Building Resources for Algerian Arabic Dialects

Salima Harrat, Karima Meftouh, Mourad Abbas, Kamel Smaïli

Building Resources for Algerian Arabic Dialects

S. Harrat[1], K. Meftouh[2], M. Abbas[3], K. Smaili[4]

[1]ENS Bouzareah, Algiers, Algeria
[2]Badji Mokhtar University-Annaba, Algeria
[3]CRSTDLA, Algiers, Algeria
[4]Campus scientifique LORIA , Nancy, France

slmhrrt@gmail.com, Karima.meftouh@univ-annaba.org, m_abbas04@yahoo.fr, smaili@loria.fr

Abstract

The Algerian Arabic dialects are under-resourced languages, which lack both corpora and Natural Language Processing (NLP) tools, although they are increasingly used in written form, especially on social media and forums. We aim through this paper, and for the first time, to build parallel corpora for Algerian dialects, because our ultimate purpose is to achieve a Machine Translation (MT) for Modern Standard Arabic (MSA) and Algerian dialects (AD), in both directions. We also propose language tools to process these dialects. First, we developed a morphological analysis model of dialects by adapting BAMA, a well-known MSA analyzer. Then we propose a diacritization system, based on a MT process which allows to restore the vowels to dialects corpora. And finally, we propose results on machine translation between MSA and Algerian dialects.

Index Terms: Algerian dialect, Modern Standard Arabic, Machine translation system, Morphological analyzer.

1. Introduction

Most of Arab people do not use MSA in their daily conversations; the result is that different Arabic dialects are spoken through more than twenty countries. In this paper, we are interested in Arabic dialects processing, their particularity is they are not written, therefore it is difficult or impossible to find available corpora for those vernacular languages. It is well known, that these corpora are necessary for statistical models based application as speech recognition and machine translation. The issue here concerns two AD, one from Algiers (capital of Algeria) and the other from Annaba (at the border of Tunisia). To the best of our knowledge, no work has been done on these dialects. Our main objective is to study several dialects and in the near future to try to adapt applications as machine translation from one dialect to another. The NLP community started to make effort for some dialects and more particularly for Egyptian [1], Levantine [2] and Iraqi [3] dialects. But, these dialects could be considered as close to MSA, whereas those of Algeria are influenced by French, Turkish and Berber which make them difficult to be processed with classical tools developed for MSA. In fact, AD use a lot of words borrowed from French, some of them are pronounced in French language and others have been altered phonologically to fit the Algerian dialect morphology.
Consequently and due to all the previous reasons, AD have no elementary resources such as monolingual or multilingual corpora, electronic dictionaries, and thesaurus. Obviously, sophisticated resources have not been either developed such as morphological analyzer, parsers, etc. In the following, and for the first time for AD, we will present preliminary corpora and tools for processing them. We will present the corpora we collected and how we enriched them. Then we develop a tool for diacritizing dialect corpora. We will present also a morphological analyzer for AD inspired from the Buckwalter one dedicated to MSA. Finally, we will present few preliminary experiments on machine translation between MSA and AD.

2. Building Algerian Dialect Corpora

2.1. Handcrafted

In Arab countries, only educated people understand MSA perfectly, for the others it is considered as a foreign language because it is not their mother language. That is why, a machine translation which translates from MSA to AD should be appreciated. In the following, we will focus on developing the needed resources for two dialects: one from Algiers (ALG) and the other from Annaba (ANB). For ANB, we recorded different conversations, while for ALG, we used Algerian movies and TV shows which most of them use Algiers's dialect. We transcribed by hand the recordings of the two dialects by adopting some writing rules [4]. Then, we extracted the words records for both dialects ALG and ANB. Finally, we built two dictionaries MSA-ALG and MSA-ANB by assigning to each extracted word, its best corresponding translation in MSA. To have more data, we translated each dialect corpus into the other, this allowed to increase both of them as presented in table 1.

Size	ANB		ALG	
	Before	**After**	**Before**	**After**
Corpus (sentences)	4000	6415	2415	6415
Vocabulary (words)	6754	9688	4545	10790

Table 1: *ANB and ALG sizes before and after enrichment.*

A deep study of the dialect corpora shows that a significant amount of dialect words have an Arabic origin. According to their origin, AD words can be split into three categories: Arabic words, Arabised borrowed words, and Unknown origin words. In order to measure the closeness between ANB/ALG and MSA, we decided to use Levenshtein distance [5]. However, some words of Arab origin have undergone some distortions. To determine the words which have been altered, we calculated the Levenshtein distance between the original words and the dialect ones (Figure 1). These deformations on Arabic words are due to a process of pronunciation simplification.

Figure 1 shows that almost 20% of the dialect words are original from Arabic and 34% of them are inspired from Arabic but with one letter which differs between the dialect word and

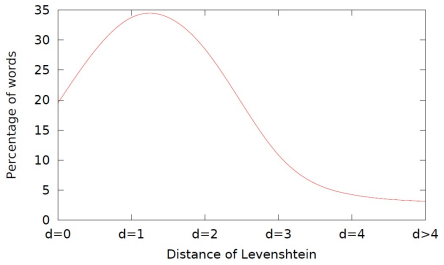

Figure 1: *Percentage of words for different values of Levenshtein distances.*

Dial. word	Origin	Leven.	Meaning
رجال *rğāl*	AR: رجال *rğāl*	0	Men
السكات *alskāt*	AR: السكوت *alskūt*	1	Silence
جودي *ğūdī*	FR: Jeudi	2	Thursday
كريدي *krīdī*	FR: Crédit	3	Credit

Table 2: *Examples of dialect words from Arabic (AR) and French (FR) origin.*

the Arabic one. We noticed also that in some cases, dialect words with a Levenshtein distance of 2 or 3 are not from Arab origin and most of them are borrowed from French (Table2).

2.2. Automatically

A significant amount of dialect words come from MSA and because Arabic corpus are available, we then propose to extract new dialect corpus by rewriting each word in Arabic sentence to its closest one on dialect. This is obviously not enough to collect a relevant Arabic dialect corpus, that is why we decided to associate to each collected sentence a measure which evaluates the human being effort necessary to correct the dialectal sentence obtained automatically.
Let A_S be the MSA sentence made-up of i words. To build the corresponding dialect sentence noted A_D, we select from the dialect vocabulary and for each word A_S^i of A_S, the word A_D^i which is the closest in terms of Levenshtein distance. Then A_D is evaluated in terms of human being effort necessary for correcting it, in accordance to formula 1.

$$d(A_S, A_D) = \frac{\frac{1}{|A_S|}\sum_{i=1}^{|A_S|}\frac{d_L(A_S^i, A_D^i)}{|A_S^i|}}{Max(\varepsilon, \sum_{i=1}^{|A_S|}\delta(A_S^i, A_D^i))} \quad (1)$$

with:

$$\delta(x,y) = \begin{cases} 1 & if\ x = y \\ 0 & otherwise \end{cases} \quad (2)$$

Where $d_L(x,y)$ is the Levenshtein distance between the words x and y. $|A_S|$ is number of words of the sentence A_S and $|A_S^i|$ represents the length of its i^{th} word. ε is to avoid a zero value. The distance d promotes sentences with many common words and penalizes words with high Levenshtein distances. d reaches its optimum for 0 which means that each word in dialect is written exactly as in MSA. But, even if the words have been selected on the basis of minimum Levenshtein distances, the sentence needs a human correction. In order to reduce at minimum the human being intervention, only the sentences which necessitate a minimum correction are kept in accordance to the following formula:

$$COR = \frac{1}{|A_D|}\sum_{j=1}^{|A_D|}\frac{W_{Eff}(A_D^j)}{|A_D^j|} \quad (3)$$

Where $W_{Eff}(A_D^j)$ is the effort expressed in terms of characters necessary to type in order to correct (or replace) the word A_D^j. The effort is up to 50% compared to the effort made when we translated by hand (100% in this case) (see figure 2). This means that we have to type at most half of characters making up the words of the dialectal sentence. This does not allow us to determine the threshold to be considered, especially if we consider the number of selected sentences for each distance; this number varying from 2 to 100 for threshold values of 0 to 0.7 respectively. So, to determine the threshold to use for selecting dialect sentences, we opted for another solution. We calculated the character error rate (CER) for different distances using Sclite. [1] We found that the lowest error rate is achieved for a distance of 0.03 which corresponds to a human being correction of 37% and allowing to enrich the initial dialect corpus by 28% pairs of sentences.

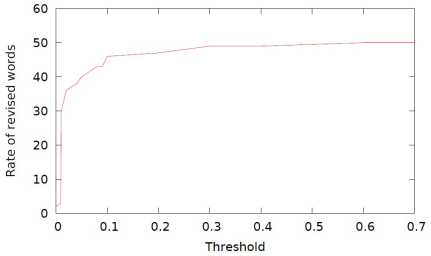

Figure 2: *The average correction effort (in %) based on the threshold.*

3. Diacritization of Algerian Dialect

Another issue concerning the production of dialects resources is to provide diacritized texts. In fact, absence of diacritics in Arabic texts produces a serious problem for many NLP applications. The AD are concerned by this problem. Several works concerning this issue have been proposed by the community using different methods as Conditional Random Fields [6], HMM [7], weighted finite state machines combined with language models [8], sequence classification using maximum entropy [9], but only for MSA, and to the best of our knowledge nothing has been done for Arabic dialects and especially for Algerian ones. In [10] we considered the diacritization as a translation issue. For that we used a training parallel corpus composed of an unvocalized source text and a vocalized target one. To train the translation model, we used classical statistical tools: Moses [11], GIZA [12], SRILM [13]. First, we experimented our solution on MSA diacritized corpora Tashkeela [2] and LDC Arabic Treebank (Part3,V1.0) [14] and then we tested it on our dialect corpora. For the dialect, we vocalized by hand a part of Algiers dialect which has been used for training. The achieved results have been corrected and re-injected in the training corpus. The whole corpus has then been diacritized in an

[1] `http://www1.icsi.berkeley.edu/Speech/docs/sctk-1.2/sclite.htm`

[2] A collection of classical Arabic books from an on-line library available on `http://sourceforge.net/projects/tashkeela`

iterative way. The main results for MSA and dialects are presented in table 3 in terms of DER (Diacritization Error Rate) and precision.

Corpus	DER (%)	Precision (%)
Tashkeela	4.1	93.1
LDC ATB	5.7	96
Algiers Dialect	12.8	98

Table 3: *Diacritization System Results(DER and Precision)*

This table shows, in spite of the fact that dialect is less diacritized than MSA, that its DER is higher than the one of MSA. This is mainly due to the small size of ALG corpus. We think that for a larger AD corpus we should achieve a smaller DER than the one get for MSA.

4. Morphological Analysis for Algerian Dialect

Compared to MSA, there are very little Morphological Analysers (MA) dedicated to Arabic dialects. Works in this area could be divided into two categories. The first one includes MA that are built from scratch as in [15] and [16], the second includes works that attempt to adapt existing MSA Morphological Analysers to Arabic dialect. This trend is adopted for several dialects since it does not consume time and effort. In [17], authors used BAMA Buckwalter Arabic Morphological Analyser [18] by extending its affixes table with Levantine/Egyptian dialectal affixes. The same approach is adopted in [19] where a list of dialectal affixes (belonging to four Arabic dialects) was added to Al-Khalil [20] affix list. Authors in [21] converted the ECAL (Egyptian Colloquial Arabic Lexicon) to SAMA (Standard Modern Arabic Analyser) representation [22]. For Tunisian dialect, authors in [23] adapted Al-Khalil MA, they create a lexicon by converting MSA patterns to Tunisian dialect patterns and then extracting specific roots and patterns from a training corpus that they created. To build a MA for Algiers's dialect, we decide to adapt BAMA, since it does not consume time and takes profit from the fact that it is widely used. BAMA is based on a dictionary of three tables containing Arabic stems, suffixes and prefixes and three compatibility tables defining relations between stems, prefixes and affixes.

4.1. Building the dialect dictionary

We built dialect dictionary by adopting the following principle: keeping from MSA affixes and stems tables all entries that belong also to ALG and deleting those which do not. For affixes tables, common affixes between MSA and ALG are kept (in prefixes and suffixes tables), whereas all other MSA affixes which do not belong to dialect were deleted. However, some dialect affixes which do not exist in MSA were added to affixes tables. Note that when an affix is deleted, all affixes where it occurs are also deleted, see table 4 for some examples.

Dialect stems table was populated by the lexicon of Algiers dialect corpus and MSA stems included in BAMA. We used a part (85%, 9170 distinct words) of our ALG corpus that we diacritized as explained in section 3 for creating dialect stems, the remaining 15% (1618 distinct words) is used for test. First, we began by extracting a list of nouns easily identifiable by affixes ة *h* and definite article الـ *āl* (used only with nouns). We deleted these two affixes from all extracted words, then from obtained

Kept Aff.	Description
ت *t*, ي *y*	Imperfect Verb Prefix(sing.,third person,masc.,fem.)
ات *āt*	Noun Suffix(fem.,plu.)
Del. Aff.	**Description**
ف *f*	Conjunction Prefix
فبال *fbāl*	Conj. Pre.+Preposition Pre.+Definite Art. Pre.
هن *hn*	Perfect/Imperfect Verb Suffix(direct object, plu., fem.)
Add. Aff.	**Description**
ش *š*	Perfect/Imperfect Verb Negation Suffix
و *w*	Per./Imp. Verb Suffix(direct object,plu.,masc.,fem.)

Table 4: *Examples of kept, deleted and added affixes in ALG affixes tables*

list of words we created stem entries according to BAMA. Next, the rest of the corpus was analysed and classified into three sets: function words, verbs and nouns (which do not include ة *h* and الـ *āl* suffixes) and converted to stems according to BAMA stems categories. Let us indicate that we added some stems categories to take into account all dialectal features. For example, in MSA the perfect verb stem category with the pattern فَعَل *faʿal* covers the three persons, the two genders, the single, the dual and plural; just relative suffixes are added to it to have its different inflected forms. In ALG, we split this stem category into two classes:فعَل *fʿal* and فَعل *faʿl* to cover all perfect verbs inflected forms, in table 5 some examples are given .

Lang.	Pron.	Verb	Stem	Meaning
MSA	هي *hy*	سَمِعَت *samiʿat*	سَمِع *samiʿ*	She heard
	هو *hw*	سَمِعَ *samiʿa*		he heard
ALG	هي *hy*	سْمْعَتْ *samʿat*	سَمع *samʿ*	She heard
	هو *hw*	سَمَع *smaʿ*	سَمَع *smaʿ*	he heard

Table 5: *Examples of splitting a MSA stem to two Dialectal stems*

As mentioned above another part of stems tables was created from MSA stems. We first process verbs, the main idea for creating ALG verb stems from MSA stems is using verbs pattern. For example ALG verbs with the pattern فعَل *fʿal* are in most cases Arabic verbs with the patterns فَعَل *faʿal*, فَعُل *faʿul* or فَعِل *faʿil*. Some other ALG verbs keep the same pattern as in MSA like verbs with the pattern فَعَّل *faʿʿal*. From MSA stems table, we extracted all perfect verbs stems having the patterns فَعَل *faʿal*, فَعُل *faʿul*, فَعِل *faʿil* and فَعَّل *faʿʿal*. After that, the verbs having the three first patterns are converted to Algiers dialect pattern فعَل *fʿal* by changing diacritic marks, while verbs corresponding to pattern فَعَّل *faʿʿal* are kept as they are (since this pattern is used in Algiers dialect). At this stage, we constructed a set of Arabic verb stems having dialect pattern, we analysed them and eliminated all stems that are not used in ALG. We give in table 6 some examples. Note that, we constructed imperfect verb stems and command verb stems from the ALG perfect verb stems that we created. We proceed as de-

Stems	ALG	MSA	Meaning
ضرب *ḍrb*	ضرَب *ḍrab*	ضَرَب *ḍarab*	He beat
كبر *kbr*	كبَر *kbar*	كَبُر *kabur*	He grew

Table 6: *Examples of converted stems from MSA to ALG*

scribed above for other patterns as تَفَعَّل *tafaʿʿal*, تَفَاعَل *tafāʿal*, فَاعَل *fāʿal*,استَفعَل *āstafʿal*. For nouns, we kept all proper nouns from MSA stems table since it contains an important number of entries related to countries, currencies,... We analysed remaining other types of words and kept from them those existing in ALG by modifying diacritics, adding or deleting one or more letters. We also deleted all function words that do not exist in ALG like relative pronouns and personal pronouns related to the dual and feminine plural, then we translated remaining ones to ALG. Note that we introduced dialect stems with non Arabic letters ڤ *G*, ف *V*, and پ *P* in stems table and we modified BAMA code to consider words containing these letters. Note that every stem entry in BAMA contains an English glossary, when creating a dialect entry, we added the Arabic word to English glossary, so for each dialect entry is associated an English and Arabic glossary. After creating affixes and stems tables for ALG, compatibility tables of BAMA were updated according to the data included in these tables.

4.2. Experiment

As mentioned above, we tested our MA on the Algiers Dialect corpus, the test set contains 1618 distinct words extracted from 600 sentences chosen randomly. We consider that a word is correctly analysed if it is correctly decomposed to prefix+stem+suffix and if all the features related to them are correct (POS, gender, number, person). We first began by testing the MA with stems extracted only from the ALG corpus lexicon, then we introduced stems created from the MSA stems table. We list in table 7 the obtained results.

Results	ALG corpus stems	MSA stems+ALG corpus stems
# Analysed words	703	1115
Percentage	43.3%	68.98%
# Unanalysed words	915	503
Percentage	56.6%	31.08%

Table 7: *Morphological Analysis on Algerian Dialect*

Unanalysed words mainly are French words which do not exist in the stem table like الجون (le jeune, the young man), انجنيور (ingénieur, engineer). Another source of unanalysed words are those written with an orthography for which no stem does exist like for example nouns written with long vowel ا in the end instead of ة such as كلاسا (classroom). We noticed also that some words are written with missed letters as قالي (he said to me) instead of قالّي or قال لي or قتلو (I said to him) instead of قلتلو or قلت لو. Some Unanalysed words also are proper nouns.

5. Machine Translation for Algerian Dialects

Before developing a speech to speech machine translation which should be extended to translate to French and English, we present for the first time in machine translation community results concerning translation from AD to MSA. All the MT systems we used are phrase-based with default settings: bidirectional phrase and lexical translation probabilities, distortion model, a word and a phrase penalty and a language model. We used GIZA++ to align sentences and the SRILM toolkit to compute tri-gram MSA language model. Because our corpus is not large enough, we did several experiments using the same methods for training, decoding and parameter tuning. We only varied the corpora used for computing the language model and the smoothing technique (Kneser-Ney and Witten-Bell) by hoping that it can achieve better performance. All experiments are performed on training corpora containing more than 6400 sentences, and tests are achieved on a corpus of 300 sentences. To calculate the language model, we first used the MSA side of the parallel corpus we built (MSA-6400). Thereafter, we tested another language model calculated on a larger corpus Tashkeela. We report in table 8 the results in terms of BLEU score. The first conclusion is that clearly the use of a smoothing technique should not be fortuitous since the performance varies in accordance to the method and the corpus used for computing the language model. Furthermore, it seems to be more difficult to translate ALG than ANB. This could be explained by the fact that more we move for east more the dialect is closer to MSA and by the fact that in ALG people use more French words.

LM corpus	MSA-6400		Tashkeela	
Smooth. / Dial.Corpus	KN	WB	KN	WB
ANB	14.22	14.57	14.26	13.88
ALG	10.04	8.53	8.70	8.71

Table 8: *MT evaluation according to LM & smoothing tech.*

6. Conclusion

In this paper we presented a work on Algerian dialects. We started by collecting a corpus of 4K sentences which has been increased automatically by 25%. To do that, we proposed a method which enrich the corpus by transforming the MSA text on dialect by reducing at maximum the errors due to the automatic process. Then we proposed a method based on machine translation to diacritize the achieved corpus, results are very satisfactory (a DER of 12.8%, a Precision of 98%). We used this corpus to build a part of the MA dictionary. The other part was built from MSA dictionary which we translated to ALG. The deal of this analyser is to recognize correctly and to cover a large portion of ALG words. A first experiment in MT for AD has been proposed. In spite of the weak score of BLEU, we showed the feasibility of machine translation on a vernacular language for which no resource was available. A positive result has been presented, when the corpus is small the choice of the smoothing method of the language model should not be chosen by default. In this work, some tasks have been handcrafted, so it was time consuming. We believe that the result is proportional to this effort since we got a substantial resource for two Algerian dialects constituted of corpus cleaned and checked and language tools to process them.

7. References

[1] H. Gadalla, H. Kilany, H. Arram, A. Yacoub, A. El-Habashi A. Shalaby, K. Karins, E. Rowson, R. MacIntyre, P. Kingsbury, D. Graff and C. McLemore, "CALLHOME Egyptian Arabic Transcripts", Linguistic Data Consortium LDC, Philadelphia, USA, 1997.

[2] M. Maamouri and T. Buckwalter and D. Graff and H. Jin, "Fisher Levantine Arabic Conversational Telephone Speech", Linguistic Data Consortium LDC, Philadelphia, USA, 2007.

[3] Appen Pty Ltd (Iraqi), Sydney, "Iraqi Arabic Conversational Telephone Speech, Transcripts", Linguistic Data Consortium LDC, Philadelphia, USA, 2006.

[4] K. Meftouh, N. Bouchemal and K. Smali, "A study of a non-resourced language: an Algerian dialect", Proceedings of the third international workshop on spoken languages technologies for under-resourced languages SLTU, Cape Town, South Africa, 2012.

[5] M. Gilleland, "Levenshtein distance, in three Flavors", Merriam Park Software, http://www.merriampark.com/ld.htm.

[6] T. Schlippe, T. Nguyen and S. Vogel, "Diacritization as a Machine Translating Problem and as a Sequence Labeling Problem", Proceedings of the Eighth Conference of the Association for Machine Translation in the Americas (AMTA), Hawai, USA, 2008.

[7] Y. Gal, "An HMM approach to vowel restoration in Arabic and Hebrew", Proceedings of the ACL-02 workshop on Computational approaches to Semitic languages, PA, USA, 2002.

[8] R. Nelken and M. Sheiber Stuart, "Arabic Diacritization Using Weighted Finite-State Transducers", Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages, Ann Arbor, Michigan, USA, 2005.

[9] I. Zitouni, J. Sorensen and S. Ruhi, "Maximum entropy based restoration of arabic diacritics, Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, Sydney, Australia, 2006.

[10] S. Harrat, M. Abbas, K. Meftouh and K. Smaili, "Diacritics Restoration for Arabic Dialect Texts", Proceedings of 14th Interspeech, Lyon, France, 2013.

[11] P. Koehn, H. Hoang,A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, E. Herbst, "Moses: Open Source Toolkit for Statistical Machine Translation", Annual Meeting of the Association for Computational Linguistics ACL, demonstration session, Prague, Czech Republic, 2007.

[12] F. Josef Och, H. Ney, A Systematic Comparison of Various Statistical Alignment Models, Computational Linguistics, volume 29, number 1, pp. 19-51, 2003.

[13] A. Stolcke, SRILM An Extensible Language Modeling Toolkit on Spoken Language Processing, Proceedings of the International Conference, volume 2, pp. 901-904, Denver, 2002.

[14] M. Maamouri, A. Bies, T. Buckwalter and H. Jin, "Arabic Treebank: Part 3 v 1.0", Linguistic Data Consortium LDC, Philadelphia, USA, 2004.

[15] N. Habash and O. Rambow,"MAGEAD: a morphological analyzer and generator for the Arabic dialects", Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics ACL44, pp. 681-688, Stroudsburg, PA, USA, 2006.

[16] M. Altantawy, N. Habash, and O. Rambow, "Fast Yet Rich Morphological Analysis", Proceedings of the 9th International Workshop on Finite-State Methods and Natural Language Processing FSMNLP, Blois, France, 2011.

[17] W. Salloum and N. Habash, "Dialectal to Standard Arabic Paraphrasing to Improve Arabic-English Statistical Machine Translation", Proceedings of EMNLP 2011, Conference on Empirical Methods in Natural Language Processing, pp. 10-21, Edinburgh, Scotland, UK, 2011.

[18] T. Buckwalter, "Buckwalter Arabic Morphological Analyzer Version 1.0", Linguistic Data Consortium LDC, University of Pennsylvania, 2002.

[19] K. Almeman and M. Lee, "Towards Developing a Multi-Dialect Morphological Analyser for Arabic", 4th International Conference on Arabic Language Processing, Rabat, Morocco, 2012.

[20] A. Boudlal, A. Lakhouaja, M. Azzeddine, and M. Abdelouafi,"Alkhalil Morpho Sys: A Morphosyntactic analysis System for Arabic texts", Proceedings of ACIT2010, Riyadh, Saudi Arabia, 2011.

[21] N. Habash, R. Eskander and A. Hawwari, "Morphological Analyzer for Egyptian Arabic", Proceedings of the Twelfth Meeting of the Special Interest Group on Computational Morphology and Phonology SIGMORPHON ,pp. 1-9, Montral, Canada, 2012.

[22] D. Graff, M. Maamouri, B. Bouziri, S. Krouna, S. Kulick, and T. Buckwalter, "Standard Arabic Morphological Analyzer (SAMA) Version 3.1", Linguistic Data Consortium LDC, Philadelphia, 2009.

[23] I. Zribi, M. Ellouze Khemakhem, L. Hadrich Belguith, "Morphological Analysis of Tunisian Dialect", International Joint Conference on Natural Language Processing, pp. 992-996, Nagoya, Japan, 2013.