

2009-11-05

Title: Proposal to encode four combining Arabic characters for Koranic use
Action: For consideration by UTC and ISO/IEC JTC1/SC2/WG2
Author: Roozbeh Pournader
Date: 2009-11-05

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 document calls for opening a new “Arabic Extended-A” block at the location that is roadmapped for it.

Requests

- The author requests the creation of a new Arabic block, named “Arabic Extended-A”, at U+08A0..08FF. (This is already roadmapped.)
- The author requests 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..77F 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 *ikhfā'*), 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.

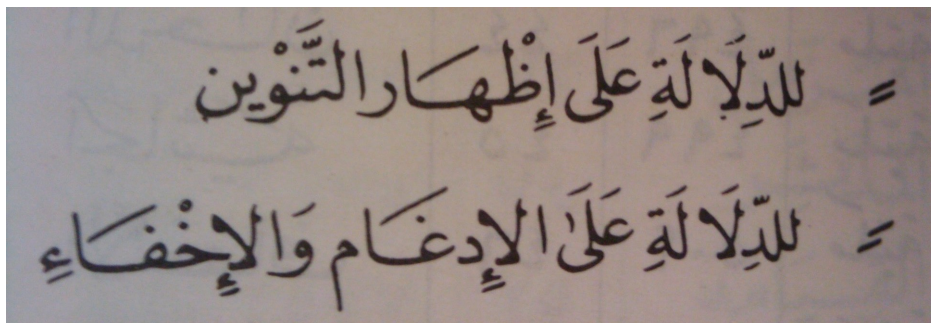


Figure 1. Symbol legend in Arabic language, appearing at the end of a Koran published in Iran. A normal *tanween* is shown first, saying it marks *izhār* (إظهار). Then an open *tanween* is shown, saying it marks *idqām* (إدغام) and *ikhfā'* (إخفاء).

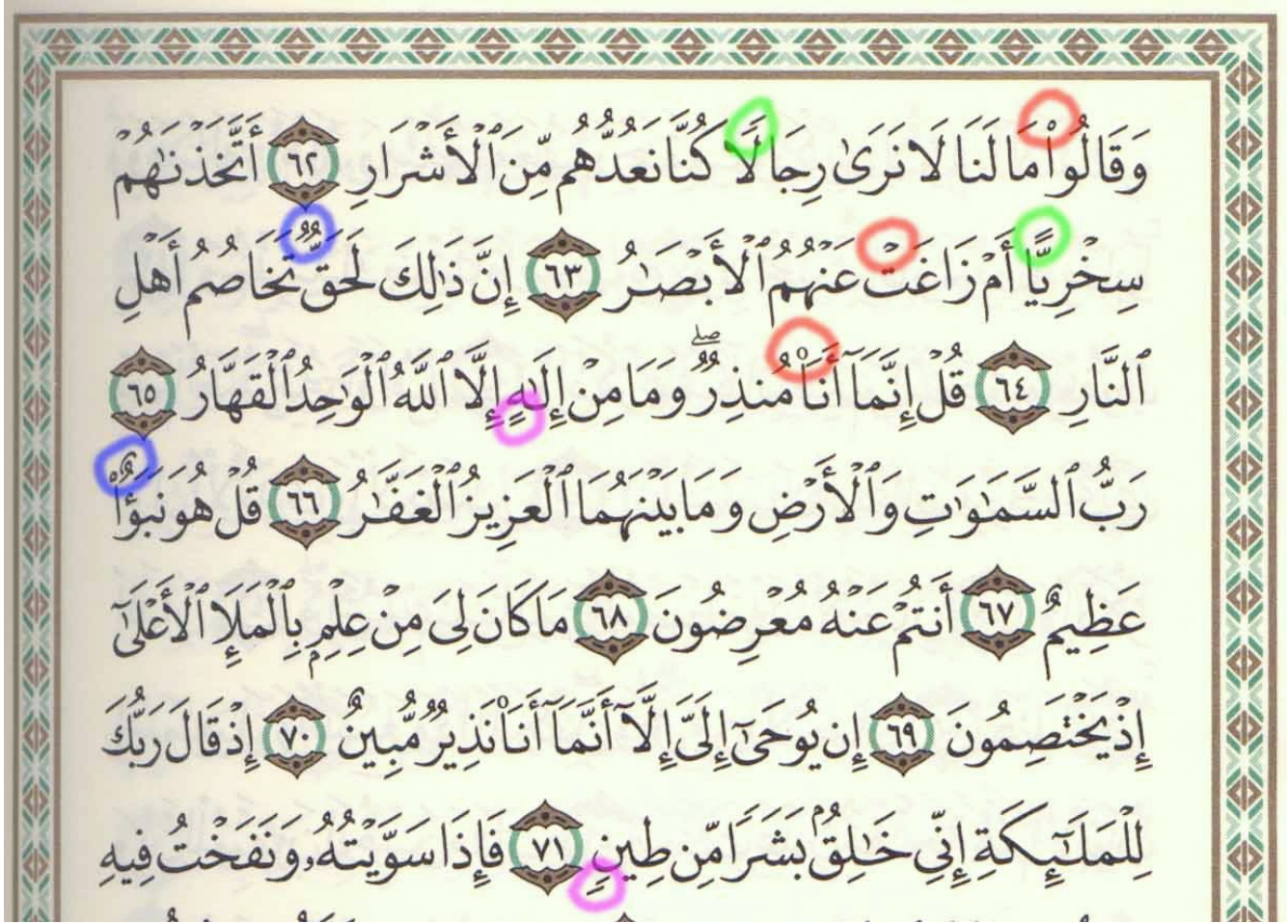


Figure 2. Examples of the three open *tanween* characters as opposed to normal *tanweens*, from L2/02-275 (green = *fathatan*, blue = *dammatan*, purple = *kasratan*, ignore red). Note that intelligent Koranic software, in order to determine the open shape of the Kasratan on the last line (if not provided), would need to skip over at least three characters that compose the marker: an End of Ayah character and two digits.

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 [u:]). A Koran reciter is aware of the visual difference and will pronounce them differently because of the visual hint.

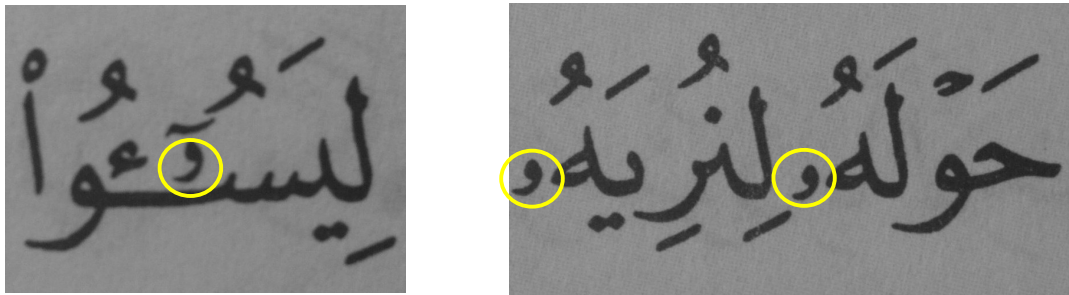






Figure 3. On the right side, there are two examples of spacing U+06E5 ARABIC SMALL WAW at the end of each word. On the left side, there is a combining version (applied to either a *seen*, or a *tatweel*, and followed by a *small madda*), not encoded yet. Note the visual difference between ARABIC SMALL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

Character names, shapes, and properties

Proposed codepoints, names and glyphs for the characters follow:

| | | |
|---|--------|-----------------------|
|  | U+08F0 | ARABIC OPEN FATHATAN |
|  | U+08F1 | ARABIC OPEN DAMMATAN |
|  | U+08F2 | ARABIC OPEN KASRATAN |
|  | U+08F3 | ARABIC SMALL HIGH WAW |

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.

Confusability

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.

**ISO/IEC JTC 1/SC 2/WG 2
PROPOSAL SUMMARY FORM TO ACCOMPANY SUBMISSIONS
FOR ADDITIONS TO THE REPERTOIRE OF ISO/IEC 10646¹.**

A. Administrative

| | |
|--|--|
| 1. Title: | Proposal to encode four combining Arabic characters for Koranic use |
| 2. Requester's name: | <i>Roozbeh Pournader</i> |
| 3. Requester type (Member body/Liaison/Individual contribution): | <i>Individual contribution</i> |
| 4. Submission date: | <i>2009-11-04</i> |
| 5. Requester's reference (if applicable): | |
| 6. Choose one of the following: | |
| This is a complete proposal: | <input checked="" type="checkbox"/> |
| (or) More information will be provided later: | <input type="checkbox"/> |

B. Technical – General

| | | |
|---|--|---|
| 1. Choose one of the following: | | |
| a. This proposal is for a new script (set of characters): | <input checked="" type="checkbox"/> | |
| Proposed name of script: | <i>Arabic Extended-A (new block, the script exists already)</i> | |
| b. The proposal is for addition of character(s) to an existing block: | <input type="checkbox"/> | |
| Name of the existing block: | | |
| 2. Number of characters in proposal: | <i>4</i> | |
| 3. Proposed category (select one from below - see section 2.2 of P&P document): | | |
| A-Contemporary <input type="checkbox"/> | B.1-Specialized (small collection) <input checked="" type="checkbox"/> | B.2-Specialized (large collection) <input type="checkbox"/> |
| C-Major extinct <input type="checkbox"/> | D-Attested extinct <input type="checkbox"/> | E-Minor extinct <input type="checkbox"/> |
| F-Archaic Hieroglyphic or Ideographic <input type="checkbox"/> | G-Obscure or questionable usage symbols <input type="checkbox"/> | |
| 4. Is a repertoire including character names provided? | <i>Yes</i> | |
| a. If YES, are the names in accordance with the “character naming guidelines” in Annex L of P&P document? | <i>Fa</i> | |
| b. Are the character shapes attached in a legible form suitable for review? | <i>Yes</i> | |
| 5. Who will provide the appropriate computerized font (ordered preference: True Type, or PostScript format) for publishing the standard? | <i>Roozbeh Pournader</i> | |
| If available now, identify source(s) for the font (include address, e-mail, ftp-site, etc.) and indicate the tools used: | <i>The font was created using FontForge. It will be provided to the editors.</i> | |
| 6. References: | | |
| a. Are references (to other character sets, dictionaries, descriptive texts etc.) provided? | <i>Yes</i> | |
| b. Are published examples of use (such as samples from newspapers, magazines, or other sources) of proposed characters attached? | <i>Yes</i> | |
| 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)? | <i>Yes</i> | |
| | <i>See the section titled Character names, shapes, and properties</i> | |

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> for such information on other scripts. Also see <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.

¹ Form number: N3152-F (Original 1994-10-14; Revised 1995-01, 1995-04, 1996-04, 1996-08, 1999-03, 2001-05, 2001-09, 2003-11, 2005-01, 2005-09, 2005-10, 2007-03, 2008-05)

C. Technical - Justification

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