A Conventional Orthography for Tunisian Arabic

Inès Zribi, Rahma Boujelbane, Abir Masmoudi, Mariem Ellouze,
Lamia Belguith and †Nizar Habash
ANLP Research group, MIRACL Lab., University of Sfax, Tunisia
†Center for Computational Learning Systems, Columbia University, USA
E-mail: ineszribi@gmail.com, rahma.boujelbane@gmail.com, masmoudiabir@gmail.com, mariem.ellouze@planet.tn,
l.belguith@fsegs.rnu.tn, habash@ccls.columbia.edu

Abstract

Tunisian Arabic is a dialect of the Arabic language spoken in Tunisia. Tunisian Arabic is an under-resourced language. It 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 Tunisian Arabic; the two exist on a continuum dominated by mixed forms. In this paper, we present a conventional orthography for Tunisian Arabic, following a previous effort on developing a conventional orthography for Dialectal Arabic (or CODA) demonstrated for Egyptian Arabic. We explain the design principles of CODA and provide a detailed description of its guidelines as applied to Tunisian Arabic.

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

1. Introduction

The Arabic language in its modern form is a collection of dialects with various degrees of differences in terms of phonology, morphology, syntax and lexicon among each other and between them and Modern Standard Arabic (MSA). While MSA is the language of official use, the media and education, Dialectal Arabic (DA) is the language of daily life, the true native form of Arabic. The dialects are not taught in schools and have no standard orthography, although they have been for a long time the carriers of rich oral traditions. Tunisian Arabic (henceforth, TUN) is the primary dialect spoken in the North African country of Tunisia. TUN has some unique features that distinguish it from its direct neighboring dialects as well as other Arabic dialects. In the last decade, as has happened for many Arabic dialects, TUN has emerged as the language of informal communication online: in emails, blogs, discussion forums, SMS, etc. However the development of natural language processing (NLP) tools and resources for TUN still lags behind other dialects and is quite behind the state-of-the-art for MSA NLP. With the increasing presence of TUN online and the increasing use of language technologies for many languages (e.g., Siri), the need for work on technologies such as speech recognition, speech synthesis, telephony, machine translation, etc., for TUN is more real than ever before. The absence of resources creates a pronounced bottleneck for processing and building robust tools and applications. Applying NLP tools designed for MSA directly to TUN yields significantly low performance, making it imperative to build resources and dedicated tools for TUN processing (Diab et al, 2010, Boujelbane et al., 2013b).

In this paper, we discuss an important basic technology that is necessary for the efficient development of, and maximized synergy between, the various ongoing and future efforts (both tools and resources) on TUN NLP: the design of an orthography to be used as a common standard convention. Our work is a continuation of the work of Habash et al, (2012a) who proposed CODA, a Conventional Orthography for Dialectal Arabic, which is designed for the purpose of developing computational models of Arabic dialects and provided a detailed description of its guidelines as applied to Egyptian Arabic (EGY). We do not expect this convention to be produced by TUN speakers as input, but it is primarily for use in development NLP systems. Spontaneously written TUN will have to be converted automatically into its CODA version (Habash et al., 2012b, Eskander et al., 2013).

In this paper, we first review some previous related work (Section 1). In Section 2, we present an overview of TUN. In Section 3, we highlight the linguistic differences between TUN and both MSA and EGY to motivate some of our TUN CODA decisions. In Section 4, we present TUN CODA guidelines. And in Section 5, we briefly discuss ongoing efforts by the authors which use TUN CODA.

2. Related Works

Efforts on modernizing Arabic orthography and developing orthographies for Arabic dialects have been going on for many years (Habash et al, 2012a). Zawaydeh et al. (2003) and Maamouri et al. (2004) developed a set of rules for orthographic transcription and annotation of Levantine dialects in order to create a Levantine Arabic corpus. The proposed transcription rules are based on two levels of transcription: MSA-based transcription for the purpose of language modeling and Arabic orthographic system based transliteration for the purpose of acoustic modeling. Zribi et al, (2013a) presented *OTTA*, the *Orthographic Transcription for Tunisian Arabic* convention. This convention proposed the use of some rules based on MSA conventions and defined another set of rules which preserved the phonetic particularities of TUN. Zribi et al, (2013a) presented also a set of rules for annotation to use while transcribing spoken TUN. Habash et al. (2013a) introduced the concept of a conventional orthography for dialectal Arabic (CODA) and defined it for EGY. They identified five goals for CODA: (i) CODA is an internally consistent and coherent convention for writing DA; (ii) CODA is created for computational purposes; (iii) CODA uses the Arabic script; (iv) CODA is intended as a unified framework for writing all DAs; and finally, (v) CODA aims to strike an optimal balance between maintaining a level of dialectal uniqueness and establishing conventions based on MSA-DA similarities. Their convention is used in many of their NLP tools and resources for EGY (Habash et al, 2012a; Habash et al., 2013; Eskander et al., 2013; Pasha et al., 2014). In this paper, we extend the CODA guidelines they created for EGY to TUN. We believe that the CODA goals, especially the unified framework for all DAs, can help in maximizing synergy between, and encouraging adaptation from, other dialects and TUN, when it comes to resource creation.

3. An Overview of Tunisian Arabic

Arabic dialects are the vernacular of all Arabic speakers. They are the native languages of peoples of various Arabic countries, and these linguistic forms are sometimes very different from one region to another. TUN is a dialect of the Arabic language spoken in Tunisia. It is often referred to as دارجة *daArijaħ*[1], عامية *ςaAm~iyaħ*, or تونسي *tuwnsiy* which is considered as the low variety given that it is neither codified nor standardized even though it is the mother tongue and the variety spoken by all the population in daily usage (Saidi, 2007). Approximately 11 million people speak one or two of the many regional varieties of TUN; including the Tunis dialect (Capital), Sahil dialect, Sfax dialect, Northwestern Tunisian dialect, Southwestern Tunisian dialect, and Southeastern Tunisian dialect (Gibson, 1998, Talmoudi, 1980).

TUN is considered an under-resourced language. It has neither a standard orthography nor large collections of written text and dictionaries. Actually, there is no strict separation between MSA and its dialects: they coexist on a continuum dominated by mixed forms (MSA-DA). In addition, TUN is distinguished by the presence of words from several other languages. The presence of these languages mainly occurred due to historical facts. Indeed, they have rendered the linguistic situation in Tunisia rather complex. Lawson and Sachdev (2000) describe the linguistic situation in Tunisia as “poly-glossic” where multiple languages and language varieties coexist. Before describing Tunisian language situation, we present a brief historical overview of the TUN dialect.

During the centuries before the Islamic conquests, the native languages of the Maghreb in general were varieties of Berber. The few Arabic words that were part of Berber then are due to trade between the non-Arabs living in North Africa and the Arabs who traveled. Later on, the Arabization of the Maghreb was connected to the Islamic conquests from the east, which introduced the Arabic language on much larger scale in North Africa (Peirera, 2011). The Ottoman Turkish political domination of North Africa roughly from the mid-fifteenth to the late nineteenth century and the French colonization from 1830 had an impact on the absorption of foreign vocabulary into the lexicon of local Arabic dialects (Holes, 2004). In addition to Turkish and French, we find numerous examples of the European lexical elements in TUN. We can identify a significant number of expressions and words from Spanish and Italian, and even Maltese (which is an Arabic dialect historically). Table 1 contains some examples of borrowed words in TUN.

Words	**Transliteration**	**Origin**	**English sense**
براكة	*bar~iAkaħ*	Italian	booth
بانكة	*baAnkaħ*	Italian	bank
داكوردو	*daAkuwrduw*	Italian	okay
فيشطة	*fiyšTaħ*	Italian	party
ماكينة	*miAkiynaħ*	Italian	machine
كروسة	*kar~uwsaħ*	Italian	stroller
كوجينة	*Kuwjiynaħ*	Italian	kitchen
بابور	*baAbuwr*	Turkish	ship
سفنارية	*sfin~aAriyaħ*	Turkish	carrots
قهواجي	*kahwaAjiy*	Turkish	waiter
برنوس	*Barnuws*	Berber	burnous
كسكسي	*Kusksiy*	Berber	couscous
بطانية	*baT~aAniy~aħ*	Berber	blanket
صباط	*Sab~aAT*	Spanish	Shoe
بوسطة	*buwsTaħ*	French	post office
بلاصة	*blaASaħ*	French	Space
باكو	*baAkuw*	French	package
سبيطار	*sbiyTaAr*	French	hospital
قطوس	*qaT~uws*	Maltese	cat

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

In addition to all these borrowed terms which have been integrated in the TUN morpho-phonology, Tunisians code switch often in daily conversations, particularly from French, e.g., “*ça va*”, “*désolé*”, “*rendez-vous*”, etc. All these expressions and words are used without being adapted to the phonology.

4. Tunisian Arabic vs. MSA and Egyptian Arabic

TUN, EGY and MSA differ at the phonological, morphological and of course orthographic levels. We present in this section the main differences between the TUN, EGY, and MSA. For further discussions of Arabic morphology and orthography, see (Habash, 2010).

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

4.1. Phonological Variations

There are many phonological differences between TUN and both MSA and EGY (Mejri et al, 2009; Mejri and Baccouche, 2003; Habash et al, 2012a; Zribi et al, 2013a). We list below the main phonological differences:

- The non-MSA phoneme /g/ is used in both TUN and EGY, although in different ways. While in EGY, /g/ is how the MSA consonant ج *j* is pronounced, in TUN, the consonant equivalent to the MSA ق *q* is pronounced /g/, especially in rural dialects. In urban dialects, the consonant ق *q* is primarily pronounced as /q/ except for some words that have rural origins, e.g., بقرة *baqraħ* /bagra/ 'cow'. A few cases even create minimal pairs in urban dialects: قرون *qruwn* /qru:n/ 'centuries' and /gru:n/ 'horns'. The phoneme /g/ is also used in many TUN words that have no MSA cognates, such as بالقدا *biAlqdA* /bilgda:/ 'very well' and قربيطة *qurbiyTaħ* /gurbi:ta/ 'ribbon'.
- Within TUN, there are commonly accepted interchangeable lexically specific pronunciations. One example is the consonant /j/ which assimilates in some cases to /z/, e.g., the words جزار *jaz~aAr* 'butcher' and جليز *jliyz* 'tile' are pronounced respectively /jazza:r/ or /zazza:r/ and /jliz/ or /zliz/. Another example is the consonant /γ/ (غ) which assimilates also to /x/ (خ), e.g., the word غسالة *γas~iAlaħ* 'washing machine' is pronounced /γasse:la/ or /xasse:la/.
- Many words in TUN and EGY whose MSA cognates had the Hamza phoneme (glottal stop, /'/) lost that phoneme. In many cases, the glottal stop becomes a long vowel or disappears altogether, e.g. (MSA→TUN) كأس *kaÂs* /ka's/→/ka:s/ كاس *kaAs* 'cup', بئر *biŷr* /bi'r/→/bi:r/ بير *biyr* 'well', مؤمن *muŵmin* /mu'min/→/mumin/ مومن *muwmin* 'believer', and سماء *samaA'* /sama:'/ →/sma:/ سما *smaA* 'sky'.
- Like most Arabic dialects, TUN changes and neglects short vowels, especially when located at the end of a syllable, e.g., شرب *šariba* /šariba/, 'he drank' in MSA is transformed into /šrib/ in TUN. Generally, deleting the first vowel changes the syllabic structure of lexical units, which tend to become monosyllabic for certain words.
- Like many Arabic dialects (but unlike EGY), TUN elides many short vowels in unstressed contexts, e.g., (MSA→TUN) شَرِبَ /šariba/ → شرِب /šrib/ 'he drank' and سماء /sama:'/ → سما /sma:/ 'sky'.
- TUN has a long vowel /e:/ which does not exist in MSA. EGY /e:/ is almost exclusively related to MSA /ay/ (and occasionally originating from foreign words). In TUN, the situation is more complex: MSA /ay/ became primarily TUN /i:/, while MSA /a:/ has become /e:/ in some TUN words and remained as /a:/ in others, e.g., (MSA→TUN) بَيْت /bayt/ → بِيتْ /bi:t/ 'house', حِرَامْ /Hira:m/→/Hre:m/ 'cover' but حَرَامْ /Hara:m/ → /Hra:m/ 'sin'.

4.2. Morphological Variations

There are many morphological differences between Arabic dialects and MSA. The overarching themes are those of simplifying inflections and introducing new clitics.

- In terms of inflections, the MSA nominal case, verbal mood, and the dual and the plural feminine in verb conjunction have disappeared in TUN and EGY. TUN goes further than EGY in this trend: TUN lost the singular and plural feminine in verbal conjugation, e.g., (MSA→TUN) شَرِبْتِ *šaribti*→ شْرِبتْ *šribt* 'you drank', خَرَجْتُنَّ *xarajtun~a* 'they fem.pl. went' →خْرَجْتُوا *xrajtuwA* 'they went'. TUN normalized, also, the first person singular and plural to follow the 2nd and 3rd persons on verb conjugation respectively, e.g., (MSA→ TUN) خَرَجْتُ *xarajtu* 'I went out' and خَرَجْتَ *xarajta* 'you went out' →خْرَجِتْ *xrajit* 'I/you went out'. Furthermore, TUN almost lost all of the nominal dual forms, which are replaced with the word زوز *zuwz* /zu:z/ 'two' with the plural form, e.g., (MSA→ TUN) أستاذين *ÂustaAðayn* →اساتذة زوز *zuwz AasAtðaħ* 'two professors'.
- As with EGY, TUN introduces new non-MSA clitics. One example is the negation circum clitic ما+ +ش *ma+ +š* which MSA expresses with various particles: ما *mA*, لا *lA*, لن *lan*, and لم *lam* 'not'. Another example is the MSA verbal interrogation clitic أ *Âa* and the particle هل *hal*, which are replaced by the clitic شي *šiy* in TUN. Like others Arabic dialect, TUN has a set of clitics that are reduced forms of MSA words, e.g., the demonstrative proclitic +هـ *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→TUN) هذا الطفل *haðaA AlTifl*→هالطفل *haAlTfil* 'this child'. Similarly, TUN has the proclitic +ع *ça+*, a reduced form of the preposition على *çalaý* 'on/upon/about/to', e.g., (MSA→TUN) على الطاولة *çalaý AlTaAwilaħ* → عالطاولة *çaAlTaAwlaħ* 'on the table', and the proclitic +م *m+*, a reduced form of the preposition من *min* 'from' or the coordination conjunction مع *maça* 'with', e.g., (MSA→TUN) من الدار *mina AldaAri* →مالدار *miAldaAr* 'from the house'; مع بعضنا *maça baEDinaA* →مبعضنا *mabEaDnaA* 'with some of us'.

4.3. Orthographic Variations

The absence of orthographic standards in Arabic dialects and the phonological differences between MSA and TUN, in addition to the variability within TUN, are responsible for a lot of orthographic variations; dialect writers are often inconsistent even within themselves choosing to write words phonologically or in deference to etymological cognate forms in MSA. The following are some basic illustrative examples: the word متاعها *mtiAçhA* /mte:çha/ 'hers' is also written متاحها *mtiAHhA* /mte:Hha/ (TUN variants); the word /bi:r/ may be written بير *byr* or بئر *bŷr* (phonological spelling vs. cognate spelling); the words كتبوا *kitbuwA* ‘they wrote’ or كتبه *kitbuh* ‘he wrote it’ are often spelled using the same form: كتبو *ktbw* which introduces some ambiguity. Finally, shortened long vowels can be spelled long or short, e.g., ‘he didn’t say’ مقالش *maqaAliš* and ماقالش *mAqaAliš*. In this particular example where the MSA particle ما *mA* is the source of the proclitic *ma-*, another spelling is possible: ما قالش *mA qaAliš* (two separate words). Some adverbs have too multiple forms, e.g., the interrogative adverb آش *Āš* ‘what’ sometimes appears as a proclitic +ش *š*+ and in certain cases is transcribed as a separate word reflecting different pronunciations, e.g., شقال *šqaAl* and آش قال *Āš qaAl*.

4.4. Lexical Variations

As discussed above, the TUN lexicon is strongly influenced by Berber and by other languages such as Maltese, Turkish, Italian, Spanish, and French for various historical reasons.

5. CODA Guidelines for Tunisian Arabic

Our goal in this paper is to present a CODA map for Tunisian dialect. In this section, we summarize the CODA goals and principles (for more details see (Habash et al, 2012b)) and present a specific CODA guideline for TUN. An example of TUN in CODA is presented in Table 5.

5.1. CODA Goals and Principles

CODA is a conventionalized orthography for Arabic dialects (Habash et al, 2012b). It has five goals.

1. CODA is an internally consistent and coherent convention for writing Dialectal Arabic (DA): every word has a single orthographic rendering.
2. CODA is created for computational purposes.
3. CODA uses the Arabic script.
4. CODA is intended as a unified framework for writing all Arabic dialects.
5. 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 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.

5.2. Tunisian CODA

We present a summary of specific CODA guidelines for TUN. We consider the dialect used in the media as default TUN – this happens to be predominantly the dialect of the capital city Tunis.

TUN follows the same orthographic rules as MSA with the following exceptions and extensions.

5.2.1. Phonological Extensions

- Unlike EGY CODA, the long vowel /e:/ will be written as *ay* or *iA* depending on its MSA cognate: *ay* or *aA*, respectively. The sequence *iA* is not possible in MSA orthography, and as such it is a good solution for TUN words with *aA* MSA cognates, since the basic non-diacritical form of the word is preserved, e.g., حرام *HraAm* /Hra:m/ ‘sin’ and *HriAm* /Hre:m/ ‘cover’.
- Similar to EGY CODA, TUN long vowels, which are shortened in certain cases such as when adding clitics, are written in long form, e.g., تقول لها *tquwl lhaA* /tqullha/ ‘you tell her’ (not تقلها *tqulhaA*);تقول له *tquwl lh* /tqulluw/ ‘you tell him’ (not تقلّو *tquwluw*); and تقول لهم *tquwl lhm* /tqullhum/ ‘you tell them’ (not تقلهم *tqulhm*).

5.2.2. Phono-Lexical Exceptions

- Similar to EGY CODA, a set of consonants will be written in the form that reflects the MSA cognate root (a sort of historical spelling). We consider two specific cases here:
    a. **The Tunisian “gaf”** The letter ق *q* is used to represent the two consonants /q/ and /g/. A list specifying some examples of exceptional pronunciation as /g/ are presented below in Table 2.

CODA		Pronunciation	English
بقرة	*baqrah̄*	/bagra/	cow
بالقدا	*biAlqdaA*	/bilgda:/	very well
قازوز	*qiAzuwz*	/ge:zu:z/	Soda
قربيطة	*qurbiyTah̄*	/gurbi:Ta/	ribbon
منقالة	*minqiAlah̄*	/minge:la/	watch
قلص	*qlaAS*	/gla:S/	closet

Table 2: Some words have an exceptional pronunciation in TUN.

b. **TUN Consonant with Multiple Pronunciations** Consonants with multiple pronunciations will be written using the form closes to the MSA cognate if an MSA cognate exists. Table 3 presents some examples. It is important to note that the phenomenon of multiple pronunciations was not addressed in the EGY CODA as that work focused on Cairene Arabic, which seems to have less variations compared to the dialect of Tunis. Furthermore, TUN, unlike EGY seems to have more

MSA-like pronunciations, e.g., for ق /q/ and ذ /ð/, where MSA spelling is simply the same as TUN, and no CODA exceptions are needed.

CODA		Multiple Pronunciations	English
جزار	*jaz~aAr*	/jazzar/ /zazzar/	butcher
ثمة	*θam~aħ*	/θamma/ /famma/	there is
رسول	*rasuwl*	/rasu:l/ /raSu:l/	prophet
متاعها	*mtiAςhaA*	/mte:ςha/ /mte:Hha/	hers
غسالة	*γas~iAlaħ*	/γasse:la/ /xasse:la/	washer
سأل	*sÂal*	/s’al/ /shal/	he asked
صايغي	*SaAyγiy*	/saAyγiy/ /SaAyγiy/	jeweler

Table 3: Some examples that have multiple pronunciations in TUN.

- As in EGY CODA, TUN words which have a Hamzated MSA cognate may not be spelled in a way corresponding to the MSA cognate, i.e., they will be spelled phonologically, e.g., يبدا *yibdaA* as opposed to the MSA form يبدأ *yabdaÂu* ‘he starts’. Word initial real glottal stops (همزة القطع) have all but disappeared in TUN. As such, word initial Hamzated Alif are not seen in TUN CODA, e.g., احمد *AaHmad* (not أحمد *ÂaHmad*), واحمد *wAaHmad* /waHmad/ (not وأحمد *wÂaHmad*), الاولاد *AlAawliAd* /lawle:d/ ‘children’ (not لولاد *lawliAd*), الاساتذة *AlAasiAtðaħ* /*lase:tða*/ ‘professors’ (not لساتذة *lasiAtðaħ*).
- **N of Number Construct**. TUN CODA writes the phoneme /n/ that is added after some numerals in construct cases, e.g., خمسطاشن راجل *xmsTaAšn raAjil* ‘15 men’ as opposed to خمسطاش *xmsTaAš*.

5.2.3. Morphological Extensions

Attached Clitics TUN shares most of MSA’s attached clitics, e.g., the definite article +ال *Al+*, the coordinating conjunction +و *w*+, etc. There are other attached clitics are defined in TUN, e.g., the interrogation proclitic شي + +*šiy*, the negation particle enclitic ش+ +*š*. TUN uses non-MSA single letter clitics, e.g., +ع *ς*+, +م *m+*, etc. As an example, consider the word وشريتوهاشي *wišriytuwhaA$iy*. ‘And have you bought it?’

Enclitics		Suffixes	Stem	Proclitics
شي	ها	تو	شري	و
Šiy	*haA*	*tuw*	*šriy*	*wi*

Table 4: Tokenization of the word وشريتوهاشي *wšriytuwhaAšiy* (Arabic is written from right to left).

Separated Clitics The negative form of a verb without the negation particle doesn’t make sense in TUN dialect, e.g., كتبش *ktibš* /ktib*š*/. But, to have a standard CODA across Arabic dialects, TUN CODA map preserves the rule of spelling of the indirect object enclitics and also the negation proclitic which requires the separation with a space between the negation particle and the indirect object enclitics, e.g., ما قال ليش *mA qaAl liyš* /ma+qal+li+*š*/ ‘he did not tell me’.

5.2.4. Lexical Exceptions

Like EGY CODA guidelines, TUN CODA includes a word list specifying the spelling of TUN words that have exceptional spelling or that are commonly spelled in different ways and thus require the CODA choice to be stated clearly. Examples include pronouns such as انتي *Aintiy* (not انت *Ainti*) ‘*you fem.sg.*’, demonstratives such as هذاكه *haðAkah* (not هاذاكة *hAðAkaħ*) ‘that’, nouns such as عالسلامة *ςaAlsliAmaħ* (not عسلامة *ςasliAmaħ*).

Furthermore, many foreign words are used and even integrated in TUN. These words containing the non-Arabic phoneme /g/, /v/, and /p/ will be written using the Arabic script characters *q*, *f* and *b*, respectively, e.g. قازوز *qiAzuwz* ‘soda’, فيستة *fiystaħ* ‘jacket’, and بورتابل *buwrtaAbl* ‘mobile phone’.

6. Ongoing Efforts using TUN CODA

The TUN CODA we propose in this paper is already used in several NLP resources and tools that we developed. Zribi et al. (2013b) respect TUN CODA rules while transcribing their Spoken Tunisian Arabic Corpus (STAC) (Zribi et al. 2013b). Also, they use TUN CODA for morphological analysis, where it defines the internal databases for word and affix forms. More particularly, Zribi et al. (2013b) used CODA for defining the structure of the derivation patterns for TUN nouns and verbs and also for the definition of the writing form of different affixes and clitics for TUN. Boujelbane et al. (2013b) use TUN CODA for creating TUN corpora to train language models for an automatic speech recognition system. They are also building lexical resources such as a bilingual lexicon MSA-TUN to convert MSA corpora to TUN corpora (Hamdi et al., 2013) to generate more language model training data. Masmoudi et al. (2014) use TUN CODA in the context of realization of a speech recognition system for TUN used in railway transport network. The realization of this system requires the usage of several resources, namely linguistic resources (texts), a pronunciation dictionary, language model, etc. However, such resources are absent in TUN, which brings the authors to transcribe them manually. Transcripts must follow standards of writing. Furthermore, the creation of a pronunciation dictionary requires studying the phonetic characteristics of TUN. A convention of writing TUN identifying the phonological variations is very important.

Raw Text	مساء الخير مرحبا بكم في المباشر في ناس نسمة سبسيال اليومة في ساعتنا الثانية باش نحكيو على تطورات جديدة . فمة تطورات في موضوع تشكيل الحكومة . نحكيو عليها مع التحالف الديموقراطي . نشوفو وين وصلت الأمور معاهم ، ونشوفو شنو سار في الأحداث متاع مقبرة زلاز . و بش نحكيو زادة على الزيادات في اسوام الڨاز و نسهلو ع انعكاسات متاعو ع المواطن . *msA' Alxyr mrHbA bkm fy AlmbAšr fy nAs nsmħ sbsyAl Alywmħ fy sAςtnA AlθAnyħ bAš nHkyw ςlý tTwrAt jdydħ . fmħ tTwrAt fy mwDwς tškyl AlHkwmħ . nHkyw ςlyhA mς AltHAlf AldymwqrATy . nšwfw wyn wSlt AlÂmwr mςAhm , wnšwfw šnw sAr fy AlÂHdAθ mtAς mqbrħ zlAz . w bš nHkyw zAdħ ςlý AlzyAdAt fy AswAm AlgAz w nshlw ς AnςkAsAt mtAςw ς AlmwATn .*
CODA	مسا الخير مرحبا بكم في المباشر في ناس نسمة سبسيال اليوم في ساعتنا الثانية باش نحكيوا على تطورات جديدة . ثمة تطورات في موضوع تشكيل الحكومة . نحكيوا عليها مع التحالف الديموقراطي . نشوفوا وين وصلت الامور معاهم ، ونشوفوا شنوة صار في الاحدا متاع مقبرة جلاز . وباش نحكيوا زادة على الزيادات في اسوام القاز ونسألوا عالانعكاسات متاعه عالمواطن . *msA Alxyr mrHbA bkm fy AlmbAšr fy nAs nsmħ sbsyAl Alywm fy sAςtnA AlθAnyħ bAš nHkywA ςlý tTwrAt jdydħ . θmħ tTwrAt fy mwDwς tškyl AlHkwmħ . nHkywA ςlyhA mς AltHAlf AldymwqrATy . nšwfwA wyn wSlt AlAmwr mςAhm , wnšwfwA šnwħ SAr fy AlAHdAθ mtAς mqbrħ jlAz . wbAš nHkywA zAdħ ςlý AlzyAdAt fy AswAm AlqAz wnsÂlwA ςAlAnςkAsAt mtAςh ςAlmwATn .*
English	Good evening. Hello. You are on the air with “Nesma People Special Program”. Today in our second hour, we'll talk about some new developments. There are developments on the subject of forming the government. We will discuss it with the Democratic Alliance. We will see where they have reached on this topic and we will discover what happened in the events at Jlaz cemetery. We will talk also about the increase in gas prices and question its impact on the citizen.

Table 5: An example of sentences in TUN.

Acknowledgment

The last author was supported by the Defense Advanced Research Projects Agency (DARPA) under 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.

References

Boujelbane, R., Khemakhem, M. E., and Belguith, L. (2013a). Mapping rules for building dialect Tunisian Lexicon and Generating Corpora. In Proceeding of International Joint Conference on Natural Language Processing (IJCNLP 2013), Nagoya, Japan.

Boujelbane, R., Khemekhem M. E., BenAyed S., Belguith L. H. (2013b). Building bilingual lexicon to create Dialect Tunisian corpora and adapt language model. In Proceedings of the Second Workshop on Hybrid Approaches to Translation.

Diab, M., Habash, N., Rambow, O., Altantawy, M., and Benajiba, Y. (2010). COLABA : Arabic Dialect Annotation and Processing, In Proceedings of LREC Workshop on Semitic Language Processing, Malta, May 2010. pp. 66–74.

Eskander, R., Habash, N., Rambow, O., and Tomeh, N. (2013). Processing Spontaneous Orthography. In Proceedings of Conference of the North American Association for Computational Linguistics (NAACL), Atlanta, Georgia.

Gibson, M. (1998). *Dialect contact in Tunisian Arabic: Sociolinguistic and structural aspects,* In Ph.D. Dissertation, University of Reading.

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

Habash, N., Diab, M., and Rambow, O. (2012a). Conventional Orthography for Dialectal Arabic. In: Proceedings of the Language Resources and Evaluation Conference (LREC), Istanbul.

Habash, N., Eskander, R., and Hawwari A. (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.

Habash, N., Roth, R., Rambow, O., Eskander, R., and Tomeh, N. (2013). Morphological Analysis and Disambiguation for Dialectal Arabic. In *Proceedings of Conference of the North American Association for Computational Linguistics (NAACL)*, Atlanta, Georgia.

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

Hamdi, A., Boujelbane, R., Habash, R., and Nasr, A. (2013). The Effects of Factorizing Root and Pattern Mapping in Translating between Tunisian Arabic and Standard Arabic. In *Proceedings of the Machine translation Summit (MT Summit)*, Nice, France.

Holes, C. (2004). *Modern Arabic: Structures, Functions, and Varieties.* (Georgetown University Press., p220). Washington.

Lawson, S., and Sachdev, I. (2000). Code Switching in Tunisia: attitudinal and behavioral dimensions. In *Journal of Pragmatics 32, (9): 1343-61.*

Maamouri, M., Buckwalter, T., Cieri, C. (2004). *Dialectal Arabic Telephone Speech Corpus: Principles, Tool Design, and Transcription Conventions*. In: NEMLAR International Conference on Arabic Language Resources and Tools, Cairo, pp. 22--23.

Masmoudi, A., Khemakhem, M. E., Estève, Y., Belguith, L. and Habash, N. (2014). The Tunisian Dialect phonetic dictionary for speech recognition. In: *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Iceland.

Mejri, S., and Baccouche, T. (2003). L'atlas linguistique de Tunisie: repères méthodologiques pour la description du système dialectal. In: Lentin, J., Lonnet, A. (eds.) Mélanges David Cohen, pp. 47–54. Maisonneuve & Larose, Paris.

Mejri, S., Said, M., and Sfar, I. (2009). Pluringuisme et diglossie en Tunisie. In: *Synergies Tunisie, vol. (1),* pp. 53--74.

Pasha, A., Al-Badrashiny, M., Diab, M., El Kholy, A., Eskander, R., Habash, N., Pooleery, M., Rambow, O. and Roth, R. (2014). "MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic." In Proceedings of the 9th International Conference on Language Resources and Evaluation, Reykjavik, Iceland.

Peirera, C. (2011). Arabic in the North African Region. In W. Stefan (Ed.), The Semitic Languages. pp. 954--969.

Saidi, D. (2007). Typology of Motion Event in Tunisian Arabic. In *LingO*. pp. 196--203.

Talmoudi, F. (1980) A morphosyntactic study of Romance verbs in the Arabic dialects of Tunis, Susa, and Sfax, In *Göteborg: Acta Universitatis Gothoburgensis*.

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

Zribi, I., Graja, M., Khemakhem, M. E., Jaoua, M., and Belguith, L. H. (2013a). Orthographic Transcription for Spoken Tunisian Arabic. CICLing 2013, Part I, LNCS 7816 (pp. 153–163).

Zribi I., Khemakhem M. E., and Belguith, L. H. (2013b). Morphological analysis of Tunisian Dialect. In *Proceeding of International Joint Conference on Natural Language Processing (IJCNLP 2013), Nagoya, Japan*.