

Roman Transliteration of Arabic Script in Unicode: A Project Report

By

Dr. Abdul Kabir Hussain Solihu
International Islamic University Malaysia

Introduction

- Transliteration refers to the act or product of representing or spelling (words, letters, or characters of one language) in the corresponding letters or characters of another language or alphabet.

- Transliteration is needed mainly for cataloging and information retrieval, for writing of personal and geographical names and for explaining terms/concepts of a language which cannot be accurately translated into another language.

Statement of the Problem

- The existing transliteration technology from Arabic script into Roman script is not made for individual users/researchers, and what is available for them is of poor quality or has become outdated. The encoding systems of letters with diacritical marks of the commonly used transliteration fonts often conflict with one another and many are not compliant with the international Unicode standard. Two encodings can use the same number for two *different* characters, or use different numbers for the *same* character.

For example, the code 003E has been used by different transliteration fonts for different characters

Font name	003E “>”
<i>Times New Arabic</i>	<i>āmil</i>
<i>TranslitLS</i>	<i>a‘mil</i>
<i>Sabondiacritic</i>	<i>AṬmil</i>
<i>Times Di</i>	<i>a>mil</i>

- On the other hand, a good command of both English and Arabic is a necessary for Romanization of Arabic script; yet not everyone who masters both languages knows how to transliterate Arabic alphabet into Roman alphabet properly.
- Transliteration is a special skill that requires special expertise in the script of the original language and the script of the target language.

- We provided a technique to enhance and accelerate the transliteration process through a manual or automatic transliteration. Automatic transliteration, requires no knowledge of the rules of Romanization, though knowledge of the Arabic grammar will improve the transliteration accuracy.

ROTAS Kit

```
graph TD; A[ROTAS Kit] --- B[1. Automatic Transliteration]; A --- C[2. Manual Transliteration]; A --- D[3. Unicode Compliant Fonts];
```

1. Automatic
Transliteration

2. Manual
Transliteration

3. Unicode
Compliant Fonts

1. Automatic Transliteration

A. Transliteration Method

- In automatic transliteration, we have developed a transliteration algorithm using the writing-based model of the Arabic orthography and have implemented it in macros for Microsoft Word users. Each Arabic character code is treated alone and in its grammatical position in a word or sentence, which is then replaced by a corresponding Roman character code based on certain rules that we defined.

Table 1:
Transliteration
Table:
Consonants

Arabic	Roman	Arabic	Roman	Arabic	Roman	Arabic	Roman
	b		dh		ṭ		l
t		r		ẓ		m	
th		z				n	
j		s		gh		h	
ḥ		sh		f		w	
kh		ṣ		q			
d		ḍ		k		y	

Table 2:
Transliteration
Table Vowels
and Diphthongs

ā	a	ā	ā	aw	uww, ū (in final position)
ī	i	ī	ī	ay	iyy, ī (in final position)
ū	u	ū	ū		

Table 3:
Difference
between
ALA-LC &
ROTAS
Transliteratio
n Schemes

Character	word	ROTAS	ALA-LC
◌ِ		islāmiyya h	islāmīyah
◌ِ◌ِ		quwwah	qūwah
◌ِ (alif maqṣūrah)		muṣṭafā	muṣṭafá
'(prime)		adham	ad'ham

B. Vocalization

- Our transliteration algorithm requires vocalization. Short vowels are rarely written; yet they are always presumed. There must be short vowels in order for the computer to understand the text; otherwise the text will remain highly ambiguous.

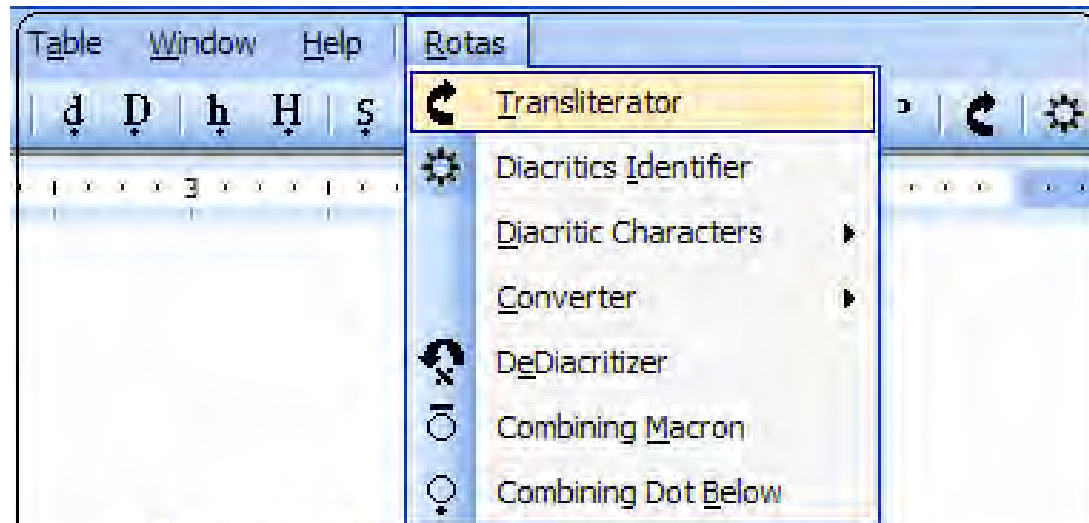
- For example, the word علم can mean different things depending on the type of short vowels it acquires. The same word can be pronounced as عِلْم (science), عَلَم (flag), or عَلِمَ he knew, عَلِمَ (it was known), عَلَّمَ (he taught) or عَلِّمَ (he was taught). Similarly, the word دين can be pronounced as دَيْن (debt) or as دِين (religion).
- It is not possible for any artificial intelligence to come out with the exact pronunciation or meaning of a word or its pronunciation without a presumed vocalization.

- While vocalization is necessary for a correct transliteration of the Arabic text, there are certain occasions where certain letters/words need not be vocalized. We divided letters/words according to their need for vocalization into four categories:

1. Letters that must be vocalized: All letters when they require any of the three short-vowels: َ ِ ُ.

2. Letters that must not be vocalized: ال (definite article); ة in non-construct state; لله, تالله, بالله, اللهم, مائة, الله, بسم الله الرحمن الرحيم, اللذين, الذين, أولئك, هؤلاء, هذه, هذا, تلك, ذلك, إله, اللاتي, الذي, التي,

- 3. Letters that make no difference whether vocalized or not:** letters when they require *sukūn* ◌.
- 4. Word/Letters that must be vocalized twice (i.e. ◌◌ , or ◌◌ ,)** : Prefixed prepositions or conjunctions, like و and بِ in **وَزَيْدٍ** and **بِالْبَيْتِ** respectively.



[Original text]

الحِكْمَةُ ضَالَّةٌ الْمُؤْمِنُ يَأْخُذُهَا حَيْثُ وَجَدَهَا. مَنْ جَدَّ
 وَجَدَ. وَيَبْقَى الْوُدُّ مَا بَقِيَ الْعِتَابُ. الْكُلِّيَّاتُ الْخَمْسُ.

[Transliterated text]

al-hikmah dāllat al-mu'min ya'khudhuhā
 haythu wajadahā. man jadda wajad. wa-
 yabqā al-wudd mā baqiya al-'itāb. al-
 kulliyāt al-khams.

Strategy for Conversion and Backward Compatibility

- Because of the conflicting code of these early fonts, we divided the converter into three macros. Fonts that have conflicting character codes were set apart.
 1. The first macro will convert four fonts, namely, *AHT Times New Roman*, *Times Di*, *Times New Arabic* and *Islamicstudies*.
 2. The second macro will convert *Sabondiacritic* and *Transliteration1* fonts.
 3. The third macro will convert the *TranslitLS* font.

- This division was based on whether or not they have conflicting codes. Although the character codes of font(s) under each template may not support one another, they must not have conflicting character codes.
- Characters that will be converted are:
“āĀīĪūŪđĐḥḤṣṢṭṬẏẐ ᶜ ᶞ”

2. Manual Transliteration

- Manual transliteration is useful for users who already know the rules of Romanization or do not have Arabic support in their Microsoft Windows. In manual transliteration, a user simply inserts the relevant Roman letters with diacritical marks

Manual Transliteration

```
graph TD; A[Manual Transliteration] --- B[Diacritical Character Buttons]; A --- C[Transliteration On-Screen Keyboard]
```

Diacritical Character
Buttons

Transliteration
On-Screen Keyboard

Diacritical Character Buttons

- Diacritical Character Buttons provide shortcut buttons for ten Roman letters with diacritical marks which are ā Ā ī Ī ū Ū đ Đ ħ Ħ š Š ṭ Ṭ ẓ Ẓ ‘ ’.
- Two supplementary buttons are provided in the ROTAS menu command, one for combining macron and the other for combining dot below.



Diacritics Character Buttons

Transliterator
Diacritic Identifier

Transliteration Keyboard

- The Rotas Transliterater Keyboard is an On-Screen keyboard that one can use to type any word in Arabic or English, copy it to the clipboard and then paste it into any application that can display them.
- This means that if one wants to type an Arabic word, the Microsoft Windows operating system and the Windows application to which the word is to be copied must support Arabic script

Rotas Tansliterator Keyboard

ء	b ب	t ت	th ث	j ج	h ح	kh خ	d د	dh ذ	r ر	z ز
s س	sh ش	s ص	d ض	t ط	z ظ	ع	gh غ	f ف	q ق	k ك
ا	m م	n ن	h ه	w و	y ي	ā ا	ī ي	ū و	a ا	i ا
u	an	in	un	g	e	c	p	x	v	o

CAPS COPY SPACE CLEAR ENG

الحِكْمَة ضَالَّة الْمُؤْمِن يَأْخُذْهَا حَيْثُ وَجَدَهَا
 al-ḥikmah ḍāllat al-mu'min ya'khudhuhā ḥaythu
 wa jadahā

Interface2
Help
About

ROTMAS
رَوْتِمَاسِي
Keyboard
لوحة المفاتيح

3. Unicode Compliant Fonts

- To make the software self-contained, we included a set of family fonts (serif, sans-serif and monospace), all of which are Unicode compliant.
- Diacritic coverage of some of these fonts includes Basic Latin, Latin-1, Latin Extended-A, Latin Extended-B and Latin Extended Additional.

- The fonts support several transliteration schemes, including *ISO 233*, *Encyclopedia of Islam* (1960), *United Nations Romanization Systems for Geographical Names* (1972), *British Standard BS 4280: 1968* and most importantly *ALA-LC Romanization Tables*.

- Each family font has Roman (regular), Bold, Italic (oblique) and Bold Italic typefaces. Rotas Gombak Naskh font has only one typeface and it is the only font in Rotas Gombak family fonts that is primarily made for Arabic script.

Conclusion

- Our primary objective in undertaking the project is to provide the transliteration facilities for researchers on Islamic studies writing for English readers.
- The software is open for further developments, especially in the field of vocalization
- Another area of improvement is in the scope of the script covered.

Thank you