UTC Document Register **L2/14-105R**
ISO/IEC JTC1/SC2/WG2 **N4589**

Proposal to encode fourteen Pakistani Quranic marks

Roozbeh Pournader, Google Inc. (for the UTC)
July 27, 2014

Background

In Shaikh 2014a (L2/14-095) and 2014b (L2/14-096), fifteen characters are proposed in order to represent the characters used in Qurans published in Pakistan. The Unicode Technical Committee accepted the characters with some modifications, as listed below. The characters are requested to be encoded in ISO/IEC 10646.

Glyph	Codepoint	Name and notes
◌ࣕ	08D5	ARABIC SMALL HIGH SAD
◌ࣖ	08D6	ARABIC SMALL HIGH AIN
◌ࣗ	08D7	ARABIC SMALL HIGH QAF
◌ࣘ	08D8	ARABIC SMALL HIGH NOON WITH KASRA
◌ࣙ	08D9	ARABIC SMALL LOW NOON WITH KASRA

الثلثة ◌	08DA	ARABIC SMALL HIGH WORD ATH-THALATHA
السجدة ◌	08DB	ARABIC SMALL HIGH WORD AS-SAJDA
النصف ◌	08DC	ARABIC SMALL HIGH WORD AN-NISF
سكتة ◌	08DD	ARABIC SMALL HIGH WORD SAKTA
قف ◌	08DE	ARABIC SMALL HIGH WORD QIF
وقفة ◌	08DF	ARABIC SMALL HIGH WORD WAQFA
ـه ◌	08E0	ARABIC SMALL HIGH FOOTNOTE MARKER
صـ ◌	08E1	ARABIC SMALL HIGH SIGN SAFHA
٥	08E2	DISPUTED END OF AYAH

The main character properties will be as follows:

```
08D5;ARABIC SMALL HIGH SAD;Mn;230;NSM;;;;;N;;;;;
08D6;ARABIC SMALL HIGH AIN;Mn;230;NSM;;;;;N;;;;;
08D7;ARABIC SMALL HIGH QAF;Mn;230;NSM;;;;;N;;;;;
08D8;ARABIC SMALL HIGH NOON WITH KASRA;Mn;230;NSM;;;;;N;;;;;
08D9;ARABIC SMALL LOW NOON WITH KASRA;Mn;220;NSM;;;;;N;;;;;
```

```
08DA;ARABIC SMALL HIGH WORD ATH-THALATHA;Mn;230;NSM;;;;;N;;;;;
08DB;ARABIC SMALL HIGH WORD AS-SAJDA;Mn;230;NSM;;;;;N;;;;;
08DC;ARABIC SMALL HIGH WORD AN-NISF;Mn;230;NSM;;;;;N;;;;;
08DD;ARABIC SMALL HIGH WORD SAKTA;Mn;230;NSM;;;;;N;;;;;
08DE;ARABIC SMALL HIGH WORD QIF;Mn;230;NSM;;;;;N;;;;;
08DF;ARABIC SMALL HIGH WORD WAQFA;Mn;230;NSM;;;;;N;;;;;
08E0;ARABIC SMALL HIGH FOOTNOTE MARKER;Mn;230;NSM;;;;;N;;;;;
08E1;ARABIC SMALL HIGH SIGN SAFHA;Mn;230;NSM;;;;;N;;;;;
08E2;DISPUTED END OF AYAH;Cf;0;AN;;;;;N;;;;;
```

No new entry will be made in ArabicShaping.txt, except for the following:

```
08E2; DISPUTED END OF AYAH; U; No_Joining_Group
```

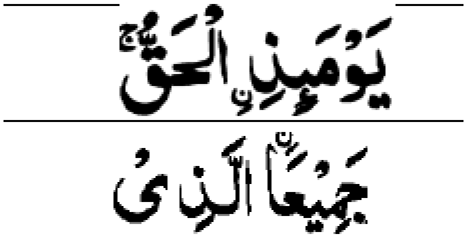
The script property for U+08D5..08DE1 will be Arabic. The script property of U+08E2 DISPUTED END OF AYAH will be Common, similar to U+06DD ARABIC END OF AYAH.

Confusability

1. The character U+08D8 ARABIC SMALL HIGH NOON WITH KASRA could be confused with the sequences <06E8, 0618>, <06E8, 064E>, <0618, 06E8>, and <064E, 06E8>.
2. The character U+08E2 DISPUTED END OF AYAH could be confused with U+06F5 EXTENDED ARABIC-INDIC DIGIT FIVE.

Differences with Shaikh proposals

1. Instead of the originally proposed character LOW NOON, two atomic characters were encoded, which contain a *kasra*. This is because base letters that preceded the newly proposed characters typically already contain another tashkil, and commonly even a main *kasra*. See the following figure (from Usmah 2013) and Figure 1 in Shaikh 2014a:



   If the sequences seen above were encoded as <**letter**, kasra, **small noon**, kasra>, as assumed envisioned by Shaikh 2014a, they would have been normalized to <**letter**, kasra, kasra, **small noon**> which would lose the order of the characters and would be considered to be in error by various rendering engines because of doubling of *kasra*.

The atomic encoding of the noon with *kasra* is also consistent with its atomic semantics of hinting towards a tanween before a wasla.

2. The END OF RUKU was not encoded, as no evidence of plain text usage was provided. Marks at the margin of the Qurans are known to contain complicated structures, which could be represented in higher level protocols.
3. The ALTERNATE DAMMATAN was not encoded. The shape is widely known as an alternative shape of the character DAMMATAN, and no evidence of contrastive usage was provided to support disunification.
4. The glyphs for the WAQFA, THALATHA, and SAKTA characters were changed to use a TEH MARBUTA GOAL, as seen in the samples provided in Shaikh 2013a. The glyph for DISPUTED END OF AYAH was changed to reflect its behavior of merging with following numbers.
5. The character names were changed to better match the existing patterns in Unicode character names. The following patterns are used:
    a. The one-letter marks follow the pattern of similar one-letter Quranic marks.
    b. The multi-letter marks follow the pattern of similar multi-letter Arabic characters, with “LIGATURE” replaced by “WORD”, as these marks are not ligatures. The semanic names proposed in Shaikh 2014a were dropped for a more graphical naming, in case the same characters could be used for other purposes. The words are transcribed with a key similar to that used in U+FDFD ARABIC LIGATURE BISMILLAH AR-RAHMAN AR-RAHEEM, as opposed to U+FDF0..FDFB, since the former pattern may be more readable for Quran reciters. (If the pattern used in U+FDF0..FDFB would be preferable, the hyphens would need to be removed and NISF and QIF would need to be replaced by NESF and QEF.)
    c. The naming of the disputed end of *ayah* character was kept semantical, as different shapes may exist for the character, similar to the normal end of *ayah* character. The character name does not include the script name “ARABIC”, as it may also be in used in translations of the Quran written in other scripts.
6. The codepoints are provided: First the one-letter marks, then the words, then the signs, and finally the ayah marker. In each class, characters have been sorted in alphabetic order. For the signs, the order of the main Unicode characters has been followed.

Bibliography

1. Lateef Sagar Shaikh. 2014a. “Proposal to encode Quranic marks used in Quran published in Pakistan.” UTC Document Register L2/14-095, The Unicode Consortium. http://www.unicode.org/L2/L2014/14095-quranic-marks.pdf
2. Lateef Sagar Shaikh. 2014b. “Proposal to encode Quranic Alternate Dammatan used in Quran published in Pakistan.” UTC Document Register L2/14-096, The

Unicode Consortium. http://www.unicode.org/L2/L2014/14096-dammatan-alt.pdf

3. Sister Usmah. 2013. “The Noon Qutni Places in the Uthmani Script of the Quraan.” Accessed April 29, 2014. http://www.scribd.com/doc/127046919/

ISO/IEC JTC 1/SC 2/WG 2
PROPOSAL SUMMARY FORM TO ACCOMPANY SUBMISSIONS
FOR ADDITIONS TO THE REPERTOIRE OF ISO/IEC 10646

A. Administrative

1. Title**:** ***Proposal to encode fourteen Pakistani Quranic marks***
2. Requester's name: **UTC**
3. Requester Type (Member body/Liaison/Individual contribution): **Liaison**
4. Submission date: **July 27, 2014**
5. Requester's reference, if applicable: **N/A**
6. Choose one of the following:
    This is a complete proposal: **Yes**
    (or) More information will be provided later: **No**

B. Technical – General

1. Choose one of the following:
    a. This proposal is for a new script (set of characters): **No**
        Proposed name of script: **N/A**
    b. The proposal is for addition of character(s) to an existing block: **Yes**
        Name of existing block: **Arabic Extended-A**
2. Number of characters in proposal: **Fourteen (14)**
3. Proposed category: **A-Contemporary**
4. Is a repertoire including character names provided? **Yes**
    a. If YES, are the names in accordance with the "character naming guidelines" in Annex L of P&P document? **Yes**
    b. Are the character shapes attached in a legible form suitable for review? **Yes**
5. Fonts related:
    a. Who will provide the appropriate computerized font to the Project Editor of 10646 for publishing the standard? **Michael Everson**
    b. Identify the party granting a license for use of the font by the editors (include address, e-mail, ftp-site, etc.): **N/A**
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. In L2/14-095.**
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 endorse information)? **No**

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 Unicode Character Database (http://www.unicode.org/reports/tr44/) and associated Unicode Technical Reports for information needed for consideration by the Unicode Technical Committee for inclusion in the Unicode Standard.

C. Technical - Justification

1. Has this proposal for addition of character(s) been submitted before? **Yes**
	If YES explain: **It was originally proposed in L2/14-095. This proposal provides suggested codepoints and properties.**

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.)? **Unknown**
	If YES, with whom? **See original proposal**
	If YES, available relevant documents: **L2/14-095**

3. Information on the user community for the proposed characters (for example: size, demographics, information technology use, or publishing use) is included? **Yes**
	Reference: **L2/14-095**

4. The context of use for the proposed characters (type of use; common or rare): **Common in Pakistani Qurans**
	Reference: **L2/14-095**

5. Are the proposed characters in current use by the user community? **Yes**
	If YES, where? Reference: **See original proposal.**

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. Need to be next to similar characters.**
	If YES, reference: **N/A**

7. Should the proposed characters be kept together in a contiguous range (rather than being scattered)? **No**

8. Can any of the proposed characters be considered a presentation form of an existing character or character sequence? **No**
	If YES, is a rationale for its inclusion provided? **N/A**
	If YES, reference: **N/A**

9. Can any of the proposed characters be encoded using a composed character sequence of either existing characters or other proposed characters? **Yes**
	If YES, is a rationale for its inclusion provided? **Yes**
	If YES, reference: **See present document**

10. Can any of the proposed character(s) be considered to be similar (in appearance or function) to, or could be confused with, an existing character? **Yes**

If YES, is a rationale for its inclusion provided? **Yes. The proposed characters have different identities.**

If YES, reference: **See present document**

11. Does the proposal include use of combining characters and/or use of composite sequences? **Yes**

If YES, is a rationale for such use provided? **Yes**

If YES, reference: **Similarity to already encoded characters**

Is a list of composite sequences and their corresponding glyph images (graphic symbols) provided? **N/A**

If YES, reference: **N/A**

12. Does the proposal contain characters with any special properties such as control function or similar semantics? **Yes**

If YES, describe in detail (include attachment if necessary):
Similar to already encoded characters

13. Does the proposal contain any Ideographic compatibility characters? **No**

If YES, are the equivalent corresponding unified ideographic characters identified? **N/A**

If YES, reference: **N/A**