Unified Guidelines and Resources for Arabic Dialect Orthography

Nizar Habash, Fadhl Eryani, Salam Khalifa, Owen Rambow,† Dana Abdulrahim,‡
Alexander Erdmann, Reem Faraj,† Wajdi Zaghouani,* Houda Bouamor,♣
Nasser Zalmout, Sara Hassan, Faisal Al-Shargi,⋆ Sakhar Alkhereyf,†
Basma Abdulkareem,† Ramy Eskander,† Mohammad Salameh,♣ Hind Saddiki
New York University Abu Dhabi, UAE, † Columbia University, USA, ‡University of Bahrain, Bahrain,
*Hamad Bin Khalifa University, Qatar, ♣Carnegie Mellon University in Qatar, Qatar, ⋆Univisität Leipzig, Germany
nizar.habash@nyu.edu, rambow@cs.columbia.edu, darahim@uob.edu.bh,
wzaghouani@hbku.edu.qa, hbouamor@cmu.edu, alshargi@informatik.uni-leipzig.de

Abstract

We present a unified set of guidelines and resources for conventional orthography of dialectal Arabic. While Standard Arabic has well defined orthographic standards, none of the Arabic dialects do today. Previous efforts on conventionalizing the dialectal orthography have focused on specific dialects and made often ad hoc decisions. In this work, we present a common set of guidelines and meta-guidelines and apply them to 28 Arab city dialects from Rabat to Muscat. These guidelines and their connected resources are being used by three large Arabic dialect processing projects in three universities.

Keywords: Dialectal Arabic, Orthography, Phonology, Morphology, Conventions

1. Introduction

Arabic dialects are linguistic varieties that are historically related to classical Arabic and co-exist with Modern Standard Arabic (MSA) in a diglossic relationship. While MSA, the official language of all Arab countries, has well-defined orthographic standards, Arabic dialects have no official orthographies. As such, besides unintentional typographic errors, no spelling of a dialectal word can be considered "incorrect." And since Arabic dialects vary from MSA and from each other in terms of phonology, morphology, lexicon and syntax (Watson, 2007), using MSA orthographic standards cannot fully address the needs of the dialects. As an example of the degree of variety in dialectal spelling, Figure 1. presents the 27 actually attested spellings of one Egyptian Arabic word online. The large number of possibilities results from independent decisions such as whether the proclitic /ma/ should be written attached or separated (+مـ *m+*[1] or ما *mA*), or whether to write the stem in a way that reflects its phonology (ؤ ŵ), or etymology (ق *q*).

Arabic Orthography	Arabic Transliteration	Frequency
مبيقولهاش	*mbyqwlhAš*	≈ 26,000
ما بيقولهاش	*mA byqwlhAš*	≈ 13,000
مابقلهاش، مبقولهاش، مبقلهاش، ما بقلهاش، مابيقولهاش	*mAbqlhAš, mbqwlhAš, mbqlhAš, mA bqlhAš, mAbyqwlhAš*	≤ 10,000
مابقولهاش، ما بقولهاش، مبيقلهاش، ما بيقلهاش	*mAbqwlhAš, mA bqwlhAš, mbyqlhAš, mA byqlhAš*	≤ 1,000
مبئلهاش، مابيئولهاش، ما بيئولهاش، مابيؤلهاش	*mbŷlhAš, mAbyŷwlhAš, mA byŷwlhAš, mAbywŵlhAš*	≤ 100
ما بيؤلهاش، مابئلهاش، مبيئولهاش، ما بيئلهاش، مابئولهاش، ما بئلهاش، ما بؤلهاش، مبئولهاش، مبيؤلهاش، مابؤلهاش، مبؤلهاش	*mA bywŵlhAš, mAbŷlhAš, mbyŷwlhAš, mA byŷlhAš, mAbŷwlhAš, mA bŷlhAš, mA bwŵlhAš, mbŷwlhAš, mbywŵlhAš, mAbwŵlhAš, mbwŵlhAš*	≤ 10

Figure 1: 27 encountered ways to write the Egyptian Arabic word /mabiʔulha:ʃ/ 'he does not say it' and their frequencies from Google Search (September 29, 2017).

Habash et al. (2012) introduced the concept of Conventional Orthography for Dialectal Arabic (CODA); and they proposed a set of guidelines and exception lists for Egyptian Arabic. Their conventions were used in the Linguistic Data Consortium for annotating Egyptian Arabic (Maamouri et al., 2014). Since then, a number of additional efforts followed suit for other dialects (Zribi et al., 2014; Saadane and Habash, 2015; Jarrar et al., 2016; Khalifa et al., 2016). While the original CODA guidelines aimed at being easy to adjust to new dialects and contained some dialect-independent components, the guidelines were not specific enough, and often open to interpretation. Furthermore, the resources supporting the process of extending CODA to new dialects were non-existent.

Previous CODA efforts approached the conventionalization problem with a focus processing Arabic dialect text as input only. They did not address the challenge of generating Arabic dialect text for human readability (e.g., as output of speech recognition, machine translation or chatbots). Considering both aspects (input and output) highlights the need of conventions that are accessible to Arabic readers.[2]

In this work, we present a common set of guidelines with enough specificity to help in creating dialect-specific conventions, and we apply them to 28 Arab city dialects.

[1] Arabic script transliteration is presented in the Habash-Soudi-Buckwalter transliteration scheme (Habash et al., 2007):
ي و ه ن م ل ك ق ف غ ع ظ ط ض ص ش س ز ر ذ د خ ح ج ث ت ب أ
Â b t θ j H x dð 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, ى ء, ħ ة, ý ى.
Phonological forms are presented in IPA or in the CAPHI scheme, which is discussed in Section 5.

[2] Most recently, the Palestinian CODA conventions have been adopted by a website for teaching Colloquial Arabic: http://www.learnpalestinianarabic.com.

We call our new version of CODA: CODA* (pronounced *CODA Star*, as in, for *any dialect*). The contributions of this paper include: (a) the definition of a phonological representation inspired by Arpabet (Shoup, 1980) for Arabic and its dialects to be used for specifying the pronunciation in computational resources; (b) a clear separation between CODA general dialect-independent rules and specification rules for organizing and managing the numerous exceptional cases presented in previous work; (c) the introduction of the concept of a multidialectal *Seed Lexicon* that is used to allow users of CODA* to have access to previous decisions when identifying spellings for new words in new dialects; and finally, (d) a set of online pages that give users easy public access to all of these resources.

The CODA* guidelines and their connected resources are being used by three large Arabic dialect processing projects in three universities: The Multi-Arabic Dialect Applications and Resources at Carnegie Mellon University Qatar and New York University Abu Dhabi (NYUAD) (Bouamor et al., 2018), The Gulf Arabic Annotated Corpus (NYUAD) (Khalifa et al., 2018), and The Columbia Arabic Dialect Annotation project (Columbia University and NYUAD). The CODA* effort is large and ongoing; the goal of this paper is to introduce the effort and some of its important contributions on how to conceptualize and address the question of orthographic decisions in dialectal Arabic computational processing.

The rest of the paper is structured as follows. We present common challenges to Arabic processing in Section 2. This is followed by related work in Section 3. We introduce CODA* in Section 4., and discuss its components in Section 5. (CAPHI), Section 6. (General Rules and Specifications), and Section 7. (Seed Lexicon).

2. Challenges to Arabic Processing

There are four distinct and orthogonal challenges to working on written Arabic natural language processing (NLP): morphological richness, orthographic ambiguity, dialectal variations, and orthographic inconsistency.

Morphological Richness Arabic words have a large number of forms. This results from a rich inflectional morphology that models gender, number, person, aspect, mood, case, state and voice, in addition to a large number of clitics such as conjunctions, negative particles, future particles, etc. The word featured in Figure 1. is only one of a few thousand forms (inflections and cliticizations) of the verbal lemma قال *qAl* 'to say'.

Orthographic Ambiguity Arabic orthography using the Arabic script employs optional diacritical marks for short vowels and consonantal gemination. The missing diacritics are not a major challenge to literate native adults. However, their absence is the main source of ambiguity in Arabic NLP. In MSA, the average ambiguity is 2.7 lemmas per word (Habash, 2010). For example, the MSA word عقد ʕqd can be diacritized as عَقَد ʕaqad 'contract' or عُقُد ʕuqud 'necklace', among other readings.

Dialectal Variations Arabic dialects are often classified regionally (such as Egyptian, Levantine, Gulf, etc.) or subregionally (e.g., Lebanese, Syrian, Jordanian, etc.). These classifications are generally problematic because of the continuous nature of language variation. In our work, we have opted to focus on specific cities that represent the different regions and sub-region, in an effort to control the degree of variation we study. Table 1 lists the names of the cities we cover in the work presented in this paper. Arabic dialects differ significantly in terms of their phonology, morphology and lexicon from one another and from MSA (Watson, 2007).

Orthographic Inconsistency Noise in written text is a common problem for NLP when working in social media and non-edited text (see Figure 1.). For MSA, Zaghouani et al. (2014) report that 32% of words in MSA comments online have spelling errors. Eskander et al. (2013) also report close to 24% of Egyptian Arabic words having non-CODA-compliant spelling. Dialectal Arabic text is also known to appear on social media in a non-standard romanization, often called Arabizi (Darwish, 2013).

The work presented in this paper focuses primarily on the issue of orthographic inconsistency although it is inseparable from all of the other challenges.

3. Related Work

Before Habash et al. (2012) introduced their Egyptian Arabic conventional orthography (CODA-Egyptian), there were many proposals such as the Asaakir system ('Asaakir, 1950) and Akl's system (Arkadiusz, 2006), neither of which are broadly used today. Various DA dictionaries used Arabic, Latin or mixed script orthographies (Badawi and Hinds, 1986). In the context of NLP, the Linguistic Data Consortium (LDC) guidelines for transcribing Levantine Arabic (Maamouri et al., 2004) and the COLABA project at Columbia University (Diab et al., 2010) were precursors to the work of Habash et al. (2012).

After the CODA-Egyptian guidelines were created and used for the creation of Egyptian Arabic resources (Maamouri et al., 2014; Diab et al., 2014; Pasha et al., 2014; Eskander et al., 2013; Al-Badrashiny et al., 2014), two additional sets of guidelines were created for CODA-Tunisian (Zribi et al., 2014) and CODA-Palestinian (Jarrar et al., 2014). These were part of projects involving morphology annotation (Palestinian) or speech recognition (Tunisian). A variant on CODA was proposed for speech recognition by Ali et al. (2014) and was shown to reduce OOV and perplexity. Since then, four more dialects followed: CODA-Algerian (Saadane and Habash, 2015), CODA-Gulf (Khalifa et al., 2016), CODA-Moroccan and CODA-Yemeni (Al-Shargi et al., 2016). The latter efforts were heavily based on earlier versions, modifying/extending the exception lists of the Egyptian and Palestinian versions while preserving the general CODA rules. These efforts focused on one dialect at a time, and a number of them were only interested in processing Arabic input – not considering the challenges of dialectal Arabic output. Some recent efforts have highlighted the value of generating Arabic dialect text in the context of speech recognition, chatbots, and machine translation (Ali et al., 2014; Meftouh et al., 2015; Abu Ali and Habash, 2016). Erdmann et al. (2017), for instance, evaluated translation output in DA, finding that 10% of tokens not found in the reference but

Maghreb				Nile Basin		Fertile Crescent			Arabian Peninsula				
Morocco	**Algeria**	**Tunisia**	**Libya**	**Egypt**	**Sudan**	**South Levant**	**North Levant**	**Iraq**	**Hijaz**	**Najd**	**Gulf**	**Oman**	**Yemen**
Rabat	Algiers	Tunis	Tripoli	Cairo	Khartoum	Jerusalem	Beirut	Mosul	Jeddah	Riyadh	Abu Dhabi	Muscat	Sana'a
Fes		Sfax	Benghazi	Alexandria		Amman	Damascus	Baghdad			Manama		Taiz
				Aswan		Salt	Aleppo	Basra			Doha		

Table 1: The different region, sub-region, and city dialects covered in CODA*.

manually judged to be correct, were in fact orthographic variants of a corresponding reference token. Zalmout et al. (2018) trained a morphological disambiguator on a CODA-based version of the data to show the upper accuracy limit noise-wise.

4. CODA* : Conventional Orthography for Multiple Arabic Dialects

In this section, we review the design goals and principles of CODA and discuss how we extend the CODA guidelines to CODA*.

4.1. CODA Goals and Design Principles

The original CODA goals as outlined by (Habash et al., 2012) in their paper on CODA-Egyptian were that: (i) CODA is an internally consistent and coherent convention; (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. The authors also describe CODA design as a consistent *ad hoc* convention that balances being MSA-like with being generally phonemic, and morphologically and syntactically faithful to the dialect. Furthermore, it aims to be easily learnable and readable.

4.2. Extending CODA to other Dialects

Despite the stated goals and design principles, the final guidelines Habash et al. (2012) presented were for Egyptian Arabic only. They included general rules and an exception list, which is to be consulted before applying any rules. The contents of the exception list are a mix of frequent closed class words (e.g., pronouns, demonstratives), frequent adverbial expressions (e.g., *today*, *tomorrow*, etc.), number words and days of the week, with the occasional odd case or not-so obvious example that applies the general rules. The protocol of consulting the exception list before applying the general rules is clear; however, there were no guidelines on what to put on the exception list and how to write words in that list. For one dialect, this perhaps is not a problem, but as we expand to other dialects, we have no guidance for how to proceed. As such, different efforts since Habash et al. (2012) interpreted the general rules as dialect independent, and the exception list as dialect specific and *translated* the exception list, which only added more ad hoc decisions.

4.3. CODA*

What we propose in this paper is an extension of CODA, which we call CODA* (as in, for *any dialect*). CODA* includes minor adjustments to the CODA general dialect-independent rules, and major clarifications of the structure of the exception list. CODA* replaces said list with a detailed set of dialect-independent *specifications* to guide the creation of all closed classes and other categories of expressions. Additionally, CODA* introduces the concept of a *seed lexicon* which contains dialect-specific examples of frequent words, closed classes, examples, etc. The seed lexicon can grow as more work on a dialect is done to become eventually a dictionary of the dialect. As part of the effort to make the rules and lexicons easily definable and usable across all dialects in CODA*, we introduce a phonological representation that can be used to discuss and compare different dialectal entries in their respective lexicons, as phonological information is often obscured or generalized by CODA orthography. We also created a website (see Section 8.) for listing the rules and specification, and an easy-to-use interface for the seed lexicon, which currently includes multiple dialects.

In the next three sections, we present a summary of the phonological representation, the general rules and specifications, and the seed lexicon.

5. CODA* Phonological Representation

In discussing the many dialects included in CODA*, it is necessary to distinguish CODA orthography from the actual phonological properties of an utterance. Because CODA may conflate some phonological variation for the sake of pan-Arabic consistency, it is not ideal for describing dialectal phonological variation. For this purpose, we present the CAMEL[3] Arabic Phonetic Inventory (CAPHI).[4] Inspired by the International Phonetic Alphabet (IPA) and Arpabet (Shoup, 1980), CAPHI is an objective system for transcribing utterances in all dialects of Arabic in a simple, user-friendly fashion. The CAPHI inventory contains only ASCII symbols–avoiding the need to switch between multiple keyboard layouts–such that symbols intuitively correspond to the representative phone and are easily memorized. This makes it a far more attractive representation for our purposes than the otherwise popular IPA. Additionally, each transcribed phone is white space-separated with optional markers available for distinguishing morpheme (+) and word boundaries (#). This enables single phones to be written with multiple characters when it is intuitive to do so, as is the case for long vowels or affricates. The infrastructure is easily extendable to cover yet undiscovered phenomena.

[3]Computational Approaches to Modeling Language (CAMEL) Lab at New York University Abu Dhabi.

[4]The Arabic word كافي *kAfy* means 'sufficient'.

			Example			
CAPHI	IPA	Letter	CODA	CAPHI	Gloss	Dialect
p	**p**	**ب *b***	بري	p r i	price	Algiers
p.	**pˤ**	**ب *b***	بمب	p. a m. p.	pump	Baghdad
b	**b**	**ب *b***	باب	b ee b	door	Sfax
b.	**bˤ**	**ب *b***	يباباني	y aa b. aa n i	Japanese	Muscat
f	**f**	**ف *f***	فردي	f a r d i	single	Tripoli
f.	**fˤ**	**ف *f***	فَكّ	f. a k k	he opened	Baghdad
v	**v**	**ف *f***	روندیفو	r o o n d ii v uu	appointment	Algiers
t	**t**	**ت *t***	تفاح	t i f f ee 7	apple	Beirut
t	t	ث *θ*	ثامن	t aa m i n	eighth	Jeddah
t.	**tˤ**	**ط *T***	طبق	t. a b a g	dish	Riyadh
t.	tˤ	ت *t*	فستان	f u s t. aa n	dress	Amman
d	**d**	**د *d***	جديد	j d ii d	new	Algiers
d	d	ت *t*	اتدشدش	2 i d d a sh d a sh	he got smashed	Cairo
d	d	ذ *ð*	ذيل	d ee l	tail	Beirut
d	d	ض *D*	ضحك	d i 7 i k	he laughed	Cairo
d.	**dˤ**	**ض *D***	ضرس	d. i r s	tooth	Jeddah
d.	dˤ	ت *t*	اتضرب	2 i d. d. a r a b	he got hit	Cairo
d.	dˤ	د *d*	ضفدعة	d. u f d. a 3 a	frog	Cairo
d.	dˤ	ظ *Ď*	ظهرية	d. u h r i y y a	shaddow	Jeddah
th	**θ**	**ث *θ***	ثياب	th y aa b	clothes	Salt
dh	**ð**	**ذ *ð***	ذراع	dh r aa 3	arm	Muscat
dh.	**ðˤ**	**ظ *Ď***	ظاهر	dh. aa h i r	appearing	Abu Dhabi
dh.	ðˤ	ذ *ð*	ذوق	dh. oo g	tasting	Baghdad
dh.	ðˤ	ض *D*	ضغط	dh. a gh a t.	he pressed	Jerusalem
s	**s**	**س *s***	اساسي	2 a s aa s i	main	Rabat
s	s	ث *θ*	ثورة	s a w r a	revolution	Cairo
s	s	ز *z*	الزعيم	2 a s s e 3 ii m	the boss	Sana'a
s	s	ص *S*	صايغ	s aa y e gh	jeweler	Cairo
s.	**sˤ**	**ص *S***	قصّة	q i s. s. a	story	Alexandria
s.	sˤ	س *s*	سلطة	s. a l. a t. a	salad	Khartoum
z	**z**	**ز *z***	زورق	z a w r a 2	boat	Aleppo
z	z	ذ *ð*	ذبيحة	z a b ii 7 a	meat	Muscat
z	z	س *s*	اسبوع	2 i z b uu 3	week	Khartoum
z	z	ص *S*	صغير	z gh ii r	small	Beirut
z.	**zˤ**	**ز *z***	جزمة	tsh i z. m. a	boot	Baghdad
z.	zˤ	ض *D*	مضبوط	m a z. b uu t.	correct	Jeddah
z.	zˤ	ظ *Ď*	عظيم	3 a z. ii m	great	Damascus
sh	**ʃ**	**ش *š***	شاذر	sh aa dh i r	blanket	Muscat
j	**ʒ**	**ج *j***	جاب	j ee b	he brought	Beirut
j	ʒ	ق *q*	طريق	t. a r ii j	road	Baghdad
ts	**ts**	**تس *ts***	فرتس	f r i ts	Fritz	MSA
ts	ts	ك *k*	سمك	s i m a ts	fish	Riyadh(B)
ts	ts	ج *j*	كتابج	k i t aa b i ts	your [2fs] book	Riyadh(B)
dz	**dz̄**	**دز *dz***	ايدز	2 ee dz	AIDS	Beirut
dz	dz̄	ج *j*	جزاير	dz aa y i r	Algeria	Algiers
dz	dz̄	ق *q*	طريق	t. a r ii dz	road	Riyadh(B)
tsh	**tʃ**	**تش *tš***	كتشب	k a tsh tsh a p	ketchup	Sana'a
tsh	tʃ	ج *j*	عيونج	3 y uu n i tsh	your eyes [2fs]	Doha
tsh	tʃ	ش *š*	شاف	tsh aa f	he saw	Doha
tsh	tʃ	ك *k*	سمك	s i m a tsh	fish	Basra
dj	**dʒ**	**ج *j***	موجود	m a w dj uu d	available	Khartoum
dj	dʒ	ق *q*	طريق	t. i r ii dj	road	Abu Dhabi
k	**k**	**ك *k***	فاكهة	f aa k h a	fruit	Doha
k	k	ق *q*	رقم	r a k a m	number	Jerusalem(R)
g	**g**	**ج *j***	جميل	g a m ii l	beautiful	Cairo
g	g	ق *q*	قال	g aa l	he said	Aswan
g	g	ك *k*	بنك	b. a n g	bank	Baghdad

			Example			
CAPHI	IPA	Letter	CODA	CAPHI	Gloss	Dialect
gy	**gʲ**	**ج *j***	جليز	gy aa y i z	possible	Khartoum
q	**q**	**ق *q***	قطار	q i t. aa r	train	Fes
qh	**ɢ**	**ق *q***	رقم	r a qh a m	number	Khartoum
kh	**x**	**خ *x***	تَخت	t a kh i t	bed	Aleppo
kh	x	غ *ɣ*	غسالة	kh a s s ee l a	washer	Tunis
gh	**ɣ**	**غ *ɣ***	غاوي	gh aa w i	beautiful	Muscat
gh	ɣ	ر *r*	اريد	2 a gh ii d	I want	Mosul
7	**ħ**	**ح *H***	حبّ	7 a b b	he loved	Sfax
3	**ʕ**	**ع ʕ**	عمّر	3 a m m a r	he filled	Rabat
2	**ʔ**	**ء *'***	غطاء	gh t. aa 2	blanket	Tunis
2	**ʔ**	**آ *Ā***	آمين	2 aa m ii n	amen	Abu Dhabi
2	**ʔ**	**أ *Â***	متأكد	m i t 2 a k k i d	sure	Amman
2	**ʔ**	**إ *Ǎ***	إجازة	2 i dj aa z e	vacation	Sana'a
2	**ʔ**	**ؤ *ŵ***	تساؤل	t a s aa 2 u l	question	Muscat
2	**ʔ**	**ئ *ŷ***	جائز	j aa 2 i z	possible	Mosul
2	ʔ	ق *q*	طريق	t. a r ii 2	road	Damascus
h	**h**	**ه *h***	مهم	m u h i m m	important	Aswan
m	**m**	**م *m***	مكن	m a k k a n	he gave	Sana'a
m	m	ن *n*	جنب	j a m b	besides	Jeddah
m.	**mˤ**	**م *m***	ميّ	m. a y y	water	Damascus
n	**n**	**ن *n***	فنّ	f a n n	art	Baghdad
n	n	ً *ã*	فعلاً	f i 3 l a n	indeed	Cairo
n	n	ٌ *ũ*	جميلٌ	dj a m ii l u n	beautiful	MSA
n	n	ٍ *ĩ*	غصبٍ	gh a s. b i n	forcefly	Abu Dhabi
n.	**nˤ**	**ن *n***	ناي	n. aa y	flute	Damascus
r	**r**	**ر *r***	روج	r uu j	red	Tunis
r.	**rˤ**	**ر *r***	فرنسي	f a r. a n s i	french	Khartoum
l	**l**	**ل *l***	لازم	l aa z i m	necessary	Riyadh
l.	**lˤ**	**ل *l***	لطش	l. a t. a sh	he stole	Khartoum
w	**w**	**و *w***	ولد	w a l d	boy	Sana'a
y	**j**	**ي *y***	هايل	h aa y i l	great	Alexandria
y	j	ج *j*	جلس	y a l a s	he sat down	Abu Dhabi
i	**i**	**ِ *i***	مرشد	m u r sh i d	guide	Mosul
i	i	ة *ħ*	سمكة	s a m a k i	fish	Beirut
i	i	ه *h*	فواكه	f w ee k i	fruits	Beirut
i	i	ي *y*	نسي	n i s i	he forgot	Sana'a
ii	**ī**	**ي *y***	مزيكة	m a z z ii k a	music	Alexandria
e	**e**	**ِ *i***	باكر	b aa k e r	tomorrow	Muscat
e	e	ة *ħ*	ستّة	s i t t e	six	Damascus
e	e	َي *ay*	غيرهم	gh e r h o m	other than them	Cairo
ee	**ē**	**َي *ay***	ستّيشن	s t ee sh a n	station	Muscat
ee	ē	ا *iA*	حاسب	7 ee s a b	he paid	Beirut
a	**a, æ**	**َ *a***	جَرَّب	dj a r r a b	he tried	Sana'a
a	a, æ	ا *A*	سما	s a m a	sky	Cairo
a	a, æ	ة *ħ*	شيبة	sh ee b a	old man	Jeddah
a	a, æ	ه *h*	شافته	sh aa f i t a	she saw him	Doha
a	a, æ	ى *ý*	حمّى	7 o m m a	fever	Muscat
aa	**ā, æ:**	**ا *A***	دار	d aa r	house	Abu Dhabi
o	**o**	**ُ *u***	قُلت	2 o l t	I said	Cairo
o	o	ه *h*	جيبته	j i b t o	I brought it	Damascus
o	o	و *w*	الو	2 a l o	hello	Amman
oo	**ō**	**و *w***	دور	d oo r	turn	Damascus
u	**u**	**ُ *u***	جُبن	dj u b i n	cheese	Doha
u	u	ه *h*	جبته	g i b t u	I brought it	Cairo
u	u	و *w*	كتابكو	k i t a b k u	your [2p] book	Cairo
u	u	وا *wA*	كنتوا	k i n t u	you [2p] were	Beirut
uu	**ū**	**و *w***	بلوزة	b l uu z a	blouse	Jeddah

Figure 2: A detailed listing of possible pairings of CAPHI and IPA sounds with CODA letters. Each pairing is accompanied with an example in terms of CODA, CAPHI, English gloss and city dialect. The dialect chosen is not meant to be exclusionary of other city dialects. The order of presentation is based on phonological features, and it positions consonants > vowels, stops > fricatives > affricates > nasals > other, labials >> glottals, voiceless > voiced, and non-emphatic > emphatic. Highlighted cells indicate CAPHI-CODA default pairing. (B) and (R) refer to *Bedouin* and *Rural* sub-dialects.

CAPHI phones include all phonemes in any dialect and any allophones which are confusable with phonemes of another dialect. For example, the voiceless alveolar stop [t] and its pharyngealized counterpart [tˤ], are both phonemes because they can be used to distinguish meaning, as demonstrated by the minimal pair, تيّار [taj:aːr] 'current' and طيّار [tˤaj:aːr] 'pilot'. Thus both phones exist in CAPHI as /t/ and /t./ respectively. Conversely, [aˤ] and [a] are not separate phonemes despite the fact that pharyngealized vowels do occur near pharyngealized consonants. Pharyngealized vowels, however, do not distinguish meaning and are not confusable with any phonemes in any known dialect. Therefore, they are not included in CAPHI, as little descriptive power would be gained from their inclusion and annotators would be required to make a challenging, error-prone distinction. Some allophones though, are included in CAPHI, like the Iraqi [pˤ] from phoneme [p]. The pharyngealization on the voiceless bilabial stop causes this allophone to sound similar to the voiced bilabial stop [b] with

which it can be confused by speakers of dialects which do not have the phoneme [p]. Thus, p^{ς} is included in CAPHI as /p./ as it is useful in describing the dialectal differences between Iraqi and other dialects. The complete CAPHI inventory is listed in Figure 2.

6. CODA* General Rules and Specifications

While the goals of the CODA* guidelines is to precisely define the CODA choices, it is unavoidable that different versions of the guidelines will need to be presented differently for specific annotators on specific tasks for specific dialects: e.g., conversion form Arabizi to Arabic script (Bies et al., 2014), or lexicon construction (Diab et al., 2014).

In this paper, we summarize and highlight specific contributions of the effort; but the full set of CODA* guidelines is described on its online page (See Section 8.). We start with a description of the technical terminology we use; then we discuss the various rules and how to use them. The border between the general rules and the specification rules is broadly drawn along the lines that general rules do not refer to any specific lexical items (morphemes or words) and pertain to the meta-mechanics of CODA; while the specification rules are lexically specific, and at times ad hoc.

6.1. Terminology

We define the various technical terms we use in the rest of this section. For more information on Arabic morphology, see (Habash et al., 2012).

Sounds, Letters and Diacritics The term *sounds* is used in the context of pronunciation (phonology), while letters and diacritics are used in the context of writing (orthography). Sounds can be consonants or vowels. They are represented using the CAPHI representation and are bounded by forward slashes. *Letters and diacritics* are symbols used in the Arabic script to write words. Letters in the Arabic language are always required to be written; while diacritics are optional.[5]

Roots, Patterns, and other Morphemes Arabic's templatic morphology makes common reference to the concept of the *root*, a typically tri-consonantal abstraction capturing a general meaning about the word. For example, the root ك.ت.ب *k.t.b* 'writing-related' appears in words like مكتب *maktab* 'office' and كتاب *kitAb* 'book'. Each sound in the root is referred to as a *radical*. The general complement of the root is the pattern, which in the examples above are *ma12a3* and *1i2A3* (here, 1, 2, 3 are slots for the root radicals). In addition to the root and pattern templatic morphemes, Arabic uses numerous other concatenative morphemes.

Words, Basewords, and Clitics We define an Arabic *baseword* to consist of a stem and the minimal number of concatenative affixes needed to specify the obligatory features for its POS. A stem can be non-templatic or it can be composed from the interdigitation of a root and a pattern. The pattern may specify the features fully, as in broken plurals. Basewords are as such the smallest fully formed words. Examples include: كتابين *kitAb+yn* 'two books' and يكتبون *y+ktb+wn* 'they write'. Clitics are syntactically independent but phonologically dependent morphemes that are attached to the word phonologically. Words can be basewords or basewords with added clitics. We use the term *word* to refer to the phonological utterance or the orthographic string, and we specify as needed. In CODA, phonological words typically map one-to-one to orthographic words; but there are many exceptions, pertaining mostly to clitics that are spelled as separate orthographic words.

6.2. Determining CODA-Compliant Orthography

To construct the spelling of a word, one must first identify all of its components: from sounds to morphemes, basewords and clitics. The morphemes should not just be identified in terms of their form, but also in terms of their meaning and POS. Different rules, general and specific, will often apply to different parts of the word. In practice, we expect some users to try to identify the baseword and look it up in the seed lexicon first; otherwise, they should form the baseword then add the clitics and follow the rules of clitic attachment.

6.3. General Rules

The general CODA* rules are for the most part a subset of the original CODA rules (Habash et al., 2012) with minor simplifications. We summarize below some of the most important general rules.

6.3.1. Basic Phonology to Orthography Mapping

These rules cover the mapping from sounds to letters. The default mapping is indicated in Figure 2 (bolded sections). All other pairings in that table follow from other general and specification rules, some of which are discussed below.

Hamza Spelling Hamza (Glottal Stop) spelling follows from the same rules as those of MSA and the original CODA. The Hamza is represented in six letters that are conditioned on its phonological context. In previous versions of CODA, and in MSA spelling, baseword initial Hamza had complex rules for deciding its form. The rule is now simplified to ا *A* and considers the Hamzation (أ *Â*, إ *Ǎ*) optional.

Diacritic Spelling While Arabic diacritics are optional in general, they can be required in certain contexts, e.g., lemmas in the work of Khalifa et al. (2018) are diacritized. In this paper, we generally omit the diacritics unless needed. Arabic diacritics are primarily used for representing short vowels, or absence of vowels. However, the Shadda diacritic is used to represent consonantal gemination, e.g., كَتَّب *kat∼ab* /k a t t a b/ 'he dictated'. As such, using the Shadda interacts with the number of letters in a word. The Shadda general rule states that it is used within the baseword, but not across word-clitic boundaries. Any exceptions must be specified in the specification rules.

Long-Short Vowel Spelling In many dialects, baseword long vowels may be shortened in certain contexts. Gener-

[5] While sounds are represented in CAPHI, letters and diacritics will be represented in Arabic script and supplemented with a romanized transliteration (Habash et al., 2007) for non-Arabic readers.

ally, the rule is to prefer the *long* letter-based spelling over the shortened diacritic spelling.

6.3.2. Baseword Spelling

Unlike the first set of general rules discussed above, the next three rules make reference to the root and pattern morphemes.

Root Radical Spelling We expand the rule on etymologically spelled consonants discussed by Habash et al. (2012): dialectal word root radicals which have MSA cognates will be spelled using the MSA cognate radical if the dialectal radical sound and the MSA radical sounds are paired according to a specific set of common sound changes. Our expanded list of pairings is presented in Figure 3. Examples of specific words can be found in Figure 2.

CODA		CAPHI: MSA Sound	CAPHI: Cognate Dialect Variant Sound
ت	*t*	t	t.
ث	*θ*	th	t, t., s
ج	*j*	dj, g	j, y, gy, tsh
د	*d*	d	t., d.
ذ	*ð*	dh	d, dh., z
ر	*r*	r	gh
ز	*z*	z	s, s.
س	*s*	s	s., z
ش	*š*	sh	tsh
ص	*S*	s.	s, z
ض	*D*	d.	d, dh., z.
ط	*T*	t.	t
ظ	*Ď*	dh.	d., z.
ق	*q*	q	j, dz, dj, k, g, qh, 2
ك	*k*	k	ts, tsh, g
ن	*n*	n	m

Figure 3: Root Radical Map

Pattern Spelling Dialectal words with patterns that are cognates of MSA patterns will retain the spelling choice of the MSA pattern if the difference in pronunciation can be expressed using diacritics (for vowel change or absence), or if the pronunciation is a shortened form of the MSA pattern vowels.

Alif Maqsura The MSA rules for spelling the Alif-Maqsura (ى *ý*), which are sometimes based on roots and sometimes on patterns, apply in CODA*.

6.3.3. Clitic Spelling

The general rule on phonological clitic spelling is that clitics that are mapped into single letters (with possible diacritics) will be spelled attached to the word, and will not interact with the spelling of the word. The specification rules identify the exceptions to this rule.

6.4. Specification Rules

The CODA* specification rules are organized along the different POS tags, such as pronouns, conjunctions, demonstratives, etc.; and other word classes, such as number words and vocative familial expressions. We present next a few iconic examples of such specification rules. The full listing is part of the online CODA* guidelines. While in this section we use examples from specific dialects, the rules are dialect-independent. They, however, make specific reference to the morpheme POS, meaning, and pronunciation as the main determinants of how it is written in CODA.

6.4.1. The Definite Article

The Arabic definite article is always written as a proclitic +ال *Al+*, regardless of how it is pronounced. Table 2 presents a number of example cases with the definite article pronunciations bolded. As with MSA spelling, general cliticization rules apply except when following the proclitic +ل *l+*, where the article is spelled without its ا *A*. The Shadda rule is overridden in the specific context of +ال+ل *l+Al+* followed by an *l*-initial baseword (see last row in Table 2).

CODA	CAPHI	Gloss	Dialect
القمر *Alqmr*	**2 i l** 2 a m a r	'the moon'	Cairo
الشمس *Alšms*	**2 i sh** sh a m e s	'the sun'	Jerusalem
الكتاب *AlktAb*	**2 i k** k i t aa b	'the book'	Cairo
البيت *Albyt*	**l** b ee t	'the house'	Jerusalem
البيوت *Albyt*	**l e** b y uu t	'the houses'	Jerusalem
بالبيت *bAlbyt*	b e **l** b ee t	'at home'	Jerusalem
بالبيوت *bAlbywt*	b **l** e b y uu t	'at the houses'	Jerusalem
للبيت *llbyt*	l a **l** b ee t	'for the house'	Jerusalem
للبيوت *llbywt*	l a **l e** b y uu t	'for the houses'	Jerusalem
للشمس *llšms*	l a **sh** sh a m e s	'to the sun'	Jerusalem
للشموس *llšmws*	l a **l e** sh m uu s	'to the suns'	Jerusalem
اللجنة *Alljnħ*	**2 e l** l a g n a	'the committee'	Cairo
للجنة *lljnħ*	l e **l** l a g n a	'for the committee'	Cairo

Table 2: Definite Article examples.

6.4.2. The Ta-Marbuta

The *Ta-Marbuta* (ة ħ) is a secondary letter of the Arabic alphabet used to represent a particular suffix morpheme that is often (but not exclusively) associated with the feminine-singular feature (Alkuhlani and Habash, 2011). This morpheme has a number of allomorphs with differing pronunciations. Most notably, it appears as a vowel at the end of nominals, and changes to a ∼ /t/ when followed by possessive pronominal enclitics. The Ta-Marbuta should be written as ة ħ in word-final positions, regardless of its pronunciation, and following general CODA rules in non-word-final positions. See Table 3 for example cases.

CODA	CAPHI	Gloss	Dialect
حاجة *HAjħ*	7 aa g **a**	something	Cairo
حاجتي *HAjty*	7 aa g **t** i	my thing	Cairo
حاجتها *HAjthA*	7 aa g **i t** h a	her thing	Cairo
طاولة *TAwlħ*	t. aa w l **e**	table	Jerusalem
غزالة *γzAlħ*	gh a z ee l **i**	gazelle	Beirut
معلمة مدرسة	m 3 a l m **i t** # m a d r a s **e**	school teacher	Jerusalem
mʕlmħ mdrsħ	m 3 a l m **e** # m a d r a s **e**	she taught a school	Jerusalem
معلمتهم *mʕlmthm*	m 3 a l m **i t** h u m	their teacher	Jerusalem
معلماهم *mʕlmAhm*	m 3 a l m **aa** h u m	she taught them	Jerusalem

Table 3: Ta-Marbuta examples.

6.4.3. The Plural Waw

Verbal suffixes that indicate the feature *plural subject* and end with the sounds (/u/, /uu/, /o/, /oo/, and /aw/) will represent those sounds as وا *wA* (واو الجماعة 'Waw of Plurality') in word-final positions, and as و *w* when followed by other attached clitics. This rule is similar to the MSA rule, except for expanding the phonetic definition. See Table 4 for example cases.

CODA	CAPHI	Gloss	Dialect
قالوا *qAlwA*	2 aa l **u**	they said	Cairo
بيقولوا *byqwlwA*	b i y 2 uu l **u**	they say	Cairo
نقولوا *nqwlwA*	n q uu l **u**	we say	Tunis
قالوا *qAlwA*	g aa l **a w**	they said	Abu Dhabi
قالوها *qAlwhA*	2 a l **uu** h a	they said it	Cairo
ما قالوش *mA qAlwš*	m a # 2 a l **uu** sh	they did not say	Cairo
قالوا له *qAlwA lh*	2 a l **uu** # l u	they said to him	Cairo

Table 4: The Plural Waw examples.

6.4.4. Negation Clitics

The negation particle (/m a/, /m aa/) has phonologically become a proclitic in many dialects. However, it is always written as a separate particle ما *mA* except when overridden by another specification rule. One example of such a rule is the case of negated pronouns, which require the ما *mA* to be attached to the pronoun stem and does not allow repeated *Alif* letters. Table 5 presents a number of example cases.

CODA	CAPHI	Gloss	Dialect
ما قال	**m aa** # 2 aa l	he did not say	Damascus
ما قالش	**m a** # 2 a l sh	he did not say	Cairo
ما بدناش	**m a** # b i d d n aa sh	we do not want	Amman
مانيش	**m a** n ii sh	I am not	Cairo
ماهياش	**m a** h i y y aa sh	she is not	Cairo

Table 5: Negation clitic examples.

6.4.5. Prepositional Enclitics

Post-verbal and post-nominal prepositions that have phonologically become enclitics will nonetheless be spelled separately from the words they follow. The most prominent such case is the preposition +ل *l+* 'to, for' which introduces indirect verb objects in a number of dialects. Table 6 shows are a number of example cases from the dialect of Cairo.

CODA	CAPHI	Gloss
قالوها لي *qAlwhA ly*	2 a l u h aa # **l** i	they said it to me
ما قالوا ليش *mA qAlwA lyš*	m a # 2 a l u # **l** ii sh	they did not say it to me
بالنسبة له *bAlnsbħ lh*	b i n n i s b aa # **l** u	as for him

Table 6: Prepositional enclitic examples.

6.4.6. Numbers

The words for numbers in Arabic dialects are amongst the most rich in phonological variety. The rules of writing number words in CODA* add the following exceptions to the general rules:

- The sometimes reduced historical Ta-Marbuta in the middle of the teens (11-19) is always written as ت *t* regardless of its pronunciation as /t/ or /t./. It is never reduced to a Shadda diacritic.
- The sometimes reduced historical ع ς /3/ in numbers such as عشر ςšr 'ten, -teen', and تسع *ts*ς 'nine' will always be spelled as ع ς even if completely elided or turned into a vowel.
- The sometimes reduced or altered final letter of عشر ςšr 'ten, -teen' will be written as pronounced. The variation in this form marks different syntactic construction in some dialects.
- The hundreds will be written as a single word only if the *hundred* part is singular in form.
- The remnant /t/ of the historical Ta-Marbuta appearing only before Alif-initial words after number words will not be written.

The above rules apply to all number words, whether ordinal, cardinal, or fractions. Number words sometimes have different masculine and feminine forms that are used according to different dialect-specific rules. CODA guidelines do not interact with these dialect-specific decisions. Table 7 presents example cases, some of which involve interactions with other specification rules and general rules.

CODA	CAPHI	Gloss	Dialect
ثمانية *θmAnyħ*	th a m a n y e	eight	Salt
ثمانية *θmAnyħ*	t a m a n y e	eight	Amman
ثمانة *θmAnħ*	t a m aa n e	eight	Damascus
ثمان *θmAn*	t a m a n	eight X	Amman
ثمان *θmAn*	t m aa n	eight X	Damascus
ثمانتعش *θmAntςš*	t a m a n t a 3 sh	eighteen	Amman
ثمانتعش *θmAntςš*	t m a n t a 3 sh	eighteen	Damascus
ثمنتعش *θmntςš*	th m u n t. a 3 i sh	eighteen	Baghdad
ثمانتعشر *θmAntςšr*	t a m a n t aa sh a r	eighteen	Cairo
ثمانتعشر *θmAntςšr*	t a m a n t a 3 sh a r	eighteen X	Amman
ثمانتعشن *θmAntςšn*	th m a n t. aa sh e n	eighteen X	Tunis
اربعمية *Arbςmyħ*	2 a r b a 3 m i y y e	400	Amman
اربعمية *Arbςmyħ*	2 a r b a 3 m ii t	400 X	Amman
ربعمية *rbςmyħ*	r u b 3 u m i y y a	400	Cairo
اربع الاف *Arbς AlAf*	2 a r b a 3 # t a l aa f	4,000	Cairo
خمس ارباع *xms ArbAς*	kh a m a s # t i r b aa 3	five-fourths	Cairo

Table 7: Number examples.

6.4.7. Pronominal Enclitics

The set of specifications for the pronominal clitics which can serve as possessive pronouns, direct objects or indirect objects are presented in Table 8. Some of the decisions follow from the general rules, but for the most part they are intended to normalize the spelling as close as possible to the MSA variety without adding unnecessary and unresolvable ambiguity (e.g., using diacritics). It is important to point out again that this list is not dialect specific, but rather, it lists all the phonological forms of the pronominal morphemes in all dialects. The CODA* spelling for a dialect will depend on the phonology-morphology pair it corresponds to. Some of these pronouns have a large number of variants that can be ambiguous cross-dialectally. An interesting example is the case of the morpheme pronunciation /a/ which can be 3rd masculine singular in Gulf Arabic, but 3rd feminine singular in North Levantine: / k t aa b + a / can correspond to كتابه *ktAb+h* 'his book' (Abu Dhabi) or to كتابها *ktAb+hA* 'her book' (Damascus). The CODA* specification does not address how a particular dialect may organize the use of the different forms in terms of morphotactics, e.g., the possessive 2nd person singular feminine pronominal clitic is always ك+ *+k* in Tunis, and always كي+ *+ky* in Mosul; however, in Amman, it is كي+ *+ky*

post-vocalically, and ك+ *+k* otherwise. The underspecification of some features is intentional as some pronominal clitics may be used with different associated genders in different dialects, e.g., كن+ *+kn* is 2nd person plural feminine in Doha, but its is gender ambiguous in Beirut.

CODA	CAPHI	Morpheme Features
ني *ny*	n i, n e, n ii, n ee	1st Person Singular
ي *y*	i, ii, e, ee, y, y a, y e	1st Person Singular
ك *k*	k, i k, e k, k a	2nd Person Singular
كي *ky*	k i, k e, k ii, k ee	2nd Person Singular Feminine
ج *j*	tsh, i tsh, ts, i ts	2nd Person Singular Feminine
ه *h*	h, h u, u, o, a, a h, u h, *length*	3rd Person Singular Masculine
ها *hA*	h a, h aa, a, aa, h e, h ee	3rd Person Singular Feminine
نا *nA*	n a, n aa, n e, n ee	1st Person Plural
كم *km*	k u m, k o m	2nd Person Plural
كن *kn*	k u n, k o n, tsh i n	2nd Person Plural
هم *hm*	h u m, h o m, u m, o m	3rd Person Plural
هن *hn*	h u n, h o n, u n, o n	3rd Person Plural

Table 8: CODA* specifications for pronominal clitics.

6.4.8. Vocative Familial Expressions

Some of the vocative expressions used primarily for familial reference have vocalic endings that are homophonous with pronominal suffixes. These endings are spelled following the general phonology-to-orthography rules. For example, the word /3 a m m o/ in the dialect of Amman can mean 'uncle!' (spelled in CODA as عمو ʿ*mw*) or 'his uncle' (spelled in CODA as عمه ʿ*mh*).

7. CODA* Seed Lexicon

The CODA* seed lexicon is a large and growing database containing verified examples of CODA* spelling for dialectal words. The seed lexicon started, as per its namesake, as a starter kit for defining CODA for new dialects by considering earlier decisions. It minimally contains the closed class words from any dialect in it, in addition to any number of examples that come out of the specification rules (e.g., numbers, familial expressions, etc.). The current CODA* lexicon has 4,819 entries from 19 cities (average 253 per city). Some city dialects have more entries than others. As part of the work on the MADAR project, we are adding all the entries from MADAR's 25 cities, which are over 47,000 entries. Table 9 shows a few examples of the CODA* seed lexicon. The different columns in the table are as follows.

- **Category** identifies the type of the entry, as phrase, word, prefix, suffix, proclitic, or enclitic.
- **Lemma** is a dialect specific lemma that abstracts over the inflectional variants of the word.
- **CODA** is the spelling of the entry according to the CODA spelling guidelines.
- **CAPHI** is the phonological transliteration of the entry following the CAPHI guidelines in Section 5.
- **English** provides the lemmatized form of the English gloss for each entry.
- **POS** identifies the entry's part-of-speech tag following the CAMEL POS guidelines (Khalifa et al., 2018).
- **Dialect** identifies the city-based dialect for each entry.

Category	Lemma	CODA	CAPHI	English	POS	Dialect
WORD	برشا *bršA*	برشا *bršA*	b a r sh a	very	ADV	Tunis
WORD	قوي *qwy*	قوي *qwy*	2 a w i	very	ADV	Alexandria
WORD	قوي *qwy*	قوي *qwy*	2 a w i	very	ADV	Cairo
WORD	قوي *qwy*	قوي *qwy*	g a w i	very	ADV	Sanaa
WORD	كثير *kθyr*	كثير *kθyr*	k a t ii r	very	ADV	Aswan
WORD	كثير *kθyr*	كثير *kθyr*	k i t ii r	very	ADV	Cairo
WORD	كثير *kθyr*	كثير *kθyr*	k th ii gh	very	ADV	Mosul
WORD	كثير *kθyr*	كثير *kθyr*	k th ii r	very	ADV	Salt
WORD	كثير *kθyr*	كثير *kθyr*	k t ii r	very	ADV	Beirut
WORD	واجد *wAjd*	واجد *wAjd*	w aa y i d	very	ADV	Abu Dhabi
WORD	واجد *wAjd*	واجد *wAjd*	w aa g i d	very	ADV	Muscat
WORD	واجد *wAjd*	واجد *wAjd*	w aa j i d	very	ADV	Benghazi
WORD	قال *qAl*	قال *qAl*	g aa l	he said	VERB.P3MS	Abu Dhabi
WORD	قال *qAl*	نقول *nqwl*	n g uu l	we say	VERB.I1P	Abu Dhabi
WORD	قال *qAl*	نقول *nqwl*	n q uu l	I say	VERB.I1S	Tunis
ENC	ك+ *+k*	ك+ *+k*	i k	you	PRON.2FS	Jerusalem
ENC	ك+ *+k*	كي+ *+ky*	k i	you	PRON.2FS	Jerusalem
ENC	ج+ *+j*	ج+ *+j*	tsh	you	PRON.2FS	Abu Dhabi
PROC	+ش *š+*	+ش *š+*	sh a	will	PART_FUT	Sanaa
PROC	+ح *H+*	+ح *H+*	7 a	will	PART_FUT	Amman
PROC	+ه *h+*	+ه *h+*	h a	will	PART_FUT	Cairo
PROC	+غ *γ+*	+غ *γ+*	gh a	will	PART_FUT	Rabat
PROC	+ب *b+*	+ب *b+*	b	will	PART_FUT	Manama

Table 9: Examples from the CODA* seed lexicon.

The table also shows how the same word, with the same lemma and CODA representation, can be mapped to multiple phonological representations from different dialects. For example, the word كثير *kθyr* 'very' is mapped to five different phonological representations. In the online browsable version of the seed lexicon there are extra comments and notes indicating which rules were used.

8. Conclusion and Outlook

We presented a unified set of guidelines and resources for conventional orthography of dialectal Arabic. These guidelines and their connected resources are being used by three large Arabic dialect processing projects in three universities working on dialects from 28 Arab cities.

The resources are all available online at the project website: `http://resources.camel-lab.com/`.

In the future, we plan to continue expanding our guidelines and resources for the Arabic dialects we are working with and add new dialects. We also plan to annotate collections of text in their CODA forms to train and benchmark systems for automatic spelling conventionalization.

Acknowledgments

This effort has been supported by the Gulf Arabic Morphological Annotation project (New York University Abu Dhabi – research enhancement fund), and the Multi-Arabic Dialect Applications and Resources (MADAR) project (grant NPRP 7-290-1-047 from the Qatar National Research Fund – a member of Qatar Foundation). All statements made herein are solely the responsibility of the authors.

Bibliographical References

Abu Ali, D. and Habash, N. (2016). Botta: An arabic dialect chatbot. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, pages 208–212, Osaka, Japan, December.

Al-Badrashiny, M., Eskander, R., Habash, N., and Rambow, O. (2014). Automatic Transliteration of Romanized Dialectal Arabic. *CoNLL-2014*, page 30.

Al-Shargi, F., Kaplan, A., Eskander, R., Habash, N., and Rambow, O. (2016). A Morphologically Annotated Corpus and a Morphological Analyzer for Moroccan and Sanaani Yemeni Arabic. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Portorož, Slovenia.

Ali, A., Mubarak, H., and Vogel, S. (2014). Advances in dialectal arabic speech recognition: A study using twitter to. In *Improve Egyptian ASR. International Workshop on Spoken Language Translation (IWSLT*. Citeseer.

Alkuhlani, S. and Habash, N. (2011). A Corpus for Modeling Morpho-Syntactic Agreement in Arabic: Gender, Number and Rationality. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL'11)*, Portland, Oregon, USA.

Arkadiusz, P. (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.

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

Badawi, E.-S. and Hinds, M. (1986). *A Dictionary of Egyptian Arabic*. Librairie du Liban.

Bies, A., Song, Z., Maamouri, M., Grimes, S., Lee, H., Wright, J., Strassel, S., Habash, N., Eskander, R., and Rambow, O. (2014). Transliteration of arabizi into arabic orthography: Developing a parallel annotated arabizi-arabic script sms/chat corpus. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, pages 93–103.

Bouamor, H., Habash, N., Salameh, M., Zaghouani, W., Rambow, O., Abdulrahim, D., Obeid, O., Khalifa, S., Eryani, F., Erdmann, A., and Oflazer, K. (2018). The MADAR Arabic Dialect Corpus and Lexicon. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, May.

Darwish, K. (2013). Arabizi Detection and Conversion to Arabic. *CoRR*.

Diab, M., Habash, N., Rambow, O., AlTantawy, M., and Benajiba, Y. (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*.

Diab, M. T., Al-Badrashiny, M., Aminian, M., Attia, M., Elfardy, H., Habash, N., Hawwari, A., Salloum, W., Dasigi, P., and Eskander, R. (2014). Tharwa: A large scale dialectal arabic-standard arabic-english lexicon. In *LREC*, pages 3782–3789.

Erdmann, A., Habash, N., Taji, D., and Bouamor, H. (2017). Low Resourced Machine Translation via Morpho-syntactic Modeling: The Case of Dialectal Arabic. In *MT Summit*.

Eskander, R., Habash, N., Rambow, O., and Tomeh, N. (2013). Processing Spontaneous Orthography. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, Atlanta, GA.

Habash, N., Soudi, A., and Buckwalter, T. (2007). On Arabic Transliteration. In A. van den Bosch et al., editors, *Arabic Computational Morphology: Knowledge-based and Empirical Methods*. Springer.

Habash, N., Diab, M., and Rabmow, O. (2012). Conventional Orthography for Dialectal Arabic. In *Proceedings of LREC*, Istanbul, Turkey.

Habash, N. Y. (2010). *Introduction to Arabic natural language processing*, volume 3. Morgan & Claypool Publishers.

Jarrar, M., Habash, N., Akra, D., and Zalmout, N. (2014). Building a Corpus for Palestinian Arabic: a Preliminary Study. *ANLP 2014*, page 18.

Jarrar, M., Habash, N., Alrimawi, F., Akra, D., and Zalmout, N. (2016). Curras: an annotated corpus for the Palestinian Arabic dialect. *Language Resources and Evaluation*, pages 1–31.

Khalifa, S., Habash, N., Abdulrahim, D., and Hassan, S. (2016). A Large Scale Corpus of Gulf Arabic. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Portorož, Slovenia.

Khalifa, S., Habash, N., Eryani, F., Obeid, O., Abdulrahim, D., and Kaabi, M. A. (2018). A morphologically annotated corpus of emirati arabic. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, May.

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

Maamouri, M., Bies, A., Kulick, S., Ciul, M., Habash, N., and Eskander, R. (2014). Developing an egyptian arabic treebank: Impact of dialectal morphology on annotation and tool development. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*. European Language Resources Association (ELRA).

Meftouh, K., Harrat, S., Jamoussi, S., Abbas, M., and Smaili, K. (2015). Machine translation experiments on PADIC: A parallel Arabic dialect corpus. In *The 29th Pacific Asia conference on language, information and computation*.

Pasha, A., Al-Badrashiny, M., Kholy, A. E., Eskander, R., Diab, M., Habash, N., Pooleery, M., Rambow, O., and Roth, R. (2014). MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic. In *In Proceedings of LREC*, Reykjavik, Iceland.

Saadane, H. and Habash, N. (2015). A Conventional Orthography for Algerian Arabic. In *ANLP Workshop 2015*, page 69.

Shoup, J. E. (1980). Phonological aspects of speech recognition. *Trends in speech recognition*, pages 125–138.

Watson, J. C. (2007). *The Phonology and Morphology of Arabic*. Oxford University Press.

Zaghouani, W., Mohit, B., Habash, N., Obeid, O., Tomeh, N., Rozovskaya, A., Farra, N., Alkuhlani, S., and Oflazer, K. (2014). Large scale Arabic error annotation: Guidelines and framework. In *International Conference on Language Resources and Evaluation (LREC 2014)*.

Zalmout, N., Erdmann, A., and Habash, N. (2018). Noise-robust morphological disambiguation for dialectal arabic. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2018)*.

Zribi, I., Boujelbane, R., Masmoudi, A., Ellouze Khmekhem, M., Hadrich Belguith, L., and Habash, N. (2014). A Conventional Orthography for Tunisian Arabic. In *In Proceedings of LREC*, Reykjavik, Iceland.