International Journal of Informatics and Applied Mathematics
e-ISSN:2667-6990 Vol. 2, No. 2, 1-14

Arabic Algerian Oranee Dialectal Language Modelling Oriented Topic

Fréha Mezzoudj, Mourad Loukam, and Fatma Zohra Belkredim

Université Hassiba Benbouali Chlef, Algérie
{f.mezzoudj,m.loukam,f.belkredim}@univ-chlef.dz

Abstract. The Modern Standard Arabic (MSA) is the formal language used in the Arab world. In Algeria, the MSA and other varieties of informal Arabic dialects are used in the everyday matter communication. These dialects are by no means subject to further regional variations: eastern, western, central or southern. The Oranee dialect is the most important and used one in the west of Algeria. However, it is an under-resourced language, which lacks both audio and textual corpora. In this paper, we present the most particularities of this western Algerian dialect and introduce a natural language processing on an Oranee textual corpus. A MSA transcribed discourse could contain some dialect vocabularies and viceversa. Therefore, we propose to interpolate dialectal language models and MSA ones with respect to some topics. The best obtained interpolation weights are related to Religion topic data.

Keywords: Algerian dialect · Oranee dialect · Modern Standard Arabic · Natural Language Processing · Language Modelling · Speech Recognition · topics.

1 Introduction

Arabic language is the most important and widely spoken and written Semitic language. It is the official language of 22 countries, spoken by more than 400 million speakers. It is recognized as the 4th most used language on Internet [1]. The term *Arabic language* is often used to refer to a collection of multiple variations: Ancient Arabic (AA), Classical Arabic (CA), Modern Standard Arabic (MSA) and spoken Arabic Dialects (AD).

The AA, or pre-Islamic Arabic, can be found in pre-islamic poems and history books. Classical Arabic (CA) is considered as the language of Quran which is the holy book of Islam. It is still used in muslim prayers and other religious activities. However, the Modern Standard Arabic (MSA) is the formal language used in the Arab world. It is the modern descendant of CA, it is in particular, the language of the written Arabic press, the formal discourse, the political speeches, the official venues and education.

In general, *dialects* are variations of the same language, specific to geographical regions or social groups. The main Arabic dialects are four: Gulf, Levantine, Eygptian and Maghrebi. The Gulf Arabic includes dialects of Kuwait, Saudi Arabia, Bahrain, Qatar, United Arab Emirates, Oman and Iraq. Levantine Arabic includes the dialects of Lebanon, Syria, Jordan and Palestine. Eygptian Arabic covers the dialects of Egypt and Sudan. However, Maghrebi Arabic includes the dialects of Tunisia, Morocco, Mauritania, Libya and Algeria [2]. These dialects differ from the standard Arabic form; though they have their particularities, they are still mutually intelligible.

Historically, the Berber tribes were the ancient local inhabitants of Algeria. Despite the successive waves of invaders including the Phoenicians, the Romans, the Vandals and the Byzantines, the Berbers have succeeded to preserve their culture and language, *Tamazight* which is not related to Arabic. With the Islam revelation and expansion around 7^{th} centruy, many arabic muslims have immigrated since early time (678) to North Africa. The Arab tribes exported their language and specially the Quran language. This important event expanded the use of Ancient and Classical Arabic as the religious, liturgical and everyday language in these areas. However, the *Tamazight* language is still used until now.

Also, the contact with other invaders in Algeria such as Spanish colonialism (15th century), the Ottoman- Turkish command (1518 - 1671) and the European inhabitants (in the nineteenth century) during French colonialism (1830-1962) have participated to create new language practices such as the use of many Algerian Arabic dialects including foreign words.

After the independence of Algeria in 1962, the Algerian authorities mainly the Nationalists tried to regain the Arab identity by establishing Arabic as the official language of the country. Arabic became the official language instead of French which was the official one during the colonisation period. However, for many reasons, this Arabisation process was not saved. The French language continued to play an important role in the Algerian society in various domains.

The MSA (written and spoken) is used only by the Intelligentsia community, in official situations. The most vocabulary of the Algerian Arabic dialects used

in the daily life is derived from CA and MSA. However, different languages which have existed in the Algerian territory in different periods of its history, also influence them. These different Algerian Arabic dialects are also used by large communities of Algerian immigrant speakers in Europe and America.

Algeria which is a big country, dialects are by no means subject to further regional variations: eastern, western, central or southern. The Oranee dialect (OranD) is the most important and used one in the west of the country. However, it is an under-resourced language which lacks both audio and textual corpora.

Investigating in Dialectal Arabic is useful for many research directions [3,4,5] such as Spoken Arabic Dialect Identification, Collecting Arabic Dialect Corpora, Arabic Dialect Machine Translation, Text Diacritiation in Arabic Dialect, etc. There are also many works made on textual data, like Arabic Dialect Word Segmentation, Arabic Textual Dialect Identification and language modelling.

In the current work, we are interested in textual monolingual corpora written resources by collecting and modelling a first version of an Arabic Algerian Oranee Dialectal corpus. We present the most particularities of this western Algerian dialect and introduce a natural language processing on an Oranee textual corpus. Some phonetic characteristics and vocabulary of the latent's dialect could affect the MSA pronunciation and oral discourse. Also, an MSA transcribed discourse could contain some dialect vocabularies and viceversa. In order to improve the language modelling and the language recognition, we propose also to interpolate a dialectal language models with MSA ones.

The reminder of this paper is organised as follows. In the next section, we review some related work that contributed to built corpora for both MSA and Algerian dialects. In section 3, we give a brief description of Algerian dialects features in general and particularly for the Oranee one. These particularities are respected during the choice of the corpus text. In section 4, first we introduce the dialectal language modelling and then their interpolation with an MSA language models. Finally, we conclude about the most important results and identify some directions for future works.

2 Related Works

In general, there is minor MSA speech and textual datasets production. In addition, they are designed for a specific research purpose and most of them are not publicly (freely) available as confirmed by Zaghouani in his critical survey [6]. For Arabic dialects and espcially Algerian ones, the situation is more dramatic. We tried in this section to present some works related to the Arabic Algerian.

The language resources are essential components in all Natural Language Processing (NLP) applications, Automatic Speech Recognition Systems, Machine Translation Systems and so on. They are divided into written and oral resources. Therefore it is necessary to provide, as much as possible, standard and free resources for the Arabic NLP community.

Few are the sociolinguistic and dialectological studies on Algerian Arabic. Some well-known linguist who has worked on Algerian dialects is the historian

Ibn Khaldoun (the 14^{th} centruy) [7]. Some phonological studies about eastern Algerian dialects can be found in [8,9,10].

For an automatic processing perspective, some natural language processing tasks based on Algerian Arabic audio corpora representing different accents and dialects have received relatively little attention. The authors in [11] presented an ALGerian Arabic Speech Database (ALGASD) recorded from 200 phonetically rich and balanced Arabic sentences (1080 utterances) by 300 Algerian native speakers selected from eleven departments (dubbed Wilayas) with different regional accents of MSA spoken in Algeria. The ALGASD corpus was used in some vocalic rhythms classification works on MSA and on the Algerian Arabic dialects [12]. The authors conclued that the Arabic language is classified as stressed language but Algerians tend to pronounce the MSA as an intermediate language between stressed and timed languages. The pronunciation of vowels by Algerian speakers increases in terms of vocalic duration when compared with Eastern Arabic countries speakers (such as Egyptian, Lebanese and Jordanian). In addition, Algerian dialect presents higher consonantal proportions and lower vocalic rhythm values than ALGASD corpus.

The authors in [13], proposed a dialect identification system of five Arabic Maghreb dialects: Moroccan, Tunisian, and 3 dialects of the western (Oranian), central (Algiersian), and eastern (Constantinian) regions of Algeria. They applied an GMM-UBM system and an improved GMM-UBM system using SVM feature selection on five datasets of about 50 houres for each dialect. The experiments show that their approaches significantly improve identification performance over purely acoustic features with an identification rate of 80.49%.

In the same context, the authors of [14] showed that Algiers and Oranee dialects can be identified by an acoustical processing and prosodic cues. The recurrences of prosodic patterns can differentiate the two dialects using narrow and interrogative focus. The narrow focus is marked by $F0$ rise and often accompanied by an increase of duration and intensity. The used acoustic database was collected from the recorders of 20 Algiers' speakers and 20 Oran's speakers and processed using Praat tool. The speakers of the two dialects were better identified in interrogative focus with an identification rate of 80%.

Considering both of audio and textual data, some pioneer studies have actively worked on eastern (Annaba's dialect) and central (Algiers's dialect) Algerian dialects [15,16,17]. These works contributed to collect large Algerian dialectal corpora from Scratch using direct recording and movies. The size of the two dialectal corpora was increased on translating each one to the other as presented in [18]. In [19], a multilingual parallel textual corpus called PADIC (for: Parallel Arabic DIalect) composed of 6400 sentences for MSA and five Arabic dialects, including two from Algeria: an eastern and central ones (Annabi and Algiers's dialects), was presented and used in machine translation experiments.

In [20], the authors extract and collect from Youtube, Algerian data by crawling 22000 videos by using 250 queries related to Algerian news with diffrent topics, such as politic, Algerian celebrity, Algerian football, etc. This corpus is called CALYOU (for: Comparable spoken ALgerian extracted from Youtube).

In [21], the authors described a preliminary version of a direct recorded Algerian dialect corpus ALG -DARIDJAH (for ALGerian Darijah) that encompasses 17 sub-dialects with 109 native speakers and more than 6 K utterances. In [22], they introduced a preliminary version of the Web based corpus KALAM'DZ of 8 Algerian dialects, crawled from some Algerian TV and Youtube resources.

Even in [23], the authors collected an Algerian Modern Colloquial Arabic Speech Corpus (AMCASC) from telephone conversations of 735 speakers, representing three Algerian dialectal varieties. The corpus was used for training an automatic regional accents' recognition system.

Any of these previous works has focused sufficiently and especially on the Oranee dialect. The paper that we present is the first ever work, as far as we know, introducing the specificities of the Oranee dialect and using an important textual corpus for automatic processing and language modelling oriented topic.

3 Arabic Algerian Oranee Dialect

The spoken Algerian Arabic dialects (called *Darija*) are rich complex languages and are well used in the everyday matter communication, movies, social networks, TV emissions, telephone conversations and so on.

Oran is a Northwestern Algerian city (in the north of Africa) and it is known as the second metropolis in Algeria (after the capital, Algiers). It has benefited from its maritime coastal attractivity and its industrial and commercial interests. The Oranee dialect is spoken in western Algeria, precisely from Ain-Temouchent until the limits of the city of Ténès (Chlef), see the figure 1.

In this section, we give a brief features description of Algerian dialects in general and particularly for the Oranee one. These syntactic and phonological particularities must be respected during the collect of the Oranee textual corpus.

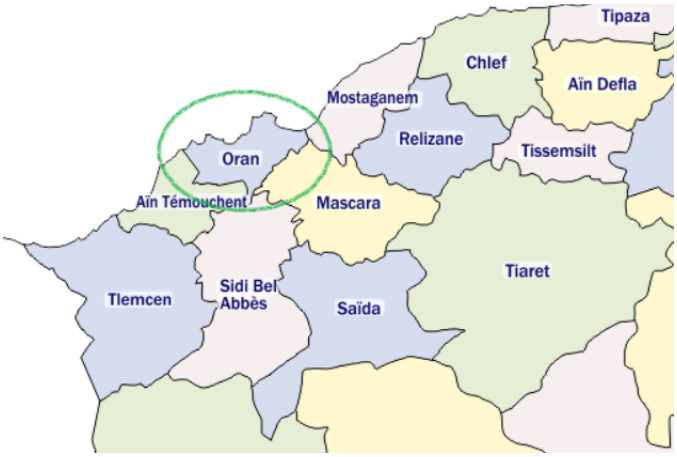

Fig. 1: Oran, the Algerian city and geographic extent of the dialect Oranee' use.

3.1 Vocabulary

Evidently, the Arabic Algerian Dialects are basically based on Classical Arabic (CA). Moreover, they contain original or code-switched words, expressions, and linguistic structures from Arabic language (CA), different Berber varieties, French, Spanish, Turkish as well as other Mediterranean Romance languages. See some samples in the table 1.

For historical reasons, the Algerian Oranee Dialect is influenced by the Spanish language whereas the Algiers' dialect is rather influenced by the Turkish language. Both of their vocabularies contain original or modified (code-switched) French words. Some future statistical studies could be helpful for an accurate categorisation. Also, an important specification of Oranee dialect is the use of

Table 1: Samples of Dialect vocabulary influenced by different languages.

MSA (Buckwalter)	Algerian Dialect	Source	English
كرسي (krsy)	كرسي (krsy)	Arabic: كرسي	Chair
ذهب (*hb)	دهب (dhb)	Arabic: ذهب	Gold
فرَاشة (frA$p)	فرطـطو (frTTw)	Berber: *Fertatoo*	Butterfly
سلحفَاة (slHfAp)	فكرون (fkrwn)	Berber: *Fekroon*	turtle
مطبخ (mTbx)	كوزينة (kwzynp)	French: *Cuisine*	Kitchen
محفظة (mHfZp)	كرطَاب (krTAb)	French: *Cartable*	Satchel
حفلة (Hflp)	فيشطة (fy$Tp)	Spain: *fiesta*	Feast
الحَامي (AlHAmy)	كرنتيكة (krntykp)	Spain: *Calentica*	Kind of Food
نقود (nqwd)	صوَاردَا (SwrdA)	Italian:*Soldo*	Penny
إفرنجي (<frnjy)	قَاوري (GAwry)	Turck: *Gavur*	Occidental
فقير (fqyr)	زوَالي (zwAly)	Turck: *Zavalli*	Poor

some vocabulary such as the samples in the non-exhaustive list in the table 2, using Buckwalter notations [24].

Note that the word أرجوك(>rjwk)in the last line, has a semantic contextual translation which means *May God bless you father* but this meaning is hidden in the everyday use.

Table 2: Samples of Dialect Oranee vocabulary.

MSA (Buckwalter)	Oranee Dialect	English
نعم (nEm)	وَاه (wAh)	yes
لَا (lA)	لّ (la)	no
مَاذَا (mA*A)	شوَالَا ($wAlA)	what
بنت (bnt)	شيرة ($yrp)	girl
ولد (wld)	شير ($yr)	boy
أَرجوك (>rjwk)	حمبوك (Hmbwk)	please

3.2 Phonetic

The phonetic system of Modern Standard Arabic (MSA) has 34 phonemes: 6 vowels (3 short vowels: (**ar**. اَ, **buck**. a)(اُ, u), (اِ, i) with 3 opposite long ones: (و, w), ي, y), (ا, A) and 28 consonants [1]. Algerian Arabic and specially the Oranee one is different from MSA mainly phonologically and morphologically. The literature confirms the Bedouin affiliation of Oranee dialect [9]. The reason behind this genealogical conflict is the co-presence of Bedouin and sedentary dialect-background migrants in Oran city. The majority constitutes the Bedouin group.

The important specification of this Bedouin affiliation is that the MSA letter [q] is pronounced, in most of the cases, as a voiced velar noted phonetically [g] (ڤ), which in not an Arabic letter. For example as in the dialectal sentences: **ar**. قَالِي (**buck**. GAly) (**en**. He told me) or **ar**.قَاع رَاحو (**buck**. GAE rAHw) (**en**. They're all gone).

In addition, there are two other non-Arabic characters which appear in Algerian Oranee Dialect, when we use foreign word to Arabic language with respect (or not) to the syntactic or morphological properties of that foreign language, such as French, for example: **ar**. مغَاديش يڤوطي **buck**. (mgAdy$ yVwTy) (**en**. He will not vote.). In the original encoding Buckwalter scheme, they don't exist. To simplify this work, we propose and use an extended annotation: G and P and V for these three non-Arabic but Dialectal characters extended from three foreign letters ڤ [g] and پ [p] and ڤ [v].

[1] Notation: **ar**. for Arabic letter, **buck**. for Standard Buckwalter notation and **en**. for english translation

Some adorable and strong MSA letters or sounds are not properly used in the Western Algerian dialects (neither pronounced nor written). The interdental fricative sounds [θ] (ث, v) and [ð] (ذ, *) are pronounced nearly to the aspired stop [t] (ت, t) and [d] (د, d) respectively. In addition, the strong phonems (ض, D) and (ظ, Z) are pronounced as a low and a weak-emphatic [d] (ض D) or close to a strong [d] (د, d) in some cases (see the table 3).

Table 3: Samples of some MSA letters' phonetic and their correspondents in Algerian Dialects (Algeris and Oranee's ones).

Letter (buck.)		MSA (IPA)	Algerian Dialect	
			Algiers	Oranee
ق	(q)	/q/	[q]	[g]
ج	(j)	/d͡ʒ/	[d͡ʒ]	[ʒ]
ث	(v)	/θ/	[θ], [t]	[t]
ذ	(*)	/ð/	[ð], [d]	[d]
ظ	(Z)	/ðˤ/	[dˤ]	[dˤ]
ض	(D)	/dˤ/	[dˤ]	[dˤ]

4 Experiments, Results and Discussions

4.1 Dialectal Corpus

PADIC (Parallel Arabic DIalectal Corpus) [19] is a multi-dialectal corpus built in the framework of the Algerian Research Project TORJMAN, led by Scientific and Technical Research Center for the Development of Arabic Language, is composed of 6 dialects: two Algerian dialects (Algiers and Annaba cities), Palestinian, Syrian, Tunisian, Moroccan) and MSA. Because of the lack of an available Oranee lexicon, an Oranee textual corpus (noted *OranD_ V1*) was obtained by a manual translation (by 3 Oranee native people with high educational level) of a sub-corpus extracted from the MSA PADIC corpus to Oranee Dialect. All the syntactic and phonological particularities of the Oran dialect have been respected as far as possible. The obtained OranD_V1 corpus is a first version of Oranee dialectal textual corpus, we are working to improve it to obtain an augmented standard version.

Before any automatic processig, it is advisable to explore the OranD_V1 dataset in terms of word counts (see the table 4).

4.2 Dialectal Language Model

Language modelling is to create statistical models that are able to capture the regularities of a natural language. So, the language model tries to capture the

Table 4: General Statistics of the Oranee Dialect corpus version 1 (OranD-V1) (D. for Dialect)

Data	#Files	#Sentences	#Words	
			Raw	Segmented
TrainD	8	2 695	16 457	23 752
DevD	1	251	1 154	1 614
TestD	1	250	1 712	2 458

structure of a language through word frequencies. It assigns a probability of a valid sentence using the sequence of its words. The most popular form of a statistical language model is the *n-gram model* which is notable for its simplicity and computational efficiency. The measure of language models complexity is the *perplexity*. Therefore, when comparing two language models, the one which gives the lowest perplexity is the appropriate model [25,26].

We note that the most important factors that influence the quality of the resulting n-gram model are the choice of the order n and of the smoothing technique. Relatively to the training dataset's quantity, we use 3-ngram model. The different LMs are smoothed using the *modified Kneser-Ney* as done in [27] which gives commonly the best results.

In this paper, all the n-gram language models were built with the SRILM (SRI Language Modelling Toolkit) [2] [28,29] a powerful used toolkit for building and applying statistical language models.

The available training Arabic dialectal data, presented in table 4, was used for training Language Models (LM). The development (DevD) and the test (TestD) data are useful to estimate the LMs perplexities. The data encoding was changed to UTF-8 coding for applying the SRILM functions directly on the Arabic texts, and avoid the transliteration.

In practice, text preprocessing for Natural Language Processing (NLP) tasks can be divided into two broad categories, noise removal and normalisation. Data components that are redundant to the core text analytics can be considered as noise. Handling multiple occurrences or representations of the same word is called normalisation. There are two kind of normalisation: stemming and lemmatisation. Note that all these steps are not necessary for language modelling.

The morphology of Arabic dialectal words shares a lot of features with MSA morphology. For the morphological processing, we use without any adaptation the Farasa toolkit [3] [30] which is a fast and accurate text processing toolkit for Arabic text. Farasa consists of the segmentation and tokenization based on SVM-rank that uses a variety of features and lexicons to rank possible segmentations of a word : likelihoods of stems, prefixes, suffixes and so on.

First, we used the raw textual Arabic Dialectal data to train the LM without any morphological processing. Second, we applied the Farasa toolkit on the whole

[2] SRILM : http://www.speech.sri.com/projects/srilm. Last accessed 01 jan 2019
[3] Farasa : http://qatsdemo.cloudapp.net/farasa/. Last accessible 20 apr 2019

data and re-trained the LM. The table 5 summarises the preliminary result. It is clear, that the segmentation and tokenisation; even without any adaptation to the Oranee dialect, reduces efficiently its complex morphology.

Table 5: Language models perplexities using Training OranD-V1: Perplexities and n-grams counts.

Data	Raw		Segmented	
LM-OranD	ppl-DevD	ppl-TestD	ppl-DevD	ppl-TestD
	572.7	997.01	169.5	260.9
#1-gram	4 419		3 359	
#2-gram	13 449		12 592	
#3-gram	627		1 152	

4.3 MSA Corpus

For training MSA language models, we used textual corpus based on many topics such as economy, religion, sport, politic, Science-technology, music and so on. This textual MSA corpus is Arabic Keyphrase Extraction Corpus (AKEC), which was selected from four sources: Arabic Newspapers Corpus, Corpus of Contemporary Arabic, Essex Arabic Summaries Corpus and Open Source Arabic Corpora [31]. It is advisable to explore the dataset in terms of word counts (see the table 6).

Table 6: General Statistics of the used training MSA sub-corpus (only 2 topics, Science-technology (noted *SciTec*) and Religion (noted *Religi*).

Data	#File	#Sentence	#Segmented Words
SciTec	18	335	15 112
Religi	17	303	25 148

Our motivation in this experimental work is that any MSA transcribed discourse in Broadcast or TV emission of any Algerian Speaker could contain some dialectal vocabularies and vice versa. Therefore, we propose to interpolate MSA language models with the dialectal ones.

4.4 Dialectal -MSA Language Model

We have interpolated the dialectal LM each time with an LM trained on a specific MSA data related to many topics such as economy, religion, sport, politic,

Science-technology, art and music. The MSA LMs, which match better with the dialectal LM are trained on two topic data: Religion then Science-technology (see the table 7).

Table 7: Best LM, interpolated from the Dialectal corpus and the MSA data.

Data	**Individual LMs**	**Interpolated LM**		
	ppl *DevD*	**weights**	**ppl** *DevD*	**ppl** *TestD*
OranD	169.47	0.9941	184.6	302.5
SciTec	1675.10	0.0001		
Religi	1896.59	0.0058		

The baseline Dialectal LM has a perplexity of 169.47 based on Dialectal data for development step. We notice that after the interpolation with the MSA LM, the final perplexity of the interpolated LM (Dialect-MSA) has increased to 184.6 which is not good for language modelling but it reflects the real difficult situation in everday communication. Also, this phenomenon may be due to the non-homogeneous quality of data (MSA and Dialect).

So, the best obtained weights are related to Religion then Science-Technology topics textual data. The weight given to the interpolated language model estimated from the textual data relative to Religion topic (Religi)is higher than the textual data of Science and technology topic (SciTec) (0.0058 instead of 0.0001). We can explain this result by the fact that MSA religion vocabulary is correctly used in dialectal conversation comparing to other topics such as Science and technology. The used vocabulary of Science and technology can be taken from MSA or French, we plan to add some French language models for this kind of topic.

5 Conclusion

In Algeria, the MSA and other varieties of informal Arabic dialects are used in the everyday matter communication. The Oranee dialect is the most important and used one in the west of Algeria. However, it is an under-resourced language.

First, we contribute in this paper to present the most important particularities of this western Algerian dialect and propose a natural language processing on an Oranee textual corpus. We have to augment the Dialectal textual Oranee, which allows us to work more deeply on its morphological and syntactic structure. We also expect also to explore other ways based on neural and deep neural network to model the dialect Corpus.

Second, an MSA transcribed discourse could contain some dialect vocabulary and viceversa. So, we propose to interpolate the dialectal language model with some MSA language models. The MSA LM that matches better with the dialectal LM is the LM based Religion Data. We relied this result to the fact that

some Dialectal Speakers could use Arabic Vocabulary in different situations, for example if the discussion topic is about religion. We have to explore this idea and other ways to improve the MSA-Dialectal language models with more important textual datasets. These LMs components can be used to enhance Arabic Algerian Automatic Speech Recognition, in future works.

Acknowledgment

We would like to thank Pr. Kamel Smaili (INRIA), Dr. Salima Harrat (ENSB), Mrs Samia Ali Larbi (USTO-MB) and Mr. Boumazza Samir for their help.

References

1. Habash, N.Y.: Introduction to arabic natural language processing. Synthesis Lectures on Human Language Technologies 3, pp.1–187 (2010)
2. Biadsy, F., Hirschberg, J., Habash, N. : Spoken Arabic dialect identification using phonotactic modelling. In: the eacl 2009 workshop on computational approache to semitic languages. Association for Computational Linguistics, 2009, pp. 53–61 (2009)
3. Shoufan, A., Alameri, S.: Natural language processing for dialectical Arabic: A Survey. In Proceedings of the Second Workshop on Arabic Natural Language Processing, pp. 36–48 (2015)
4. Harrat, S., Meftouh, K., Smaili, K.: Maghrebi Arabic dialect processing: an overview. Journal of International Science and General Applications (ISGA), (2018)
5. Guellil, I., Saâdane, H. et al.: Arabic natural language processing: An overview, Journal of King Saud University Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.2019.02.006, (2019)
6. Zaghouani, W.: Critical survey of the freely available Arabic corpora. arXiv preprint arXiv:1702.07835, (2017)
7. Ibn Khaldun, A.:The Muqaddimah. Translated by Franz Rosenthal. http://www.muslimphilosophy.com/ik/Muqaddimah/. Last accessed 20 jan 2020
8. Miller, C., Caubet, D.: Arabic sociolinguistics in the Middle East and North Africa (MENA). MJ Ball (Ed.), The Routledge Handbook of Sociolinguistics Around the World, pp. 238–256. (2009)
9. Labed, Z. : Genealogical koineisation in Oran speech community: the case of young university oranees. Phd Thesis, University of Oran, (2014)
10. Ghouthi Zazoua, F.: Sociophonetic Variation among Tlemcen and Oran Speakers. PhD thesis.Tlemcen University, Algeria. (2016)
11. Droua-Hamdani, G., Selouani S.A. and Boudraa, M.: Algerian Arabic Speech Database (ALGASD): Corpus Design and Automatic Speech Recognition Application. In The Arabian Journal for Science and Engineering, **35**(2C), pp.157–166, (2010)
12. Droua-Hamdani, G., Alotaibi, Y. A., Selouani S.A. and Boudraa, M.: Rhythmic Feature across Modern Standard Arabic and Arabic Dialects. In Proceedings of Workshop on free/Open Source Arabic corpora and corpora processing tools, pp.43–46, (2014)

13. Lachachi, N., Adla A.: GMM-Based Maghreb Dialect Identification System. Journal of Information Processing Systems (JIPS), **11**.(1): pp. 22–38.(2015). DOI: https://doi.org/10.3745/JIPS.02.0015
14. Benali, I.: The identification of two Algerian Arabic dialects by prosodic focus. In Proceedings of 7th ExLing 16 pp. 37–40, Saint Petersburg, Russia. (2016)
15. Meftouh, N., Bouchemal, S., Smaili, K.: A study of a non-resourced language: an algerian dialect. 3rd workshop on spoken language technologies of under-resourced languages. Cape Town, South Africa. (2012)
16. Harrat, S., Meftouh, K., Abbes, M., Smaili, K.: Building resourced for Algerian Arabic dialects. In proceedings of annual conference of the international communication association (interspeech), Singapore. (2014)
17. Harrat, S., Meftouh, K., Abbas, M., Hidouci, K. W., Smaili, K.: An Algerian dialect: Study and Resources. In International Journal of Advanced Computer Science and Applications, pp. 384–395, (2016)
18. Harrat, S., Meftouh, M., Smaili, K. : Creating parallel Arabic Dialect Corpus: pitfalls to avoid. In proceedings of international conference on computational Linguistics and intelligent text processing. Budapest, Hungary, (2017)
19. Meftouh, K., Harrat, S., Jamoussi, S., Abbes, M., Smaili, K. Machine translation experiments on PADIC: a parallel Arabic dialectl corpus. In proceedings of 29th pacific Asia conference on language, information and computation, Shanghai, China, (2015)
20. Abidi, K., Menacer, M. A., Smaili, K.: CALYOU: a comparable spoken Algerian corpus harvested from Youtube. In proceedings of 18th annual conference of the international communication association (interspeech). (2017)
21. Bougrine, S., Cherroun, H., Ziadi, D. Lakhdari, A., and Chorana, A. (2016). Toward a rich Arabic Speech Parallel Corpus for Algerian sub-Dialects. In proceedings of the 2nd Workshop on Arabic Corpora and Processing Tools. Theme: Social Media, pp. 2–10, (2016)
22. Bougrine, S., Chorana, A., Lakhdari, A., and Cherroun, H. Toward a Web-based Speech Corpus for Algerian Arabic Dialectal Varieties. In proceedings of the 3rd Arabic Natural Language Processing. Workshop WANLP, Spain, pp. 138–146, (2017)
23. Djellab, M., Amrouche, A., Bouridane, A., Mehallegue, N.: Algerian Modern Colloquial Arabic Speech Corpus (AMCASC): regional accents recognition within complex socio-linguistic environments. In Language Resources and Evaluation, **51** (3), pp. 613–641, (2017)
24. Habash, N., Soudi, A., Buckwalter, T.: On arabic transliteration. In: Arabic Computational Morphology. Springer, pp. 15–22. (2007)
25. Goodman, J.T.:A bit of progress in language modeling. In Computer Speech & Language, **15** (4), pp.403–434. (2001)
26. Mezzoudj, F., Benyettou, A. An empirical study of statistical language models: n-gram language models vs. neural network language models. In International Journal of Innovative Computing and Applications, **9**(4), pp. 189–202, (2018)
27. Mezzoudj, F., Langlois, D., Jouvet, D., Benyettou, A. : Textual data selection for language modelling in the scope of automatic speech recognition. Procedia Computer Science, pp 55–64, (2018)
28. Stolcke A. : SRILM-an extensible language modelling toolkit. In InterSpeech, (2002)
29. Stolcke, J. Zheng, W. Wang, and V. Abrash. Srilm at sixteen: Update and outlook. In Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop, pp. 5,(2011)

30. Ahmed Abdelali, A., Darwish, K., Durrani, N., Mubarak, H. Farasa: A Fast and Furious Segmenter for Arabic. NAACL, (2016)
31. Helmy, M., Basaldella, M., Maddalena, E., Mizzaro, S., Demartini, G.: Towards building a standard dataset for arabic keyphrase extraction evaluation. In 2016 International Conference on Asian Language Processing (IALP), pp. 26–29, (2016)