Technical Notes on Arabic Romanization (2007)[1]

Note. In this document, the **UN 1972 System** refers to the amended Beirut system adopted by Arabic experts at a conference in Beirut 1971, and subsequently approved by the Second United Nations Conference on the Standardization of Geographical Names in 1972 (Resolution 8). **ADEGN 2007 System** refers to the unified Arabic romanization system that was adopted by the Third Arabic Conference in Beirut in 2007, based on the agreement by Arabic experts at the Eighth United Nations Conference on the Standardization of Geographical Names in 2002 in Berlin. The ADEGN 2007 System is published on the website of the Arabic Division of UNGEGN (*www.adegn.org*).

1. Romanization equivalents in Unicode[2]

Romanization equivalents in the ADEGN 2007 System consist of basic Roman letters and those with diacritical marks. While some of the letters with diacritical marks are readily available in Unicode (so-called precomposed characters), others need to be entered using the code of the basic letter and the code of a combining diacritical mark. It is not always clear which combining diacritical mark should be used. Although some of the romanization equivalents may be written using formatting commands in text editors, such as underline, it is strongly recommended to avoid such usage as these characters can be easily distorted in various text editing stages, and most of all, are not suitable as entries in names databases.

1.1. Macron (above)

There are three vowel letters with macrons, and as these are available in Unicode as precomposed characters, there should be no problems using the letters:

Ā (U+0100), **ā** (U+0101), **Ī** (U+012A), **ī** (U+012B), **Ū** (U+016A), **ū** (U+016B).

1.2. Underline

Underline is used for four consonant letters and one digraph. Most of the underlined letters are available as precomposed characters:

D̲ (U+1E0E), **d̲** (U+1E0F), **h̲** (U+1E96)[3], **T̲** (U+1E6E), **t̲** (U+1E6F).

[1] Submitted by Peeter Päll, Convenor, UNGEGN Working Group on Romanization Systems.
[2] Unicode is a universal standard of character codes covering the world's scripts. It is also known as ISO/IEC 10646.
[3] Note that a precomposed character is given only for the small letter h̲. For capital letter a combining diacritical mark has to be used.

For underlined **H**, **S** and **s** there is a choice of two combining diacritical marks: combining macron below (U+0331) or combining low line (U+0332). The results may look different, as shown in the examples below[4].

Combinations with the combining macron below:

H̲ (U+0048, U+0331), **S̲** (U+0053, U+0331), **s̲** (U+0073, U+0331).

Combinations with the combining low line:

H̲ (U+0048, U+0332), **S̲** (U+0053, U+0332), **s̲** (U+0073, U+0332).

Here are some names and words, containing both precomposed characters and those with combining diacritical marks. First, examples with the combining macron below: Fus̲h̲á, Jiwār Al Ḫawz, Umm Qaṣ̲ir, Ṣ̲ūr, Ṭ̲arāblus. The same words with the combining low line: Fus̲h̲á, Jiwār Al Ḫawz, Umm Qaṣ̲ir, Ṣ̲ūr, Ṭ̲arāblus. Judging by the appearance, the combining macron below will give a more aesthetic look which is also more in line with the other precomposed characters, so a combining macron below is to be recommended.

For the underlined digraph **dh** there is a special code in Unicode: combining double macron below (U+035F), which has to be used between the two characters:

D͟h (U+0044, U+035F, U+0068), **d͟h** (U+0064, U+035F, U+0068).[5]

1. 3. Sign for the *alif maqṣ̲ūrah*.

This letter is not available in Unicode as precomposed character. The ADEGN 2007 System shows a sign added to the basic letter **a** which in Unicode terms can be described variously, either:

a) combining comma above right (U+0315): **a̓**
or
b) combining horn (U+031B): **a̛**

Neither of these signs resemble closely the actual shape of the sign given in the ADEGN document: a combining acute shifted to the upper right edge of the basic letter (looks like **a´** if the acute is moved closer to the basic letter).

The ADEGN letter for *alif maqṣ̲ūrah* looks very similar to a combination of **a** plus the sign for *hamzah* (’) when using the nomenclature of signs available in Unicode, besides has various options, so it is strongly advised to replace the letter with another one which

[4] In this document all letters are entered using Unicode codes and an ordinary Times New Roman font (version 5.01, 2006). This would give an idea of what the actual results could look like.
[5] Note that even if the right code for the double macron below has been entered in both cases, and the combining mark is present in the font, Word does not show it on screen, although on printout it is there. Some browsers, e.g. Firefox, do show it properly.

is available in Unicode. For example, the letter **á** used in the UN 1972 System, is easy to use.

1.4. Other special symbols

The signs for *hamzah* (’) and *‘ayn* (‘) are similar to each other but their use is long established and should not cause too much confusion. It must be noted that in Unicode there are two available codes for these signs: either ’ (U+2019, right single quotation mark), ‘ (U+2018, left single quotation mark) or ’ (U+02BC, modifier letter apostrophe), ‘ (U+02BB, modifier letter turned comma). Although the first pair of codes is much more readily available, it would be more consistent in Unicode to use the second pair of codes.

2. Comparison of UN 1972 and ADEGN 2007

2.1. Characters

In the ADEGN 2007 System the combining cedillas of the UN 1972 System have been replaced by underlines. Thus

ḍ (ض) becomes **d̲**,
ḥ (ح) becomes **h̲**,
ṣ (ص) becomes **s̲**,
ṭ (ط) becomes **t̲**.

The romanization of ẓ (ظ) is changed to the underlined **dh**[6], and that of *alif maqṣūrah* (á, ى) becomes **a´** (for the actual shape of this letter see 1.3).

While it is easy to use underlining in text editors, it should again be stated that such a usage will not be practical for databases and is likely to cause distortion when transferring names from one application to another, so it should be avoided[7]. Letters with underlines would need special care if for example names in maps are underlined for distinction (cf. the names mentioned earlier, when underlined: Fushá, Jiwār Al Hawz, Umm Qasir, Sūr, Tarāblus). In some romanization systems used for Arabic, underlined letters might have different meanings, e.g. d̲ is also used for the romanization of *dh* (ذ), and t̲ is used for *th* (ث).

[6] For comments on this change see Paul Woodman’s paper presented to the 22nd UNGEGN Session in 2004 (working paper 23).

[7] It will be a good idea to have a font that would suit the needs of Arabic romanization, as some of the combinations using an ordinary font may look clumsy. However, any fonts should strictly follow the Unicode encoding scheme described above. The font currently supplied by ADEGN uses for several letters code points that in Unicode are reserved for other letters or symbols, e.g. the code point for **ā** is the same as [, **h̲** equals to $, **t̲** to @, etc. This should be discouraged as names entered using non-Unicode fonts are not interchangeable in worldwide communication.

2.2. Articles

The UN 1972 System does not have rules on the romanization of articles but contains examples of names where the definite article is always written with a small initial and connected by the hyphen to the main part of the name, e.g. البصرة *al-Başrah*, الرياض *ar-Riyāḍ*. In a modification of the system, the BGN/PCGN 1956 System, it is customary not to use hyphens between articles and names and to capitalize the first definite article in a name, e.g. *Al Başrah, Ar Riyāḍ, Minyah aḍ Ḍinnīyah*. In the ADEGN 2007 System, articles are always written with a capital initial and hyphen is not used: منية الضنية *Minyah Aḍ Ḍinniyyah*, جوار الحوز *Jiwār Al Ḥawz*.

2.3. Adjectival ending

The UN 1972 System does not specify the romanization of adjectival endings but the BGN/PCGN 1956 System uses *-īyah*, e.g. الجمهورية اليمنية *Al Jumhūrīyah al Yamanīyah*. The ADEGN 2007 System prefers *-iyyah*, e.g. *Al Jumhūriyyah Al Yamaniyyah*.

2.4. Special rules

The UN 1972 System does not specify any special rules but in the BGN/PCGN 1956 System possible ambiguities in romanization are solved by adding a middle dot (·): *t·h, k·h, d·h* and *s·h* reflect the romanization of two consecutive Arabic consonants while *th, kh, dh* and *sh* refer to single Arabic characters. In ADEGN 2007 this same distinction is made by a slash (/): cf. أدهم *Ad/ham* – أذم *Adham*, أسهم *As/hum* – أشم *Ashum*. The use of a slash for this purpose is unusual, most often slash is used to show alternative names, such as bilingual names in maps.

3. Conclusion

The ADEGN 2007 System is consistent and sufficiently detailed in order to be implemented successfully. Implementation of the system is a necessary precondition for approving the system as the United Nations system. But several technical aspects of the romanization system need to be reconsidered to facilitate its implementation.