In Arabic Computational Morphology: Knowledge-based and Empirical Methods. Soudi, Abdelhadi; van den Bosch, Antal; Neumann, Günter (Eds.), 2007. ISBN: 978-1-4020-6045-8

On Arabic Transliteration

Nizar Habash, Abdelhadi Soudi and Tim Buckwalter

1. Introduction

In this chapter, we introduce the transliteration scheme used in this book to represent Arabic words for readers who cannot read the Arabic script. We follow the definition of the terms *transcription* and *transliteration* given by Beesley [1]: the term *transcription* denotes an orthography that characterizes the phonology or morpho-phonology of a language, whereas the term *transliteration* denotes an orthography using carefully substituted orthographical symbols in a one-to-one, fully reversible mapping with that language's customary orthography. This specific definition of transliteration is sometimes called a "strict transliteration" or "orthographical transliteration" [1]. [1]

For Arabic script, as used in writing Modern Standard Arabic, there are many ways to define the orthographic symbol set. The basic Arabic Alphabet has 28 letters and eight diacritical marks. However, there are eight additional symbols that can be treated as separate letters and/or special combinations of letter and additional diacritics. One example is the Hamza (همزة hamzaħ) which can be a separate letter (ء) or can combine with other letters: أ, ؤ, ئ. As a result, it is possible to define an orthographic symbol set of Arabic where the Hamza is not just a letter but also a diacritic with a limited number of combinations. In fact, standard computer encodings of Arabic, such as CP1256, ISO-8859, and Unicode[2], do not do that. They all consider the additional eight symbols as separate letters.[3]

The Buckwalter Arabic transliteration [2] is a transliteration system that follows the standard encoding choices made for representing Arabic characters for computers. The Buckwalter transliteration has been used in many publications in natural language processing and in resources developed at the Linguistic Data Consortium (LDC). The main advantages of the Buckwalter transliteration are that it is a strict transliteration (i.e., one-to-one) and that it is written in ASCII characters. However, the Buckwalter transliteration is not intuitively easy to read. We avoid this

[1] We do not consider non-one-to-one schemes because of their potential to add ambiguity [1] or become excessively cumbersome to deal with, e.g., SATTS transliteration [6].

[2] Unicode actually implements both approaches, but the use of Hamza as a diacritic in Unicode is not that common to our knowledge.

[3] Arabic letters also have multiple shapes (allographs) that fully depend on their position in the word, e.g. the letter ع has the forms عـ, ـعـ, ـع, and ع in its initial, middle, final and standalone positions, respectively. There are also additional ligatures such as لا for ل+ا . We do not discuss the possibility of defining an orthographic symbol set that considers allographs and ligatures as base symbols since Arabic's simple graphotactic rules are abstracted away in all of its standard encodings. Considering sub-letter dots in Arabic as separate symbols is also not discussed for the same reason.

problem in our transliteration by extending the Buckwalter transliteration scheme to include non-ASCII characters of which the pronunciation is easier to remember. Since non-ASCII characters are less accessible through standard American keyboards (as opposed to ASCII characters), this is a clear trade-off between typing/coding simplicity and ease of readability. In terms of choices made, the transliteration used in this book is a modified Buckwalter transliteration that keeps most of Buckwalter's choices, and only modifies those characters deemed most problematic to read, e.g., * for ذ /ð/, v for ث /θ/, and symbols for nunation.

To our knowledge, there has been no earlier attempt to create a one-to-one transliteration of the Arabic script that is complete and easy-to-read, and that is consistent with Arabic computer encodings. Almost all of the previously created schemes to represent Arabic characters for western readers have focused on representing phonology and morphology (transcription) or mixing between phonology and orthography, making exceptions for transliteration of morphemes such as the definite article [1,4]. One transliteration scheme, ISO 233, gets close to achieving our goal except that it is not consistent with computer encodings, e.g., the Hamza (همزة hamzaħ) is treated as a diacritic that combines with other characters [5].

In the next section, we introduce our transliteration scheme for Arabic script as used in Modern Standard Arabic. In Section 3, we present guidelines for pronunciation of Arabic using this transliteration scheme and address various idiosyncrasies of Arabic orthography.

2. This Book's Arabic Transliteration Scheme

Table 1 provides the full list of Arabic transliterations used in this book. The first three columns contain the symbols for the characters in Arabic script and contrast our transliteration with the Buckwalter transliteration. Cases where our transliteration differs from Buckwalter's are highlighted. The last four columns present English-glossed examples in Arabic script, our transliteration, and a phonological transcription. Since Arabic words can be written fully diacritized, partially diacritized or non-diacritized, a transliteration should preserve fully how an Arabic word is constructed. This includes preserving all possible ambiguities. For example, the Arabic word كتب ktb could represent one of many diacritized words with different meanings and pronunciations: the noun كُتُب kutub '*books*' or the verb كَتَب katab '*he wrote*' among others. Of course, most naturally occurring Arabic text is not diacritized; however, in this book, diacritized transliterations are always used for readability unless the point is to discuss diacritization ambiguity. In Table 1, we show examples in fully diacritized transliteration to contrast with the corresponding transcriptions, but the Arabic text examples are not fully diacritized.[4]

[4] In this book, there are very few cases that slightly deviate from our transliteration scheme:

(i) A special transcription variant is used where the one-to-one transliteration of the Arabic script interferes with the author's explanation of some linguistic phenomena. For example, in chapter 4 "A syllable-based account of Arabic morphology", the author represents vowel lengthening and gemination by vowel doubling and consonant doubling, respectively. The use of transliteration to represent these phenomena would interfere with the syllabification process. In some cases, the authors use a phonetic transcription.

(ii) Snapshots from LDC resources that use the Buckwalter transliteration are presented in the Buckwalter transliteration. This is done in some of the chapters in the empirical part of the book.

In Arabic Computational Morphology: Knowledge-based and Empirical Methods. Soudi, Abdelhadi; van den Bosch, Antal; Neumann, Günter (Eds.), 2007. ISBN: 978-1-4020-6045-8

Table 1: This Book's Arabic Transliteration Scheme with Examples

Characters			**Examples**			
Arabic	**Transliteration**	**Buckwalter**	**Arabic**	**Transliteration**	**Transcription**	**Gloss**
ء	**'**	**'**	سماء	samaA'	/samā'/	*sky*
آ	**Ā**	**|**	آمن	Āmana	/'āmana/	*he believed*
أ	**Â**	**>**	سَأل	saÂala	/sa'ala/	*he asked*
ؤ	**ŵ**	**&**	مؤتمر	muŵtamar	/mu'tamar/	*conference*
إ	**Ă**	**<**	إنترنت	Ăintarnit	/'intarnit/	*internet*
ئ	**ŷ**	**}**	سائل	saAŷil	/sā'il/	*liquid*
ا	**A**	**A**	كان	kaAna	/kāna/	*he was*
ب	**b**	**b**	بريد	bariyd	/barīd/	*mail*
ة	**ħ**	**p**	مكتبة	maktabaħ maktabaħũ	/maktaba/ /maktabatun/	*library* *a library* [nom.]
ت	**t**	**t**	تنافس	tanaAfus	/tanāfus/	*competition*
ث	**θ**	**v**	ثلاثة	θalaAθaħ	/θalāθa/	*three*
ج	**j**	**j**	جميل	jamiyl	/jamīl/	*beautiful*
ح	**H**	**H**	حاد	HaAd	/Hād/	*sharp*
خ	**x**	**x**	خوذة	xawðaħ	/xawða/	*helmet*
د	**d**	**d**	دليل	daliyl	/dalīl/	*guide*
ذ	**ð**	*****	ذهب	ðahab	/ðahab/	*gold*
ر	**r**	**r**	رفيع	rafiyς	/rafīς/	*thin*
ز	**z**	**z**	زينة	ziynaħ	/zīna/	*decoration*
س	**s**	**s**	سماء	samaA'	/samā'/	*sky*
ش	**š**	**$**	شريف	šariyf	/šarīf/	*honest*
ص	**S**	**S**	صوت	Sawt	/Sawt/	*sound*
ض	**D**	**D**	ضرير	Dariyr	/Darīr/	*blind*
ط	**T**	**T**	طويل	Tawiyl	/Tawīl/	*tall*
ظ	**Ď**	**Z**	ظلم	Ďulm	/Ďulm/	*injustice*
ع	**ς**	**E**	عمل	ςamal	/ςamal/	*work*
غ	**γ**	**g**	غريب	γariyb	/γarīb/	*strange*
ف	**f**	**f**	فيلم	fiylm	/fīlm/	*movie*
ق	**q**	**q**	قادر	qaAdir	/qādir/	*capable*
ك	**k**	**k**	كريم	kariym	/karīm/	*generous*
ل	**l**	**l**	لذيذ	laðiyð	/laðīð/	*delicious*
م	**m**	**m**	مدير	mudiyr	/mudīr/	*manager*
ن	**n**	**n**	نور	nuwr	/nūr/	*light*
ه	**h**	**h**	هول	hawl	/hawl/	*devastation*

و	**w**	**w**	وصل	waSl	/waSl/	*receipt*
ى	**ý**	**Y**	على	ςalaý	/ςala/	*on*
ي	**y**	**y**	تين	tiyn	/tīn/	*figs*
َ	**a**	**a**	دَهَنَ	dahana	/dahana/	*he painted*
ُ	**u**	**u**	دُهِنَ	duhina	/duhina/	*it was painted*
ِ	**i**	**i**	دُهِنَ	duhina	/duhina/	*it was painted*
ً	**ã**	**F**	كتاباً	kitaAbAã	/kitāban/	*a book* [nom.]
ٌ	**ũ**	**N**	كتابٌ	kitaAbũ	/kitābun/	*a book* [acc.]
ٍ	**ĩ**	**K**	كتابٍ	kitaAbĩ	/kitābin/	*a book* [gen.]
ّ †	~	~	كَسَّرَ	kas~ara	/kassara/	*he smashed*
ْ ‡	**.**	**o**	مَسْجِد	mas.jid *or* masjid	/masjid/	*mosque*
ـ §	–	–	مَسْـــجِد	mas.____jid	/masjid/	*mosque*

† Shadda (شدة šad~aħ) is a symbol marking consonant doubling.
‡ Sukun (سكون sukuwn) is a symbol marking lack of vowel. It can be used for contrastive purposes in the transliteration. However, it is not required in this book for the purpose of improving readability.
§ Tatweel (تطويل taTwiyl) or Kashida (كشيدة kašiydaħ) is an orthographic elongation symbol with no phonetic value.

3. Pronunciation Guidelines

Arabic script, as is used in Modern Standard Arabic, is *mostly* a phonemic system with one-to-one mappings of sounds to letters and diacritics. When fully diacritized, Arabic is *almost* perfectly phonologically reproducible by readers given the following few rules and exceptions:

1. **For most *consonants*,** there is no issue in mapping letters to sounds. Some are easier for English speakers than others. The transcription and transliteration are the same for these cases. Table 2 below describes how to pronounce these consonants.

Table 2: Arabic Consonant Pronunciation

Arabic	Transliteration	Pronunciation
ب	**b**	<u>B</u>oy
ت	**t**	<u>T</u>oy
ث	**θ**	<u>Th</u>ree
ج	**j**	<u>J</u>ordan
ح	**H**	Voiceless pharyngeal fricative. Sounds like a sharp <u>h</u>.
خ	**x**	Scottish Lo<u>ch</u>; Yiddish <u>Ch</u>utzpa;
د	**d**	<u>D</u>oor
ذ	**ð**	<u>Th</u>e
ر	**r**	<u>R</u>oad

ز	**z**	Zoo	
س	**s**	Sue	
ش	**š**	Shoe	
ص	**S**	Emphatic s	*Emphasis is a bass effect giving an acoustic impression of hollow resonance to the basic sounds* [3].
ض	**D**	Emphatic d	
ط	**T**	Emphatic t	
ظ	**Ď**	Emphatic ð	
ع	**ς**	Voiced pharyngeal fricative. Sounds like a sharp a.	
غ	**γ**	Parisian French r	
ف	**f**	Film	
ق	**q**	Uvular stop. Sounds like a deep k.	
ك	**k**	Kite	
ل	**l**	Cool	
م	**m**	Man	
ن	**n**	Man	
ه	**h**	Hot	
و	**w**	Would	
ي	**y**	Yoke	

2. **The consonant Hamza** (همزة hamzaħ) has multiple forms in Arabic script. There are complex rules for Hamza spelling that depend on its vocalic context. For a reader, however, all of these forms are pronounced the same way: a glottal stop as in the value of 'tt' in the London Cockney pronunciation of bottle. Table 3 relates the different forms of Hamza in Arabic script and our transliteration. The form of the transliteration is intended to evoke the form used in the Arabic variant as much as possible. For instance, a circumflex is used with A (ا), w (و) and y (ي) to create their corresponding hamzated forms Â (أ), ŵ (ؤ) and ŷ (ئ).

Table 3: Hamza (Glottal Stop) Forms

Arabic	ء	آ	أ	ؤ	إ	ئ
Transliteration	**'**	**Ā**	**Â**	**ŵ**	**Ǎ**	**ŷ**

3. Arabic has three short vowel **diacritics** that are represented using a, u and i. Arabic also has three *nunation* diacritics. These are short vowels pronounced followed by an /n/. They are not nasalized vowels. Nunation is a mark of nominal indefiniteness in Standard Arabic. Finally, Arabic has a consonant doubling diacritic which repeats previous consonant and also a diacritic for marking when there is no diacritic. Table 4 lists these diacritics, their names, and corresponding transliteration and transcription values. Diacritics are largely restricted to religious texts and Arabic language school textbooks. In this respect, the Arabic writing system depends on the background knowledge of the reader to accurately pronounce the written word—much as a reader in English needs to decide on the basis of context whether "read" is pronounced /rīd/ (present tense) or /rɛd/ (past tense).

Table 4: Arabic Diacritics

Diacritic	Name	Transliteration	Transcription
َ	فتحة fatHaħ	a	/a/
ُ	ضمة Dam~aħ	u	/u/
ِ	كسرة kasraħ	i	/i/
ً	تنوين فتح tanwiyn fatH	ã	/an/
ٌ	تنوين ضم tanwiyn Dam~	ũ	/un/
ٍ	تنوين كسر tanwiyn kasr	ĩ	/in/
ّ	شدة šad~aħ	~	Double previous consonant
ْ	سكون sukuwn	.	No vowel

4. **Long vowels and diphthongs** in Arabic are indicated by a combination of a short vowel and a consonant. Table 5 lists the various Arabic long vowels and diphthongs together with their transliteration and transcription.

Table 5: Long vowels and diphthongs

Arabic	ـَا	ـُو	ـِي	ـَو	ـَي
Transliteration	aA	uw	iy	Aw	ay
Transcription	/ā/ (long a)	/ū/ (long u)	/ī/ (long i)	/aw/	/ay/

5. **The letter Alif** (ا A) is used to (a.) hold vowels at the beginning of words, (b.) represent the long vowel /ā/, and (c.) mark a couple of morphophonemic symbols in which Alif is not pronounced (See note 8 below).

6. **The /tā' marbūTa/** (تاء مربوطة tA' marbuwTaħ), ة ħ, is typically a feminine ending. It can only appear at the end of a word and can only be followed by a diacritic. In standard Arabic it is always pronounced as /t/ unless it is not followed by a diacritic, in which case it is silent.[5]

7. **The /alif maqSūra/** (ألف مقصورة Âlif maqSuwraħ), ى ý, is a dotless Ya (ي y). In standard Arabic, it is silent and always follows a short vowel **a** at the end of a word. For example, روى rawaý '*to tell a story*' is pronounced /rawa/.[6]

8. There are few **exceptions** to the guidelines above:
   a. The Arabic **definite article**, ال Al /al/, is a prefix that assimilates to the first consonant in the noun it modifies if this consonant is an alveolar or dental sound (except for ج j). This set of letters is called Sun Letters. They include ت t, ث θ, د d, ذ ð, ر r, ز z, س s, ش š, ص S, ض D, ط T, ظ Ď, ل l, and ن n. For example, the word الشمس Alšams '*the sun*' is pronounced /aššams/ not */alšams/. The rest of the consonants are called Moon

[5] In modern dialects of Arabic, the /*tā' marbūTa*/ is always silent except when the noun ending with it is part of an (/*'idāfa*/ إضافة ĂiDAfaħ) compound, in which case it is pronounced as /t/.

[6] In some Arab countries such as Egypt, a common orthographic variation is to use ى ý for the letter ي y in word-final position. Orthographic variation in Arabic is further discussed in chapter 3.

Letters; the definite article is not assimilated with them. For example, the word القمر Alqamar ‘*the moon*’ is pronounced /alqamar/ not */aqqamar/.

b. A silent Alif appears in the morpheme +وا +uwA /ū/ which indicates masculine plural conjugation in verbs. Another silent Alif appears after some *nunated* nouns, e.g., كتاباً kitaAbAã /kitāban/. In some poetic readings, this Alif can be produced as the long vowel /ā/: /kitābā/.
c. A common odd spelling is that of the proper name عمرو ςamrw /ςamr/ ‘*Amr*’ where the final w is silent.

3. Conclusion

In this chapter, we presented the transliteration scheme used in the rest of this book. This transliteration is a one-to-one easy-to-read complete transliteration of the Arabic script consistent with Arabic computer encodings. We also presented guidelines for pronouncing Arabic given this transliteration. We hope that this transliteration scheme will become a standard to follow in the natural language processing research community working on Arabic.

Acknowledgements

We would like to thank Ali Farghaly, Joseph Dichy and Owen Rambow for helpful discussions.

References

1. Beesley, Kenneth. Romanization, Transcription and Transliteration. <http://www.xrce.xerox.com/competencies/content-analysis/arabic/info/romanization.html>
2. Buckwalter, Timothy. Arabic Transliteration. <http://www.qamus.org/transliteration.htm>
3. Holes, Clive. Modern Arabic: Structures Functions and Varieties. Georgetown University Press, Washington D.C., 2004.
4. Wikipedia contributors. Arabic Transliteration. *Wikipedia, The Free Encyclopedia*, <http://en.wikipedia.org/wiki/Arabic_transliteration> [accessed June 19, 2006]
5. Wikipedia contributors. ISO 233. *Wikipedia, The Free Encyclopedia*, <http://en.wikipedia.org/wiki/ISO_233> [accessed June 19, 2006]
6. Wikipedia contributors. Standard Arabic Technical Transliteration System. *Wikipedia, The Free Encyclopedia*, <http://en.wikipedia.org/wiki/SATTS> [accessed June 19, 2006]