A Conventional Orthography for Maghrebi Arabic

Houcemeddine Turki, *Emad Adel, **Tariq Daouda, ***Nassim Regragui
Faculty of Medicine of Sfax, University of Sfax, Sfax, Tunisia
* Sbikha 1979 High School, Sbikha, Kairouan, Tunisia
** Institut de Recherche en Immunologie et en Cancérologie, Université de Montréal, Montréal, Québec, Canada
*** Copenhagen Business School, Copenhagen, Denmark
E-mail: turkiabdelwaheb@hotmail.fr, geekemad@gmail.com, tariq.daouda@umontreal.ca, nassreg@gmail.com

Abstract

Maghrebi Arabic is the set of dialects of the Arabic language spoken in the Maghreb (Tunisia, Algeria, Morocco, Libya and Mauritania). This set of dialects is under-resourced and has neither a standard orthography nor large collections of written text and dictionaries. Actually, there is no strict separation between Modern Standard Arabic, the official language of the government, media and education, and Maghrebi Arabic; the two exist on a continuum dominated by mixed forms. In this paper, we present a conventional orthography for Maghrebi Arabic, following a previous effort on developing a conventional orthography for Dialectal Arabic (or CODA) demonstrated for Egyptian, Tunisian and Algerian Arabic. We explain the design principles of CODA and provide a detailed description of its guidelines as applied to Maghrebi Arabic.

Keywords: Maghrebi Arabic, Arabic Dialect, Orthography, CODA.

Notes: This paper was not presented in LREC 2016 even if it was accepted in this international conference due to a lack of funding and time. However, we have posted the paper online as we are convinced that the paper can provide significant contributions to the development of the CODA research.

1. Introduction

The Arabic language is currently characterized by a kind of polyglossia. Further than the Modern Standard Arabic that is considered as the official written and spoken variety of Arabic, a considerable number of localized Arabic dialects exist all over the Arab World (Volk, 2015; Watson, 2007). These dialects are mainly used in daily communication and social discussions and differ in their morphology, phonology, semantics and pragmatics from Modern Standard Arabic due to a series of dialect contact and leveling (Watson, 2007; Singer, 1984; Volk, 2015; Souag L. , 2005; Heath, 1997). They are the native, main spoken and most intelligible varieties of Arabic in the countries they are used in (Volk, 2015; Lewis, Simons, & Fennig, 2009; Sayahi, 2014) and that is why it is worthy to study them and promote their use in sensitizing people and education (Maamouri M. , 1973; Maamouri M. , 1977; Ayari, 1996; Maamouri M. , 1983; Maamouri M. , 1998). This would not be possible without creating a Conventional Orthography for these dialects (Habash, Diab, & Rambow, 2012). That is why Habash, Diab and Rambow have defined in 2012 several principles for doing this. These principles involve the use of the rule of orthography of the Modern Standard Arabic and the use of the original words in Standard Arabic to differentiate short vowels in the words in Dialectal Arabic (Habash, Diab, & Rambow, 2012). Unlike other existing orthography, this orthography could be used in Morphological Analysis of the Arabic dialects because it is convertible to the Buckwalter Transliteration (Habash, Diab, & Rambow, 2012; Diab & Habash, 2007).

In this paper, we discuss an important basic technology that is necessary for the efficient development of the various current and future efforts on NLP of Maghrebi Arabic Dialects: the design of orthography to be used as a common standard convention. Our work is a continuation of the work of Habash. We do not expect this convention to be produced by Maghrebi speakers as input, but it is primarily for use in development NLP systems. Spontaneously written Maghrebi Arabic Dialects will have to be converted automatically into its CODA version (Habash, Eskander, & Hawwari, 2012; Eskander, Habash, Rambow, & Tomeh, 2013).

In this paper, we first review some previous related work (Section 1). In Section 2, we present an overview of Maghrebi Arabic. In Section 3, we highlight the linguistic differences between Maghrebi Arabic and both MSA and Egyptian Arabic to motivate some of our Maghrebi Arabic CODA decisions. And in Section 4, we present Maghrebi Arabic CODA guidelines.

2. Related Works

In the 20th century, all the Latin transcriptions for Maghrebi Arabic were phonological. In fact, transcriptions using Deutsche Morgenländische Gesellschaft Umschrift (Singer, 1984; Dallaji-Hichri, 2010) and even using other transcriptions (Jourdan, 1952; Younes & Souissi, 2014) seemed to transcribe Arabic dialects as they are pronounced by people without any consideration to the existence of phonological simplifications like Assimilation within them. These transcriptions did not consequently give a full overview of the morphology of Maghrebi Arabic (Maamouri, Graff, Jin, Cieri, & Buckwalter, 2004). That is why some linguists had the idea of creating a morphology-based orthography for the Arabic dialects that is influenced from Arabic orthography to eliminate the phonological simplifications

existing in the spoken Arabic dialects (Maamouri, Graff, Jin, Cieri, & Buckwalter, 2004).

The idea of inspiring the orthography patterns of the Modern Standard Arabic to create a conventional orthography for an Arabic dialect was created in the first years of the 21st century (Habash, Diab, & Rambow, 2012). In fact, by getting inspired from an original idea of the Tunisian *Taht Essour* team that had appeared in the 1920s and particularly of Ali Douagi (Dhaoudi & Lahmar, 2004, p. 144), the teams of Zawaydeh and Maamouri had developed a set of rules for orthographic transcription and annotation of Levantine dialects that were inspired from the orthography of MSA in order to create a Levantine Arabic corpus respectively in 2003 and 2004 (Zawaydeh, Stallard, & Makhoul, 2003; Maamouri, Graff, Jin, Cieri, & Buckwalter, 2004).

In 2012, Habash and Diab had generalized this work and created CODA that is a generalized set of orthography rules to be used for transcribing any Arabic Dialect (Habash, Diab, & Rambow, 2012).

After its creation in 2012 until 2015, CODA had been used as a reference in 42 different works. 9 of them were about the Maghrebi Arabic dialects (Google Corporation, 2015). In fact, Zribi et al. had created an adaptation of CODA to Tunisian Arabic (ISO 639-3: aeb) in 2014 which takes into consideration several facts of Tunisian Arabic (Zribi, et al., 2014), Tachicart et al. had developed an adaptation of CODA to Moroccan Arabic in 2014 in order to create a translator between Moroccan Arabic (ISO 639-3: ary) and MSA (Tachicart & Bouzoubaa, 2014) and Saadane et al. had created an adaptation of CODA to Algerian Arabic (ISO 639-3: arq) in 2015 (Saadane & Habash, 2015). All of these works had recognized some deficiencies in using CODA for Maghrebi Arabic dialects. However, only some of them had been solved (Zribi, et al., 2014; Tachicart & Bouzoubaa, 2014; Saadane & Habash, 2015).

In this paper, we will continue the previous works about the Conventional Orthography for Dialectal Arabic by extending the CODA guidelines to let it more adapted for Maghrebi Arabic dialects (Reviewing and adjusting the CODA writing principles). We believe that the CODA goals, especially the unified framework for all Arabic dialects, can help in maximizing synergy between and encouraging adaptation from other dialects and Maghrebi ones, when it comes to resource creation either by Maghrebi people (People would like to use a more adapted Arabic Script to practice their native Maghrebi dialects) or within linguistic studies (Scientists would be interested in using the ASCII based Buckwalter transliteration when creating their corpuses as its output would be less voluminous than the one of the one created using Arabic Script).

3. An overview of Maghrebi Arabic

Also known as Darija, Maghrebi Arabic is the family of the Arabic dialects of North Africa (Zaidan & Callison-Burch, 2014; Sayahi, 2014). It is mainly constituted of Tunisian Arabic, Algerian Arabic, Moroccan Arabic, Libyan Arabic, Saharan Arabic and Hassaniya Arabic as well as Judeo-Maghebi Arabic dialects (Lewis, Simons, & Fennig, 2009). Because of the language contact they had experienced from Berber dialects and Romance languages, they are known to have a different phonology and morphology from the other Arabic dialects (Lewis, Simons, & Fennig, 2009; Zaidan & Callison-Burch, 2014; Sayahi, 2014; Tilmatine, 1999).That is why they are not intelligible to the speakers of Eastern Arabic Dialects (Sayahi, 2014) and that is why a conventional orthography that is efficient on Eastern Arabic dialects can be deficient for Maghrebi Arabic dialects.

4. Maghrebi Arabic vs. the Eastern Arabic dialects and MSA

In general, the use of loanwords is more common in the Maghrebi Arabic Dialects than in Eastern Arabic and MSA due to the language contact they had experienced with Berber dialects and several common foreign languages like French, Italian, Turkish and Spanish (Singer, 1984; Sayahi, 2014). That is why Maghrebi Arabic includes several European phonemes further than the main Arabic phonemes (Zribi, et al., 2014; Singer, 1984; Saadane & Habash, 2015; Tachicart & Bouzoubaa, 2014). In fact, Maghrebi Arabic also involves [p], [v], [yː], [œː], [ɔː], [ɛ̃], [ɔ̃], and [ɑ̃] (Singer, 1984; Sayahi, 2014). These phonemes are mainly used for loanwords. The [p] and [v] in this situation should not be confused with [p] and [v] obtained due to assimilation phenomenon (Chekili, 1982).

Furthermore, differently from Eastern Arabic, [g] in Maghrebi Arabic is not always a Nomadic pronunciation of [q] or a substitute for [ʒ] (Singer, 1984; Tachicart & Bouzoubaa, 2014; Zribi, et al., 2014; Saadane & Habash, 2015). It is true that [q] is always pronounced as [g] in rural cities (Zribi, et al., 2014; Baccouche, 1972). However, there are several words in Maghrebi Arabic in which /g/ is pronounced as [g] even in Urban cities (Zribi, et al., 2014; Baccouche, 1972). For example, بقرة is pronounced as [bægræ] in both Urban and rural areas (Baccouche, 1972). Sometimes, the substitution of [g] by [q] can influence the meaning of the given word. In fact, [gruːn] for example means horns and [qruːn] means centuries (Zribi, et al., 2014). That is why [q] and [g] are distinct letters in Maghrebi Arabic even if they are spoken as [g] in Nomadic areas.

Unlike Egyptian Arabic and MSA, Maghrebi Arabic is also characterized by the simplification of several letters when coming in the first position in a word (Zribi, et al., 2014; Heath, 1997). In fact, it is widely known that [ʒ], [θ] and [ʁ] are simplified respectively as [z] or [dʒ], [t] and

[χ]. It is common that ظ and ض are pronounced as [ðˤ] (Singer, 1984; Heath, 1997; Souag L. , 2005). Moreover, All the Maghrebi Arabic dialects excepting Tunisian Arabic are known for the pronunciation of د and ذ as [d], ت and ث as [t] (Singer, 1984; Saadane & Habash, 2015; Tachicart & Bouzoubaa, 2014). The situation becomes more complicated for Judeo-Maghrebi dialects that also substitute Arabic phonemes in certain situations by Hebrew phonemes so that their pronunciation can be simpler for the Jewish communities of Maghreb (Cohen, 1985).

Maghrebi Arabic is known as well by the existence of several emphatic consonants. In fact, the use of [mˤ], [bˤ], [zˤ], [nˤ], [lˤ] and [rˤ] in all Maghrebi Arabic dialects is significantly more important than in MSA and Eastern Arabic (Singer, 1984; Souag L. , 2005; Heath, 1997; Ghazeli, 1981; Gouma, 2013). These minimal pairs influence the pronunciation of long vowels near them just like the pharyngeal, emphatic or uvular letters in Arabic dialects as shown in Table 1 (Singer, 1984; Souag L. , 2005; Heath, 1997; Gouma, 2013).

Phoneme	**Allophones**	
	Near a pharyngeal, emphatic or uvular letter	**In all other situations**
/a./	[ɑː]	[ɛː]
/u./	[oː] ([uː] in Tunisian, Libyan and Judeo-Tunisian)	[uː]
/i./	[eː] ([iː] in Tunisian, Libyan and Judeo-Tunisian)	[iː]

Table 1: The pronunciation of the long vowels in Maghrebi Arabic (Singer, 1984; Souag L. , 2005; Heath, 1997; Abumdas, 1985)

These minimal pairs should be taken into consideration when transcribing texts in the Maghrebi Arabic dialects mainly because these phonemes can change the meaning of a word (Gibson, 2009). For example, [dɛːr] in Tunisian Arabic is to turn and [dɑːrˤ] is house.

It is also common that Maghrebi Arabic tend to not to include nunation (Pronouncing [n] after an indefinite noun) just like other Arabic dialects (Singer, 1984; Souag L. , 2005; Zribi, et al., 2014) and that it simplifies the Short Vowel with glottal stop as a long vowel and the glottal stop after a long vowel or a given consonant as an [h] (Singer, 1984; Souag L. , 2005; Zribi, et al., 2014). In general, the glottal stops are only kept when the word is a result of a code change to MSA (Singer, 1984; Zribi, et al., 2014).

Further than these characteristics, Maghrebi Arabic pronounciation is known for the simplification of sounds. In fact (Singer, 1984; Souag L. , 2005; Maamouri M. , 1967; Gibson, 2009; Heath, 1997; Chekili, 1982; Abumdas, 1985; Cohen & el-Chennafi, 1968),

- If they are in the end of a word, [iː] and [ɪ] are pronounced as [ɪ], [uː] and [u] are pronounced as [u], and [aː], [ɛː], [a] and [æ] are pronounced as [æ]. This is what explains the lack of accuracy of the grammar specification of Tunisian. For example, none of the works had made an interest to explain why the present of /mʃæ/ is /yimʃi/ and the present of /bdæ/ is /yibdæ/...
- Elision: If a word finishes with a vowel and the next word begins with a short vowel, this short vowel and the space between the two words are not pronounced.
- Epenthesis: If a word begins with two successive consonants, an [e] is pronounced in its beginning.
- Assimilation: Two successive consonants are substituted by two other successive consonants that are easier to speak like /ʕh/ that becomes /ħħ/
- Centralization of short vowels: short vowels are mostly centralized as schwa in all Maghrebi dialects expecting the Tunisian, Libyan and Judeo-Tunisian ones. Tunisian, Libyan and Judeo-Tunisian dialects mostly centralize short damma as [o] and short kasra as [e].

4.2 Morphological and Orthographic Variations

There are many morphological and Orthographic differences between Maghrebi Arabic, Eastern dialects and MSA (Zribi, et al., 2014; Souag L. , 2005; Tachicart & Bouzoubaa, 2014). These variations are explained by the fact that the current transcription of Maghrebi Arabic is influenced by the Phonological Characteristics of the variety like the simplification of sounds (Watson, 2007; Volk, 2015; Zribi, et al., 2014; Saadane & Habash, 2015; Tachicart & Bouzoubaa, 2014). The most important variations are the lack of differentiation between the singular third person direct object pronoun and the suffix of the conjugation of the verb in Plural Form in Present and the important use of proclitics. More details about these variations can be found in the work of Zribi et al. about COTA (Zribi, et al., 2014) and in the work of Saadane et al. about COAA (Saadane & Habash, 2015).

4.3 Further Reading

Further information about the particularities of Maghrebi Arabic dialects can be found in (Turki, Zribi, Gibson, & Adel, 2015), in (Souag L. , 2005), in (Souag M. L., 2006), in (Harrell, 2004), in (Cohen & el-Chennafi, 1968), in (Sounkalo, 2008), in (Abumdas, 1985), and in (Owens, 1984).

5. CODA guidelines for Maghrebi Arabic

Our goal in this paper is to present a CODA map for

Maghrebi Arabic. In this section, we summarize the CODA goals and principles and present specific CODA guidelines for Maghrebi Arabic.

5.1 CODA Goals and Principles

According to Ines Zribi (Zribi, et al., 2014),
« CODA is a conventionalized orthography for Arabic dialects (Habash, Diab, & Rambow, 2012). It has five goals.

- CODA is an internally consistent and coherent convention for writing Dialectal Arabic (DA): every word has a single orthographic rendering.
- CODA is designed for NLP processes.
- CODA uses the Arabic script. It does not use Persian script. That is why it can be simply converted to ASCII characters using Buckwalter transliteration and easily processed. Further information about the principle of Buckwalter transliteration can be found in (Habash, Soudi, & Buckwalter, 2007).
- CODA is intended as a unified method 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

The design of the original CODA respects several principles. Firstly, CODA is an ad hoc convention. It uses only the Arabic characters, including the diacritics for writing Arabic dialects. Secondly, CODA is consistent. A unique orthographic form that represents the phonology and morphology for each word is used. CODA uses the MSA orthographic decisions (rules, exceptions and ad hoc choices) and generally preserves the phonological form of dialectal words given the unique phonological rules of each dialect, and the limitations of Arabic script. CODA also preserves dialectal morphology and dialectal syntax. CODA is easily learnable and readable. All Arabic dialects generally share the same CODA principles; each dialect will have its unique CODA map by respecting the phonology and the morphology of each dialect. However, CODA is not a purely phonological representation. Text in CODA can be read perfectly in DA given the specific dialect and its CODA map. In fact, CODA differentiates between letters having the same pronunciation in DA according to their Classical Arabic etymology to avoid significantly speech ambiguity and fights phonological simplification phenomena like Assimilation and restore the letters suffering from such facts to their original Classical Arabic spelling. . »

5.2 CODA General Matters and their solutions

5.2.1. Phonological and Lexical Matters

- The *Ta Marbuta* (p in Buckwalter transliteration) is not always silent. It can be pronounced as [t] when coming before a *Mudhaf ilayh* (Biadsy, Habash, & Hirschberg, 2009). To avoid this matter that can cause a deficiency in the results provided by rule-based text to speech converters for CODA Arabic script inputs in Arabic dialects, we propose to put *Mudhaf ilayh* between [and]. For example, [eʒmɛːʕet elːɑʕruːsæ] (meaning *the family of the bride*) is transcribed as جماعة [العروسة] (jmaAEap [Al_Eruwsa] in Buckwalter transliteration).
- CODA has also failed to decide the transcription of several structures of Arabic dialects that were created through the simplification of Modern Standard Arabic like the Egyptian Arabic [di] (meaning *This (f.)*) that was created through the abbreviation of Classical Arabic هذه ha*ihi. In fact, this example is transcribed as دي diy or ذي *iy. Another example of this controversy is the Egyptian Arabic [ha] proclitics (meaning the modal verb will) that is transcribed as حـ, as ها and as هـ and that was created by the simplification of Classical Arabic راح. In these situations, we propose to adopt the transcription that conserves the most the Modern Standard Arabic phonemes. In fact, [di] is ذي and [ha] is حـ.
- When doing an automated tokenization of CODA based texts and particularly of undiacritized ones, some matters of the recognition of the [el] deteminant can occur (Habash N. Y., 2010; Habash & Rambow, 2005). Furthermore, the [l] of the [el] determinant can have different pronunciations than an ordinary [l] when it comes before a Sun consonant and consequently, a lack of differentiation between both [l] can constitute a deficiency for a text to speech converter for Arabic dialects (Masmoudi, Khmekhem, Estève, Belguith, & Habash, 2014). That is why we advise to add a *Tatweel* (U+0640) after it to solve the problem.

5.2.2. Morphological and Orthographic Matters

- Although the transcription of glots (Hamza) using CODA guidelines follow the same rules for the transcription of Hamza in Modern Standard Arabic, no paper about CODA guidelines has given these rules in details. Citing such rules is important mainly because there are six graphs in CODA Arabic script that are used in transcribing glots (Habash, Diab, & Rambow, 2012; Habash, Soudi, & Buckwalter, 2007). As a solution, we adopt the algorithm that was described in (Younes M. A., 2005) when writing Hamza using CODA guidelines.
- In CODA, there is a lack of differentiation between the singular third person direct object pronoun and the suffix of the verb conjugation in plural form and in Present. CODA had solved the problem when the [u:] particle is in the end of the word by respectively transcribing them as uh and

uwA. However, it did not solve the problem when [u:] is in the middle of the word (Habash, Diab, & Rambow, 2012; Zribi, et al., 2014; Saadane & Habash, 2015; Tachicart & Bouzoubaa, 2014). Maghrebi CODA solved the problem by writing the singular third person direct object pronoun as w and the suffix of the verb conjugation in plural form and in Present as uw.

5.3 Maghrebi CODA

Here is an overview of specific CODA guidelines for Maghrebi Arabic. Maghrebi Arabic follows the same orthographic rules as MSA with the following exceptions and extensions. A practical example of the application of Maghrebi Arabic CODA is proposed in Table 3.

5.3.1. Phonological and Lexical Extensions

- As [mˤ], [bˤ], [zˤ], [nˤ], [lˤ] and [rˤ] are respectively transcribed using CODA guidelines as m, b, z, n, l, and r (Zribi, et al., 2014; Saadane & Habash, 2015), we have to transcribe long vowels differently according to their pronunciation so that the efficiency of CODA based text to speech converters and CODA based machine translation systems would not be limited. This is just what Zribi et al. and Saadane et al. have done when they got inspired from the Maltese ie and had respectively transcribed [ɑ:] and [ɛ:] as ـَا aA and as ـِا iA. Similarly, we propose to respectively transcribe [o:], [u:], [e:] and [i:] as ـُا Nw, ـُا uw, ـِا Ky, and ـِا iy.
- As [g] is kept in some words when converting them from Nomadic to sedentary accents, [g] and [q] constitute distinct phonemes that are pronounced the same in rural areas (Singer, 1984; Saadane & Habash, 2015; Tachicart & Bouzoubaa, 2014; Baccouche, 1972). As already done by most of the Maghrebi people, this geographically conserved [g] is transcribed as q ق in Tunisian Arabic, Libyan Arabic and Judeo-Tunisian Arabic and as k ك in all the other Maghrebi dialects (Volk, 2015; Aguadé, 2006).
- The phonemes [p] and [v] existing in the loanwords in Maghrebi Arabic (Singer, 1984; Saadane & Habash, 2015; Tachicart & Bouzoubaa, 2014). As transcribed using the Tunisian and Algerian CODA guidelines, they are written as b ب and f ف (Zribi, et al., 2014; Saadane & Habash, 2015).
- As mentioned by a number of linguists, the use of [œ:], [ɔ:], [y:], [œ], [ɔ], [y], [ɑ̃], [ɔ̃], and [ɛ̃] is limited in loanwords as they are quickly substituted by regular Maghrebi Arabic vowels in pronunciation (Singer, 1984; Gibson, 2009; Saadane & Habash, 2015). So, it is useless to consider them. We can consider them as regular Maghrebi Arabic vowels as shown in Table 2.

Regular Maghrebi Arabic vowels in Buckwalter Transcription	**Foreign European Phonemes**
uw	[œ:], [ɔ:]
iy	[y:]
an	[ɑ̃]
un	[ɔ̃]
in	[ɛ̃]
u	[œ], [ɔ]
i	[y]
aA	[ɑ:]
iA	[ɛ:]
a	[ɑ], [ɛ]

Table 2: Correspondence between Buckwalter Transliteration and European Vowel Phonemes used for Maghrebi Arabic

- However, each two Arabic letters should not be reduced to a unique Arabic letter even if these two graphemes are pronounced the same in Maghrebi Arabic just as clearly specified in the original CODA (Habash, Diab, & Rambow, 2012). For example, ت and ث are both pronounced as [t] in Moroccan Arabic (Heath, 1997). If we drop ث, we will obtain several nouns with the same spelling and different meanings just like تمر (dates) and ثمر (fruit) and this will cause reading difficulties and confusion.

5.3.2. Morphological and Orthographic Extensions

- Due to pronunciation simplification, Separated prepositions are sometimes abbreviated as clitics by users. Maghrebi CODA recovers the clitics as their original MSA equivalent when the abbreviation did not drop any consonant from the original MSA preposition. This is what is done in Tunisian and Algerian CODA but was not explicitly explained (Zribi, et al., 2014; Saadane & Habash, 2015). For example, ma becomes maA, However, Ea is kept as an l is dropped from its MSA equivalent that is ElY when it was created.
- [n] or [r] of Number Construct: Maghrebi CODA writes the phoneme /n/ or /r/ that is added after some numerals in construct cases, e.g., خمسطاشن راجل „15 men“ as opposed to خمسطاش „15 “ (Singer, 1984; Souag L. , 2005; Gibson, 2009; Owens, 1984).
- The CODA extensions for Tunisian and Algerian Arabic are adopted de facto (Zribi, et al., 2014; Saadane & Habash, 2015).

Raw Text	مساء الخير، مرحبا بيكم فالمباشر في ناس نسمة سباسيال. اليومة في ساعتنا الثانية، باش نحكيو ع التطورات الجديدة فالبلاد
Original Tunisian and Algerian Arabic CODA	مساء الخير، مرحبا بيكم في المباشر في ناس نسمة سباسيال. اليومة في ساعتنا الثانية، باش نحكيوا عالتطورات الجديدة في البلاد masa’ Alxiyr, marHbiA biykum fiy AlmubiA$ir fiy niAs nismap sbiAsyaAl. Alyuwmap fiy siAEitniA AlviAnyap, biA$ naHkiyuwA EAltaTawraAt Aljdiyda fiy AlbliAd.
Maghrebi CODA	مساء الـخير، مرحبا بيكم في الـمباشر في ناس نسمة سباسيال. الـيومة في ساعتنا الـثانية، باش نحكيوا عالـتطورات الـجديدة في الـبلاد masa’ Al_xiyr, marHbiA biykum fiy Al_mubiA$ir fiy niAs nismap sbiAsyaAl. Al_yuwmap fiy siAEitniA Al_viAnyap, biA$ naHkiyuwA EaAl_taTawraAt Al_jdiyda fiy Al_bliAd.
English Translation	Good evening. Hello. You are in a live broadcasting of “Nessma’s People Special Program”. Today in our second hour, we will be dealing with the latest news in the country.

Table 3: A Practical Application of the Maghrebi CODA

6. Conclusion

CODA is already used for main Maghrebi Arabic dialects (Zribi, et al., 2014; Saadane & Habash, 2015; Tachicart & Bouzoubaa, 2014). However, each CODA used for each of these dialects has its distinctive features. A unified Maghrebi CODA would let doing comparative studies between the main Maghrebi Arabic dialects possible. Furthermore, a specific CODA for Maghrebi Arabic that solves the problems of the original one can be used for doing more precise and detailed NLP for the Maghrebi Arabic dialects.

7. Acknowledgements

I have to thank Mr. Mohamed Maamouri and Mr. Nizar Habash for his helpful discussions and clarifications and the community of Wikimedia Incubator for testing the CODA in Tunisian, Algerian and Moroccan Arabic Wikipedias and Wiktionaries for one month.

8. References

Abumdas, A. H. A. (1985). *Libyan Arabic Phonology*. University of Michigan.

Aguadé, J. (2006). Writing dialect in Morocco. *Estudios de dialectología norteafricana y andalusí*, 10, 253-274.

Ayari, S. (1996). Diglossia and illiteracy in the Arab world 1. *Language, Culture and Curriculum*, 9(3), 243-253.

Baccouche, T. (1972). Le phonème g dans les parlers arabes citadins de Tunisie. *Revue tunisienne de sciences sociales*, 9(30-31), 103-137.

Biadsy, F., Habash, N., & Hirschberg, J. (2009). Improving the Arabic pronunciation dictionary for phone and word recognition with linguistically-based pronunciation rules. *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 397-405). Association for Computational Linguistics.

Chekili, F. (1982). *The morphology of the Arabic dialect of Tunis*. London: University of London.

Cohen, D., & el-Chennafi, M. (1963). *Le dialecte arabe Hassaniya de Mauritanie* (Vol. 5). C. Klincksieck.

Cohen, D. (1985). Some historical and sociolinguistic observations on the arabic dialects spoken by north african Jews. *Readings in the sociology of Jewish languages*, 246-260.

Dallaji-Hichri, I. (2010). *Hochzeitsbräuche in Nābil (Tunesien)*. (Doctoral dissertation, uniwien).

Dhaoudi, R., & Lahmar, M. (2004). Ali Douagi, The Ghalba Artist and the Taht Essour Troupe. In *The Taht Essour Troupe* (pp. 143-145). Cairo: General Egyptian Book Organization.

Diab, M., & Habash, N. (2007). Arabic dialect processing tutorial. *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Tutorial Abstracts* (pp. 5-6). Association for Computational Linguistics.

Eskander, R., Habash, N., Rambow, O., & Tomeh, N. (2013). Processing Spontaneous Orthography. *HLT-NAACL*, (pp. 585-595).

Ghazeli, S. (1981). La coarticulation de l'emphase en arabe. *Arabica*, 251-277.

Gibson, M. (2009). Tunis Arabic. *Encyclopedia of Arabic Language and Linguistics*, 4, 563–571.

Google Corporation. (2015, October 16). *Conventional Orthography for Dialectal Arabic*. Retrieved from Google Scholars: https://scholar.google.com/scholar?start=10&hl=fr&as_sdt=2005&sciodt=0,5&cites=10202889320135106055&scipsc=

Gouma, T. (2013). *L'emphase en arabe marocain: vers une analyse autosegmentale*. (Doctoral dissertation, Paris 8).

Habash, N. Y. (2010). Introduction to Arabic natural language processing. *Synthesis Lectures on Human Language Technologies*, 3(1), 1-187.

Habash, N., & Rambow, O. (2005). Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop. *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics* (pp. 573-580). Association for Computational Linguistics.

Habash, N., Diab, M. T., & Rambow, O. (2012). Conventional Orthography for Dialectal Arabic. *LREC*, (pp. 711-718).

Habash, N., Eskander, R., & Hawwari, A. (2012). A morphological analyzer for Egyptian Arabic. *Proceedings of the Twelfth Meeting of the Special Interest Group on Computational Morphology and Phonology* (pp. 1-9). Association for Computational Linguistics.

Habash, N., Soudi, A., & Buckwalter, T. (2007). On arabic transliteration. In *Arabic computational morphology* (pp. 15-22). Springer Netherlands.

Harrell, R. S. (2004). *A short reference grammar of Moroccan Arabic: With audio CD*. Georgetown University Press.

Heath, J. (1997). Moroccan Arabic phonology. *Phonologies of Asia and Africa (including the Caucasus)*, 1, 205-217.

Jourdan, J. (1952). *Cours pratique et complet d'arabe vulgaire, grammaire et vocabulaire: dialecte tunisien, 1. année*. C. Abela.

Lewis, M. P., Simons, G. F., & Fennig, C. D. (2009). *Ethnologue: Languages of the world* (Vol. 9). Dallas, TX: SIL international.

Maamouri, M. (1967). *The Phonology of Tunisian Arabic*. Ithaca: Cornell University.

Maamouri, M. (1973). The linguistic situation in independent Tunisia. *The American Journal of Arabic Studies*, 1, 50-65.

Maamouri, M. (1977). Illiteracy in Tunisia: An evaluation. In T. P. Gorman, *Language and literacy: Current issues and research*. Teherán, Irán: International Institute for Adult Literacy Methods.

Maamouri, M. (1983). Illiteracy in Tunisia. *Language in Tunisia*, 149-58.

Maamouri, M. (1998). *Language Education and Human Development: Arabic Diglossia and Its Impact on the Quality of Education in the Arab Region*. Philadelphia, PA: International Literacy Institute

Maamouri, M., Graff, D., Jin, H., Cieri, C., & Buckwalter, T. (2004). Dialectal Arabic Orthography-based Transcription. *EARS RT-04 Workshop*.

Masmoudi, A., Khmekhem, M. E., Estève, Y., Belguith, L. H., & Habash, N. (2014). A Corpus and Phonetic Dictionary for Tunisian Arabic Speech Recognition. *LREC*, 306-310.

Owens, J. (1984). *A short reference grammar of Eastern Libyan Arabic*. O. Harrassowitz.

Saadane, H., & Habash, N. (2015). A Conventional Orthography for Algerian Arabic. *ANLP Workshop 2015*, (p. 69).

Sayahi, L. (2014). *Diglossia and language contact: Language variation and change in North Africa*. Cambridge University Press.

Singer, H.-R. (1984). *Grammatik der arabischen Mundart der Medina von Tunis*. Berlin: Walter de Gruyter.

Sounkalo, J. (2008). *Spoken Hassaniya Arabic*. Dunwoody Press.

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

Souag, M. L. (2006). *Explorations in the Syntactic Cartography of Algerian Arabic*. School of Oriental and African Studies (University of London).

Tachicart, R., & Bouzoubaa, K. (2014). A hybrid approach to translate Moroccan Arabic dialect. *Intelligent Systems: Theories and Applications (SITA-14), 2014 9th International Conference on* (pp. 1-5). IEEE.

Tilmatine, M. (1999). Substrat et convergences: le berbère et l'arabe nord-africain. *Estudios de dialectología norteafricana y andalusí, EDNA*, 4, 99-120.

Turki, H., Zribi, R., Gibson, M., & Adel, E. (2015). Tunisian Arabic. *Wikipedia*. Wikimedia Foundation.

Volk, L. (2015). *The Middle East in the World: An Introduction*. Routledge.

Watson, J. C. (2007). *The phonology and morphology of Arabic*. Oxford university press.

Younes, J., & Souissi, E. (2014). A quantitative view of Tunisian dialect electronic writing. *5th International Conference on Arabic Language Processing, CITALA 2014*.

Younes, M. A. (2005). BRINGING HAMZA UNDER CONTROL: A PROPOSAL FOR SIMPLIFYING HAMZA-WRITING RULES IN ARABIC. *al-'Arabiyya*, 38-39, 99-127.

Zaidan, O. F., & Callison-Burch, C. (2014). Arabic dialect identification. *Computational Linguistics*, 40(1), 171-202.

Zawaydeh, B., Stallard, D., & Makhoul, J. (2003). *Babylon Transcription Guidelines*. Retrieved from http://ldc.upenn.edu/Catalog/docs/LDC2005S08/BBN-Babylontranscription-guidelines.pdf

Zribi, I., Boujelbane, R., Masmoudi, A., Ellouze, M., Belguith, L. H., & Habash, N. (2014). A Conventional Orthography for Tunisian Arabic. In *LREC* (pp. 2355-2361).