Palestinian Arabic Conventional Orthography Guidelines - Technical Report

Nizar Habash[1], Mustafa Jarrar[2], Faeq Alrimawi[2], Diyam Akra[2], Nasser Zalmout[1,2], Eric Bartolotti[?] and Mahdi Arar[2]

[1] New York University Abu Dhabi, United Arab Emirates
{nizar.habash}@nyu.edu

[2] Birzeit University, Palestine
{mjarrar,falrimawi}@birzeit.edu

Abstract This technical report describes a general template for Conventional Orthography for Dialectial Arabic creation regardless of dialect and specific details and guidelines for Palestinian Arabic creation. The technical report serves two purposes: defining the general and specific Conventional Orthography for Dialectial Arabic information as well as being the base for a Palestinian CODA.

The technical report describes the variations generated by Palestinian Arabic speakers and the sub-dialects of the Palestinian Arabic, and sets guidelines on how to deal with these variations. It also describes the structure of a word and states the clitics and affixs associated with the base word. Moreover, it adds clitics and affixs which are specific to the Palestinain Arabic. the technical report also provides a list of exceptional list that has words with its variations and which variation to be used as the CODA form.

Keywords Palestinian Arabic, Conventional Orthography for Dialectal Arabic, Dialectal Arabic

1. Introduction

In this technical report we provide guidelines, and examples to provide a conventional orthography for dialectal Arabic (CODA) [1] words, especially Palestinain dialectal [2] words.

There are different scenarios for using CODA:

a. to directly write in to it (such as when creating lexicon entries or morphology analysis examples)
b. to correct spontaneously written Arabic text to CODA
c. or to map Arabizi (Arabic written in Roman script) to CODA

Each of these scenarios has its own restrictions, which should be specified in additional guidelines. But as a general principle, correct CODA spelling should be typo-free and may use punctuation marks in a way consistent with MSA's use of punctuation marks. CODA has been implemented for the Tunisian dialect [3] and Algerian [4] dialect.

2. Basic Phonology-to-Orthography Map

This section specifies the mapping of phonemes to graphemes (letters and diacritics). These map rules are generic and apply to all dialects. These rules can be used to write words "phonetically". The morphology and lexicon based exceptions to these general rules within CODA will be specified in later sections.

2.1. Consonantal Phonemes

2.1.1. Consonants

The following consonants phonemes are mapped to consonants graphemes as is done in MSA:
/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)

Emphasis (**تفخيم**) Emphatic spread can confuse the choice of a phoneme, e.g. س/ص or ت/ط; in such cases, we consider the non-emphatic form as the default. Exceptions discussed later in this document can overwrite this choice. Some dialects have additional emphatic phonemes that are written (e.g., the /b/ in bAbA 'daddy').

Allophones Allophonic version of these consonants should not be made explicit in writing, e.g., جنبmay be pronounced /jamb/ but not written as جمب.

2.1.2. Shadda

The Shadda diacritic replaces the second letter in a repeated letter sequence:
e.g., /kallam/ => (kal~am) (كَلَّم)

2.1.3. Hamza Spelling

The glottal stop consonant (Hamza) is mapped to different Hamza graphemes determined by the vowel context as is done in MSA (details omitted here):

/'/ => (>,<,|,&,},') (أ،إ،آ،ؤ،ئ،ء)

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

2.2. Vocalic Phonemes

2.2.1. Vowels

The vowels are mapped as follows:

/a,i,u/ => (a, u, i) (ـَ, ـُ, ـِ)
/a:, i:, u:, e:, o:/ => (aA, iy, uw, ay, aw) (ا, ـِي, ـُو, ـَي, ـَو)
/aw, ay/ => (aw, ay) (ـَو, ـَي)

This is the same as MSA with the exception of /e:/ and /o:/ which do not appear in MSA.
CODA conflates /aw/ and /o:/, as well as /ay/ and /e:/.

Allophones As in MSA, emphatic versions of the vowels are allophonic variants (not written).

Vowel Length Some dialects (unlike MSA) allow vowels to shorten phonetically in different contexts. Vowel allophones involving shortening are written phonemically, i.e., phonetically shortened long vowels are still written long. The reverse is assumed to not happen (short vowels that are lengthened phonetically).

2.2.2. Sukun

The sukun symbol marks the absence of a vowel after a consonant. (In fully diacritized writing it can be ignored as it can be inserted automatically).

2.2.3. Hamzat Wasl (Temporary Hamza)

Vowels at the beginning of words result in word-initial glottal stop pronunciation. This is a phonetic phenomenon and does not reflect the presence of a real word-initial Hamza (همزة قطع). This phenomenon appears in MSA and all dialects, although with different distributions. To determine if the word starts with a real Hamza phoneme or a temporary Hamza, add the clitic “w” و or “b” بـ before it: real Hamza (همزة قطع) remains, temporary Hamza (همزة وصل) disappears.

e.g., from Egyptian Arabic
/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)

Compare the behavior between MSA and Egyptian on some Egyptian words:

	Egyptian same as MSA				Egyptian different from MSA				
Correct	اسم	ابن	انكتب	اتكسر	ابراهيم	اسرائيل	اسود	ابيض	اعزب
Incorrect	إسم	إبن	إنكتب	إتكسر	إبراهيم	إسرائيل	إسود	أبيض	أعزب

Table 1 comparison between MSA and Egyptation for same words.

2.2.4. Epenthesis and Elision

Phonetic epenthesis is adding a short vowel (schwa) to break up consonant clusters (الكسر لمنع التقاء الساكنين/السواكن). Phonetic elision is deleting short unstressed vowels in certain contexts. These two phenomena will not be written when resulting from interaction between the base word and its clitics or the word and other words; the specifics pronunciation of each dialect determines the rules of adding/deleting them. E.g., اِبْن بْلادِي
/ibn bla:di/ => [ib.nib.la:.di].

2.3. Suboptimal Spelling

In addition to avoiding typographical errors, the following very common sub-optimal spelling choices must be avoided:

a. Alif Maqsura/Ya (ي / ى): spell final ى/ي correctly: ى for /a:/ and ي for /i:/, /e:/, /ay/ and /y/
b. Alif-Hamza forms (آأإ) must be spelled with the Hamza or Madda, and not be confused with bare Alif (ا). The Alif Wasla (for Hamzat Wasl) (ٱ) will be written in CODA as bare Alif.

2.4. Foreign Phonemes

2.4.1. Consonants

For commonly used foreign consonants (in foreign words), use the following:
/g/, /ž/, → ج
/p/ → ب
/v/ → ف
/č/ → تش

2.4.2. Vowels

For foreign vowels, map them to the closest Arabic vowel. If a long vowel reading is possible, prefer it over a short vowel reading (e.g., "Mirage" ميراج not مراج). Foreign words ending with long vowels will be rendered with an extra silent (h) word finally:
e.g., [mayo:] => /ma:yo:/ => (mAyawh) (مايوه) ‘swimsuit’

2.5. Palestinian Arabic Phonology Observations and Pronunciation Rules

2.5.1. Sub-dialectal variations

There are a number of variations within Palestinian Arabic. We consider the Urban dialect(مدني), the unmarked base dialect. Accent variations from it will not be written. Below is a table of some of these cases:

MSA	PAL-Urban(مدني)	PAL-Rural(فلّاحي)	PAL-Bedouin	PAL-Druze
/q/	/'/	/k/	/g/	/q/
/k/	/k/	/č/	/k/	/k/

Table 2 examples of accent variations in PAL.

2.5.2. Phonotactic Vowel Shortening and Stress

PAL expresses regular stress (specify details); adding affixes may cause stress movement.

In some cases one long vowel allowed per word; Long vowel receives stress; unstressed long vowels shorten (phonotactic). However, Some Palestinian Arabic speakers keep the first long vowel of a word without shortening it, others shorten it. For example:
/$a:ku:$/ [شاكوش] and /$aku:$/ [شكوش]

Long vowels shorten in word final positions unless they are followed by some morpheme (which may simply realize as nil): /za:r+u:/ [za:ru] but /za:r+u:+h/ [zaru:]

2.5.3. Epenthesis

All words ending with CC clusters (that are not geminates) allow for a CiC epenthesized pronunciation: /kalb/ and /kalib/; /Darb/ and /Darib/; /katabt/ and /katabit/; We will consider the non-epenthetic version the base and use it in CODA.

Consonant clusters across words are broken up with an epenthetic vowel (الكسر لمنع التقاء السواكن). The vowel can be /i/ or /u/ in round contexts. E.g., /ibn/+/bla:d/ => /ib-nib-la:d/. The epenthetic vowel will not be written.

2.5.4. Vowel Allophones that are not written.

(a) The short vowel /i/ has two allophones [i] and [e].
(b) The short vowel /u/ has two allophones [u] and [o].
(c) All vowels have emphatic forms.

2.5.5. ??? Reference to other Levantine dialects?

3. Word Structure

CODA, like MSA orthography, is a morphophonemic writing system with some exceptions. To be able to explain and apply this system, we must understand the structure of the words we are writing. The first step to write a word is to decompose it into its components:

1) Pronunciation – how is the word pronounced?
2) Meaning – what does it mean?
3) Morphology – what are the basic units the make it up?

In terms of morphology, we break up the word into:

1) **Clitics**: proclitics (prefixing) and enclitics (suffixing). Clitics, which are all *optional* to word formation, include all the particles (الحروف المتصلة) and object/possessive pronouns (ضمائر النصب والجر المتصلة).
2) **Base Word**, which can be further broken into
   a) **Affixes**: *obligatory* prefixes and suffixes include all affixes other than clitics, e.g. ة+ات،+وا،+، etc.
   b) **Stem**, which can be further broken into
      i) Root
      ii) Pattern

It is important that the meaning and the morphology be consistent. All affixes and clitics are provided in Appendix A.

Example (MSA):

Pronunciation	/wasayaktubu:naha:/					
Meaning	'and they will write it'					
Morphology	**Proclitics**	**Base Word**				**Enclitics**
	wa+ sa+	yaktubu:n				+ha:
		Prefixes	**Stem**		**Suffixes**	
		ya-	aktub		-u:na	
			Root	**Pattern**		
			k.t.b	a12u3		

Table 3 Word structure.

The rest of this document will refer to these levels of representing the structure of a word. First we discuss the spelling of Base Words independently of clitics. Then we discuss the addition of clitics. The list of clitics and affixes needs to be specified for each dialect, although some general principles will be followed across all dialects (details below).

4. Base Word Spelling

For spelling of the base word, we start with the lexical exceptions, which override any other set of decisions. Then we discuss the rules for spelling the word using its components (affixes, stem, root and pattern).

4.1. Lexical Exceptions

Some words are spelled in a particular *ad hoc* way. A list is provided in Appendix B. The appendix should not include clitics. It is only for Base Word spelling.

General Thoughts on Exceptional Spelling Choices The exception list should be easy to remember by CODA users. As such, in selecting the exceptional spelling, the CODA designers will try to obey as many of the basic phonology-orthography rules and affix spelling as possible. The option of using an exceptional spelling from the CODA of another dialect is encouraged if the pronunciation is similar. Considering Google counts is also encouraged. However, consistency in spelling related words (same POS, same class of phenomena) is preferred over pure statistics.

e.g.,	/intu/	=> (Aintuw) (إِنْتو) NOT (AintuwA) (إِنْتوا)
	/barduh/	=> (barDuh) (بَرْضُه) NOT (barDuw) (بَرضو)

4.2. Affix Spelling

Appendix A specifies the list of affixes and clitics. The list should be updated for each dialect. As with other dialect specific decisions, we try to follow some general principles that carry across different dialects as much as possible. Most affixes are spelled as pronounced. The following are some exceptional affix rules that are dialect independent:

- Ta Marbuta is a particular affix that is usually associated with the feminine singular inflection of a noun. The test for whether a word ends in Ta Marbuta is whether the ending changes from some vowel to /t/ (or /it/) in the context of an Idafa construction: e.g., say~Arap mnY /sayya:rit muna/ as opposed to samA falasTiyn /sama falasTi:n/.
  - 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:**ra** jidi:**da**/

    - سيارتي // سيارتنا
    - معلمته `his teacher/boss’
    - معلماه `she taught him ’ شايفة ‘she is seeing’ the gerund deverbal form using the active participle
    - In Levantine there are different pronunciations of Ta Marbuta, we only write the version with +ap and consider the rest accent variants. For example the word معلمة can be written as: معلمه or معلمي.
- Plural affixes of verb subjects (+/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 as ه+.
- Feminine affixes (+/i:/ and +/ti:/) are spelled with extra ي+ reflecting the underlying long vowel.
    - كتبتي / تكتبي
- Nunation in adverbial constructions should be spelled using the nunation diacritics.
    - عملياً، فعلاً، غصبٍ

4.3. Stem Root Spelling

4.3.1. Root Consonants

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

MSA/CODA orthography	Pronunciation variants	Correct	Incorrect
ق	ء ك	طريق برتقان قال	طريء برتئان كال
ك	تش	كيف حالك	تشيف حالتش
ث	س ت ط	كثير ام كلثوم ثور	كتير ام كلسوم طور
ذ	د ز ظ	كذب ذل	كدب زل
ض/د	ظ ز د	ضابط	ظابط
ظ/ز	ض ز ذ	ظل	ضل
ص/س	س ص	صاقع	ساقع
ط/ت	ت ط	اللطف فستان	اللتف فسطان

Table 4 letters’ variations in PAL

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 /t’i:l/ ‘heavy’ or /sa’i:l/ ‘annoying [metaphorically heavy]’ (note that the two words have different diacritizations).

4.3.2. Disappearing Hamza

Hamzas are only written when pronounced. Many words without the Hamza sounds have cognates in MSA that have Hamzas. These Hamzas are not written. Some examples:

Correct	هوا	سما	بير	مايل	راس	ولاد
Incorrect	هواء	سماء	بئر	مائل	رأس	أولاد

Table 5 Examples of correct and incorrect Hamaz writings.

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)
/’ara/   => (qarA) (قرا) NOT (qaraY) (قرى)

4.3.3. Initial Hamza

Hamzas that appear at the beginning of a word are always dropped, and replaced with a bare Alif. This decision was made since dropping the Hamaza will not, in most cases, create ambiguity. For example:
/’x*/ **أخذ**=> (Ax*) **اخذ**
/’Hmd/**أحمد** => (AHmd) **احمد**

4.3.4. Alif Maqsura (Type 1 - Root based)

A word final /a,a:/ vowel is spelled as ىY if the word has a root radical يy that changed to /a/. Final root radicals other than y that turn into /a,a:/ (Such as w) are written with Alif ا. This rule is the same in MSA.

4.4. Stem Pattern Spelling

4.4.1. Pattern Consonants

Follow the same rules of spellings as MSA.

Correct	نفترض	استعطى	ازدهر
Incorrect	نفطرض	اصطعطى	ازتهر (افتعل)

Table 6 Examples of correct and incorrect pattern spellings

4.4.2. Vowels

The general rule is to preserve long vowels 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

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:

- فاعول قانون NOT قنون
- مفاعيل مجانين NOT مجنين

It is also important to avoid making short vowels long in context where the long vowel cannot appear and has not pattern-based evidence. A common case is the conjugation of hollow verbs (same rule as in MSA spelling):

- قلت NOT قولت

4.4.3. Alif Maqsura (Type 2 – Pattern based)

Some Alif Maqsuras are pattern of the pattern in MSA. If the pronunciation remains the same in the dialect, we use the Alif Maqsura. For words that end with /a:,a/ and whose pattern in MSA had a Hamza, the CODA spelling preserves the Alif (and not the Hamza).

- فعلى قتلى NOT قتلا
- فعلا حمرا NOT حمرى ، حمرة

4.4.4. Putting the Base Word Together

When integrating the components of the base word (root, pattern, affixes) to compose it, we apply the general rules discussed above. In particular, the Shadda rule may be needed when a suffix starts with the same consonant as the end of the stem: e.g., jan~an+nA => jan~an~A not jan~annA (which is a different word that we discuss in the next section).

5. Clitic Spelling

5.1. Clitic Forms

The forms of the clitics in a dialect should be specified in Appendix A (including morphotactics). Many clitics are the same as in MSA and other dialects. But there are some unique clitics in some dialects that are not shared. Some linguistic clitics are not written as clitics in CODA.

5.1.1. Basic Clitic Rules

- We follow MSA morphophonemic writing as in writing ال+ always as ال+ regardless if the stem starts with a sun/moon letter and regardless of any reduced pronunciations:
  i. القمر /il'amar/
  ii. الشمس /i$$ams/
  iii. الجبال /lijba:l/

- All single letter particles are cliticized (attached) while multi-letter clitics are not attached. The only exception is the definite article ال+
  iv. البيت / بالطول / عالبيت
  v. ب+تكتب = بتكتب
  vi. ح+امشي = حامشي
  vii. كتب+ش = ما كتبش

- Pronominal enclitics (ضمائرالنصب والجر المتصلة)
  viii. بيتنا / بيتكم
  ix. شافوكو NOT شافوكوا

- Dialectal clitics

5.1.2. Linguistic clitics that are not written attached

- mA of Negation. Always add a space between it and word even though the -$ is cliticized.
  i. ما كتب+ش // ما كتبش NOT ماكتبش ، مكتبش
  ii. ما شربش NOT مشربش، ماشربش
- Indirect object pronouns (ل+ pronoun) is also separated
  iii. قل له NOT قلله

iv. قلت لك NOT قلتلك

- Any particles longer than one letter. The only exception is Al.

v. رح يمشي NOT رحيمشي

5.2. Spelling Interactions as a result of Cliticization

5.2.1. No Change

The default rule is that cliticization does not change the spelling of the base word.
/bilmaka:tib/ => /bi+Al+makAtib/ => (biAlmakAtib) (بِالمَكاتِب)
ب+اكتب // باكتب NOT بكتب even if pronounced as such;

Consistent with this rule is the disabling of the Shadda rule across base-word-clitic boundaries (except for +ya):

(bArik+kum) (باركْكُم), (wAH$iyn+nA) (واحشيننا), (jan~an+nA) (جننّنا)
but (Ealy+~a) (عَليّ)

5.2.2. Clitic Changes

Exceptional to the above rule are changes in spelling specific clitics

- The definite article Al loses its Alif after the preposition l+
  ل+البيت = للبيت

- Some pronominal clitics have different allomorphic forms. Examples:

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

PRON_3MS /o/ => /o/, /:/, /ho/
5. /$a:fo/ => ($Afuh) (شافُه)
6. /$afu:/ => ($Afuwh) (شَافُوه)
7. /$afo:$/ => ($Afhuw$) (شَافهُوش)
8. /$afuho:$/ => ($Afuwhuw$) (شافُوهُوش)

5.2.3. Base Word Changes

Exceptional to the above rule are changes in spelling specific to base word endings

- Ta Marbuta changes to t or A before an enclitic:
  - معلمة+هم // معلمتهم // معلماهم

- The Alif of the subject ending wA is dropped before an enclitic:
  - كتبوا+ش // كتبوش
  - كتبوا+ها // كتبوها

- Alif Maqsura is turned into Alif or Ya depending on the word
  - حكى+هم // حكاهم
  - على+هم // عليهم

6. Examples

Raw sentence	لبست الفسطان وراحت جري عالكصر وشافها الأمير وحبها وركضو سو وفجـــــــأة دكت الساعة طنعش وجراي وصارت ترمح و ترمح ووكّعت ببوجها عالدرج سندريلا راحت عالبيت واتمنت لو انو الساعة ما دكتش على الطنعش،

Buckwalter	بس شو تسوي بحظها المعتر مثل حظ هاظا الشعب الأمير حط فردة البابوج على مخدة يمكن لونها زركة أو خضرة، المهم صار ينادي بالصوت على كل البنات عشان يكيسن البابوح، كل البنات كاسوا البابوج إلا سندريلا. lbst AlfsTAn wrAHt jry EAlkSr w$AfhA Al>myr wHbhA wrkDw sw wfj>p dkt AlsAEp TnE$ wjrAy wSArt trmH w trmH wwk~Et bbwjhA EAldrj sndrylA rAHt EAlbyt wAtmnt lw Anw AlsAEp mA dkt$ ElY AlTnE$, bs $w tswy bHZhA AlmEtr mvl HZ hAZA Al$Eb Al>myr HT frdp AlbAbwj ElY mxdp ymkn lwnhA zrkp >w xDrp, Almhm SAr ynAdy bAlSwt ElY kl AlbnAt E$An ykysn AlbAbwH, kl AlbnAt kAswA AlbAbwj <lA sndrylA.
CODA Buckwalter	لبست الفستان وراحت جري عالقصر وشافها الامير وحبها وركضوا سوا وفجأة دقت الساعة اطنعش وجراي وصارت ترمح و ترمح ووقعت ببوجها عالدرج سندريلا راحت عالبيت واتمنت لو انه الساعة ما دقتش على الاطنعش، بس شو تسوي بحظها المعتر مثل حظ هاذا الشعب الامير حط فردة البابوج على مخدة يمكن لونها زرقة او خضرة، المهم صار ينادي بالصوت على كل البنات عشان يقيسن البابوج، كل البنات قاسوا البابوج الا سندريلا. lbst AlfstAn wrAHt jry EAlqSr w$AfhA AlAmyr wHbhA wrkDwA swA wfj>p dqt AlsAEp ATnE$ wjrAy wSArt trmH w trmH wwqEt bbwjhA EAldrj sndrylA rAHt EAlbyt wAtmnt lw Anh AlsAEp mA dqt$ ElY AlATnE$, bs $w tswy bHZhA AlmEtr mvl HZ hA*A Al$Eb AlAmyr HT frdp AlbAbwj ElY mxdp ymkn lwnhA zrqp Aw xDrp, Almhm SAr ynAdy bAlSwt ElY kl AlbnAt E$An yqysn AlbAbwj, kl AlbnAt qAswA AlbAbwj AlA sndrylA.
English	She wore the dress and started running towards the castle then the prince saw her and fell in love with her, then they ran together and suddenly the clock pointed to twelve so she started running and running and she dropped her shoe on the stairs Cinderella went home and wished that the clock did not reach twelve but what she can do with her bad luck as this people's luck the prince put the shoe on a pillow maybe its color was blue or green however he started shouting for all the girls to try the shoe, all the girls tried the shoe except for Cinderella.

Table 7 Example 1: Using PAL-CODA guidelines

Raw sentence Buckwalter	أنا حاسس انو بيوم عرسي أمي راح تقلي خد الزبالة معك وانت طالع >nA HAss Anw bywm Ersy >my rAH tqly xd AlzbAlp mEk wAnt TAlE
CODA Buckwalter	انا حاسس انه بيوم عرسي امي رح تقول لي خذ الزبالة معك وانت طالع AnA HAss Anh bywm Ersy Amy rH tqwl ly x* AlzbAlp mEk wAnt TAlE
English	I feel that on my wedding day my mother would say to me take the garbage with you when you go out.

Table 8 Example 2: Using PAL-CODA guidelines

7. List of Basic Morphemes and POS Tags

Morp heme	Morpheme	Cliticiz-ation Allom-orph	POS	gloss	Example	Notes
	Proclitics					
الـ	Al		DET		Al+ktAb	
و	w		CONJ	And	AsmE w+Afhm	
			SUB_CONJ	While	jA’ w+hw ybtsm	
			PREP	with	AstwY AHmd w+Ely	
فـ	f		CONJ	And, so	jA’ AHmd f+Ely	
			CONNEC_PART	And, so		
			RC_PART	So, then		
			SUB_CONJ	So, that	lA thml f+trsb	
بـ	b		PROG_PART		b+ylEb	PAL
			PREP	By, with	bi+yAdy	
كَـ	ka		PREP	like	ka+ AlzhrA	
سَـ	sa		FUT_PART	will	sa+yElm	
لِـ	li		PREP	To, for	hA*A l+Ely	
			EMPHATIC_PART	Will certainly	Aryd l+AnsY *krhA	
			RC_PART	So, then		
	Enclitics					
	Nominal Enclitics					
ي	yi		POSS_PRON_1S			
نا	nA		POSS_PRON_1P			
كَ	ka		POSS_PRON_2MS			
كِ	ki		POSS_PRON_2FS			
كي	ky		POSS_PRON_2FS			PAL
كما	kmA		POSS_PRON_2D			
كم	km		POSS_PRON_2MP			
كو	kw		POSS_PRON_2MP			
كن	kn		POSS_PRON_2FP			
ه	h		POSS_PRON_3MS			
ها	hA		POSS_PRON_3FS			
هما	hmA		POSS_PRON_3D			
هم	Hm		POSS_PRON_3MP			
هن	hn		POSS_PRON_3FP			
	Verbal Enclitics					
ني	ny		PVSUFF_DO:1S IVSUFF_DO:1S CVSUFF_DO:1S			
نا	nA		PVSUFF_DO:1P			

			IVSUFF_DO:1P CVSUFF_DO:1P			
كَ	ka		PVSUFF_DO:2MS IVSUFF_DO:2MS			
كِ	ki		PVSUFF_DO:2FS IVSUFF_DO:2FS			
كي	ky		PVSUFF_DO:2FS IVSUFF_DO:2FS			PAL
كما	kmA		PVSUFF_DO:2D IVSUFF_DO:2D			
كم	km		PVSUFF_DO:2MP IVSUFF_DO:2MP			
كو	kw		PVSUFF_DO:2MP IVSUFF_DO:2MP			PAL
كن	kn		PVSUFF_DO:2FP IVSUFF_DO:2FP			
ه	h		PVSUFF_DO:3MS IVSUFF_DO:3MS CVSUFF_DO:3MS			
ها	hA		PVSUFF_DO:3FS IVSUFF_DO:3FS CVSUFF_DO:3FS			
هما	hmA		PVSUFF_DO:3D IVSUFF_DO:3D CVSUFF_DO:3D			
هم	Hm		PVSUFF_DO:3MP IVSUFF_DO:3MP CVSUFF_DO:3MP			
هن	hn		PVSUFF_DO:3FP IVSUFF_DO:3FP CVSUFF_DO:3FP			
	Prefixes					
أَ	Aa		IV1S	I	Aa+Drub	
آ	|		IV1S	I	|+kul	
نـ	n		IV1P	we	n+HAwil	
نِـ	ni		IV1P	we	ni+tEal~am	
نُـ	nu		IV1P	we	nu+Drub	
نَـ	na		IV1P	we	na+ETiy	
تـ	t		IV2MS	you		
تِـ	ti		IV2MS	you		
تُـ	tu		IV2MS	you		
تَـ	ta		IV2MS	you		
تـ	t		IV2FS	you		
تِـ	ti		IV2FS	you		
تُـ	tu		IV2FS	you		
تَـ	ta		IV2FS	you		
تـ	t		IV2P	you		
تِـ	ti		IV2P	you		
تُـ	tu		IV2P	you		

تَـ	ta		IV2P	you		
تـ	t		IV3FS	it/they/she		
تِـ	ti		IV3FS	it/they/she		
تُـ	tu		IV3FS	it/they/she		
تَـ	ta		IV3FS	it/they/she		
يـ	y		IV3MS	he/it	y+HAwil	
يِـ	yi		IV3MS	he/it	yi+tEal~am	
يُـ	yu		IV3MS	he/it	yu+Drub	
يَـ	ya		IV3MS	he/it	ya+ETiy	
يـ	y		IV3P	they	y+HAwl+uwA	
يِـ	yi		IV3P	they	yi+tEal~am+uwA	
يُـ	yu		IV3P	they	yu+Durb+uwA	
يَـ	ya		IV3P	they	ya+ET+uwA	
	Suffixes					
	Nominal Suffixes					
اً	AF		CASE_INDEF_ACC	[indef.acc.]	Eamaliy~+AF	
◌ٍ	K		CASE_INDEF_GEN	[indef.gen.]	gaSb+K	
◌َة	ap	it,t	NSUFF_FEM_SG	[fem.sg.]	Eulbap, EaD~ap, jArithA, jArtuh	Alway use ap; do not use ip
◌َة	ap	A	NSUFF_FEM_SG	[fem.sg.]	sAmEap AlSawt => SAmEAh	
تَين	tayn		NSUFF_FEM_DU	two		
◌َين	ayn		NSUFF_MASC_DU	two		
ات	At		NSUFF_FEM_PL	[fem.pl.]		
◌ِين	iyn		NSUFF_MASC_PL	two		
	Verbal Suffixes with PV verbs					
ت	t		PVSUFF_SUBJ:1S	I	katabt	
◌َيت	ayt		PVSUFF_SUBJ:1S	I	Hab~ayt	
◌ّ	~		PVSUFF_SUBJ:1S	I	fut~	Pronunciation issue -> fut~ / futit
نا	nA		PVSUFF_SUBJ:1P	we		
◌َينا	aynA		PVSUFF_SUBJ:1P	we		
◌ّا	~A		PVSUFF_SUBJ:1P	we	jan~an~A	compare: jan~an~A

						(we drove others crazy) with jan~annA (he drove us crazy)
تِي	tiy		PVSUFF_SUBJ:2FS	you		
َيتِي	aytiy		PVSUFF_SUBJ:2FS	you		
ِّي	~iy		PVSUFF_SUBJ:2FS	you		
ت	t		PVSUFF_SUBJ:2MS	you		
َيت	ayt		PVSUFF_SUBJ:2MS	you		
ّ	~		PVSUFF_SUBJ:2MS	you		
تُوا	tuwA		PVSUFF_SUBJ:2P	you		
َيتُوا	aytuwA		PVSUFF_SUBJ:2P	you		
ُّوا	~uwA		PVSUFF_SUBJ:2P	you		
َت	at		PVSUFF_SUBJ:3FS	it/they/she		
(مستتر)	(null)		PVSUFF_SUBJ:3MS	he/it		
ُوا	uwA		PVSUFF_SUBJ:3P	they		
	Verbal Suffixes with IV verbs					
ِي	iy		IVSUFF_SUBJ:2FS	you	t+HAwl+iy	
ُوا	uwA	uw+	IVSUFF_SUBJ:P	we	t+HAwl+uwA	
	Verbal Suffixes with CV verbs					
(مستتر)	(null)		CVSUFF_SUBJ:2M S	you	saj~il+(null)	
ِي	iy		CVSUFF_SUBJ:2FS	you	saj~l+iy	
ُوا	uwA	uw+	CVSUFF_SUBJ:2P	they	saj~l+uwA	

Table 9 Clitics and prefixes

8. Exceptional Spelling Choices

This section includes words with exceptional spelling, as well as common words whose CODA spelling follows the rules but not in a necessarily obvious way.

The appendix should not include clitics. It is only for Base Word spelling.

General Thoughts on Exceptional Spelling Choices: The exception list should be easy to remember by CODA users. As such, in selecting the exceptional spelling, the CODA designers will try to obey as many of the basic phonology-orthography rules and affix spelling as possible. The option of using an exceptional spelling from the CODA of another dialect is encouraged if the pronunciation is similar. Considering Google counts is also encouraged. However, consistency in spelling related words (same POS, same class of phenomena) is preferred over pure statistics.

	CODA	Non-CODA Variants	English	Example
Subject Pronouns	انا مانيش	منيش	I	
positive/negative	احنا ماحناش	نحنا - إحنا محناش	we	
	انت - مانتاش	انتا – إنتا منتاش	you [2ms]	
	انتي مانتيش	انت – إنت منتيش	you [2fs]	
	انتو مانتوش	انتوا - إنتوا – إنتو منتوش	you [2p]	
	هو ماهواش- ماهوش	هوه - هوة - هوا- هوتا مهواش مهوش	he, it [3ms]	
	هي ماهيش	هيه - هوة - هيا- هيتا مهيش	She, it [3fs]	
	هم ماهماش	همه - همة – هما-همي مهيش	they [3p]	
Demonstrative Pronouns	هاذا	هاظا-هادا- هاذ-هاض-هاضا	this, that [3ms]	
	هذهو	هظهو-هضهو	This, that [3ms]	
	هاذي	هاظي-هادي-هاضي	this, that [3fs]	
	هاي		this, that [3fs]	
	هذول	هدول-هظول-هضول	these, those [3p]	
	هذولاك	هدولاك-هظولاك-هضولاك	these, those [3p]	
	هيني	هايني-هيوني	Here I am	
	هيو	هيه-هيوتو-هيتا	There he/it is	
	هيي	هيتها-هيوتها	There she/it is	
	هينا	هاينا-هيتنا-هيوتنا	Here we are	
	هيكو	هيكوا-هيكم	There you are	
	هيهم	هيومه-هيوتهم- هيتهم	There they are	
Relative Pronouns	اللي	الي - الي - اللى	who, which, whom	The Alif-Lam is treated as a definite article in cliticization: باللي - عاللي - للي
	يا اللي	يللي - ياللي - يللى - ياللى	who, which, whom	/yalli/ can be one of two things, a variant of illi or ya+ill. They are spelled differently
	تاع تبع		of	
Interrogatives	وين		where	

	من وين		where	
	مين	من	who	
	ايش شو	إيش– اش	what	
	ليه ليش	لية	why	
	امتى وينتا	امته - امتة - امتا - إمته - إمتة - إمتى - امتىً - إيمتىً - امتن - إمتن	when	
	اني	أني-انى	which	
	انو	أنو-انه-انوه	which	
	انهي	أنهي-أنوهي-انهه	which	
Location expressions	هان	هانا	here	
	هون	هونا		
	هناك	هيناك هناكا	there	
	برا	بره - برة	outside, outside of	كان بيستناني برا البيت
	جوا	جوه - جوة	inside, inside of	كان بيستنى جوا البيت
	دغري	دغرى - دوغرى - دوغري	straightforward	
Existential vs Location Fi	في	فى	In[prep]	
	فيه	في - فى	there is, in [prep] + it	two readings!
	فش-مفش	مفيش - مافيش - ما فيش - مافيهش - ما فيهش - مفيهش	there is not	only existential
	ما فيهوش	مافيهوش – مفيهوش	not in it	only prep
Time expressions	هسا	هسى-هس-هسع-هساع	now	
	هلقيت	هأيتي-هالقيت		
	هالحين	الحين-إلحين		
	هلق	هلأ-هلا		
	امبارح	إمبارح	yesterday	
	بكرة	بكره - بكرا	tomorrow	
Conjunctions	برضه – برضك	برضو - بردو - برده - بردك	also	
	بس	باس	only, enough, just	
	كمان	كمانا - كمانه	also	
	عشان مشان	علشان منشان	in order to	
	والا	واللا - وإلا - والا - وللا	or (in questions)	هاذا حرام ولا حلال؟....ولا شو رأيك؟
Negation particles	مش	موش	not	
Yes/ no	أيوه	ايوه - ايوة - ايوا - أيوة - أيوا - ايواً - ايون - أيواً - أيون	yes, indeed	
	لأ	لأا	no	

	لا	ل	not, neither, nor	No Hamza
Verb particles			will	
	رح	راح	will	a separate particle
	ب+	ب+	progressive particle	
Other	نيالكو	نيالكم-نيالكوا	how lucky [2mp]	
	نياله	نيالو	How luck [3ms]	
	زي	زى	like	
	هيك	هايك-هيكا	like that	
	ابصر	أبصر	I Don’t know, not known	ابصر شو عمل؟ شو صار مع احمد؟ ابصر!
	معلش		never mind	
	يالله	يلا - يلله		yalla (hurry up): do not confuse with “ يا الله” (oh, my god/godness)
	إن شاء الله	ان شاء الله - انشالله - انشاله - إنشالله - إنشاله	in God’s will	
	يا الله	ياالله	oh God	
	ألو	الو	hello (on phone)	
	أوكيه	أوكي - اوكيه - اوكي	OK	
	أوك	اوك	O.K.	EGY young people talk- pronounced like 'oak' not o-kay
	طيب	طب	OK, so	
	لسه	لسة – لسا- إسا	still	
	حد – حدش حدا-حداش	حاد حادا	somebody, someone, nobody	
Numbers	صفر		0	
	نص		½	
	واحد		1	
	اثنين	إثنين - اتنين - تنين	2	
	ثلاثة	تلاتة	3	ثلاثة شباب راحوا السوق
	ثلاث	تلات	3	ثلاث حاجات
	اربعة	أربعة - اربعه - أربعه	4	
	اربع	أربع	4	
	خمسة	خمسه	5	
	خمس		6	
	ستة	سته	6	
	ست		6	
	سبعة	سبعه	7	
	سبع		7	
	ثمنية	تمنية - تمانية - تمنيه - تمانيه	8	
	ثمن	تمن	8	
	تسعة	تسعه	9	

	تسع		9	
	عشرة	عشره	10	
	عشر		10	
	احدعش		11	
	اطنعش	اتناشر - إثناشر – إتناشر اتنعش-اثنعش	12	
	ثلطعش	تلتاشر ثلاثطعش-ثلثتعش	13	
	اربعطعش	أربعتاشر اربعتعش	14	
	خمسطعش	خمستعش	15	
	ستطعش	ستعش	16	
	سبعطعش	سبعتعش	17	
	ثمنطعش	ثمانتاشر - تمانتاشر – تمنتاشر ثمانطعش- ثمنتعش	18	
	تسعطعش	تسعتعش	19	
	عشرين		20	
	ثلثين	ثلاثين - تلاتين - تلتين	30	
	اربعين	أربعين	40	
	خمسين		50	
	ستين		60	
	سبعين		70	
	ثمنين	ثمانين - تمانين - تمنين	80	
	تسعين		90	
	مية	ميه - ميت - مئة	100	
	ميتين متين	ميتين	200	
	ثلثمية	تلتميت - تلتمية - تلتميه - ثلثميت – ثلثميه-ثلاثمية	300	
	اربعمية	أربعمية	400	
	خمسمية	خمسميه	500	
	ستمية	ستميه	600	
	سبعمية	سبعميه	700	
	ثمنمية	تمنميه	800	
	تسعمية	تسعميه	900	
	ألف	الف	1000	
	آلاف	الاف - تالاف	thousands	ثمن الاف /tamantalaaf/
Days of the Week:	الاحد	الأحد	Sunday	
	الاثنين	الاتنين - الإتنين - الإثنين	Monday	
	الثلاثا	التلات التلاتا الثلاثه	Tuesday	
	الاربعا	الأربعا- الاربعه	Wednesday	
	الخميس		Thursday	
	الجمعة	الجمعه	Friday	

	السبت		Saturday	
Exceptional Verbs	أخذ	خذ - اخد – خد اخذ	took [3ms]	Hamza appears
	ياخذ يوخذ	ياخد – يأخذ-يوخد	take [3ms]	Hamza disappears
	أكل	كل - اكل	ate [3ms]	Hamza appears
	ياكل يوكل	يأكل	eat [3ms]	Hamza disappears
	كانت		was [3fs]	
	اجا	اجه-أجا	came [3ms]	
	اجت	اجات-أجت	came [3fs]	
	اجوا	اجو - أجوا	came [3p]	
	اجيت	إجيت	came [2ms]	
	اجيتي	إجيتي	came [2fs]	
	اجيتوا	اجيتو-إجيتوا	came [2p]	
	اجيت	إجيت	came [1s]	
	اجينا	إجينا	came [1p]	
	ما جوش	مجوش - ماجوش	did not come [3p]	
	اعطى انطى	أعطى	gave [3ms]	
	يعطي ينطي		give [3ms]	
	بدكو	بدكم-بدكوا	Want [2mp]	
	بده	بدو-بدا	Want [3ms]	
	بدك	بدكي	Want [2fs]	
Special Words	خالو	خاله - خالوا	Uncle	
	عمو	عموا - عمه	Uncle	
	جدو سيدو	جده - جدوا	grandfather	
	خاله		uncle + his	
	عمه		uncle + his	
	جده سيده		grandfather + his	

Table 10 Exceptional spelling choices

References

1. Habash, N., M. Diab, and O. Rambow. *Conventional Orthography for Dialectal Arabic*. in *Language Resources and Evaluation Conference*. 2012.
2. Jarrar, M., et al., *Building a Corpus for Palestinian Arabic: a Preliminary Study.* ANLP 2014, 2014: p. 18.
3. Zribi, I., et al. *A Conventional Orthography for Tunisian Arabic*. in *Language Resources and Evaluation Conference, Reykjavik, Iceland*. 2014.
4. Saadane, H. and N. Habash. *A Conventional Orthography for Algerian Arabic*. in *ANLP Workshop 2015*. 2015.