

TO: UTC

TITLE: Proposal to add two Kashmiri characters and one annotation to the Arabic block

SOURCE: Muzaffar Aazim, Kamal Mansour, and Roozbeh Pournader

DATE: 30 April 2009

This is a request to add two characters to the Arabic block and add one annotation in order to more fully represent the Kashmiri language in the Arabic script.


0. Background

Kashmiri is a Dardic language, a member of the Indo-European family of languages, spoken in Indian-administered state of Jammu and Kashmir and the Pakistani-administered state of Azad Kashmir. The Arabic script is the traditional and official orthography for Kashmiri; it has been in use since the fifteenth century and is used currently by the people of Kashmir and the federal and state agencies of India, including the Kashmir State's department of education. (The Devanagari script is also used by parts of the Hindu community to write the Kashmiri language.)

Kashmiri orthography has only been standardized in the past fifty years. It is currently taught in all schools in Kashmir, including colleges. Kashmir University has both masters and doctorate courses in Kashmiri.

The proposed characters appear in newspapers and other publications, such as the *Weekly Sangarmal* (<http://www.sangarmal.net/home.html>)

1. Addition of Annotation to U+0658

We request an annotation be added to U+0658 ARABIC MARK NOON GHUNNA. In Kashmiri, this character is used to indicate nasalization, just as in Urdu. However, the glyph is different from that of Urdu. In Kashmiri, it appears an inverted small v above. The glyph appears to be similar in shape to U+065B ARABIC VOWEL SIGN INVERTED SMALL V ABOVE: .

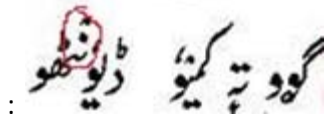
We request the following annotation be added to U+0658:

0658  ARABIC MARK NOON GHUNNA

- glyph is an inverted v-shape in Kashmiri


(Note: 0658 character now has an annotation, "Kashmiri and Baluchi," so while it is used in Kashmiri, the glyph currently shown in 0658 for ARABIC MARK NOON GHUNNA is inappropriate for Kashmiri.)


Example from Kashmiri book *Kuliyate Sheikhul-Alam* published by the Jammu and Kashmir Academy of Art Culture and Languages, Srinagar (India), 1985 (p. 31):



2. Two Proposed characters

a. 0620 ARABIC LETTER KASHMIRI YEH

This character is used to indicate palatalization. ARABIC LETTER KASHMIRI YEH  may occur in all positions, but is especially found in initially and medially.

This character has a “half yeh” variant, , that appears commonly in final or isolated contexts.

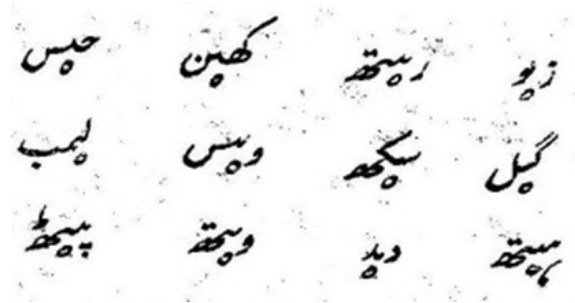
The form with the ring below is included with its own codepoint in the Indian standard PASCII:

187		LETTER YE (CIRCLE BELOW) • Kashmiri
-----	---	--

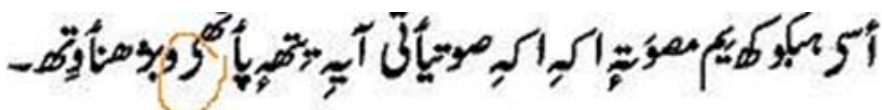
The proposed location for this character is U+0620.

The name of this character is proposed to be ARABIC LETTER KASHMIRI YEH (instead of ARABIC LETTER YEH WITH RING) to hint to the character's identity, since the letter may appear without any ring and with a very different shape in final and isolated forms.

Example of yeh from *Kaeshir Acchar Zaan*, a primer written by Amin Kamil (1866):



Example of the “half yeh” from *Ilm-O-Adab*, the official Magazine of Kashmiri Department of the Kashmir University (p. 220):



b. 06XX ARABIC WAVY HAMZA BELOW

This combining mark is used to indicate a common vowel in Kashmiri, and can appear under many base characters. It appears under alef in U+0673 ARABIC LETTER ALEF WITH WAVY HAMZA BELOW.

The wavy hamza below, used for Kashmiri, is contained in the Indian standard PASCII:

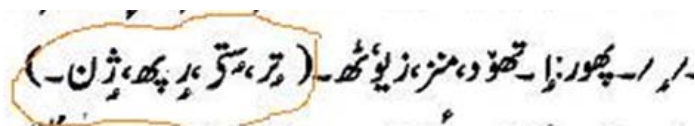
197	~	Diacritic Mark (Hamza Below) Kashmiri
-----	---	---------------------------------------

Proposed glyph and name:

⦿ 06XX ARABIC WAVY HAMZA BELOW
~

A possible location for this character would be 065F.

Example from *Ilm-O-Adab*, the official Magazine of Kashmiri Department of the Kashmir University (p. 220):



WAVY HAMZA BELOW raises normalization questions. The options on how to deal with normalization are described below.

Normalization Issues for ARABIC WAVY HAMZA BELOW

There are issues regarding normalization for this character.

A precomposed form of U+0627 ARABIC LETTER ALEF and ARABIC WAVY HAMZA BELOW already exists in Unicode: U+0673 ARABIC LETTER ALEF WITH WAVY HAMZA BELOW (according to character annotations, this character is used in Baluchi and Kashmiri). This has implications on normalization and canonical equivalence.

In Kashmiri orthography, *wavy hamza below* is used with several letters, including Alef. After encoding a combining ARABIC WAVY HAMZA BELOW in the standard, the abstract character (sequence) *alef with wavy hamza below* may be represented in two ways: <0627, 06xx> or <0673>.

In a perfect world, the canonical decomposition mapping for U+0673 would be changed to accommodate for the encoding of a new combining character. But according to Unicode stability policies, “Once a character is assigned, its decomposition mapping will not change”.

Allowing both representations to coexist, with no explanation, will result in text getting encoded two ways with no equivalency relation to each other, causing various problems, including security issues. The problems would not be limited to Kashmiri, as other languages may use any of the characters mentioned.

Instead, one of the following solutions should be chosen, and the text of the Unicode Standard should

be amended to reflect the solution. The authors prefer option A.

A) Deprecate U+0673 ARABIC LETTER ALEF WITH WAVY HAMZA BELOW and suggest that everyone uses <0627, 06xx> in the future.

Pros: Future text in Unicode will only have one recommended way to represent the abstract entity.

Cons: The precomposed character may have been used in existing data (including data in other languages), which needs to be converted. Users who ignore the deprecation will cause text processing problems.

B) Recommend against using the sequence <0627, 06xx>, and recommend use of U+0673.

This is comparable to recommendations in various Indic scripts to use atomic vowel letters, and not the sequence of codepoints resulting from visual analysis (see *The Unicode Standard 5.0*, page 299, Table 9-1). If this solution is chosen, text rendering processes would be recommended to show an error visually, like rendering the sequence <0627, 06xx> with a *wavy hamza below* under a dotted circle that follows the Alef, instead of rendering it under the Alef.

This is a weaker deprecation compared to deprecations of the precomposed character (option A), passing the problem to processes like rendering, spell checking, etc. If this solution is chosen, Kashmiri users would be asked to use a precomposed Unicode character when the abstract character Wavy Hamza appears with Alef, but a sequence of Unicode characters when the abstract character appears with other letters.

Pros: Future text in Unicode will only have one recommended way to represent the abstract entity. Old data does not need to be converted.

Cons: Text processing in Kashmiri would require a few extra steps to detect and possibly correct the sequence. Users who ignore the recommendation will cause text processing problems. (Considering that Kashmiri users will have the combining character available on their keyboards anyway, this would happen frequently.)

B') Recommend against using the sequence <0627, 06xx>, but only in Kashmiri text. This is even weaker than B, and will have less implications on general purpose text rendering processes.

Additional Pro to B: There would be less requirements for general purpose processes.

Additional Con to B: Processes that don't discriminate based on language will not be recommended to do anything, resulting in everybody except specialized Kashmiri software ignore the recommendation.

C) Do not encode a combining ARABIC WAVY HAMZA BELOW. Instead, encode a set of precomposed characters that correspond to a sequence of two abstract characters, for example: ARABIC LETTER BEH WITH WAVY HAMZA BELOW, ARABIC LETTER JEEM WITH WAVY HAMZA BELOW, etc.

This would need about 35 new precomposed characters to be encoded.

Pros: There is no need to recommend against the usage of any character or sequence or change any algorithm.

Cons: Several new characters need to be encoded. Kashmiri information processing would become quite complex, requiring special input methods, special decomposition for linguistic processing, etc.

D) Expand the processes specified in the standard for canonical decomposition (in section 3.7), to add a precomposition level, to be applied before applying the canonical mappings. In this suggested precomposition process, the sequence <0627, 06xx> will be mapped to <0673> before any other processing is done.

This option would also open a window of opportunity to encode various combining dots, an act that would reduce future Arabic letter proposals significantly.

Pros: Stability promises of Unicode will be kept. No character or sequence would be deprecated or recommended against. Unicode can switch to a generative Arabic letter model.

Cons: Normalization and canonical equivalency algorithms need to be changed and will become more complex.

3. Unicode Character Properties

0620;ARABIC LETTER KASHMIRI YEH;Lo;0;AL;;;;;N;;;;;

06XX;ARABIC WAVY HAMZA BELOW;Mn;220;NSM;;;;;N;;;;;

4. Joining type and group for ArabicShaping.txt

0620; YEH WITH RING; D; YEH

5. Bibliography

Koul, Omkar N. *An Intensive Course in Kashmiri*. CIIL Intensive Course Series, 7. Mysore: Central Institute of Indian Language. 1985.

Library of Congress Romanization Table: <http://www.loc.gov/catdir/cpsd/romanization/kashmiri.pdf>

Munnawar, Naji, and Shafi Shauq. *Kaeshur Grammar*. Kulgam: Bazme Adab, Kapren, 1973

Perso-Arabic Script Code for Information Interchange (PASCII). http://parc.cdac.in/PASCII_V10.pdf

Acknowledgements

Shakeel Ahmed (Assistant Professor in the Department of Kashmir Studies, Oriental College, University of the Punjab, Lahore, Pakistan) was consulted on this proposal, as well as Prof. Omkar Koul. The Universal Scripts Project, with support from the National Endowment of the Humanities, also contributed to this proposal.