

Annotations for Kashmiri Yeh

To: Unicode Technical Committee

Author: Manish Goregaokar (मनीष गोरेगांवकर) <manish@unicode.org>









Date: Oct 16, 2023

This proposal recommends adding some annotations to the names list, discouraging the use of [U+06CD](#) ARABIC LETTER YEH WITH TAIL to represent Taler Ye in Kashmiri, recommending [U+0620](#) ARABIC LETTER KASHMIRI YEH instead.

Background

Kashmiri is an Indo-Aryan language that has traditional orthographies in multiple scripts, including Perso-Arabic. The Perso-Arabic orthography for Kashmiri is structured similarly to that of Urdu with some additional letters, including an additional letter to mark palatalization.

This letter was requested in [L2/09-176](#), subsequently encoded as [U+0620](#) ARABIC LETTER KASHMIRI YEH in Unicode 6.0. It currently looks like this (in Noto fonts):

U+0620	Final	Medial	Initial	Isolated
Nastaliq				
Naskh				

The form with the ring is usually called “Gol Ye” (“round yeh”), whilst the form above without it is called “Taler Ye” (“half yeh”).

Kashmiri is typically written in the Nastaliq style. It’s not quite clear if the Gol Ye form is valid for this letter in final/isolated position in the Naskh style, or if this is a bug in fonts. One can find fonts of both kinds; to avoid confusion these forms have been marked in pink.

Kashmiri also uses [U+06CC](#) ARABIC LETTER FARSI YEH, which looks like this:

	Final	Medial	Initial	Isolated
Nastaliq	ی	یہ	یہ	یہ
Naskh	یہ	یہ	یہ	یہ

(It additionally uses [U+06D2](#) ARABIC LETTER YEH BARREE which has a very distinct form and is not relevant here)

Gol vs Talur (Round vs Half)

While both the Gol and Talur (round and half, respectively) forms of KASHMIRI YEH mark palatalization, they do have slightly different effects, typically the Taler Ye will be a pure palatalization, for example in شری

/ʃuri/, but the Gol Ye will have an associated vowel value, for example in دیہ /diːd/.

This does lead to a perception of them being separate letters.

There is a *potential* case for disunification: It appears that the Taler Ye form can be found inside a word, at certain morpheme boundaries and also before another Ye. Some examples of the former include گوری باے

(“milkwoman”) and ٹرونی کُل (“walnut tree”). Some examples of the latter include دری یاو (“river”, “ocean”)

and بنی یان (“sweater”). This is currently representable with a Zero-Width Non-Joiner character, but that is suboptimal. The author of this proposal is interested in pursuing this avenue in the future, but wishes to keep this proposal scoped towards improving annotations for now.

Yeh with Tail

Unicode has a letter [U+06CD](#) ARABIC LETTER YEH WITH TAIL for use in Pashto (since 1.1), only found at the end of words for the feminine marker.

	Final	Isolated
Nastaliq	ۛ	ۛ
Naskh	ۛ	ۛ

It was not encoded for Kashmiri, nor does it look particularly like the Taler form of KASHMIRI YEH. It does look “different from FARSI YEH” which makes it a tempting way to represent Taler Ye when there is no font/keyboard support.

Usage in the wild

[CLDR-16595](#) reported that CLDR uses U+06CD YEH WITH TAIL in Kashmiri. Upon investigation, it was discovered that U+06CD YEH WITH TAIL is sporadically found in digital Kashmiri-language texts, only in the final position, as a replacement for U+0620 KASHMIRI YEH. This incorrect usage is in part due to slow Unicode and font support for the letter. Things seem to largely be better now, but broken text is still out there in the wild.

Bing Translate will also produce this code point. For example, it uses it instead of KASHMIRI YEH in the word

شُری [here](#).

Gboard also exposes this alongside KASHMIRI YEH on long press, and since KASHMIRI YEH in the Noto Naskh font used in the keyboard shows the Gol form only, it easily misleads people into using the YEH WITH TAIL to represent Taler Ye.



Proposal

Given that this confusion exists in the wild; it is worth it for Unicode to give a clearer authoritative signal on character identity here. I would like to propose the following annotations:

On [U+0620](#) ARABIC LETTER KASHMIRI YEH:

- Used in Kashmiri to represent Half Ye / Taler Ye

On [U+06CD](#) ARABIC LETTER YEH WITH TAIL:

- Should not be used in Kashmiri to represent Half Ye / Taler Ye (0620 ي)

In the long run we can consider whether things need to be disunified here, but until then we should avoid having more sources produce Kashmiri text with an incorrect Pashto letter especially when at least some fonts support this correctly now.

Further reading

- [CLDR-16595](#) Multiple issues in CLDR ks.xml
- [L2/09-176](#) Proposal to add two Kashmiri characters and one annotation to the Arabic block
- [Kashmiri orthography notes](#)
- [W3C alreq #249](#) Lack of font support for Kashmiri characters makes text deviate from true semantics