L2/09-146

2009-04-15

Title:	Moving dots and Arabic script shaping: Farsi Yeh's and Jawi Nya
Source:	Roozbeh Pournader
Date:	2009-04-15

The Unicode Standard, as of version 5.1, does not provide enough information regarding the contextual shaping behavior of some Arabic letters. Although there is good machine-readable information provided as a part of Unicode data file ArabicShaping.txt, that information is not sufficient for determining whether a letter should be displayed with or without dots in initial and medial positions, or where and how should those dots appear.

Font and software vendors planning to support certain Arabic letters would need extra information to implement them. This is true even for very basic non-typographic support, like simple fonts used in user interfaces (e.g., Tahoma for Microsoft Windows or DejaVu Sans for GNU/Linux).

The author believes this was not the intention of the authors of the Unicode Standard, and that this could be easily remedied by splitting a couple of Arabic joining groups into two.

The expectation

As of Unicode 5.1, a font/software vendor planning to support an Arabic script letter with limited access to extra information or local experts, would depend on what is already provided in the Unicode standard:

- The text and tables of Section 8.2 of The Unicode Standard 5.0, titled Arabic;
- Character shapes as provided in the code charts;
- Information provided in the data file ArabicShaping.txt.

Putting those together, the simple strategy for a vendor planning to support an Arabic letter would be the following strategy.

Strategy A:

1. Find the Joining Group for the letter from ArabicShaping.txt;
2. Find the skeletal shape for the group from Tables 8-7 and 8-8 of TUS 5.0;
3. Find the differences between the glyph shape provided in the code chart and the isolated shape of the letter from the column labeled Xn;[1]
4. Apply the differences found in step 3 to the other shapes in the table row, to get the contextual shapes for the letter in question.

1 Except for HEH, where initial/Xl would be used, and for HEH GOAL and HAMZA ON HEH GOAL, where final/Xr would be used. This is because those forms are traditionally considered more representative of the characters in these groups, and the code charts follow the tradition of the users of the Arabic script.

For example, for U+0686 ARABIC LETTER TCHEH, the vendor would find that the Joining Group is HAH, and would then find the following row in Table 8-7 of TUS 5.0 (page 280):

Group	Xn	Xr	Xm	Xl	Notes
HAH	ڃ	ڃ	ـڃـ	ـڃ	Includes KHAH and JEEM.

The vendor would then check the glyph from the charts for U+0686, and compare it to the HAH in Xn form:

Xn form for HAH (from Table 8-7)	Glyph shape for U+0686 (from the charts)
ڃ	ڃ

Looking at the differences, the vendor would assume that U+0686 ARABIC LETTER TCHEH adds three dots pointing downwards under the letter. He would correctly conclude that the four main shapes should be displayed like these:

Letter	Xn	Xr	Xm	Xl
TCHEH	ڃ	ڃ	ـڃـ	ـڃ

The differences can be a bit more complex than addition or removal of marks or dot patterns. For example, in the case of U+06C0 ARABIC LETTER HEH WITH YEH ABOVE, the difference with the glyphs from Table 8-8 (the TEH MARBUTA group) is the replacement of two horizontal dots with a small *hamza* mark. Still, the replacements often follow the same pattern: some mark is added, some mark is removed, or some mark is replaced by another. The new marks are assumed to be added to and removed from the same relative position.

The author believes that this has been the intention of the authors of the Unicode Standard. But Strategy A does not result in the right shapes for eight Arabic letters (out of the 178 encoded in Unicode 5.1).

The problematic cases

The problem appears when the dot patterns do not follow the standard pattern of other extended forms. The letters with this problem fall into two classes: the Farsi Yeh's and the Jawi Nya.

The Farsi Yeh's

U+06CC ARABIC LETTER FARSI YEH is used in Persian, Urdu, Pashto, Azerbaijani, Kurdish, and various minority languages written in the Arabic script, and also Koranic Arabic. It behaves differently from most Arabic letters, in a way surprising to native Arabic language speakers. The letter has two horizontal dots below the skeleton in initial and medial forms, but no dots in final and isolated forms.

Compared to the two Arabic language Yeh forms, FARSI YEH is exactly like U+0649 ARABIC LETTER ALEF MAKSURA in final and isolated forms, but exactly like U+064A ARABIC LETTER YEH in initial and medial forms:

Code	Letter	Xn	Xr	Xm	Xl
U+0649	ALEF MAKSURA	ى	ى	ـى	ـى
U+064A	YEH	ي	ي	ـيـ	ـي
U+06CC	FARSI YEH	ى	ى	ـيـ	ـي

When a vendor tries to follow Strategy A for FARSI YEH, he first looks at the file ArabicShaping.txt:

```
06CC; DOTLESS YEH; D; YEH
```

This leads him to the following row from Table 8-7:

Group	Xn	Xr	Xm	Xl	Notes
YEH	ي	ي	ـيـ	ـي	Includes ALEF MAKSURA.

He then compares the glyph from the charts to the glyph in the Xn column:

Xn form for YEH (from Table 8-7)	Glyph shape for U+06CC (from the charts)
ي	ى

From the comparison, and the information in ArabicShaping.txt (DOTLESS YEH), he sees that the difference is the removal of the dots. He may conclude the contextual glyphs should look like these:

Letter	Xn	Xr	Xm	Xl
FARSI YEH	ى	ى	ـى*	ـى*

The only information to the contrary in the Unicode Standard is the character notes for this letter from the code charts/NamesList.txt: “initial and medial forms of this letter have dots”. Still, there is no mention of how many dots, or in which relative position and arrangement.

Unfortunately, even that is not provided for other letters with the same behavior, i.e. the Yeh forms with dots in initial and medial forms but no dots in final and isolated forms. These letters are:

- U+063D ARABIC LETTER FARSI YEH WITH INVERTED V (used in Azerbaijani)
- U+063E ARABIC LETTER FARSI YEH WITH TWO DOTS ABOVE (used in early Persian)
- U+063F ARABIC LETTER FARSI YEH WITH THREE DOTS ABOVE (used in early

Persian)

- U+06CE ARABIC LETTER YEH WITH SMALL V (used in Kurdish)
- U+0775 ARABIC LETTER FARSI YEH WITH EXTENDED ARABIC-INDIC DIGIT TWO ABOVE (used in Burushaski)
- U+0776 ARABIC LETTER FARSI YEH WITH EXTENDED ARABIC-INDIC DIGIT THREE ABOVE (used in Burushaski)

Interestingly enough, U+0777 ARABIC LETTER FARSI YEH WITH EXTENDED ARABIC-INDIC DIGIT FOUR BELOW, does not belong to the above group, although its character name includes "FARSI YEH".[2]

The Jawi Nya

The letter U+06BD ARABIC LETTER NOON WITH THREE DOTS ABOVE, used in Jawi, is more interesting. Called Nya in Jawi, in final and isolated forms it has three dots above the skeleton, but in initial and medial forms the dots move below the skeleton and invert. The Jawi orthography does this in order to avoid confusion with U+062B ARABIC LETTER THEH, which is used in Jawi for some loanwords from the Arabic language.

This means that in initial and medial forms, NOON WITH THREE DOTS ABOVE has the same shape as U+067E ARABIC LETTER PEH. The following table compares the three letters mentioned above, together with the shapes for Joining Group NOON (that Nya currently belongs to) from Table 8-7 of TUS 5.0:

Code	**Letter**	**Xn**	**Xr**	**Xm**	**Xl**
U+062B	THEH	ﺙ	ﺚ	ﺜ	ﺛ
(U+0646)	NOON	ﻥ	ﻦ	ﻨ	ﻧ
U+067E	PEH	ﭖ	ﭗ	ﭙ	ﭘ
U+06BD	NOON WITH THREE DOTS ABOVE	ڽ	ڼ	ڽ	ڼ

But following Strategy A, one would arrive at the following shapes:

Letter	**Xn**	**Xr**	**Xm**	**Xl**
NOON WITH THREE DOTS ABOVE	ڽ	ڼ	ڽ*	ڼ*

2 This can be verified by checking the original Unicode proposal for the letter, L2/06-149, page 20. In the same table that shows a sample of this letter in isolated form, the letter is also seen in medial form, with no dots.

Fortunately, the desired behavior for Nya is explained in TUS 5.0, page 281, which happens to be next to Tables 8-7 and 8-8. Quoting from page 281 of TUS 5.0:

> ***Jawi.*** U+06BD ARABIC LETTER NOON WITH THREE DOTS ABOVE is used for Jawi, which is Malay written using the Arabic script. Malay users know the character as *Jawi Nya*. Contrary to what is suggested by its Unicode character name, U+06BD displays with the three dots below the letter when it is in the initial or medial position. This is done to avoid confusion with U+062B ARABIC LETTER THEH, which appears in words of Arabic origin, and which has the same base letter shapes in initial or medial position, but with three dots above in all positions.

If the vendor is lucky enough, he will run into the information while checking the tables and will find about the relative position of the dots. But unfortunately, no glyphs are shown in the Unicode Standard for initial and medial forms of the Jawi Nya. This means that the vendor can not find if the dot pattern that moves below the skeleton does so pointing upwards, or downwards.

For example, after reading the current text of the Unicode Standard, a vendor may incorrectly assume that the dot pattern for initial and medial forms of the Jawi Nya should be kept in the same direction as they were in the isolated form of the letter, pointing upwards:

Letter	**Xn**	**Xr**	**Xm**	**Xl**
NOON WITH THREE DOTS ABOVE	ں	ں	*ہ	*ہ

The proposal

Presently, users of the Unicode Standard have no way of determining the exact arrangement of dot patterns or existence of dots in initial and medial forms for the Arabic letters under discussion.

A vendor trying to implement these letters would need to research the special shaping using the archives of L2/UTC documents (if available to them) and from local experts. Experts may be very hard to find, considering that most of these letters are only used in minority languages or in historical orthographies. The lack of information would result in vendors implementing these letters incorrectly, in turn leading to content providers using non-standard tricks to get their text to display correctly.[3]

The author believes that the two Arabic Joining Groups containing these letters should be split. These groups are YEH (that contains the various Farsi Yeh forms) and NOON (that contains the Jawi Nya).

Two new Arabic Joining Groups will be created. These will be named “FARSI YEH” and “NYA”, and will be added to Table 8-7 of the Unicode Standard, with the following contextual shapes:

Group	**Xn**	**Xr**	**Xm**	**Xl**	**Notes**
FARSI YEH	ى	ى	ـيـ	ـي	
NYA	ڻ	ڻ	ـڹـ	ـڹ	

Finally, the following lines in the file ArabicShaping.txt will be changed to reflect these splits, using improved schematic names that hint towards correct shapes for all contextual forms (important changes and non-changes are in boldface):

```
063D; FARSI YEH WITH INVERTED V; D; FARSI YEH
063E; FARSI YEH WITH 2 DOTS ABOVE; D; FARSI YEH
063F; FARSI YEH WITH 3 DOTS ABOVE; D; FARSI YEH
06BD; NYA; D; NYA
06CC; FARSI YEH; D; FARSI YEH
06CE; FARSI YEH WITH SMALL V; D; FARSI YEH
0775; FARSI YEH WITH DIGIT TWO ABOVE; D; FARSI YEH
0776; FARSI YEH WITH DIGIT THREE ABOVE; D; FARSI YEH
0777; YEH WITH DIGIT FOUR BELOW; D; YEH
```

Other derived data files (like DerivedJoiningGroup.txt) would need to be changed accordingly.

3 For example, lack of information in the Unicode Standard may have led to incorrect font support for U+06CC ARABIC LETTER FARSI YEH in various Microsoft fonts, first shipped with Internet Explorer 5. That bug resulted in several early adopters of Unicode opting to use U+064A ARABIC LETTER YEH instead of U+06CC in Persian text, when the Persian letter Yeh appeared in initial and medial forms. This was to make sure the text always displayed correctly, sacrificing other aspects of Persian computing, especially searching. Although Microsoft was notified of the Persian Yeh bug in early 1999, it took years to correct the problem in all Arabic script fonts shipped with Microsoft products. The first Windows version (even including service packs) that had the Persian Yeh bug fixed in all Arabic fonts was released in 2007.

Editorial corrections

During the research leading to this proposal, the author found some errors in the comments parts of the data file ArabicShaping.txt (version 5.1.0). He would appreciate the correction of these:

- There is a paragraph saying:

  ```
  # This file defines the shaping classes for Arabic and Syriac
  # positional shaping, repeating in machine readable form the
  # information printed in Tables 8-3, 8-7, 8-8, 8-11, 8-12, and
  # 8-13 of The Unicode Standard, Version 5.0.
  ```

  The above paragraph needs to be updated.

  First, there is also information available about N'Ko in the file, so the scripts mentioned should be "Arabic, Syriac, and N'Ko", and the table list should include Table 13-5.

  Then, the mentioned Tables do not contain all the information provided in ArabicShaping.txt, so it is misleading to consider the file to be "repeating" the same information in machine readable form. That was only true with TUS 4.0, that included a list of character codes in its Tables 8-7 and 8-8 (pages 203–204).

- There is another paragraph saying:

  ```
  # Each line contains four fields, separated by a semicolon.
  ```

  It may be better to say "separated by a semicolon **and a space**."

Also, it appears that Tables 8-7 and 8-8 of the Unicode book need to be updated to reflect the change of Joining Type and Joining Group for U+06C2 ARABIC LETTER HEH GOAL WITH HAMZA ABOVE that happened with Unicode 4.1:

- In Table 8-7, in the notes for the joining group HEH GOAL, is is said that it "Excludes HAMZA ON HEH GOAL". But that letter, U+06C2, now belongs to the group. The author believes that the note should be replaced with "**Includes** HEH GOAL WITH HAMZA ABOVE".
- The author wishes to ask the editorial committee to consider replacing the the representative character for the group HAMZA ON HEH GOAL in Table 8-8 of the Unicode book. The character currently used in the table, U+06C2 ARABIC LETTER HEH GOAL WITH HAMZA ABOVE, no longer belongs to that group.[4] The author suggests using U+06C3 ARABIC LETTER TEH MARBUTA GOAL as the representative character for the group. This would help minimize the possible confusion of font/software vendors about the standard joining type and group of U+06C2 ARABIC LETTER HEH GOAL WITH HAMZA ABOVE.

4 The old group name has only been kept for stability reasons. See the header of the file ArabicShaping.txt for more information.