Proposal to encode Devanagari Sign High Spacing Dot

Jonathan Kew, Steve Smith
SIL International
April 20, 2006

1. Introduction

In several language communities of Nepal, the Devanagari script has been adapted to represent additional phonological features not found in major languages such as Hindi, Marathi, or Nepali, or historically in Sanskrit. One such adaptation is the use of a modifier dot under letters, including vowels, where it is not traditionally used. This can be represented in Unicode/ISO10646 using the existing character U+093C DEVANAGARI SIGN NUKTA, provided fonts and rendering engines support the productive use of this mark; there is no fundamental character encoding problem here.

Another form of script modification, however, seen in several languages, is the use of a dot (similar in design to the NUKTA or ANUSVARA dots, often diamond-shaped in typical fonts) appearing as a spacing character at or very slightly above the level of the connecting bar across the top of Devanagari letters. Although this dot shares the same basic glyph shape as both U+0902 DEVANAGARI SIGN ANUSVARA and U+093C DEVANAGARI SIGN NUKTA, it is clearly distinct in both positioning (at the “hanging baseline” of the text, not either above or below other letters) and behavior (it is not a combining mark but a spacing character, seen word-initially as well as between other letters).

Such a character is known to have been used in orthographies of at least three different languages: Yohlmo (also known as Helambu Sherpa, http://www.ethnologue.com/show_language.asp?code=scp), where it indicates a high falling tone on the following suffix; Lhomi (http://www.ethnologue.com/show_language.asp?code=lhm), where it is written word-initially to distinguish words with ‘tense’ or ‘clear’ vowels from those with ‘lax’ vowels; and Takale Kham (Western Parbate, http://www.ethnologue.com/show_language.asp?code=kjl), to indicate high tone on breathy vowels. As these are small language communities with limited literacy as yet, it is possible that some conventions may change over time, but in each case there are existing publications and readers using this mark.

2. Proposed character

To support the character encoding requirements of these extended Devanagari writing systems, the following character is proposed. The representative glyph is shown between two typical Devanagari consonants to make its relative size and positioning clear:

\[
\text{पंय}
\]

0971;DEVANAGARI SIGN HIGH SPACING DOT;Lm;0;L;;;;;N;;;;;

The codepoint may of course be changed to a different position in the Devanagari block (U+0900 might be another reasonable possibility). The proposed character is named using SIGN rather than LETTER as it is not regarded as a full-fledged letter of the alphabet, but rather a sign that indicates a modification of the syllable or word. Other properties are the same as for typical Devanagari consonants, or the analogous spacing sign U+093D DEVANAGARI SIGN AVAGRAHA, except that a General Category of Lm seems more appropriate than Lo to the known usage of this character.

The linebreak class of the new character should be AL, as it is treated just like a Devanagari letter for line-break purposes.

We have seen little evidence relating to collation, but the one source available [3] treats the HIGH SPACING DOT as ignorable at the primary level. No minimal pairs that would have forced the compilers to make a clear decision regarding secondary or tertiary collation weight have been observed.

Regarding rendering behavior, this character is always used at the beginning of an orthographic syllable or cluster. Its presence in the text explicitly begins a new cluster; therefore, in a sequence such as <RA, VIRAMA, DOT, KA>, the ra-virama should be rendered with a visible halant, not as reph: रँक, not रःक. The dot also remains in initial position in the presence of the short i vowel; therefore, <DOT, KA, VOWEL SIGN I> is rendered रःिक, not रःिक.
3. Examples

[1], page 513

[2], page 2

[3], page 313

[3], page 354

[3], page 363
4. References


ISO/IEC JTC 1/SC 2/WG 2

PROPOSAL SUMMARY FORM TO ACCOMPANY SUBMISSIONS FOR ADDITIONS TO THE REPERTOIRE OF ISO/IEC 10646.


### A. Administrative

1. **Title:** Proposal to encode Devanagari Sign High Spacing Dot
2. **Requester's name:** SIL International (contact: Jonathan Kew)
3. **Requester type (Member body/Liaison/Individual contribution):** Individual contribution
4. **Submission date:** 2006-04-20
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
   - b. The proposal is for addition of character(s) to an existing block: yes
   - Name of the existing block: Devanagari
2. **Number of characters in proposal:** 1
3. **Proposed category (select one from below - see section 2.2 of P&P document):**
   - A-Contemporary
   - 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. **Proposed Level of Implementation (1, 2 or 3) (see Annex K in P&P document):** 1
   - Is a rationale provided for the choice? yes
   - If Yes, reference: Simple non-combining, non-contextual character
5. **Is a repertoire including character names provided?** yes
   - 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
6. **Who will provide the appropriate computerized font (ordered preference: True Type, or PostScript format) for publishing the standard?** Jonathan Kew, SIL International
   - If available now, identify source(s) for the font (include address, e-mail, ftp-site, etc.) and indicate the tools used: Contact jonathan_kew@sil.org when required
7. **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
8. **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

9. **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.

---

## C. Technical - Justification

1. Has this proposal for addition of character(s) been submitted before?
   - Yes: 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: yes
   - If YES, with whom?
   - Linguists researching languages of Nepal
   - 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?
   - Yes: yes
   - Reference: Total population of language communities ca. 60,000 (Ethnologue), but low mother-tongue literacy

4. The context of use for the proposed characters (type of use; common or rare)
   - Yes: common
   - Reference: Used in several minority languages, although not in national languages using the script

5. Are the proposed characters in current use by the user community?
   - Yes: yes
   - Reference: Published books in the concerned languages (see bibliography)

6. After giving due considerations to the principles in the P&P document must the proposed characters be entirely in the BMP?
   - Yes: yes
   - If YES, a rationale provided?
   - If YES, reference: Keep with other Devanagari characters

7. Should the proposed characters be kept together in a contiguous range (rather than being scattered)?
   - Yes: N/A

8. Can any of the proposed characters be considered a presentation form of an existing character or character sequence?
   - Yes: no
   - If YES, 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?
   - Yes: no
   - If YES, 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?
    - Yes: no
    - If YES, a rationale for such use provided?
    - If YES, reference:

11. Does the proposal include use of combining characters and/or use of composite sequences?
    - Yes: no
    - If YES, is a rationale for such use provided?
    - If YES, reference:
    - Is a list of composite sequences and their corresponding glyph images (graphic symbols) provided?
    - If YES, reference:

12. Does the proposal contain any ideographic compatibility character(s)?
    - Yes: no
    - If YES, describe in detail (include attachment if necessary)

13. If YES, is the equivalent corresponding unified ideographic character(s) identified?
    - Yes: no
    - If YES, reference: