ISO/IEC JTC1/SC2/WG2
Universal Multiple-Octet Coded Character Set

Title: **Tangut Background**
Reference: **L2/07-143** "Proposal to encode Tangut characters in UCS Plane 1"
Doc Type: Working Group Document
Source: UC Berkeley Script Encoding Initiative (Universal Scripts Project)
Author: Richard Cook
Status: Liaison Contribution
Action: For consideration by UTC; provided for information to JTC1/SC2/WG2
Date: 2007-09-01

No action by WG2 is requested on *this* document — it is only provided for information.

The current document provides background information relating to

> **L2/07-143** (**WG2/N3297**) "Proposal to encode Tangut characters in UCS Plane 1".

L2/07-143 (p.2,3) refers to the following "Research Notes" link, for scanned examples, and for details on Tangut history, writing, and for background information relating to the mapping database and code charts:

> *http://unicode.org/~rscook/Xixia/index.html*

These notes (included below in PDF snapshot), go into some depth on points summarized in **L2/07-143**, and also reflect on-going revision of the property documentation in **L2/07-290** (== L2/07-158, "Proposed Draft Unicode Technical Report #43: A User's Guide to the UniTangut Database" [PDUTR #43]). These notes are now being provided to the technical committees for use in preparing text for a *Tangut Block Introduction,* for a future version of the standard.

The online database look-up tool (also mentioned in **L2/07-143**), has been updated with images for the full set of representative glyphs (in four styles, and in three sizes each) from the *Multi-Column Code Chart* (**L2/07-144**), and accesses data from a draft of "UniTangut.txt" (**L2/07-291**):

> *http://linguistics.berkeley.edu/~rscook/cgi/ztangut.html*

Copies of the latest revision of *PDUTR #43* (**L2/07-290**) and latest "UniTangut.txt" (**L2/07-291**) draft are also linked on that page.

Tangut (X xià) Orthography and Unicode

Research notes toward a Unicode encoding of Tangut

by **Richard Cook**

STEDT, **SEI**, **Unicode**

"   ,   ."

(, 1997:4)

'A Tangut character generally consists of many strokes,
and so the Tangut written language is one of the most difficult in the world.'

L Fànwén (*Tangut/Chinese Dictionary*, 1997:11)

X xià (a.k.a. [Tangut](); 　　　　　, 　　　　, 　　　　　, 　　; CH:2063; 　　　　　, cf. 　　　　 [1997:302.1572]()) is an extinct Sino-Tibetan (ST) language of central China (　　　　, 　　　　　 modern Níngxià Yínchu n; 　　 GMR 1979.2:41; [[N38.4, E106.3]()]). Conquered by the Mongols in 1227, there were 10 emperors total, spanning some 190 years. Tangut is sometimes thought to bear an especially close affiliation to Qiangic (　　　　; cp. 　　　　, CH:1875; [JAM:2004](), "'Brightening' and the place of X xià in the Qiangic branch of Tibeto-Burman"), and vestiges of Tangut speech are said to be found in minority languages of 　　 G nsù and 　　 Sìchu n provinces (　　 = Muya = Mi-nyag; 　, 　, 　 → 　　　; 　 → 　　　; 1997:2,11). The script and language have been studied in growing detail since the beginning of the 20th century, by Russian, Chinese, Swedish, Japanese, and American paleographers and phonologists (paleography is a criterion for phonology here), following [discoveries]() in the early 1900s, and gradual publication of a number of manuscripts (some only published in the late 1990s).

The 12th century Tangut rulers, asserting their independence from the Chinese, began to use Tangut as the official language, and decreed that a writing system should be developed, and that classical Buddhist and Chinese texts should be translated. Like the Chinese writing system which served as its conceptual model, the Tangut script (a.k.a. 　　　 X xià Wén, 　　　 Héx Zì, 　　 F n Wén, 　　 Tángg tè Wén, 　　　　　, 　　　　, ...) is a heterographic syllabary, which is to say that the writing is phonographic and syllabic (syllabographic), each element of the syllable canon having multiple representations, the representation of a specific member of any given homophone class being semantically motivated. Both cursive and square forms of Tangut writing are known, though the standard writing is the latter. There are also ornamental styles of Tangut writing, for example, there is a [Seal Script form]() styled after Chinese 　　 zhuàn . As in modern Chinese writing, each Tangut syllabograph is confined to a roughly (depending on the style or font) uniform em-square and conjoins elements drawn from a set of stroke primitives, this set being an [extension]() of that refined in the 　 Sòng Chinese orthographic and lexicographic [traditions]().

The Tangut script is componential, which is to say that similar assemblages of strokes recur in different characters, and these recurrent elements are variously

used as classifers for indexing (several competing   bùsh u *radical/stroke* and also *component* schemes have been developed by modern scholars). Thus, Tangut script elements may be described using a stroke-based [CDL](#), and in fact, structural analysis (as by Nishida and Kychanov) of this complex script would benefit greatly from future use of slightly augmented [CJK](#) CDL. Unlike Chinese writing, Tangut script elements (characters and components) are said to be wholly lacking in pictographic basis, which is to say that there do not appear to be native traditions seeing the characters or components as stylized depictions of real-world objects. In this respect Tangut writing is the opposite of its ST cousin   Nàx Tomba (¹Na-²khi ²Dto-¹mba), which is primarily pictographic (and rather non-phonographic).

All Tangut characters are highly reminiscent of Chinese characters (more or less so, depending upon the calligrapher), and give the general impression that squinting hard might bring them into focus as some form of ancient Chinese writing. But even to native Chinese readers the Tangut script is strange and impenitrable. A story was told to me by an historian at Academia Sinica of the discovery in   Míng-Q ng times of a stela bearing Tangut writing: the people were convinced that it was the work of the devil, and walled it up to protect themselves from its evil influence (they were too afraid to simply destroy it). Tangut characters are not simply non-standard (heretical) Chinese characters: rather, they write a non-Chinese language, and constitute a unique and innovative offshoot of sinitic graphological traditions. Unlike Vietnamese *Chu Nom* and Japanese *Kanji* (also writing non-Chinese languages), Tangut characters use several non-CJK stroke types and have mostly unanalysable (from a CJKV perspective) components. Tangut script lies completely outside CJK lexicographic traditions and outside the scope of UCS CJK unification, and Tangut characters are *not* treated as CJK "[ideographs](#)".

In contrast with Chinese and Chinese-derived scripts, which have well-defined graphological traditions filtered through standardizing reanalyses in the Eastern Hàn Dynasty (121 A.D.   ; see [Cook 2003](#)), there seems to be no surviving comparable native analysis of the Tangut script elements (cp. *Wén H i*). The "components" (according to some analysis) of Tangut characters are sometimes independent characters, but the contribution to (role in) the character composition (aside from contributing to the formation of a unique sign) is not always clear. Some few Tangut characters and components seem in fact to

be abbreviations or variant forms of Chinese characters, though the vast majority do not.

For example, the Tangut character 𖿀 *wuo* 'round' (LFW 1997:4743) seems to be a simplified writing of Chinese   (   ) yuán (< [SBGY](#):110.14, 142.15, 396.06 /Gju n/, /Gjuan/, /Gju ns/) 'round'; it is used as the left-side component in several other characters with 'round' meanings, and so seems to serve as semantic determiner in those compounds. As another example, the Tangut characters 𖿁 (LFW 1997:3745) *dzja* 'surname' and 𖿂 *dzji* 'male' ('   ,   '; 1997:3746) both seem to be variant writings of Chinese   xióng (< SBGY 026.06 /Gju /) 'male (bird)': the phonetic component   g ng (<   SBGY 201.46 /ku / 'arm') of Chinese   is given unique form in the Tangut (unattested in known Chinese variants), such that   is miswritten (from the Chinese perspective) rather like Chinese   d ng (< SBGY 032.30 /tuo /) 'winter'; the zhu 'bird' component written on the right-side of both, in the latter writing (3746) has a rather conservative form of the Chinese (clearly derivative of Eastern Hàn traditions, giving an explicit loop for the head of the bird, as in the Hàn Small Seal Script), whereas the former writing (3745) used for the surname is the normal simple Chinese form   . The 'bird' component of the Tangut 'male' writing is productive in the script (e.g. on the left side of LFW 1997:5567..9, 5543, 5282), but the contribution (phonetic or semantic) in those writings is unclear.

Such examples of relatively obvious parallels to Chinese characters are the exception: most Tangut characters make use of unanalysable (from a CJKV perspective) components, themselves comprised of decidedly [non-CJK stroke types](#). It may turn out that a comprehensive and persuasive graphological analysis of Tangut will appear in the future, but such work has not yet (to my knowledge) been accomplished, and a great deal of groundwork remains to be done. Nevertheless, the character repertory itself is very clearly delimited, as we shall see below.

The   Zh ngy ng Yánji yuàn [Academia Sinica Linguistics Institute](#)   X xià type face and mapping database, which played a key part in the early stages of the [UCS Proposal](#) development, were developed by

researchers under the direction of the historical phonologist　　　　Prof. Gong Hwang-cherng (Gōng Huángchéng). The primary source for this work was a native 12th century Tangut phonological text with the Chinese title *Tóngyīn* 'Homophones' (TY; Xīxià name /Ge_1 ̣u/ 'sound same', TYYJ:417-43B48, 469-53A72). A database was constructed on the basis of the character forms catalogued in collation of editions of this text analyzed by　　　　Lǐ Fànwén (1932- ; a.k.a.　　, LFW) in his study　　　　　*Tóngyīn Yánjiū* ['Homophony' Research; TYYJ] (　　　　　　　, 1986). This database originally contained mappings for a total of 5,805 Tangut characters: in the proofing process, a total of 4 omissions were found, bringing the total for this data set (and for the "W" column TY source font in the [Multi-column Code Chart](#)) to 5,809 TY Tangut characters (the following 4 missing TY characters were added, and assigned virtual TY indices [TY9990..TY9903]: LFW 1997: 2585, 3007, 3044, 4480).

In his introduction,　　　　Lǐ Fànwén wrote ([1986](#):1) [and I translate, inserting my notes in square brackets]:

"　　　　　　　　　　　　　，
　　　．　　　　　　　　　　　（　　1125　），
　（　　1132　）　　　　，　　　　．　　　　　　　（　　1176
）　，　　　　　　　　　，　　（　　1187　）
　　，　　　　　　．"

'The *Tóngyīn* (TY) text is an imperial Xīxià rhyme book compilation, important for the study of Tangut language, literature and especially phonology. The earliest printing of this book dates to the year 1125, and reprinted in 1132, these are known as the old editions of the text. Around the year 1176 the famous Tangut scholar Liáng Déyǎng revised the text, and in the year 1187 reprinted what is known today as the new or reprint version of the text.'

"　　　　　　　，　　　　　　，　　　　（
1132　）　　　．
　，　1909　　　　　（P. K. Kozlov）　　　（
）　　，　　　　　，　　　　（M. V. Sofronov）
　　　（1968）　　　　　　　　　　."

'The current Chinese editions of the text derive from the hand copy of Luó Fúchéng [made by his father   Luó Zhènyù (1866-1940) in 1919, first published in Japan, and published in 1935 in China; see L Fànwén 1986:926]. This is a copy of the old text of 1132. The old and new texts were excavated by [Kozlov](#) [   ] in 1909 at ancient H ishu (modern Éjìnàqí [a.k.a. [Khara-Khoto](#)], in inner Mongolia) and taken back to Russia. The originals up to the present have not been published [** this is no longer true, as of 1997; see   :
H ishu chéng Manuscripts Collected in the St. Petersburg Branch of the Institute of Oriental Studies of the Russian Academy of Sciences, Vol. 7; Tangut Secular Ms.; Shanghai Chinese Classics Publishing House, 1997; ISBN: 7-5325-2213 X/Z-307; Sinica Ling. Lib.: 797.9 2652 6:7], and are only known from Sofronov's 1968 *Tangut Grammar* [,   . .,
                                    ].'

"   ,
          ,   ,   ;   ,
      . [...] "

'According to analysis of the published material it can be seen that the major difference between the old and new editions of the TY text relate to their respective tonal classifications: the old text in large part does not distinguish the even and rising tone classes, which are seen as homophones; the new text distinguishes the two tones, putting them in separate classes. ...'

(1986:14): "   ,   ,
    .   .
        ,   ,   .   ,
                    ,   .   ,
            [:   ,   102-273   ,   484-
655   .]   [:   ,   499-607   ,
            559-668   .],   [:
                            1984   .]   ,
            ,   ,   (   ) 7.3% . [...]"

'The academic contribution which Luó Fúchéng made with his hand-copied edition of TY is generally recognized. A full fifty years have gone by since the

book was first published in 1935. Because Luó had not seen the *Wén Hǎi* [native Tangut dictionary] manuscript, and had not seen the new text of TY, but only had access to photos of the old TY text, materials at that time were lacking. He was unable to collate the different editions with complete accuracy, and slips of the pen and miswritten characters were difficult to avoid. Now, we have access to both old and new TY texts (thanks to Sofronov's *Grammar*), the *Wén Hǎi* manuscript (thanks to Kepping et al. 1969 [　．　．　　　　　　　　，　．　．　　　　　　　　，　．　．　　　　　　　　：　　　　　　　　　　　　], and Shǐ Jīnbō [1983], as well as the Tangut tomb inscription fragments (see Lǐ Fànwén 1984), and other primary resources. Collating the Luó Fúchéng text against these, we find a total of 864 errors, accounting for roughly 7.3% of the total text (including both large and small characters). [Tabulation of these errors follows.] [...]'

Some ten years later, in the *Introduction* to his 　　　　　　　　*Tangut-Chinese Dictionary* Lǐ Fànwén (1997:15) says more about these numbers:

"　　　　　　　　6,000　　　　（　　　　　　　）．
　　　　　　　：　　'　　6,133，　6,230'．　　　　　，　　　　　　　　300
　，　　　　　，　　　　　　　　'　　　　　6,000　　'，
　　　　，　　　　5,800　　，　　　　　　　　．"

'The present dictionary collects altogether 6,000 characters (including variants [and duplicates]). The author of the Tangut national book *Tóng Yīn* says in his preface that he wrote 6,133 main characters, and 6,230 annotation characters. The total number of *annotation characters* is not incorrect, but the count of *main characters* is over-estimated by about 300, and thereafter, as error breeds error, in academic circles down to the present day it is mistakenly reckoned that Tangut has about 6,000 characters. From extant materials we are certain that Tangut writing has only about 5,800 characters, and the rest are variants and mis-spellings.'

According to 　　　　Hán Xiǎománg (HXM), Xīxià characters were in use for a total of less than 500 years (1036-1502) [　　　　，2004.5:iii;
　　(*On Tangut Orthography*; Ph.D. dissertation K246.3 H211.7, directed by
　　Lǐ Fànwén)]; their use extended beyond the Mongol conquest (1227) for less

than 300 years. HXM undertakes a comprehensive and systematic collation of Tangut characters, based on nine Tangut dictionaries (,   ,   ,   ,   ,   ,   ,   ), and catalogues a total of 6,066 forms, including 169 variants, 36 errors, and 5,861 unique 'standard-style characters' ("   " zhèngzì 'orthography'). Because of the relatively short period of usage, and invention by decree, there is fairly little variation in character forms (L Fànwén 1997:11; especially relative to the Chinese script, which by comparison is written over perhaps 3,000 years, and reinvented, reanalyzed, and redefined periodically). HXM gives a complete set of mappings to the various primary manuscripts, to the *Xià-Hàn Dictionary* (L 1997), and to Sofronov (1968). HXM's total of 5,861 unique forms (larger by 21 than the Sinica TY inventory of 5,805 [+4; see above] characters), is divided into 3 broad categories (given below as **1**, **2**, **3**).

- **Category 1** with the bulk of the graphs has 5,981 forms (tokens, including variants), distributed over 5,784 types (there are 197 variants; 5,784 + 197 = 5,981); the characters in this category are rather well attested in the surviving literature (occurring in two or more sources, including TY and   WH), and have mappings to L (1997) and Sofronov (1968). Only **2** characters in this class lack mappings to L (1997), HXM 1066 = varclass 1015 (47B21) and 1877 = varclass 1805 (06B6,   0916.01).
- **Category 2** contains only 22 poorly attested "   " characters, occuring only once in a version of a primary lexical source (,   ,); 7 of these have mappings to L (1997), 2 of these 7 also having Sofronov (1968) mappings; there are no variant sets in this class. A total of **15** characters in this class lack mappings to L (1997): HXM varclasses: 5786, 5788, 5790, 5792, 5794, 5796, 5797, 5798, 5799, 5800, 5802, 5803, 5804, 5805, 5806).
- **Category 3** includes a total of 55 rare graphs, attested only in secondary lexical sources, or in commentaries; only 3 of these graphs have mappings to L (1997), and none has a mapping to Sofronov (1968); there are 7 variant pairs in this class, including mis-spellings. A total of **52** characters in this class lack mappings to L (1997): HXM varclasses: 5807, 5808, 5809, 5810, 5811, 5812, 5813, 5814, 5815, 5816, 5818, 5819, 5820, 5821, 5822, 5823, 5824, 5825, 5826, 5827, 5828, 5829, 5830, 5831, 5833, 5834,

5835, 5836, 5837, 5838, 5839, 5840, 5841, 5842, 5843, 5844, 5845, 5846, 5847, 5848, 5849, 5850, 5851, 5852, 5853, 5854, 5855, 5857, 5858, 5859, 5860, 5861).

HXM's multi-column text-based variant typology makes it clear that the Sinica character set includes all but two of the 5,784 Class 1 elements of the script, and that the Sinica character set also admits a total of 21 other forms which are distributed over Classes 1, 2 and 3. That is, by his analysis, each of these 21 forms is either a variant of a primary class member, or a secondary or tertiary class member. (The precise relation of the HXM varclasses to the Sinica character set is set forth in the online Tangut mapping data.) As with CJK Unified Ideographs, the set of Tangut characters is somewhat ill-defined, due to manuscript deficiencies and scribal variation, and yet, it is anticipated that (barring unexpected manuscript discoveries) the total number of candidates for future Tangut encoding (beyond the proposed repertory) will be small. As with CJK, issues relating to distinctive features and variant mapping can be handled by combination of variation selector (VS) and higher-level protocol (CDL). In contrast with CJK, although component-based lexical classifiers (   ) are in use for Tangut, the classifier systems are not standardized. Though it may be appropriate in the future to encode a block of Tangut Radicals (and perhaps even some stroke types), the present proposal does not include these: instead, we provide radical assignments and residual stroke-counts in the mapping data, following several such systems (e.g. LFW 1986, 1997; HXM 2004; documented in TR43). Character components or recurrent stroke patterns are best identified using CDL, the preferred method for indexing the members of this character set.

As mentioned in the proposal document, the first attempt to define a standard electronic encoding of Tangut was Grinstead's "Tangut Telecode" (*Analysis of Tangut Script*, Copenhagen: Scandinavian Institute of Asian Studies Monograph, 1971 [UCB DS 3 A2 S4 M6 No.8-10; ISBN 91-44-09191-5]). This early system, which assigns serial numbers to 5,819 characters and maps them to *Wén H i* (Kepping et al. 1969), apparently never became widely used. In more recent times, electronic Tangut data has been in use among Japanese and Russian scholars, based (it seems) primarily on the Mojikyo (   ) character collection, which applies a Shift-JIS encoding to all 6,000   L Fànwén (1997) entries

(Mojikyo numbers 570001 .. 576000 = LFW:1997: 0001 .. 6000; see Arakawa Shintaro below; this outline font [augmented by a bitmapped radical & component font] was used by   .   .   [E.I. Kychanov], *Tangut-Russian-English-Chinese Dictionary*, Kyoto Univ., 2006; cf.   Ikeda Takumi). See also the recent [work](#) on X xià undertaken at   (*B ij ng Zh ng Yì* Electronics Ltd.).

In addition to the primary lexical sources mentioned above, primary manuscript sources of Tangut are translations of Buddhist (cf. Kychanov:1999 *Catalogue*;   Nishida Tatsuo, *[Tangut Lotus Sutra](#)*, 2004) and classical Chinese texts (e. g.   Lín Y ngj n,   *The Tangut text of 'The Art of War'*, 1994). For a general introduction to Tangut primary and secondary documents, see "   " [Readers' guide to Tangut literature] (,   , 2004.2). For a recent bibliography of Tangut research in Chinese, Japanese, Russian, English, etc., see: *X xià gu nlián yánji   wénxiàn mùlù* [A catalogue of Tangut-related research literature] (2002: ISBN: 4-902325-00-4). In addition to the recent computerization work by   Hán Xi ománg (2004), see e.g.   NAKAJIMA Motoki et al. (,   :   1998;   ,   :   1997). For further information on Tangut, see also: Chinese Academy of Social Sciences (CASS), Institute of Ethnology and Anthropology [_____](#) [_____](#).

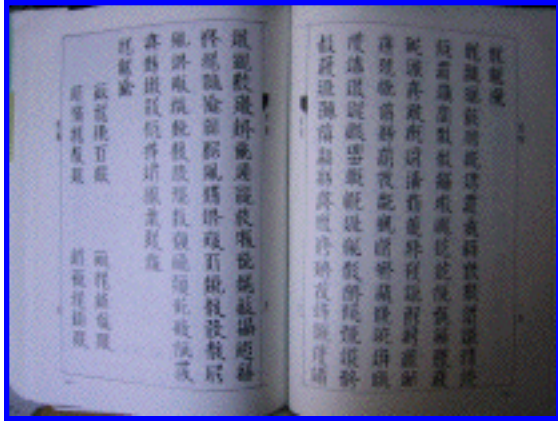# Images of some Tangut (X xià) texts and manuscripts

(JPG images, will open in a new window)
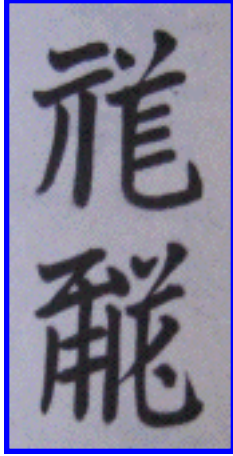
X xià /Ge_ l u/ 'sound same': Tangut title of the ancient Tangut phonology book called　　　　　　*Tóngy n* (*'Homophones'*) in Chinese ([TY](#)).
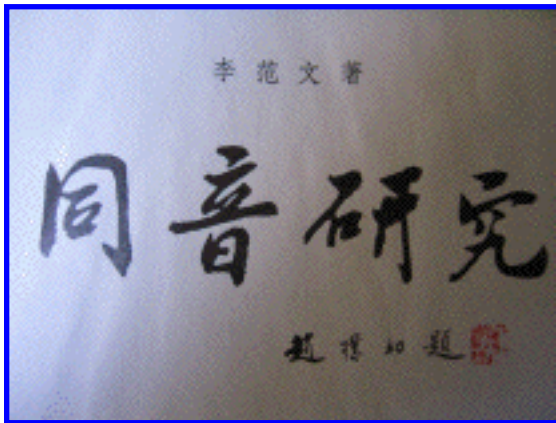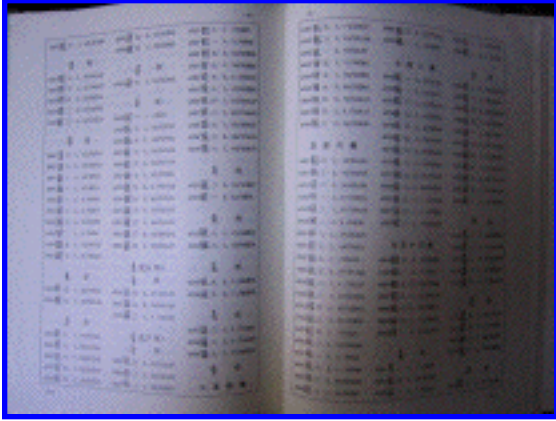

Title page of　　　　　　　　*Tóngy n Yánji*　(*'Homophones' Research*,　　　L Fànwén, 1986,　　　).

First page of the 1986 recension (title characters are at upper right, reading top-to-bottom in columns).

Random inner page (1986:53A,B), showing "large" ( ) and "small" ('note') Tangut characters.

Random index pages (1986:820-1), exemplifying primary mapping data print source.

Sample manuscript pages reproduced in   (1997; ISBN: 7-5004-2113-3):   ,   ,   ,   .

Another page from　　　　　1997, showing (lower right) examples of Tangut in
Seal Script (　　　Zhuànt) style.

A page from the main body of the dictionary (　　　　　1997:302.1572), showing
the entry for the native name of the Tangut state (glossed "　　　　　　").

The set of Tangut stroke types, Hán Xi ománg (2004).

For assistance in preparation of the present notes and materials for the encoding proposal, I am indebted to the following people at Linguistics Institute, Academia Sinica, Taiwan: Gong Hwang-cherng, Chuang Derming, Gau Yeachyi, C.C. Cheng, Lin Ying-chin, Jonathan Evans; and thanks also to Hwang Ming-chorng of Institute of History and Philology, Academia Sinica.

Special thanks also to Ikeda Takumi, Jim Matisoff, Ken Lunde, Martin Heijdra, Andrew West, Atsuhiko Kato, Anan Yasuhiro, and Arakawa Shintaro. See the full acknowledgments in the encoding proposal.

Page Generated in Perl:
Sat Sep 1 13:04:17 2007
Berkeley, California, USA