

ISO/IEC JTC1/SC2/WG2 N3244

Title	Review of CJK-C Repertoire
Source	UK National Body
Document Type	National Body Contribution
Date	2007-04-14

The UK national body has carried out a review of the CJK-C repertoire, and would like to draw to the attention of WG2 a number of characters which may be unifiable with existing characters. Analysis of the component structure and component features of the characters in the table below suggests that the pairs of characters may be unifiable, and should be unified unless they are determined to be non-cognate pairs.

Possible Unifiable Characters

CJK-C Character				Existing Character			Annex S Unification Examples
Codepoint	Source Reference	Glyph	Reading	Codepoint	Glyph	Reading	
U+2A744	TC-367C	俊	jun	U+3453	俊	jùn	U+9AEA 髮 = U+9AEE 髮 U+62D4 拔 = U+629C 拔
U+2A988	TC-553A	媯	kua	U+2177B	媯	kuā	U+6C5A 汚 = U+6C61 汚
U+2ACF5	TD-4D43	梲	gao	U+69D4	梲	gāo	U+5932 卒 = U+672C 本
U+2B017	TD-4161	磽	qiang	U+254CC	磽	qiǎng	U+79BF 秃 = U+79C3 秃
U+2B1B4	TC-5E70	聃	pan	U+2655E	聃		U+514C 兌 = U+5151 兑 U+6C5A 汚 = U+6C61 汚
U+2B6B8	TE-435A	𩇛	bie	U+9C49	𩇛	biē	U+514C 兌 = U+5151 兑

The readings for the above CJK-C characters are all taken from IRG N1232:

- http://www.cse.cuhk.edu.hk/~irg/irg26/IRGN1232_C1_evidences_TCA-Part1.pdf
- http://www.cse.cuhk.edu.hk/~irg/irg26/IRGN1232_C1_evidences_TCA-Part2.pdf
- http://www.cse.cuhk.edu.hk/~irg/irg26/IRGN1232_C1_evidences_TCA-Part3.pdf

Note that the definition in the TCA evidence for all six of these CJK-C characters is "character used in Person's name".

Notes

U+2A744 U+3453

俊俊

The components 友 and 发 are often unifiable (e.g. the pairs U+62D4 拔 & U+629C 拔 and U+9AEA 髮 & U+9AEE 髮 given in Annex S), but in some cases they are used in pairs of non-cognate characters, and so are not unifiable (e.g. the pair U+8307 芟 and U+82C3 芟 given in Annex S). There are a number of possible duplicate CJK-B characters with the same type of component variation as U+2A744/U+3453:

- U+216C0 媠 vs. U+59AD 媠 (bá)
- U+221EC 废 vs. U+221F7 废 (bá)
- U+2504A 盍 vs. U+76CB 盍 (bō)
- U+26413 斐 vs. U+7FC7 斐 (fú)
- U+29D7C 鮒 vs. U+9B81 鮒 (bà, bō)

U+2A988 U+2177B

媠媠

Example of characters from CJK-B and CJK-C where the righthand components of U+2A988 and U+2177B 媠 have been unified:

U+28706

28706 57042 57042 1273.271 63778.060 邑 8	G_HZ T6-4564	媠 媠 媠 媠
---	-----------------	------------

Super CJK Version 14.0 page 1729

U+2A746

2A746	倭 TC-4375	倭 V04-4126
--------------	--------------	---------------

ISO/IEC 10646 PDAM4 (SC2 N3914) page 30

U+2ACF5 U+69D4

𣎵 𣎵

The righthand component of U+2ACF5 is unifiable with the righthand component of U+69D4 𣎵 (the element 𣎵 is a common hand-written form of 本). For example, the compatibility ideograph U+2F988 𣎵 is canonically equivalent to U+267B5 𣎵 (the font used does not show the difference, but in the code charts U+2F988 is written with 𣎵白本, whereas U+267B5 is written with 𣎵白𣎵).

U+2B017 U+254CC

𣎵 𣎵

The righthand component of U+2B017 only occurs in one encoded character, U+22BB2 𣎵, and is probably a glyph variant of the common phonetic element U+7F8C 羌 (qiāng).

U+2B1B4 U+2655E

𣎵 𣎵

The righthand components of U+2B017 and U+2655E are both very uncommon, only occurring in four encoded characters, U+2514E 𣎵, U+2655E 𣎵, U+28232 𣎵 and U+298A3 𣎵. From a component structure perspective, the righthand components of U+2B017 and U+2655E should be unifiable as it normally does not matter whether the top two strokes are written pointing up or pointing down, and whether the vertical stroke passes through both horizontal strokes or not (cf. U+2A988).

U+2B6B8 U+9C49

𣎵 𣎵

The bottom component of U+2B6B8 (U+29D4B 𣎵) is a calligraphic glyph variant of the fish radical (U+9B5A 魚), as shown in the following examples:



There is only one encoded character that uses this form of the fish radical, U+29E3A 𩺰.

Appendix

The six characters mentioned in the report above were selected from a computer generated list of possible duplicates. The search was made based on the IDS of each character and was deliberately fuzzy. The result of only 14 hits for 4,219 characters was very low — a similar search of 940 characters submitted to the IRG for Extension D produced 27 hits.

A table showing the 14 hits is given below, in which the first two hits though obviously spurious give some insight into the fuzziness of the search.

Matches from a Fuzzy IDS Search on the CJK-C Repertoire

CJK-C Character		Fuzzy IDS Matches				
U+2A702	𠄎	𠄎 U+4E4A	𠄎 U+5140	𠄎 U+51E2	𠄎 U+51E3	𠄎 U+5C22
U+2A721	𠄎	𠄎 U+4EA2	六 U+51CD			
U+2A744	俊	俊 U+3453				
U+2A956	奄	奄 U+5944	奄 U+FA7E			
U+2A988	嫖	嫖 U+2177B				
U+2AAF6	怵	怵 U+6035				
U+2ACF5	榘	榘 U+69D4				
U+2AE3F	炆	炆 U+241A6				
U+2AEEF	獺	獺 U+24814	獺 U+2F927			

CJK-C Character		Fuzzy IDS Matches	
U+2B017	磽	磽 U+254CC	
U+2B1B4	聒	聒 U+2655E	
U+2B523	钋	钋 U+9492	
U+2B556	閤	閤 U+9581	閤 U+9582
U+2B6B8	鰲	鰲 U+9C49	

Notes

1. Although U+2AEEF and U+24814 share a unifiable component (cf. U+508A 偵 vs. U+202CF 偵 and other examples in IRG N1155), they are non-cognate according to the Kangxi Dictionary (U+2AEEF is defined as a vulgar form of U+7328 , whereas U+24814 is defined as a vulgar form of U+733F 猿), and so have been excluded from the list of possible unifiable characters.
2. U+2A956 occurs as a component in one encoded character U+29100 𪛀 (an archaic variant of U+9748 靈), which is not cognate with the very similar U+290D7 𪛀. Therefore, despite the correspondence in pronunciation between U+2A956 (yan according to the evidence for TC-3953) and U+5944 奄 (yān, yǎn), they are probably non-cognate.