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

L2/08-337

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. Kyčanov, 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 Zìdiǎn*). 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 and Lǐ Fànwén 3880 , which are unified as U+17F88 :

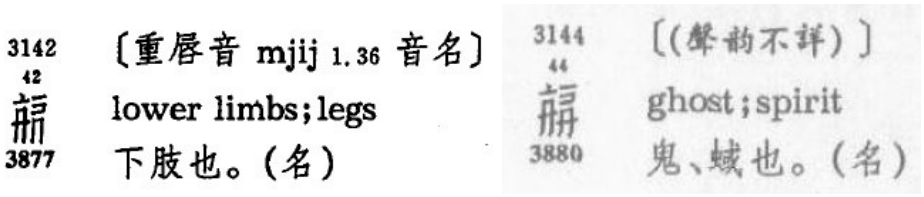


Figure 1 : Lǐ Fànwén 3877 and 3880

The implicit justification for this unification is Hán Xiǎománg 2004:

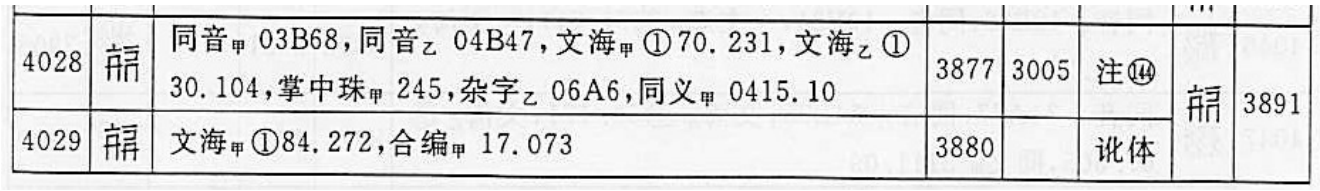


Figure 2 : Hán Xiǎománg 2004 page 231

Note 144 (page 341) referred to above simply states " is the correct form, is the corrupt form" 为正体，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) and Lǐ Fànwén 0542 (Kyčanov 2880) 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) :

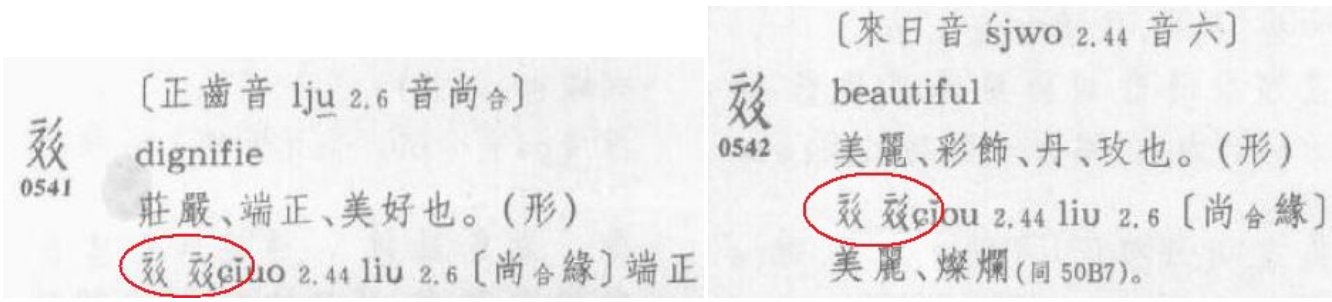
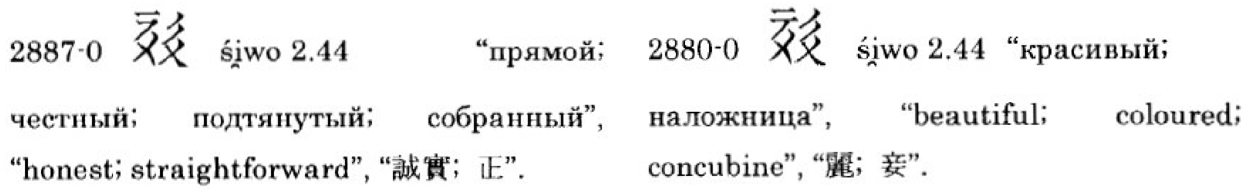


Figure 3 : Lǐ Fànwén 0541 and 0542



2887-0 ṣ̂jwo 2.44 "прямой; честный; подтянутый; собранный", "honest; straightforward", "誠實; 正". 2880-0 ṣ̂jwo 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 (Lǐ 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	[Tangut glyph]	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	[Tangut glyph]	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

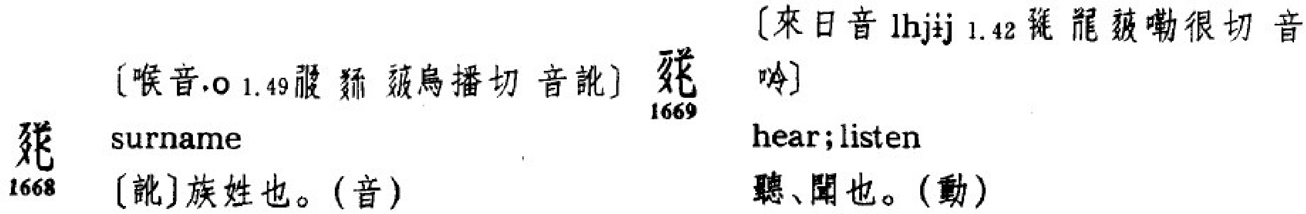
Another similar example is the pair of characters Lǐ Fànwén 1666 (left side element written as three strokes) and Lǐ Fànwén 1667 (left side element written as four strokes), which have different glyph shapes, different meanings and different reconstructed pronunciations, but which are unified as U+177E1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

1841 00 〔舌頭音 nə(聲调不詳)音能〕
fox
1666 狐也。(名)

1667 〔舌頭音 ta 1.17 都臘切 音怛〕
tail
①尾也。(名)

Figure 6 : Lǐ Fànwén 1666 and 1667

The difference in glyph shape between these two characters is exactly analagous 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):



〔喉音.o 1.49 烏播切 音訛〕
surname
1668 〔訛〕族姓也。(音)

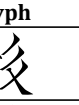
1669 〔來日音 lhjɨj 1.42 嘞很切 音吟〕
hear; listen
聽、聞也。(動)

Figure 7 : Lǐ 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		L0017 K4240		L0486 K4241	*  *		
U+1701C		L0042 K0998		L0215	*  *	L4537 K0998	
U+17022		L0057 K2389		L0232 K2388	*  *	L4548	
U+171D3		L0064 K5739	*  *	L1147 K5741		L3832 K5740	*  *
U+18172		L0076 K4026	*  *	L0786 K4027			
U+186B7		L0079 K5507	*  *	L0676 K5508			
U+175C0		L0162 K5394		L0202 K1334	*  *		
U+1817B		L0184		L1919	*  *	L4510 K4462	*  *
U+183E2		L0187 K4735		L0647 K4742	*  *		
U+17D15		L0213 K0279		L1957 K0280	*  *		
U+17D69		L0256 K5681	*  *	L0316			
U+180CD		L0259 K0005		L0345	*  *		
U+1731E		L0266 K5390		L0325	* *		
U+177D7		L0271 K3715	* *	L0534 K4235		K4236	* *
U+17DA4		L0276 K4858		L2099 K4857	* *		
U+17CF0		L0346 K2430 K2431		L0347	* *		
U+17D8B		L0406 K5297		L0407 K5297			
U+1756E		L0430 K0011		L0452	* *		
U+17247		L0431 K0012		L0453	* *		
U+180EE		L0490 K5604	* *	L0987 K5579			
U+17D14		L0514		L2388 K2668	* *		
U+18176		L0515 K2684		L2390	* *		
U+185CA		L0523 K3500		L4549 K3501	* *	L4565	* *

Unicode		Lǐ Fànwén 1997 (Kyčanov 2006)					
Code	Glyph	Code	Glyph	Code	Glyph	Code	Glyph
U+17D67		L0533 K5684	*  *	L0555			
U+17CED		L0541 K2887		L0542 K2880	*  *		
U+17525		L0583 K2776	*  *	L1362 K2775			
U+17BC3		L0608 K5594	*  *	L1878 K5578			
U+1817C		L0644 (K4579)		L2553	*  *		
U+1817F		L0671 K0287		L2572 K0288	*  *	L4578 K0295	*  *
U+17248		L0735		L1521 K3418	*  *		
U+17D3C		L0895 K5566	*  *	L4491 K5560			
U+17D13		L1039 K1652		L2943	*  *		
U+171E8		L1068 K3207		L3827 (K3206)	*  *		
U+18408		L1106 K3448		L1107 K3448			
U+17D12		L1131 K0597		L2985	*  *		
U+1817D		L1134 K0610		L2988	*  *	L4624	* *
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		L1947 K0212	*  *	L2354 K0149			
U+1704E		L1960 K5721		L1975 K1039	*  *		
U+1704D		L1972 K1931	*  *	L2971 K1930		L4544	*  *
U+1714C		L1981 K1144		L3517 K1144			
U+185C9		L1984	*  *	L4554 K1258			
U+17814		L2027 K0969		L2350	*  *		
U+179F5		L2078 K0007		L2227	*  *		
U+17249		L2080 K5487		L4023	*  *		
U+17A22		L2124 K5580	*  *	L2455 K5606	*  *	L3141	
U+17F21		L2146	*  *	L2175 K4505			
U+179C2		L2221 K0250		L2807 K0250			
U+17EF3		L2252 K3562		L2253 K3562			
U+18640		L2298 K5459		L2299 K5459			
U+17E7D		L2342 K0213		L2545	* *		
U+17821		L2578 K5070		L2608 K5071	* *		
U+17826		L2592 K3324		L2659 K3323	* *		
U+170EE		L2597 K0294		L2695	* *		
U+178BF		L2610	* *	L2630 K0830			
U+1797A		L2619	* *	L3147 K4069			
U+17E30		L2622 K5436	* *	L2687 K1352			
U+17EC5		L2773 K3763		L3137 K3750	* *		
U+17084		L2831 K1409		L4159 K1410	* *		
U+17072		L2832 K1740		L4162 K1741	* *		

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

---

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		0589		0590		

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

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

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		211-041		211-043		Lǐ Fànwén 2252 and 2253 (same glyph) Kyčanov 3562
U+17CED		220-041		221-042	*  *	Lǐ Fànwén 0541 and 0542 (different glyphs) Kyčanov 2887 and 2880 (different glyphs)
U+17CF4		220-052		221-051	*  *	Lǐ Fànwén 1871 and 1872 (same glyph) Kyčanov 1415

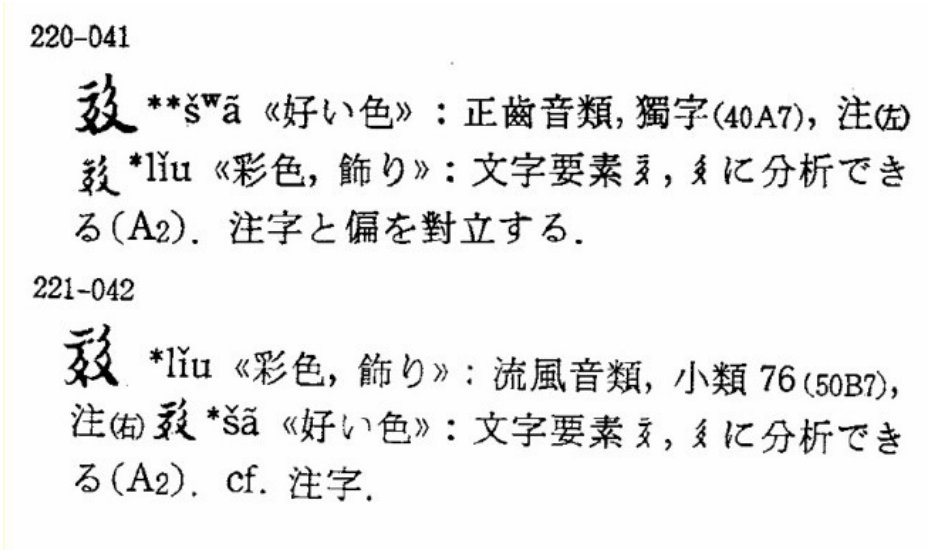


Figure 8 : Nishida 220-041 and 221-042

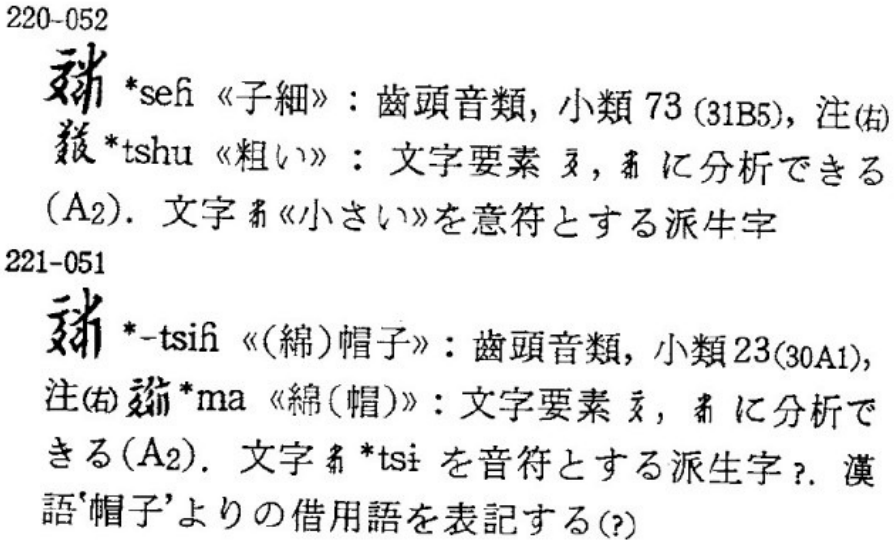


Figure 9 : Nishida 220-052 and 221-051

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 ⿰ (Radical 220, 4 strokes), whereas the left-hand element of 221-042 and 221-051 is ⿰ (Radical 221, 5 strokes).

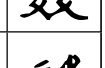
Lǐ Fànwén 1997 and Kyčanov 2006 follow Nishida in differentiating 220-041 and 221-042, although both Lǐ Fànwén and Kyčanov use the same glyph form for 220-052 and 221-051. However, it is still necessary to encode 220-052 and 221-051 separately in order to be able to represent the usage in Nishida's dictionary.

---

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		0021		0022		
U+17CF0		0138	*  *	1170		Different left-hand side : ⿰ vs. ⿰
U+17CF4		0147	*  *	1174		
U+17CED		0153	*  *	1177		
U+17D8B		0381		0382		
U+18408		0465		0466		

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 Lǐ 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

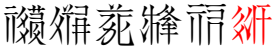
Unicode		Wenhai				Notes
Code	Glyph	Code	Glyph	Code	Glyph	
U+17E75		10.253		30.141		Distinguished in Sofronov (3308 and 3339) and Lǐ Fànwén 1986 (2905 and 3348), but not in Lǐ Fànwén 1997 (3435 and 3436) or Kyčanov (3388 and 3389)
U+180A1		16.262		44.112		

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 *oṃ maṇi padme hūṃ* written in six different scripts (Lantsa, Tibetan, Uyghur, Phags-pa, Tangut and Chinese). The Tangut inscription reads :



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 𗣺 (a transliteration character used to represent Sanskrit *hūṃ*). But although 𗣿 and 𗣺 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 Lǐ Fànwén into ordinary Tangut script as 彔, which is an otherwise unknown character.

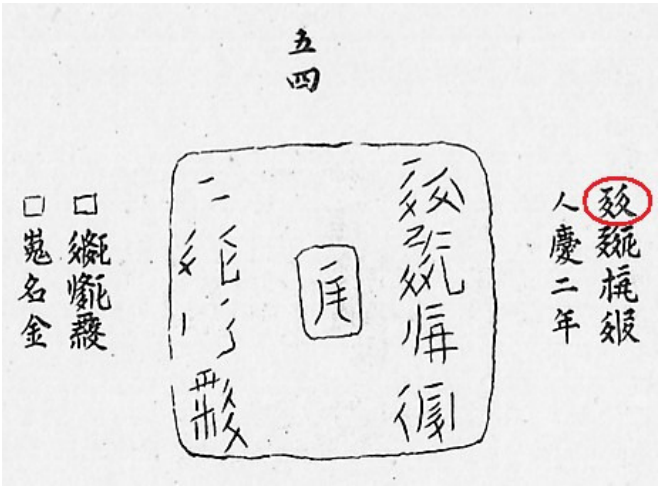


Figure 12 : Luo Fuyi, *Xi Xia Guanyin Huikao* [Collection of Official Seals of the Western Xia] No.54

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 𗍫 is here written as 𗍬 (i.e. with the two components swapped horizontally and an additional dot),