

Universal Multiple-Octet Coded Character Set
UCS

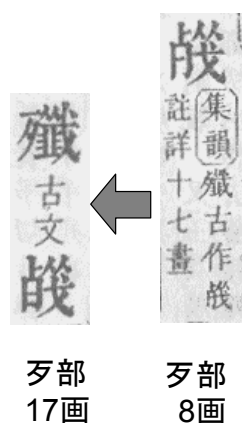
ISO/IEC JTC 1/SC 2/WG 2 IRG N1859

Date: 2012-06-07

Source:	SUZUKI Toshiya, Faculty of Integrated Arts and Science, Hiroshima University
Title:	Proposal for the Discussion How to Handle the Mistakenly Differentiated Glyphs in Huge Dictionaries
Status:	Individual Contribution
Distribution:	IRG Members
Medium :	Electronic

Abstract

In G_K characters proposed to CJK E, some glyphs were suspected to be mistakenly designed by the Kangxi editors, but the glyphic differences cannot be handled by the precedents, like, U+23A26 (GKX-0582.04) versus G_K584.12.



Because these characters (or glyphs) are often used for the digitization of the dictionaries, the separated encoding of them may introduce the inconsistency between the referential glyphs and their descriptions, and incompatibilities among the digitization before & after of CJK E. It is expected that the submitters check whether the separately encoding of them are necessary or not, with concrete use cases.

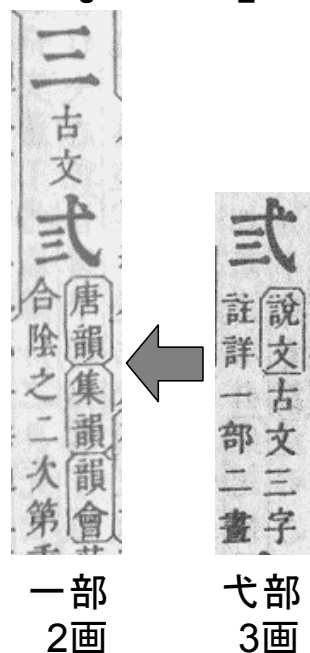
Proposal for the Discussion How to Handle the Mistakenly Differentiated Glyphs in Huge Dictionaries

suzuki toshiya, Hiroshima University, Japan

Abstract

In G_K characters proposed to CJK E, some glyphs were suspected to be mistakenly designed by the Kangxi editors, but the glyphic differences cannot be handled by the precedents. Because these characters (or glyphs) are often used for the digitization of the dictionaries, the separated encoding of them may introduce the inconsistency between the referential glyphs and their descriptions, and incompatibilities among the digitization before & after of CJK E. It is expected that the experts and the submitters discuss about how to handle the mistakenly differentiated glyphs, with some concrete usecases.

Background of G_KX and G_K glyphs

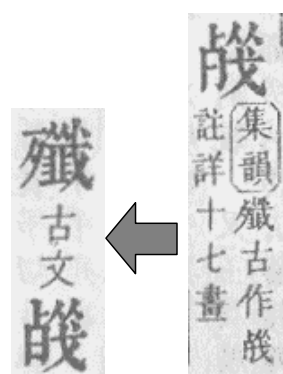


The CJK B characters taken from Kangxi Zidian are tagged as G_KX, and it is expected that most Kangxi characters are already coded by CJK B. But a bunch of Kangxi characters was proposed to CJK C, with G_K tag (and now IRG is working for them to include them in CJK E). It seems that G_KX characters were taken from the individual item, and the characters in the supplement are not proposed. It seems that G_K characters are proposed to improve the coverage of Kangxi characters. They are taken from the alternative form list or the supplement.

In Kangxi, some characters have “ancient forms” (古文) after the standard character. Some of the ancient forms are based on different origins, thus they don’t include the radical of the standard character. In such case, Kangxi shows “ancient form” glyph in 2 places; one place is in the list after the standard character, another place is in the radical that the “ancient form” glyph seems to have. For example, “弌” is found in radical “一” (as ancient form of “三”) and

radical “弌” as an individual item. In the description of such duplicated characters, just the source and the current form are informed, and the detailed description is omitted.

Basically these characters should have same glyphs, but sometimes they have different shapes and suspected to be mistakenly differentiated, as U+23A26 (GKX-0582.04) versus G_K584.12. By checking original reference, it is possible to detect which glyph is mistakenly designed. Several examples that are already found are summarized in the end of this document.



Concern on the Impact of the Disunification of the Mistakenly Designed Variants

Most “ancient variants” are never used widely, thus their shapes have not been stabilized (thus Kangxi editors had designed them inconsistently). Because of the lack of stabilized shapes, the glyphic differences between existing G_KX and proposed G_K are sometimes difficult to unify

殳部
17画

殳部
8画

by the precedents. In fact, some versions of Kangxi do not distinguish them. Taking an example “𢦏” (U+23A26), following evidences are found. Even for the publishing of Kangxi in different formats, these differences are not regarded as the shapes to be distinguished.

- The handwritten copy of Kangxi in Siqu Quanshu uses U+23A26 shape at the place corresponding to G_K0584.12
- The digitally typesetted version of Kangxi published by Shanghai Lexicographical Publishing House uses U+23A26 shape at the place corresponding to G_K0584.12
- The digitally typesetted index added to Kangxi published by Zhonghua Bookstore includes U+23A26 but not G_K0584.12

If these shape differences are regarded as different base characters, the impact may introduce the incompatibility among the existing and future versions of digitized dictionaries.

Proposed Item for the Discussion

Because the glyphic differences are not found in the precedent unifications, it is questionable if they should be dealt as generally unifiable difference. In the case of the difference between U+21156 versus G_K1612.27, the difference looks like “名” and “各” and difficult to take as generally unifiable. Therefore, “when two variants are found to be caused by the mistake in the dictionary compilation, and their difference is difficult to take as generally unifiable, how the variants should be handled?” is expected to be discussed by the experts. The possible options would be following:

- (A) apply non-generic unification, and code at CJK Compatibility Ideograph.
- (B) apply non-generic unification, and register the shape in IVD.
- (C) classify the evidence taken from the dictionary as unreliable, and postpone the discussion until yet-another evidence to justify the separated encoding of the variants.
- (D) code the variants separately, but make a record that these variants are cognate.

Considering that there are many users who don't care the original reference of CJK Ideographs (i.e. most users may use the character because their shapes are looking like what they want. The users thinking as “this is taken from Kangxi, so I use”, “this is not found in Kangxi, so I should avoid” would not be majority), the option (C) would be most tolerant, because it just postpones the decision.

The option (D) looks like as if it is the easiest, but it may request the information on existing characters, like, Extension B which was standardized without the evidences. Some members may have the difficulty to excavate the source materials of them.

In summary, my proposed option is (C), and I want to receive the feedback about which option is the best for the submitters, with the concrete use cases.

(end of document)

Investigation on G_K Glyphs and the Glyph Shapes in Referenced Materials

CJK E G_K glyph	Evidence for CJK E	Reference in Evidence	Similar character	Existing	Evidence of Existing character	Scanned images from references in evidence
𦵑 G_K1393.07		唐韻、 集韻、 韻會、 說文、 広韻、 他	𦵑 GKX-0505.05 U+23344			
𦵒 G_K1393.26		唐韻、 集韻、 韻會、 正韻、 說文、 広韻、 集韻、 玉篇、 字彙、 五音集韻、 他	𦵒 GKX-0559.06 U+23756			
𦵓 G_K0589.07		唐韻、 広韻、 集韻、 類篇、 韻會、 他	𦵓 GKX-1054.15 U+26E15			Not found in 大広益会玉篇・宮内庁本、澤存堂本

 G_K0349.26		唐韻、広韻、集韻、類篇、韻會、玉篇、他	𪛗 GKX-0152.34 U+20919		 (大広益会玉篇・宮内庁本)  (大広益会玉篇・元刊本)  (大広益会玉篇・澤存堂本)  (集韻・北京図書館本)
 G_K0081.14		他 唐韻、集韻、韻會、正韻、玉篇、字彙、正字通、	𪛘 GKX-0158.21 U+209E4		 (集韻・北京図書館本)  (四声篇海・萬曆本)  (字彙補)
 G_K0584.12		他 唐韻、広韻、集韻、類篇、韻會、	𪛙 GKX-0582.04 U+23A26		 (集韻・北京図書館本)  (類篇・汲古閣本)

<div data-bbox="113 141 253 239" data-label="Text"> <div>𩇛</div> <div>G_K1612.27</div> </div>	<div data-bbox="288 125 397 389" data-label="Text"> <div>𩇛</div> <div>五音篇海 名養切</div> </div>	<div data-bbox="501 125 534 239" data-label="Text"> 五音篇海 </div>	<div data-bbox="571 125 762 295" data-label="Text"> <div>𩇛</div> <div>GHZ-20867.10 U+21156</div> </div>	<div data-bbox="962 125 1299 333" data-label="Text"> <div>𩇛</div> <div>名養反 (龍龕手鏡・高麗本)</div> </div> <div data-bbox="962 367 1299 575" data-label="Text"> <div>𩇛</div> <div>名養切 (五音篇海・萬曆本)</div> </div> <div data-bbox="962 609 1412 725" data-label="Text"> <div>𩇛</div> <div>mǎng 《龍龕手鑑・名部》：“𩇛，名養反。” 漢語大字典</div> </div>
<div data-bbox="113 766 253 864" data-label="Text"> <div>𩇛</div> <div>G_K0369.24</div> </div>	<div data-bbox="288 745 397 1160" data-label="Text"> <div>復</div> <div>古文復復</div> </div>	<div data-bbox="453 745 534 1919" data-label="Text"> 他 唐韻、集韻、韻會、正韻、說文、正字通、書彙典、書說命、禮曲禮、周札天官、諸葛亮出師表 </div>	<div data-bbox="571 745 748 918" data-label="Text"> <div>𩇛</div> <div>GKX-0152.02 U+208F8</div> </div>	<div data-bbox="821 745 936 1391" data-label="Text"> <div>𩇛</div> <div>集韻 𩇛 古作𩇛 玉篇 同 𩇛 今文作復 註見 彳部 九畫</div> </div> <div data-bbox="962 745 1412 909" data-label="Text"> <div>𩇛</div> <div>上同 (大庋益会玉篇・宮内庁本)</div> </div> <div data-bbox="962 943 1152 1099" data-label="Text"> <div>𩇛</div> <div>上同 (元刊本)</div> </div> <div data-bbox="962 1120 1206 1261" data-label="Text"> <div>𩇛</div> <div>上同 (澤存堂本)</div> </div> <div data-bbox="962 1276 1361 1727" data-label="Text"> <div>𩇛</div> <div>說文 𩇛 古文重也 (集韻・北京圖書館本)</div> </div> <div data-bbox="962 1749 1158 1926" data-label="Text"> <div>𩇛</div> <div>說文 𩇛 (正字通)</div> </div>

Korea JTC1/SC2, Committee on Character Codes

Author: JEONG, Hyeongdo; KIM, Kyongsok

Date: 2012. 11. 06.

Status: National Body Position, Rep. of KOREA

Subject: comments RE: IRG N1859, mistakenly differentiated glyphs in
dictionaries

1. Background

1.1 a related document

ISO/IEC JTC 1/SC 2/WG 2 IRG N1859


Date: 2012-06-07

Source: SUZUKI Toshiya,

Title: Proposal for the Discussion How to Handle the Mistakenly
Differentiated Glyphs in Huge Dictionaries

1.2 Relevant portion extracted from IRG N1859

...

Taking an example “” (U+23A26), following evidences are found. Even for the publishing of Kangxi in different formats, these differences are not regarded as the shapes to be distinguished.

- *The handwritten copy of Kangxi in Siqu Quanshu uses U+23A26 shape at the place corresponding to G_K0584. 12*

- *The digitally typesetted version of Kangxi published by Shanghai Lexicographical Publishing House uses U+23A26 shape at the place corresponding to G_K0584. 12*

- *The digitally typesetted index added to Kangxi published by Zhonghua Bookstore includes U+23A26 but not G_K0584. 12 If these shape differences are regarded as different base characters, the impact may introduce the incompatibility among the existing and future versions of digitized dictionaries.*

Proposed Item for the Discussion

Because the glyphic differences are not found in the precedent unifications, it is questionable if they should be dealt as generally unifiable difference. In the case of the difference ... difficult to take as generally unifiable. Therefore, “when two variants are found to be caused by the mistake in the dictionary compilation, and their difference is difficult to take as generally unifiable, how the variants should be handled?” is expected to be discussed by the experts. The possible options would be following:

- (A) apply non-generic unification, and code at CJK Compatibility Ideograph.
- (B) apply non-generic unification, and register the shape in IVD.
- (C) classify the evidence taken from the dictionary as unreliable, and postpone the discussion until yet-another evidence to justify the separated encoding of the variants.
- (D) code the variants separately, but make a record that these variants are cognate.

...

In summary, my proposed option is (C), and I want to receive the feedback about which option is the best for the submitters, with the concrete use cases.

..

		他 唐韻、 廣韻、 集韻、 類篇、 韻會、	 GKX-0582.04 U+23A26		 (集韻・北京圖書館本)  (類篇・汲古閣本)
---	---	-----------------------------------	---	---	--

2. ROK comments

- In principle, ROK agrees with Mr. Suzuki's proposed option C.
- ROK suggests that "yet-another evidence" be elaborated as follows:

1) Glyphs in other dictionaries are not to be considered as yet-another evidence in general.

- It seems desirable that actual usage in ordinary books or documents (other than dictionaries) is considered as yet-another evidence.

2) However, suppose that none of two mistakenly differentiated glyphs in a Hanzi dictionary (for example, Kangxi) are encoded in UCS (this situation will occur rarely); furthermore, no actual usage other than dictionaries exists,

In this situation, if we want to encode one of them in UCS, then we can probably accept usage in other dictionaries as yet-another evidence.

* * *