L2/02-402

COMMENTS ON SOME PROPOSALS FOR REVISION OF DEVANAGARI ENCODING IN UNICODE

by
Peri Bhaskararao
Professor
Research Institute for Languages & Cultures of Asia & Africa
Tokyo University of Foreign Studies

I. Background: The characters that are newly proposed for inclusion in the revision can be grouped under nine sets as below:

1. Four characters for Sindhii language to be written in Devanaagarii script: 0979 ग़ 097A ज़ 097B ड़ 097C ब़
2. Three characters for Devanaagarii script: 0972 क्ष 0973 ज्ञ 0974 श्र
3. One character for the 'Soft' RA used in MaraaThii writing: 0978 ऱ
4. Two characters for a specific way of writing in MaraaThii: 0976 श॒ 0977 ल॒
5. One character for the abbreviation of RUPAYAA in Devanaagarii: 0971 ऱु
6. One character for the Devanaagarii Repha sign: 0975 ◌ॎ
7. Two characters representing two types of Anuswaara in Yajurveda: Given at 0955 and 0956
8. One character representing the Jihvaa-muuliiya sign: Given at 0957
9. One character to represent the 'invisible letter' of Devanaagarii: Given at 093A

II. Discussion:

II.1. Definitions:

For a scientific discussion on linguistic and orthographical issues of a script, the following points need to be reiterated.

A. **Script and Orthography:** Script needs to be differentiated from Orthography. Devanaagarii script is used by Hindi, Marathi, Sanskrit, Nepali, Newari, and Konkani etc. Consequently, we have Hindi orthography, Nepali orthography etc. Orthography does two things to the letters of the script: it utilizes some or all of the letters from the script and it assigns phonemic value ranges to letters from the script. All the orthographies that use a script will be utilizing a maximum number of the letters from that script. However, each of them might not use a few of the letters available in the script. For instance, the Devanaagarii letter ळ (U+0933) is used in the orthographies of Marathi and Sanskrit but not in the orthography of Hindi.

B. **Graphemes and allographs:** In Indic scripts, there is a distinction between a grapheme and its allographs. Let us represent graphemes enclosed in { } and allographs in < >.

   a. **Context-sensitive Allographs:** These are variants of a single grapheme in the orthography of a particular language. For instance, the grapheme {r} (at U+0930) has the following allographs (and more): The primary allograph < र > and the secondary allographs <◌̲> < ◌́ > <◌ॎ> (and a few more secondary allographs). Some of the distributive patterns of these allographs are:

      1. <◌̲> is the realization of the grapheme when it is preceded by the consonant graphemes ट ठ ड ढ (all with a bottom meniscus) (e.g. {Tra} = ट्र)
      2. < ◌́ > is the realization of the grapheme when it is preceded by a consonant grapheme other than ट ठ ड ढ and र (e.g. {kra} = क्र)
      3. <◌ॎ> is the realization of the grapheme when it is followed by a consonant (e.g. {rka} = र्क)
      4. The primary allograph < र > occurs in a context not covered by the secondary allographs (e.g. {rii} = री {rra} = र्र).

   b. **Orthography-dependent Allographs:** Besides context-sensitive allographs, we get language-dependent allographs. For instance झ of U+091D has the allograph झ़ that was used in earlier

printing of Devanaagarii in India but still is currently used in Nepali orthography. Another set of language-dependent allographs are U+090B ऋ in Hindi orthography and its variant ऋ in Nepali orthography .

c. **Allographs inside Grapheme-compounds:** Two allographs of two different graphemes might be fused to form a grapheme-compound. The shapes of the allographs of the constituent graphemes cannot be clearly demarcated – hence it is a compound (parallel to compounds and mixtures in chemistry). क्ष is a grapheme-compound. Its constituents, the allographs of { क } and { ष } cannot be visually demarcated inside this grapheme-compound.

Note that the above listing of allographs and the contexts of their occurrence are illustrative but not exhaustive.

II.2. Comments: With the above points in mind, let us look at the 9 revision proposals.

1. The four new characters for Sindhi orthography in Devanaagarii script (0979 ग़ 097A ज़ 097B ड़ 097C ब़) are valid candidates for inclusion. Sindhi is the only Indian language (and also Multani, when it needs to be encoded for Devanaagarii), which has a set of four voiced implosives (glottalic ingressive sounds) that are phonemic and are written distinct from the regular non-implosive stops. Hence these four new characters need to be encoded. We cannot treat them as underlined counterparts of some other characters. Though their shapes can be derived by underlining, they have a distinct phonemic value as well as graphemic value and thus require independent status.

2.The three characters (0972 क्ष, 0973 ज्ञ , 0974 श्र) are grapheme-compounds. All of them are derived from underlying sequences of graphemes. Confusion usually arises about their status because they are separately listed at the end of the conventional Devanaagarii alphabet (varNamaalaa). Though they are grapheme-compounds in Devanaagarii, their counterparts are not grapheme-compounds in several other Indic scripts. For instance, in Telugu, the components are distinctly visible even after the clusters are formed (note that they are grapheme-clusters in Telugu, not grapheme-compounds). Observe the following comparative chart:

	Devanaagarii	Telugu
KA+SSA	क + ष > क्ष	క + ్ష > క్ష
TA+RA	त + र > त्र	త + ్ర > త్ర
JA+NYA	ज + ञ > ज्ञ	జ + ్ఞ > జ్ఞ

The fact that these are clusters of consonants was recognized by the ancient grammarians and in any scientific lexicographical work on languages that use Devanaagarii script, words containing these three grapheme-compounds are sorted under the respective underlying first consonants (e.g., क्ष is sorted under { क } as a sequence of { क } followed by { ष }. For these reasons, they need not be assigned independent character-status.

3. The proposal under discussion asks for giving a character-status to 'Soft' RA used in MaraaThii writing: 0978 ॅ. This character was included in ISCII1991 (D0 208). However, there it was given the name 'Consonant Hard RA (Southern Scripts)'. Later in the same document (on page 12 under a note), its Halanta-equivalent was described as 'the half र form ॅ as in वाऱ्या. This is an example of its usage in Marathi orthography. It should be emphasized that ॅ is NOT a context-sensitive allograph of { र } as its occurrence is not predictable. There are contrastive sets of words from Marathi (आचार्यास 'to the teacher' आचाऱ्यास 'to the cook' – an example from Naik, B.S. *Typography of Devanagari-1.* Bombay: Directorate of Languages 1971. 186pp.). Another point about this special 'soft' Ra seemed to have escaped the notice of the proposers. It is used distinctively in Nepali and Newari orthographies. In these orthographies too, it cannot be considered as an allograph of { र }. Hence, it should be given a character-status.

4. The proposal to give character-status to ऌ ॡ (the so-called Bombay letters) is not supportable. Both these are Orthography-dependent allographs. In fact, they are used not only in Marathi orthography of Devanaagarii but also in the South-Indian variant of Devanaagarii usage. By encoding them differently, we will get two different strings for

the same underlying word across these languages (e.g., a Sanskrit word used in both Hindi and Marathi). They are orthography-dependent variants of ल and श. They need not be given character-status.

5. The proposal to have a separate character for RUPAYAA in Devanaagarii: 0971 ॱ is quite valid. However, there are some other abbreviatory symbols used in Devanaagarii (for instance the abbreviations for 'paisa', kilomiiTar, kilograam). All these abbreviations should be seen together as one set and a consolidated proposal needs to be made to include them. We have to look at other Indic scripts that have similar abbreviatory notations. Hence, this proposal may be considered along with a more comprehensive proposal with all the other abbreviatory symbols in Devanaagarii.

6. The proposal to have Devanaagarii Repha sign: 0975 ◌ॎ is not clearly explained. Besides the Repha, there are some more context-sensitive allographs of the grapheme र (U+0930). All of them are derivable by rule and represent only one character. Hence, the Repha sign need not be given a character-status unless a clear justification is forthcoming.

7 & 8. The proposal to include two characters representing two types of Anuswaara in Yajurveda given at 0955 and 0956 and one character representing the Jihvaa-muuliiya sign given at 0957 is valid. However, similar to the Rupayaa sign explained above, it is not just three signs that are there especially in Vedic. There are many more signs that have to be included for Vedic. A comprehensive proposal to include all the Vedic signs should be put forward. Till then, inclusion of these three signs may be deferred.

9. The proposal to include a character to represent the 'invisible letter' of Devanaagarii given at 093A is also debatable. Prima facie it seems whatever is to be achieved by this 'invisible letter' can be achieved by a 'grapheme joiner' that is already encoded. A detailed justification for encoding the 'invisible letter' is necessary.

☼

Colophon: The Indic scripts are fortunate enough in having a foundation of a rich grammatical tradition of the three sages (muni-traya) – PaaNini, Patañjali and Kaatyaayana (and their followers). Professor J.R.Firth, the famous British linguist tells about Sir William Jones: 'Without the Indian grammarians and phoneticians whom he [=William Jones] introduced and recommended to us, it is difficult to imagine our nineteenth century school of phonetics' (*Transactions of the Philological Society* 1946: pp.92ff.). One of the important dictums of SaMskRita (=Sanskrit) grammatical tradition is 'brevity' (called '*laaghava*' as opposed to complexity '*gaurava*' which is detested). Indian grammarians are trained to look for atomistic and minimal elements that make up larger parts of the language. Strict encoding principles will also adhere to such minimalist approach to scripts.

☼