plural of بنت bant (girl) is بنت bant. For some words the broken plural is used: like طواب طواب tavālah which is the plural of طابة tavālah (table).

We have listed in the foregoing, the main features of the dialect of Annaba. We will now present how we proceeded to develop corpora for use in a statistical translation system.

5. COLLECTING CORPORAS

The statistical translation approach and availability of tools ready-to-use allow us to build quickly a machine translation system with sufficient parallel training data. For the translation to (or from) an under-resourced language, this type of parallel corpora does not always exist, or exist with only a small amount of insufficient data for learning robust probabilistic models. In the case of Annaba’s dialect, there is no corpus that can be used to develop a translation system. We start this project from scratch. For the construction of such corpus, a first step is to establish a standard bilingual dictionary Arabic - arabic dialect. Thus the dictionary will contain entries like this: أعر َلر اذرب Azrib - اذرب. This entry is the word corresponding to “act quickly” which is translated into the dialect of Annaba by: آعر أعر. The constitution of this dictionary is the first stone of the building which will subsequently build the corpus.

To build the dictionary and consequently the corpus, we made recordings of discussions “in live” in different environments (medical offices, cafes, markets, ...) to ensure a large variety of vocabulary used. Afterward we performed a manual transcription of these recordings and extracted all words. Subsequently, we have assigned, to each extracted word, the arabic form which best fits. This resulted in a dictionary MSA-Annaba’s dialect and a written dialectal corpus. We give in table5 a sample of this dictionary. To complete the construction of the parallel corpus, we performed the translation of the dialect of Annaba to MSA based on the developed dictionary. A sample of this corpus is given in figure 1.

Table 5. A sample of the dictionary MSA-Annaba’s dialect.

<table>
<thead>
<tr>
<th>Annaba Dialect</th>
<th>MSA</th>
</tr>
</thead>
<tbody>
<tr>
<td>جريت girit</td>
<td>جريت garaytu</td>
</tr>
<tr>
<td>لبخان ignān</td>
<td>لمستان al-bustān</td>
</tr>
<tr>
<td>لبخان ignān</td>
<td>الديقة al-hadīqah</td>
</tr>
<tr>
<td>ورا wrā</td>
<td>ورو warāa</td>
</tr>
<tr>
<td>خكس hallas</td>
<td>مصد مصد saddada</td>
</tr>
<tr>
<td>خليك مهيا</td>
<td>دعاء مهيا darka minhā</td>
</tr>
</tbody>
</table>
6. ENRICHING CORPORA

As noted above, a machine translation system requires a large amount of data. However, in order to increase the size of our corpora, we propose to produce new sentences from the initial corpus. Producing new sentences is done by replacing each word in the original sentence by its different synonyms. Each time a word is replaced by its synonym will produce a new sentence which is added to the initial corpus. For the development of such a tool, we must necessarily start by the development of two dictionaries: one containing synonyms in AD and the other synonyms in MSA. To this end, we used the MSA-AD dictionary. We have assigned to each entry (of dialect or MSA) one or several synonymous words if they exist. This tool uses the dictionaries of synonyms to produce all possible sentences by combination. Once the sentences are generated, they are added to the appropriate corpus.

7. IS THE MSA-AD CORPUS PARALLEL?

In this section we show that the corpus we have built is really parallel. To this end, we selected the most commonly used measure in this area called "cosine similarity"[7]. The cosine of null angle is 1, and less than 1 for any other angle; the lowest value of the cosine is −1. The cosine of the angle between two vectors thus determines whether two vectors are pointing in roughly the same direction. This is often used to compare documents in text mining. The vectors used in this case consist of normalized frequencies of words. So, we have computed and normalized word frequencies for each of the corpora⁵ to constitute the vectors. We have taken vectors of different size each time to determine from what size the corpora became very close. In order to interpret these values, we compared to those obtained with the BAF corpus⁶. The values in terms of cosine, for our corpus and the BAF one, were very close, so the curves were juxtaposed. In order to have more demonstrative curves, we chose to use their respective angles (see figure 2).

We note that the curves are similar. The more we increase the size of the vectors the more the angles tend to zero. We can therefore confirm that MSA and AD corpora are parallel.

8. THE DIALECT'S VOCABULARY

In this section we focus on the study of dialect’s vocabulary. We notice that there are three types of words:

- Arabized borrowed words: are words belonging to foreign vocabulary (most of them are words borrowed from French), which were introduced in the dialect after having been naturalized phonetically and/or morphologically. Examples of such words are given in table 6.

<table>
<thead>
<tr>
<th>English</th>
<th>Annaba Dialect</th>
<th>Origin</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nurse</td>
<td>فرَمْلَي، einfirmer</td>
<td>French &quot;Infirmier&quot;</td>
</tr>
<tr>
<td>Place</td>
<td>بلاص، bilasah</td>
<td>French &quot;Place&quot;</td>
</tr>
<tr>
<td>That’s enough</td>
<td>يَجِبُ</td>
<td>Berber</td>
</tr>
<tr>
<td>Ship</td>
<td>بور، babûr</td>
<td>Turk</td>
</tr>
</tbody>
</table>

- Words that have unknown origin like صوارد şwârad, صوارد Money, مهبول mahbûl, مرعب Crazy...

- Arabic words: The dialect of Annaba is largely based on the standard Arabic. However, the words of arab origin have undergone some distortions. In order to determine these distortions, we computed the Levenshtein distance. The results showed that the deformations performed on arabic word are:
  - In pronunciation: all consonants occur in the dialectal word but the short vowels are changed. In such cases, the Levenshtein distance is zero.
- By insertion, deletion or substitution of consonants.

Table 7 provides examples of dialect words, their equivalents in standard Arabic and their corresponding Levenshtein distance.

Table 7. Levenshtein distance for dialect words and their equivalents in MSA

<table>
<thead>
<tr>
<th>MSA</th>
<th>Annaba Dialect</th>
<th>Lev. dist.</th>
</tr>
</thead>
<tbody>
<tr>
<td>تعرف</td>
<td>تعرف taraf</td>
<td>0</td>
</tr>
<tr>
<td>تكون</td>
<td>تكون tkān</td>
<td>0</td>
</tr>
<tr>
<td>لبحر</td>
<td>لبحر lbḥar</td>
<td>1</td>
</tr>
<tr>
<td>يحاسبونه</td>
<td>يحاسبونه yḥāsbūnahu</td>
<td>1</td>
</tr>
<tr>
<td>الأيام</td>
<td>أيام liyām</td>
<td>2</td>
</tr>
<tr>
<td>أشتريه</td>
<td>شرته nišrīh</td>
<td>2</td>
</tr>
<tr>
<td>يعینه</td>
<td>يعیونو yāwnū</td>
<td>3</td>
</tr>
<tr>
<td>وكل</td>
<td>ماکله māklaḥ</td>
<td>3</td>
</tr>
</tbody>
</table>

9. CONCLUSION

In this paper, we have presented the main features of the dialect of Annaba through a linguistic study. We believe we are the first to do this study. As we have already specified above, this work is part of a project TORJMAN which is dedicated to translating standard Arabic to Algerian Arabic dialect. To build a machine translation system a sufficient parallel training data is necessary. In the case of Annaba’s dialect, there is no corpus that can be used. So, to build the corpus, we performed recordings of dialect we transcribed. We subsequently developed AD-MSA dictionary that we used to translate the dialect corpus in standard Arabic. We demonstrated that the built corpus is parallel using cosine similarity measure. We have also presented a study of the dialect’s vocabulary which has shown that it is mainly inspired from standard Arabic. The development of a machine translation system is subject to our future work.

10. REFERENCES


