Title **Problems with the joining behavior of Arabic Letter Yeh Barree (U+06D2)**

Author Kamal Mansour, Monotype Imaging Inc.

Date 2010-05-04

The Urdu language, along with other South Asian languages, uses the Farsi Yeh (U+06CC) to write /i/ and the Yeh Barree (U+06D2) for /e/. Just as the /i/ can appear in initial, medial, and final position in an Urdu word, so can /e/. In terms of common writing practice, the word-initial and -medial representations of /e/ are usually identical to those of /i/. Whenever it is important to distinguish between /e/ and /i/ in an initial or medial context, a combining mark (Arabic Subscript Alef U+0656) is added to the glyph representing /i/. The name 'Benazir'—pronounced as /benazi:r/— turns out to be an ideal way to demonstrate these visual and phonetic distinctions. Usually, it is written as بینظیر where the same medial glyph for /e/ is used also for /i/. If one wanted to ensure correct pronunciation, the ambiguity would be clarified by adding the mark Subscript Alef (U+0656) under the glyph representing the /i/ as follows: بینظٖیر

In this spelling, one can correctly infer that the unmarked glyph represents /e/. As further evidence that the speakers of Urdu are consciously aware of this phonetic distinction, 'Benazir' can be alternatively spelled as two morphemes بےنظیر where the /e/ is written unambiguously with the Yeh Barree. A simple search on the internet will provide ample evidence for this alternation.

What is current encoding practice?

The Unicode attributes of Yeh Barree (U+06D2) indicate that it is a right-linking character which implies that it has only isolated and final shapes, and appears only in word-final context. Consequently, when 'Benazir' is represented as a single morpheme, people use Farsi Yeh (U+06CC) to represent the medial /e/.

The fact that both Yeh Barree and Yeh share the same initial and medial forms led to the wrong conclusion that only the Yeh appears in those positions. By inference, the Yeh Barree was erroneously classified as right-linking.

Recent additions for Burushaski U+077A and 077B (both characters consist of Yeh

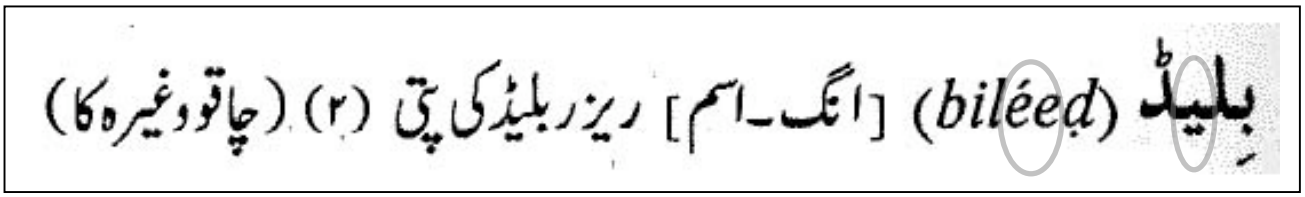
Barree with a modifying mark) were classified under a new joining group of “Burushaski Yeh Barree” to allow them to appear in any word position. In recent comments from Elaine Bashir, collaborator on the Urdu-Burushaski Dictionary project at University of Chicago, the dictionary makers were similarly expecting to be able to represent the unmarked Yeh Barree (U+06D2) in any word position, but were surprised to find that such use always resulted in word break. So, it has become apparent that Burushaski —and other languages— requires a Yeh Barree character in the same joining group as U+077A and 077B, but such a character does not currently exist. Figure 1 shows samples of the unmarked Yeh Barree (U+06D2) in initial, medial, and final position in entries of the Urdu-Burushaski Dictionary. The pertinent glyphs of Arabic script, as well as their Latin transcription, are indicated with grey ovals.

Where to from here?

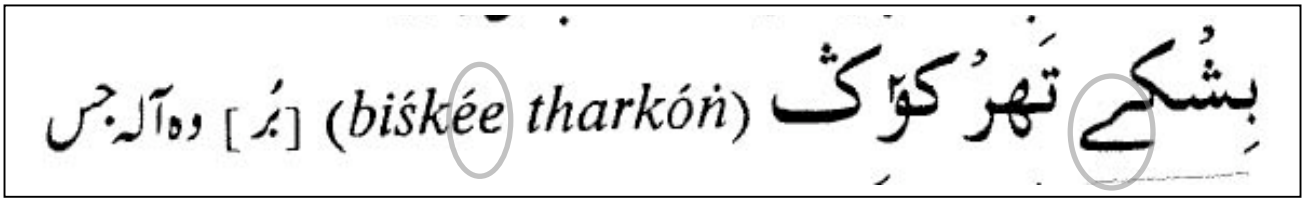
The easiest way forward would be to simply encode a new Yeh Barree character that is both right- and left-linking; i.e., with the joining group “Burushaski Yeh Barree”. Such an addition would result in new, alternative ways to encode words spelled with a Yeh Barree, but would satisfy the requirements of appropriate lexical and graphic representation. I would also recommend the renaming of the joining group “Burushaski Yeh Barree” to reflect its use in multiple languages.

Alternatively, the UTC could revise the joining group of the original Yeh Barree (U+06D2) to be both right- and left-linking. What would be the practical consequences of such a change? The rendering of texts currently encoded to display an initial or medial Yeh form to represent /e/ would remain unchanged because they use either a Yeh (U+64A) or Farsi Yeh (U+06CC). However, texts currently encoded with a Yeh Barree in initial or medial position, would render in newly joined form. For instance, assuming no space or other control character appears between the two morphemes, بےنظیر would render as بینظیر. One would need to interject a control character between the two morphemes in order to keep them visually separate as in بےنظیر.

Yeh Barree in initial position



Yeh Barree in medial position



Yeh Barree in final position

Figure 1