Maghrebi Arabic dialect processing: an overview

Salima Harrat, Karima Meftouh, Kamel Smaïli

Maghrebi Arabic dialect processing: an overview

SALIMA HARRAT[1], KARIMA MEFTOUH[2], AND KAMEL SMAÏLI[3]

[1] *Ecole Supérieure d'Informatique (ESI), Ecole Normale Supérieure de Bouzareah (ENSB), Algiers, Algeria*
[2] *Badji mokhtar University - Annaba, Algeria*
[3] *Loria University of Lorrraine - France*
[1] *slmhrrt@gmail.com*
[2] *karima.meftouh@univ-annaba.org*
[3] *smaili@loria.fr*

Compiled February 23, 2018

Natural Language Processing for Arabic dialects has grown widely these last years. Indeed, several works were proposed dealing with all aspects of Natural Language Processing. However, some AD varieties have received more attention and have a growing collection of resources. Others varieties, such as Maghrebi, still lag behind in that respect. Maghrebi Arabic is the family of Arabic dialects spoken in the Maghreb region (principally Algeria, Tunisia and Morocco). In this work we are interested in these three languages. This paper presents a review of natural language processing for Maghrebi Arabic dialects.

1. INTRODUCTION

The Arabic language is characterized by its plurality. It consists of a wide variety of languages, which includes the Modern Standard Arabic (MSA), and a set of various dialects differing according to regions and countries. The varieties of Arabic dialects (AD) are distributed over the 22 countries in the Arab World. Geographically, Arabic dialects are classified in two main blocs, namely Middle East (Mashriq) and North Africa (Maghreb) dialects. Maghrebi dialects are the languages that are spoken in this geographical area (Maghreb). They are characterized by the coexistence of several languages: MSA, dialectal Arabic, Berber and French. The Berber dialects constitute the oldest linguistic substratum of this region and are, therefore, the mother tongue of a part of the population. Since the Islamic conquest of the Maghreb, several Arab tribes have intermingled, especially in pastoral areas, because of the similarity of their way of life. This coexistence reinforced the Arabization of the Berber tribes. The influence of the Arabic language on the Berber world spread fairly rapidly, and this practically all over the Maghreb [1]. The French language was introduced by the colonial occupation. First, as the language of the colonial administration, this language has spread to a large part of the population through education and administration. This language spread in its written and oral uses, it influenced the spoken languages (Berber and AD) by the borrowings that these made to it [2].
The Maghreb is composed, in its central part, of Algeria, Tunisia and Morocco. In this paper, we are interested in the Arabic spoken in these three countries. This interest is justified by the fact that these countries have in common a lot of socio-historical similarities and an identical linguistic situation. We therefore present in this work an overview of these dialects, first on several levels of linguistic representation (section 2) and then in terms of research work dealing with these languages (section 3). We believe that such a study is very useful for the scientific community working in the field of Natural Language Processing (NLP) in general and more specifically those working on NLP of Maghrebi Arabic dialects.

2. LINGUISTIC OVERVIEW

Maghrebi Arabic dialects include principally Algerian Arabic, Moroccan Arabic and Tunisian Arabic. In this section, we give an overview of these three languages regarding phonological, lexical, morphological and syntactic level.

A. At phonological level

The three Maghrebi Arabic dialects share the most features of standard Arabic. Besides the 28 Arabic consonants phonemes, the three dialects of the Maghreb use non Arabic phonemes /g/, /p/ and /v/ which are mainly used in words borrowed from foreign languages as French. Also, the (ظ) is uttered as /dʿ/(ض), whereas (ذ) and (ث) are mostly pronounced as /d/(د) and /t/(ت) for both Algerian[1] and Moroccan dialects and not for Tunisian where the utterance of these two consonants is the same as in MSA. Furthermore, the letter (ق) is particular in the way that it has different pronunciations. For the three dialects it is uttered as /q/ and /g/. It should be noted that the use of /g/ is observed not only in rural places but also in urban cities. In addition, the (ق) is uttered as the glottal stop /?/ as in Tlemcen (west of Algeria) and Fes (Morocco), just like in Egyptian dialect. In some eastern cities of Algeria a particular pronunciation of the (ق) is /k/ (this phenomenon does not exist in Tunisian and Moroccan). Also, The consonant (ج) has different pronunciations /dj/, /j/ or /z/ (for Tunisian dialect and the dialect of Tlemcen and other cities in the east of Algeria). Other notable features of Maghrebi dialects are the collapse of short vowels both in nouns and verbs and the glottal stop (Hamza) omission particularly in the middle and the end of words.

B. At Lexical level

Maghrebi dialects' vocabulary is mostly inspired from Arabic but it is phonologically altered, with significant Berber substrates, and many loanwords from French, Italian, Turkish and Spanish. Like for Arabic vocabulary, these dialects' vocabularies include verbs, nouns, pronouns and particles.

C. At Morphological level

The morphology of Arabic dialectal words shares a lot of features with MSA morphology. Furthermore, dialect inflection system is simpler in some aspects than MSA, whereas affixation system seems to be more complicated than MSA. Indeed inflection system is simplified by the elimination of a wide range of rules. In fact, as in all Arabic dialects, Algerian, Moroccan and Tunisian do not accept the singular word declension which corresponds to the nominative, the genitive, and the accusative cases which take the short vowels ـُ, ـِ and ـَ respectively in the end of the word. Similarly, the three doubled case endings expressing nominal indefiniteness are also dropped. It should be noted that for the three dialects, the singular nouns declension to the plural (feminine/masculine regular plural and broken plural) follows MSA rules but with the difference that the three cases enumerated above are not distinguished.[2] In addition, the three dialects do not bave the nominal dual which is a distinctive feature of standard Arabic. The verb conjugation of the three dialects uses a set of affixes slightly different with MSA ones besides a variation in vocalization. We mention that the dual and feminine plural of MSA are lost in the dialects. Moreover, the negation in the three dialects seems to be more complex than in MSA, the circumfix negation (مَا + ش) surrounds the verbs with all its affixed direct and indirect object pronouns.

D. At Syntactic level

The words order of a declarative sentence in the three dialects is relatively flexible but the most commonly used order is the SVO order (Subject-Verb-Object)[3],[4],[5]. The Other orders are also allowed, the speaker generally begins his sentence with the item that he wants to highlight.

3. NLP OF THE THREE DIALECTS

In this section, we are interested in the research work developed for these dialects in various NLP issues.

A. Corpora and lexicons

A dictionary containing 18K MSA and Moroccan dialects entries was built in [6]. The authors used manual translation from MSA dictionary to Moroccan dialect and vice-versa.

In [7] authors created an annotated corpus of 223K that they collected from Moroccan social media sources. The corpus has been annotated on token-level by three native speakers of Moroccan dialect.

In [8] a focus was made on Tunisian dialect processing. The authors extracted textual user-generated contents from social networks that they filtered and classified automatically. From the built corpora they drew a picture of the main features related to Tunisian dialect.

The authors in [9] presented a bilingual lexicon of deverbal nouns between MSA and Tunisian dialect that has been created automatically. They extended an existing Tunisian verbal lexicon by using a table of deverbal patterns in order to generate pairs of Tunisian and MSA deverbal nouns.

The work presented in [10] is related to the construction of a railway domain ontology from a Tunisian speech corpus created for this purpose within this study. The authors used a statistical method for term and concept extraction whereas for semantic relation extraction they choose a linguistic approach.

In [11], authors generated automatically phonetic dictionaries for Tunisian dialect by using a rule approach. The work is part of an automatic speech recognition framework of the Tunisian Arabic in the particular field of railway transport.

In [12] is presented STAC (Spoken Tunisian Arabic Corpus), 5 transcribed hours of spontaneous Tunisian Arabic speech enriched with morpho-syntactic and disfluencies annotations.

[1] In some rural dialects they are pronounced as in MSA.

[2] Example: Depending on its function in the sentence, the masculine regular plural of MSA word مُسلِم (Muslim) could be مسلمون (nominative case) or مسلمين (accusative or genitive). In contrast, for the dialect word مَسلَم (Muslim) always takes مسلمين for the regular plural whatever its grammatical category.

For Algerian dialect, in [13], the authors crawled an Algerian newspaper to extract comments that they used to build a romanized code-switched Algerian Arabic-French corpus. In this study, the authors highlighted the particular Algerian linguistic situation by discussing its main features. It should be noted that the corpus is annotated by language identification at word-level.

KALAM'DZ, An Arabic Spoken corpus dedicated to Algerian dialectal varieties was built in [14] by exploiting Web resources such as Youtube and other Social Media, Online Radio and TV. The dataset covers a large number of Algerian dialects with 4881 native speakers and more than 104 hours.

An other Speech corpus dedicated to Algerian dialect, AMCASC (Algerian Modern Colloquial Arabic Speech Corpus) was presented in [15]. Authors used this corpus for the purpose of evaluating their automatic regional accent recognition approaches based on GMM-UBM and i-vectors frameworks.

In the same vain, authors of [16] presented their methodology to build an Arabic Speech Corpus for Algerian dialects. The authors proceeded by recording speeches uttered by 109 native speakers from 17 different regions in Algeria.

In [17], CALYOU, a Comparable Corpus of the spoken Algerian was built from Youtube comments. It consists of 853K comments including a total of 12.7M words. This work deals with the issue of comparability of comments extracted from Youtube. It presents a Word2Vec based method of alignment which achieves the best comparability results among the other methods that the authors experimented.

B. Identification

Several efforts dealing with Maghrebi Arabic dialects are those dedicated to the identification and recognition. In fact, Arabic dialects differ from one country to another and even in the same Arab country there is a lot of dialect varieties. In this context, authors of [18] addressed the problem of spoken Algerian dialect identification by using prosodic speech information (intonation and rhythm). They performed an experiment of their approach on six dialects from different Algerian departments. An other study [19] showed that Algiers and Oran dialects can be identified by prosodic cues.

In [20], for the classification of Tunisian and Moroccan dialects, two methods were used namely the feed forward back propagation neural network (FFBPNN) and the support vector machine (SVM). The former (FFBPNN) performs better than the later in terms of recognition rates.

In the context of dialect identification within social media (Facebook comments), authors of [21] used an Algiers dialect lexicon and perform different ways of identification: total (word matching), partial (prefix and suffix matching) and by applying improved Levenshtein distance.

The work cited in [22] presents DATOOL a graphical tool for annotating tweets. A native speaker of Moroccan dialect annotated an average of 250 (mixed-language and mixed-script) tweets per hour. The obtained corpus has been used for the purpose of dialect identification.

C. Orthography

A particular attention is devoted to dialect orthography because of their spoken nature and thus a total absence of standard writing rules. Some efforts were made to resolve this issue. The authors of [23] presented orthography guidelines for transcribing Tunisian speech corpora based on the standard Arabic transcription conventions. Later, the CODA map (Conventional Orthography for Dialectal Arabic) described in [24] was adapted to Tunisian dialect [25], Algerian dialect [26] and finally in general for Maghrebi dialects [25].

D. Morphological analysis

In [27], a morphological analyzer for the Tunisian dialect based on a MSA analyzer was proposed. Furthermore, as an expansion of a MSA lexicon, a lexicon for the Tunisian dialect was built. This last lexicon has been used in [28] to convert a standard Arabic corpus for creating a large Tunisian dialect corpus, in order to train a POS tagger. A similar approach was adopted in [29] where the authors exploited also the closeness between standard Arabic and Tunisian dialect. They developed a POS tagger by converting a Tunisian sentence to MSA lattice, after a disambiguation step, a MSA target sentence is then produced and tagged simply with a MSA tagger.

For Moroccan dialect, in [30] a morphological analyzer has been developed in addition of an annotated corpus that has been created within this work. It should be noted that specific CODA guidelines for Moroccan dialect has been also created (inspired from [24] cited above).

For Algerian dialect, a morphological analyzer was developed in [31]. Authors adapted the well-known morphological analyzer BAMA dedicated for MSA.

E. Sentiment analysis

Sentiment analysis is a promising and challenging direction research in the area of dialect NLP. Indeed, Arab people use their dialects on social media and discussion forums to express their opinions. Sentiment analysis for Maghrebi dialects is still in a earlier stage. Most of the work are recent compared to contributions related to MSA or a relatively more-resourced dialect such as Egyptian dialect.

In [32], the authors proposed a lexicon-based approach for sentiment analysis of Algerian dialect. They used a manually annotated dataset and three Algerian Arabic lexicons.

Authors of [33] presented an approach for emotion analysis of Tunisian Facebook pages. They introduced a new method to create emotion dictionaries by using emotion symbols as sentiment polarity indicators. Recently, in [34] the focus was also made on Tunisian dialect sentiment analysis. Their approach is based on machine learning techniques for determining comments polarity. Within this research, a corpus of 17K Facebook comments has been created and annotated.

F. Machine translation

Machine translation is an other issue related to Arabic dialects and Maghrebi ones particularly. In fact, Machine translation requires specific resources like parallel corpora in the context of data-based approach and strong linguistic studies in the case of rule-based approach, while this dialects suffer from a lack of resources especially parallel corpora. Few efforts have been deployed to deal with machine translation of Maghrebi dialects,

most issues are not yet solved. There is still much work to be done in this area.

In [35] is proposed a machine translation system between MSA and Tunisian dialect verbal forms (in both directions). It is based on deep morphological representations of roots and patterns (a specific feature of Arabic). Another work dedicated to Tunisian dialect is described in [36]. The authors attempted to translate Tunisian dialect text of social media into MSA by using a bilingual lexicon and a set of grammatical mapping rules and a disambiguation step.

In [37], a machine translation system from Moroccan dialect to MSA is presented. The work used a rule-based approach in addition to a language model. The system used transfer rules based on a morphological analysis (with Alkhalil morphological analyzer [38] which the authors adapted to Moroccan dialect).

In [39] a hybrid machine translation system combining statistical and rule-based approaches is presented. It translated from Arabic dialects to English. Dialects concerned by this study were those of the middle-east in addition to Tunisian, Moroccan and Libyan. MSA was as a pivot language. This system showed that the hybridization of statistical and rule-based approaches performs better than using each approach separately.

Authors of [40] presented PADIC a multi-dialect Arabic corpus that includes MSA, Maghrebi dialects (Algerian and Tunisian and in the last version Moroccan) and Levantine dialects (Palestinian and Syrian). They conducted several experiments on different Statistical Machine Translation (SMT) systems between all pairs of languages (MSA and dialects). They studied the impact of the language model on machine translation by varying the smoothing techniques and by language model interpolation.

G. Other resources

In [41] authors dealt with the detection of sentence boundary in transcribed spoken Tunisian Arabic. They proposed a rule-based method and a statistical method, in addition to a third method which combines these two last. Their detection system has been used to improve the accuracy of a POS tagger of transcribed Tunisian dialect.

In [42] an automatic diacritics restoration system was built for Algiers dialect. The system was based on a statistical approach and allowed to vocalize the Algerian part of PADIC [40]. This vocalized corpus has been used in [43] for the purpose of grapheme to phoneme conversion. This last combined a rule-based and a statistical approaches.

In [44], the authors proposed a method to disambiguate the output of a morphological analyzer of the Tunisian dialect (cited in [27]) by using machine-learning techniques.

4. CONCLUSION

We focused in this paper on Maghrebi Arabic dialects particularly Algerian, Moroccan and Tunisian Arabic. After a linguistic overview, we provided a survey of the research work dealing with these languages. Several comments can be made based on this work. In view of the various published works, we can see that the research efforts dealing with Maghrebi Arabic dialects are at an early stage. Most of the research work dealing with these dialects has been devoted to the construction of corpora and lexicon. This is mainly due to the fact that these languages are under-resourced. The identification task has also been researched. While the morphology of the Maghrebi Arabic dialects has been addressed in few papers, the syntactic analysis remains totally ignored. It is also worth noting the small number of works devoted to machine translation of these dialects. In addition, these few existing contributions are dedicated to the translation between dialects and MSA, no work has considered the French language.

REFERENCES

1. F. Khelef and R. Kebieche, "Evolution ethnique et dialectes du maghreb," Synergies Monde Arabe no 8 (2011).
2. G. Grandguillaume, "L'arabisation du maghreb," Revue d'Aménagement linguistique, Aménagement linguistique au Maghreb, Off. Québécois de la langue française pp. 15–40 (2004).
3. A. Mahfoudhi, "Agreement lost, agreement regained: A minimalist account of word order and agreement variation in Arabic," California Linguist. Notes **27**, 1–28 (2002).
4. M. L. Souag, "Explorations in the syntactic cartography of Algerian Arabic," Ph.D. thesis, School of Oriental and African Studies (University of London (2006).
5. M. Ennaji, *Multilingualism, cultural identity, and education in Morocco* (Springer Science & Business Media, 2005).
6. R. Tachicart, K. Bouzoubaa, and H. Jaafar, "Building a Moroccan dialect electronic dictionary (MDED)," in "5th International Conference on Arabic Language Processing," (2014), pp. 216–221.
7. Y. Samih and W. Maier, "An Arabic–Moroccan darija code–switched corpus." in "Proceedings of 10th Language Resources and Evaluation Conference (LREC 2016)," (2016).
8. J. Younes, H. Achour, and E. Souissi, *Constructing Linguistic Resources for the Tunisian Dialect Using Textual User-Generated Contents on the Social Web* (Springer International Publishing, Cham, 2015), pp. 3–14.
9. A. Hamdi, N. Gala, and A. Nasr, "Automatically building a Tunisian lexicon for deverbal nouns," in "Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects," (2014), pp. 95–102.
10. J. Karoui, M. Graja, M. Boudabous, and L. H. Belguith, "Domain ontology construction from a Tunisian spoken dialogue corpus," in "International Conference on Web and Information Technologies, ICWIT'2013," (2013).
11. A. Masmoudi, Y. Estève, M. E. Khmekhem, F. Bougares, and L. H. Belguith, "Phonetic tool for the Tunisian Arabic," in "Spoken Language Technologies for Under-Resourced Languages," (2014).
12. I. Zribi, M. Ellouze, L. H. Belguith, and P. Blache, "Spoken Tunisian Arabic corpus" STAC": Transcription and annotation." Res. computing science **90**, 123–135 (2015).
13. R. Cotterell, A. Renduchintala, N. Saphra, and C. Callison-Burch, "An Algerian Arabic-French code-switched corpus," in "Workshop on Free/Open-Source Arabic Corpora and Corpora Processing Tools Workshop Programme," (2014), p. 34.
14. S. Bougrine, A. Chorana, A. Lakhdari, and H. Cherroun,

"Toward a web-based speech corpus for Algerian Arabic dialectal varieties," WANLP 2017 (co-located with EACL 2017) p. 138 (2017).
15. M. Djellab, A. Amrouche, A. Bouridane, and N. Mehallegue, "Algerian modern colloquial Arabic speech corpus (AMCASC): regional accents recognition within complex socio-linguistic environments," Lang. Resour. Eval. **51**, 613–641 (2017).
16. S. Bougrine, H. Cherroun, D. Ziadi, A. Lakhdari, and A. Chorana, "Toward a rich Arabic speech parallel corpus for Algerian sub-dialects," in "LREC'16 Workshop on Free/Open-Source Arabic Corpora and Corpora Processing Tools (OSACT)," (2016), pp. 2–10.
17. K. Abidi, M. Menacer, and K. Smaïli, "CALYOU: A comparable spoken Algerian corpus harvested from youtube," Proc. Annu. Conf. Int. Speech Commun. Assoc. (Interspeech) pp. 3742–3746 (2017).
18. S. Bougrine, H. Cherroun, and D. Ziadi, "Prosody-based spoken Algerian Arabic dialect identification," in "International Conference on Natural Language and Speech Processing, ICNLSP'2015," (2015).
19. I. Benali, "The identification of two Algerian Arabic dialects by prosodic focus," in "7th Tutorial and Research Workshop on Experimental Linguistics, ExLing 2016," (2016), p. 37.
20. M. Hassine, L. Boussaid, and H. Messaoud, "Maghrebian dialect recognition based on support vector machines and neural network classifiers," Int. J. Speech Technol. **19**, 687–695 (2016).
21. I. Guellil and F. Azouaou, "Arabic dialect identification with an unsupervised learning (based on a lexicon). application case: Algerian dialect," in "Computational Science and Engineering (CSE) and IEEE Intl Conference on Embedded and Ubiquitous Computing (EUC) and 15th Intl Symposium on Distributed Computing and Applications for Business Engineering (DCABES), 2016 IEEE Intl Conference on," (IEEE, 2016), pp. 724–731.
22. S. Tratz, D. M. Briesch, J. Laoudi, and C. R. Voss, "Tweet conversation annotation tool with a focus on an Arabic dialect, Moroccan darija." in "LAW@ ACL," (2013), pp. 135–139.
23. I. Zribi, M. Graja, M. E. Khmekhem, M. Jaoua, and L. H. Belguith, "Orthographic transcription for spoken Tunisian Arabic," in "International Conference on Intelligent Text Processing and Computational Linguistics," (Springer, 2013), pp. 153–163.
24. N. Habash, M. T. Diab, and O. Rambow, "Conventional orthography for dialectal Arabic." in "Proceedings of the International Conference on Language Resources and Evaluation (LREC)," (2012), pp. 711–718.
25. I. Zribi, R. Boujelbane, A. Masmoudi, M. Ellouze, L. H. Belguith, and N. Habash, "A conventional orthography for Tunisian Arabic." in "Proceedings of the International Conference on Language Resources and Evaluation (LREC)," (2014), pp. 2355–2361.
26. H. Saadane and N. Habash, "A conventional orthography for Algerian Arabic," in "ANLP Workshop 2015," (2015), p. 69.
27. I. Zribi, M. E. Khemakhem, and L. H. Belguith, "Morphological analysis of Tunisian dialect," in "International Joint Conference on Natural Language Processing," (2013), pp. 992—996.
28. R. Boujelbane, M. Mallek, M. Ellouze, and L. H. Belguith, "Fine-grained pos tagging of spoken Tunisian dialect corpora," in "International Conference on Applications of Natural Language to Data Bases/Information Systems," (Springer, 2014), pp. 59–62.
29. A. Hamdi, A. Nasr, N. Habash, and N. Gala, "POS-tagging of Tunisian Dialect Using Standard Arabic Resources and Tools," in "Workshop on Arabic Natural Language Processing," (Beijing, China, 2015), pp. 59 – 68.
30. F. Al-Shargi, A. Kaplan, R. Eskander, N. Habash, and O. Rambow, "Morphologically annotated corpora and morphological analyzers for Moroccan and sanaani yemeni Arabic," in "10th Language Resources and Evaluation Conference (LREC 2016)," (2016).
31. S. Harrat, K. Meftouh, M. Abbas, and K. Smaïli, "Building resources for Algerian Arabic dialects," in "Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)," (2014), pp. 2123–2127.
32. M. Mataoui, O. Zelmati, and M. Boumechache, "A proposed lexicon-based sentiment analysis approach for the vernacular Algerian Arabic," Res. Comput. Sci. **110**, 55–70 (2016).
33. H. Ameur, S. Jamoussi, and A. B. Hamadou, "Exploiting emoticons to generate emotional dictionaries from facebook pages," in "Intelligent Decision Technologies 2016," (Springer, 2016), pp. 39–49.
34. S. Medhaffar, F. Bougares, Y. Estève, and L. Hadrich-Belguith, "Sentiment analysis of Tunisian dialects: Linguistic ressources and experiments," (2017).
35. A. Hamdi, R. Boujelbane, N. Habash, and A. Nasr, "The effects of factorizing root and pattern mapping in bidirectional Tunisian-standard Arabic machine translation," in "MT Summit," (2013).
36. F. Sadat, F. Mallek, M. Boudabous, R. Sellami, and A. Farzindar, "Collaboratively constructed linguistic resources for language variants and their exploitation in nlp application, the case of Tunisian Arabic and the social media," in "Proceedings of the Workshop on Lexical and Grammatical Resources for Language Processing," (Association for Computational Linguistics and Dublin City University, 2014), pp. 102–110.
37. R. Tachicart and K. Bouzoubaa, "A hybrid approach to translate Moroccan Arabic dialect," in "Proceedings of the 9th International Conference on Intelligent Systems: Theories and Applications (SITA-14)," (2014), pp. 1–5.
38. A. Boudlal, A. Lakhouaja, A. Mazroui, A. Meziane, M. O. A. o. Bebah, and M. Shoul, "Alkhalil morpho sys1: A morphosyntactic analysis system for Arabic texts," in "Proceedings of the International Arab Conference on Information Technology, ACIT," (2010).
39. H. Sawaf, "Arabic dialect handling in hybrid machine translation," in "Proceedings of the Conference of the Association for Machine Translation in the Americas (AMTA), Denver, Colorado," (2010).
40. K. Meftouh, S. Harrat, S. Jamoussi, M. Abbas, and K. Smaïli,

"Machine translation experiments on PADIC: A Parallel Arabic Dialect Corpus," in "Proceedings of the 29th Asia Conference on Language, Information and Computation (PACLIC)," (2015), pp. 26–34.

41. I. Zribi, I. Kammoun, M. Ellouze, L. Belguith, and P. Blache, "Sentence boundary detection for transcribed Tunisian Arabic," Bochumer Linguist. Arbeitsberichte p. 323 (2016).
42. S. Harrat, M. Abbas, K. Meftouh, and K. Smaïli, "Diacritics restoration for Arabic dialect texts," in "Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)," (2013), pp. 125–132.
43. S. Harrat, K. Meftouh, M. Abbas, and K. Smaïli, "Grapheme to phoneme conversion: An Arabic dialect case," in "Proceedings of 4th International Workshop On Spoken Language Technologies For Under-resourced Languages SLTU," (2014), pp. 257–262.
44. I. Zribi, M. Ellouze, L. H. Belguith, and P. Blache, "Morphological disambiguation of Tunisian dialect," J. King Saud Univ. - Comput. Inf. Sci. **29**, 147 – 155 (2017).

Salima Harrat is an assistant professor in Computer Science at Ecole Normale Supérieure de Bouzaréah. Her research interests are topics related to Natural language processing, mainly for Arabic language. Recently, She focused on Arabic dialects processing in the context of machine translation. She published in several international conferences (Interspeech' 2013 and 2014, CICLING'2015 and 2017, PACLIC'2015, SLTU'2014) and journals (Information Processing and Management (IPM' 2017), International Journal of Advanced Computer Science and Applications (IJACSA' 2016). She was a member of the research group of TORJMAN which is a research national project (PNR) from 2011 to 2013.

Karima Meftouh studied at Badji Mokhtar - Annaba University where she received the degree of computer science engineer in 1992 and the Master degree in 2000. Since 2001, she is a teacher researcher at the computer science department of Badji Mokhtar University and a member of the research group in artificial intelligence within the LRI laboratory. She defended her PHD in 2010. She was interested particularly in Arabic statistical language modeling. Her research was the object of several publication in various conferences: CITALA' 2007, SIIE' 2008, JADT' 2008, ICAART' 2009 and journals: IRECOS' 2009 , AJSE' 2010. In the last years, her research interest concerns machine translation of Arabic dialects. As part of this work, she published in several international conferences (Interspeech' 2013 and 2014, CICLING'2015 and 2017, PACLIC'2015, SLTU'2012 and 2014) and journals (Information Processing and Management (IPM' 2017), International Journal of Advanced Computer Science and Applications (IJACSA' 2016). She was a member of the research group of TORJMAN which is a research national project (PNR) from 2011 to 2013. She is involved in the program committee of SETIT, Language Resources and Evaluation journal, ISGA journal.

Kamel Smaïli is Professor at the university of Lorraine since 2002, he obtained a PhD from the university of Nancy 1 in 1991 on automatic speech recognition. He defended his HDR in 2001 (Statistical language modelling: from speech recognition to machine translation). His research interest since more than 20 years concerns statistical language modelling for automatic speech recognition. Since 2000 he oriented his research towards speech to speech translation. He participated to several European and national projects concerning automatic speech recognition: COCOS, MULTIWORKS, COST, MIAMM, IVOMOB (RNRT project) and CMCU. He advised 14 PhD and HDR students and he participated to more than 35 PhD committees in France, Germany, Spain and Algeria. He took part to several program committees: Interspeech, Eurospeech, ICSLP, ICASSP, TALN, ICWMI, SIIE, TAIMA, Machine Translation, Computer speech and language, Speech communication, Journal of Natural Language Engineering,… He has been invited several times to give talks as invited speakers in Japan, France, Tunisia, Algeria and Morocco. He

published 90 papers in international conferences and journals and 20 papers in francophone conferences. Furthermore, Mr Smaili was the head of MIAGE (Master and Bachelor) department for 7 years and the head of UFR (equivalent to a faculty) Mathematics and Informatics for 5 years where he managed more than 30 permanent people and 120 temporary positions, 500 students.