A Conventional Orthography for Algerian Arabic

Houda Saadane[1] and **Nizar Habash**[2]

(1) Univ. Grenoble Alpes, LIDILEM, Grenoble, France
GEOLSemantics & Consulting, Paris, France
houda.saadane@e.u-grenoble3.fr
(2) New York University Abu Dhabi, United Arab Emirates
nizar.habash@nyu.edu

Abstract

Algerian Arabic is an Arabic dialect spoken in Algeria characterized by the absence of writing resources and standardization, hence it is considered as an under-resourced language. It differs from Modern Standard Arabic on all levels of linguistic representation, from phonology and morphology to lexicon and syntax. In this paper, we present a conventional orthography for Algerian Arabic, following a previous effort on developing a conventional orthography for Dialectal Arabic (or CODA), demonstrated for Egyptian and Tunisian Arabic. We explain the design principles of Algerian CODA and provide a detailed description of its guidelines.

1 Introduction

The Arabic language today is characterized by a complex state of polyglossia. Modern Standard Arabic (MSA) is the official variety of Arabic used primarily in written literal contexts. There is also a large number of dialects whose dominant features are noticeable to Arab-speaking people. The Arabic dialects differ from Modern Standard Arabic (MSA) on all levels of linguistic representation, from phonology and morphology to lexicon and syntax. MSA is classified as a high variety as is contains lot of normalization and standardization. It is generally considered as a prestigious, valued and official language; hence it is used for training (media and education). Arabic Dialects (DA) are considered a low variety which includes languages with less normalization and standardization. These languages are used in daily life, interviews and for informal conversations. Algerian Arabic (henceforth, ALG) is one of the Western group of Arabic dialects spoken in Algeria. ALG differs from other Arabic dialects, neighboring or far ones by having some specific features. In addition to MSA and DA, foreign languages, particularly French and English have been increasingly part of the Arabic spoken in daily basis.

With the emergence of Internet and social media, ALG (and other DAs) have become the language of informal online communication, for instance emails, blogs, discussion forums, SMS, etc. Most Arabic natural language processing (NLP) tools and resources were developed specially to treat MSA. Corresponding tools processing ALG are not as mature and sophisticated as those for MSA. This is due to the recent involvement of works on ALG dialect and the limit quantity of results and resources generated till today. To address this problem, some solutions propose to apply NLP tools designed for MSA directly to ALG. This proposition is interesting but yields to significantly low performance. This is why it is necessary to develop solutions and build resources for ALG treatment.

In this paper, we present a basic layout of ALG processing which is necessary to build efficient NLP tools and applications. This layout is a design of standard common convention orthography dedicated to ALG dialect. The proposed standard is an extension of that proposed in the work of Habash et al, (2012a) who proposed a Conventional Orthography for Dialectal Arabic (CODA). CODA is designed in order to develop computational models of Arabic dialects and provided a detailed description of its guidelines as applied to Egyptian Arabic (EGY).

In this paper, we present a conventional orthography for Algerian Arabic. The paper is organized as follows. Section 2, discusses related works. In Section 3, we present an historical overview of ALG. In Section 4, we highlight the

Proceedings of the Second Workshop on Arabic Natural Language Processing, pages 69–79,
Beijing, China, July 26-31, 2015.

linguistic differences between ALG and the languages MSA, EGY and TUN in order to motivate some of our ALG CODA decisions. In Section 5, we present ALG CODA guidelines.

2 Related works

Studying and processing dialects is an interesting recent research area which took progressively a big attention, especially with the explosion of internet public communications. Hence, there is actually a big interest to develop new tools to process an exploit the huge quantities of resources established using dialects (oral communications, web, social networks, etc.). However, Arabic dialects are languages without standardization or normalization, these why much efforts are necessary to modernize Arabic orthography and develop orthographies for Arabic dialects.

Maamouri et al. (2004) have developed a set of rules for Levantine dialects. These rules define the conversational Levantine Arabic transcription guidelines and annotation conventions. Habsh et al.(2012a) have proposed a conventional orthography for Egyptian dialectal (CODA). This work is inspired by the Linguistic Data Consortium (LDC) guidelines for transcribing. However, CODA is intended for general purpose writing allowing many abstracts from these variations, whereas the LDC guideline are dedicated for transcription, and thus focus more on phonological variations in sub-dialects. A proposition for transcription Algerain dialect are developed in (Harrat et., 2014) where a set of rules for transcription Algerain dialect are defined and a grapheme-to-phoneme converter for this dialect was presented. Grapheme-to-Phoneme (G2P) conversion or phonetic transcription is the process which converts a written form of a word to its pronunciation form; hence this technique focuses only on phonological variations.

To remedy the lack of building resources and tools dedicated to the treatment of ALG issue, (Harrat et al., 2014) built parallel corpora for Algerian dialects, because their ultimate purpose is to achieve a Machine Translation (MT) for Modern Standard Arabic (MSA) and Algerian dialects (AD), in both directions. They also propose language tools to process these dialects. First, they developed a morphological analysis model of dialects by adapting BAMA, a well-known MSA analyzer. Then they propose a diacritization system, based on a MT process which allows restoring the vowels to dialects corpora. And finally, they propose results on machine translation between MSA and Algerian dialects.

In the same way, (Harrat et al., 2015) present an Arabic multi-dialect study including dialects from both the Maghreb and the Middle-east that they compare to the Modern Standard Arabic (MSA). Three dialects from Maghreb are concerned by this study: two from Algeria : Annaba's dialect (ANB), the language spoken in the east of Algeria, on Algiers's dialect (ALG), the language used in the capital of Algeria, and one from Tunisia, on Sfax's dialect (TUN) spoken in the south of Tunisia and two dialects from Middle-east (Syria and Palestine). The resources which have been built from scratch have lead to a collection of a multi-dialect parallel resource.

Furthermore, (Zribi, et al., 2014) extend the CODA guidelines to take into account to Tunisian dialect and (Jarrar, et al., 2014) have adapted it to the Palestinian dialect. In addition, authors of Egyptian and Tunisian CODA encourage the adaptation of CODA to other Arabic dialects in order to create linguistic resources. Following this council, we extend in this paper CODA guidelines to ALG.

3 Algerian Arabic: Historical Overview

Arabic speakers have Arabic dialects or vernacular as their mother tongues. These dialects can be stratified in two big families of dialects: the Western group (the Maghreb) or North African group and the Eastern group (the Mashriq). Algerian dialect, noted ALG, is one of the Western group which is spoken in Algeria. This dialect is also called دارجة *daArjaħ*[1] or جزايري *jazaAyriy* or دزيري *dziyriy* simply meaning "Algerian". These variations do not create generally barriers to understand the dialect. In addition to ALG, the Algerian's population speaks also Berber but with different ratios: ALG is used by 70 to 80% of the population however; the Berber language is the mother tongue of 25% to 30% of population. Berber is used mainly in center of Algeria (Algiers and Kabylie), East of Algeria (Béjaia and Sétif), in Aures (chaoui), the Mzab (north of the

[1] Arabic transliteration is presented in the Habash-Soudi-Buckwalter (HSB) scheme (Habash et al., 2007). Phonological transcriptions will be presented between /…/ but we will use the HSB consonant forms when possible to minimize confusion from different symbol sets.

Sahara) and it is used by the Twaregue based in south of the Sahara (Hoggar mountains). Even if ALG is spoken by Algeria's population, estimated to 40 million of persons, it is characterized by variation of this same dialect according to geographic location of ALG's speakers.

This dialect cannot be presented as homogeneous linguistic system but it has many varieties. According to (Derradji et al., 2002) we distinguish four varieties for ALG as follow: I) the Oranais: is the variety spoken in the Western of Algeria, precisely from Moroccan frontiers to the limit of Ténès, ii) Algérois: this variety covers the central zones of Algeria to Béjaia and it is widely spread, iii) Rural: the speakers of this variety are located in the East of Algeria like Constantine, Annaba or Sétif, and iv) Sahara: is the dialect of the south of Algeria population. ALG is also the language used in press, television, social communication, internet exchanges, SMS, etc. Only in official communications, both reading and writing ones, where ALG is not used.

Furthermore, we note that ALG is enriched by the languages of the groups colonized or managed the Algerian population during the history of the country. Among these group's languages we can cite: Turkish, Spanish, Italian and more recently French. This enrichment, materialized by the presence of foreign words in the dialect, has contributed to create many varieties of ALG from one region to another one, with a quite complex linguistic situation resulting from this language mixture. Indeed, this language mixture has been studied by many socio-linguistic like (Morsly, 1986; Ibrahimi, 1997; Benrabah, 1999; Arezki, 2008). They described the linguistic landscape of Algeria as '*multilingualism*' or "poly-glossic" where multiple languages and language varieties coexist. In other words, the ALG is a suitable example of a complex sociolinguistic situation (Morsly, 1986).

Historically, Berber was the native language of the population of the Maghreb in general and Algeria in particular before the Islamic conquest, which introduced Arabic in all aspects of life. Centuries of various foreign powers introduced vocabulary from Turkish, Spanish and finally (and most dominantly today) French. French colonization tried to impose the French language as the only way of communication during its 132 year control of Algeria. This situation caused a significant decline in the Arabic language, characterized by increased French influence and the introduction of some other languages like Italian and Spanish due to migratory flow from Europe (Ibrahimi, 2006). The influence of these languages on ALG realizes in frequent code-switching without any phonology adaptation in daily conversations, particularly from French, e.g., "*lycée*", "salon", "*quartier*", "*normal*", etc.

4 Comparison among Algerian, Egyptian, Tunisian and Standard Arabic

There are many differences among ALG, EGY, TUN and MSA regarding many levels: phonological, morphological and orthographic. In this section we present some of these differences that are important and determinant of the distinction between these Arabic flavors. We refer the reader to (Habash, 2010) for further elements and discussions.

4.1 Phonological Variations

We give in the following list the major phonological differences between ALG and both MSA and EGY:

The consonant equivalent the MSA (ق) /q/ is one of the sounds that deserve special attention. This sound has many varieties of pronunciation in Algerian Arabic dialects that we can find in the different regions, cities and localities of Algeria. Hence, the pronunciation of "q" can be realized as *q*, *g*, *ʔ*, or *k*.

- *uvular stop* "ق" [q]: like Moroccan and Tunisian dialects, this pronunciation is present in ALG in different localities as in some urban cities like Algiers or Constantine.
- *palatal sound* "ڤ" [g]: this sound is also used in both Moroccan and Tunisian dialects in addition to the ALG one. In Algeria, this sound is used in some cities like Annaba and Sétif, in addition to the Bedouin dialects where this sound is widely employed.
- *glottal stop* [ʔ]: this sound is used in Tlemcen city in the same manner we find it in the Egyptian dialect.
- *k postpalatal:* this sound is a particularity of the ALG dialect that we do not find it in the other north African dialects. This sound is used in the rural localities and some cities like Kabylia, Jijel, Msirda and Trara.

We note that in the case of dialects not using glottal stop consonant, there are some exceptions where the pronunciation is the same way regardless of the dialect. This is the case of the word

بقرة *bagraħ*‘ cow’ which is pronounced in the same way using the palatal sound *bagra*.

The pronunciation of the consonant (ج) /j/ has also different from specific for a location or a group of speakers in the north of Africa. It is pronounced [dj] in Algiers and most of central Algeria as in the word نجاح *ndjaH* ‘success’, but when the consonant (ج) /j/ precedes a (د) /d/ consonant it will be pronounced with the allophone [j] like in the word جديد *jdid* ‘new’. In Egypt this consonant is pronounced as /g/. For Tunisian, Tlemcenian and east Algerian speakers, 'ج' is realized as /j/ or /z/ when the word contains the consonant (س) /s/ or (ز) /z/ like in the words جبس *ğibs* or *djibs* ‘plaster’ become زبس *zebs*; and عجوز *ϛadjuwz* ‘old women’ become عزوز ϛzuwz.

The MSA consonant (غ) /ɣ/ is assimilated in different manner according to some categories of speakers. In the eastern Algerian Sahara, like M'sila and BouSaâda, /ɣ/ is assimilated to (ق) /*q*/, for instance, the words غالي*ɣaAliy* ‘expensive’ and صغيرة *sɣayraħ* ‘small’, are pronounced respectively /*qaAliy*/, and /*sqayra*/. Sometimes, it is assimilated to (خ) /*x*/, like Tunisian and eastern Algeria speakers, e.g., the word غسل‘washed’ is pronounced /xssel/ or /ɣssel/.

The interdental MSA consonant (ث) /θ/ can be pronounced as (ت) /t/, in both ALG and EGY dialects like for the word ' *uwmθ* ثومgarlic’ is pronounced as توم /tuwm/. But it is also pronounced /θ/ in some urban Algerian dialects as in the word ثوم *uwmθ*, (ف) /f/ like in nomadic dialects of Mostaganem where for instance the word ثاني *θaAniy* ‘also’ is pronounced فاني *faAniy*; or (س) /s/ in some cases in EGY dialect, for example, the word ثابت *θaAbit ‘fixe’* is pronounced سابت *saabit.* Another MSA interdental consonant has also special pronunciations; it is the consonant (ذ) /ð/. In the EGY dialect, it can be pronounced (د) /*d*/, like the word ذهب *ðhab* ‘gold’ pronounced دهب *dhab*, or (ز) /z/ for instance the word ذكي ‘clever’ is realized *zakiy*. However, in the ALG dialect, the consonant (ذ) /ð/ has one of the following pronunciations: (ذ) /ð/ or (د) /d/. For instance the word ذراع ‘arm’ can be pronounced *ðraAç* or *draAç*. Moreover, in some regions in Algeria, like Mostaganem, this consonant is realized as (ب) /*v*/, like for the word ذهب *ðhab*' gold’ pronounced فهب *vhab*.

The pronunciation of the glottal stop phoneme that appears in many MSA words in ALG dialect has different forms:

- *The glottal stop becomes longue:* this pronunciation is also present in TUN and EGY dialects. We can give as example the words : فأس *faÂs* /fa’s/ → /fa:s/ فاس *faAs* ‘pickaxe', ذئب *Diŷb* /Di’b/ → /Di:b/ ذيب *diyb* ‘wolf’, and مؤمن *muŵmin* /mu’men/ → /mumin/ مومن *muwmin* ‘beliver’.
- *The glottal stop disappears:* it consists on simply removing the glottal when pronouncing the word. This form is also used in TUN and EGY dialects. For instance, let us take the following word: زرقاء *zarqaA’* /zarqa:’/ → /zarqa:/ زرقا *zarqA* ‘blue’.
- *The glottal stop is replaced by a semi-vowel /w/ or /y/:* this pronunciation is found in ALG and TUN dialects and not in EGY one. It is used for instance in the case of the wordsأَكَّلَ /*Âak~al* /‘to give eating’ → وكّل *wuk~al*, أمس /*Âa*ms/ ‘yesterday’ → يامس *yaAmas*
- *The glottal stop is replaced by the letter /l/:* This form is also used uniquely in the ALG and TUN dialects unlike the EGY one. Let us take the following examples of using of this form: أفعى */Âafça/ ‘snake’* → لفعى /*lafça*/, أرض /*ÂaarD*/ ‘earth’ → لرض /*larD*/. We note that the given examples are also exceptions where we use the same form for both definite and indefinite.
- *The glottal stop is replaced by the letter /h/:* opposite to the EGY dialect, the ALG and TUN ones use this form to pronounce in some cases the glottal stop, like in the words أَجَّالة *Âaj~aAlaħ /Âajja:la/* ‘widow’ → هجَّالة *hajjaAlaħ* /hajja:la/,أمّالا *Âam~aAlaA* /*Âamma:laA*/ ‘however’ → همّالا *ham~aAlaA* /hamma:laA/.

Unlike the Egyptian dialect, the Algerian dialect elides many short vowels in unstressed contexts. This feature characterizes also the other Maghreb dialects. This is the case of the following words: MSA جمل *jamal* ‘Camel’ (and EGY /*gamal*/) becomes ALG /*jmal*/. In addition, this feature introduce an interesting element to distinguish the Maghreb dialects from the EGY one, this element is the presence of a succession of two consonants at the beginning of the word which introduces a specific particularity in the verb scheme ‘*fçal’* in ALG instead of ‘*façal’* in EGY, like in the verb MSA قتل /*qatal*/ ‘he killed’ (and EGY /*’atal*/) becomes ALG /*qtal*/.

The MSA diphthongs *ay* and *aw* are generally reduced uniformly to /i:/ and /u:/. For example, let us take the words: حيط /*HayT*/ 'wall' becomes ALG /*Hi:T*/, لون /*lawn*/ 'color' becomes ALG /lu:n/. We note that this particularity is found in the younger generation speakers; however, older speakers still retain them in some words and contexts, for instance the word عود stills pronounced /ʕawd/ 'horse' by some old speakers.

Another feature of ALG dialect, shared with the TUN one, is the pronunciation of the MSA /a:/: in some words it is realized as /e:/ and in others remains /a:/. For example, the word جَمَالْ /jam:al/ 'beauty' with this signification is pronounced with /a:/ but it is realized with /e:/ in the word جْمَالْ /jme:l/ meaning 'camels'.

4.2 Morphological Variations

ALG dialect has also some morphological aspects that are different from that of the MSA, and closer to that of Maghreb dialects. These aspects consist essentially on a simplification of some inflexions and inclusion of new clitics as follow:

As regards the inflexion, in ALG dialect, like other Arabic ones, the casual endings in nouns and verbs mood are lost. We note that the indicative mood is the one which is used as default unlike the other moods that are not used. Moreover, the dual and the feminine plural disappeared; they are assimilated to the masculine in the plural form. For example, the word شَكَرْتُنَّ *šakartun~a* 'they (fem.pl.) thank' is normalized in the ALG dialect in شْكَرْتُوا *škar-tuwA* 'they thank'. In addition, the first and the second person of the singular form are conjugated in the same way in the dialect, e.g., in MSA we say شَكَرْتُ *šakartu* 'I thank' and شَكَرْتَ *šakarta* 'you thanks', these two forms are normalized in ALG dialect in the following unique form: شْكَرْتْ *škart* 'I/you thank'. This simplification can lead to some ambiguities in ALG.

The ALG dialect modifies the interne form of the verbs when it does their flexion in imperfective form. It introduces a gemination in the first radical letter and moving to this radical the vowel of the second one. This modification is applied only in the plural form and the 2nd person of feminine singular. For example, in ALG the verb 'to thank' in 3rd person of masculine singular is يُشْكُرْ *yu-škur* (he is thanking) and in 3rd person of masculine plural we have: يُشُّكْرُوا *yuš~ukr-uwA* (they are thanking) but in EGY the same case have the form: يُشْكُرُوا *yuškur-uwA*. To enforce this statement we refer to (Souag, 2005) work where they defend that: "*As is common in Algeria, when normal short vowel elision would lead to another short vowel being in an open syllable, we have slight lengthening on the first member so as to change the stress:* يضرب *yaDṛab 'he hits'* → *yaD~arbuwA* يضربوا *"they hit"*, ركبة *rukba 'knee'* → ركبتي *ruk~ubtiy 'my knee'*; this gemination need not occur, however, if the consonant to be geminated is one of the sonorants r, ṛ, l, n, although for younger speakers it often does. I have the impression that these compensatory geminates are not held as long as normal geminates; this needs further investigation."

Otherwise, ALG dialect uses, like the other Arabic dialects, only the suffix ين /yn/ to form the regular plural. However, the ALG elides the short vowels in plural forms like in the following examples: مُلْحَدْ *mulHad* 'unbeliever', in the plural form مُلْحْدِين *mulHdiyn*, مُهَنْدِسْ *muhandis* 'engineer', pl. مُهَنْدْسِينْ *muhandsiyn*. But in some dialects, like the EGY one, they don't elide the short vowel, for instance the plural of مُهَنْدِسْ *muhandis* 'engineer' in EGY is مُهَنْدِسِينْ *muhandisiyn*. But for some exception, like for the active participle [1A2i3] → [1A23-iyn] (Gadalla, 2000), this elision is maintained whatever the dialect like for the word صَايِمْ *SaAyim* 'fasting' → صَايْمِينْ *SaAymiyn*.

Cohen (1912) describes the emphatic suffix تيك /–tiyk/ as a characteristics of the Muslim Algiers dialect that is used to express adverbs ending with –a like in for the words قانا *gana* 'also' which becomes *ganaAtiyk*, زعما *zaçma* 'supposedly' which becomes *zaçmaAtiyk*.

For the form استفعل [Aista12a3] which exists in the different dialects, the ALG introduces in addition a new variant of this form. This variant is سفعل [*ssa-12a3*] and it is used essentially by the speakers of the west of Algeria (Marçais, 1902). For example, let us take the verb اِسْتَكْلَفْ *Aistaklaf* 'take care of' can be also used like سَّكْلَفْ *ssaklaf* or سَكْلَفْ *saklaf*.

Another feature of the ALG dialect is the insertion of vowel /i:/ between the stem and the consonantal suffixes of the perfect form of the primary geminate verb, e.g in MSA the verb شدّ/شددت *šad~a/šadadtu* 'he/I pulled' becomes in ALG شدّ/شديت *šad~/šad~iyt*. This feature is also present in the other Arabic dialects.

The passive voice in classical Arabic uses vowel changes and not verb derivation but in ALG as in many Arabic dialects, the passive form is obtained by prefixing the verb with one the following elements:

- t- / tt-, for example : تبنى *tabnaý* 'it was built', ترفد *ttarfad* 'it was lifted'
- n-, for instance : نفتح *nftah* 'it opened'
- /tn- / or /nt/, e.g., نتكل *ntkal* 'was edible', تنقتل *tnaqtal* 'to be killed'. We note that this last element is specific for the ALG dialect.

The ALG dialect uses the particle «n» for the first person of singular like the other Maghreb dialects. This particle is generally absent from the Mashreq dialects like EGY one. In those dialects the «n» is substituted by the «a» like shown in the following example: نكتب /naktab/ 'I write' in ALG while the equivalent of it in EGY is اكتب /*Aaktib*/.

Like several dialects (EGY and TUN), ALG include the clictics, that are reduced forms of the MSA words, e.g., the demonstrative proclitic +ه *ha+* which strictly precedes with the definite article +ال *Al+* is related to the MSA demonstrative pronouns هذا *haðaA* and هذه *haðihi*, e.g.; (MSA → ALG) هذه الدنيا *haðihi AldunyaA* → *haAldinyaA* 'this life'.

Several dialects include the proclitic +ع, *ç*a+ a reduced form of the preposition على/*çalaý*/ 'on/upon/about/to'. For example, (MSA → ALG) على الطاولة/*çalaý AlTaAwilaħ*/ → عالمايدة *çaAlmaAydaħ* 'on the table'. The same interpretation is valid for the proclitics +ف *+f*a and +م *m+*; which are the reduced form of the prepositions في *fiy* 'in' and من *min* 'from' respectively. Also, several dialect include the non-MSA negation circum-clitic+ما *mA+* +ش *+š*. For example ما قريتش *mA qriyteš* 'I haven't read'.

Furthermore, ALG almost lost all of the nominal dual forms, which are replaced with the word **زوج** *zudwj* /zu:dj/ 'two' with the plural form, e.g., (MSA→ ALG) كتابين *kitaAbayn* → زوج كتب *zuwdj ktub* 'two books'

4.3 Orthographic Variations

The orthographic variation in writing of Arabic dialects words is due to two reasons: i) the non-existence of an orthographic standard for Arabic dialects because these varieties are not codified and normalized, and ii) the phonological differences between MSA and Algerian dialect (ALG). For these dialects words can be spelled phonologically or etymologically using their corresponding MSA form. This fact creates some inconsistency among dialect writers. For example, the corresponding word to 'gold' can be written دهب *dhab* or ذهب *ðhab*. In addition, in some cases the phonology or underlying morphology is reflected by some regular phonological assimilation writing, e.g. طوموبيل *Tuwmuwbiyl* 'cars' is also written as طونوبيل *Tuwnuwbiyl*, إسماعيل *AismaAçiyl*, 'Ismaël' is also written as إسماعين *AismaAçiyn*, من بعد *min baçd* 'after' is also written as مم بعد *mim baçd*. Furthermore, these different spelling can conduce to some semantic confusion, like for شربو *šrbw* may be شربوا *šarbuwA* 'they drank' or شربه *šarbuh* 'he drank it'. Finally, the shortened long vowels, can be spelled long or short, for instance, شافوها/شفوها *šAfw+hA/ šfw+hA* 'they saw her, and مجابش *majaAbaš* 'he didn't bring' ماجابش *mAjaAbaš*.

4.4 Lexical Variations

As presented in Section 3, the Algerian dialect, like other Arabic dialects, has been influenced, over centuries, by other languages like Berber, Turkish, Italian, Spanish and French. Table 1 shows some examples of borrowed words[2] in ALG.

Words	Translation	Transliteration	Origin
فكرون	a tortoise	*Fakruwn*	Berber
شلاغم	Moustache	*šliAγam*	
قرجومة	a throat	*Qarjuwmaħ*	
تقاشير	Socks	*tqaAšiyr*	Turkish
سكارجي	a drunkard	*sukaArjiy*	
زردة	Feast	*Zardaħ*	
فيشطة	Party	*fiyšTaħ*	Italian
زبلة	Foul	*Zablaħ*	
صوردي	Money	*Suwrdiy*	
سيمانة	a week	*siymaAnaħ*	Spanish
سبردينة	Snickers	*Spardiynaħ*	
سُكويلة	a school	*Sukwiylaħ*	
طابلة	Table	*TaAblaħ*	French
تيليفون	Phone	*Tiyliyfuwn*	
فرملي	Nurse	*Farmliy*	

Table 1: The origin and the meaning of some borrowed words used in ALG.

5 Algerian Arabic CODA Guidelines

In this section we present a mapping of the CODA convention for the Algerian dialect. The CODA convention is presented and its goals and

[2] We refer to (Guella, 2011) for more examples.

principals are described in details in (Habash et al, 2012a). An example of Algerian CODA is presented in Table 5.

5.1 CODA Guiding Principles

We summarize the main CODA design elements (Habash et al., 2012a, Eskander et al., 2013):

- CODA is an internally consistent and coherent convention for writing Dialectal Arabic.
- CODA is created for computational purposes.
- CODA uses the Arabic script.
- CODA is intended as a unified framework for writing all Arabic dialects.
- CODA aims to strike an optimal balance between maintaining a level of dialectal uniqueness and establishing conventions based on MSA-DA similarities.

CODA is designed respecting many principles:

1. CODA is an Ad Hoc convention which uses only the Arabic script characters including the diacritics used for writing MSA.
2. CODA is consistent as it associates to each DA word a unique orthographic form that represents its phonology and morphology.
3. CODA uses and extends the basic MSA orthographic decisions (rules, exceptions and ad hoc choices), e.g., using Shadda for phonological gemination or spelling the definite article morphemically.
4. CODA generally preserves the phonological form of dialectal words given the unique phonological rules of each dialect (e.g., vowel shortening), and the limitations of Arabic script (e.g., using a diacritic and a glide consonant to write a long vowel).
5. CODA preserves DA morphology and syntax.
6. CODA is easy to learn and write.
7. The CODA principles are the same for all the dialects, however each dialect will have its proper CODA map. This unique map respects the phonology and the morphology of the considered dialect.
8. CODA is not a purely phonological representation. Text in CODA can be read perfectly in dialect given the specific dialect and its CODA map.

5.2 Algerian CODA

As we said above, CODA principles are applicable for all dialects but with a specific map for each dialect. Hence, in this section we present the map of the Algerian dialect (ALG) to CODA by summarizing the specific CODA guidelines for ALG. Firstly we chose a variant of the ALG which is the one used in the media as default. This variant represents the dialect of the capital city Algiers and follows the same orthographic rules as MSA by taking into accounts all the following exceptions and extensions.

5.3 Phonological Extensions

Long Vowels In ALG CODA the long vowel /e:/, which do not exist in MSA, will be written as *ay* or *iA* depending on its MSA cognate: *ay* or *aA*, respectively. In MSA orthography, the sequence *iA* is not possible, hence using words with *aA* MSA cognates can be a good solution for ALG. This orientation is suitable since the basic non-diacritical form of the word is preserved, for instance, دار *daAr* /da:r/ ‘turn’ and *diAr* /de:r/ ‘do’. This extension is present also in Tunisian CODA unlike the Egyptian one.

Vowel Shortening Like the EGY and TUN CODA, the ALG long vowels are written in long form. In some cases, which are shortened in certain cases such as when adding affixes and clitics even if it is writing long. For example, ماجابهاش *mA jAb+hA+š* ‘he did not forghets for her’ and تقول لهم *tquwl lhm* /tqullhum/ ‘you tell them’ (not تقلهم *tqulhm*). This vowel shorting can be also considered in words with two long vowels. Phonologically, in DA, even if the two long vowels are written, only one is allowed in a word, in other terms, it should be only one stressed syllable in each phonological word. For instance, صايمين *SaAymiyn* ‘fasting’ (not صيمين *Saymiyn*).

5.4 Phono-Lexical Exceptions

The Algerian "qaf" The letter (ق) /q/ is used to represent the four following consonants: /q/, /g/ (like TUN), /k/ and (') (like EGY). The table 2 gives some examples of exceptional pronunciation for /g/.

CODA		Pronunciation	English
بقرة	*baqraħ*	/bagra/	Cow
قاناتيك	*qaAnaAtiyk*	/ga :na :ti :k/	so …
قاوري	*qiAwriy*	/ge:wriy/	foreign

Table 2: ALG exceptional pronunciation examples

Consonant with Multiple Pronunciations
In ALG we use the MSA forms to write consonants with multiple pronunciations. The used MSA form has to be closer to the considerate

consonant if it has a corresponding MSA cognate. We give in Table 3 some examples. Like TUN CODA, the ALG one has more variations than the ones addressed in EGY CODA as for the former the efforts were focused on Cairene Arabic. Hence, ALG seems to have more MSA-like pronunciations where MSA spelling is simply the same as ALG.

Hamza Spelling Hamzated MSA cognate may not be spelled in ALG CODA in a way corresponding to the MSA cognate. In other words, the glottal stop will be spelled phonologically. This feature is also present in EGY and TUN CODA. However, when Hamza is pronounced in ALG, we apply the same MSA spelling rules. Furthermore, the glottal stop phoneme, appearing in many MSA words, has disappeared in ALG, like in the words: فاس *fAs* 'pickaxe' (not like MSA فأس *faÂs*), ذيب *Diyb* 'wolf' (not like MSA ذئب *Diŷr*). In addition, words starting with Hamzated Alif are not seen in ALG CODA, e.g, الارض *AlAarD* /larD/ 'earth' (not لرض *larD*).

CODA		Pronunciations	English
عجوز	*ʒjuwz*	/ʒadju:z/, /ʒzu:z/ /ʒju:z/	old women
ثاني	*θaAniy*	/fa:niy/, /θa:niy/	Also
صدر	*Sadr*	/sadr/, /Sadr/	Chest
قهوة	*qahwaħ*	/qahwa/, /gahwa/, /kahwa/, /'ahwa/	Coffee
غسل	*γsal*	/γsal/,/xsal/	he washed
غالي	*γaliy*	/γa*a:*li/, /qaa:li/	Expensive
فاسدة	*faAsdaħ*	/fa:zda/, /fa:sda/	Corrupt
ذهب	*ðhab*	*/ðhab/, /dhab/ /vhab/*	Gold
هبط	*hbaT*	/hbaT/, /HbaT/	he descended

Table3: examples of multiple pronunciations in ALG.

Definite Article If the word contains the article Al (ال), we must distinguish between the sun and the moon letters. In the case of the sun letters, the "L" is silent and the letter that follows is doubled (gemination) in pronunciation and in writing, e.g., النّهار *AlnnhAr* 'day' (not انّهار *AnnhAr*). Conversely, with the moon letters, the 'A' is not pronounced, the "L" of the article is pronounced and the letter that follows is not doubled, neither in pronunciation nor in writing, e.g., القمر *Alqmar* 'the moon' (not لقمر *lqmar*) (Saadane and Semmar, 2012; Biadsy et al., 2009).

N of Number Construct The ALG CODA adds the phoneme /n/ after some numerals in construct cases, e.g.,سطاشن طابلة *sTaAšn TaAblaħ* '16 tables' whereas the number 16 is pronounced alone سطاش *sTaAš*. This exception is valid for Number Construct forms with number between 11 and 19 preceding a noun in the singular. This property is also valid in TUN CODA.

5.5 Morphological Extensions

Attached clitics ALG dialect, as many other dialects, uses almost all the attached clitics in MSA, the definite article +ال *Al+*, the future particle proclitic +ح *Ha+* (expressed in east of Algeria like Annaba city), the coordinating conjunction و + *w+*, the negation particle enclitic ش+ +*š*. In addition ALG uses the new attached clitics reduced forms of the MSA, e.g., +ع ʒ+, +م *m+*, +ه *h+*, +ف *f+*. The following table illustrates some examples of these clitics where we consider the word وكليناهالكم *wikliynaAhaAlkum* 'and we have eaten your food'

Enclitics			Suffixes	Stem	Proclitics
كم	ل	ها	نا	كلي	و
kum	*l*	*haA*	*naA*	*kliy*	*wi*

Table 4: Tokenization of the word وكليناهالكم *wikliynaAhaAlkum*

Separated Clitics The spelling rule for the indirect object enclitics and the negation proclitic ما mA is preserved in the ALG CODA map. This map puts a separation using a space between the negation particle and the indirect object, e.g., ما جاب لكمش *mA jAb lkumš* /ma+jab+lkum+š/ 'he did not give/com you'.

5.6 Lexical Exceptions

The ALG CODA, like the TUN and EGY ones, contains a list of Algerian dialect words that have a specific ad hoc spelling. This specific spelling may be inconsistent with the map of CODA introduced above and can be spelled commonly in different ways. These exceptions include for instance:

- The demonstratives هذوك *haðuwk* (not هاذوكة *haðukaħ*) 'that', هكذا *hakðaA* 'like this' (not هاكذا *haAkðaA*, or هكدا *hakdaA* or هاكدا *haAkdaA*)
- The preposition 'I know' is expressed with the phrase عْلَى بَالِي ʒlayÝ *baAliy* (not عمبالي ʒambaAliy, or عن بالي ʒan *baAliy*, or علبالي ʒlabaAliy)

Raw Text	مرحبا بكم في بلاتو حصة برنامج الخط لحمر لنهار اليومة والي يتزامن مع عيد المرأة. إنشاء الله قاع النساء الي راهم يشوفو فينا إنشاء الله أيام سعيده وجميلة فحياتهم. إنشاء الله يتهناو ب ماليهم، ب والديهم وولادهم. قبل منروحو للموضوع نتاع اليومة والي خصصناه للمرأة ف الجزاير وكيفاش راهي عايشة خلونا نرحبو بالضيوف تع لبرنامج.
	mrHbA bkm fy plAtw HSћ brnAmj AlxT lHmr lnhAr Alywmћ wlly ytzAmn mç çyd AlmrÂћ. ĂnšA' Allh gAç AlnsA' Aly rAhm yšwfw fynA ĂnšA' Allh ÂyAm sçydh wjmylћ fHyAthm. ĂnšA' Allh ythnAw b mAlyhm, b wAldyhm wwlAdhm. qbl mnrwhw llmwDwç ntAç Alywmћ wAly xSSnAh llmrÂћ f AljzAyr wkyfAš rAhy çAyšћ xlwnA nrhbw bAlDywf tç lbrnAmj.
CODA	مرحبا بكم في بلاتو حصة برنامج الخط الحمر لنهار اليوم واللي يتزامن مع عيد المراة. انشا الله قاع النسا اللي راهم يشوفوا فينا انشا الله ايام سعيدة وجميلة فحياتهم. انشا الله يتهناوا بماليهم، بوالديهم وولادهم. قبل ما نروحوا للموضوع تاع اليوم واللي خصصناه للمراة فالجزاير وكفاش راهي عايشة خلونا نرحبوا بالضيوف تاع البرنامج.
	mrHbA bkm fy blAtw HSћ brnAmj AlxT AlHmr lnhAr Alywm wAlly ytzAmn mç çyd AlmrAћ, AnšA Allh qAç AlnsA Aly rAhm yšwfwA fynA AnšA Allh AyAm sçydћ wjmylћ fHyAthm. AnšA Allh ythnAwA bmAlyhm, bwAldyhm wwlAdhm. qbl mA nrwhwA llmwDwç tAç Alywm wAlly xSSnAh llmrAћ fAljzAyr wkfAš rAhy çAyšћ xlwnA nrhbwA bAlDywf tAç AlbrnAmj.
English	Hello everyone, in « The Red Line » daily show, which coincides with the Women's Day. God willing, for all the women who watch this show, they may have happy and beautiful days in their lives. God willing, and they will rejoice in their families, parents and children. Before addressing the topic of the day, where we focus on women in Algeria and how they are living, let's welcome to our program's guests.

Table 5: An example sentence in ALG

- The adverbs زعمة *zaçmaћ* (not زعما *zaçma*) 'supposedly', ضركة *Durkaћ* (not ضركا *Durka*) 'now', قانة *gaAnaћ* (not قانا *gaAna*) 'also'

In addition, in influence and integration of foreign words from other languages, like French, Berber or Italian, have emerged new phonemes like /g/, /p/ or /v/. These phonemes are used to express sounds that do not exist in MSA, but in CODA we will use the following Arabic characters: /q/, /b/ and /f/ to express respectively g, p and v. For example, جافال *jaAfiAl* 'detergent', كافي kaAvi 'stupid', بوبية *puwpiyaћ* 'doll', قيدون *qiyduwn* 'handlebar'.

6 Conclusions and Future Work

We presented in this paper a set of guidelines towards a conventional orthography for Algerian Arabic. We discussed the various challenges of working with Algerian Arabic and how we address them. In the future, we plan to use the developed guidelines to annotated collections of Algerian Arabic texts, in a first step towards developing resources and tools for Algerian Arabic processing.

Acknowledgment

The first author was supported by the DGCIS (Ministry of Industry) and DGA (Ministry of Defense): RAPID Project 'ORELO', referenced by N°142906001. The second author was supported by DARPA Contract No. HR0011-12-C-0014. Any opinions, findings and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of DARPA. We would like to thank Bilel Gueni, Emad Mohamed and Djamel Belarbi for helpful feedback.

References

Abdenour Arezki. 2008. *Le rôle et la place du français dans le système éducatif algérien*. Revue du Réseau des Observatoires du Français Contemporain en Afrique, (23), 21-31.

Mohamed Benrabah. 1999. *Langue et pouvoir en Algérie: Histoire d'un traumatisme linguistique*. Seguier Editions.

Fadi Biadsy, Nizar Habash and Julia Hirschberg. 2009, *Improving the Arabic Pronunciation Dictionary for Phone and Word Recognition with Linguistically-Based Pronunciation Rules*, The 2009 Annual Conference of the North American Chapter of the ACL, pages 397–405, Boulder, Colorado.

Marcel Cohen. 1912. *Le parler arabe des Juifs d'Alger*. Champion :Paris.

Yacine Derradji, Valéry Debov, Ambroise Queffélec, Dalila S. Dekdouk and Yasmina C. Benchefra. 2002. *Le français en Algérie : lexique et dynamique des langues*, Ed. Duclot, AUF, 2002, 590 p.

Ramy Eskander, Nizar Habash, Owen Rambow and Nadi Tomeh. 2013. Processing Spontaneous Orthography. In Proceedings of Conference of the North American Association for Computational Linguistics (NAACL), Atlanta, Georgia.

Charles A. Ferguson. 1959. *Diglossia*. Word-Journal of the International Linguistic Association, 1959, vol. 15, no 2, p. 325-340.

Hassan A. Gadalla. 2000. *Comparative Morphology of Standard and Egyptian Arabic* (Vol. 5). Lincom Europa.

Noureddine Guella. 2011. *Emprunts lexicaux dans des dialectes arabes algériens*. Synergies Monde arabe, 8, 81-88.

Nizar Habash, Abdelhadi Soudi and Tim Buckwalter. 2007. *On Arabic Transliteration*. Book Chapter. In *Arabic Computational Morphology: Knowledge-based and Empirical Methods*. Editors Antal van den Bosch and Abdelhadi Soudi.

Nizar Habash. 2010. *Introduction to Arabic Natural Language Processing*. Synthesis Lectures on Human Language Technologies, Graeme Hirst, editor. Morgan & Claypool Publishers.

Nizar Habash, Mona Diab and Owen Rambow. 2012a. *Conventional Orthography for Dialectal Arabic*. In: Proceedings of the Language Resources and Evaluation Conference (LREC), Istanbul.

Nizar Habash, Ramy Eskander and Abdelati Hawwari. 2012b. *A Morphological Analyzer for Egyptian Arabic*. In the Proceedings of the Workshop on Computational Research in Phonetics, Phonology, and Morphology (SIGMORPHON) in the North American chapter of the Association for Computational Linguistics (NAACL), Montreal, Canada.

Salima Harrat, Karima Meftouh, Mourad Abbas and Kamel Smaïli. 2014. *Grapheme To Phoneme Conversion-An Arabic Dialect Case*. In Spoken Language Technologies for Under-resourced Languages.

Salima Harrat, Karima Meftouh, Mourad Abbas and Kamel Smaili. 2014. *Building Resources for Algerian Arabic Dialects*. Corpus (sentences), 4000(6415), 2415.

Salima Harrat, Karima Meftouh, Mourad Abbas, Salma Jamoussi, Motaz Saad, and Kamel Smaili. 2015. *Cross-Dialectal Arabic Processing.* In Computational Linguistics and Intelligent Text Processing (pp. 620-632). Springer International Publishing.

Khawla T. Ibrahimi. 1997. *Les Algériens et leur (s) langue (s): éléments pour une approche sociolinguistique de la société algérienne*. Éds. El Hikma.

Khawla T. Ibrahimi, K. 2006. *L'Algérie: coexistence et concurrence des langues*. L'Année du Maghreb, (I), 207-218.

Mustafa Jarrar, Nizar Habash, Diyam Akra and Nasser Zalmout. 2014. *Building a Corpus for Palestinian Arabic: a Preliminary Study*. *ANLP 2014*, 18.

Mohamed Maamouri, Tim Buckwalter and Christopher Cieri. 2004. *Dialectal Arabic telephone speech corpus: Principles, tool design, and transcription conventions*. In NEMLAR In-

ternational Conference on Arabic Language Resources and Tools, Cairo (pp. 22-23).

William Marçais. 1902. *Le dialecte arabe parlé à Tlemcen: grammaire, textes et glossaire* (Vol. 26). E. Leroux.

Philippe Marçais. 1956. *Le parler arabe de Djidjelli: Nord constantinois, Algérie* (Vol. 16). Librairie d'Amérique et d'Orient Adrien-Maisonneuve.

Dalila Morsly. 1986. *Multilingualism in Algeria*. The Fergusonian Impact: In Honor of Charles A. Ferguson on the Occasion of His, 65.

Houda Saadane, Aurélie Rossi, Christian Fluhr and Mathieu Guidère. 2012. *Transcription of Arabic names into Latin*. In Sciences of Electronics, Technologies of Information and Telecommunications (SETIT), 2012 6th International Conference on (pp. 857-866). IEEE.

Houda Saadane and Nasredine Semma. 2013. *Transcription des noms arabes en écriture latine*. Revue RIST| Vol, 20(2), 57.

Lameen Souag. 2005. *Notes on the Algerian Arabic dialect of Dellys*. Estudios de dialectología norteafricana y andalusí, 9, 1-30.

Ines Zribi, Rahma Boujelbane, Abir Masmoudi, Mariem Ellouze, Lamia Belguith, and Nizar Habash. 2014. *A Conventional Orthography for Tunisian Arabic*. In Proceedings of the Language Resources and Evaluation Conference (LREC), Reykjavik, Iceland.