L2/06-137

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:

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

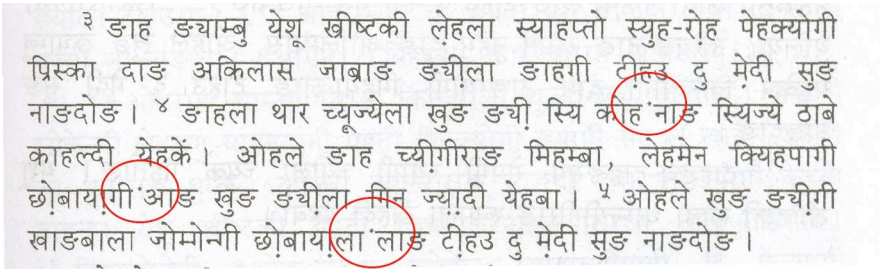
The codepoint may of course be changed to a different position in the Devanagari block (U+0900 might be another reasonable possibility). The proposed character is named using SIGN rather than LETTER as it is not regarded as a full-fledged letter of the alphabet, but rather a sign that indicates a modification 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.

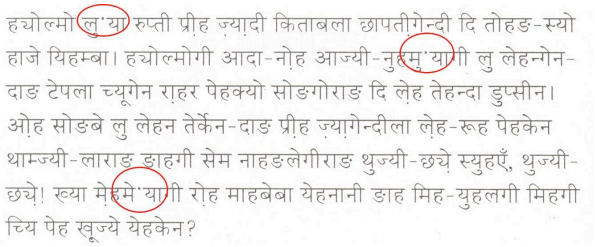
Regarding rendering behavior, this character is always used at the beginning of an orthographic syllable or cluster. Its presence in the text explicitly begins a new cluster; therefore, in a sequence such as <RA, VIRAMA, DOT, KA>, the *ra-virama* should be rendered with a visible halant, not as *reph*: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

3. Examples



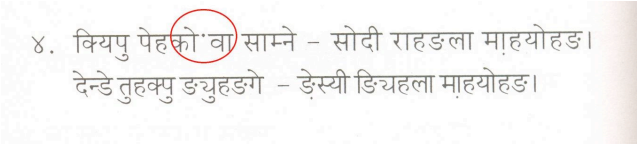
३ ङाह ङ्याम्बु येशू ख्रीष्टकी लेहला स्याहप्तो स्युह-रोह पेहक्योगी प्रिस्का दाङ अकिलास जाब्राङ ङ्यीला ङाहगी टीह्ङ दु मेदी सुङ नाङदोङ। ४ ङाहला थार च्यूज्येला खुङ ङ्यी स्यि कोह नाङ स्यिज्ये ठाबे काहल्दी येहकें। ओहले ङाह च्यीगीराङ मिहम्बा, लेहमेन क्यिहपागी छोबायगी आङ खुङ ङ्यीला मिन ज्यादी येहबा। ५ ओहले खुङ ङ्यीगी खाङबाला जोम्मोन्गी छोबायला लाङ टीहउ दु मेदी सुङ नाङदोङ।

[1], page 513



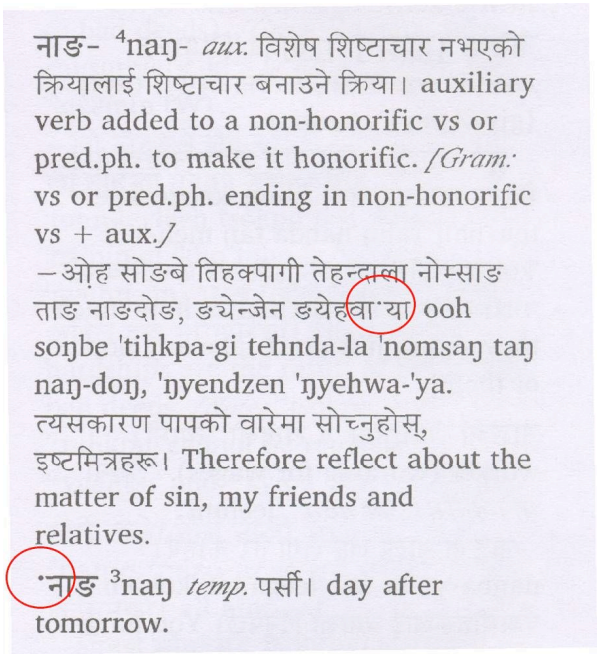
हच्योल्मो लु या रुप्ती प्रीह ज्यादी किताबला छापतीगेन्दी दि तोहङ-स्यो हाजे यिहम्बा। हच्योल्मोगी आदा-नोह आज्यी-नुसमु यागी लु लेहन्गेन-दाङ टेपला च्यूगेन राहर पेहक्यो सोङगोराङ दि लेह तेहन्दा डुप्सीन। ओह सोङबे लु लेहन तेर्केन-दाङ प्रीह ज्यागेन्दीला लेह-रूह पेहकेन थाम्ज्यी-लाराङ ङाहगी सेम नाहङलेगीराङ थुज्यी-छचे स्युहएँ, थुज्यी-छचे! रूया मेहमे यागी रोह माहबेबा येहनानी ङाह मिह-युहलगी मिहगी च्यि पेह खूज्ये येहकेन?

[2], page 2



४. क्यिपु पेहको वा साम्ने – सोदी राहङला माहयोहङ।
देन्डे तुहक्पु ङ्युहङगे – डेस्यी ङ्यिहला माहयोहङ।

[2], page 10

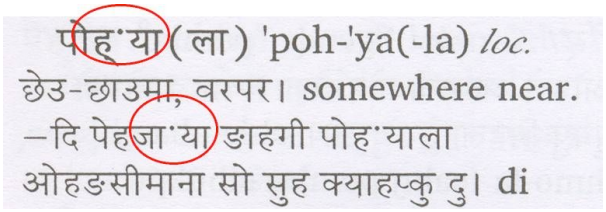


नाङ- [4]naŋ- *aux.* विशेष शिष्टाचार नभएको क्रियालाई शिष्टाचार बनाउने क्रिया। auxiliary verb added to a non-honorific vs or pred.ph. to make it honorific. *[Gram.* vs or pred.ph. ending in non-honorific vs + aux.*]*

–ओह सोङबे तिहक्पागी तेहन्दाला नोम्साङ ताङ नाङदोङ, ङ्येन्जेन ङ्येहवा या ooh soŋbe 'tihkpa-gi tehnda-la 'nomsaŋ taŋ naŋ-doŋ, 'ŋyendzen 'ŋyehwa-'ya. त्यसकारण पापको वारेमा सोच्नुहोस्, इष्टमित्रहरू। Therefore reflect about the matter of sin, my friends and relatives.

ˈनाङ [3]naŋ *temp.* पर्सी। day after tomorrow.

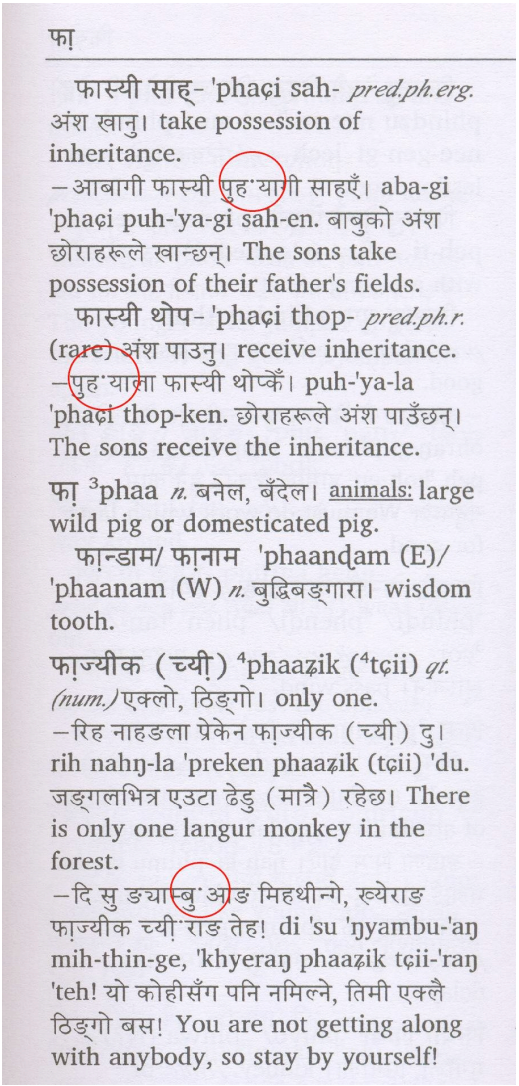
[3], page 313



पोह या (ला) 'poh-'ya(-la) *loc.* छेउ-छाउमा, वरपर। somewhere near.

–दि पेह जा या ङाहगी पोह याला ओहङसीमाना सो सुह क्याहप्कु दु। di

[3], page 354



फा

फास्यी साह- 'phaçi sah- *pred.ph.erg.* अंश खानु। take possession of inheritance.

–आबागी फास्यी पुह यागी साहएँ। aba-gi 'phaçi puh-'ya-gi sah-en. बाबुको अंश छोराहरूले खान्छन्। The sons take possession of their father's fields.

फास्यी थोप- 'phaçi thop- *pred.ph.r.* (rare) अंश पाउनु। receive inheritance.

–पुह याला फास्यी थोप्कें। puh-'ya-la 'phaçi thop-ken. छोराहरूले अंश पाउँछन्। The sons receive the inheritance.

फा [3]phaa *n.* बनेल, बँदेल। animals: large wild pig or domesticated pig.

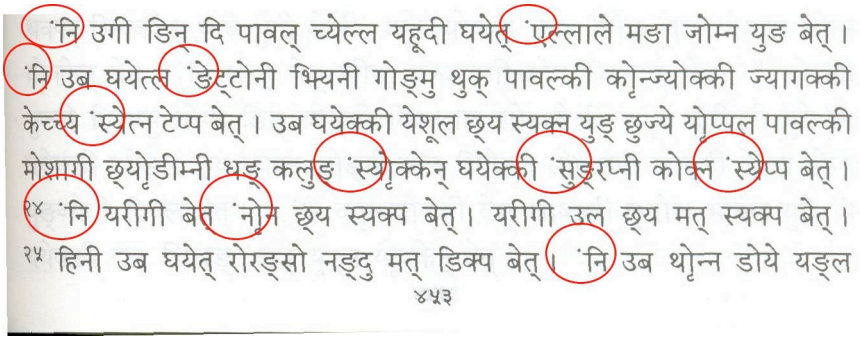
फान्डाम/ फानाम 'phaandam (E)/ 'phaanam (W) *n.* बुद्धिबङ्गारा। wisdom tooth.

फाज्यीक (च्यी) [4]phaaʑik ([4]tçii) *qt. (num.)* एक्लो, ठिङ्गो। only one.

–रिह नाहङला प्रेकेन फाज्यीक (च्यी) दु। rih nahŋ-la 'preken phaaʑik (tçii) 'du. जङ्गलभित्र एउटा ढेडु (मात्रै) रहेछ। There is only one langur monkey in the forest.

–दि सु ङ्याम्बु आङ मिहथीन्गे, ख्येराङ फाज्यीक च्यी राङ तेह! di 'su 'ŋyambu-'aŋ mih-thin-ge, 'khyeraŋ phaaʑik tçii-'raŋ 'teh! यो कोहीसँग पनि नमिल्ने, तिमी एक्लै ठिङ्गो बस! You are not getting along with anybody, so stay by yourself!

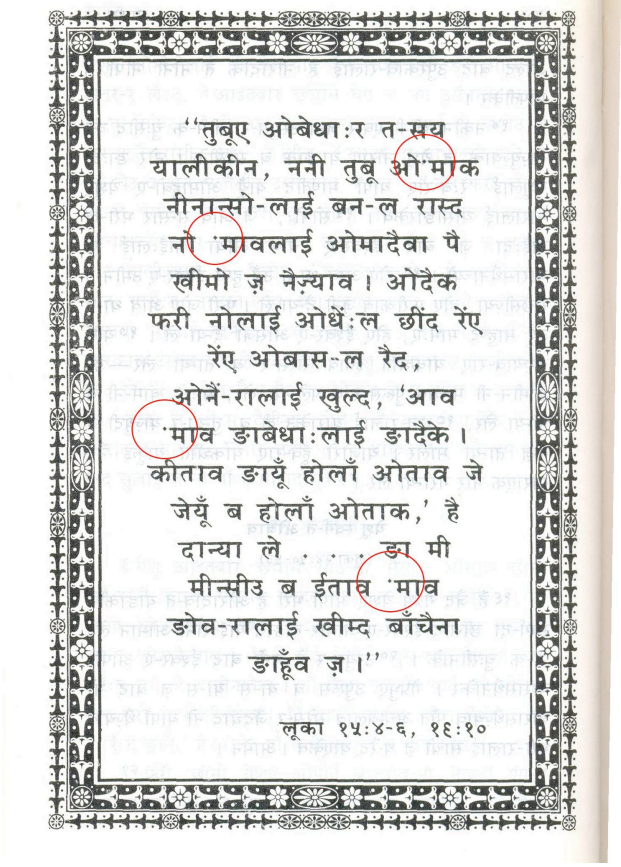
[3], page 363



ंनि उगी डिन् दि पावल् च्येल्ल यहूदी घयेत् ंएल्लाले मडा जोम्न युङ बेत् ।
ंनि उब घयेत्स ंडेट्टोनी भ्यिनी गोङ्मु थुक् पावल्की कोन्ज्योक्की ज्यागक्की
केच्च्य ंस्येत्न टेप्प बेत् । उब घयेक्की येशूल छ्य स्यक्प युङ् छुज्ये योप्प्ल पावल्की
मोशागी छ्योंडीम्नी धङ् कलुङ ंस्योंक्केन् घयेक्की ंसुङ्रप्नी कोक्न ंस्येप्प बेत् ।
२४ ंनि यरीगी बेत् ंनोंन छ्य स्यक्प बेत् । यरीगी उल छ्य मत् स्यक्प बेत् ।
२५ हिनी उब घयेत् रोरङ्सो नङ्दु मत् डिक्प बेत् । ंनि उब थोंन्न डोये यङ्ल

४५३

[4], page 453; the dot is used only in word-initial position in this language



"तुबूए ओबेधाःर त-सय
यालीकीन, ननी तुबू ओ ंमाक
नीनान्सो-लाई बन-ल रास्द
नों ंमावलाई ओमादैवा पै
खीमो ज़ नैज़्याव । ओदैक
फरी नोलाई ओधेंःल छींद रेंए
रेंए ओबास-ल रैद,
ओनैं-रालाई खुल्द, 'आव
ंमाव डाबेधाःलाई डादैके ।
कीताव डायूँ होलाँ ओताव जे
जेयूँ ब होलाँ ओताक,' है
दान्या ले . . . डा मी
मीन्सीउ ब ईताव ंमाव
डोव-रालाई खीम्द बाँचैना
डाहूँव ज़ ।"

लूका १५:४-६, १९:१०

[5], page 206

ताके । १७ है जैद नँ डालाई ननैं मीताव नरँःनाकीन नोलाई
डा डा ंरँःद रैव ताके । १८ नोए नउपुर-त मादान्य ंएन-र
ओदोव ताकीन, सोनोः काता-र हाइद उईव ताकीन, डा ज़
नँलाई ओलोइ डाईया । १९ है जैद डा पाउल-ए ज़ डाफो·या
लीद डासही छाप ब भाःद डाईके । (खाली नँलाई डाबाँचैनीउ
डागुण उलीदी ब डागुण लीज़्या है डामादींए ।) २० है जैद नँ
ईश्वर-ए ओपाँताव डाभाई, नँ डाबीन्ती-र चीउ·द स्रीसलाई
सम्जीद नो मी·लाई फरी नैद डायूँ होलाँ जैदय्याव ताके ।
२१ आव डासरो-लाव है डादींव-कीस ंकुथूः ज़ नदोया लीद
डायूँ हुबो ज़ लीज़्या ।

[5], page 827

4. References

[1] *lehangu yahbu'ya* (New Testament in Yohlmo language, Nepal). Samdan Publishers, Kathmandu, 2000.

[2] Anna Maria Hari. *yohlmo lu*. A collection of Yohlmo (Helambu Sherpa) Folksongs. 2003.

[3] Anna Maria Hari and Chhegu Lama (compilers). *yohlmo – nepali – angreji shabdkosh* (Yohlmo – Nepali – English Dictionary). Central Department of Linguistics, Tribhuvan University, Kathmandu, 2004.

[4] *sungrap samba* (New Written Word) The New Testament in Lhomi. Nepal Bible Society, Kathmandu. 1995.

[5] *iishwar-e sahro yahka-law opa* (New Testament, Kham language, Nepal). World Home Bible League. 1985.

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 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*
     - Proposed name of script:
   - 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	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. 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.

---

[1] Form number: N3002-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)

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? *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*
   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) *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
   If YES, where? 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*
   If YES, is a rationale provided? *yes*
   If YES, reference: *Keep with other Devanagari characters*
7. Should the proposed characters be kept together in a contiguous range (rather than being scattered)? *N/A*
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? *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 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: