# A Study of a Non-Resourced Language: The Case of one of the Algerian Dialects

Karima Meftouh, Najette Bouchemal, Kamel Smaïli

# 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 *TORJMAN*[1] 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 arabic 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*

[1] *TORJMAN* is a national research project which is totally financed by the algerian research ministry

*spectrum of variation, with the dialects spoken in the eastern and western extremes of the Arab-speaking world being mutually unintelligible*" [2]. Effectively, while middle easterners can generally understand one another, they often have trouble understanding Maghrebis[2]. Although the converse is not true, due to the popularity of middle eastern, especially egyptian, films and other media. In some cases people from these countries are unable to understand each other, at most few words are unknown for them [3]. In other cases, people from one of the concerned country could find the grammatical structure of the neighbor country bit understandable. Table1 provides a simple, yet interesting, example of how spoken varieties of Arabic differ in intelligibility. The English sentence *I am going now* is given in the syrian, egyptian, tunisian and algerian dialects and in MSA with their respective transliteration.

**Table 1**. Variants of arabic dialects expressing the English sentence *I am going now*

| | |
|---|---|
| MSA | أنا ذاهب الآن *ʾanā ḏāhibun ālʿān* |
| Egyptian | أنا رايح دلوءتي *ʾanā rāyiḥ dilwʿty* |
| Syrian | راح روح هلّ *rāḥ rūḥ halla* |
| Tunisian | باش نمشي توى *bāš nimšy tawā* |
| Algerian | راح نروح درك *rāḥ nrūḥ durk* |
| Moroccan | أنا غادي دب *ʾanā ġādy daba* |

These examples reflect clearly the distance between dialectal sentences expressing the same idea. If we consider only the word الآن *ālʿān* (*Now*) in MSA, we remark that its equivalent in each of the considered dialects differs from that used in the others: دلوءتي *dilwʿty* in egyptian, هلّ *halla* in syrian, توى *tawā* in tunisian, درك *durk* in algerian and دب *daba* in moroccan.

Now let us consider maghreb spoken languages. There are clearly two native languages in Morocco and Algeria, algerian or moroccan Arabic and Berber[3] (respectively 40 to 50% of Berbers in Morocco, and 25 to 30% in Algeria). In Tunisia, there are only few Berbers (1 or 2%). In addition, the number of monolingual berbers in rural areas is not negligible. On the other hand, the most optimistic estimates of illiteracy is 50% in Morocco, Algeria 26% and 23% in Tunisia [4]. MSA is therefore still possessed by a small minority. So, much of the population is monolingual in Arabic moroccan, algerian or tunisian or bilingual berber/arab moroccan or algerian, with snippets of standard Arabic and French.

[2]People from Tunisia, Algeria and Morocco

[3]Berber or berber languages are a family of similar or closely related languages and dialects indigenous to North Africa.

## 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 كِتَابْ *kitāb* (*book*) in MSA correspond to كْتَابْ *ktāb*: the short vowel إِ *i kasra* on the first consonant ك *k-* in MSA is deleted in dialectal and replaced by the *sukūn* .

In AD, the consonant ق *q* is generally pronounced ڤ *v*. For example قال *qāl* (to say) is pronounced ڤال *vāl*. For some words both alternatives exist like the word قطع *qṭaʿ* which can be also pronounced ڤطع *vṭaʿ*. We give in Table 2 a list of other consonants which pronunciation differs from standard Arabic, and their respective pronunciation.

**Table 2**. Arabic consonant and their dialectal pronunciation

| **Consonant** | **pronunciation** |
|---|---|
| ذ *ḏ* | د *d* |
| ث *ṯ* | ت *t* |
| ظ *ẓ* | ض *ḍ* |

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 *y* like in مَائِدَة *mā'idah* or عَائِلَة *'ā'ilah* in MSA which correspond respectively to مَيْدَهْ *maydah* and عَايْلَهْ *'aylah* in dialect form. At the beginning of a word, the Hamza can be preserved as in the case of imperative form, for example أُدْخُلْ *'udẖul* (*enter*). However, it disappears automatically if it is preceded by the article ال *āl* (*the*), لتنين *ltnīn* (الإثنين *āl-'iṯnayn* in MSA). We give in the following other dialectal characteristics and we begin with the personal pronouns used.

### 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. أنا *'anā*, أني *'anī* (*I*).
  2. masc. نتَ *nta*; fem. نتِ *nti* (*You*).
  3. masc. هو *huwa* (*He*); fem. هي *hiya* (*She*).

  Plural.

  1. حنايا *ḥnāyā*, إحنا *'iḥnā* (*We*).
  2. نتوما *ntūmā* or إنتم *'intum* (*You*) is said to both plural masculine and feminine.
  3. هوما *hūmā* (*They*) also is said to both plural masculine and feminine.

  It is generally possible to omit the personal pronoun when it is obvious, thus when we ask someone "are you thirsty ?", we will just say عطشان؟ *'aṭšān?* = "thirsty ?".
  Very often a personal pronoun is added to a word already defined, and this added pronoun may become necessary when the predicate is also defined. Thus used the pronoun seems to be an equivalent to the verb "to be"[6], أنا هو لحفاف *'anā huwa lḥafāf* (*I am the hairdresser*).

- The personal pronoun as suffixes.
  We have already mention that for possessive such as "my", "his", "our", etc., or objective such as "me", "him", "us",etc., a different system is employed and the pronoun is expressed by a shortened form which is added to the end of a noun, verb, or certain particles. The suffixes thus used are:
  Singular.

  1. ي *y* is used for "My", for example كتابي *ktābī* (*My book*).
  2. ك *k* is used for "Your", as كتابك *ktābik* (*Your book*).
  3. masc. و *ū* or ه *h* "His" as كتابو *ktābū* (*His book*), خوه *ẖūh* (*his brother*); fem. ها *hā* "Her", as كتابها *ktābhā* (*Her book*).

  Plural.

  1. نا *nā* is used for "Our", دارنا *dārnā* (*Our house*).
  2. كم *kum* is used for "Your", as داركم *dārkum* (*Your house*).
  3. هم *hum* is used for "Their", as دارهم *dārhum* (*Their house*).

  In the case of feminine nouns ending with ة *h ta' marbūta* as شمعة *šam'ah*, the suffixes are تي *tī*, تك *tk*, تها *thā*, ...
  The form تاع *tā'* 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 becoming necessary that that noun be defined by the addition of the defining article, as لكتاب تاعي *lktāb tāī* (*My book*), الدّار تاعكم *ad-dār tā'kum* (*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.

| English | Annaba dial. | MSA |
|---|---|---|
| Who | شكون *škūn* | من *man* |
| Which | ونا *wanā* | أي *ʾayu* |
| Where | وين *wayn* | أين *ʾayna* |
| What | وشيا *wšiyā*<br>وش *wš* | ماذا *māḏā* |
| When | وقتاش *waqtāš* | متى *matā* |
| Why | وعلاش *waʿlāš* | لماذا *limāḏā* |
| How | كفاش *kifāš* | كيف *kayfa* |

1. It may be spoken in an interrogative tone of voice, like راح تقرا؟ *rāḥ taqrā?* (*Will you revise?*).

2. An interrogative pronoun or compound derived from a pronoun may be used, as ويني داركم ؟ *wayniya dārkum?* (*where is your house?*).

### 4.4. The negative sentence

The form مش *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 [4] to get negatives: مشني *mašnī* (*I am not*); مشك *mašk*, مشكم *maškum* (*You are not*); مشنا *mašnā* (*We are not*); مشو *mašū* (*He is not*); مشي *mašī* (*She is not*) and مشهم *mašhum* (*They are not*). The negative sentence can also be obtained by adding affixes ما *mā* (as a prefix) and ش *š* (as a suffix) to verbs. Table 4 gives examples of negative sentences.

**Table 4**. Negative sentences

| English | Annaba Dialect |
|---|---|
| I do not go | مش رايح *maš rāyaḥ* |
| I do not remember | مشني متفكر *mašnī matfakar* |
| You did not eat | ماكلتيش *māklitīš* |

### 4.5. Pluralization

Algerian Arabic uses broken and regular plural. Like all other arabic dialects, suffix ون *wn* used for the nominative in classical Arabic is no longer in use in regular plural. Suffix ين *yn* used in classical Arabic for the accusative and the genitive is used for all cases. For example the plural of مومن *mūman* (*believer*) is مومنين *mūmnīn*.

[4] We are referring here to personal pronouns as suffixes

For feminine nouns, the plural is mostly regular (obtained by postfixing ات *-at*): the plural of بنت *bant* (*girl*) is بنات *bn-at*. For some words the broken plural is used: like طوابل *ṭw-abl* which is the plural of طابلة *ṭāblah* (*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 CORPORA

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: أسرع *ʾsriʿ* → إزرب *ʾizrib*. This entry is the word corresponding to "act quickly" which is translated into the dialect of Annaba by: إزرب *ʾizrib*. 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.

| Annaba Dialect | MSA |
|---|---|
| جريت *ǧrīt* | جريت *ǧaraytu* |
| لجنان *lǧnān* | البستان *āl-bustān* |
| لجنان *lǧnān* | الحديقة *āl-ḥadīqah* |
| ورا *wrā* | وراء *warāʾa* |
| خلّص *ẖallaṣ* | سدّد *saddada* |
| خلّيك منها *ẖallīk minhā* | دعك منها *daʿka minhā* |

| | |
|---|---|
| راكي ترهـيزي من صباح و ماغلوبتيش. | أنت تراجعين منذ الصباح و لم تتعب بعد. |
| نحس في راسي حيتـفلـــــق. | أحس أنّ رأسي سينفجر. |
| غدوا عندي إمتحان و مزلت ماكملتش. | غدا لديّ إمتحان و لم أكمل بعد. |
| خدمت في صبيطار ڤريب من دارنا الحمدوله | إشتغلت في مستشفى قريب من بيتنا الحمد لله أنا |
| راني لاباس و نعيش معا بابا. | بخير و أعيش مع والدي. |
| كنت نخدم كوافور | كنت أعمل حلاقا |
| في الدوّار لي نسكن فيه أنا ما كان حتى حاجه | في الريف الذي أسكن فيه لا يوجد أي شيء |
| ديفولي باها على روحك. | ترفه به عن نفسك. |

**Fig. 1**. A sample of parallel corpus MSA-Annaba's dialect.

## 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 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 corpora[5] 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 corpus[8]. 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.

[5]Here we are referring to AD and MSA corpora.

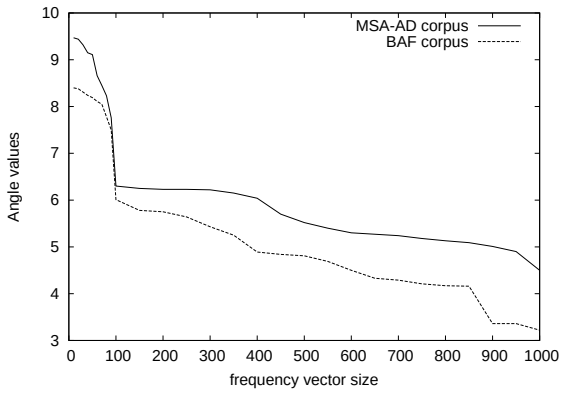


**Fig. 2**. The cosine similarity for BAF and MSA-AD corpora

## 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 6**. Examples of Arabized Foreign words

| English | Annaba Dialect | Origin |
|---|---|---|
| Nurse | فرملي *farmlī* *infirmier* | French "Infirmier" |
| Place | بلاصه *blāṣah* | French "Place" |
| That's enough | يزّي *yizzī* | Berber |
| Ship | ببور *babūr* | Turk |

- Words that have unknown origin like صوارد *ṣwārad Money*, مهبول *mahbūl 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

| MSA | Annaba Dialect | Lev. dist. |
|---|---|---|
| تعرف *taʕrif* | تعرف *taʕraf* | 0 |
| تكون *takūn* | تكون *tkūn* | 0 |
| البحر *āl-baḥr* | لبحر *lbḥar* | 1 |
| يحاسبونه *yuḥāsibūnahu* | يحاسبوه *yḥāsbūh* | 1 |
| الأيام *āl-ʔayām* | ليام *liyām* | 2 |
| أشتريه *ʔaštarīh* | نشريه *nišrīh* | 2 |
| يعينه *yuʕīnuh* | يعاونو *yʕāwnū* | 3 |
| أكل *ʔakl* | ماكله *māklah* | 3 |

## 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


[1] Abdel Monem A., Shaalan K., Rafea A., and Baraka H., "Generating arabic text in multilingual speech-to-speech machine translation framework," in *Machine Translation, Springer*, 2009.

[2] Jeremy Palmer, "Arabic diglossia: Teaching only the standard variety is a disservice to students," http://w3.coh.arizona.edu/AWP/AWP14/Palmer.pdf, 2007.

[3] Barkat-Defradas M., Al-Tamimi J., and Benkirane T., "Phonetic variation in production and perception of speech : a comparative study of two arabic dialects.," in *proc. of the 15th International Congress of Phonetic Sciences (ICPhS)*, 2003.

[4] Dominique Caubet, "Arabe maghrebin," http://corpusdelaparole.in2p3.fr/spip.php.

[5] Boucherit A., *L'Arabe parlé à Alger*, ANEP Edition, 2002.

[6] De Lacy O'Leary, "Colloquial arabic," http://www.archive.org/details/colloquialarabic00oleauoft, Digitized by the Internet Archive in 2007 with funding from Microsoft corporation.

[7] Salton G. and Mac Gill M.J, *Introduction to modern information retrieval*, International student Edition, 1983.

[8] Langlais P., Simard M., and Veronis J. et al., "Arcade: A cooperative research project on parallel text alignment evaluation," http: www.lpl.univ-aix.fr/projects/arcade, 1998.