UTC Document Register **L2/14-207R**

Implications of the Unicode Arabic model for Warsh-based orthographies

Roozbeh Pournader, Google Inc.
August 9, 2014

Summary

In Evans 2014, the Warsh-based orthographies of Arabic is discussed, and three alternatives are suggested for an standard way to support the orthographies in Unicode. This document documents the model followed by existing Arabic characters and recommends a way forward.

The recent history of the present model

Most of the Arabic letters encoded in Unicode follow a relatively simple model. Although the original model wasn't very clear, the model was clarified, expanded and codified based on Pournader 2009, Mansour 2010, Priest and Hosken 2010, Pournader 2011, Pournader 2013 (among various other proposals), and the related UTC discussions happening around them.

Pournader 2009 led to a clarified Arabic model that made the *i'jam* patterns of Arabic letters normative by splitting Joining_Group classes that had different dot patterns into different Joining_Groups, creating two new Joining_Groups FARSI YEH and NYA. The author was awarded the Bulldog Award in 2009 partially based on that work.

The non-acceptance of the Mansour 2010 proposal about Yeh Barree, the various proposals for encoding Arabic digraphs considered letters in Turkic languages of China, and the various proposals asking for a unified hamza character with complex language-dependent contextual shaping was another piece of the puzzle. The UTC decisions made it clear that the Arabic letters in Unicode are intended to be stable, and do not always reflect the users' or the linguists' perception of a pedagogical, orthographic, or linguistic "letter" in a one-to-one mapping.

Priest and Hosken 2010 and Pournader 2011 led to a clarification of what the UTC would encode as new Arabic letters, and what it would consider already encoded as a sequence of already existing characters. Pournader 2011 was followed except for hamzas, where it was decided to encode new letters when hamza was an *i'jam* (and not a glottal stop or an *ezafe*). This was an expansion of the model used for the Pashto U+0681 ARABIC LETTER HAH WITH HAMZA ABOVE and the Ormuri U+076C ARABIC LETTER REH WITH HAMZA ABOVE, which led to the acceptance of Priest and

Hosken 2010 and the encoding of U+08A1 ARABIC LETTER BEH WITH HAMZA ABOVE for Fulfulde.

The latest piece of the puzzle was the acceptance of Pournader 2012, which asked for confirmation that the normative and stable *i'jam* model holds for every character, including U+06BA ARABIC LETTER NOON GHUNNA which was suffering from inconsistent implementation in different fonts and products. UTC's confirmation of the proposal strengthened the model.

Features and problems in the Arabic model

The Unicode Standard has provided a certain model for the Arabic script which is not perfect. The author would join several critiques in claiming that if we were re-encoding the Arabic script in Unicode today, we would have approached things very differently. Still, it is a very good model and has worked very well to serve hundreds of millions of users around the world.

Following are some of the features of the current model:

1. The phoneme-grapheme mapping is not one-to-one. For example:
    a. The glottal stop in Arabic and various other languages could be represented as U+0621, U+0622, U+0623, U+0624, U+0625, U+0626, U+0627, U+0653, U+0654, or U+0655.
    b. The vowel /u/ in Persian is represented as <U+0627, U+0648> in word-initial positions, but as <U+0648> in other positions.
    c. In Persian, the vowels /o/ and /u/, the diphthong /ow/, and the consonant /v/ could be represented as <U+0648>.
    d. The vowel /e/ in Urdu is represented as <U+0627, U+06CC> in word-initial positions, <U+06CC> in mid-word positions, and <U+06D2> in word-final positions.
    e. Both the /i/ and the /e/ vowels in Urdu could be represented as <U+06CC> in mid-word positions, while they would use different characters in word-final positions (respectively <U+06CC> and <U+06D2>).
    f. The /i/ vowel and the /j/ consonant in Afghan Pashto in mid-word positions could be represented by either U+064A or U+06CC, which look identical in medial form (Everson and Pournader 2003 recommends that only U+06CC be used medially and initially in Pashto, while Wikipedia 2014 suggests U+064A instead).
2. Characters could be specified to look exactly the same in some forms but not others. For example, the initial and medial shapes of U+0754 ARABIC LETTER BEH WITH TWO DOTS BELOW AND DOT ABOVE and U+08A9 ARABIC LETTER YEH WITH TWO DOTS BELOW AND DOT ABOVE are exactly the same, even when they are from different Joining_Groups. The same is true with the much more common U+064A ARABIC LETTER YEH and U+06CC ARABIC LETTER FARSI

YEH, which belong to the sister groups of YEH and FARSI YEH. The pattern happens with final and isolated forms too.

3. The dot patterns (*i'jam*) on letters is specified exactly and normatively by their Joining_Group and the difference between the representative letter in the Joining_Group and the chart shape of the character. (Note that there are still open issues about the exact model for the Heh class of characters, which are being studied in Pournader 2014 and follow-up documents, but those do not affect the *i'jam*.)
4. Characters that are considered different forms of the same letter in some languages are encoded separately if their Joining_Group would be different. For example, compare U+0643 ARABIC LETTER KAF with U+06A9 ARABIC LETTER KEHEH and U+06AA ARABIC LETTER SWASH KAF. Or more interestingly, U+06AC ARABIC LETTER KAF WITH DOT ABOVE and U+0762 ARABIC LETTER KEHEH WITH DOT ABOVE, which would not be considered different letters in any language, but are encoded separately simply because they would need to have a slightly different Joining_Group and one form is preferred over the other in Old Malay.

The implications for Warsh-based orthographies

The Unicode model–compatible approaches for the Warsh-based orthographies become clear when one considers the development of the Arabic writing system: The various letters started as dotless skeletons, roughly corresponding to the Unicode's basic Joining_Groups. These were later disunified in different ways by either adding *i'jam* (Beh vs Teh vs Theh vs Peh vs …), splitting Joining_Groups (Kaf vs Gaf vs Swash Kaf, Yeh vs Yeh Barree, Waw vs Straight Waw, …), or sometimes creating new Joining_Groups from scratch (Rohingya Yeh).

The Warsh-based orthographies of Africa simply added different dot patterns to skeletons to differentiate letters. While the mainstream Arabic script writing systems usually added dots to all the positional forms of the letters, the Warsh-based orthographies added them to some positional forms only, since they were considered "unnecessary" in other forms since as skeleton already differentiated the letters.

Note that the Warsh pattern has already been used in many existing Arabic letters in Unicode, specifically all characters of the Joining_Group FARSI YEH (which is used in Hafs-style Qurans, as well as several non-Arabic languages), BURUSHASKI YEH BARREE, and NYA.

The model of fixed and normative *i'jam* in the Unicode Standard make only the solutions 1 and 3 in Evans 2014 compatible with the Unicode Standard:

- Evans 2014's solution 1 (use different codepoints) would be generally compatible with the Unicode model, although it's not very desirable, as it leads to usability headaches. This would be similar to the encoding of the /e/

vowel in Urdu, which mixes U+06CC and U+06D2; or the /i/ vowel in Afghan Pashto (according to Everson and Pournader 2003 recommendations), which mixes U+06CC and U+064A. Such a solution should be generally avoided, unless features of the writing system itself or stability reasons enforce it.

- Evans 2014's solution 2 (orthography-dependent forms) would be the choice most incompatible with the present Unicode Arabic model. If chosen, the Unicode Arabic model would need to be modified and explained. While the author has tried his best over the years to clarify and document the existing Arabic model in Unicode, he doesn't know to reconcile such a solution with the existing model. Also, no higher-level markup system has a tag for Warsh-based orthographies, which would delay the implementation of such a solution.
- Evans 2014's solution 3 (encoding new characters) would be the model compatible with ten already-encoded letters, including the very common U+06CC ARABIC LETTER FARSI YEH. The rest of the characters following that model include U+063D, U+063E, U+063F, U+06BD, U+06CE, U+0755, U+0776, U+077A, U+077B. These characters are normatively defined to have different *i'jam* patterns in medial/initial vs final/isolated forms.

Considering all this, the author recommends three new characters to be encoded in the Arabic Extended-A block, with the following properties (the glyphs are provided in Evans 2014):

```
08BB;ARABIC LETTER AFRICAN FEH;Lo;0;AL;;;;;N;;;;;
08BC;ARABIC LETTER AFRICAN QAF;Lo;0;AL;;;;;N;;;;;
08BD;ARABIC LETTER AFRICAN NOON;Lo;0;AL;;;;;N;;;;;
```

The properties for ArabicShaping.txt would be as follows:

```
08BB; AFRICAN FEH; D; AFRICAN FEH
08BC; AFRICAN QAF; D; AFRICAN QAF
08BD; AFRICAN NOON; D; AFRICAN NOON
```

The three new joining groups created would join FARSI YEH, NYA, and BURUSHASKI YEH BARREE in having different *i'jam* in different contextual positions.

The following annotations are recommended for NamesList.txt:

```
@               Additions for African orthographies
08BB    ARABIC LETTER AFRICAN FEH
        * initial and medial forms of this letter have one dot below
        x (arabic letter dotless feh - 06A1)
        x (arabic letter feh with dot moved below - 06A2)
08BC    ARABIC LETTER AFRICAN QAF
        * initial and medial forms of this letter have one dot above
        x (arabic letter dotless qaf - 066F)
        x (arabic letter qaf with dot above - 06A7)
```

```
08BD    ARABIC LETTER AFRICAN NOON
        * initial and medial forms of this letter have one dot above
        x (arabic letter noon ghunna - 06BA)
        x (arabic letter noon - 0646)
```

Bibliography

1. Lorna Evans. 2014. “Supporting the Warsh orthography for Arabic script.” UTC Document Register L2/14-104. The Unicode Consortium. http://www.unicode.org/L2/L2014/14104-warsh.pdf
2. Michael Everson and Roozbeh Pournader. 2003. “Computer Locale Requirements for Afghanistan.” Afghan Transitional Islamic Administration Ministry of Communication and United Nations Development Programme Afghanistan. http://evertype.com/standards/af/af-locales.pdf (accessed August 7, 2014).
3. Kamal Mansour. 2010. “Problems with the joining behavior of Arabic letter Yeh Barree (U+06D2).” UTC Document Register L2/10-168. The Unicode Consortium. http://www.unicode.org/L2/L2010/10168-yb-problems.pdf
4. Roozbeh Pournader. 2009. “Moving dots and Arabic script shaping: Farsi Yeh’s and Jawi Nya.” UTC Document Register L2/09-146. The Unicode Consortium. http://www.unicode.org/L2/L2009/09146-moving-dots.pdf
5. Roozbeh Pournader. 2010. “Of hamza and other harakat.” UTC Docment register L2/10-455. The Unicode Consortium. http://www.unicode.org/L2/L2010/10455-of-hamza.pdf
6. Roozbeh Pournader. 2011. “Letter-making Arabic harakat.” UTC Document Register L2/11-069. The Unicode Consortium. http://www.unicode.org/L2/L2011/11069-arabic-harakat.pdf
7. Roozbeh Pounader. 2013. “Initial and medial forms of Arabic Letter Noon Ghunna.” UTC Document Register L2/12-381. The Unicode Consortium. http://www.unicode.org/L2/L2012/12381-dotless-noon.pdf
8. Roozbeh Pournader. 2014. “The right *heh*s for Arabic script orthographies of Sorani Kurdish and Uighur.” UTC Document Register L2/14-136. The Unicode Consortium. http://www.unicode.org/L2/L2014/14136-hehs-sorani-uighur.pdf
9. Lorna A. Priest and Martin Hosken. 2010. “Proposal to add ARABIC LETTER BEH WITH HAMZA ABOVE.” UTC Document Register L2/10-442R. The Unicode Consortium. http://www.unicode.org/L2/L2010/10442r-hamza-on-beh.pdf
10. Wikipedia contributors. 2014. “Pashto alphabet.” Wikipedia, The Free Encyclopedia. http://en.wikipedia.org/w/index.php?title=Pashto_alphabet&oldid=618144658 (accessed August 7, 2014).