Conventional Orthography for Dialectal Arabic

Nizar Habash, Mona Diab, Owen Rambow

Center for Computational Learning Systems
Columbia University
New York, NY, USA
{habash,mdiab,rambow}@ccls.columbia.edu

Abstract

Dialectal Arabic (DA) refers to the day-to-day vernaculars spoken in the Arab world. DA lives side-by-side with the official language, Modern Standard Arabic (MSA). DA differs from MSA on all levels of linguistic representation, from phonology and morphology to lexicon and syntax. Unlike MSA, DA has no standard orthography since there are no Arabic dialect academies, nor is there a large edited body of dialectal literature that follows the same spelling standard. In this paper, we present CODA, a conventional orthography for dialectal Arabic; it is designed primarily for the purpose of developing computational models of Arabic dialects. We explain the design principles of CODA and provide a detailed description of its guidelines as applied to Egyptian Arabic.

Keywords: Arabic, Dialects, Orthography

1. Introduction

Dialectal Arabic (DA) refers to the day to day vernaculars spoken in the Arab world. DA lives side by side with Modern Standard Arabic (MSA). As spoken varieties of Arabic, DAs differ from MSA on all levels of linguistic representation, from phonology and morphology to lexicon and syntax. Most differences are at the phonological, morphological and lexical levels. MSA is the language of education in the Arab world, while DA is perceived as a lower form of expression; this has implications on the way DA is used in daily written venues. On the other hand, being the natively spoken language, DAs have been the object of many efforts to study their patterns and regularities (Erwin, 1963; Cowell, 1964; Abdel-Massih et al., 1979; Holes, 2004). Most of such studies have been field work or theoretical in nature with limited transcribed data.

In current statistical Natural Language Processing (NLP) there is an inherent need for large-scale annotated resources. For DA, the absence of such resources creates a pronounced bottleneck for processing and building robust tools and applications. Applying NLP tools designed for MSA directly to DA yields significantly low performance, making it imperative to build resources and dedicated tools for DA processing.

In recent years, DA has emerged as the language of informal communication online, in emails, blogs, discussion forums, SMS, etc. These genres pose significant challenges to NLP in general for any language including English. The challenge arises from the fact that the language is less controlled and more speech-like while many of the textually oriented NLP techniques are designed for processing edited text. The problem is compounded for Arabic precisely because of the use of DA in these genres. Unlike MSA, DAs have no standard published orthographies since there are no Arabic dialect academies nor is there a large body of edited dialectal literature that follows the same spelling standard. There is a wide range of conventions used by native speakers in naturally occurring text and by creators of various DA computational resources (tools, transcript collections). These conventions are often inconsistent, a problem for efforts in DA computational processing.

In this paper, we present CODA, a conventional orthography for dialectal Arabic that aims at filling this gap; it is designed primarily for the purpose of developing computational models of Arabic dialects. The paper is organized as follows. Section 2. discusses previous efforts. Section 3. presents a sketch of MSA orthography. Section 4. outlines relevant differences between MSA and DA. Section 5. highlights the goals and principles of CODA. Section 6. details CODA decisions for one dialect, Egyptian Arabic (EGY).

2. Previous Work

The issue of standardization of DA orthography is politically loaded, since it is seen by many as an attack on MSA hegemony and Arab nationalism. One extreme example is that of the Lebanese poet Said Akl, who proposed a Latin-based orthography for Lebanese (Arabic) in the 1960s (Arkadiusz, 2006). On the other end of the spectrum, the Asaakir system, which is the only approach to Arabic dialect orthography approved by the Arabic Language Academy of Egypt, utilizes additional diacritics to add on top of standard Arabic words to produce their dialectal forms (‘Asaakir, 1950). This standard is not used outside of very limited circles (Al-Tonsi and Al-Sawi, 1990). Various DA dictionaries utilize Arabic, Latin or mixed script orthographies (Badawi and Hinds, 1986). These resources often focus on lemmatized (uninflected) forms. Resources developed for DA automatic speech recognition are typically phonological transcriptions that are not readily usable for modeling written text (Kilany et al., 2002; Maamouri et al., 2004). Our CODA guidelines are inspired by the Linguistic Data Consortium (LDC) guidelines for transcribing Levantine (LEV) and Iraqi (IRQ) Arabic (Maamouri et al., 2004). They differ from them in that, whereas the LDC guidelines are for transcription, and thus focus more on phonological variations in sub-dialects, CODA is intended for general purpose writing in a way that abstracts from these variations when possible. CODA is intended and designed as a common convention for all DAs, making choices that minimize differences among them. We extend the LDC guidelines to cover EGY in detail – for which we profited from the work on CallHome Egyptian (Kilany et al., 2002).

In a previous publication (Diab et al., 2010), we presented a different conventional orthography (CCO: COLABA Conventional Orthography). CCO differs from CODA in many respects, the most important of which is that CCO is intended to capture specifics of dialectal phonology and morphology. This goal, however, is very hard to achieve as the annotator/transcriber training process was long and tedious and annotators had a very hard time learning what some described as a “foreign” system of writing. Also, inter-annotator agreement was rather low, especially over short vowels that are often ignored in Arabic orthography.

3. A Sketch of MSA Orthography

We present a general sketch of Arabic orthography starting with a brief description of MSA phonology followed by a presentation of Arabic script and MSA orthographic rules. For more details, see Habash (2010).

3.1. MSA Phonology

Consonants and Vowels MSA’s phonological profile includes 28 consonants, three short vowels, three long vowels and two diphthongs (/ay/ and /aw/). Some of the consonants are emphatic versions of other consonantal phonemes. Emphasis (التفخيم *Altafxiym*)[1] is a bass effect giving an acoustic impression of hollow resonance to the basic sounds (Holes, 2004). MSA vowel phonemes are limited in number compared to English or French; however, there are many allophones to each of them depending on the consonantal context, such as becoming emphatic near emphatic consonants. Another interesting phenomenon, called *Waqf*, allows for optionally dropping the word-final short vowels marking syntactic case in utterance-final words.

Morphotactics There are numerous additional phonological variations that are limited to specific morphological contexts, i.e., they are constrained morpho-phonemically as opposed to phonologically. The most common example of such phenomena is the assimilation of the Arabic definite article proclitic +ال *Al+* to the first consonant in the noun or adjective it modifies if this consonant is an alveolar, dental or inter-dental phoneme (except for /j/). This set of 14 consonants is called *the Sun Letters*. It includes among others, ت *t*, ث *θ*, ز *z*, and ش *š*. For example, the word الشمس *Al+šams* ‘the sun’ is pronounced /aššams/ not */alšams/. The rest of the consonants are called the *Moon Letters*. A less common example is the phoneme /t/ in verbal pattern VIII (Ai1ta2a3)[2] which becomes voiced (/d/) when adjacent to specific root consonants such as /z/: *Aiztahar* becomes *Aizdahar* ‘it flourished’.

Syllabic Structure and Stress Syllabically, MSA is rather simple having mostly CV and CVC syllables and a few CVCC syllables in some word final positions. Stress is not phonemic in Arabic.

3.2. Arabic Script

The Arabic script is a right-to-left alphabet. There are two types of symbols in the Arabic script for writing words: letters and diacritics. Arabic letters are written in cursive style in both print and script (handwriting). Diacritics are additional zero-width symbols that appear above or below the letters. MSA uses 36 letters and nine diacritics. We discuss the different types of letters and diacritics in more detail below as part of the orthography of MSA. There are a few additional letters that are not officially part of Arabic script for MSA. Most commonly seen are پ *p*, چ *c* ڤ *v* and گ *g*. These are borrowings from other languages typically used to represent sounds not in MSA.

3.3. MSA Orthography

An orthography is a specification of how the sounds of a language are mapped to/from a particular script. We present an account of standard MSA orthography using the Arabic script. The correspondence between writing and pronunciation in MSA falls somewhere between that of languages such as Spanish and Finnish, which have an almost one-to-one mapping between letters and sounds, and languages such as English and French, which exhibit a more complex letter-to-sound mapping (El-Imam, 2004). Most Arabic letters and diacritics have a one-to-one mapping to MSA phonemes. However, there is a number of common important exceptions (El-Imam, 2004; Habash et al., 2007; Biadsy et al., 2009).

3.3.1. Basic Phonemic Map

Consonants All of the consonants except for the glottal stop (aka, *Hamza*) have a unique mapping into an Arabic letter.

Short Vowels The three short vowels /a/, /u/ and /i/ are written using the three short-vowel diacritics, ـَ *a*, ـُ *u*, and ـِ *i*, respectively.

Long Vowels Long vowels are written as a combination of a short vowel and a glide consonant. The long vowels /ū/, /ī/ and /ā/ are written as ـُو *uw*, ـِي *iy* and ـَا *aA*, respectively. The diphthongs /ay/ and /aw/ are written as ـَو *aw* and ـَي *ay*.

No Vowels The Sukun ـْ. diacritic marks vowel absence. It is typically used to mark syllable boundaries. In the case of two identical consecutive consonants with no vowel between them, the second repeated consonant is replaced with the *Shadda*, the consonant doubling diacritic, e.g., بّ *b∼* (/bb/).

Vowels at the Beginning of Words Arabic diacritics can only appear after a letter. As such, word-initial vowels are preceded with an extra silent Alif (ا *A*) called *Hamzat-Wasl*. The following are some examples: the word /kattaba/ ‘he dictated’ is written as كَتَّبَ *kat∼aba*, the word /maktūb/ ‘letter/written’ is written as مَكْتُوب *mak.tuwb*, and the word /inkataba/ ‘it was written’ is written as اِنْكَتَبَ *Ain.kataba*.

3.3.2. Hamza Spelling

The consonant Hamza (glottal stop /’/) has multiple forms in Arabic script: ء ’, آ Ā, أ Â, ؤ ŵ, إ Ǎ and ئ ŷ. The different

[1] Arabic transliteration is presented in the Habash-Soudi-Buckwalter scheme (Habash et al., 2007): (in alphabetical order)
أ ب ت ث ج ح خ د ذ ر ز س ش ص ض ط ظ ع غ ف ق ك ل م ن ه و ي
Â b t θ j H x d ð r z s š S D T Ď ς γ f q k l m n h w y
and the additional symbols: ’ ء, Â أ, Ǎ إ, Ā آ, ŵ ؤ , ŷ ئ , ħ ة, ý ى.

[2] The digits 1/2/3 refer to root radicals.

forms are governed by a set of complex spelling rules that reflect word position, vocalic context and neighboring letter forms (Habash and Rambow, 2007). For example, consider the different Hamza forms in the following word meaning 'his glory' when its case marker changes: بهاءه *bahA'ahu* /bahā'ahu/ (accusative), بهاؤه *bahAŵuhu* /bahā'uhu/ (nominative), and بهائه *bahAŷihi* /bahā'ihi/ (genitive).

Arabic orthography distinguishes between two types of Hamzas. The Real Hamza (همزة قطع) is always pronounced as a glottal stop regardless of whether it is at the beginning or in the middle of a word. The Temporary Hamza, or Hamzat-Wasl (همزة وصل) – see above, is a word-initial glottal-stop vowel allophone that only appears if the word is at the beginning of a sentence/utterance.

3.3.3. Clitic Spelling

A clitic is a morpheme that has the syntactic characteristics of a word, but shows evidence of being phonologically bound to another word (Loos et al., 2004). In this respect, a clitic is distinctly different from an affix, which is phonologically and syntactically part of the word. MSA has a small number of such clitics which are written attached to the word. Proclitics (prefixing clitics) are typically single-letter particles, such as the conjunction +و *wa*+ 'and', the preposition +ب *bi*+ 'in/with', the future particle +س *sa*+ 'will' and the definite article +ال *Al*+ 'the'. Enclitics (suffixing clitics) are generally object/possessive pronouns, e.g. هم+ +*hum* 'them/their'. Multiple clitics can appear in a word. For example, the word وسيكتبونها *wa+sa+yaktubuwna+hA* 'and they will write it' has two proclitics and one enclitic. Clitics generally do not modify the spelling of the word base they attach to, although there are a few exceptions, which are presented below.

3.3.4. Morpho-phonemic Spelling

The Arabic script contains a small number of common morphophonemic spellings. These are cases that spell a morpheme with multiple allomorphs using a form that reflects the phonology of the most common allomorph or that of some combination of allomorphs.

Definite Article The Arabic definite article is always spelled as +ال *Al*+ even though it phonologically assimilates to the first consonant in the noun or adjective it attaches to (as discussed above). The Alif of the definite article remains written when additional proclitics are added to the word except with the prepositional proclitic +ل *li*+, e.g., compare كالكتاب *ka+Al+kitAb* /kalkitāb/ 'like the book' and للكتاب *li+l+kitAb* /kilkitāb/ 'for the book'.

Ta-Marbuta The Ta-Marbuta (ة *ħ*) is typically a feminine ending. It can only appear at the end of a word. In MSA, it is pronounced as /t/ unless it is not followed by a vowel (as in *Waqf*), in which case it is silent. For example, المكتبة *Almaktabaħu* 'the library' is pronounced /'almaktabatu/ (normal) or /'almaktaba/ (Waqf). When the morpheme it represents is in word-medial position, such as before an enclitic, it is written using the letter Ta (ت). For example, مكتبة+هم *mktbħ+hm* 'library+their' is written as متكبتهم *mtkbthm* 'their-library'.

Alif-Maqsura The Alif-Maqsura (ى ý) is a silent derivational marker marking a range of morphological information from feminine endings to underlying word roots. Alif-Maqsura always follows a short vowel /a/ at the end of a word. In word-medial positions, it may be written using the letters Alif (ا) or a Ya (ي). For example, مستشفى+هم *mstšfý+hm* 'hospital+their' is written مستشفاهم *mstšfAhm* 'their-hospital'; however, إلى+هم *Ǎlý+hm* 'to+them' is written إليهم *Ǎlyhm* 'to-them'.

Waw of Plurality A silent Alif appears in the morpheme +وا +*uwA* /ū/ (واو الجماعة *wAw AljamAςaħ*) which indicates a masculine plural conjugation in verbs. For example, كتبوا *katabuwA* 'they wrote' is pronounced /katabū/. This Alif is deleted if followed by an enclitic, e.g., كتبوها *katabuwhA* 'they wrote it'.

Nunation Nunation is a nominal indefiniteness morpheme in MSA. It has the form of a word-final /n/, which is written using the *nunation* diacritics ـً ã, ـٌ ũ and ـٍ ĩ. These diacritics combine the short vowel (case marker) preceding the nominal indefiniteness morpheme: they are pronounced /an/, /un/ and /in/, respectively. For example, كِتابٌ *kitAbũ* is pronounced /kitābun/. A silent Alif appears word finally with some nunated nouns (before or after the diacritic), e.g., كتاباً *kitAbAã* or *kitAbãA* /kitāban/.

3.3.5. Exceptional Spelling

There are few cases of exceptional spelling that are outside the rules presented above. Archaic spellings of some common words, e.g., اللّٰه *All∼áh* 'Allah' and هٰذا *háðA* 'this', use a diacritic called the Dagger Alif (الألف الخنجرية á), which represents a *long /a/ vowel* (/ā/). Another common odd spelling is that of the proper name عمرو *ςamrw* /ςamr/ 'Amr' where the final و *w* is silent.

3.3.6. Notes on Consistency and Standardization

Diacritic Optionality Whereas letters are always written, diacritics are optional: written Arabic can be fully diacritized, partially diacritized, or entirely undiacritized. Over 98% of written Arabic words are diacritic free (Habash, 2010). This is not so much a problem when mapping from phonology to script but it poses a challenge in the other direction.

Suboptimal Spelling A few letters are not spelled consistently. Arabic writers often replace hamzated letters with the un-hamzated form, e.g., أ Â ⇔ ا *A*, or through two-letter spelling, e.g., ئ ŷ ⇔ ىء *ý'*. And the word-final letters ي *y* and ى *ý* are often used interchangeably (Buckwalter, 2007).

Regional Standards MSA orthography is largely standardized. However, a few variations remain across and within different Arab countries. For example, there are two common spellings for names of geographic entities ending with an /a/ vowel: /sūrya/ 'Syria' appears as سوريا *swryA* and سورية *swryħ*. Hamza spelling rules may have some exceptions also. For example, the word for 'official/responsible' appears as مسؤول *masŵuwl* (common in the Levant) and مسئول *masŷuwl* (common in Egypt).

4. Dialectal Arabic vs. MSA

We present below a listing of important differences between DAs and MSA. For more information on Arabic dialects, see (Harrell, 1962; Erwin, 1963; Cowell, 1964; Abdel-Massih et al., 1979).

4.1. Phonological Variations

Arabic dialects vary phonologically from MSA and from each other. Some of the common variations include the following (Holes, 2004; Habash, 2006; Biadsy et al., 2009; Habash, 2010):

- The MSA alveolar affricate ج /j/ is realized as /g/ in EGY, as /ž/ in LEV and as /y/ in Gulf Arabic (GLF). For example, جميل 'handsome' is pronounced /jamīl/ (MSA, IRQ), /gamīl/ (EGY), /žamīl/ (LEV) and /yamīl/ (GLF). The EGY and LEV pronunciations are used for MSA in those regions.
- The MSA consonant ق /q/ is realized as a glottal stop /'/ in EGY and LEV and as /g/ in GLF and IRQ. For example, طريق 'road' appears as /Tarīq/ (MSA), /Tarī'/ (EGY and LEV) and /Tarīg/ (GLF and IRQ). These changes do not apply to modern and religious borrowings from MSA. For instance, قرآن 'Qur'an' is never pronounced anything but /qur'ān/.
- The MSA consonant ث /θ/ is pronounced as /t/ in LEV and EGY (or /s/ in more recent borrowings from MSA), e.g., ثلاثة 'three' is pronounced /θalāθa/ in MSA versus /talāta/ in EGY.
- The MSA consonant ذ /ð/ is pronounced as /d/ in LEV and EGY (or /z/ in more recent borrowings from MSA), e.g., MSA ذنب *ðanb* 'fault' and كذب *kiðb* 'lies' are pronounced /zanb/ and /kidb/, respectively.
- The MSA consonants ض /D/ (emphatic d) and ظ /Ď/ (emphatic /ð/) are both normalized to /D/ in EGY and LEV and to /Ď/ in GLF and IRQ. In modern borrowings from MSA, /Ď/ is pronounced /Z/ (emphatic z) in EGY and LEV. For instance, ظابط 'police officer' is /ĎābiT/ in MSA but /ZābiT/ in EGY and LEV.
- Change in or complete drop of short vowels, e.g., يكتب 'he writes' is pronounced /yaktubu/ MSA versus /yiktib/ (EGY and IRQ) or /yuktub/ (LEV). MSA diphthongs /aw/ and /ay/ have mostly become /ō/ and /ē/, respectively.
- Predictable shortening of long vowels under certain conditions such as word-final position, loss of stress or syllabic constraints. For example, compare the following forms of the same verb (stress vowel is bolded: *šAf* /š**ā**f/ 'he saw', *šAf+hA* /š**a**fha/ 'he saw her', and *mA+šAf+hA+š* /mašafh**ā**š/ 'he did not see her'.

4.2. Morphological Variations

There are a lot of differences between MSA and DAs morphologically. Some of these differences are a result of a simplification of complex MSA paradigms. Others are the opposite: more complex structures arising in the dialects with no correlates in MSA. Some examples of the simplifying direction are the disappearance of the nominal case marking system altogether in DAs. This is an important change that has syntactic consequences. Similarly, verbal mood and voice have disappeared. It is interesting to note that the form of the indicative mood still survives as the default form in some dialects, whereas the subjunctive/jussive mood form is used in others. Other simplification phenomena include the loss of the dual form in verb conjugation in the dialects and the consolidation of feminine and masculine in the plural form.

In the rest of this section, we present some of important specific examples of morphological differences. A verbal progressive particle, which has no correspondence in MSA, appears as +ب *bi+* in EGY and LEV, as +د *da+* in IRQ and +ك *ka+* in Moroccan Arabic (MOR). The MSA future proclitic +س *sa+* is replaced by +ح *Ha+* in EGY and LEV (appearing also as +هـ *ha+* occasionally in EGY) and +غ *γa* in MOR. LEV, IRQ and GLF have a demonstrative proclitic +هـ *ha+* which strictly precedes with the definite article +ال *Al+*. Several dialects include the proclitic +ع *ςa+*, a reduced form of the preposition على *ςalaý* 'on/upon/about/to'. Also, several dialect include the non-MSA negation circum-clitic ش+ +ما *mA+ +š*.

There are also specific patterns that appear in some dialects but not in MSA, e.g., *it1a2a3* as in اتكتب *Aitkatab* 'it was written'. The form of some pronominal clitics and affix has also changed. For example MSA تُم+/ كُم+ *+tum/+kum* 'you [nominative]/[accusative]' becomes EGY توا+/كو+ *+tuwA/+kuw*. Some sub-paradigm changes also occur, e.g., MSA مدّ/مددت *mad~a/madadtu* 'he/I extended' becomes مدّ/مديت *mad~/mad~ayt* in EGY and LEV.

4.3. Lexical Variations

Lexically, the number of differences is quite significant. The following are a few examples: EGY بس *bas* 'only', طربيزة *tarabayzaħ* 'table', مرات *mirAt* 'wife [of]' and دول *dawl* 'these', correspond to MSA فقط *faqaT*, طاولة *TAwilaħ*, زوجة *zawjaħ* and هؤلاء *hawŵlA'*, respectively. For comparison, the LEV forms of the above words are *bas* (like EGY), *TAwliħ* (closer to MSA), *mart* and *hadawl*.

4.4. Orthographic Variations

Given the lack of an orthographic standard, there is a lot of orthographic variation in DA. DA writers are often inconsistent even with themselves. The differences in phonology between MSA and EGY are often responsible: words can be spelled phonologically or etymologically (using their related MSA form), e.g., كدب *kidb* or كذب *kiðb*. Furthermore, some cases of regular phonological assimilation are written to reflect their phonology or underlying morphology, EGY جنب *janb* 'side' is also written as جمب *jamb*; while the plural form جناب *jinAb* does not have a similar alternate form. Some clitics have multiple common forms, e.g., the future particle ح *Ha* appears as a separate word or as a proclitic حـ/هـ *Ha+/ha+*, reflecting different pronunciations. The different spellings may add some confusion, e.g., كتبو *ktbw* may be كتبوا *katabuwA* 'they wrote' or كتبه *katabuh* 'he wrote it'. Finally, shortened long vowels can be spelled long or short, e.g., شافها/شفها *šAf+hA/šf+hA* 'he saw her'.

5. CODA Goals and Principles

In this section, we outline CODA goals and principles and discuss some relevant practical considerations for the creation of a CODA annotated corpus.

5.1. CODA Goals

We identify 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.

5.2. CODA Design Principles

An Ad Hoc Convention CODA is an ad hoc convention. There are numerous decisions that could have been made differently especially when it comes to the phonology/orthography interface. These principles make CODA comparable to English spelling (a bit phonological, a bit historical, with some exceptions). In some cases, we followed decisions that have been made by previously published efforts.

Arabic Script CODA uses only the inventory of Arabic script characters including the diacritics used for writing MSA. CODA does not use extended Arabic characters, e.g., from Persian or Urdu. Just like MSA, CODA can be written undiacritized or diacritized.

Consistent Each DA word has a unique orthographic form in CODA that represents its phonology and morphology.

MSA-like As a general rule, CODA uses MSA-like orthographic decisions (rules, exceptions and ad hoc choices), e.g., cliticizing single letter particles, using Shadda for phonological gemination, using Ta-Marbuta and Alif-Maqsura, and spelling the definite article morphemically.

Generally Phonemic 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). Two examples of important ad hoc exceptions pertain to specific root radical letters that happen to be highly variant across dialects, e.g., ق *q*, and to long pattern vowels that can be shortened deterministically in the dialects, e.g., the pattern فواعيل *1awA2iy3*. For these cases, the word is written using the MSA cognate root radicals or pattern. The following are idiosyncratic examples from EGY:

- كتاب *kitAb* 'book' is written the same way it is in MSA since the word does not vary in phonology or morphology.
- راجل *rAjil* 'man' is not written using the MSA variant رجل *rajul*.
- اتكتب *Aitkatab* 'was written' is not written using the MSA form كتب *kutib*.
- قصر *qaSr* 'palace' is written using MSA root radicals even through it is pronounced /'aSr/ (and can be spelled more phonologically as أصر).
- طابور *TAbuwr* 'line/queue' (pronounced /Tabūr/) is written using the MSA pattern /1A2uw3/ and not as طبور *Tabuwr*.
- برتقان *burtuqAn* 'oranges' (pronounced /burtu'ān/) is the EGY word for برتقال *burtuqAl* in MSA. CODA spells this word using the MSA root radical for the q/' but not for n/l.
- مش *miš* 'not' is uniquely dialectal and is not replaced by one of its MSA equivalents: ما/ لم/ لن *mA/lm/ln*.

Morphologically Faithful CODA preserves dialectal morphology (e.g., dialectal clitics حتقول *Ha+tiquwl* 'she will say' instead of the MSA variant ستقول *sa+taquwlu*). The only exception here is separating the negation and indirect object pronouns although they are part of the word's phonological utterance, e.g., EGY ما قلت لهاش *mA qult lihAš* /ma'ultilhAš/ 'I did not tell her'.

Syntactically Faithful CODA preserves dialectal syntax, i.e., there is no change in word order.

Easily Learnable CODA is easy to learn and write. The more CODA looks like what a dialect speaker may write, the better.

Pan-Arabic but Specific Although most of the principles of CODA are the same for all DAs, each dialect will have its unique CODA Map (a list of rules and exceptions) where the relevant phonology and morphology of the dialect are outlined with the full diacritized inventory together with a list of idiosyncratic exception cases.

Easily Readable CODA is not a purely phonological representation; however, text in CODA can be read perfectly in DA given the specific dialect and its CODA Map.

5.3. Practical Considerations

The following are some consideration that arise when working on creating annotated text with CODA, where the raw text is converted manually to a CODA form to create data that can be used to train automatic CODAfication.

Code Switching Since MSA and DA coexist, we often find a lot of code switching between the two. CODA for MSA text is the accepted MSA Arabic spelling. In the example in Figure 1, a joke is set up in MSA but the punchline is in DA. DA words that are not CODA-compliant but happen to mimic MSA spelling of a cognate word in context should be changed to a CODA-compliant form. For example, in the following EGY sentence الرجل ده محترم *Alrjl dh mHtrm* 'this man is respectable', the MSA-spelled word الرجل *Alrjl* should be changed to الراجل *AlrAjl*. It may not always be easy to distinguish between the two cases. For a discussion of the issues of dialect identification, see (Habash et al., 2008; Zaidan and Callison-Burch, 2011; Diab and Elfardy, 2012).

CODA Diacritization We expect CODA to be rendered diacritized for morphological representations and be rendered undiacritized for large-scale creation of orthographically normalized training data (annotation).

Consistency CODA idiosyncratic decisions must be followed strictly. There is no room for improvisation by annotators and tool creators. New cases that are not handled can be identified and added to the CODA exception lists and CODA Map as needed.

Typographical Errors Typos such as split or merged words (e.g., يارب *yArb* instead of يا رب *yA rb* 'oh Lord'), misspelled words where some letters are missing, added or transposed (e.g., كيبر *kybr* vs. كبير *kbyr* 'big'), should be corrected as part of the annotation process. The directive is to render them in a CODA-compliant orthography.

Other Issues The data to annotate may have other types of issues due to the nature of the noisy input stream such as URLs, html markup, speech effects (such as كتيييير *ktyyyyr* 'very'), internet language, emoticons. These phenomena, though they do touch on CODA, are considered outside the scope of CODA definition. That said, these phenomena need to be handled as part of an initial preprocessing round following guidelines specific to the general task.

6. CODA Guidelines for Egyptian Arabic

In this section, we present a summary of specific CODA guidelines for EGY as an example of CODA guidelines. For the full guidelines, see (Habash et al., 2012). An example of EGY in CODA is presented in Figure 1. In the rest of this section, we consider Cairene the default EGY. Generally, EGY follows the same orthographic rules as MSA (Section 3.3.) with the following exceptions and extensions.

6.1. Phonological Exceptions

The Egyptian "Geem" The phoneme symbol /j/ and the corresponding letter ج *j* are used to represent the voiced velar stop [g] in both MSA and dialect in Egypt.

Long Vowels The long vowels /ē/ and /ō/ which do not exist in MSA are spelled as ـَي *ay* and ـَو *aw*. There are a few cases that will be ambiguous, but are all a result of MSA influences, e.g., دَولَت *dawlat* is pronounced /dawlat/ 'Dawlat' (a proper name from standard Arabic through Turkish) or /dōlat/ 'these'.

Vowel Shortening EGY long vowels shorten under certain conditions such as being word final, losing stress or being followed by two consonants. Only one long vowel is maximally allowed per word. Adding affixes and clitics changes stress patterns and interacts with vowel length. Long vowel phonemes have short allophones, but short vowel phonemes do not have long allophones. Vowel allophones involving shortening or emphasis are written phonemically, i.e., phonetically shortened long vowels are still written long, e.g., ما شافهاش *mA šAf+hA+š* /ma šafh**ā**š/ 'he did not see her'.

6.2. Phono-Lexical Exceptions

Etymologically Spelled Consonants A limited number of consonants may be spelled differently from their phonology if the following two conditions are met: (1) the consonant must be an EGY root radical and (2) the EGY root must have a cognate MSA root. If the conditions are met, then we spell the consonant using the corresponding radical from the cognate MSA root of the dialectal word's root. Only mapping into MSA *q*, *ð*, *θ* and the emphatic consonants *S*, *T*, *D* and *Ď* are allowed. These cases are chosen because they are often variable across DAs. The following are some illustrative examples:

CODA	EGY Pronunc.	Example
ق *q*	/'/	قلب *qalb* /'alb/ 'heart'
ث *θ*	/t/ or /s/	كثير *kiθiyr* /kitīr/ 'lots'
ذ *ð*	/d/ or /z/	ذلّ *ðul∼* /zull/ 'oppression'
ض *D*	/Z/ or /d/ or /z/	ضابط *DAbiT* /ZābiT/ 'officer'
ظ *Ď*	/D/ or /z/	ظلمة *Ďalmaħ* /Dalma/ 'darkness'
ص *S*	/s/	صايغ *SAyig* /sāyig/ 'jeweler'

All other phonological differences from MSA are written phonologically even though there are cases where there are shared cognates, e.g., كحك *kaHk* 'cookies' (not using the MSA form كعك *kaςk*). In some cases, some consonants are spelled etymologically but others are not, e.g., برتقان *burtuqAn* 'oranges' (not using the MSA برتقال *burtuqAl*).

MSA Pattern Vowels and Consonants A number of patterns in MSA have multiple long vowels which are not allowed in EGY. However since EGY phonology shortens some of these vowels regularly, we write the word with the MSA pattern, e.g., قانون *qAnuwn* 'law' (pronounced /qanūn/ not like MSA /qānūn/). The same principle applies to pattern consonants (except for pattern Ai1ta2a3 as in MSA), e.g., نفترض *naftariD* 'we suppose' has the inflected pattern *na+1ta2i3* and is pronounced /nafTariD/ (with the *t* becoming emphatic).

Hamza Spelling The Hamza spelling rules for EGY are the same as MSA (Section 3.3.2.), when the Hamza is pronounced in EGY. However, EGY words that have hamzated MSA cognates but no Hamza in EGY are written as pronounced in EGY, e.g., راس *rAs* 'head' (not like MSA رأس *raÂs*), بير *biyr* 'well' (not like MSA بئر *biŷr*), قرا *qarA* 'he read' (not like MSA قرأ *qaraÂa*), مايل *mAyil* 'leaning' (not like MSA مائل *mAŷil*), and ولاد *wilAd* 'children' (not like MSA أولاد *ÂawlAd*).

Alif-Maqsura The letter ى *ý* is often used in Egypt to write word-final ى/ي *ý/y* (even when writing MSA). This is not allowed in CODA. All rules for using Alif-Maqsura are the same as MSA.

6.3. Morphological Extensions

Attached Clitics EGY uses almost all the attached clitics in MSA, e.g. the definite article ال+ *Al+*. EGY also has a few additional attached clitics not in MSA, e.g., the progressive particle proclitic ب+ *bi+*, the future particle proclitic ح+ *Ha+* and the negation particle enclitic +ش *+š*. Some of the EGY pronominal enclitics have multiple contextual forms (allomorphs). Clitics are generally written in their allomorphic phonemic form, with a few exceptions, e.g., compare شافك *šAf+ik* 'he saw you', and شافوكي *šAfuw+kiy* 'they saw you'; and شفته *šuft+uh* 'I saw him', شافوه *šAfuw+h* 'they saw him'; and ما شافوهوش *mA šA-*

fuw+huw+š 'they didn't see him'. The Shadda rule is disabled across stem-clitic boundaries (except for ي+ *+ya*), e.g., واحشيننا *wAHšiynnA* 'we miss you' and عليّ *ςalay∼a*.

Separated Clitics The indirect object enclitics and the negation proclitic ما *mA* are written separately, e.g., ما قال ليش *mA qAl liyš* /ma+'al+lī+š/ 'he did not tell me'.

Ta-Marbuta The Ta-Marbuta has four forms in EGY. Three are similar to MSA: word-final non-construct ـة *aħ* /a/, word-final construct ـِة *iħ* /it/, and word-medial construct ـِت *it* /it/, e.g., عجلة *ςajalaħ* 'bicycle', عجلة *ςajaliħ* 'bicycle of', and عجلتها *ςajalithA* 'her bicycle'. The fourth case is dialectal: word-medial non-construct ـا *A* /ā/, e.g., دارسة الكتاب *dArsaħ AlkitAb* 'she studied the book' / دارساه *dArsA+h* 'she studied it'.

Waw of Plurality The silent Alif added at the end of the 3rd person plural affix وا+ *+uwA* in MSA is also used in EGY. It is also added to the 2nd person plural affix توا+ *+tuwA* which is not in MSA. This silent Alif is not added to the pronominal enclitic كو *kuw* 'your/you [acc]' or the pronoun انتو *Aintuw* 'you [nom]'.

Nunation Nunation has disappeared from EGY as a productive inflection. It remains as an adverbial derivational morpheme, e.g., عمليًا *ςamaliy∼Aã* 'practically'.

6.4. Lexical Exceptions

EGY CODA guidelines include a word list specifying ad hoc spellings of EGY words that may be inconsistent with the default mapping outlined above or that have multiple commonly used spellings. Examples include pronouns such as انتو *Aintuw* 'you' (not انتوا *AintuwA*), demonstratives such as ده *dah* 'this' (not دا *dA*), limited cliticizations such as داحنا *d+AHnA* 'so+we' (not دحنا *daHnA*), adverbs such as برضه *barDuh* 'also' (not برده *barduh*, or برضو *barDaw*), and special partly ambiguous cases such as the existential فيه *fiyh* 'there is' and its negative مفيش *mafiyš* 'there is not' contrasted with the closely related preposition+pronoun فيه *fiyh* 'in it' and its negative ما فيهوش *mA fiyhuwš* 'not in it'.

7. Future Directions

More Details, More Dialects We plan on continuously improving the CODA guidelines. There will naturally be additional issues to address in EGY. We are also working on developing CODA guidelines for other dialects.

Resources In terms of developing resources that are annotated for CODA, we will use a graphical user interface tool developed under the COLABA project for annotation (Benajiba and Diab, 2010) to annotate a large collection of EGY text.[3]

Tools We plan to use the annotations we will create to develop automatic CODAfication tools that can be used as part of general preprocessing of DA data for a variety of NLP applications. For a published early attempt at this effort, see (Dasigi and Diab, 2011).

Acknowledgments

We would like to thank Abdelati Hawwari, Sondos Krouna, Mohamed Maamouri, Reem Faraj, Ahmed El Kholy and Ramy Eskander for helpful feedback. This work was partially funded under DARPA project number HR0011-12-C-0014.

8. References

Ernest T. Abdel-Massih, Zaki N. Abdel-Malek, and El-Said M. Badawi. 1979. *A Reference Grammar of Egyptian Arabic*. Georgetown University Press.

Abbas Al-Tonsi and Laila Al-Sawi. 1990. *An Intensive Course in Egyptial Colloquial Arabic*. American University in Cairo.

Plonka Arkadiusz. 2006. Le nationalisme linguistique au liban autour de sa'id 'aql et l'idée de langue libanaise dans la revue "lebnaan" en nouvel alphabet. *Arabica*, 53(4):423–471.

Khalil 'Asaakir. 1950. A Method for Writing Modern Arabic Dialects with Arabic Letters. (in Arabic). *The Arab Academy Magazine*, 8(181–192), January.

El-Said Badawi and Martin Hinds. 1986. *A Dictionary of Egyptian Arabic*. Librairie du Liban.

Yassine Benajiba and Mona Diab. 2010. A web application for dialectal arabic text annotation. In *Proceedings of the LREC Workshop for Language Resources (LRs) and Human Language Technologies (HLT) for Semitic Languages: Status, Updates, and Prospects*.

Fadi Biadsy, Nizar Habash, and Julia Hirschberg. 2009. Improving the Arabic Pronunciation Dictionary for Phone and Word Recognition with Linguistically-Based Pronunciation Rules. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 397–405, Boulder, Colorado.

Tim Buckwalter. 2007. Issues in Arabic Morphological Analysis. In A. van den Bosch and A. Soudi, editors, *Arabic Computational Morphology: Knowledge-based and Empirical Methods*. Springer.

Mark W. Cowell. 1964. *A Reference Grammar of Syrian Arabic*. Georgetown University Press.

Pradeep Dasigi and Mona Diab. 2011. Codact: Towards identifying orthographic variants in dialectal arabic. In *Proceedings of the International Joint Conference on Natural Language Processing*, Chiang Mai, Thailand.

[3] We conducted a preliminary annotation experiment using four annotators trained with an earlier version of the guidelines. We do not expect major differences in the statistics we report here. The annotation covered over 110K words. 15.1% of all words were changed. 12.1% of the changes involved a merge action (removing incorrect space between two words) and 13.2% involved a split action (adding a space to separate two incorrectly attached words). Among character substitutions, changing the Alif form into one of its variants is the most common change (22.1%) followed by cases involving the Ta-Marbuta (14.4%) and Alif-Maqsura/Ya (8.5%); these are expected results given Arabic orthography (Habash, 2010). Among the less common but interesting cases linguistically, we find that 1.7% of the words have a ت *t* → ث *θ* change and 0.8% of all words have a change involving the letter ق *q*. Inter-annotator agreement is about 98%.

Raw Text	هه هه واللة العظيم فطست من الضحك اية ياعبير المواضيع الجامدة دى دحنا كدة بقى عندنا بنك كامل متكامل على العموم انا اجتهدت وجبت شوية نكت واكيد طبعا منقولين بس يارب يعجبوكوا اسيبكوا مع النكت . هنا رقد الرجل على فراشة يغالب الغيبوبة وكلما افاق وجد زوجتة بجانبة وتنظر الية بحنان فامسك بيديها قائلا : لما اترفدت وقفتى معايا . ولما شركتى فلست كنتى جمبى . ولما بيتنا اتحرق كنتى جمبى . ودلوقتى انتى بردة جمبى . مش عارف لية انا حاشش انك نحس *hh hh wAllħ AlςĎym fTst mn AlDHk Ayħ yAςbyr AlmwADyς AljAmdħ dý dHnA kdħ bqý ςndnA bnk kAml mtkAml ςlý Alςmwm AnA Ajthdt wjbt šwyħ nkt wAkyd TbςA mnqwlyn bs yArb yςjbwkwA AsybkwA mς Alnkt . hnA rqd Alrjl ςlý frAšħ yγAlb Alγybwbħ wklmA AfAq wjd zwjtħ bjAnbħ wtnĎr Alyħ bHnAn fAmsk bydyhA qAŷlA : lmA Atrfdt wqftý mςAyA . wlmA šrktý flst kntý jmbý . wlmA bytnA AtHrq kntý jmbý . wdlwqtý Antý brdħ jmbý . mš ςArf lyħ AnA HAšš Ank nHs*
CODA	هه هه والله العظيم فطست من الضحك إيه يا عبير المواضيع الجامدة دي داحنا كده بقى عندنا بنك كامل متكامل على العموم انا اجتهدت وجبت شوية نكت وأكيد طبعا منقولين بس يا رب يعجبوكو اسيبكو مع النكت . هنا رقد الرجل على فراشه يغالب الغيبوبة وكلما أفاق وجد زوجته بجانبه وتنظر إليه بحنان فأمسك بيديها قائلا : لما اترفدت وقفتي معاي . ولما شركتي فلست كنتي جنبي . ولما بيتنا اتحرق كنتي جنبي . ودلوقتي انتي برضه جنبي . مش عارف ليه انا حاسس انك نحس *hh hh* **wAllh** *AlςĎym fTst mn AlDHk* **Ǎyh yA ςbyr** *AlmwADyς AljAmdħ* **dy dAHnA kdh** *bqý ςndnA bnk kAml mtkAml ςlý Alςmwm AnA Ajthdt wjbt šwyħ nkt* **wÂkyd** *TbςA mnqwlyn bs* **yA rb yςjbwkw Asybkw** *mς Alnkt . hnA rqd Alrjl ςlý* **frAšh** *yγAlb Alγybwbħ wklmA* **ÂfAq wjd zwjth bjAnbh** *wtnĎr* **Ǎlyh** *bHnAn* **fÂmsk** *bydyhA qAŷlA : lmA Atrfdt* **wqfty mςAy** . *wlmA* **šrkty** *flst* **knty jnby** . *wlmA bytnA AtHrq* **knty jnby** . **wdlwqty Anty brDh jnby** . *mš ςArf* **lyh** *AnA* **HAss** *Ank nHs*
English	ha ha [,] I swear to God [,] I died from laughter [,] Abeer [,] what cool topics [!] we now have a complete comprehensive bank [.] any way [,] I put some effort and got some jokes that are of course copied [,] but hopefully you will like them [.] I leave you with the jokes . [MSA] There lied a man on his bed coming in and out of a coma [;] and every time he woke up he found his wife by his side looking at him lovingly [.] so he held her hands and said [/MSA]: when I got fired [,] you stood by me . And when my company went bankrupt you were by my side . And when our house burnt down you were by my side . And now also you are by my side . I don't know why I feel you're bad luck [.]

Figure 1: An Egyptian Arabic snippet in raw and CODA orthography. Bracketed punctuation and comments are added in the English translation to help the reader. The region between [MSA] and [/MSA] is in MSA. Bolding in the CODA row marks modified words.

Mona Diab and Heba Elfardy. 2012. Simplified guidelines for the creation of large scale dialectal arabic annotations. In *Language Resources and Evaluation Conference (LREC)*, Istanbul.

Mona Diab, Nizar Habash, Owen Rambow, Mohamed Al-Tantawy, and Yassine Benajiba. 2010. Colaba: Arabic dialect annotation and processing. In *Proceedings of the LREC Workshop for Language Resources (LRs) and Human Language Technologies (HLT) for Semitic Languages: Status, Updates, and Prospects*.

Y. A. El-Imam. 2004. Phonetization of Arabic: Rules and Algorithms. In *Computer Speech and Language 18*, pages 339–373.

Wallace Erwin. 1963. *A Short Reference Grammar of Iraqi Arabic*. Georgetown University Press.

Nizar Habash and Owen Rambow. 2007. Morphophonemic and Orthographic Rules in a Multi- Dialectal Morphological Analyzer and Generator for Arabic Verbs. In *International Symposium on Computer and Arabic Language (ISCAL)*, Riyadh, Saudi Arabia.

Nizar Habash, Abdelhadi Soudi, and Tim Buckwalter. 2007. On Arabic Transliteration. In A. van den Bosch and A. Soudi, editors, *Arabic Computational Morphology: Knowledge-based and Empirical Methods*. Springer.

Nizar Habash, Owen Rambow, Mona Diab, and Reem Kanjawi-Faraj. 2008. Guidelines for Annotation of Arabic Dialectness. In *Proceedings of the LREC Workshop on HLT & NLP within the Arabic world*.

Nizar Habash, Mona Diab, and Owen Rambow. 2012. Conventional Orthography for Dialectal Arabic: Principles and Guidelines – Egyptian Arabic. Technical Report CCLS-12-02, Columbia University Center for Computational Learning Systems.

Nizar Habash. 2006. On Arabic and its Dialects. *Multilingual Magazine*, 17(81).

Nizar Habash. 2010. *Introduction to Arabic Natural Language Processing*. Morgan & Claypool Publishers.

Richard Harrell. 1962. *A Short Reference Grammar of Moroccan Arabic*. Georgetown University Press.

Clive Holes. 2004. *Modern Arabic: Structures, Functions, and Varieties*. Georgetown Classics in Arabic Language and Linguistics. Georgetown University Press.

H. Kilany, H. Gadalla, H. Arram, A. Yacoub, A. El-Habashi, and C. McLemore. 2002. Egyptian Colloquial Arabic Lexicon. LDC catalog number LDC99L22.

Eugene E. Loos, Susan Anderson, Jr. Dwight H., Day, Paul C. Jordan, and J. Douglas Wingate. 2004. Glossary of Linguistic Terms.

Mohamed Maamouri, Tim Buckwalter, and Christopher Cieri. 2004. Dialectal Arabic Telephone Speech Corpus: Principles, Tool design, and Transcription Conventions. In *NEMLAR International Conference on Arabic Language Resources and Tools*.

Omar F. Zaidan and Chris Callison-Burch. 2011. The arabic online commentary dataset: an annotated dataset of informal arabic with high dialectal content. In *Proceedings of the Association for Computational Linguistics*, Portland, Oregon, USA.