Introduction

Quick Summary. This is a discussion paper on the behavior for U+06CC ARABIC LETTER FARSI YEH when followed by U+0654 ARABIC HAMZA ABOVE.

Background.
A user of SIL’s Arabic fonts stated that U+06CC should lose the dots on U+06CC when followed by a hamza above (U+0654).

Annotations for:

U+064A has:
- loses its dots when used in combination with 0654
- retains its dots when used in combination with other combining marks

while U+06CC just has this:
- initial and medial forms of this letter have dots

Additionally, TUS (page 395) says:

U+0654 ARABIC HAMZA ABOVE should not be used with U+0649 ARABIC LETTER ALEF MAKSURA. Instead, the precomposed U+0626 ARABIC LETTER YEH WITH HAMZA ABOVE should be used to represent a yeh-shaped base with no dots in any positional form, and with a hamza above. Because U+0626 is canonically equivalent to the sequence <U+064A ARABIC LETTER YEH, U+0654 ARABIC HAMZA ABOVE>, when U+0654 is applied to U+064A ARABIC LETTER YEH, the yeh should lose its dots in all positional forms, even though yeh retains its dots when combined with other marks.

So, XB Niloofar (a popular Persian font), and the SIL fonts have implemented U+064A to lose the dots with U+0654. Although most industry fonts do lose the dots when followed by the hamza above some software and operating systems unfortunately override the behavior specified in the font.
In SIL fonts U+06CC does not lose the dots in combination with U+0654. The author looked at various Arabic script fonts and they all seem to retain the dot for U+06CC U+0654.

<table>
<thead>
<tr>
<th></th>
<th>U+06CC (farsi yeh) U+0654</th>
<th>U+064A (yeh) U+0654</th>
<th>U+0649 (alef maksura) U+0654</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scheherazade New</td>
<td>وُسْبُيِّيَّ َكِ ِّبِضْيَيِّ َي</td>
<td>وُسْبُيِّيَّ َكِ ِّبِضْيَيِّ َي</td>
<td>وُسْبُيِّيَّ َكِ ِّبِضْيَيِّ َي</td>
</tr>
<tr>
<td>XB Niloofar (Persian font)</td>
<td>وُسْبُيِّيَّ َكِ ِّبِضْيَيِّ َي</td>
<td>وُسْبُيِّيَّ َكِ ِّبِضْيَيِّ َي</td>
<td>وُسْبُيِّيَّ َكِ ِّبِضْيَيِّ َي</td>
</tr>
<tr>
<td>Times New Roman</td>
<td>وُسْبُيِّيَّ َكِ ِّبِضْيَيِّ َي</td>
<td>وُسْبُيِّيَّ َكِ ِّبِضْيَيِّ َي</td>
<td>وُسْبُيِّيَّ َكِ ِّبِضْيَيِّ َي</td>
</tr>
<tr>
<td>Arabic Typesetting</td>
<td>وُسْبُيِّيَّ َكِ ِّبِضْيَيِّ َي</td>
<td>وُسْبُيِّيَّ َكِ ِّبِضْيَيِّ َي</td>
<td>وُسْبُيِّيَّ َكِ ِّبِضْيَيِّ َي</td>
</tr>
<tr>
<td>Calibri</td>
<td>وُسْبُيِّيَّ َكِ ِّبِضْيَيِّ َي</td>
<td>وُسْبُيِّيَّ َكِ ِّبِضْيَيِّ َي</td>
<td>وُسْبُيِّيَّ َكِ ِّبِضْيَيِّ َي</td>
</tr>
<tr>
<td>Adobe Arabic</td>
<td>وُسْبُيِّيَّ َكِ ِّبِضْيَيِّ َي</td>
<td>وُسْبُيِّيَّ َكِ ِّبِضْيَيِّ َي</td>
<td>وُسْبُيِّيَّ َكِ ِّبِضْيَيِّ َي</td>
</tr>
<tr>
<td>Bressay Arabic</td>
<td>وُسْبُيِّيَّ َكِ ِّبِضْيَيِّ َي</td>
<td>وُسْبُيِّيَّ َكِ ِّبِضْيَيِّ َي</td>
<td>وُسْبُيِّيَّ َكِ ِّبِضْيَيِّ َي</td>
</tr>
<tr>
<td>ArabicUIText</td>
<td>وُسْبُيِّيَّ َكِ ِّبِضْيَيِّ َي</td>
<td>وُسْبُيِّيَّ َكِ ِّبِضْيَيِّ َي</td>
<td>وُسْبُيِّيَّ َكِ ِّبِضْيَيِّ َي</td>
</tr>
<tr>
<td>Noto Naskh Arabic</td>
<td>وُسْبُيِّيَّ َكِ ِّبِضْيَيِّ َي</td>
<td>وُسْبُيِّيَّ َكِ ِّبِضْيَيِّ َي</td>
<td>وُسْبُيِّيَّ َكِ ِّبِضْيَيِّ َي</td>
</tr>
</tbody>
</table>

Pournader’s document on “moving dots” led to the current documentation in Unicode chapter 9 (page 391) which says:

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, as shown in Table 9-10.

Neither Pournader’s document nor TUS address anything regarding dots on U+06CC (farsi yeh) plus hamza above. However, when it says “but exactly like U+064A ARABIC LETTER YEH in initial and medial forms” it might lead us to the conclusion that it does lose its dots in combination with U+0654.

So, my question began with “Should U+06CC lose its dots in combination with U+0654 and if so, should it be documented? Perhaps it should be documented even if it does not lose its dots.”

Pournader (personal communication) says:

The Iranian standard ISIRI 6219, which is co-authored by me, says¹ U+06CC should lose its dots when combined with hamza above, but since then, I have arrived at the conclusion that it should not lose its dots when combined with hamza above.

The reason is that some Azerbaijani orthographies have a yeh-hamza form that has dots and a hamza above in initial and medial forms, but just the hamza in final and initial forms. For those orthographies, I don't want us to encode a new character, or tell them to use different characters. I want to advise them to use U+06CC+hamza above for their letter.

---

¹ ISIRI 6219, page 14, note 2: “if the characters HAMZA ABOVE or HAMZA BELOW are combined with FARSI YEH or ‘DOTTED ARABIC’ YEH, the base character loses its dots.”

Dot behavior for U+06CC ARABIC LETTER FARSI YEH followed by U+0654 ARABIC HAMZA ABOVE (Page 2)
The Iranian standard isn’t widely implemented, and I’m sure nobody even remembers that obscure section of it. I don’t think Unicode has ever explicitly said that U+06CC should lose its dots when combined with hamza above, and I think it was wrong of the Iranian standard (i.e. me) to specify such a thing without asking Unicode to do that first.

Pournader provided a sample from an authoritative Azerbaijani dictionary (Behdazi, page 74) demonstrating that farsi yeh retains its dots in initial and medial positions in the Azerbaijani language:

If there is agreement that U+06CC followed by U+0654 should not lose its dots, this should be documented since it is unclear.

Suggested actions

Since there is no indication of how many dots are on U+06CC, the annotation for U+06CC should be updated to include information on how many dots and to include the information on retaining the dots when used in combination with 0654:

- initial and medial forms of this letter have two horizontal dots below
- retains its dots in initial and medial forms when used in combination with 0654

Additionally, chapter 9 should add this sentence:

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, as shown in Table 9-10. However, U+06CC ARABIC LETTER Farsi Yeh followed by U+0654 ARABIC HAMZA ABOVE should retain its dots in initial and medial forms.
References


