Comments on R. K. Joshi's documents L2/07-386 and L2/07-388

Peter M. Scharf
Brown University
18 October 2007

Development

Scharf and Everson's N3366, submitted to the UTC 17 October 2007 in San Jose, California, is a revision of N3290 which itself was a revision of N3235. N3235 was submitted to the WG2 for consideration at its meeting 23-27 April in Frankfurt, Germany. N3290 was submitted to the UTC for consideration at its meeting 6-10 August 2007 in Redmond, Washington. 7-8 August 2007 Scharf drafted the documents L2/07-262 A history of the development of Scharf and Everson (eds.) Vedic Unicode Proposal, and L2/07-271 Comparison of proposed characters in Lata 2006 (L2/06-185) with Scharf and Everson WG2/n3290 (L2/07-230). R. K. Joshi drafted two documents submitted to the UTC 13 October 2006. L2/07-388 comments on Scharf and Everson N3235, and L2/07-386 comments on Scharf and Everson L2/07-271. The current document comments on Joshi's last two documents for the purpose of determining courses of action to encode Vedic.

Comments on L2/07-388 Joshi 11 July 2006 "Following observations have been made with reference to the document No. L2/07-095 dated 2007-04-13 WG2 N3235."

Observation 1

N3366 notes 5 characters that have been encoded. Joshi's observation on the "sixth" perhaps is meant to address "VEDIC TONE RIGVEDIC KASHMIRI INDEPENDENT SVARITA" proposed under number 3 in N3366 with additinal evidence that confirms and clarifies its depiction.

Concerning KANNADA SIGN JIHVAMULIYA, the two different images of juhvāmūlīya in N3366 figure 2C show two glyph variants of a single character.

No further comment is required.

Observation 2

Joshi's observation notes the convergence between n3235 on the one hand and L2/03-066 and L2/06-185 regarding 35 characters proposed. These convergences are pointed out in detail by Scharf in L2/07-271 on which Joshi comments in L2/07-388 and on which I will comment below.

Let it be observed that it is not the intention of Scharf and Everson to usurp credit for proposing certain characters. Rather it is to submit an actionable proposal to the WG2 and UTC to allow Vedic characters to be included in the Unicode Standard. We openly welcome collaboration with Lata, Joshi, and others to achieve this end. Our efforts to promote the broadest possible collaboration and participation of all interested parties and the scholarly community to this end are documented in L2/07-262.

Observation 3

The proposal to encode pr̥ṣṭhamātrā has been removed, accepting the recommendation of Ken Whistler that it be handled as an alternate rendering. A note on how font designers ought to implement pr̥ṣṭhamātrā notation has been added to N3366, section 2.1.

Observation 4

The proposed DEVANAGARI VOWEL SIGN CANDRA LONG E is a combining mark that can go as a superscript over the Devanagari e to indicate initial long E; Since initial long E can be encoded as a sequence of Devanagari e + candra long E, there is no need to encode a stand-alone initial long e. Michael Witzel, an expert in Avestan and Vedic has confirmed the use of Candra E for schwa. Many consonant characters and nukta have been added to the Devanagari block in order to represent perso-arabic consonants. The proposal of others does not conflict with the present proposal should evidence warrant it. The same is pertinent to the transcription into Devanagari of Kashmiri vowels.

Observation 5

The Telugu table is indeed missing dependent vowel signs vocalic l and vocalic long l. If not present in an amendment, they should be included in a subsequent proposal.

Joshi concurs with the WG2's acceptance of N3235 item 10 into amendment 5, and suggests the possibility of the necessity for additions to accommodate peculiar nasalizations and seminasalizations in Bengali, Malayalam, and Konkani in Devanagari anusvaras and ardha anusvaras. It is hoped that Joshi will include these items in a subsequent proposal.

Observation 11 Symbols

The svastikas were removed from N3235 and placed in a separate proposal in May. The proposal N3268 was accepted into amendment 6 at the WG2 meeting 21 September in Hangzhou, China.

The symbol om is included in the Devanagari code page as 0950. Joshi suggests the addition of a symbol for the word siddham. It is hoped that he will include this in a subsequent proposal.

The symbol flower 0974 in N3235 was removed from the revised proposal N3290 because a flower symbol is included as U+2055 in the General Punctuation page [2000-206F] of the Unicode Standard.

Joshi suggests that there are many decorative flower symbols for which a separate area should be defined. If his subsequent proposal evidences serveral such and WG2 or UTC deem it advisable to move the proposed DEVANAGARI SIGN PUSHPIKA, now 0973 in N3366, to such a new area, the authors of N3366 have no objection.

Observation 7 Precomposed diacritics

The encoding of precomposed diacritics, which Joshi points out is prevelant in Latin and suggests be adopted similarly for Devanagari, has been deprecated. If the characters can be reasonably and economically formed by a sequence, the sequence is to be preferred over the precomposed character.

Observation 8 Visarga and anusvara

N3366 section 7 presents motivations and evidence that support a separate encoding of the accent diacritics attached to visarga. These accents were often added by a separate hand in different colored ink. Encoding them separately will permit accurate representation of this different coloration which would otherwise be impossible.

Joshi argues for precomposed characters in the case of N3366 1CE5, 1CE6, and 1CE7 combined with bindu, candrabindu, and virāma just as N3366 A8F3, A8F4, A8F5, A8F6, and A8F7 are proposed as combined characters. [In N3235 the characters in question were proposed to be A4E5, A4E6, and A4E7, and A8F5, A8F6, A8F7, A8F8, and A8F9.] Conversely, the editors of N3366 deem it in accord with the current policy of favoring sequences over precomposed characters to attempt to encode as a sequence when reasonable and economical. A4E5, A4E6, and A4E7 combine easily with combining bindu, candrabindu, and virāma to generate the range of characters Joshi proposes. Rendering the components A8F6, A8F6, A8F7, A8F8, and A8F9 by sequences would likewise be neater.

In deliberation over how to encode these characters Scharf suggested the following sequences be used to encode the five characters in section 8 Nasals.

DEVANAGARI SIGN CANDRABINDU VIRAMA will be produced by the sequence DEVANAGARI SIGN SPACING CANDRABINDU [A8F2] + DEVANAGARI SIGN VIRAMA [094D].

DEVANAGARI SIGN DOUBLE CANDRABINDU VIRAMA will be produced by the sequence DEVANAGARI SIGN SPACING CANDRABINDU [A8F2] + DEVANAGARI SIGN SPACING CANDRABINDU [A8F2] + DEVANAGARI SIGN VIRAMA [094D].

(Alternatively, if DEVANAGARI SIGN CANDRABINDU VIRAMA were accepted as precomposed, DEVANAGARI SIGN DOUBLE CANDRABINDU VIRAMA will be produced by the sequence DEVANAGARI SIGN SPACING CANDRABINDU [A8F2] + DEVANAGARI SIGN CANDRABINDU VIRAMA [A8F3].)

DEVANAGARI SIGN CANDRABINDU TWO will be produced by the sequence DEVANAGARI SIGN SPACING CANDRABINDU [A8F2] + DEVANAGARI DIGIT 2 [0968].

DEVANAGARI SIGN CANDRABINDU THREE will be produced by the sequence DEVANAGARI SIGN SPACING CANDRABINDU [A8F2] + DEVANAGARI DIGIT 3 [0969].

DEVANAGARI SIGN CANDRABINDU AVAGRAHA will be produced by the sequence DEVANAGARI SIGN SPACING CANDRABINDU [A8F2] + DEVANAGARI SIGN AVAGRAHA [093D]."

However, Everson is of the opinion that this is not feasible or advisable. N3366, pp. 8-9, section 8 and its Note argue that special circumstances prevent facile decomposition of the proposed characters in these cases. If it should prove feasible to overcome the special circumstances mentioned there, the editors would not maintain any objection to the rendering of these by sequences as well.

Joshi's suggestion to encode a character for each Devanāgarī consonant sign with virāma attached would have merit in an encoding for the Sanskrit language that departs from the principals hitherto adopted by ISCII and Unicode for Devanagari and other Indic scripts. In the current ISCII and Unicode, virāma followed by another consonant sign is used to indicate that the first consonant sign combines with the subsequent sign. An alternative scheme, which Joshi has described in an unpublished paper draft seen by the current author and has utilized in software development, would not encode consonant signs without virāma at all. In this way the encoding of Devanāgarī would closely model Sanskrit phonology. The virāma would not be needed to indicate conjunction of consonants; consonants in sequence (already marked with virāma) would automatically be assumed to conjoin. To indicate the presence of the vowel *a*, an *a*-vowel character would have to follow the consonant sign. In this scheme there would be no need for two sets of vowel vowel characters: initial (i.e. DEVANAGARI LETTER AA, etc.) and dependent (DEVANAGARI VOWEL SIGN AA, etc.). One set of vowel characters would suffice. A phonetic encoding scheme such as this is desirable for Sanskrit. Similar phonetic encodings for Sanskrit have been devised and utilized in software by others such as Amba Kulkarni, the present author, and others. It is recommended that parties that have a stake in such a phonetic encoding for Sanskrit devise a commonly agreed upon character set and submit it to the ISO and to Unicode.

Comments on L2/07-386 Joshi 13 October 2007 "Comments on Comparison of proposed characters in Lata 2006 (L2/06-185) with Scharf and Everson WG2/n3290 (L2/07-230)"

The following characters are recognized by Joshi (L2/06-185 and L2/07-386) and Scharf (n3366) as identical:

Scharf	N3366 page	Joshi	L2/07-386 page
097A	30	08E7	7
1CD1	32	08E1	6
1CD2	32	08B1	4
1CD3	32	08B4	4
1CD4	32	08B2	4
1CD5	32	08B3	4
1CD6	32	08B7	4
1CD7	32	08B6	4
1CD8	32	08B8	4
1CDA	32	08B5	4
1CDB	32	08B0	4
1CDD	32	08EB	7 (miscategorized and misnamed), and part of 08AD
1CE8	32	0896	2, with glyph variants at 088F, 0890, 0891, 0899
A8E1	33	08C7	5
A8E2	33	08C8	5
A8E3	33	08C9	5

A8E4	33	08CA	5
A8E5	33	08CB	5
A8E6	33	08CC	5
A8E7	33	08CD	5

A8EA	33	08CE	6
A8EB	33	08DD	6
A8EC	33	08DF	6
A8ED	33	08CF	6
A8EF	33	08D1	6

A8F1	33	08E2	6
A8F4	33	0889	1
A8F6	33	088A	1

The following characters proposed by Scharf have at least one glyph variant recognized by Joshi

1CE4	32	089B, 08A9, 08AA, 08AB

The following characters proposed by Scharf constitute analyzed portions of precomposed characters proposed by Joshi

1CDD	32	08AD	
1CDF	32	08BF	5
1CE1	32	089D	
1CE2	32	089E, 08A0, 08A2, 08A3	
1CE3	32	089F, 08A1, 08A2, 08A4	
1CE5	32	088B, 088C	
1CE6	32	0893, 0894, 0895	
1CE7	32	088D, 088E	

In sec. 9, DEVANAGARI VOWEL SIGN CANDRA LONG E

Although the CANDRA LONG E could be composed of the sequence of U+ 0304 COMBINING MACRON and U+ 0306 COMBINING BREVE, it would be very strange to have a Devanagari letter followed by a Generic European macron followed by a Devanagari candra.

Acknowledgements

This document was drafted as part of the International Digital Sanskrit Library Integration Project funded by a grant to Brown University from the U.S. National Science Foundation (grant no. 0535207). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.