ISCA Archive
http://www.isca-speech.org/archive

Third Workshop on Spoken Language
Technologies for Under-resourced Languages
Cape Town, South Africa
May 7-9, 2012

A STUDY OF A NON-RESOURCED LANGUAGE: AN ALGERIAN DIALECT

K. Meftouh, N. Bouchemal

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

K. Smaïli

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

ABSTRACT

The objective of this paper is to present an under-resourced language related to Arabic. In fact, in several countries through the Arabic world, no one speaks the modern standard Arabic language. People speak something which is inspired from Arabic but could be very different from the modern standard Arabic. This one is reserved for the official broadcast news, official discourses and so on. The study of dialect is more difficult than any other natural language because it should be noted that this language is not written. This paper presents a linguistic study of an Algerian Arabic dialect, namely the dialect of Annaba (AD). In our knowledge, this is the first study made on Algerian dialect. It also presents the methodology used for building a parallel corpus: modern standard Arabic versus Arabic Dialect in order to achieve a machine translation for this pair of languages. This preliminary work is presented to try to attract the attention of the scientific community to this difficult and challenging problem. A realistic machine translation on Arabic should be done principally on dialect. This is our objective at a medium term.

Index Terms— Standard Arabic, Algerian Arabic dialect, parallel corpus, dialect of Annaba, distance of Levenshtein, Machine translation system.

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 which was retained as the official language in all Arab countries, and as a common language between them [2]. It is 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 [3].

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)[4]. Differences between various spoken Arabic can be large enough to make them incomprehensible from one region to another one. 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.

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.

The work presented here is part of the project *TORJMAN*[1] which is dedicated to translating standard Arabic to Algerian Arabic dialect. Interest in such 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, to make this issue more difficult, 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 user to understand the standard Arabic.

[1] *TORJMAN* is supported by the Algerian research ministry.

2. WHY ARE WE INTERESTED IN COLLOQUIAL ARABIC?

The existence of dialects of the language is a challenge for Natural Language Processing (NLP) in general, as it adds another set of dimensions of variation from a known standard. The problem is particularly interesting in Arabic and its dialects. Any practical and realistic approach of the treatment of Arabic should report the use of dialect, because it is ubiquitous [5].

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 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 spectrum of variation, with the dialects spoken in the eastern and western extremes of the Arab-speaking world being mutually unintelligible*" [6]. 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, movies and other media. In some cases, people from these countries are unable to understand each other, at most few words are unknown for them [7]. 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 Syrian, Egyptian, Tunisian, Algerian and Moroccan 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 ālʾān*
Egyptian	أنا رايح دلوءتي *ʾanā rāyiḥ dilwʾty*
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 phonetic and lexical distance between dialectal sentences expressing exactly the same idea. If we consider only the word الآن *ālʾān* (*Now*) in MSA, we remark that its equivalent in each of the considered dialects differs from that used in the others: دلوءتي *dilwʾty* 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 [8]. 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.

3. ALGERIAN SPOKEN 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 songs if not just heard in Algerian homes and on the street. Algerian Arabic is spoken daily by the vast majority of Algerians [9].

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 and practice. We present in section 4 its peculiarities.

4. SPECIFICITIES OF ANNABA'S DIALECT

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

Annaba's dialect (AD) is spoken in the city of Annaba located at the east of Algeria. It is spoken by more than one million people. Like for Maghreb Arabic dialects, the most

[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.

notable features of this dialect, is the collapse of short vowels[4] in some positions. The word كِتَابْ *kitāb* (*book*) in MSA corresponds to كْتَابْ *ktāb*: the short vowel إِ *i kasra* on the first consonant ک *k-* in MSA is deleted in dialectal and replaced by the *sukūn*[5] .
In AD, the consonant ق *q* is generally pronounced ڨ *v*. For example قال *qāl* (to say) is pronounced ڨال *vāl*. For some words both alternatives exist like the word قطع *qṭaʿ* which can be also pronounced ڨطع *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[6], 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 *māʾidah* or عَائِلَة *ʿāʾilah* in MSA which correspond respectively to مَيْدَةْ *maydah* and عَايْلَةْ *ʿāylah* in dialect form. At the beginning of a word, the Hamza can be preserved as in the case of imperative form, for example أُدْخُلْ *ʾudḫul* (*enter*). However, it disappears automatically if it is preceded by the article ال *āl* (*the*), لتنين *ltnīn* (الإثنين *āl-ʾiṯnayn* in MSA). We give in the following, other dialectal characteristics.

4.1. Personal pronouns

The personal pronoun appears in two forms. The separate pronouns which are used in the nominative form (the equivalent of ”I”, ”he”, etc) are separated from other words. The suffixed pronouns which are used in the possessive (the equivalent of ”my”, ”his”, etc), or in the objective form (the equivalent of ”me”, ”him”) are attached to nouns, verbs or certain particles.

4.1.1. The separate personal pronouns

Let’s distinguish in the following their use in singular and plural forms.

- Singular form
  1. أنا *ʾanā*, أني *ʾanī* (*I*).
  2. masc. نتَ *nta*; fem. نتِ *nti* (*You*).
  3. masc. هو *huwa* (*He*); fem. هي *hiya* (*She*).
- Plural form
  1. حنايا *ḥnāyā*, إحنا *ʾiḥnā* (*We*).
  2. نتوما *ntūmā* or إنتم *ʾintum* (*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 there is no ambiguity about the speaker, as when we ask someone ”are you thirsty ?”, we will just say عطشان؟ *ʿaṭšān?* = ”thirsty ?”.
Sometimes, a personal pronoun is added to a word already defined, 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”[11], أنا هو لحفاف *ʾanā huwa lḥafāf* (*I am the hairdresser*).

4.2. The suffixed pronouns

In this case the pronoun is expressed by a shortened form which is added to the end of a noun, verb or certain particles. The used suffixes in singular and plural forms are:

- Singular form
  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. For masculine form و *ū* or ه *h* ”His” as كتابو *ktābū* (*His book*), خوه *ḫūh* (*his brother*) and for feminine ها *hā* ”Her”, as كتابها *ktābhā* (*Her book*).
- Plural form
  1. نا *nā* is used for ”Our”, دارنا *dārnā* (*Our house*).

[4] Arabic has three short vowels (a, i, u) which are not part of Arabic alphabet. When anyone of these vowels is coupled with a consonant, the consonant develop a unique sound (see [10] for more details).

[5] A consonant can carry no vowels for the default sound of the letter. No vowels means the consonant carries a *sukūn*.

[6] The Hamza is a letter in the Arabic alphabet, representing the glottal stop.

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 taʾmarbūta* as شمعة *šamʿah*, the suffixes are تي *tī*, تك *tk*, تها *thā*, ...
The form تاع *tāʿ* combined with personal pronouns as suffixes is used to insist that something belongs to someone. It is introduced after the noun to which the possessive refers as in لكتاب تاعي *lktāb tāʿī* (*My book*), الدّار تاعكم *ad-dār tāʿkum* (*Your house*).

4.3. Interrogative form

We list in table 3 the most common forms of interrogative particles and pronouns used in the dialect of Annaba. We can notice that in comparison to MSA no dialect word corresponds to the original pronoun, except the third one وين *wayn* which is a modified form of ayna أين *ʾayna*.
In this dialect, any sentence can be turned into a question, in any one of two ways.

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 وين جات داركم ؟ *wayn ǧāt dārkum?* (*where is your house?*).

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*
Where	منين *mnīn*	من أين *min ʾayn*
What	وشيا *wšiyā* وش *wš*	ماذا *mādā*
With what	باش *bāš*	بماذا *bimādā*
When	وقتاش *waqtāš*	متى *matā*
Why	وعلاش *waʿlāš*	لماذا *limādā*
How	كفاش *kifāš*	كيف *kayfa*
How many	قداش *vidāš*	كم *kam*

4.4. The negative sentence

The form مش *maš* (*Not*) is in general used as a negative particle, and may be found with all person form. It can also be combined with the personal pronouns [7] to get negative sentences: 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 illustrates some 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. The plural form

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*.
For feminine nouns, the plural is mostly regular (obtained by postfixing ات *-at*): the plural of بنت *bant* (*girl*) is بنات *bn-at*. Here, the grammatical rule used to constitute plural forms for our dialect is the same one used in MSA. For some words the broken plural is used: like طوابل *ṭw-abl* which is the plural of طابلة *ṭāblah* (*table*). Let's remark here that, the word طابلة *ṭāblah* is loaned from the French Language and correspond to the word ”Table”.
We listed in the foregoing, the main features of the dialect of Annaba in order to understand the particularity of this language before doing any particular processing. In the following we will present how we proceed for developing a parallel corpus in order to use it in a future work in statistical translation machine.

5. COLLECTING CORPORA

Statistical translation approach and availability of tools ready-to-use, allow us to build quickly a machine translation system when sufficient parallel training data are available. Unfortunately for an under-resourced language, this kind of parallel corpora does not always exist, or does exist with only a small amount of insufficient data. For Annaba's dialect and for any

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

other Algerian dialect, there is no corpus that can be used to develop a translation system. In this project, we start then from scratch. To build a such corpus, a first step is to establish a standard bilingual dictionary Arabic - Arabic dialect. The dictionary will contain entries as: أسرع *ʾsriʿ* → إزرب *ʾizrib* which corresponds to ”hurry up”. Developing a dictionary is the first stone which will allow us to get a corpus.

To build the dictionary and consequently the corpus, we recorded discussions ”in live” in different environments (medical offices, pubs, markets, ...) to ensure a large variety of the vocabulary. These recordings contained very noisy segments that have been removed. Ultimately, we selected the equivalent of 10 hours of speech transcribable.

This approach could be surprising, but we would like to remind that this language has never been written, then no reference to a list of words nor sentences is available somewhere in our knowledge. Afterward we performed a manual transcription of these recordings and extracted all words which will be analyzed. Subsequently, we assigned, to each extracted word, the corresponding standard Arabic form which best fits. This achieves a dictionary MSA-Annaba's dialect and a written dialect corpus. In table 5, a sample of this dictionary is given. We built a parallel corpus dialect-standard Arabic by translating by hand the transcribed corpus. Manual transcription and translation are very costly in time. We have managed to transcribe for the moment, only 30% of recordings.

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 *ḫallaṣ*	سدّد *saddada*
خلّيك منها *ḫallīk minhā*	دعك منها *daʿka minhā*

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. We have assigned to each entry (of dialect or MSA) one or several synonymous words if they exist. Then all the possible sentences obtained by replacing each word by its synonym are produced. the number of sentences generated depends on the number of words having synonyms and the number of synonyms for each word. Once the sentences are obtained, they are added to the appropriate corpus. A similar operation is then launched but instead replacing each word by its synonyms we replaced each feminine pronoun by a masculine pronoun, a singular by plural and so on.

7. 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:

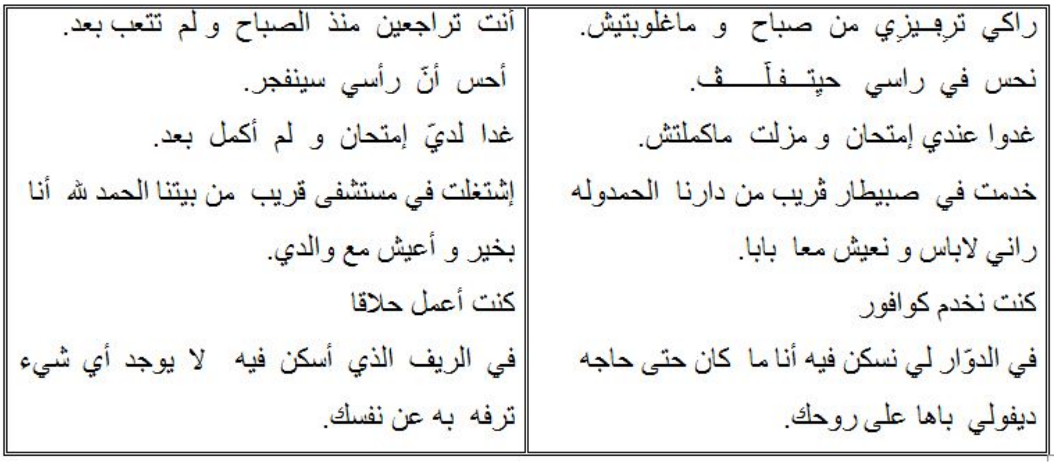
- Arabized loaned words: are words belonging to foreign vocabulary (most of them are French words), which were introduced in the dialect after having been naturalized phonetically and/or morphologically. Examples of such words are given in table 6.

Table 6. Examples of Arabized Foreign words

English	Annaba Dialect	Origin
Nurse	فرملي *farmlī*	French ”Infirmier”
Place	بلاصه *blāṣah*	French ”Place”
That's enough	يزّي *yizzī*	Berber
Ship	ببور *babūr*	Turkish

- Words for which we do not find any origins like صوارد *ṣwārad (Money), ...*

- 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 [12] between the original words and the dialectal ones. The results showed that the deformations performed on Arabic words concern principally the pronunciation. In fact, in general all consonants of the original word are kept in the dialect but the short vowels are modified. In such cases, the Levenshtein distance is zero because actually the vowels are not written. For other, original Arabic words are altered by inserting, omitting or by substituting consonants. In general the results achieved showed that the Levenshtein distance is less or equal to 3. For distances greater than three, words can be totally different, with the exception of one or two letters of Arabic words that are saved in some dialectal words. The dialect word can also be a concatenation of several words in standard Arabic, as نشاللّه *nšāl-lah* which correspond to إن شاء اللّه *ʾin šāʾa 'l-lah*. Appendix A provides several examples of



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

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

dialect words, their equivalents in standard Arabic and their corresponding Levenshtein distance.

We summarize in figure 2 the results of the analysis performed on the vocabulary constituting the corpus.

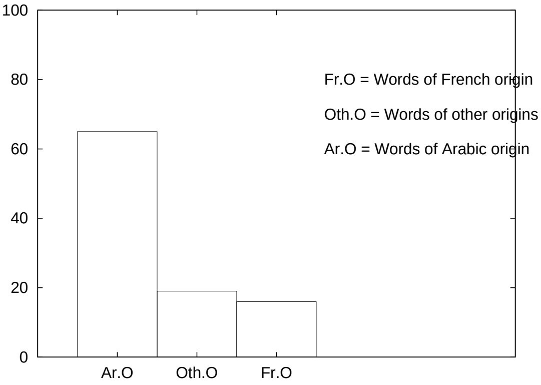


Fig. 2. Results of corpus analysis in terms of percentage.

We can therefore confirm that the vocabulary of the dialect of Annaba is a mixture of words from Standard Arabic, French and the other words borrowed from other languages such as Berber and Turkish or of unknown origin.

8. CONCLUSION

In this paper, we presented the main features of the dialect of Annaba through a linguistic study. In our knowledge no other corresponding work have been done before.
As we have already specified above, this work is part of a project *TORJMAN* which is dedicated to translating standard Arabic to Algerian Arabic dialects. To build a statistical machine translation system, a parallel training data is necessary. In the case of Annaba's dialect, there is no corpus that can be used. So, to achieve that, we recorded real discussions between people that we transcribed. We subsequently developed AD-MSA dictionary that we used to translate the dialect corpus in standard Arabic. We also presented a study of the dialect's vocabulary which has shown that it is mainly inspired from standard Arabic (65%). The remaining words are either of French origin (19%) or of Turkish or Berber (16%).The future work, before developing a statistical machine translation system, consists in increasing the available corpus. This will be done by adapting to Annaba's dialect the subtitles of some available Algerian movies. In fact, these ones are mostly expressed with the dialect of Algiers (Capital city of Algeria).

9. 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] Dina Al-Kassas, *Une étude contrastive de l'Arabe et du Français dans une perspective de génération multilingue*, Ph.D. thesis, Université PARIS 7, DENIS DIDEROT, UFR Linguistique, 2005.

[3] Kirchhoff K., Bilmes J., Das S., Duta N., Egan M., Ji G., He F., Henderson J., Liu D., Noamany M., Schone P., Schwartz R., and Vergyri D., "Novel approaches to arabic speech recognition: Report from the 2002 johns-hopkins workshop," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Hong Kong, 2003.

[4] Vergyri D. and Kirchhoff K., "Automatic diacritization of arabic for acoustic modeling in speech recognition," in *Proceedings of the COLING Workshop on*

Arabic-script Based Languages, Geneva, Switzerland, 2004.

[5] Belgacem M., "Construction dun corpus robuste de différents dialectes arabes," in *proc. of Actes des VIIIémes RJC Parole*, Avignon, France, 2009.

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

[7] 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.

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

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

[10] Debili F., Achour H., and Souissi E., "De l'étiquetage grammatical à la voyellation automatique de l'arabe.," in *Correspondances*, 2002.

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

[12] Michael Gilleland, "Levenshtein distance, in three flavors," http://www.merriampark.com/ld.htm, 2012.

A. LEVENSHTEIN DISTANCE FOR DIALECT WORDS AND THEIR EQUIVALENTS IN MSA

MSA	Annaba Dialect	Lev. dist.
تعرف *taʕrif*	تعرف *taʕraf*	0
تمرض *tamraḍu*	تمرض *tumriḍ*	
خرجت *ḫaraǧtu*	خرجت *ḫriǧt*	
يلبس *yalbasu*	يلبس *yilbas*	
يلعب *yalʕabu*	يلعب *yilʕab*	
يسهل *yusahilu*	يسهل *ysahil*	
.قصير *.qaṣīr*	.قصير *.qṣīr*	
جارتنا *ǧāratunā*	جارتنا *ǧāritnā*	
البحر *āl-baḥr*	لبحر *lbḥar*	1
يحاسبونه *yuḥāsibūnahu*	يحاسبوه *yḥāsbūh*	
يطمئن *yaṭmaʔin*	يطمن *yṭamin*	
يقرأ *ya.qraʔu*	يقرا *ya.qrā*	
يقعد *ya.qʕud*	يقعد *yuvʕud*	
وراءه *warāʔahu*	وراه *wrāh*	
ملايين *malāyīn*	ملاين *mlāyin*	
الأيام *āl-ʔayām*	ليام *liyām*	2
أشتريه *ʔaštarīh*	نشريه *nišrīh*	
أذني *ʔuḏunī*	ودني *wadnī*	
كيف *kayfa*	كفاش *kifāš*	
الأحوال *al-ʔaḥwāl*	لحوال *laḥwāl*	
معي *maʕī*	معايا *mʕāyā*	
يعينه *yuʕīnuh*	يعاونو *yʕāwnū*	3
الخضار *al-ḫuḍār*	لخضرة *lḫuḍrah*	
لتر *litr*	إترا *ʔitrā*	
أكل *ʔakl*	ماكله *māklah*	
الشاي *al-šāy*	إتاي *ʔitāy*	
لاشيء *lāšayʾ*	حتشي *ḥatašay*	
العا.قبه *al-ʕā.qibah*	لعقوبه *laʕvūbah*	
الجزائر *al-ǧazāʔir*	دزاير *dzāyir*	> 3
أورا.ق *ʔawrā.q*	ور.قات *war.qāt*	
يدعك *yadaʕka*	يخليك *yḫalīk*	
لم تجف *lam taǧif*	ماشاحتش *māšāḥitš*	
إتصلي بها *ʔitaṣilī bihā*	عيطيلها *ʕayṭīlhā*	