ISO/IEC JTC1/SC2/WG2 N5314

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 encode one newly-identified Tangut ideograph
Source: Andrew West
Status: Individual Contribution
Action: For consideration by JTC1/SC2/WG2 and UTC
Date: 2025-05-26

1. Introduction

This is a proposal to encode one newly-identified Tangut ideograph that is attested in a unique fragment of the Tangut *Homophones* text held at the British Library in London. The character is only partially complete on the fragment, but based on its gloss it is possible to reconstruct its glyph with a high degree of confidence. There is potentially also a second attestation for this character in the Tangut *Synonyms* text.

2. Background

The *Homophones* 同音 *·ẹi² lew²* (known in Chinese as *Tóngyīn* 同音 or *Yīntóng* 音同) is an important Tangut language phonetic text that lists Tangut characters according to nine phonetic initial classes, with homophonous characters grouped together. As such it is a major primary source for the repertoire of Tangut ideographs. There are two recensions of *Homophones*, both woodblock editions published during the Western Xia (1038–1227).[1]

The A version is represented by a single incomplete copy of a woodblock printed edition with a postface dated 1132 which is held at the Institute of Oriental Manuscripts (IOM) of the Russian Academy of Sciences (RAS) in Saint Petersburg, Russia. This version is missing most of folios 37B and 38A (about 84 entries in the Initial Class VII section), and parts of folios 54A, 54B, 55A, 55B, and 56A (at least 40 entries in the Initial Class IX section).

The B version is represented by six incomplete copies of woodblock editions held at IOM, as well as quite a large number of fragments held at the IOM, British Library, and various collections in China. Combining the various copies, this version has complete entries up to and including folio 53A, but folios 54B, 55A, and 55B are entirely missing, and folios 53B, 54A, 56A, and 56B are only partially preserved in fragments held at the IOM and British Library. In total, the B Version is missing about 224 entries in the Initial Class IX section.

The fact that both the A and B versions of *Homophones* are missing quite a large number of entries under Initial Class IX means that there are potentially unidentified and unencoded characters waiting to be discovered in the lost parts of *Homophones*.

[1] See WG2 N5305 p. 8 for detailed descriptions of the copies of *Homophones* held at the IOM.

3. Evidence for a newly-identified Tangut character

A fragment of *Homophones* B folio 56A held at the British Library includes a hitherto unidentified Tangut ideograph, which this document proposes for encoding.

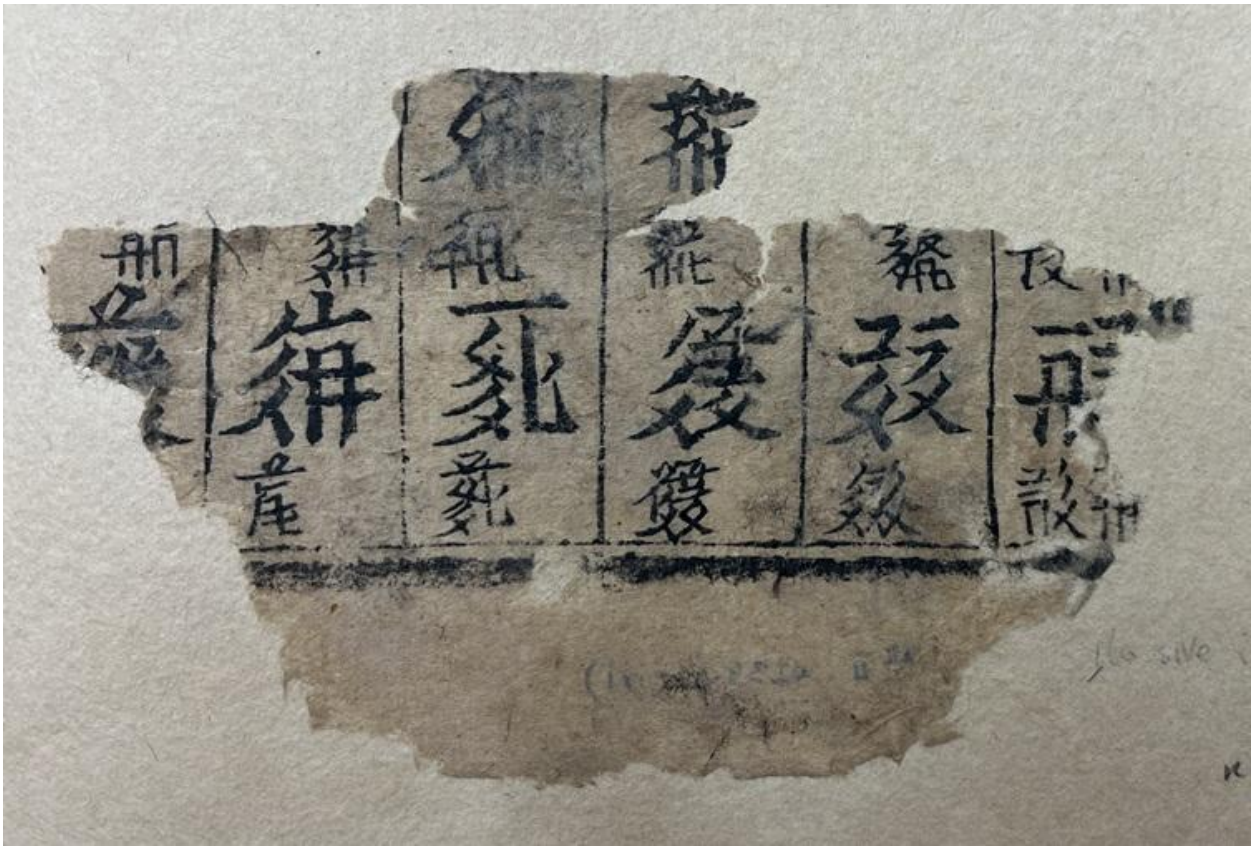

Fig. 1: British Library Or. 12380/3110.25.1[2]

British Library Or. 12380/3110.25.1 (Fig. 1) is a small fragment preserving twelve partial entries (eight head characters and thirteen complete or partial small character glosses) from *Homophones* B folio 56A (bottom two entries for cols. 2 through 7).[3] The entries must be under Initial Class IX (which starts on folio 46B), and should be under the 'single characters' subsection (characters with a unique pronunciation, so not homophonous with any other character). The entries in the two rightmost columns are not preserved in any other known fragment of *Homophones* A or B. Seven of the preserved head characters in this fragment can be identified, but the head character at the bottom right of the fragment, designated 56B28 (folio 56B column 2 position 8), is only partially preserved, and presents some difficulties in identification.

[2] This fragment is published in *Documents from Khara-Khoto in Britain* vol. 5 (Shanghai, 2010) p. 379 with the artificial pressmark Or.12380/3110.25.1. The actual British Library pressmark is currently undetermined, and the item has not yet been digitized, so is not available on the International Dunhuang Project (IDP) website (https://idp.bl.uk/). The colour photograph here has been kindly provided by Ms Han-Lin Hsieh, curator for the Chinese Collections at the British Library.

[3] The contents of this fragment overlap with IOM Tang 18/16, fr. 2 which preserves folio 56A cols. 4 through 7 and folio 56B col. 1 (see reproduction in Shǐ Jīnbō 2015 p. 121). The characters at the bottom of folio 56a cols. 2 and 3 are only preserved in the British Library fragment.

Fig. 2 shows a close-up of this entry, together with my proposed reconstruction.

Fig. 2: Detail of entry for the proposed Tangut ideograph

The left side of the head character is clearly Component 108 𗤓, but only the upper part of the right side component is preserved, apparently 𗂧. There are seventeen encoded Tangut ideographs with Component 108 on the left side, listed in Table 1.

Table 1: Tangut ideographs with Component 108 𗤓 on the left side

Code Point	Glyph	Sofronov Reading	Initial Class	*Homophones* B Entry
177E0	𗥃	*ndźi̯u*^{1}	VII	38A38
177E1	𗥄		IX	
177E2	𗥅	*ngi*^{2}	V	23B66
177E3	𗥆	*ldi̯ə*^{2}	IX	52A26
177E4	𗥇	*tśêu*^{1}	VII	38A22
177E5	𗥈	*ti̯ẹ*^{1}	III	14A16
177E6	𗥉	*man*^{1}	I	10A53
177E7	𗥊	*ri̯ẹ*^{2}	IX	47A62
177E8	𗥋	*lhwi*^{1}	IX	54A43

Code Point	Glyph	Sofronov Reading	Initial Class	*Homophones* B Entry
177E9	𗧩	*ra*	IX	
177EA	𗧪	*ngi²*	V	23B68
177EB	𗧫	*nịẹ²*	III	19A21
177EC	𗧬	*lị̂wẹ²*	IX	51B44
177ED	𗧭	*wo²*	II	12A63
177EE	𗧮	*kị²*	V	26A76
177EF	𗧯	*tśhị̂o²*	VII	36B66
177F0	𗧰	*lhwi¹*	IX	54A44
177E0	𗧱	*ndźị̂u¹*	VII	38A38

Ten of the characters in Table 1 are Initial Classes I, II, III, V, or VII, and have existing entries in *Homophones* B, so can be excluded from consideration. Five of the seven Initial Class IX characters already have head entries in *Homophones* B, and so can also be excluded from consideration. This leaves just two possible characters, U+177E1 𗧡 and U+177E9 𗧩, but neither have a right side component that matches the partial right side of 56B28. The only conclusion that can be drawn is that 56B28 is a new unencoded character.

Fortunately, the partially preserved two-character gloss underneath the head character allows the reconstruction of the glyph shape and reading for 56B28. The second gloss character (character on the bottom left of 56B28) is completely preserved, and can be identified as U+180FB 𘃻 *·ɹ̂en¹*. This is used as a gloss character for two other entries in *Homophones* A and B, where it and the preceding gloss character are together used to represent both the component construction of the head character and the *fanqie* 反切 reading for the head character (the first gloss character provides the initial consonant, and the second gloss character provides the rime and tone), as shown in Table 2.

Table 2: *Homophone* glosses with [Tangut]

***Homophones* Entry**	**Head Character**	**Gloss**	***Fanqie* Reading**	**Head Character Construction**
A 19B34 B 21A48	𘔗 *nên*^{1}	[Tangut][Tangut] *nɪn*^{1} *ên*^{1}	*n+ên*^{1}	Left side of [Tangut] and left side of [Tangut]
A 28B35 B 29A41	𘝰 *ngên*^{1}	[Tangut][Tangut] *ngi*^{1} *ên*^{1}	*ng+ên*^{1}	Right side of [Tangut] and left side of [Tangut]

Each of the head characters glossed using [Tangut] and another character are phonetic transcription characters, and so it can be deduced that 56B28 is also a phonetic transcription character riming with [Tangut] *ên*^{1}, and that the right side component of 56B28 is the same as the left side component of [Tangut], i.e. [Tangut]. This reconstruction exactly matches the partially preserved character in the British Library fragment.

The first gloss character (character on the bottom right of 56B28) should be used to indicate the initial consonant of the head character (which must be Initial Class IX), and provide the left side component of the head character (i.e. 𘡫). Although damaged, the gloss character does indeed appear to have Component 108 𘡫 on the left side, with a vertical stroke to its right, although unfortunately the rest of the character is lost. There are seven Initial Class IX characters with 𘡫 on the left: U+177E3 𗟣 *ldi̯ə*^{2}, U+177E8 𗟨 *lhwi*^{1}, U+177F0 𗟰 *lhwi*^{1}, U+177EC 𗟬 *li̯wẹ*^{2}, U+177E9 𗟩 *ra*, U+177E7 𗟧 *ri̯ẹ*^{2}, and U+177E1 𗟡. Of these, 𗟩 *ra* is the closest match for the partially preserved character, and it is also a Sanskrit transcription character which makes it suitable for use as a phonetic gloss. Therefore the first gloss character can be tentatively identified as 𗟩 *ra*, and the *fanqie* reading for the head character as 𗟩 *ra* + [Tangut] *ên*^{1} = [Tangut] *rên*^{1}.

That the newly-identified character [Tangut] is not attested in any other known source is not unusual, because the usage of transcription characters is very specialized, and many phonetic transcription characters are poorly attested. For example U+18517 𘔗 *nên*^{1} and U+18770 𘝰 *ngên*^{1} in Table 2 are both only attested once each in *Homophones*, and once each in the manuscript text of *Synonyms* [Tangut][Tangut] *wo*^{2} *leɯ*^{2} (known in Chinese as *Tóngyì* 同義), at 01B3 and 01B4 respectively.[4]

Transcription characters for Sanskrit mostly occur in *Synonyms* 01A5 through 01B6, and among these dedicated phonetic transcription characters, at 01B6.11, is a character that Lǐ Fànwén 李范文 and Hán Xiǎománg 韩小忙 (2005, p. 39) identify as U+17D76 𗵶 'broad', but which could possibly be the similar-looking [Tangut] identified above. The character 𗵶 is not used

[4] Lǐ Fànwén 2008, entry #1235 also gives two citations for 𘔗 meaning 'red', but these are very likely to be mistakes for U+1851B 𘔛 'red'.

specifically for phonetic transcription, so is anomalous in this position. It also occurs at 04B6.11, among a group of synonymous characters (with meanings such as 'vast', 'wide', 'broad', 'coarse', 'shallow'), which is suspicious as *Synonyms* generally only gives one entry for each character. The character at 04B6.11 (Fig. 4) is definitely U+17D76 𗵶 without a doubt, but unfortunately the character at 01B6.11 (Fig. 3) is damaged, and cannot be easily identified. It does not look particularly like U+17D76 𗵶, as the left side appears to be [illegible] rather than [illegible], but on the other hand it is not certainly [illegible]. All that can be said at this stage is that the character at 01B6.11 could conceivably be [illegible], but new discoveries of the *Synonyms* text would be needed to confirm this.[5]

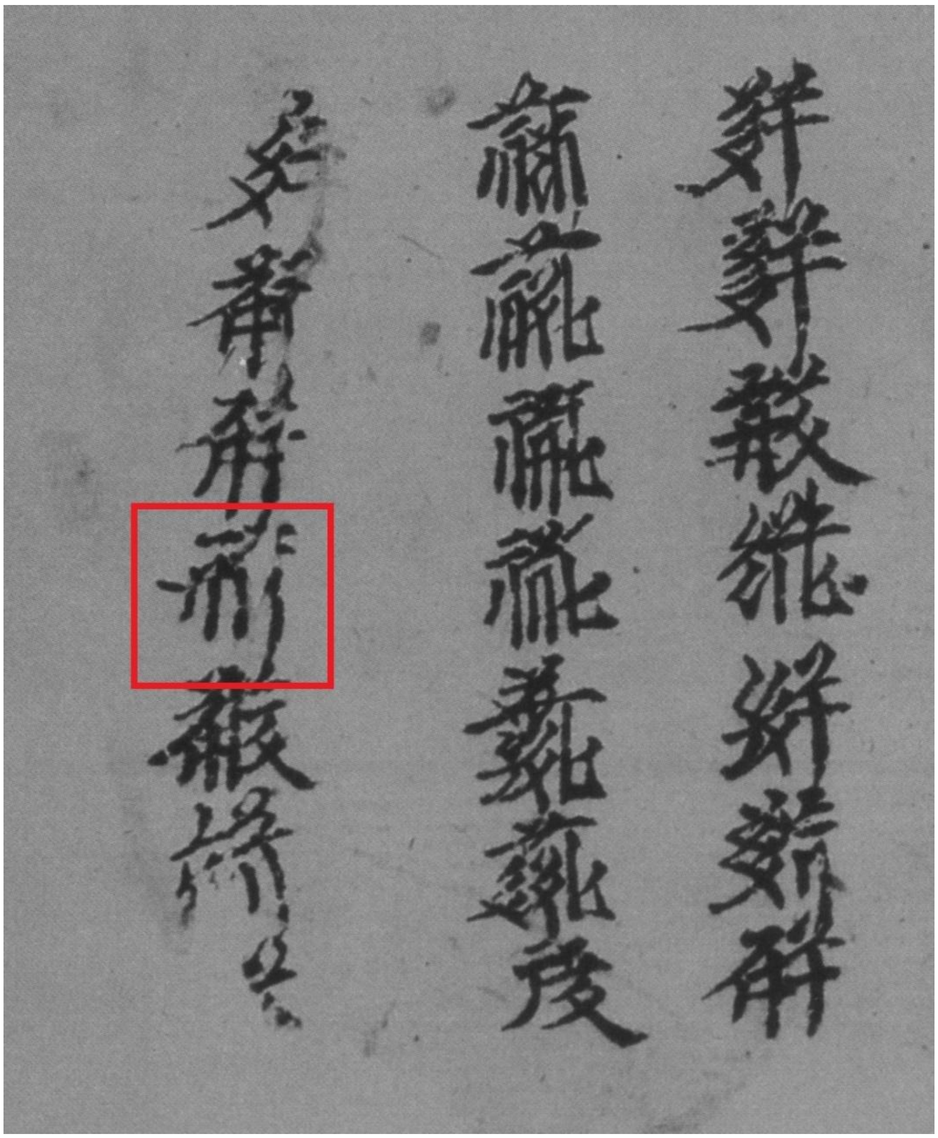

Fig. 3: *Synonyms* MS A folio 1B (IOM Tang 24/1, old inventory no. 2539)

[5] The IOM holds an almost complete manuscript text of *Synonyms* (Tang 24/1, old inventory no. 2539), as well as fragments from a different manuscript text (old inventory no. 2345) corresponding to folios 9B through 13A of the A manuscript. Additionally, two small fragments from a previously unknown woodblock printed edition of *Synonyms*, corresponding to folios 11B and 29A in the manuscript version, were discovered at the Shanzuigou site (Shānzuǐgōu shíkū 山嘴沟石窟) in 2005.

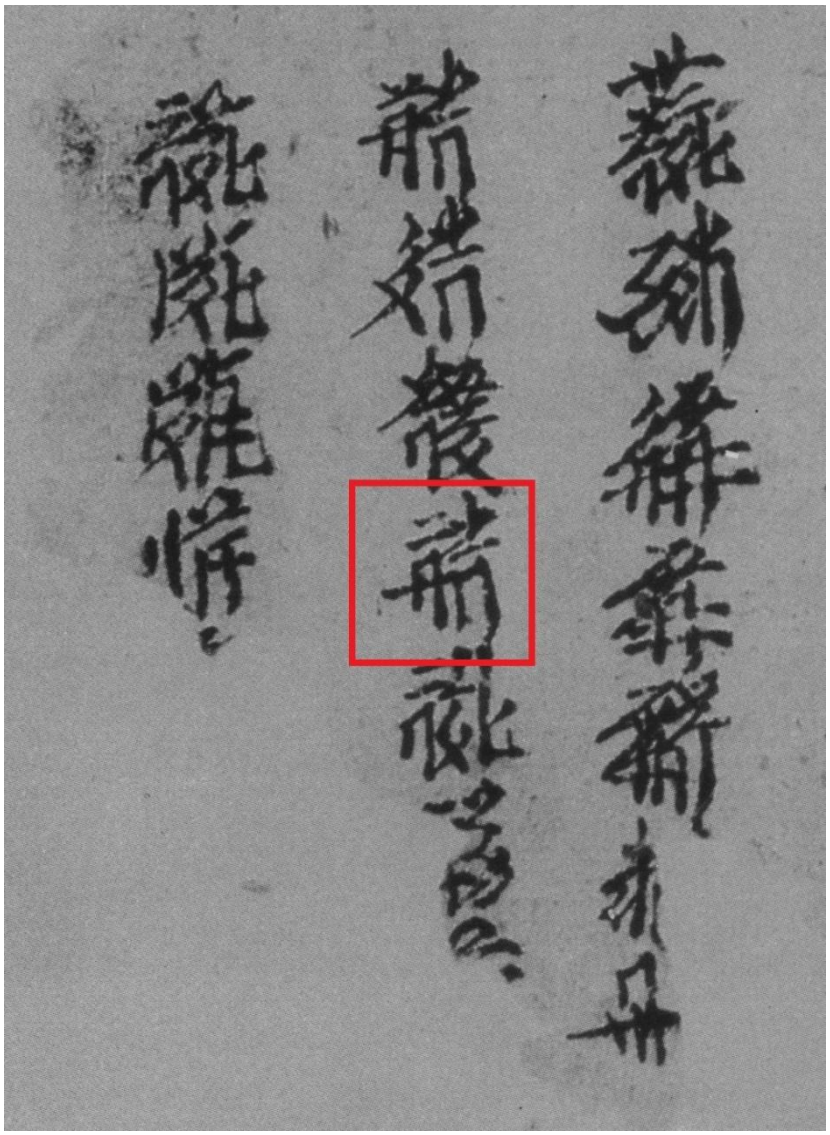

Fig. 4: *Synonyms* MS A folio 4B (IOM Tang 24/1, old inventory no. 2539)

4. Code Point Allocation and Character Properties

4.1. Code Point Allocation

The newly-identified Tangut ideograph can be encoded at U+18D20 in the Tangut Supplement block, following the Tangut character proposed for encoding at U+18D1F in WG2 N5303 = L2/25-144.

Table 3: Proposed New Tangut Ideograph

Code Point	Glyph	IDS	kTGT_RSUnicode	Source Reference
18D20			108.9	N5314-01

4.2. Unicode Character Properties

The entries for Tangut ideographs in the UnicodeData.txt file should be amended to:

```
18D00;<Tangut Ideograph Supplement, First>;Lo;0;L;;;;;N;;;;;
18D20;<Tangut Ideograph Supplement, Last>;Lo;0;L;;;;;N;;;;;
```

4.3. Tangut Sources

The following entries should be added to the TangutSources.txt file:

```
U+18D20   kTGT_RSUnicode N5314-01
U+18D20   kRSTUnicode    108.9
```

5. Bibliography

Documents from Khara-Khoto in Britain = Yīngcáng Hēishuǐchéng wénxiàn 英藏黑水城文獻 vol. 5. 2010. Shànghǎi: Shanghai Chinese Classics Publishing House. ISBN 978-7-5325-5621-2

Lǐ Fànwén 李范文 and Hán Xiǎománg 韩小忙. 2005. *Tóngyì yánjiū* 同义研究 [Study of the *Synonyms*]. In *Xīxià yánjiū* 西夏研究 [Xixia studies], Part I. Běijīng: Zhōngguó shèhuì kēxué chūbǎnshè 中国社会科学出版社. ISBN 7-5004-5204-7

Lǐ Fànwén 李範文 (comp.) and Jiǎ Chángyè 賈常業 (rev. and exp.). 2008. *Xià-Hàn zìdiǎn* 夏漢字典 [Tangut-Chinese dictionary (revised ed.)]. Běijīng: Zhōngguó shèhuì kēxué chūbǎnshè 中國社會科學出版社. ISBN 978-7-5004-2113-9

Shǐ Jīnbō 史金波. 2015. *Xīxià wénhuà yánjiū* 西夏文化研究 [Studies of Western Xia culture]. Běijīng: Zhōngguó shèhuì kēxué chūbǎnshè 中國社會科學出版社. ISBN 978-7-5161-3907-3

Proposal Summary Form

SO/IEC JTC 1/SC 2/WG 2
PROPOSAL SUMMARY FORM TO ACCOMPANY SUBMISSIONS
FOR ADDITIONS TO THE REPERTOIRE OF ISO/IEC 10646[6]
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 one newly-identified Tangut ideograph***
2. Requester's name: *Andrew West*
3. Requester type (Member body/Liaison/Individual contribution): *Individual contribution*
4. Submission date: *2025-05-26*
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: *Tangut Supplement*
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 X ; G-Obscure or questionable usage symbols
4. 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*
5. Fonts related:
   a. Who will provide the appropriate computerized font to the Project Editor of 10646 for publishing the standard?
   Andrew West
   b. Identify the party granting a license for use of the font by the editors (include address, e-mail, ftp-site, etc.):
   Andrew West
6. 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? *NO*
7. 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)? *NO*

8. 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 Unicode Character Database (http://www.unicode.org/reports/tr44/) and associated Unicode Technical Reports for information needed for consideration by the Unicode Technical Committee for inclusion in the Unicode Standard.

[6] Form number: N4102-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, 2007-03, 2008-05, 2009-11, 2011-03, 2012-01)

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? *Other experts*
   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? *NO*
   Reference:
4. The context of use for the proposed characters (type of use; common or rare) *rare*
   Reference:
5. Are the proposed characters in current use by the user community? *YES*
   If YES, where? Reference:
6. After giving due considerations to the principles in the P&P document must the proposed characters be entirely in the BMP? *NO*
   If YES, is a rationale provided?
   If YES, reference:
7. Should the proposed characters be kept together in a contiguous range (rather than being scattered)? *YES*
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, or could be confused with, 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 characters? *NO*
    If YES, are the equivalent corresponding unified ideographic characters identified?
    If YES, reference: