

JTC1/SC2/WG2 N3496
2008-09-01

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

Doc Type: Working Group Document
Title: Review of Proposed Tangut Repertoire
Source: United Kingdom
Status: National Body Contribution
Action: For consideration by JTC1/SC2/WG2 and UTC
Date: 2008-09-01 (revised 2008-09-07)

1. Introduction

As part of its review of PDAM6.2 (currently under ballot as SC2 N4028) the UK national body has carried out a review of the proposed set of 5,910 Tangut characters at 17000..18715, as proposed in N3297. The purpose of this review is the determine whether the proposed repertoire is complete and will satisfy the needs of Tangut scholars.

We have reviewed the proposed repertoire against the characters included in the following sources:

1. Kučanov, E. I. (Е. И. Кычанов). 2006. *Словарь тангутского (Си Ся) языка* (= *Slovar' tangutskogo (Si Sja) jazyka*) [Tangut-Russian-English-Chinese Dictionary]. St. Petersburg and Kyoto.
2. Lǐ Fànwén (李範文). 1997. 夏漢字典 (= *Xià-Hàn Zìdiàn*). Beijing.
3. Lǐ Fànwén (李範文). 1986. 同音研究 (= *Tóngyīn Yánjiū*). Yinchuan.
4. Nishida Tatsuo (西田龍雄). 1966. 西夏文字小字典 (= *Seika Moji Shōjiten*). In 西夏語の研究 (= *Seikago no kenkyū*) [A Study of the Hsi-Hsia Language] (1964-1966) vol.2. Tokyo.
5. Shǐ Jīnbō (史金波) *et al.*. 1983. 文海研究 (= *Wénhǎi Yánjiū*). Beijing.
6. Sofronov M. V. (М. В. Софронов). 1968. *Грамматика тангутского языка* (= *Grammatika tangutskogo jazyka*). Moscow.

We have also referred to the 2004 dissertation by Hán Xiǎománg 韓小忙 (*Xīxiàwén Zhèngzì Yánjiū* 西夏文正字研究), which is the main source for the proposed Tangut repertoire, supplemented by additional characters from Lǐ Fànwén 1986 and Lǐ Fànwén 1997 that are not in Hán Xiǎománg 2004.

2. Conclusion

It is evident from our review that the set of 5,910 Tangut characters proposed for encoding represents a restricted subset of the complete set of Tangut characters used in contemporary Tangut texts and required by modern scholars of the Tangut language. This is because the Tangut repertoire proposed in N3297 is largely based on Hán Xiǎománg's list of the "correct" forms of Tangut characters, and therefore excludes many character forms that are deemed "incorrect". Whilst this is a useful scholarly exercise, it is far from an ideal basis for defining the repertoire of Tangut characters to be encoded, as it results in a prescriptive rather than a descriptive encoding model.

If the defective set of characters currently proposed is accepted then Tangut scholars will be unable to represent the forms of characters that are actually used in contemporary Tangut texts, but will have to limit themselves to those character forms that Hán Xiǎománg believes *should have been* used by Tangut authors. This is totally unacceptable. Even worse, the proposed character set is insufficient to represent *all* of the characters in *any* of the modern dictionaries of Tangut, with the result that it will be impossible to fully represent the complete text or even indexes of Tangut dictionaries in Unicode without resorting to the PUA, which defeats the purpose of encoding Tangut in the first place. Consequently, there is a high risk that Tangut scholars will resist migrating from a Mojikyo-based solution (which has complete coverage of Lǐ Fànwén 1997) to a UCS solution (which has less than 98% coverage of Lǐ Fànwén 1997) unless the proposed repertoire of Tangut characters to be encoded is expanded to fully cover all of the characters used in the major modern dictionaries of Tangut.

In summary, we believe that the proposed set of Tangut characters is **wholly inadequate**, and does not meet the needs of the user community.

3. Lǐ Fànwén 1997

Lǐ Fànwén's Tangut-Chinese dictionary published in 1997 is the most comprehensive Tangut dictionary to have been published to date, and includes 6,000 entries for individual Tangut characters. Based on the information provided in UniTangut.txt we have listed in Table 1 all cases where more than one character in Lǐ Fànwén's dictionary corresponds to a single character in the proposed Tangut repertoire (Lǐ Fànwén character indexes prefixed by 'L', and the corresponding character index from Kyčanov 2006 prefixed by 'K'). These unifications, which derive from Hán Xiǎománg 2004, are not explicitly discussed or justified in the proposal (N3297), but only become evident from a detailed analysis of the proposed characters and the accompanying UniTangut "database". We would have expected major issues such as unification to have been thoroughly dealt with in the proposal, and a set of unification principles to have been proposed.

The 133 characters marked with asterisks and highlighted in red in Table 1 have a different actual shape to the corresponding character proposed for encoding, and are used contrastively with the proposed character, and therefore we do not believe that they should be unified (likewise for Tables 2, 3, 4 and 5). Whilst there is considerable stylistic variation in the way that Tangut characters are written in different sources, and unification of insignificant glyph variants is to be expected, where variant forms of the same character have *different structural compositions*, and are *used contrastively in the same source*, they cannot be unified, and must be encoded separately. Even though most of the 133 highlighted characters are variant forms of characters that are included in the proposed repertoire, they cannot simply be ignored, and it is still necessary to encode them separately so as to enable Lǐ Fànwén's dictionary to be fully representable using Unicode.

Although, as we have said, the vast majority of characters unified in Table 1 are variant forms of the same character, there are a few examples of unified characters that are totally unrelated to each other, for example Lǐ Fànwén 3877 𐰇 and Lǐ Fànwén 3880 𐰈, which are unified as U+17F88 𐰇:

3142 42 𐰇 3877	〔重唇音 mjij 1.36 音名〕 lower limbs; legs 下肢也。(名)
3144 44 𐰈 3880	〔(聲韻不詳)〕 ghost; spirit 鬼、蜮也。(名)

Figure 1 : Lǐ Fànwén 3877 and 3880

The implicit justification for this unification is Hán Xiǎománg 2004:

4028	𐰇	同音 _甲 03B68, 同音 _乙 04B47, 文海 _甲 ①70. 231, 文海 _乙 ①30. 104, 掌中珠 _甲 245, 杂字 _乙 06A6, 同义 _甲 0415. 10	3877	3005	注⑭	𐰇	3891
4029	𐰈	文海 _甲 ①84. 272, 合编 _甲 17. 073	3880		讹体		

Figure 2 : Hán Xiǎománg 2004 page 231

Note 144 (page 341) referred to above simply states "𐰇 is the correct form, 𐰈 is the corrupt form" 𐰇为正体, 𐰈为讹体. Whether or not one of these characters is a corrupt form of the other is totally irrelevant at the character encoding level. The only fact that needs to be considered is that these two different character forms are both used in contemporary Tangut sources, and modern scholars such as Lǐ Fànwén recognise them as different characters—this must be reflected in the character encoding.

In other cases the difference between glyphs is subtle but important, for example Lǐ Fànwén 0541 (Kyčanov 2887) 𐰇 and Lǐ Fànwén 0542 (Kyčanov 2880) 𐰇 differ only slightly, the left side element being written as four strokes in the one character and as five strokes in the other (see also Figure 8 for Nishida's treatment of these two characters) :

𐰇 0541	〔正齒音 lju 2.6 音尚合〕 dignifie 莊嚴、端正、美好也。(形) 𐰇 𐰇 _{liuo} 2.44 liu 2.6 〔尚合緣〕端正	𐰇 0542	〔來日音 śjwo 2.44 音六〕 beautiful 美麗、彩飾、丹、玫也。(形) 𐰇 𐰇 _{liou} 2.44 liu 2.6 〔尚合緣〕 美麗、燦爛(同 50B7)。
-----------	--	-----------	---

Figure 3 : Lǐ Fànwén 0541 and 0542

2887-0 𢇛 síwo 2.44 “прямой; честный; подтянутый; собранный”, “honest; straightforward”, “誠實; 正”.
 2880-0 𢇛 síwo 2.44 “красивый; наложница”, “beautiful; coloured; concubine”, “麗; 妾”.

Figure 4 : Kyčanov 2880 and 2887

Although the difference between these two characters is easily lost in casual handwriting, all three of the most distinguished living experts on Tangut (Li Fànwén, Kyčanov and Nishida) agree that these are separate characters, and differentiate the two characters at the glyph level in their dictionaries (note especially the compound word formed from 0541 and 0542 highlighted in Figure 3). However, Hán Xiǎománg 2004 does not distinguish between these two characters at the glyph level, and so these two characters have been unified as U+17CED. This is clearly wrong.

3364	𢇛	同音 _甲 40A73, 同音 _乙 40B52, 文海 _甲 ①67. 131, 文海 _乙 ②63. 608, 掌中珠 _甲 133, 杂字 _乙 15B5, 同义 _甲 2223. 03, 切韵 _甲 26A4	0541	0153	多音 多义	𢇛	3244
		同音 _甲 50B76, 同音 _乙 51A72, 文海 _甲 ①59. 161, 文海 _乙 ②80. 408, 掌中珠 _甲 133, 杂字 _乙 15B5, 同义 _甲 2223. 04	0542	1177			

Figure 5 : Hán Xiǎománg 2004 page 195

Another similar example is the pair of characters Li Fànwén 1666 𢇛 (left side element written as three strokes) and Li Fànwén 1667 𢇛 (left side element written as four strokes), which have different glyph shapes, different meanings and different reconstructed pronunciations, but which are unified as U+177E1 in the proposal:

1841 00	[舌頭音 nə(聲調不詳)音能]	𢇛	[舌頭音 ta 1. 17 𢇛 𢇛 𢇛 都臘切 音 恒]
1666	fox	1667	tail
	狐也。(名)		①尾也。(名)

Figure 6 : Li Fànwén 1666 and 1667

The difference in glyph shape between these two characters is exactly analogous to the difference in glyph shape between the following pair of characters that are not unified in the proposal (1668 = U+17C59 ; 1669 = U+17D9F):

1668	[喉音. o 1. 49 𢇛 𢇛 𢇛 烏播切 音 訛]	𢇛	[來日音 lhjij 1. 42 𢇛 龍 𢇛 嘞 很切 音 吟]
	surname	1669	hear ; listen
	[訛]族姓也。(音)		聽、聞也。(動)

Figure 7 : Li Fànwén 1668 and 1669

In both of the above pairs of characters the first stroke in the one character is a horizontal stroke with an oblique bend (1666 and 1668) whereas the first stroke in the other character is a simple horizontal stroke followed by a separate slanting stroke (1667 and 1669). Yet, in the one case the two characters are unified, and in the other case the two characters are not unified. This example supports our contention that there is no reason in principle why any of the characters marked in Table 1 should not be encoded separately.

Table 1 : Lǐ Fànwén 1997 Unifications

Unicode		Lǐ Fànwén 1997 (Kyčanov 2006)					
Code	Glyph	Code	Glyph	Code	Glyph	Code	Glyph
U+1700F	𠂔	L0017 K4240	𠂔	L0486 K4241	* 𠂔 *		
U+1701C	𠂕	L0042 K0998	𠂕	L0215	* 𠂕 *	L4537 K0998	𠂕
U+17022	𠂖	L0057 K2389	𠂖	L0232 K2388	* 𠂖 *	L4548	𠂖
U+171D3	𠂗	L0064 K5739	* 𠂗 *	L1147 K5741	𠂗	L3832 K5740	* 𠂗 *
U+18172	𠂘	L0076 K4026	* 𠂘 *	L0786 K4027	𠂘		
U+186B7	𠂙	L0079 K5507	* 𠂙 *	L0676 K5508	𠂙		
U+175C0	𠂚	L0162 K5394	𠂚	L0202 K1334	* 𠂚 *		
U+1817B	𠂛	L0184	𠂛	L1919	* 𠂛 *	L4510 K4462	* 𠂛 *
U+183E2	𠂜	L0187 K4735	𠂜	L0647 K4742	* 𠂜 *		
U+17D15	𠂝	L0213 K0279	𠂝	L1957 K0280	* 𠂝 *		
U+17D69	𠂞	L0256 K5681	* 𠂞 *	L0316	𠂞		
U+180CD	𠂟	L0259 K0005	𠂟	L0345	* 𠂟 *		
U+1731E	𠂠	L0266 K5390	𠂠	L0325	* 𠂠 *		
U+177D7	𠂡	L0271 K3715	* 𠂡 *	L0534 K4235	𠂡	K4236	* 𠂡 *
U+17DA4	𠂢	L0276 K4858	𠂢	L2099 K4857	* 𠂢 *		
U+17CF0	𠂣	L0346 K2430 K2431	𠂣	L0347	* 𠂣 *		
U+17D8B	𠂤	L0406 K5297	𠂤	L0407 K5297	𠂤		
U+1756E	𠂥	L0430 K0011	𠂥	L0452	* 𠂥 *		
U+17247	𠂦	L0431 K0012	𠂦	L0453	* 𠂦 *		
U+180EE	𠂧	L0490 K5604	* 𠂧 *	L0987 K5579	𠂧		
U+17D14	𠂨	L0514	𠂨	L2388 K2668	* 𠂨 *		
U+18176	𠂩	L0515 K2684	𠂩	L2390	* 𠂩 *		
U+185CA	𠂪	L0523 K3500	𠂪	L4549 K3501	* 𠂪 *	L4565	* 𠂪 *

Unicode		Lǐ Fànwén 1997 (Kyčanov 2006)					
Code	Glyph	Code	Glyph	Code	Glyph	Code	Glyph
U+17D67	𪛓	L0533 K5684	* 𪛓 *	L0555	𪛓		
U+17CED	𪛔	L0541 K2887	𪛔	L0542 K2880	* 𪛔 *		
U+17525	𪛕	L0583 K2776	* 𪛕 *	L1362 K2775	𪛕		
U+17BC3	𪛖	L0608 K5594	* 𪛖 *	L1878 K5578	𪛖		
U+1817C	𪛗	L0644 (K4579)	𪛗	L2553	* 𪛗 *		
U+1817F	𪛘	L0671 K0287	𪛘	L2572 K0288	* 𪛘 *	L4578 K0295	* 𪛘 *
U+17248	𪛙	L0735	𪛙	L1521 K3418	* 𪛙 *		
U+17D3C	𪛚	L0895 K5566	* 𪛚 *	L4491 K5560	𪛚		
U+17D13	𪛛	L1039 K1652	𪛛	L2943	* 𪛛 *		
U+171E8	𪛜	L1068 K3207	𪛜	L3827 (K3206)	* 𪛜 *		
U+18408	𪛝	L1106 K3448	𪛝	L1107 K3448	𪛝		
U+17D12	𪛞	L1131 K0597	𪛞	L2985	* 𪛞 *		
U+1817D	𪛟	L1134 K0610	𪛟	L2988	* 𪛟 *	L4624	* 𪛟 *
U+180A4	𪛠	L1151 K5711	* 𪛠 *	L1296 K5572	𪛠		
U+1809A	𪛡	L1217 K5772	* 𪛡 *	L1355 K2318	𪛡		
U+18095	𪛢	L1317 K2374	𪛢	L1318 K2374	𪛢		
U+17CE6	𪛣	L1383 K2478	𪛣	L1384 K2478	𪛣		
U+17C99	𪛤	L1493 (K0125)	𪛤	L1505 K0138	* 𪛤 *		
U+17563	𪛥	L1538 K2633	𪛥	L3069 K2632	* 𪛥 *		
U+17D63	𪛦	L1605 K5351	𪛦	L1665 K1308	* 𪛦 *		
U+177E1	𪛧	L1666 K4680	𪛧	L1667	* 𪛧 *		
U+180A1	𪛨	L1734 K3262	𪛨	L1735 K3263	𪛨		
U+17CF4	𪛩	L1871 K1415	𪛩	L1872 K1415	𪛩		

Unicode		Lǐ Fànwén 1997 (Kyčanov 2006)					
Code	Glyph	Code	Glyph	Code	Glyph	Code	Glyph
U+17122	𠂔	L1947 K0212	* 𠂔 *	L2354 K0149	𠂔		
U+1704E	𠂔	L1960 K5721	𠂔	L1975 K1039	* 𠂔 *		
U+1704D	𠂔	L1972 K1931	* 𠂔 *	L2971 K1930	𠂔	L4544	* 𠂔 *
U+1714C	𠂔	L1981 K1144	𠂔	L3517 K1144	𠂔		
U+185C9	𠂔	L1984	* 𠂔 *	L4554 K1258	𠂔		
U+17814	𠂔	L2027 K0969	𠂔	L2350	* 𠂔 *		
U+179F5	𠂔	L2078 K0007	𠂔	L2227	* 𠂔 *		
U+17249	𠂔	L2080 K5487	𠂔	L4023	* 𠂔 *		
U+17A22	𠂔	L2124 K5580	* 𠂔 *	L2455 K5606	* 𠂔 *	L3141	𠂔
U+17F21	𠂔	L2146	* 𠂔 *	L2175 K4505	𠂔		
U+179C2	𠂔	L2221 K0250	𠂔	L2807 K0250	𠂔		
U+17EF3	𠂔	L2252 K3562	𠂔	L2253 K3562	𠂔		
U+18640	𠂔	L2298 K5459	𠂔	L2299 K5459	𠂔		
U+17E7D	𠂔	L2342 K0213	𠂔	L2545	* 𠂔 *		
U+17821	𠂔	L2578 K5070	𠂔	L2608 K5071	* 𠂔 *		
U+17826	𠂔	L2592 K3324	𠂔	L2659 K3323	* 𠂔 *		
U+170EE	𠂔	L2597 K0294	𠂔	L2695	* 𠂔 *		
U+178BF	𠂔	L2610	* 𠂔 *	L2630 K0830	𠂔		
U+1797A	𠂔	L2619	* 𠂔 *	L3147 K4069	𠂔		
U+17E30	𠂔	L2622 K5436	* 𠂔 *	L2687 K1352	𠂔		
U+17EC5	𠂔	L2773 K3763	𠂔	L3137 K3750	* 𠂔 *		
U+17084	𠂔	L2831 K1409	𠂔	L4159 K1410	* 𠂔 *		
U+17072	𠂔	L2832 K1740	𠂔	L4162 K1741	* 𠂔 *		

Unicode		Lǐ Fànwén 1997 (Kyčanov 2006)					
Code	Glyph	Code	Glyph	Code	Glyph	Code	Glyph
U+17806	𪛦	L2840	* 𪛦 *	L4591 K1463	𪛦		
U+170BD	𪛧	L2921 K0014	𪛧	L2953	* 𪛧 *		
U+170AD	𪛨	L2922 K5345	* 𪛨 *	L2987 K1255	𪛨		
U+17076	𪛩	L2940 K0272	* 𪛩 *	L2957 K5716	𪛩		
U+172F0	𪛪	L3002 K0065	𪛪	L3524	* 𪛪 *		
U+1781B	𪛫	L3003 K4886	𪛫	L3112 K4912	* 𪛫 *		
U+1780B	𪛬	L3029 K1884	𪛬	L3225 K1885	* 𪛬 *		
U+1781A	𪛭	L3054 K1119	𪛭	L3325 K1121	* 𪛭 *		
U+1780C	𪛮	L3072 K0311	𪛮	L3363 K0312	* 𪛮 *		
U+173C3	𪛯	L3088 K5693	𪛯	L3233 K0686	* 𪛯 *		
U+173C1	𪛰	L3089 K5704	𪛰	L3350 K1960	* 𪛰 *		
U+17A66	𪛱	L3093 K5746	𪛱	L3181 K1692	* 𪛱 *		
U+17A6C	𪛲	L3096 K5751	𪛲	L3276 K2375	* 𪛲 *		
U+17A70	𪛳	L3098 K5715	𪛳	L3182 K0271	* 𪛳 *		
U+17841	𪛴	L3114 K4956	* 𪛴 *	L3191 K0962	𪛴		
U+17A17	𪛵	L3145 K5679	* 𪛵 *	L3326 K2358	𪛵		
U+173BD	𪛶	L3156 K0015	𪛶	L3271	* 𪛶 *		
U+17A03	𪛷	L3232 K1792	𪛷	L6000 K1792	𪛷		
U+17E12	𪛸	L3337 K2539	𪛸	L3338 K2539	𪛸		
U+17C9B	𪛹	L3396 K1200	𪛹	L3428 K1200	𪛹		
U+17E75	𪛺	L3435 K3388	𪛺	L3436 K3389	𪛺		
U+1819D	𪛻	L3444	* 𪛻 *	L3447 K1184	𪛻		
U+17085	𪛼	L3462 K5686	𪛼	L3496 K3239	* 𪛼 *		

Unicode		Lǐ Fànwén 1997 (Kyčanov 2006)					
Code	Glyph	Code	Glyph	Code	Glyph	Code	Glyph
U+1719F	𪛻	L3464 K5687	* 𪛻 *	L3476 K0561	𪛻		
U+17053	𪛻	L3466 K3996	𪛻	L4720 K3980	* 𪛻 *		
U+170A6	𪛻	L3467 K3961	𪛻	L4722 K3962	* 𪛻 *		
U+1711E	𪛻	L3488 K2976	𪛻	L3489 K2977	𪛻		
U+173C4	𪛻	L3558 K5706	𪛻	L3658 K2513	* 𪛻 *		
U+1783F	𪛻	L3560 (K5720)	* 𪛻 *	L3619 K0534	𪛻		
U+17A71	𪛻	L3561 K5722	𪛻	L3691 K1053	* 𪛻 *		
U+17F0D	𪛻	L3683 K2109	𪛻	L3684 K2109	𪛻		
U+17F50	𪛻	L3694 K3269	* 𪛻 *	L3709 K3183	𪛻		
U+17867	𪛻	L3797 K1421	𪛻	L3798 K1421	𪛻		
U+173DD	𪛻	L3814 K4982	* 𪛻 *	L3820 K0865	𪛻		
U+173D9	𪛻	L3822 K1560	𪛻	L3953 K1561	* 𪛻 *		
U+171BB	𪛻	L3837 K5434	* 𪛻 *	L3841 K1347	𪛻		
U+17F88	𪛻	L3877 K0987	𪛻	L3880	* 𪛻 *		
U+17ACD	𪛻	L3922 K3713	𪛻	L3930 K5591	* 𪛻 *		
U+17AC0	𪛻	L3944 K4995	* 𪛻 *	L3945 K4986	𪛻		
U+17AD6	𪛻	L3988	* 𪛻 *	L4009	𪛻		
U+17F98	𪛻	L4003 K3253	𪛻	L4005 K3251	* 𪛻 *		
U+1757E	𪛻	L4066 K1671	𪛻	L4069 K1671	𪛻		
U+1760C	𪛻	L4206 K2552	* 𪛻 *	L4383 K0446	𪛻		
U+17635	𪛻	L4224 K4994	* 𪛻 *	L4384 K4985	𪛻		
U+1773C	𪛻	L4260 K5482	𪛻	L4261 K5482	* 𪛻 *		
U+18647	𪛻	L4278	* 𪛻 *	L4423 K0176	𪛻		

Unicode		Lǐ Fànwén 1997 (Kyčanov 2006)					
Code	Glyph	Code	Glyph	Code	Glyph	Code	Glyph
U+1766A	𪛦	L4332 K2885	𪛦	L4333 K2885	𪛦		
U+17692	𪛨	L4408 K5267	𪛨	L4415 K1127	* 𪛨 *		
U+18650	𪛪	L4456 K3410	𪛪	L4457 K3411	𪛪		
U+177A6	𪛬	L4514 K4941	* 𪛬 *	L4534 K0766	𪛬		
U+17463	𪛭	L4807 K3206	𪛭	L4970 K3165	* 𪛭 *		
U+174F1	𪛮	L4852 K1142	𪛮	L5153 K1143	* 𪛮 *		
U+1743B	𪛯	L4862 K5749	* 𪛯 *	L4951 K2060	𪛯		
U+1741A	𪛰	L4863 K5744	* 𪛰 *	L4943 K1640	𪛰		
U+17498	𪛱	L4864 K5717	* 𪛱 *	L4944 K0276	𪛱		
U+17418	𪛲	L5001 (K0444)	𪛲	L5002 (K0444)	𪛲		
U+1826C	𪛳	L5096 K5336	𪛳	L5152 K1208	* 𪛳 *		
U+17B30	𪛴	L5110 K5677	𪛴	L5361	* 𪛴 *		
U+18358	𪛵	L5172 K0008	𪛵	L5231	* 𪛵 *		
U+184B5	𪛶	L5173 K0006	𪛶	L5232	* 𪛶 *		
U+17B3F	𪛷	L5174 K5490	* 𪛷 *	L5834	𪛷	L5835 K5463	𪛷
U+17B77	𪛸	L5185 K5258	𪛸	L5240 K1128	* 𪛸 *		
U+17B80	𪛹	L5187	𪛹	L5597	* 𪛹 *		
U+1824A	𪛺	L5190 K4448	𪛺	L5191 K4448	𪛺		
U+17B3B	𪛻	L5197 K5299	* 𪛻 *	L5841 K5292	𪛻		
U+17B57	𪛼	L5198 K5500	* 𪛼 *	L5842 K5498	𪛼		
U+1827C	𪛽	L5266 K5605	* 𪛽 *	L5506 K5581	𪛽		
U+184A3	𪛾	L5348 K2859	* 𪛾 *	L5777 K2854	𪛾		
U+18319	𪛿	L5373 K5435	* 𪛿 *	L5404 K1353	𪛿		

Unicode		Lǐ Fànwén 1997 (Kyčanov 2006)					
Code	Glyph	Code	Glyph	Code	Glyph	Code	Glyph
U+17B4E	𐰇	L5446 K4175	* 𐰇 *	L5912 K4169	𐰇		
U+17B07	𐰇	L5496 K4595	* 𐰇 *	L5510 K5671	𐰇		
U+18335	𐰇	L5729 K5128	* 𐰇 *	L5764 K1051	𐰇		
U+18711	𐰇	L5787 K5585	𐰇	L5794 K1395	* 𐰇 *		
U+186CB	𐰇	L5805 K5577	𐰇	L5863 K5593	* 𐰇 *		
U+1821E	𐰇	L5819 K0070	𐰇	L5826 K5592	* 𐰇 *		
U+17B37	𐰇	L5836 K5124	* 𐰇 *	L5845 K1027	𐰇		
U+17B40	𐰇	L5837 K5178	* 𐰇 *	L5846 K1108	𐰇		
U+17FF8	𐰇	L5878 K4631	* 𐰇 *	L5899 K4479	𐰇		
U+17FF7	𐰇	L5880 K2843	𐰇	L5881 K2844	𐰇		
U+1848D	𐰇	L5920 K4965	* 𐰇 *	L5925 K0852	𐰇		
U+17634	𐰇	L5944 K0023	𐰇	L5947 K1516	* 𐰇 *		

4. Kyčanov 2006

Kyčanov's recent Tangut-Russian-English-Chinese dictionary is a very important source, and we would expect any proposed repertoire of Tangut characters to be able to fully represent the contents of this dictionary. Unfortunately the version of UniTangut.txt that we have had access to does not provide any mappings to Kyčanov's dictionary, and so we have been unable to properly review the coverage of Kyčanov's dictionary. Nevertheless, it is clear that the currently proposed repertoire is not sufficient to represent all the characters in this dictionary, as 90 of the unified characters in Table 1 above are also in Kyčanov's dictionary. It is quite possible that Kyčanov also includes some other characters that are not included in the proposed Tangut repertoire (thusfar we have noticed that K4236 does not correspond exactly to any character in Lǐ Fànwén 1997 or the proposed repertoire—see entry for U+177D7 in Table 1). Mappings to Kyčanov 2006 should be supplied by the proposal's author at the earliest opportunity, so that coverage of this source can be more accurately assessed.

5. Lǐ Fànwén 1986

Table 2 lists all the unifications of characters in Lǐ Fànwén's 1986 study of the *Tóngyīn* 同音. The seven characters highlighted in red are significantly different from the character with which they are unified in UniTangut.txt, and should be encoded separately. Five of these seven characters do not occur in either Lǐ Fànwén 1997 or Kyčanov 2006.

Table 2 : Lǐ Fànwén 1986 Unifications

Unicode		Lǐ Fànwén 1986				Notes
Code	Glyph	Code	Glyph	Code	Glyph	
U+1711E	𐰇	0589	𐰇	0590	𐰇	

Unicode		Lǐ Fànwén 1986				Notes
Code	Glyph	Code	Glyph	Code	Glyph	
U+17418	𪗇	0975	𪗇	0976	𪗇	Possibly distinct, although not distinguished in Lǐ Fànwén 1997 (5001 and 5002)
U+177E1	𪗇	1439	𪗇	1440	* 𪗇 *	Distinguished in Lǐ Fànwén 1997 (1666 and 1667)
U+17B3F	𪗇	1495	𪗇	1496	𪗇	Possibly distinct, although not distinguished in Lǐ Fànwén 1997 (5834 and 5835)
U+17BA1	𪗇	1906	𪗇	3906	* 𪗇 *	3906 is not found in either Lǐ Fànwén 1997 or Kyčanov
U+1766A	𪗇	2345	𪗇	2385	𪗇	
U+1773C	𪗇	2587	𪗇	2588	𪗇	
U+18650	𪗇	2621	𪗇	2622	𪗇	
U+17867	𪗇	2816	𪗇	2817	𪗇	
U+17E75	𪗇	2905	* 𪗇 *	3348	𪗇	Distinguished in Sofronov (3308 and 3339), but not in Lǐ Fànwén 1997 (3435 and 3436) or Kyčanov (3388 and 3389)
U+1795E	𪗇	3056	𪗇	3141	𪗇	
U+17EF3	𪗇	3471	𪗇	3472	𪗇	
U+17F0D	𪗇	3491	𪗇	3492	𪗇	
U+17FF7	𪗇	3802	𪗇	3803	𪗇	
U+17CE6	𪗇	3870	𪗇	3871	𪗇	
U+17D8B	𪗇	3944	𪗇	3945	𪗇	
U+17CF0	𪗇	4333	𪗇	4334	* 𪗇 *	Distinguished in Lǐ Fànwén 1997 (0346 and 0347) and Sofronov (0138 and 1170)
U+17CED	𪗇	4339	𪗇	4340	𪗇	Distinguished in Lǐ Fànwén 1997 (0541 and 0542), Kyčanov (2887 and 2880), Sofronov (0153 and 1177) and Nishida (220-041 and 221-042)
U+17CF4	𪗇	4343	* 𪗇 *	4362	𪗇	Distinguished in Nishida (220-052 and 221-051), but not in Lǐ Fànwén 1997 (1871 and 1872)
U+18095	𪗇	4374	𪗇	4378	𪗇	

Unicode		Li Fànwén 1986				Notes
Code	Glyph	Code	Glyph	Code	Glyph	
U+180A1	𪗇	4387	𪗇	4388	𪗇	
U+17E12	𪗇	4499 * 𪗇 *	4500	𪗇		Distinguished in Sofronov (3277 and 3279), but not in Li Fànwén 1997 (3337 and 3338)
U+1824A	𪗇	4670	𪗇	4671	𪗇	
U+18342	𪗇	4805	𪗇	4840 * 𪗇 *		4840 is not found in either Li Fànwén 1997 or Kyčanov
U+18352	𪗇	4817	𪗇	4837	𪗇	
U+18408	𪗇	5537	𪗇	5538	𪗇	

6. Nishida 1966

Table 3 shows the three unifications of characters in Nishida's 1966 dictionary according to UniTangut.txt.

Table 3 : Nishida 1966 Unifications

Unicode		Nishida				Notes
Code	Glyph	Code	Glyph	Code	Glyph	
U+17EF3	𪗇	211-041	𪗇	211-043	𪗇	Li Fànwén 2252 and 2253 (same glyph) Kyčanov 3562
U+17CED	𪗇	220-041	𪗇	221-042	* 𪗇 *	Li Fànwén 0541 and 0542 (different glyphs) Kyčanov 2887 and 2880 (different glyphs)
U+17CF4	𪗇	220-052	𪗇	221-051	* 𪗇 *	Li Fànwén 1871 and 1872 (same glyph) Kyčanov 1415

220-041

𪗇 **š^wā «好い色」：正齒音類，獨字(40A7)，注(𪗇)
 𪗇 *lǐu «彩色，飾り」：文字要素𪗇，𪗇に分析できる(A₂)。注字と偏を對立する。

221-042

𪗇 *lǐu «彩色，飾り」：流風音類，小類 76 (50B7)，
 注(𪗇) 𪗇 *šā «好い色」：文字要素𪗇，𪗇に分析できる(A₂)。cf. 注字。

Figure 8 : Nishida 220-041 and 221-042

220-052

𪛗 *sefi «子細」：齒頭音類，小類 73 (31B5)，注(𪛗)
 𪛗 *tshu «粗い」：文字要素 𪛗，𪛗 に分析できる
 (A₂)。文字 𪛗 «小さい」を意符とする派生字

221-051

𪛗 *-tsifi «(綿)帽子」：齒頭音類，小類 23(30A1)，
 注(𪛗) 𪛗 *ma «綿(帽)」：文字要素 𪛗，𪛗 に分析で
 きる(A₂)。文字 𪛗 *tsi を音符とする派生字？。漢
 語‘帽子’よりの借用語を表記する(?)

Figure 9 : Nishida 220-052 and 221-051

Whereas 211-041/211-043 have the same glyph shape, and are therefore unifiable, 220-041/221-042 and 220-052/221-051 do not have the same glyph shapes, and are not unifiable. Nishida explicitly notes that the left-hand element of 220-041 and 220-052 is 𪛗 (Radical 220, 4 strokes), whereas the left-hand element of 221-042 and 221-051 is 𪛗 (Radical 221, 5 strokes).

Li Fànwén 1997 and Kyčanov 2006 follow Nishida in differentiating 220-041 and 221-042, although both Li Fànwén and Kyčanov use the same glyph form for 220-052 and 221-051. However, it is still necessary to encode 220-052 and 221-051 separately in order to be able to represent the usage in Nishida's dictionary.

7. Sofronov 1968

Table 4 shows the twenty-one unifications of characters in Sofronov 1968 according to UniTangut.txt. The eleven characters highlighted in red should not be unified, but encoded separately.

Table 4 : Sofronov 1968 Unifications

Unicode		Sofronov				Notes
Code	Glyph	Code	Glyph	Code	Glyph	
U+17CE6	𪛗	0021	𪛗	0022	𪛗	Different left-hand side : 𪛗 vs. 𪛗
U+17CF0	𪛗	0138	* 𪛗 *	1170	𪛗	
U+17CF4	𪛗	0147	* 𪛗 *	1174	𪛗	
U+17CED	𪛗	0153	* 𪛗 *	1177	𪛗	
U+17D8B	𪛗	0381	𪛗	0382	𪛗	
U+18408	𪛗	0465	𪛗	0466	𪛗	

Unicode		Sofronov				Notes
Code	Glyph	Code	Glyph	Code	Glyph	
U+1766A	𪛦	0722	𪛦	0840	* 𪛦 *	Different lower left : 𪛦 vs. 𪛦
U+18650	𪛰	0971	𪛰	0972	𪛰	
U+1711E	𪛮	1401	𪛮	1402	𪛮	
U+17B3F	𪛯	1674	𪛯	5474	* 𪛯 *	Different left-hand side : 𪛯 vs. 𪛯
U+17FF7	𪛱	1748	𪛱	5343	* 𪛱 *	Different left-hand side : 𪛱 vs. 𪛱
U+17418	𪛲	2016	* 𪛲 *	2029	𪛲	Different lower right : 𪛲 vs. 𪛲
U+1824A	𪛴	2131	* 𪛴 *	2636	𪛴	Different lower left : 𪛴 vs. 𪛴
U+17E12	𪛶	3277	𪛶	3279	* 𪛶 *	
U+17E75	𪛸	3308	* 𪛸 *	3339	𪛸	Sofronov classifies these under different radicals (cf. also Li Fànwén 1986)
U+17867	𪛺	3584	𪛺	3908	* 𪛺 *	Different rig hand side : 𪛺 vs. 𪛺
U+180A1	𪛼	4732	𪛼	4743	𪛼	
U+18095	𪛽	4766	𪛽	4767	𪛽	
U+177E1	𪛿	5351	𪛿	5352	𪛿	
U+17EF3	𪛻	5669	𪛻	5670	𪛻	
U+17F0D	𪛽	5705	𪛽	5706	𪛽	

8. Wen Hai

Table 5 shows the two unifications of characters in the *Sea of Characters* (*Wen Hai* 文海) according to UniTangut.txt.

Table 5 : Wen Hai Unifications

Unicode		Wenhai				Notes
Code	Glyph	Code	Glyph	Code	Glyph	
U+17E75	𐰇𐰆	10.253	𐰇𐰆	30.141	𐰇𐰆	Distinguished in Sofronov (3308 and 3339) and Lǐ Fànwén 1986 (2905 and 3348), but not in Lǐ Fànwén 1997 (3435 and 3436) or Kyčanov (3388 and 3389)
U+180A1	𐰇𐰆	16.262	𐰇𐰆	44.112	𐰇𐰆	

9. Future Tangut Additions

It is absolutely necessary that the initial set of Tangut characters to be encoded includes all graphically distinct characters in the major sources, such as Nishida 1966, Sofronov 1968, Lǐ Fànwén 1986, Lǐ Fànwén 1997 and Kyčanov 2006. But even if all such characters are added, there are many variant forms of Tangut characters used in contemporary Tangut sources and by modern scholars that are neither in the PDAM6.2 repertoire or in the standard Tangut dictionaries, and it is anticipated that such characters will need to be encoded in the future. Some examples of such characters are given below.



Figure 10 : Monumental Inscription dated 1348

The above is a famous monumental inscription from Dunhuang dated 1348 that has the mantra *om maṇi padme hūṃ* written in six different scripts (Lantsa, Tibetan, Uyghur, Phags-pa, Tangut and Chinese). The Tangut inscription reads :

𐰇𐰆𐰇𐰆𐰇𐰆𐰇𐰆𐰇𐰆

The first five characters are in the proposed Tangut repertoire (U+175F2, U+173B3, U+173EA, U+17C8F, U+175C1), but the last character is not in either the proposed repertoire or any of the major Tangut dictionaries. However, from context it is easily identifiable as a variant form of the character U+178FA 𐰇𐰆 (a transliteration character used to represent Sanskrit *hūṃ*). But although 𐰇𐰆 and 𐰇𐰆 are variant forms of the same character, the former cannot be considered to be a unifiable glyph variant of the other.



Figure 11 : Luo Fuyi 羅福頤, *Xi Xia Guanyin Huikao* 西夏官印匯考 [Collection of Official Seals of the Western Xia] (Yinchuan, 1982) No.7

The first character on the seal script inscription on the face of this Western Xia seal is transcribed by Li Fànwén into ordinary Tangut script as 𐰚, which is an otherwise unknown character.

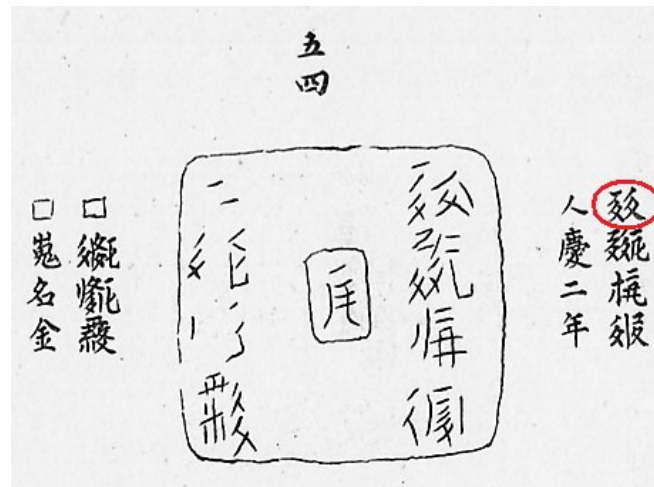


Figure 12 : Luo Fuyi, *Xi Xia Guanyin Huikao* [Collection of Official Seals of the Western Xia] No.54

The inscription on the back of this Western Xia seal gives the date of issue as the 2nd year of the Ren Qing 人慶 period (1145), but the Tangut character for Ren 人, which should be written 𐰚 is here written as 𐰚 (i.e. with the two components swapped horizontally and an additional dot),