L2/09-215

TO: UTC and ISO/IEC JTC1/SC2 WG2
TITLE: Proposal to add two Kashmiri characters
SOURCE: Deborah Anderson (SEI, UC Berkeley), Roozbeh Pournader, Muzaffar Aazim, and Kamal Mansour
DATE: 14 May 2009

This is a request to add two characters to the Arabic block in order to fully represent the Kashmiri language in the Arabic script. The proposal draws on an earlier document, L2/09-176.

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 CE and is used currently by the people of Kashmir and the federal and state agencies of India, including the department of education of the state of Jammu and Kashmir. (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)

Two Proposed characters

a. 0620 ARABIC LETTER KASHMIRI YEH

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

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

	Isolated	Final	Medial	Initial
Proposed shapes	ؠ	ؠ	ؠ	ؠ
Variant shapes, usually used in Nastaliq style	ؽ	ؽ	ؠ	ؠ

The form with the ring below is included with its own codepoint in the Indian standard PASCII at position 187:

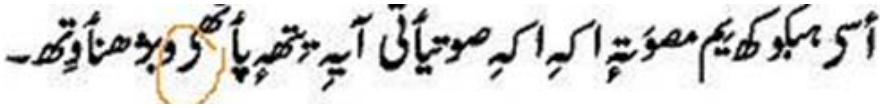
187	ۍ	LETTER YE (CIRCLE BELOW) • Kashmiri

The proposed location for this character is 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 in final and isolated forms without any ring and with a very different shape.

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

Example of the “half yeh” form of KASHMIRI YEH from *Ilm-0-Adab*, the official Magazine of Kashmiri Department of the Kashmir University (p. 220):



b. U+065F 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 also contained in the Indian standard PASCII at position 197:

197	◌ٜ	Diacritic Mark (Hamza Below) Kashmiri

Proposed glyph and name:

◌ٜ 065F ARABIC WAVY HAMZA BELOW

The proposed location for this character is 065F.

Example from *Ilm-0-Adab*, the official Magazine of Kashmiri Department of the Kashmir University (p. 220):

Normalization Issues
WAVY HAMZA BELOW raises normalization questions, because 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).

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* could be represented in two ways: <0627, 065F> 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 that would be contrary to the Unicode stability policies, which states “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.

In order to handle normalization issues, we ask U+0673 ARABIC LETTER ALEF WITH WAVY HAMZA BELOW be deprecated, and recommend that everyone use the sequence <0627, 065F> in the future.

The advantage of this approach is that future text encoded in Unicode/UCS will only have one recommended way to represent the abstract entity. However, the precomposed character may have been used in existing data (including data in other languages), which needs to be converted. Users of the standards who ignore the deprecation may cause text processing problems.

2. Unicode Character Properties
0620;ARABIC LETTER KASHMIRI YEH;Lo;0;AL;;;;;N;;;;;
065F;ARABIC WAVY HAMZA BELOW;Mn;220;NSM;;;;;N;;;;;

3. Joining type and group for ArabicShaping.txt
0620; YEH WITH RING; D; YEH

4. 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/cpso/romanization/kashmiri.pdf

Munnawar, Naji, and Shafi Shauq. *Kaeshur Gramar*. 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.

ISO/IEC JTC 1/SC 2/WG 2
PROPOSAL SUMMARY FORM TO ACCOMPANY SUBMISSIONS
FOR ADDITIONS TO THE REPERTOIRE OF ISO/IEC 10646[1]
Please fill all the sections A, B and C below.
Please read Principles and Procedures Document (P & P) from http://www.dkuug.dk/JTC1/SC2/WG2/docs/principles.html **for guidelines and details before filling this form.**
Please ensure you are using the latest Form from http://www.dkuug.dk/JTC1/SC2/WG2/docs/summaryform.html**.**
See also http://www.dkuug.dk/JTC1/SC2/WG2/docs/roadmaps.html **for latest *Roadmaps*.**

A. Administrative

1. **Title:** Proposal for two Kashmiri characters
2. Requester's name: *Muzaffar Aazim, Kamal Mansour, Roozbeh Pournader, and Deborah Anderson (SEI)*
3. Requester type (Member body/Liaison/Individual contribution): *Liaison contribution*
4. Submission date: *14 May 2009*
5. Requester's reference (if applicable):
6. Choose one of the following:
   This is a complete proposal: *Yes*
   (or) More information will be provided later:

B. Technical – General

1. Choose one of the following:
   a. This proposal is for a new script (set of characters): *No*
      Proposed name of script:
   b. The proposal is for addition of character(s) to an existing block: *Yes*
      Name of the existing block: *Arabic*
2. Number of characters in proposal: *2*
3. Proposed category (select one from below - see section 2.2 of P&P document):
   A-Contemporary X B.1-Specialized (small collection) B.2-Specialized (large collection)
   C-Major extinct D-Attested extinct E-Minor extinct
   F-Archaic Hieroglyphic or Ideographic G-Obscure or questionable usage symbols
4. Is a repertoire including character names provided?
   a. If YES, are the names in accordance with the "character naming guidelines" in Annex L of P&P document? *Yes*
   b. Are the character shapes attached in a legible form suitable for review? *Yes*
5. Who will provide the appropriate computerized font (ordered preference: True Type, or PostScript format) for publishing the standard? *Michael Everson or Kamal Mansour*
   If available now, identify source(s) for the font (include address, e-mail, ftp-site, etc.) and indicate the tools used: *Fontographer for font by Everson*
6. References:
   a. Are references (to other character sets, dictionaries, descriptive texts etc.) provided? *Yes*
   b. Are published examples of use (such as samples from newspapers, magazines, or other sources) of proposed characters attached? *Yes*
7. Special encoding issues:
   Does the proposal address other aspects of character data processing (if applicable) such as input, presentation, sorting, searching, indexing, transliteration etc. (if yes please enclose information)? *Yes*

8. Additional Information:

Submitters are invited to provide any additional information about Properties of the proposed Character(s) or Script that will assist in correct understanding of and correct linguistic processing of the proposed character(s) or script. Examples of such properties are: Casing information, Numeric information, Currency information, Display behaviour information such as line breaks, widths etc., Combining behaviour, Spacing behaviour, Directional behaviour, Default Collation behaviour, relevance in Mark Up contexts, Compatibility equivalence and other Unicode normalization related information. See the Unicode standard at http://www.unicode.org for such information on other scripts. Also see http://www.unicode.org/Public/UNIDATA/UCD.html and associated Unicode Technical Reports for information needed for consideration by the Unicode Technical Committee for inclusion in the Unicode Standard.

[1] Form number: N3152-F (Original 1994-10-14; Revised 1995-01, 1995-04, 1996-04, 1996-08, 1999-03, 2001-05, 2001-09, 2003-11, 2005-01, 2005-09, 2005-10, 2007-03, 2008-05)

C. Technical - Justification

1. Has this proposal for addition of character(s) been submitted before? *No*
   If YES explain
2. Has contact been made to members of the user community (for example: National Body, user groups of the script or characters, other experts, etc.)? *Yes*
   If YES, with whom? *Shakeel Ahmed (Univ of the Punjab, Lahore, Pakistan) and Prof. Omkar Koul*
   If YES, available relevant documents:
3. Information on the user community for the proposed characters (for example: size, demographics, information technology use, or publishing use) is included? *See next line*
   Reference: *4.5million users of Kashmiri language (source:Ethnologue 14)*
4. The context of use for the proposed characters (type of use; common or rare) *Common*
   Reference:
5. Are the proposed characters in current use by the user community? *Yes*
   If YES, where? Reference: *See references in document*
6. After giving due considerations to the principles in the P&P document must the proposed characters be entirely in the BMP? *Yes*
   If YES, is a rationale provided? *Yes*
   If YES, reference: *Used for a widely used language (see above, 3)*
7. Should the proposed characters be kept together in a contiguous range (rather than being scattered)? *No*
8. Can any of the proposed characters be considered a presentation form of an existing character or character sequence? *No*
   If YES, is a rationale for its inclusion provided?
   If YES, reference:
9. Can any of the proposed characters be encoded using a composed character sequence of either existing characters or other proposed characters? *No*
   If YES, is a rationale for its inclusion provided?
   If YES, reference:
10. Can any of the proposed character(s) be considered to be similar (in appearance or function) to an existing character? *No*
    If YES, is a rationale for its inclusion provided?
    If YES, reference:
11. Does the proposal include use of combining characters and/or use of composite sequences? *Yes*
    If YES, is a rationale for such use provided? *Yes*
    If YES, reference: *See proposal*
    Is a list of composite sequences and their corresponding glyph images (graphic symbols) provided?
    If YES, reference:
12. Does the proposal contain characters with any special properties such as control function or similar semantics? *No*
    If YES, describe in detail (include attachment if necessary)
13. Does the proposal contain any Ideographic compatibility character(s)? *No*
    If YES, is the equivalent corresponding unified ideographic character(s) identified?
    If YES, reference: