ISO/IEC JTC1/SC2/WG2 N2934
L2/05-082
2005-03-30

Universal Multiple-Octet Coded Character Set
International Organization for Standardization
Organisation Internationale de Normalisation
Международная организация по стандартизации

Doc Type: **Working Group Document**
Title: **Proposal to add four characters for Sindhi to the BMP of the UCS**
Source: **Michael Everson**
Status: **Individual Contribution**
Action: **For consideration by JTC1/SC2/WG2 and UTC**
Date: **2005-03-30**

This document requests four additional characters to be added to the UCS and contains the proposal summary form. These letters are used in Devanagari orthography for the Sindhi language, to represent implosive consonants. The sounds have unique letters in both Devanagari and Arabic orthography for Sindhi: [ɠ] *g̈* is written ग and U+06B3 گ ARABIC LETTER GUEH, [ʃ] *j̈* is written ज and U+0684 ڄ ARABIC LETTER DYEH, [ɗ] *d̤* is written ड and U+068F ڏ ARABIC LETTER DAL WITH THREE DOTS ABOVE DOWNWARDS, and [ɓ] *b̤* is written ब and U+067B ب ARABIC LETTER BEEH. (I suppose it ought to be said that these are some pretty unbelievable Arabic character names, with regard to iconicity between their names and the sounds of the letters they represent. But… oh, well.) These four Devanagari letters were not encoded in previous versions of the standard because it was thought that they bore a diacritical mark which could be unified with U+0952 DEVANAGARI STRESS SIGN ANUDATTA. That character, however, is not identical with the mark which distinguishes Sindhi implosive consonants; the unification was false. The two graphs behave quite differently in the Devanagari writing system: the underbar in Sindhi is often (and I suggest, best) fused with the stem of the letter, and the vowel signs U and UU are drawn beneath it (see figures 3, 4, and 5). The ANUDATTA, on the other hand, is a stress accent applied to the entire syllable, and accordingly is placed below, not above, U and UU (see figure 6). No "combining implosive" diacritic is proposed here for the four Sindhi letters, for simplicity in encoding.

097B	ग	DEVANAGARI LETTER GGA
097C	ज	DEVANAGARI LETTER JJA
097E	ड	DEVANAGARI LETTER DDDA
097F	ब	DEVANAGARI LETTER BBA

Unicode Character Properties

```
097B;DEVANAGARI LETTER GGA;Lo;0;L;;;;;N;;;;;
097C;DEVANAGARI LETTER JJA;Lo;0;L;;;;;N;;;;;
097E;DEVANAGARI LETTER DDDA;Lo;0;L;;;;;N;;;;;
097F;DEVANAGARI LETTER BBA;Lo;0;L;;;;;N;;;;;
```

Bibliography

Lekhwani, Kanhaiyalal. 1987 (1909). *An intensive course in Sindhi*. Mysore: Central Institute of Indian Languages; [New York]: Hippocrene Books. ISBN 0-7818-9289-6

Whitney, William Dwight. 1960. *Sanskrit grammar: including both the classical language, and the older dialects, of Veda and Brahmana*. Cambridge: Harvard University Press; London: Oxford University Press.

Figures

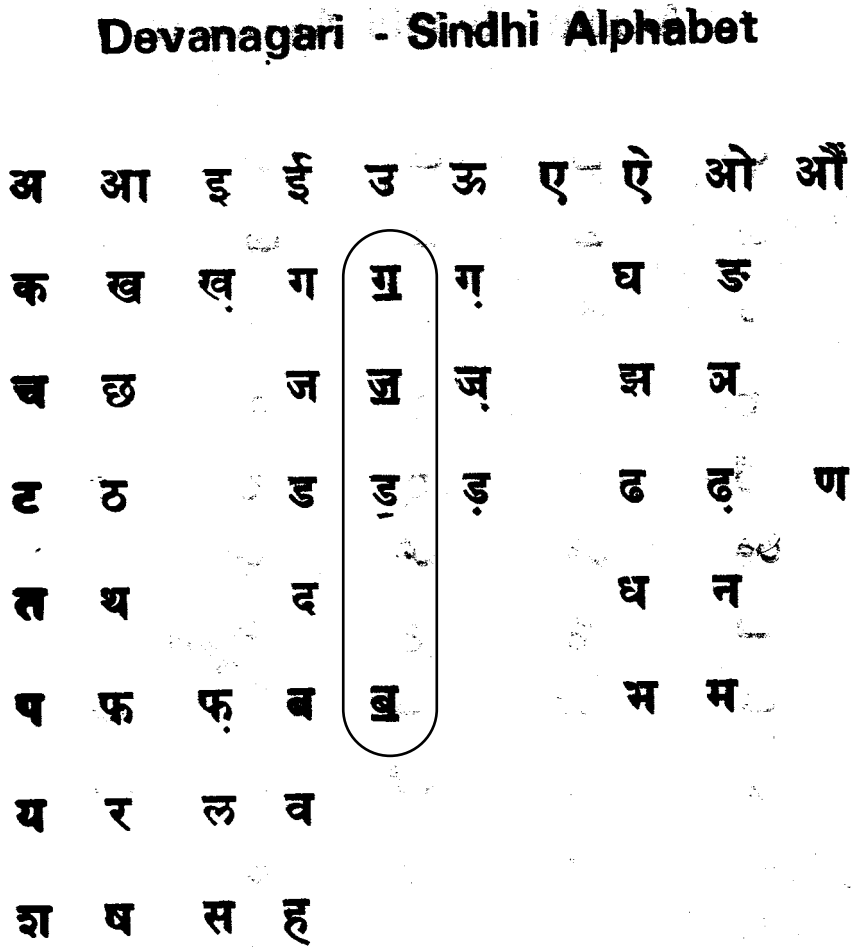


Figure 1. From Lekhwani 1987 (1909). Illustration of the Sindhi alphabet showing the letters *g̈* GGA, *j̈* JJA, *ḍ̲* DDDA, and *b̲* BBA. The underbar is attached to the vertical stem of the three characters which have it. Note the nuktated consonants beside the first three.

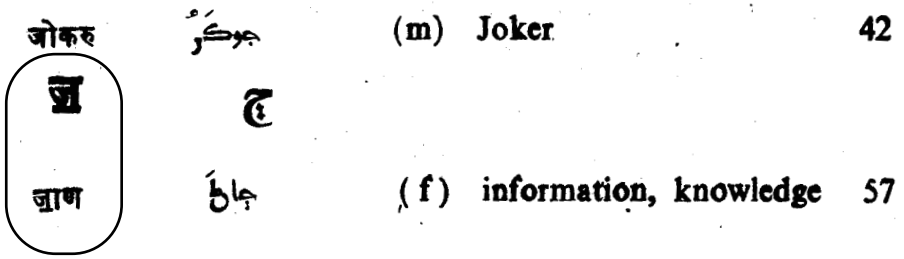
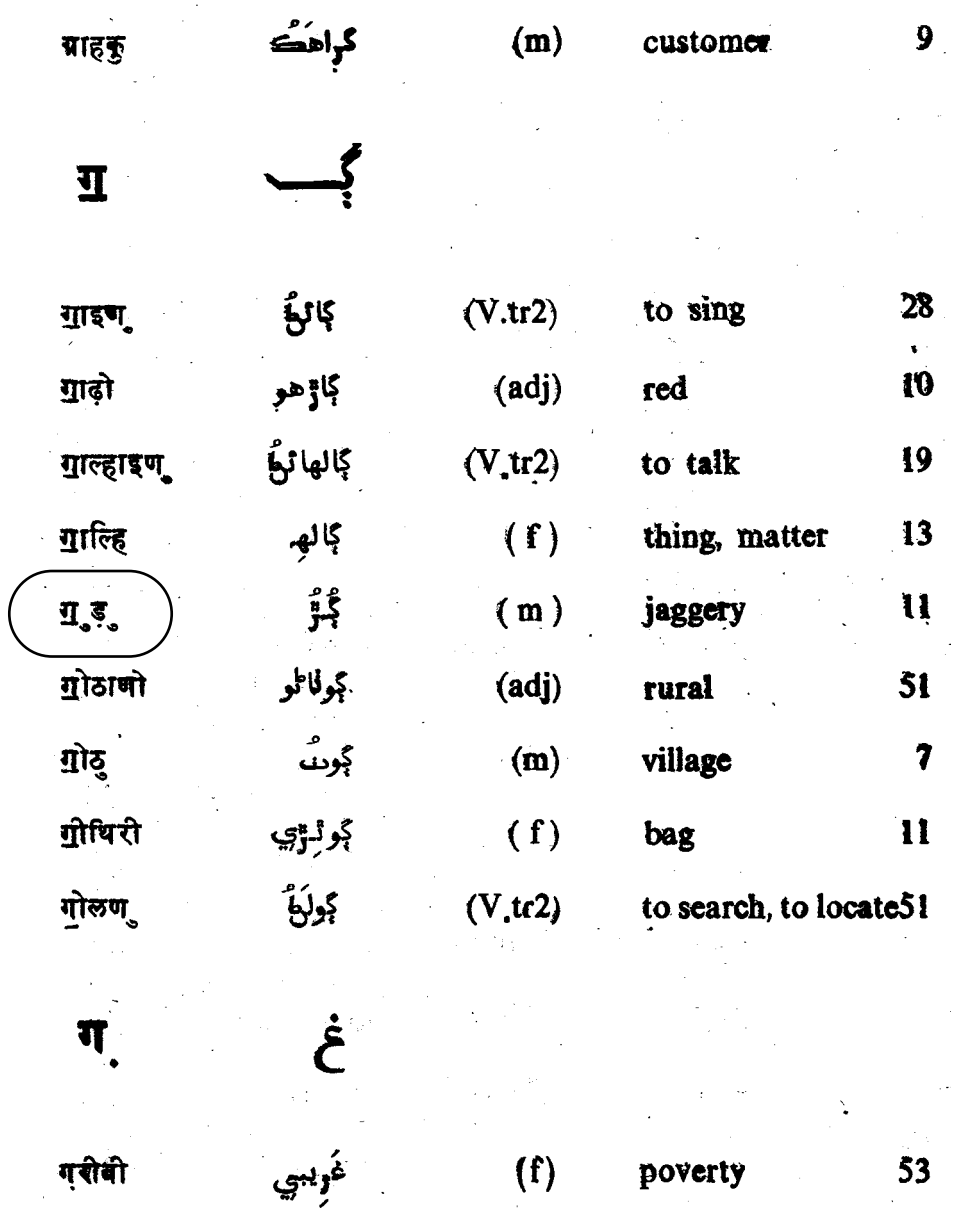
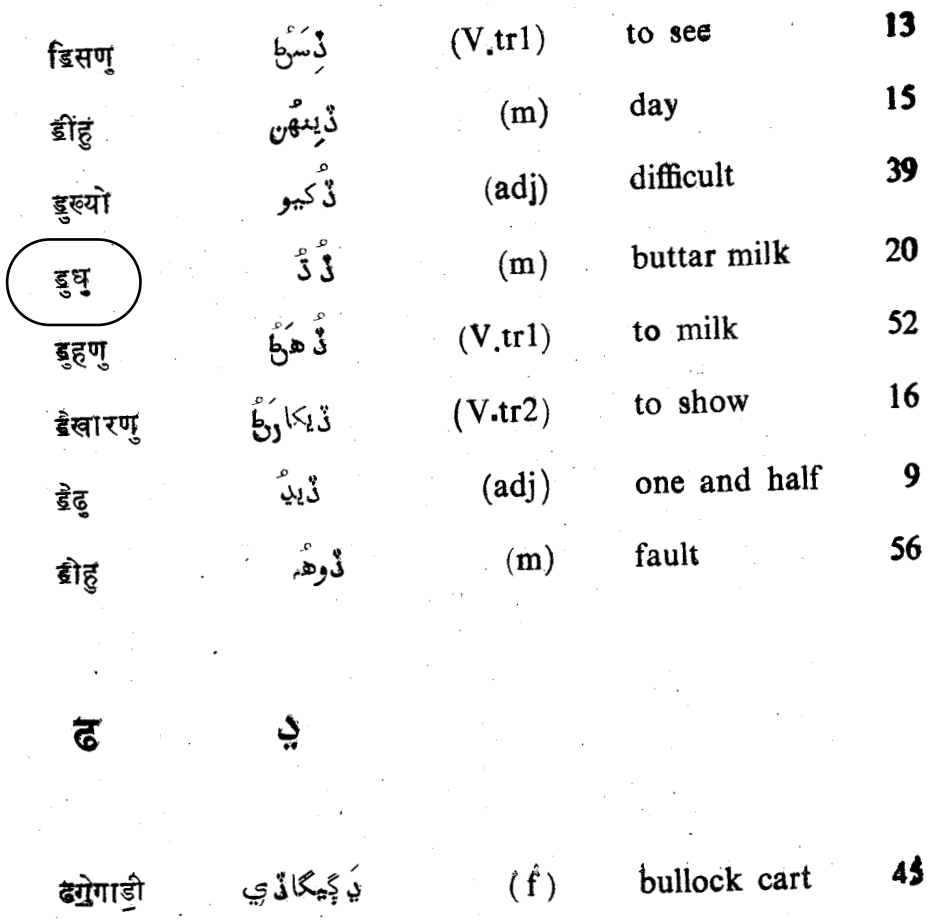


Figure 2. From Lekhwani 1987 (1909). Illustration from the glossary showing the letter *j̈* JJA, a separate letter following the letter *j* JA. The sample word shown is ज्ञाण *jäṇa* ‘information’.



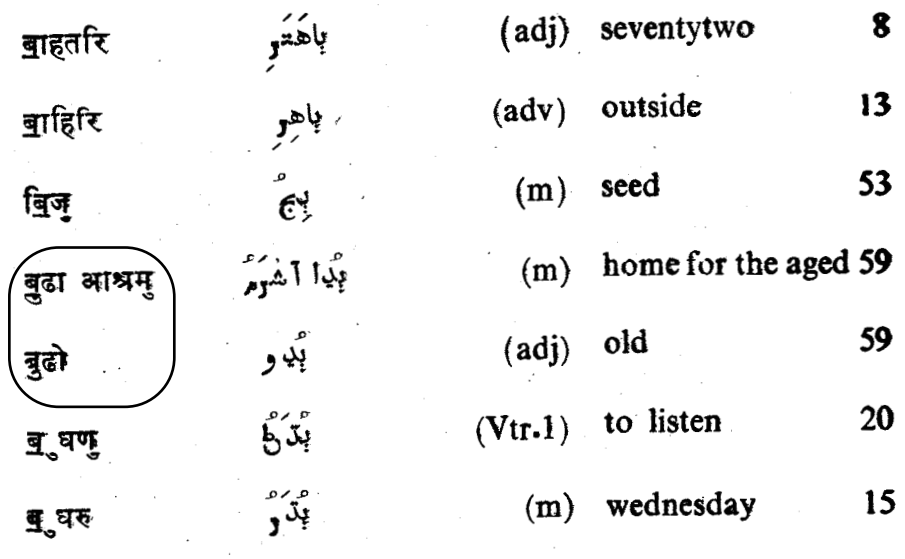
ग्राहकु	گراهڪ	(m)	customer	9
ग़	**غ**			
ग़ाइण्	غائڻ	(V.tr2)	to sing	28
ग़ाढ़ो	غاڙهو	(adj)	red	10
ग़ाल्हाइण्	غالهائڻ	(V.tr2)	to talk	19
ग़ाल्हि	غالهه	(f)	thing, matter	13
ग़ुड़ु	غُڙُ	(m)	jaggery	11
ग़ोठाणो	غوٺاڻو	(adj)	rural	51
ग़ोठु	غوٺ	(m)	village	7
ग़ोथिरी	غوٿڑي	(f)	bag	11
ग़ोलण्	غولڻ	(V.tr2)	to search, to locate	51
ग.	**غ**			
ग.रीबी	غريبي	(f)	poverty	53

Figure 3. From Lekhwani 1987 (1909). Illustration from the glossary showing the letter GGA, a separate letter after the letter GA and before the letter GHHA. The lead-type typography here is pretty bad, so the word *ġuṛu* ‘jaggery’ (an unrefined sugar) is not correctly written ग़ुड़ु. The same sort of error occurs on the following page of the glossary, where the matra is not drawn in the correct position after nuktated GA (so ग़ुसो instead of the correct ग़ुसो in *ɣuso* ‘anger’.)



डिसणु	ڈِسَݨُ	(V.tr1)	to see	13
डींहुं	ڈِينھُن	(m)	day	15
डुख्यो	ڈُکيو	(adj)	difficult	39
डुधु	ڈُڈُ	(m)	buttar milk	20
डुहणु	ڈُهَݨُ	(V.tr1)	to milk	52
डेखारणु	ڈيکارَݨُ	(V.tr2)	to show	16
डेढु	ڈيڍُ	(adj)	one and half	9
डोहु	ڈوهُ	(m)	fault	56
ढ	ڍ			
ढंगेगाड़ी	ڍَڱيگاڙي	(f)	bullock cart	45

Figure 4. From Lekhwani 1987 (1909). Illustration from the glossary showing the letter DDDA, a separate letter after the letter DDA and before the letter DDHA. Here the lead-type typography here is better, so the word डुधु *ḍudhu* ‘buttermilk’ is correctly written.



ब्राहतरि	بِاهَتَرِ	(adj)	seventytwo	8
ब्राहिरि	بِاهِرِ	(adv)	outside	13
ब्रिजु	بِرِجُ	(m)	seed	53
बुढा आश्रमु	بُڍا آشرَمُ	(m)	home for the aged	59
बुढो	بُڍو	(adj)	old	59
बुधणु	بُڌَݨُ	(Vtr.1)	to listen	20
बुधरु	بُڌَرُ	(m)	wednesday	15

Figure 5. From Lekhwani 1987 (1909). Illustration from the glossary showing the letter BBA, a separate letter after the letter BA and before the letter BHA. Here the lead-type typography here varies from line to line, but the word बुढा आश्रमु *ḅuḍhā āśramu* ‘home for the aged’ is correctly written; in बुढो *ḅuḍho* ‘old’ the stroke seems to have been omitted.

86. The essential difference of the two kinds of circumflex is shown clearly enough by these facts: 1. the independent circumflex takes the place of the acute as the proper accent of a word, while the enclitic is the mere shadow following an acute, and following it in another word precisely as in the same word; 2. the independent circumflex maintains its character in all situations, while the enclitic before a following circumflex or acute loses its circumflex character, and becomes grave; moreover, 3. in many of the systems of marking accent (below, **88**), the two are quite differently indicated.

87. The accentuation is marked in manuscripts only of the older literature: namely, in the primary Vedic texts, or **saṁhitās**, in two of the Brāhmaṇas (Tāittirīya and Çatapatha), in the Tāittirīya-Araṇyaka, in certain passages of the Āitareya-Āraṇyaka, and in the Suparṇādhyāya. There are a number of methods of writing accent, more or less different from one another; the one found in manuscripts of the Rig-Veda, which is most widely known, and of which most of the others are only slight modifications, is as follows.

a. The acute syllable is left unmarked; the circumflex, whether independent or enclitic, has a short perpendicular stroke above; and the grave next preceding an acute or (independent) circumflex has a short horizontal stroke below. Thus,

अग्निम् **agnim̀**; जुहोति **juhóti**; तन्वा **tanvà**; क्व **kvà**.

b. But the introductory grave stroke below cannot be given if an acute syllable is initial; hence an unmarked syllable at the beginning of a word is to be understood as acute; and hence also, if several grave syllables precede an acute at the beginning of a sentence, they must all alike have the grave sign. Thus,

इन्द्र **índraḥ**; ते **té**; करिष्यसि **kariṣyási**; तुविजाता **tuvijātá̀**.

c. All the grave syllables, however, which follow a marked circumflex are left unmarked, until the occurrence of another accented syllable causes the one which precedes it to take the preparatory stroke below. Thus,

सुदृशीकसंदृक् **sudṛ́çīkasaṁdṛk**;

but सुदृशीकसंदृग्गवाम् **sudṛ́çīkasaṁdṛg gávām**.

d. If an independent circumflex be followed by an acute (or by another independent circumflex), a figure 1 is set after the former circumflexed vowel if it be short, or a figure 3 if it be long, and the signs of accent are applied as in the following examples:

अप्स्वन्तः **apsv àntáḥ** (from **apsú antáḥ**);

रायोऽवनिः **rāyò3 vániḥ** from **rāyó avánih**.

Figure 6. From Whitney 1960. Illustration of Vedic Sanskrit text with U vowel and ANUDATTA. Unlike the mark which distinguishes Sindhi implosive consonants, the ANUDATTA is placed below the syllable it modifies. The *ju* जु shown here is not *jù* जु̀.

A. Administrative

1. Title

Proposal to add four characters for Sindhi to the BMP of the UCS.

2. Requester's name

Michael Everson

3. Requester type (Member body/Liaison/Individual contribution)

Individual contribution.

4. Submission date

2005-03-30

5. Requester's reference (if applicable)

6. Choose one of the following:

6a. This is a complete proposal

Yes.

6b. More information will be provided later
No.

B. Technical – General

1. Choose one of the following:
1a. This proposal is for a new script (set of characters)
No.
Proposed name of script
1b. The proposal is for addition of character(s) to an existing block
Yes.
1b. Name of the existing block
Devanagari
2. Number of characters in proposal
4
3. Proposed category (see section II, Character Categories)
Category A
4a. Proposed Level of Implementation (1, 2 or 3) (see clause 14, ISO/IEC 10646-1: 2000)
Level 1.
4b. Is a rationale provided for the choice?
Yes.
4c. If YES, reference
Brahmic spacing letters.
5a. Is a repertoire including character names provided?
Yes.
5b. If YES, are the names in accordance with the character naming guidelines in Annex L of ISO/IEC 10646-1: 2000?
Yes.
5c. Are the character shapes attached in a legible form suitable for review?
Yes.
6a. Who will provide the appropriate computerized font (ordered preference: True Type, or PostScript format) for publishing the standard?
Michael Everson. TrueType.
6b. If available now, identify source(s) for the font (include address, e-mail, ftp-site, etc.) and indicate the tools used:
Michael Everson. Fontographer.
7a. Are references (to other character sets, dictionaries, descriptive texts etc.) provided?
Yes, see bibliography above.
7b. Are published examples of use (such as samples from newspapers, magazines, or other sources) of proposed characters attached?
Yes.
8. 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, see above.
9. 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 Unicode Character Database http://www.unicode.org/Public/UNIDATA/UnicodeCharacterDatabase.html and associated Unicode Technical Reports for information needed for consideration by the Unicode Technical Committee for inclusion in the Unicode Standard.
Yes, see Unicode properties above.

C. Technical – Justification

1. Has this proposal for addition of character(s) been submitted before? If YES, explain.
No.
2a. 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.

2b. If YES, with whom?
Discussion on the indic@unicode.org list has shown support for the encoding of the characters.
2c. 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. Sindhi speakers writing Sindhi in Devanagari script.
4a. The context of use for the proposed characters (type of use; common or rare)
Commonly used Sindhi characters.
4b. Reference
See examples above.
5a. Are the proposed characters in current use by the user community?
Yes.
5b. If YES, where?
In Sindhi texts.
6a. After giving due considerations to the principles in Principles and Procedures document (a WG 2 standing document) must the proposed characters be entirely in the BMP?
Yes.
6b. If YES, is a rationale provided?
Yes.
6c. If YES, reference
All Devanagari points are in the BMP.
7. Should the proposed characters be kept together in a contiguous range (rather than being scattered)?
No.
8a. Can any of the proposed characters be considered a presentation form of an existing character or character sequence?
No. The similar character DEVANAGARI STRESS SIGN ANUDATTA has different behaviour with respect to the vowel matras, and does not in any case combine with the stem of its base characters.
8b. If YES, is a rationale for its inclusion provided?
8c. If YES, reference
9a. Can any of the proposed characters be encoded using a composed character sequence of either existing characters or other proposed characters?
No.
9b. If YES, is a rationale for its inclusion provided?
9c. If YES, reference
10a. Can any of the proposed character(s) be considered to be similar (in appearance or function) to an existing character?
No.
10b. If YES, is a rationale for its inclusion provided?
10c. If YES, reference
11a. Does the proposal include use of combining characters and/or use of composite sequences (see clauses 4.12 and 4.14 in ISO/IEC 10646-1: 2000)?
No.
11b. If YES, is a rationale for such use provided?
No.
11c. If YES, reference
12a. Is a list of composite sequences and their corresponding glyph images (graphic symbols) provided?
No.
12b. If YES, reference
13a. Does the proposal contain characters with any special properties such as control function or similar semantics?
No.
13b. If YES, describe in detail (include attachment if necessary)
14a. Does the proposal contain any Ideographic compatibility character(s)?
No.
14b. If YES, is the equivalent corresponding unified ideographic character(s) identified?
14c. If YES, reference