# L2/11-069

2011-02-09

| | |
|---|---|
| **Title:** | Letter-making Arabic harakat |
| **Author:** | Roozbeh Pournader (HighTech Passport) |
| **Action:** | For consideration by the UTC |
| **Date:** | 2011-02-09 |

## Introduction

Document L2/10-455, “Of hamza and other harakat”, called for a clarification of the model used for encoding new letters in the Arabic script blocks, in order to reduce the confusion about “letters” that could be represented two ways, either by using an existing base letter and an existing combining mark, or by encoding a new Arabic letter. *(Readers are strongly encouraged to read that document first.)*

In the UTC meeting #125, the author’s analysis of the existing model was generally agreed upon, it was confirmed that there seems to be two different classes of Arabic harakat regarding them being permitted to make letters, and it was suggested that more documentation could be provided about the issue, which the present document tries to do.

## Analysis

It was agreed at the UTC meeting #125 that U+0654 ARABIC HAMZA ABOVE is already such a letter-making character (see consensus 125-C29, accepting Recommendation 3 in L2/10-455, which included “Not restricted to hamza semantics, but also to be used for hamza-like shapes, including for creating letters that contain such a shape. Exceptions are the two characters U+0681 and U+076C.”)

It also follows that the two other characters that were encoded with Hamza Above, U+0653 ARABIC MADDAH ABOVE and U+0655 ARABIC HAMZA BELOW and had already-encoded letters canonically decompose to them are considered letter-making.

Also it is obvious that some other harakat are not intended to be letter-making. For example, U+065A ARABIC VOWEL SIGN SMALL V ABOVE and U+065B ARABIC VOWEL SIGN INVERTED SMALL V ABOVE are even named in a way to suggest that they may not be used to make letters, or it would create a real headache if people suddenly started representing common Kurdish letters like U+0692 ARABIC LETTER REH WITH SMALL V as a sequence of <REH, VOWEL SIGN SMALL V ABOVE>.

A similar case is the Koranic character U+0615 ARABIC SMALL HIGH TAH with various Urdu letters, where the chart comments explicitly state “should not be confused with the small TAH sign used as a diacritic for some letters such as 0679”. That makes the character non-letter-making.

## Breakdown

Based on similar arguments and some personal judgment, here is a breakdown of all Arabic harakat of Unicode 6.0 into the two classes. The non-letter-making classes is made of three categories: the Koranic marks, South Asian and Persian raised text, and the combining marks that were explicitly encoded to be non-letter-making. Everything else would be considered letter-making.

The author can also provide such a breakdown for the Arabic combining characters accepted by UTC for a future version of the standard.

### *Non-letter-making*

# South Asian raised text

0610 ARABIC SIGN SALLALLAHOU ALAYHE WASSALLAM

0611 ARABIC SIGN ALAYHE ASSALLAM

0612 ARABIC SIGN RAHMATULLAH ALAYHE

0613 ARABIC SIGN RADI ALLAHOU ANHU

0614 ARABIC SIGN TAKHALLUS

# Koranic

**0615 ARABIC SMALL HIGH TAH**

# Early Persian raised text

0616 ARABIC SMALL HIGH LIGATURE ALEF WITH LAM WITH YEH

# Koranic

0617 ARABIC SMALL HIGH ZAIN

0618 ARABIC SMALL FATHA

0619 ARABIC SMALL DAMMA

061A ARABIC SMALL KASRA

# Encoded after letters existed with similar diacritics

**065A ARABIC VOWEL SIGN SMALL V ABOVE**

**065B ARABIC VOWEL SIGN INVERTED SMALL V ABOVE**

065C ARABIC VOWEL SIGN DOT BELOW

# Koranic

06D6 ARABIC SMALL HIGH LIGATURE SAD WITH LAM WITH ALEF MAKSURA

06D7 ARABIC SMALL HIGH LIGATURE QAF WITH LAM WITH ALEF MAKSURA

06D8 ARABIC SMALL HIGH MEEM INITIAL FORM

06D9 ARABIC SMALL HIGH LAM ALEF

06DA ARABIC SMALL HIGH JEEM

**06DB ARABIC SMALL HIGH THREE DOTS**

06DC ARABIC SMALL HIGH SEEN

06DF ARABIC SMALL HIGH ROUNDED ZERO

06E0 ARABIC SMALL HIGH UPRIGHT RECTANGULAR ZERO

06E1 ARABIC SMALL HIGH DOTLESS HEAD OF KHAH

06E2 ARABIC SMALL HIGH MEEM ISOLATED FORM

06E3 ARABIC SMALL LOW SEEN

06E4 ARABIC SMALL HIGH MADDA

06E7 ARABIC SMALL HIGH YEH

06E8 ARABIC SMALL HIGH NOON

**06EA ARABIC EMPTY CENTRE LOW STOP**

**06EB ARABIC EMPTY CENTRE HIGH STOP**

**06EC ARABIC ROUNDED HIGH STOP WITH FILLED CENTRE**

06ED ARABIC SMALL LOW MEEM

## *Letter-making*

# Main harakat

064B ARABIC FATHATAN

064C ARABIC DAMMATAN

064D ARABIC KASRATAN

064E ARABIC FATHA

064F ARABIC DAMMA

0650 ARABIC KASRA

0651 ARABIC SHADDA

0652 ARABIC SUKUN

# Originally encoded to unify Arabic and Syriac diacritics, made letter-making when Arabic letters were canonically-decomposed to include them

0653 ARABIC MADDAH ABOVE

0654 ARABIC HAMZA ABOVE

0655 ARABIC HAMZA BELOW

# Less-common harakat

0656 ARABIC SUBSCRIPT ALEF

0657 ARABIC INVERTED DAMMA

0658 ARABIC MARK NOON GHUNNA

0659 ARABIC ZWARAKAY

065D ARABIC REVERSED DAMMA

065E ARABIC FATHA WITH TWO DOTS

065F ARABIC WAVY HAMZA BELOW

0670 ARABIC LETTER SUPERSCRIPT ALEF

## Implications

Drawing the clear line between letter-making and non-letter-making diacritics would help the users of the standard know if a certain “letter” is already encoded as a two-character . This would bring Arabic in par with scripts such as Latin, where it is clear that the Latin Capital Letter L With Tilde is already encoded as <U+004C, U+0303>, and there is no need to propose it to the committees or represent it with a private-use character.

### *Already representable “characters”*

There exist various monolithic Arabic script “letters” that are not encoded as one character, but two. The following are some examples:

- Beh With Heh Doachashmee, Peh With Heh Doachasmee, ..., Gaf With Heh Doachashmee: aspirated forms of Beh, Peh, etc. Easily encodable using Beh+Heh Doachashmee, Peh+Heh Doachashmee, ...
- Farsi Yeh With Hamza Above: Used as one letter in some Azerbaijani orthographies, a variant of Yeh With Hamza Above that has two dots in medial and initial positions, like Farsi Yeh. Easily represented using Farsi Yeh+Hamza Above.
- Alef Maksura With Hamza Below: Used as one letter in Koranic texts, a tooth with a hamza below, that appears in medial and initial positions only. Easily represented using Alef Maksura+Hamza Below.
- Alef Maksura With Damma Above: Used as one letter in some central Asian orthographies. Appears only in medial and initial forms. Easily represented using Alef Maksura+Damma.
- Alef Maksura With Subscript Alef: Used as one letter in some central Asian orthographies. Appears only in medial and initial forms. Easily represented using Alef Maksura+Subscript Alef.
- Waw With Madda Above: Used as one letter in some central Asian orthographies. Easily represented using Waw+Madda Above.
- Beh With Hamza Above: Used as one letter in Fulfulde (see L2/10-442). Easily represented using Beh+Hamza Above.

If proposals for such characters are presented to the committee, they should be rejected by mentioning that these can already representable in Unicode.

It could be argued that the recently approved character 08A8 Yeh With Two Dots Below And Hamza Above has been representable as the sequence <Yeh, CGJ, Hamza Above>. But that was considered too complicated for one letter, and the UTC agreed to keep the encoded character.

### *Letters with missed decompositions*

There are some inconsistencies in the standard regarding already existing letters that could be decomposed when decompositions were being frozen, but were not. More explanation is provided in document L2/10-455, but here is a short list of existing letters that live on the borders:

- U+0673 ARABIC LETTER ALEF WITH WAVY HAMZA BELOW, used in Kashmiri, deprecated because a combining Wavy Hamza Below was later encoded at U+065F;
- U+0681 ARABIC LETTER HAH WITH HAMZA ABOVE, used in Pashto;
- U+06C7 ARABIC LETTER U, basically a Waw With Damma, used in Kirghiz and (Iranian) Azerbaijani;
- U+06C8 ARABIC LETTER YU, basically a Waw With Superscript Alef, used in Uighur;
- U+076C ARABIC LETTER REH WITH HAMZA ABOVE, mistakenly encoded following the pattern

of HAH WITH HAMZA ABOVE.

- 08A8 ARABIC LETTER YEH WITH TWO DOTS BELOW AND HAMZA ABOVE, used in Fulfulde, encoded because the combination of YEH+HAMZA ABOVE would lose the two dots below it in Unicode-complaint rendering engines.

The UTC already agreed that these should be documented properly for security reasons and other confusions they will cause. It may be a good idea to mention that we plan to keep the list closed to these six characters.

## Recommendations

1. Document that some Arabic combining characters MAY NOT be used for creating new letters, while others could and SHOULD be used for representing letters that contain them as diacritics, even if they are just used as modifiers to create new letters and do not have separate phonemical value.
2. Document the six exceptions to pre-composed letters with no decomposition, and the intention to keep the list limited.
3. Update the answers to the last two questions in the Unicode FAQ on "Middle Eastern Scripts and Languages" to make clear the difference between encoded letter-making diacritics (like *Hamza Above*), encoded non-letter-making diacritics (like *Vowel Sign Small V Above*) and non-encoded ones (like *Three Dots Below*).