Title: Proposal to encode four combining Arabic characters for Koranic use
Source: The Unicode Consortium
Status: Liaison Contribution
Author: Roozbeh Pournader
Action: For consideration by ISO/IEC JTC1/SC2/WG2
Date: 2010-03-30

Introduction

Although we are almost there, Unicode and UCS still miss a few characters for properly encoding the modern day representation of Koranic text. This document tries to fill the gap by proposing four important missing characters.

Also, considering that the two existing Arabic blocks (“Arabic” and “Arabic Supplement”) are almost full, this documents calls for opening a new “Arabic Extended-A” block at the location that is roadmapped for it.

Requests

• The creation of a new Arabic block, named “Arabic Extended-A”, at U+08A0..08FF. (This is already roadmapped.)

• The encoding of four Koranic characters (three open tanweens and one combining version of small waw) in the “Arabic Extended-A” block at positions U+08F0..08F3.

Background

New Arabic block

As of November 2009, the “Arabic Supplement” block at U+0750..077F is full, and there are 4 empty spaces in the “Arabic” block, U+0600..06F0 (the spaces at U+0620 and U+065F are already filled by characters accepted for Kashmiri).

At the same time, there are existing UCS and Unicode proposals for at least 8 characters: four in this document, three for Arwi in L2/09-143, and one for Samvat Date Sign in L2/09-144.

The author expects the four holes in the U+06xx block to be filled by characters similar to their neighboring characters, which are symbols and punctuation. For example, the Samvat Date Sign is a good candidate for getting encoded in the existing four holes of the “Arabic” block.

Thus, the author is requesting the Koranic characters to be encoded in the new block. This would also make it possible to keep the three open tanweens, that form a set, next to each other.
The missing characters

For encoding the most common modern form of published Koranic text, four characters are missing. Three are open forms of tanween, which mark a pronunciation difference with normal tanweens, already encoded at U+064B..064D. The other is a combining companion of U+06E5 ARABIC SMALL WAW, needed for encoding some words (compare to the pair U+06E6 ARABIC SMALL YEH and U+06E7 ARABIC SMALL HIGH YEH).

The three open tanweens have been proposed to the UTC at least twice, by Thomas Milo in L2/01-325 and by Jonathan Kew in L2/02-275. There is no indication that the first document was ever on the UTC table, and although the second document appears on the agenda for UTC #92, there is no mention of it in the minutes. These may not have ever been discussed in the UTC.

There is also a third document which appears like a Unicode proposal but was never submitted for consideration of the committee. It’s Arabeyes’s “Proposal to add four Arabic characters to the BMP of the UCS and fix four characters” written by Mohammad Yousif and Nadim Shaikli, publicly available at http://arabeyes.org/~nadim/tmp/unicode_quran_prop.pdf

The open tanweens point toward a different pronunciation of tanween in Koranic text. While a normal tanween would be pronounced as [an], [in], or [un] with a clear [n] sound (called izhār), an open tanween specifies either nasalization of the [n] sound (called ikhfāʿ), or its total disappearance, geminating the next consonant (called idqām).

Theoretically, the shape of the tanweens could be determined by finding the next “pronounced” consonant, and advanced reciters even learn the rules to determine it themselves. But determination of the next pronounced consonant would need skipping over various characters to be able to determine the next pronounced consonant. The skipped-over characters may include spaces, unpronounced alefs (possibly with some combining marks), symbols like START OF RUB EL HIZB and PLACE OF SAJDA, END OF AYAH symbols with their pack of following digits, page breaks, and sura breaks. The author believes that while this is achievable in software, it is overkill for text rendering engines who want to reflect the actual textual content of Koranic text.

It should also be noted that similarly automatically determinable symbols are already encoded in the standard. For example, U+06E2 ARABIC SMALL HIGH MEEM ISOLATED FORM and U+06ED ARABIC SMALL LOW MEEM are also automatically determinable by finding the next “pronounced” consonant. They are actually parts of tanweens too: together with a fatha, kasra, or damma, they are specifying that the [n] sound in a tanween is changed to a [m] sound.

Figure 1. Symbol legend in Arabic language, appearing at the end of a Koran published in Iran. A normal tanween is shown first, saying it marks izhār (إظهار). Then an open tanween is shown, saying it marks idqām (إدغام) and ikhfāʿ (إخفاء). From [1], page 606.
Figure 2. Examples of the three open tanween characters as opposed to normal tanweens, from L2/02-275 (green = fathatan, blue = dammatan, purple = kasratan, ignore red). Note that intelligent Koranic software, in order to determine the open shape of the Kasratan on the last line (if not provided), would need to skip over at least three characters that compose the marker: an End of Ayah character and two digits. The text is from Sad (38):62–72.

Also, encoding three open tanweens will have the additional benefit of making verbatim copies of text possible when copy and pasting, and would make the discussion of Koranic text possible without being worried that a tanween may suddenly change shape because it’s followed by some text in another language in running text.

The status of the open tanweens and this request for their encoding is based on similar already-encoded characters that are variants of other Arabic characters but were encoded to make plain text encoding of the Koran possible. Examples include small version of the basic harakat (U+0618..061A vs U+064E..0650) and the alternative versions of sukun (U+06DF..06D1 vs U+0652).

The other missing character, ARABIC SMALL HIGH WAW, is exemplified in the Arabeyes document, although not as a new character. The character appears mid-word in Koranic text, as shown in Figure 3 (left-hand).

U+06E5 ARABIC SMALL WAW cannot be used for this purpose, as it is a non-joining spacing character, and would result in the first part of the word getting disconnected from the second part.

Although similar, U+064C ARABIC DAMMA cannot be used for this purpose either, as there are clear semantic
and visual differences. Apart from having slightly different shapes, *damma* and small high *waw* are used for writing two different sounds ([u] vs [uː]). A Koran reciter is aware of the visual difference and will pronounce them differently because of the visual hint.

![Image](image.png)

**Figure 3.** On the right side, there are two examples of spacing U+06E5 ARABIC SMALL WAW at the end of each word. On the left side, there is a combining version (applied to either a seen, or a *tatweel*, and followed by a *small madda*), not encoded yet. Note the visual difference between ARABIC SMALL HIGH WAW (typically indicating a long [uː] sound) and a normal U+064C ARABIC DAMMA (typically indicating a short [u]). Both samples from [1], page 282.

Suggested character sequence for the examples are below:

Right side: `<062D, 064E, 0648, 06E1, 0644, 064E, 0647, 064F, 06E5, 0020, 0644, 0650, 0646, 064F, 0631, 0650, 06CC, 064E, 0647, 064F, 06E5>`, from *Al-Isra* (17):1;

Left side: `<0644, 0650, 06CC, 064E, 0633, 064F, 0640, 08F3, 06E4, 0640, 0654, 064F, 0648, 0627, 06DF>`, from *Al-Isra* (17):7.

Note that the proposed ARABIC SMALL HIGH WAW is functionally similar to the already-encoded U+06E7 ARABIC SMALL HIGH YEH. It would complete a pair with U+06E5 ARABIC SMALL WAW, in exactly the same way that U+06E6 ARABIC SMALL YEH is completed by U+06E7 ARABIC SMALL HIGH YEH.

**Character names, shapes, and properties**

Proposed codepoints, names and glyphs for the characters follow:

<table>
<thead>
<tr>
<th>Character</th>
<th>Codepoint</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>ⵜ ⫤ ⫦ ⫯ ⫹</td>
<td>U+08F0</td>
<td>ARABIC OPEN FATHATAN</td>
</tr>
<tr>
<td>⫯ ⫹ ⫵ ⫺ ⫺</td>
<td>U+08F1</td>
<td>ARABIC OPEN DAMMATAN</td>
</tr>
<tr>
<td>⫹ ⫺ ⫺ ⫼</td>
<td>U+08F2</td>
<td>ARABIC OPEN KASRATAN</td>
</tr>
</tbody>
</table>

Proposal to encode four Arabic characters for Koranic use
Data for UnicodeData.txt follows:

08F0;ARABIC OPEN FATHATAN;Mn;27;NSM;;;;N;;;;;
08F1;ARABIC OPEN DAMMATAN;Mn;28;NSM;;;;N;;;;;
08F2;ARABIC OPEN KASRATAN;Mn;29;NSM;;;;N;;;;;
08F3;ARABIC SMALL HIGH WAW;Mn;230;NSM;;;;N;;;;;

Other properties could be copied from existing similar characters. The three open tanweens will be similar to their counterparts (U+064B..064D), and the small high waw will be similar to U+06E7 ARABIC SMALL HIGH YEH.

For collation purposes, the open tanweens should perhaps be considered variants of the normal tanweens, and the small high waw could be treated similarly to other combining Koranic characters, like U+06E7 ARABIC SMALL HIGH YEH.

It is not expected that the proposed characters would be useful in identifiers.

**Confusablity**

In terms of UTS #39 Unicode Security Mechanism, the proposed characters could participate in forming the following confusable pairs:

08F0 ; 064B # Open Fathatan vs Fathatan
08F1 ; 064C # Open Dammatan vs Dammatan
08F2 ; 064D # Open Kasratan vs Kasratan
08F0 ; 064E 064E # Open Fathatan vs two Fatha’s
08F1 ; 064F 064F # Open Dammatan vs two Damma’s
08F2 ; 0650 0650 # Open Kasratan vs two Kasra’s
08F3 ; 0619 # Small High Waw vs Small Damma
08F3 ; 064F # Small High Waw vs Damma
06E5 ; 00A0 08F3 # Small Waw vs NBSP+Small High Waw

**Acknowledgments**

The author wishes to thank Adil Allawi, Peter Constable, Mark Davis, Behdad Esfahbod, Thomas Milo, Eric Muller, Ken Whistler, and the Arabeyes free software community for fruitful discussions. Some samples are taken from papers and documents by Thomas Milo and Jonathan Kew.

**References**

### A. Administrative

1. **Title:** Proposal to encode four combining Arabic characters for Koranic use

2. **Requester's name:** Roozbeh Pournader

3. **Requester type (Member body/Liaison/Individual contribution):** Individual contribution

4. **Submission date:** 2009-11-04

5. **Requester's reference (if applicable):**

6. **Choose one of the following:**
   - This is a complete proposal: ✓
   - (or) More information will be provided later:

### B. Technical – General

1. **Choose one of the following:**
   - a. This proposal is for a new script (set of characters):
     - Proposed name of script: Arabic Extended-A (new block, the script exists already)
   - b. The proposal is for addition of character(s) to an existing block:
     - Name of the existing block:

2. **Number of characters in proposal:** 4

3. **Proposed category (select one from below - see section 2.2 of P&P document):**
   - A-Contemporary
   - B.1-Specialized (small collection)
   - B.2-Specialized (large collection)
   - C-Major extinct
   - D-Attested extinct
   - E-Minor extinct
   - F-Archaic Hieroglyphic or Ideographic
   - G-Obscure or questionable usage symbols

4. **Is a repertoire including character names provided?** Yes

5. **Who will provide the appropriate computerized font (ordered preference: True Type, or PostScript format) for publishing the standard?** Thomas Milo and Michael Everson

    If available now, identify source(s) for the font (include address, e-mail, ftp-site, etc.) and indicate the tools used:

    A font has already been provided to the editors by Michael Everson.

6. **References:**
   - a. Are references (to other character sets, dictionaries, descriptive texts etc.) provided? Yes
   - b. Are published examples of use (such as samples from newspapers, magazines, or other sources) of proposed characters attached? Yes

7. **Special encoding issues:**
   - Does the proposal address other aspects of character data processing (if applicable) such as input, presentation, sorting, searching, indexing, transliteration etc. (if yes please enclose information)? Yes

8. **Additional Information:**
   - Submitters are invited to provide any additional information about Properties of the proposed Character(s) or Script that will assist in correct understanding of and correct linguistic processing of the proposed character(s) or script. Examples of such properties are: Casing information, Numeric information, Currency information, Display behaviour information such as line breaks, widths etc., Combining behaviour, Spacing behaviour, Directional behaviour, Default Collation behaviour, relevance in Mark Up contexts, Compatibility equivalence and other Unicode normalization related information. See the Unicode standard at [http://www.unicode.org](http://www.unicode.org) for such information on other scripts. Also see [http://www.unicode.org/Public/UNIDATA/UCD.html](http://www.unicode.org/Public/UNIDATA/UCD.html) and associated Unicode Technical Reports for information needed for consideration by the Unicode Technical Committee for inclusion in the Unicode Standard.

---

### C. Technical - Justification

<table>
<thead>
<tr>
<th>Question</th>
<th>Yes/No Reason</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Has this proposal for addition of character(s) been submitted before?</td>
<td>Yes. The open tanweens have been proposed before in two documents, L2/01-325 and L2/02-275.</td>
</tr>
<tr>
<td>If YES explain</td>
<td></td>
</tr>
<tr>
<td>2. Has contact been made to members of the user community (for example: National Body, user groups of the script or characters, other experts, etc.)?</td>
<td>Yes. The experts contacted include Thomas Milo, Adil Allawi, and Behdad Esfahbod.</td>
</tr>
<tr>
<td>If YES, with whom?</td>
<td></td>
</tr>
<tr>
<td>If YES, available relevant documents:</td>
<td></td>
</tr>
<tr>
<td>3. Information on the user community for the proposed characters (for example: size, demographics, information technology use, or publishing use) is included?</td>
<td>Yes. Muslims who want to recite the Koran correctly. There are ≈ 1.5×10⁹ Muslims living.</td>
</tr>
<tr>
<td>Reference:</td>
<td></td>
</tr>
<tr>
<td>4. The context of use for the proposed characters (type of use; common or rare)</td>
<td>Koranic. The open tanweens are very common in Koranic text. The small high waw is rare.</td>
</tr>
<tr>
<td>Reference:</td>
<td></td>
</tr>
<tr>
<td>5. Are the proposed characters in current use by the user community?</td>
<td>Yes. Most of recently published Korans use all the four characters.</td>
</tr>
<tr>
<td>If YES, where? Reference:</td>
<td></td>
</tr>
<tr>
<td>6. After giving due considerations to the principles in the P&amp;P document must the proposed characters be entirely in the BMP?</td>
<td>Yes. Characters could be kept with their counterparts in Arabic block, and space is roadmapped already.</td>
</tr>
<tr>
<td>If YES, is a rationale provided?</td>
<td>Yes, see below</td>
</tr>
<tr>
<td>If YES, reference:</td>
<td></td>
</tr>
<tr>
<td>7. Should the proposed characters be kept together in a contiguous range (rather than being scattered)?</td>
<td>Yes</td>
</tr>
<tr>
<td>8. Can any of the proposed characters be considered a presentation form of an existing character or character sequence?</td>
<td>Yes. See the section titled “missing characters”</td>
</tr>
<tr>
<td>If YES, is a rationale for its inclusion provided?</td>
<td>Yes</td>
</tr>
<tr>
<td>If YES, reference:</td>
<td></td>
</tr>
<tr>
<td>9. Can any of the proposed characters be encoded using a composed character sequence of either existing characters or other proposed characters?</td>
<td>No</td>
</tr>
<tr>
<td>If YES, is a rationale for its inclusion provided?</td>
<td></td>
</tr>
<tr>
<td>If YES, reference:</td>
<td></td>
</tr>
<tr>
<td>10. Can any of the proposed character(s) be considered to be similar (in appearance or function) to an existing character?</td>
<td>Yes</td>
</tr>
<tr>
<td>If YES, is a rationale for its inclusion provided?</td>
<td>Yes</td>
</tr>
<tr>
<td>If YES, reference:</td>
<td>See the section titled “missing characters” and “Confusability”</td>
</tr>
<tr>
<td>11. Does the proposal include use of combining characters and/or use of composite sequences?</td>
<td>Yes, combining</td>
</tr>
<tr>
<td>If YES, is a rationale for such use provided?</td>
<td>Yes</td>
</tr>
<tr>
<td>If YES, reference:</td>
<td>All characters are combining, as can be seen in samples.</td>
</tr>
<tr>
<td>Is a list of composite sequences and their corresponding glyph images (graphic symbols) provided?</td>
<td>No</td>
</tr>
<tr>
<td>If YES, reference:</td>
<td></td>
</tr>
<tr>
<td>12. Does the proposal contain characters with any special properties such as control function or similar semantics?</td>
<td>No</td>
</tr>
<tr>
<td>If YES, describe in detail (include attachment if necessary)</td>
<td></td>
</tr>
<tr>
<td>13. Does the proposal contain any Ideographic compatibility character(s)?</td>
<td>No</td>
</tr>
<tr>
<td>If YES, is the equivalent corresponding unified ideographic character(s) identified?</td>
<td></td>
</tr>
<tr>
<td>If YES, reference:</td>
<td></td>
</tr>
</tbody>
</table>