TECHNICAL REPORT

No. CCLS-14-01

Title: Conventional Orthography for Dialectal Arabic (CODA): Principles and Guidelines — Egyptian Arabic - Version 0.7 – March 2012

Authors: Nizar Habash, Mona Diab and Owen Rambow

Conventional Orthography for Dialectal Arabic: Principles and Guidelines – Egyptian Arabic

Nizar Habash, Mona Diab and Owen Rambow
cadim@ccls.columbia.edu
Version 0.7 - March 2012

We thank Abdelati Hawwari, Sondos Krouna and Mohamed Maamouri for helpful feedback

---

This document introduces CODA (Conventional Orthography for Dialectal Arabic) and presents specifications and detailed guidelines for Egyptian Arabic CODA. CODA addresses the problem of inconsistent orthographic choices in raw (naturally occurring) written dialectal Arabic text. The specifications are a succinct summary, while the guidelines contain details and examples. The document has three parts that are ordered from most general to the more specific. In Part 1, we define CODA and present its general goals, principles and considerations in a non-dialect specific manner. In Part 2, we present a high level CODA specification for Egyptian Arabic (EGY). And in Part 3, we present detailed guidelines for EGY CODA.

Important Note on How to Read this Document

This document is intended only as a guideline to the orthographic conventions of full orthographic words -- not tokenized or segmented words, not morphologically segmented/ tokenized text and not part-of-speech tags. For tokenization and POS tagging guidelines, see the Linguistic Data Consortium (LDC) POS Guidelines for Egyptian Arabic. As a guide to the orthographic convention that will be used at the LDC and Columbia University for annotation and tool creation, this convention must be followed strictly. Any issues or confusion should brought to the attention of the authors immediately. Variations in spelling that often occur in raw data (i.e. naturally occurring data) are not tolerated within CODA. As an illustrative example, the raw word (ktbw,كتبو) will be rewritten in CODA by CODA annotators as (ktbwA,كتبوا) `they wrote' or (ktbh,كتبه) `he wrote it' or `his books' depending on the context it appears in.

As part of the automatic processing tools, the conversion from the raw word into its CODA form will be done automatically. Subsequently, the morphological analyzer will further determine the possible POS tags and possible tokenization (katab/PV+uwA/PVSuff:3P or kutub/NOUN+uh/ POSS_3MS or katab/PV+(null)/PVSuff:3MS+uh/DO_3MS). The morphological disambiguation tools and treebank annotators will then pick the correct reading/analysis in context.

PART 1: General Principles

1.1 What is CODA?

Unlike Modern Standard Arabic (MSA), Arabic dialects, henceforth Dialectal Arabic (DA), have no standard published orthographies since there are no DA academies nor is there a large edited body of dialectal literature that follows the same spelling standard. CODA is a conventional orthography for Arabic dialects that aims at filling this gap. It is designed for the purpose of developing computational models of DA.

1.2 CODA Goals

1. CODA is an internally consistent and coherent convention for writing DA.
2. CODA is created for computational purposes.
3. CODA uses the Arabic script.
4. CODA is a unified framework for writing all DAs.
5. CODA aims to strike an optimal balance by maintaining a level of dialectal uniqueness yet establish conventions based on MSA-DA similarities.

1.3 CODA Design Principles

1. 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.[1]
2. 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. CODA can be written undiacritized or diacritized.
3. Each DA word has a unique orthographic form in CODA that represents its phonology, morphology, and lexical semantics [meaning].
4. 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, Alif Maqsura, silent Alif in Waw-Alif of plurality, and spelling the definite article Al morphemically.
5. 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 important ad hoc exceptions** pertain to specific root radical letters that happen to be highly variant across dialects, e.g. ق، ث، ذ، ظ، ج , etc. and to long pattern vowels that can be shortened deterministically in the dialects, e.g., the pattern 1awA2iy3 فواعيل. For these cases, the word is written using the MSA cognate root radicals or pattern. The following are iconic examples from EGY:

---

[1] E. Badawi & M. Hinds. A Dictionary of Egyptian Arabic. Librairie Du Liban. 1986.
A. Al-Tonsi & L. Al-Sawi. An Intensive Course in Egyptian Colloquial Arabic. American University in Cairo. 1990.
H. Kilany, H. Gadalla, H. Arram, A. Yacoub, A. El-Habashi, and C. McLemore. Egyptian Colloquial Arabic Lexicon. LDC catalog number LDC99L22, 2002.

a. راجِل rAjil ‘man’ (not using the MSA variant رَجُل rajul)
b. اِتكَتَب Aitkatab ‘was written’ (not the MSA كُتِب kutib)
c. قَصْر qaSr ‘palace’ using MSA root radicals even through it is pronounced /’aSr/.
d. طابور TAbuwr ‘line/queue’ (pronounced /Tabu:r/) is written using the MSA pattern /1A2u:3/ and not as طبور Tabuwr.
e. بُرتُقان burtuqAn ‘oranges’ (pronounced / burtu’An/) using MSA root radical for the q/’ but not for n/l.
f. مِش mi$ ‘not’ is uniquely dialectal and is not replaced by MSA مالمالن.

6. CODA preserves dialectal **morphology** (e.g., dialectal clitics حتقول instead of ستقول). The **only exception** here is separating the negation and indirect object pronouns although they are part of the word: e.g., EGY ما قلت لهاش mA qult lihA$ /ma’ultilhA$/ ‘I did not tell her’.
7. CODA preserves dialectal **syntax**, i.e. there is no change in word order.
8. CODA is easy to learn and **write** to achieve high inter-annotator agreement; the more CODA looks like what a dialect speaker may write, the better.
9. 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.
10. 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.

1.4 Practical Considerations

1. Basic CODA issues:
   a. CODA for MSA text is the accepted MSA Arabic spelling.
   b. CODA can be rendered in both Arabic script and/or BW transliteration as needed.
   c. CODA is rendered diacritized for morphological representations and is rendered undiacritized for large-scale creation of orthographic normalization training data (annotation).
   d. CODA idiosyncratic decisions must be followed strictly. There is no room for improvisation by annotators. New cases that are not handled can be identified and added to the CODA map as needed.
   e. Dialectal Arabic 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, (الرجل ده مش محترم خالص) should be changed to (الراجل ده مش محترم خالص).
2. Non-CODA issues:
   a. The data to annotate (i.e., convert from raw form to CODA) will have other types of issues due to the nature of the noisy input stream such as URLs, html markup, speech effects, internet language, emoticons. These phenomena, though they do touch on CODA, are considered outside the scope of this document but will be handled as part of an initial preprocessing round and will follow guidelines to be specified by the LDC.
   b. Typographical errors such as split/merge words (انتهنا vs. انتهنا), misspelled words where some letters are missing, added or transposed (such as كيير vs. كبير or كثيييير vs. كثير), will be handled in the annotation process. The directive is to render them in a CODA compliant orthography.

1.5 CODA Guidelines and the LDC Guidelines

These guidelines are inspired by the LDC guidelines for transcribing Levantine and Iraqi Arabic (Buckwalter and Maamouri, 2004). They differ from them in the following ways:

1. Whereas the LDC guidelines are for transcription, and thus focus more on phonological variants of sub-dialect in some cases, CODA is intended for general purpose writing in a way that abstracts from these variations when possible.
2. CODA is designed as a pan-Arabic convention. We extend the LDC guidelines to cover Egyptian Arabic (we profit from the LDC work on Call Home Egyptian).
3. CODA fills in some issues not addressed in the LDC guidelines, specifically related to morphology.

PART 2: Egyptian Arabic CODA Specifications

We present here a high-level specification of the Egyptian Arabic CODA map. This is intended as a complete summary of all decisions on Egyptian CODA.

We start with a few notes on Egyptian phonology, highlighting differences from MSA phonology. These notes are not a complete description, but address relevant issues in this section. The core of the CODA map is the default phonology-to-orthography mapping (section 2.2). There are three types of exceptions to this default: phono-lexical (section 2.3), morphological (2.4) and lexical exceptions (2.5). Section 2.6 presents a short guide on how to apply these rules and exceptions.

In all examples below, the orthography is represented inside parentheses (). Phonology is represented inside forward slashes //. Phonetic form is represented inside square brackets []. Buckwalter transliteration is used for the orthography. Phonology is represented using a Buckwalter-like set of symbols whose meaning is determined by their MSA pronunciation, with the following exceptions: long vowels are represented using the short vowels followed by ':', the symbol [G] is used for voiced /k/, Hamza is represented only as /'/. Hamza variants, Alif Maqsura, Ta Marbuta and Alif are not used in the phonological/phonetic representation. /w/ and /y/ are the glides and not the vowels.

2.1 A few Notes on Egyptian Arabic Phonology

These notes are relevant to the discussion in this section and are not meant to be comprehensive. We are considering here that Cairene phonology is the default EGY.

1. We use the following phoneme symbol exceptionally for Egyptian
    a. /j/ is pronounced [G] (voiced [k])

2. Egyptian vowel phonology rules include
    a. Word final long vowels typically shorten (unless foreign).
    b. One long vowel maximally allowed per word.
    c. Unstressed vowels cannot be long.
    d. Adding affixes and clitics changes stress patterns and interacts with vowel length.
    e. Long vowel phonemes have short allophones, but short vowel phonemes do not

have long allophones.

f. Elided/epenthesized vowels are considered part of the phonemic form of the word.
g. The short vowel /i/ has two allophones [i] and [e].
h. The short vowel /u/ has two allophones [u] and [o].

2.2 Default Phonology-to-Orthography Mapping

1. The following is a pairing of phonological and orthographic symbols (phonemes and letters+diacritics):

a. Consonants:

The following set is an identity map:

/b,t,v,j,H,x,d,*,r,z,s,$,S,D,T,Z,E,g,f,q,k,l,m,n,h,w,y/
=> (b,t,v,j,H,x,d,*,r,z,s,$,S, D,T,Z,E,g,f,q,k,l,m,n,h,w,y)
=> (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)
(for Hamza see below)

b. Vowels

/a,i,u/ => (a, u, i) (-َ, -ُ, -ِ)
/a:,i:,u:, e:, o:/ => (A, iy, uw, ay, aw) (ـَا, ـِي , ـُو, ـِي, ـَو)
/aw, ay/ => (aw, ay) (ـَو, ـَي)

The sukun symbol is not used in this document. It is deterministically added word-medially after consonants not followed by vowels.

Vowel allophones involving shortening or emphasis are written phonemically, i.e., phonetically shortened long vowels are still written long

e.g.,	/kalb/	=> (kalb) (كَلب)	
	/kalbe:n/	=> (kalbayn) (كَلبين)	
	/$Af/	=> ($Af) (شاف)	
	/$AfhA/ [$afha]	=> ($AfhA) (شافْها)	
	/$u:f/	=> ($uwf) (شُوف)	
	/$o:f/	=> ($awf) (شَوف)	
	/ro:H/	=> (rawH) (رَوح)	
	/ru:H/	=> (ruwH) (رُوح)	
	/rawwaH/	=> (raw~aH) (رَوَّح)	
	/kaHk/	=> (kaHk) (كَحْك)	
	/dawA/ [dawa]	=> (dawA) (دَوَا)	
	/dawlat/	=> (dawlat)	‘Dawlat’ * (دَولَت)
	/do:lat/	=> (dawlat)	‘these’ *(دَولَت)
	/dawlit/	=> (dawlip)	‘state [construct]’(دَولَة)

* Note: these are ambiguous cases.

2. A bare Alif (unhamzated) is added at the beginning of words starting with vowels: this is the so-called “Hamzat Wasl”, which is a Hamzated allophone of the vowel phoneme in utterance initial contexts. To determine if the word has a real Hamza phoneme, add the

clitic “w” و or “b” بـ before it: real Hamza remains, Hamzat Wasl disappears.

e.g., /aktib/ (أكتِب) => (Aaktib) /baktib/ NOT /bi’aktib/ (باكْتِب)
/ilfikr/ (الفِكر) => (Alfikr) /wilfikr/ NOT /wi’ilfikr/ (والفِكر)
/’aHla:m/ (أحلام) => (>aHlAm) /wi’aHla:m/ NOT /wiHla:m/ (وأحلام)

3. The Hamza grapheme has many forms that are determined by the vowel context. The rules for choosing the correct form are the same as in MSA.

/’/ => (>,<,|,&,},’) (أ،إ،آ،ؤ،ئ،ء)

e.g., /fu’a:d/ => (fu&Ad) (فُؤَاد)
/’a:nis/ => (|nis) (آنِس)

4. The Shadda symbol replaces the second letter in a repeated letter sequence:
e.g., /kallim/ => (kal~im)(كَلِّم)

5. Alif Maqsura spelling: The ambiguous use of Y to mean y/Y is not CODA. A word final /a,a:/ vowel is spelled as Y following the same rules as MSA: the word has a radical y that changed to /a/ or it has a complex pattern (not made up from root and short vowels only). We recognize that this is a morpho-phonemic decision, we included it because of its pervasiveness.

6. Ta-Marbuta is spelled (ap) in non-construct cases and (ip) in construct cases (word final) or (it) (construct word medial) or (A) in non-construct word medial cases:
i.(Eajalap) vs. (Eajalip AHmad) vs. (Eajalituh)
(عَجَلَة) vs. (عَجَلِة أحمد) vs. (عَجَلِتُه)
ii.(xATbap Albint) vs. (xATbAhA)
(خاطْبَة الْبِنْت) vs. (خاطْباها)
We recognize that this is a morpho-phonemic decision, we included it because of its pervasiveness.

7. Minor emphatic cases that are accounted for in the literature but have no Arabic script representation will be treated as lexical exceptions, e.g., /<u>b</u>a:<u>b</u>a/ (bAbA) (بابا) ‘daddy’. The symbol /<u>b</u>/ is an emphatic /b/.

8. Foreign words ending with long vowels will be rendered with an extra silent (h) word finally:

e.g., [$ale:] => /$a:le:/ => ($Alayh) (شالَيه) ‘chalet’
[mayo:] => /ma:yo:/ => (mAyawh) (مايَوه) ‘swimsuit’

2.3 Phono-Lexical Exceptions

1. The following 9 DA consonants may be spelled differently from their phonology. If the following two conditions are met:

   - the consonant must be a DA root radical and
   - the DA root must have a cognate MSA root,

then spell the consonant using the corresponding radical from the cognate MSA root of the

dialectal word's root. Only these mappings are allowed.

/'/	=>	(q) or (Hamza)	(ق)	or	(همزة)
/t/	=>	(v) or (T) or (t)			(ث) or (ط) or (ت)
/s/	=>	(v) or (S) or (s)			(ث) or (ص) or (س)
/z/	=>	(*) or (Z) or (D) or (z)			(ذ) or (ظ) or (ض) or (ز)
/d/	=>	(*) or (D) or (Z) or (d)			(ذ) or (ض) or (ظ) or (د)
/S/	=>	(S) or (s)			(ص) or (س)
/T/	=>	(T) or (t)			(ط) or (ت)
/D/	=>	(D) or (Z) or (d)			(ض) or (ظ) or (د)
/Z/	=>	(Z) or (D) or (z)			(ظ) or (ض) or (ز)

e.g., /'alb/ => (qalb) (قَلْب) NOT (>alb) (ألب)
/'alam/ => (qalam) (قَلَم) 'pen' or (>alam) 'pain' (أَلَم)
/sawa:b/ => (vawAb) (ثَواب) NOT (sawAb) (سواب)
/kiti:r/ => (kiviyr) (كِثِير) NOT (kitiyr) (كتير)
/kidb/ => (ki*b)(كِذْب) NOT (kidb) (كدب)
/zall/ => (*al~) (ذَل) NOT (zul~) (زل)
/Dalma/ => (Zalmap) (ظَلْمة) NOT (Dalmap) (ضلمة)
/Za:biT/ => (DAbiT) (ضابِط) NOT (ZAbiT) (ظابط)
/samg/ => (Samg) (صَمْغ) NOT (samg) (سمغ)

The same principle applies to pattern letters (except pattern Ai1ta2a3) as in MSA:

/nafTariD/ => (naftariD) (نَفْتَرِض) NOT (nafTariD) (نفطرض)
/iSTaETa/ => (AistaETaY) (استعطى) NOT (AiSTaETaT) (اصطعطى)
/izdahar/ => (Aizdahar) (ازدهر) NOT (Aiztahar) (ازتهر)

Note that all other phonological differences from MSA are written phonologically (even though there are cases where there are shared cognates), most notably the cases corresponding to Hamza changes:

/bi:r/ => (biyr) (بِير) NOT (bi}r) (بِئْر)
/sAyil/ => (sAyil) (سايِل) NOT (sA}il) (سائِل)
/burtu'a:n/ => (burtuqAn) (بُرْتَقَان) NOT (burtuqAl) (بُرْتَقَال)
/$aHat/ => ($aHat) (شحت) NOT ($aHa*) (شحذ)
[In Egyptian only. Levantine is different since the word is pronounced /$aHad/]

Beware of Hamza confusion (from q ق or Hamza):

/rAyi'/ => (rAyiq) (رايق) NOT (rA}iq) (رائِق)
/'ara''/ => (>araq~) (أَرَق) `finer/finest; thiner/thinest'
(not to be confused with /'araq/ => (>araq) (أَرَق) `insomnia'
which is pronounced as in MSA)
/'ala'/ => (qalaq) (قَلَق)

2. One exception to Hamza spelling is the case of an MSA Hamza written as Alif-Hamza-Above that turned into an /a:/ at the end of a word in the dialect. CODA will write the final letter as Alif not Alif Maqsura. This can be thought of as an extension to the rule about root-influenced spelling of final /a:/ which in MSA covers only w/y root radicals.

/bada, yibda/ => (badA, yibdA) (بدا، يبدا) NOT (badaY, yibdaY) (بدى، يبدى)
/ibtada / => (AibtabadA) (ابتدا) NOT (AibtadaY) (ابتدى)

3. A number of patterns in MSA have multiple long vowels which are not allowed in Egyptian. However since Egyptian phonology shortens some of these vowels regularly, we write the word with the MSA pattern:

   e.g., [qanu:n] => /qa:nu:n/ => (qAnuwn) (قَانون) NOT (qanuwn) (قَنون)
   [surya] => /su:rya:/ => (suwryA) (سُوريا) NOT (surya) (سُرْيا)

4. A number of foreign words in MSA have multiple long vowels which are not allowed in Egyptian. We write the words using the MSA long vowels as in 2 above:

   e.g., [kumbiyu:tar] => /ku:mbi:yu:tar/ => (kuwmbiyuwtar)(كُومبِيُوتَر)
   NOT (kumbyuwtar) (كُمبيُوتَر)

2.4 Morphological Exceptions

1. All affixes and clitics are added to the word without changing the spelling of the stem.

   /bilmaka:tib/ => /bi+Al+makAtib/ => (biAlmakAtib) (بِالمَكاتِب)

2. The Shadda rule is disabled across stem-clitic boundaries (except for +ya): (bArik+kum) (بارِكْكُم), (wAH$iyn+nA) (واحشينا), (xaz~An+nA) (خزانا) but (Ealy+~a) (عَلي)

3. Indirect object clitics and the negation (mA) clitic are written separately:
   /wima'alli:$/ => (wimA qAl liy$) (وِما قال لِيش) NOT (wimaqal~iy$) (وِمَقَّلِّيش)

4. All afixes and clitics are written in their allomorphic phonemic form. We include tables that list all conditions for reference. E.g.

   PRON_2FS /ik/ => /ik/,/ki/,/ki:/, /iki:/
   /$Afik/ => ($Afik) (شافِك)
   /$Afu:ki/ => ($Afuwkiy) (شافُوكِي)
   /$Afuki:$/ => ($Afuwkiy$) (شافُوكِيش)
   /$uftiki:$/ => ($uftikiy$) (شُفْتِكِيش)

5. Exceptions to the allomorphic phonemic spelling are:

   a. subject-3P /u:/ in word final positions is (uwA) (وا)
   b. subject-2P /tu:/ in word final positions is (tuwA) (تُوا)
   c. clitic PRON-2P /ku:/ in word final positions is (kuw) (كو)
   d. Al (the definite article) is spelled morphemically: i.e., spell sun/moon letters as is done in MSA and consider the /i/ vowel in it epenthetic:
      /i$$ams/ => (Al$~ams) (الشَّمْس) NOT (Ail$ams) (اِلشَمس) and NOT (A$$ams) (اِشَّمْس)
   e. PRON_3MS, which has numerous forms is spelled morphemically in some cases (two forms instead of four):
      i. /$a:fu/ => ($Afuh) (شافُه)
      ii. /$afu:/ => ($Afuwh) (شَافُوه)

iii. /$afu:$/ => ($Afhuw$) (شَافهُوش)
iv. /$afuhu:$/ => ($Afuwhuw$) (شافُوهُوش)

f. (li+Al) => (lil) (للـ)

6. The Alif Maqsura and Ta-Marbuta are discussed in section 2.2.5 and 2.2.6.

2.5 Lexical Exceptions

This section will include words that are spelled in a particular *ad hoc* way:

1. MSA ad hoc spelling:

   e.g., proper names like /Eamr/ => (Eamrw) (عَمْرو) `Amr'

2. Egyptian word lists specify all types of Egyptian only words that may be spelled in ways inconsistent with the default phonology-orthography mapping. Some of these words have multiple variant sub-dialectal pronunciations. We consider Cairene the default. The full list of exceptions is provided in Part 3.

   e.g., /intu/ => (Aintuw) (إنْتو) NOT (AintuwA) (إنْتوا)
   /barduh/ => (barDuh) (بَرْضُه) NOT (barDuw) (بَرضو)
   /dah/ => (dah) (دَه) NOT (*ah) (دَه)

2.6 How to Use a CODA Map?

The CODA map assumes the CODA writer knows the input word's phonology, morphology and meaning, although she may not actually use all of this information in writing the word. For example, a word that triggers no exceptions will simply be written using the direct phonology-to-orthography mapping: the word (maktuwb) (مكتوب) has a couple of meanings, but this is irrelevant to how it is written.

The following is a possible order of applying the rules specified in section 2.2-2.5:

1. Determine the word's phonology, morphology and meaning.
2. Identify the affixes and clitics and write them according to the morphological exceptions (section 2.4).
3. Write the rest of the word according to the phonological-orthographic (2.4) default rules unless it triggers a lexical exception (2.5) or a phono-lexical exception (2.3).

Examples:

/wilkatba/ 'and the writer [fs]' (1)
=> wi+Al+ … +ap (2)
=> kAtb (3)
=> wiAlkAtbap (والكاتْبة)

/wima$afuhA$/ 'and they did not see it' (1)
=> wi+mA_ … +uw+hA+$ (2)
=> $Af (3)
=> wimA $AfuwhA$ (وما شافُوهاش)

/qanunhA/	‘her law’	(1)
=>	….. +hA	(2)
=>	qAnuwn	(3)
=>	qAnuwnhA (قَانُونها)	

PART 3: Egyptian Arabic CODA Detailed Guidelines

In the rest of this document the color red is used to mark a non-CODA compliant spelling that needs to be addressed. And green is used to highlight some corrections to render it CODA compliant.

The guidelines in this section are for undiacritized CODA and they are intended for large-scale annotation to produce a CODA compliant corpus.

We assume the Cairene accent in our CODA Map for Egyptian Arabic.

The guidelines will not refer to particles using their Egyptian POS (as specified by the LDC POS guidelines for Egyptian Arabic) unless necessary.

3.1 General Spelling

1. Correct all typos. These include missing, added, repeated, and transposed letters; and merged words and split words. Beware of letters that do not connect as they may appear unconnected.
   - رئيس الوزرا -> رئيس الورزا
   - رئيسالوزرا -> رئيس الوزرا
   - الفلسطينين -> الفلسطينين
   - الفلسطينين -> الفلسط ينين
   - الفلسطينين -> الفلسطبنين
   - يا حلاوة يا ولاد -> يا حلاوى يا ولاد
   - محمدعبدالغفارمحمد -> محمد عبد الغفار محمد

2. Do not change punctuation by adding/deleting or replacing any punctuation marks.

3.2 Letters and Sounds

1. **Diacritics (الشكل)**: ignore all diacritics, including dialectal nunation (see last three examples below). All dialectal variations in short/no vowels (that are never long under any context) are ignored. Do not add diacritics or delete existing ones.

   - بيكتب
   - بلاد
   - مثلا -> مثلن
   - غصب -> غصبن
   - على كل -> على كلن

2. **Basic Spelling**

   a. The general rule is to write words as they are pronounced unless there are morphological/phonological exceptions (section 3.3) or dialectal lexical exceptions (section 3.4).

b. The rest of this section lists the exceptions to this simple rule.

3. **Consonant-Consonant difference**

If a word's root is a cognate of an MSA root, then the root radicals are written using the corresponding MSA root radicals. This is only allowed for the following subset of root radicals:

MSA/CODA orthography	Dialectal pronunciation variants	Examples of dialect words written in CODA
ق	ء	طريق /Tari:'/
ث	س ت	كثير /kti:r/ (NOT كتير) أمّ كلثوم /ummu kulsu:m/
ذ	د ز	كذب /kidb/ ذل /zull/
ض	ظ ز د ض	ضابط /Za:bit/ (NOT ظابط)
ظ	ض ز ذ	ظلمة /Dalma/ (NOT ضلمة)
ص	س \ اص	صايغ pronounced /sa:yig/
ط	ت/ط	يا لطيف if pronounced /ya lati:f/

Table 1

Words like كحك are not written as كعك although there is an etymological link to MSA. This is because this transformation is limited in its applicability compared to the more systematic mappings observed for the phonological cases listed in Table 1.

Some words will have part of the stem written according to the default and part according to the above rule: e.g., برتقان not برتقال or برتآن.

Some words will have two letters changing and may be hard to recognize especially if they involve multiple readings depending on the semantics of the word: e.g., ثقيل can be pronounced /ti'i:l/ 'heavy' or /sa'i:l/ 'annoying [metaphorically heavy]' (note that the two words have different diacritizations).

4. **Vowel-Vowel difference**

a. The general rule is to preserve long vowels in base words in Egyptian Arabic (words without additional affixes/clitics) even if they shorten in different contexts.

- كاتب /ka:tib/
- كاتبين /katbi:n/ كتبين
- كاتبينها /katbinha/ كتبنها
- تقولها /tqu:lha/ `you say it' تقلّها
- تقول لها /tqulha/ `you tell her' تقلّها

b. Some patterns that have two long vowels in MSA and only one in EGY will be written in their MSA form since EGY phonology disallows multiple long vowels and will force the shortening of the first of the two long vowels anyway in a manner similar to what happens in EGY after the cliticization mentioned above:

- قانون فاعول NOT قنون
- مجانين مفاعيل NOT مجنين

c. Do not mark the distinction among the diphthong /ay/, the long high front /i:/ and the long mid front /e:/ vowels:

- دير
  - دير الزور/راهبات /de:r/ (/e:/ is pronounced like English *d**a**re*)
  - ديب كبير /di:b/ (/i:/ is pronounced like English *d**ee**r*)

d. Similarly, do not mark diphthong /aw/, the long high back /u:/ and the long mid back /o:/ vowels:

- دور
  - دور حوالين البيت /du:r/ (/u:/ is pronounced like English *f**oo**d*) 'turn'
  - اوقف عالدور /do:r/ (/o:/ is pronounced like English *d**oo**r*) 'queue up'
  - جاب لي دور /dawar/ role
  - دور عالدوا /dawwar/ look for

5. **Hamza**

We distinguish three types of Hamza:

a. Real Hamza (همزة قطع)

This is a hamza that is always pronounced as a glottal stop regardless of whether it is at the beginning or in the middle of a word.

e.g., مآسي, فؤاد، الأحلام، يأمرها، إمام

The rules of spelling this Hamza follow MSA guidelines. However they are disabled at clitic boundaries where MSA has case/mood diacritics and the dialects don't. For example, MSA may have بهائها // بهاؤها // بهاءها but the dialects only have بهاءها.

b. Temporary Hamza (همزة وصل)

This is the hamza sound associated with the Alif letter that is added for words starting with a vowel. This hamza will disappear if a clitic is inserted before it such as the conjunction /w/ or preposition /b/ although the Alif letter will be kept (and will be silent):

e.g., اسم، ابن، استأذن، التقى، انكتب، انكسر,اتقتل، اتكسر، اتهان، انقتل، انظلم، اتظلم

Words that have this hamza in MSA but do not in the dialect due to pattern change are written as in the dialect.

c. Missing (MSA) Hamza

EGY words that have hamzated MSA cognates will be written as pronounced in EGY, i.e. not using the MSA cognate. Word-final missing Hamza that can be written as Alif or Alif Maqsura should be spelled in a form that is closest to MSA: Alif-Hamza-Above or Alif-followed-by-Hamza becomes Alif, and Hamzated Ya becomes Alif Maqsura.

Examples:

لولي ، هوا ، سما ، بير ، مايل ، راس، ولاد، مرا/مراة
NOT لؤلؤ ، هواء ، سماء ، بئر ، مائل ، رأس, أولاد، إمرأة

Make sure not to confuse EGY hamza sounds and the hamza that is a result of a phonological transformation of the MSA cognate /q/: e.g., قرا / قريت /قريناها NOT أرا/أريت/ أريناها

6. **Alif Maqsura/Ya (ى / ي):** spell final ى/ي correctly: ى for /a:/ and ي for /i:/
   - السيد على محمد ابو شنب <- السيد علي محمد ابو شنب
   - الكلام ده على مين؟ <- الكلام ده علي مين؟

7. **Foreign words and place names (Modified from LDC)**
   - Most foreign words and place names already have established MSA spellings (e.g., Washington واشنطن, Los Angeles لوسانجلوس). In cases where the MSA spelling has regional variants, follow the Egyptian spelling since it confuses g/j/dj and does not depend on Perso-Arabic extensions as in Iraqi:
     - "garage" (Levantine: كراج -- contrast with Egyptian: جراج)
     - "congress" (Levantine: كونغرس -- contrast with Egyptian: كونجرس)
     - for /p/ use ب (e.g., "Pam" بام),
     - for /č/ use تش (e.g., "Chang" تشانج),
     - for /v/ use ف (e.g., “Vivian” فيفيان),
     - for /ž/ use ج (e.g., "Mirage" ميراج),
     - for /g/ use ج (e.g., "Gilbert" جيلبرت).

3.3. Word: Stems, Affixes and Clitics

The basic word consists of stem and obligatory affixes (some of which are nil). To the basic words, clitics can be added (all optional). Example: فحتراسليها `so you will correspond with her’.

enclitics (optional)	suffixes (obligatory)	Base Stem (obligatory)	prefixes (obligatory)	proclitics (optional)
ـها	ـيه	راسل	تـ	فـحـ

1. Stem and Affixes

a. Nouns and Adjectives have specific rules for affixes:
   - Ta Marbuta is always ة and never ه at the end of the word. Inside a word (after clitics) it may be ت or ا. The spelling of Ta Marbuta is not affected by pronunciation:
     - عربية سميرة جديدة in EGY /Earabiyyit sami:r**a** jidi:d**a**/
     - سيارتي // سيارتنا
     - معلمته `his teacher/boss’
     - معلماه `she taught him ’ شايفة ‘she is seeing’ the gerund deverbal form using the active participle
     - Note that some words introduce a Ta-marbuta-like variant that is always

pronounced /t/. This is written as ت, e.g., معنى/معنات ؛ مدام/مدامت
Egyptian Arabic only has مدامته NOT مدامه for `his wife'

- Dialectal noun suffixes include ين+ (both dual and plural) and ات+ also. Note that ين+ does not vary depending on case and construct state as in MSA (ين // ون // ان // ا // و).

b. Important EGY Verbs spelling decisions that are phonological and different from MSA:

- Added vowels for geminates (مضاعف):
  - مد / مديت / مديتوا
- Plural affixes (+/u:/, +/tu:/) are spelled with a silent Alif in word final spelling: وا ، توا
  - كتبوا / كتبتوا / كتبوها / كتبتوها

  This should not be confused with the 3rd masculine singular clitic /uh/ which is written differently (see below).

- Feminine affixes (+/i:/ and +/ti:/) are spelled with extra ي+ reflecting the underlying long vowel.
  - كتبتي / تكتبي

2. Clitics

There are **four** general rules for spelling clitics:

a. All single letter particles are cliticized (attached) while multi-letter clitics are not attached. The only exception is the definite article ال+

- البيت / بالطول / عالبيت

b. Pronominal enclitics (الضمائر المتصلة) and the negation particle (ش+) attach word finally but nothing else. CODA does not attach the prepositional phrase (ل+ضمير) to the word preceding it.

- بيتنا / بيتكم
- ما كتبناش
- ضرب له سلام // حكى لها قصة
- قلها should be written قالها 'he said it' or قال+لها 'he said to it/her' depending on context.

c. We follow MSA morphophonemic writing as in writing ال+ almost always as ال+ regardless if the stem starts with a sun/moon letter and regardless of any reduced pronunciations:

- القمر /il'amar/
- الشمس/i$$ams/
- الجبال /iljiba:l/
- الأولاد/il'awla:d/
- الولاد /ilwila:d/

d. The general rule of clitics is that when we attach the clitics to the word, we do not change the base word spelling even if there is a phonology change in either stem or clitic (for example due stress shifts), e.g., ب+ال+بيت is بالبيت. There are some exceptions:

- ل+ال become لل (but ب+ال remains بال)
  i. ل+البيت // للبيت

- ة becomes ت or ا
  i. معلمة+هم // معلمتهم // معلماهم
- الف واو الجماعة is deleted
  i. كتبوا+ها // كتبوها
- Alif Maqsura is turned into Alif or Ya depending on the word
  i. حكى+هم // حكاهم
  ii. على+هم // عليهم

- A common error is to apply the shadda spelling rule across a clitic-stem boundary.
  i. مسك+كم // مسككم NOT مسكم
  ii. شبه+ها // شبهها NOT شبها
  iii. سمن+ني // سمنني NOT سمني

3. Dialectal Clitics not in MSA

The following is a list of dialectal (non-MSA) clitics and some comments:

a. The progressive b+ attaches to imperfect verbs. Make sure to always spell the base word correctly (the example with pluses is only to explain the prefixes; do not write the pluses in CODA)

- ب+تكتب // بتكتب
- ب+يكتب // بيكتب NOT بكتب even if pronounced as such
- ب+اكتب // باكتب NOT بكتب even if pronounced as such; Note Hamza!
- ب+نكتب // بنكتب NOT بنكتب even if pronounced as such\

b. The future marker H+ attaches to imperfect verbs. Make sure to always spell the base word correctly:

- ح+امشي // حامشي NOT حمشي or همشي
- NOTE: Egyptian Arabic has two common variants: +ح and +هـ. ONLY USE +ح.

c. Second person plural /ku:/ clitic is spelled with +كو NOT (+كوا)

- شافوكو NOT شافوكوا

d. Feminine /i:/ clitics are spelled with extra +ي

- شافوكي
- This particular pronoun (+كي) has another form (+ك):
  - شافوكي // شافكي /$a:fik/ /$afu:ki/
  - ما شافوكيش // ما شافكيش

e. The 3rd person pronouns (ه ها همهن) are always written with an هـ even if not pronounced or assimilated

- كتابه/kita:buh/
- كتابها/kita:bha/
- كتبوها/katabu:ha/
- كتبوه/katabu:h/ or /katabu:/
- كتبتيها/katabti:ha/

f. Other clitics (PREP,PART, CONJ, SUB_CONJ…)

- Shared with MSA: ل / ب / ك / ف / و / etc.
  i. +و is ALWAYS attached: والبيت NOT وا البيت
- +ع 'on/at'
  i. عالبيت // ع+البيت
- +ش 'not'
  i. ما كتبش /katab$/ `he did not write'
  ii. ما كتبوش /katabu:$/ `**they** did not write'

  Note: the 3rd person masculine singular pronominal clitic (+ه) changes form before +ش to هوش:
  i. كتبه /katabuh/ `he wrote **it**'
  ii. كتبوه /katabu:h/ `they wrote **it**'
  iii. ما كتبهوش /katabhu:$/ `he did not write **it**'
  iv. ما كتبوهوش /katabuhu:$/ `they did not write **it**'

  Note: The Egyptian variant +شي of +ش should be written as +ش.

g. Particles that SHOULD NOT BE cliticized
- mA of Negation. Always add a space between it and word even though the -$ is cliticized.
  i. ما كتبش // ما كتب+ش NOT ماكتبش ، مكتبش
  ii. ما شربش NOT مشربش
- Indirect object pronouns (+ل pronoun) is also separated
  i. قول له NOT قلله
  ii. قلت لك NOT قلتلك
- Any particles longer than one letter. The only exception is Al.
  i. رح يمشي NOT رحيمشي

4. Notes on Diacritics

The following are some clarifications on writing diacritics. This is relevant only to the CALIMA project, and not to general CODA annotation (which is undiacritized).

a. Write the word as it is pronounced out of sentential context. Any epenthesis or elision (i.e. vowel added or deleted) only in context are not written.
  i. mA tiquwl$ /ma tqu:l$/ (NOT mA tquwl$)

b. Clitics may loose or gain vowels as part of word formation (word internal). These cases are written reflecting their pronunciation.
  i. bi+ni+m$iy, bi+ti+m$iy but b+A+m$iy (NOT bi+A+m$iy)
  ii. katab+uh, katab+uw+h, katab+huw+$, katab+uw+huw+$

c. Some verbs allow two forms of prefix vowels optionally: yi+quwl and yu+quwl. In all such cases, we prefer i over u: yi+quwl (NOT yu+quwl).

3.4. Exception List for Common Dialectal Words

- **Please refer to the CODA exception table document https://docs.google.com/spreadsheet/ccc?key=0Ah1Fn1dEmA95dDkxOG9hTW16QjUtY29DeVlLRHpMNFE#gid=0**
- **These tables will be continuously updated.**