IRGN2211	
Title	Proposal to Reduce Amount of "Unnecessary Variants".
Author	Henry Chan
Date	2017/06/11
Type	Individual Contribution to IRG

Recommendation M64.11 (Review of CJK Unification Rules):

WG2 recommends that IRG reviews its CJK unification rules to minimize the number of glyph variants that are coded as separate characters.

It is suggested that the following solutions be implemented:

A. Expanding UCV to cover more cases.

This is covered by IRGN2176.

B. Rejection of Certain Classes of Glyphs

Currently, IRG faces three classes of glyphs which I believe should not be encoded as new CJK Unified Ideographs. They should be encoded as variants **via IVD** if it is necessary to reproducing their glyph shape.

1) Unorthodox Transcription: Geometrically strict transcription of Oracle Bone, Bronze Inscriptions, Seal Script and Clerical Script when a universally accepted / common form already exists:

Geometrically strict transcriptions of pre-Kaishu characters should not be accepted into IRG without any proof that such transcription is well accepted and common.

For sources backed by Seal Script / Clerical Script glyphs, we should regard that the authoritativeness of the evidence only applies to the glyph in the non-Kaishu script. The Kaishu shape as a transcription of the original script should require further proof that it is indeed well accepted and common.

Example 1:

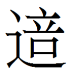
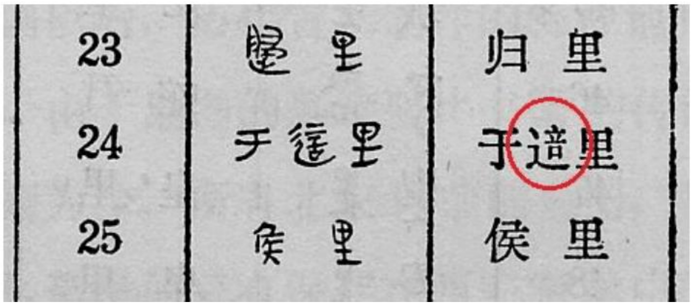
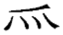
<table>
<tr><td rowspan="3">04290</td><td colspan="2">辵 162.8</td><td rowspan="3"></td><td rowspan="3"></td><td rowspan="3"></td><td rowspan="3"></td><td rowspan="3">
UTC-01478 (UK)</td><td rowspan="3"></td></tr>
<tr><td colspan="2">⿺辶音</td></tr>
<tr><td>點</td><td>N/A</td></tr>
</table>

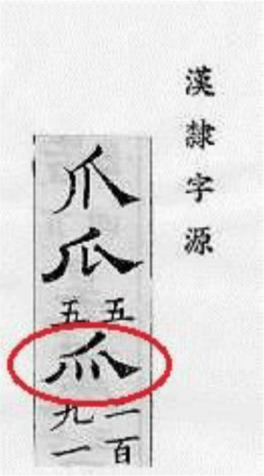
Fig. 168. ***Guditu Lunwenji*** **(1977) p. 11**

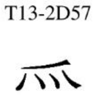

23
归 里
24
于 適 里
25
侯 里

This glyph should not be accepted for encoding because it is a geometrically strict transcription of the character in the middle of row 24, but the more common transcription is 這 U+9019. Encoding of this glyph as a separate character would generate confusion because present day 音 and 言 are different in historical derivation. There is no evidence to suggest that the transcription of UTC-01478 is authoritative.

Example 2:

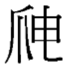
<table>
<tr><td rowspan="3">02305</td><td colspan="2">爪 87.1</td><td rowspan="3"></td><td rowspan="3">
T13-2D57</td><td rowspan="3"></td><td rowspan="3"></td><td rowspan="3"></td><td rowspan="3"></td></tr>
<tr><td colspan="2">⿻爫丶</td></tr>
<tr><td>折</td><td>N/A</td></tr>
</table>

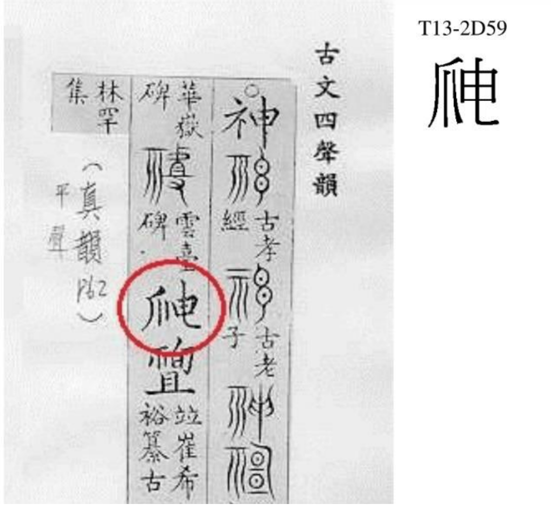

漢隸字源
五五
九一
一百



Evidence only proves existence of a four dot variation of 爪 in clerical script. It does not prove that T12-2D57 is a Kaishu glyph that exists naturally; it is a geometrically script transcription by the MOE Dictionary. The evidence should not be accepted as sufficient proof. Furthermore, the common transcription of “claw” is 爪. T13-2D57 should not be accepted as a new CJK Unified Ideograph.

Example 3:

<table>
<tr><td rowspan="3">02309</td><td colspan="2">爪 87.5</td><td rowspan="3"></td><td rowspan="3">
T13-2D59</td><td rowspan="3"></td><td rowspan="3"></td><td rowspan="3"></td><td rowspan="3"></td></tr>
<tr><td colspan="2">⿰爪电</td></tr>
<tr><td>横</td><td>N/A</td></tr>
</table>

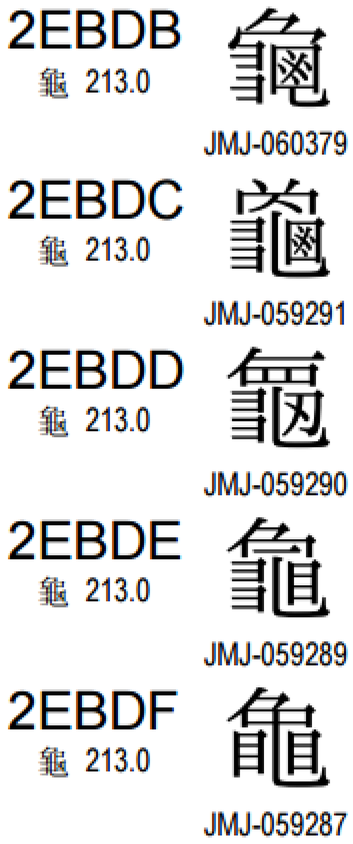

古文四聲韻
神
古孝經
古老子
華嶽碑
雲臺碑
崔希裕纂古
林罕集
真韻附
平聲
T13-2D59

This is a transcription of some kind of ancient glyph (崔希裕纂古). There is no evidence of the original glyph that this transcription was based on, nor is there any proof to suggest that this transcription is well accepted or common.

There is no evidence to suggest that god (神) can be composed of “Claw 爪” nor is the use of 电 a common variation of 申 when written on the right. It is highly probable that this “Claw 爪” is an alternative transcription of 𡭕/ 𫝀 (U+21B55), which is a strict transcription of 示 (U+793A).

A more common transcription of god is 神 (U+795E) and a stricter, less common transcription is 𥙍 (U+2564D). T13-2D59 should be rejected for encoding as a new CJK Unified Ideograph, and should be unified with either 神 (U+795E) or 𥙍 (U+2564D).

Example 4:



2EBDB
龜 213.0
JMJ-060379

2EBDC
龜 213.0
JMJ-059291

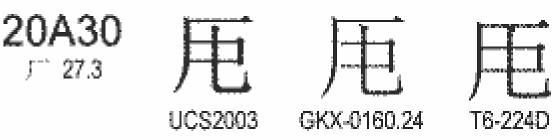
2EBDD
龜 213.0
JMJ-059290

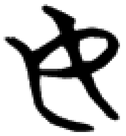
2EBDE
龜 213.0
JMJ-059289

2EBDF
龜 213.0
JMJ-059287

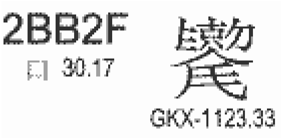
These are all various transcripted forms of the same character (龜). Only 龜 (U+9F9C) needs to be encoded, all other transcripted forms should not be coded as a CJK Unified Ideograph but via IVD.

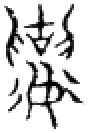
Example 5:

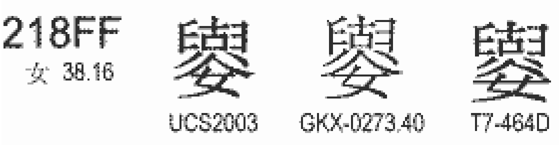

20A30
厂 27.3
UCS2003
GKX-0160.24
T6-224D

This is Kangxi’s unorthodox transcription of the ancient form of 女 (U+5973). The original form in 簡帛文字 () should be encoded somewhere instead, while the transcripted form (U+20A30) could have been encoded via IVD to 女 (U+5973). Of course, IRG usually determines the presence of Kangxi letterheads as “too significant to unify”

Example 6:


2BB2F
口 30.17
GKX-1123.33

This is Kangxi’s unorthodox transcription of the ancient form of 襄 (孃) – ().
A more common transcription is:


218FF
女 38.16
UCS2003
GKX-0273.40
T7-464D

Thus, U+2BB2F needed not be encoded as a CJK Unified Ideograph, but should be encoded via IVD to U+218FF. Of course, IRG usually determines the presence of Kangxi letterheads as “too significant to unify”.

--

At this stage, I think it is not necessary for IRG to define what is regarded as “Orthodox Transcription”. Instead, submitters should determine the “Orthodox Transcription” based on their respective national standards (such as that used for education purposes, or follow the “traditional” glyph shape), and back up their glyph by other authoritative sources if challenged.

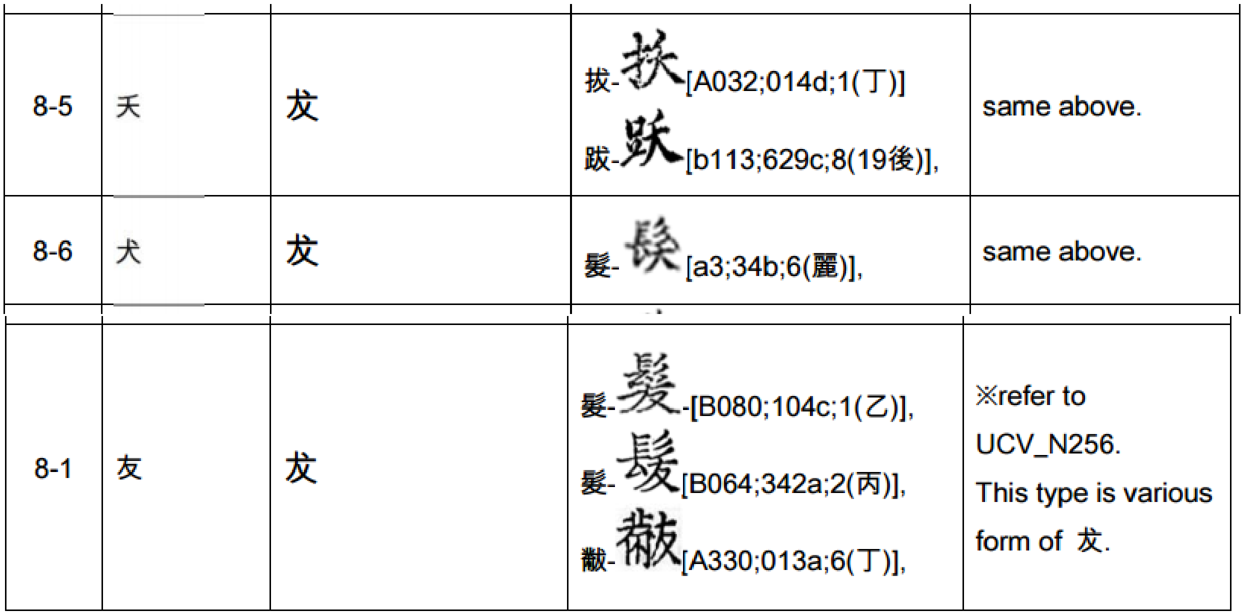
2) Pre-normalized forms: Systematic corruptions of similar-shaped, historically-unrelated components

These types of characters are well covered by IRGN2154 ROK Glyph Normalization, which deals with the normalization of handwritten characters. Handwriting is a rather “at-will” process where similar components are often exchanged with each another. Such variations may be systematically found for a given dynasty or regional area.

In IRGN2154, ROK’s glyph normalization applies only to the handwriting script. However, many historical printed materials are carved by typesetters who directly preserved the variations in the handwriting. As such, the normalization rules should be applied to the printed forms too.

Glyphs using these swapped out components should be rejected from being encoding as new characters. Only their normalized form should be encoded.

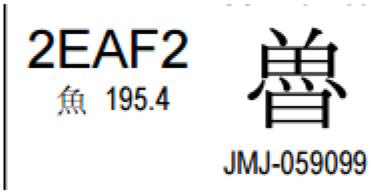
Example:



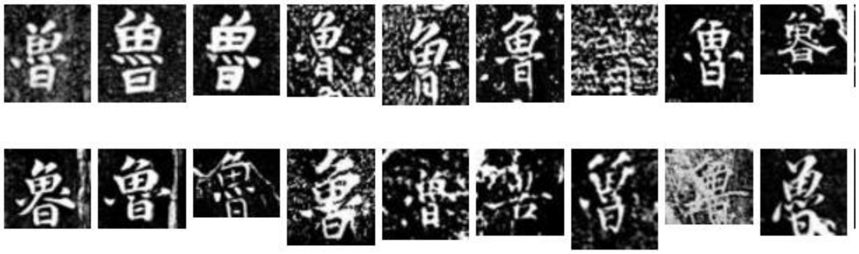
8-5	夭	犮	拔-[A032;014d;1(丁)] 䟦-[b113;629c;8(19後)],	same above.
8-6	犬	犮	髮-[a3;34b;6(麗)],	same above.
8-1	友	犮	髮--[B080;104c;1(乙)], 髮-[B064;342a;2(丙)], 黻-[A330;013a;6(丁)],	※refer to UCV_N256. This type is various form of 犮.

犬、犮、夭、友 are historically unrelated components but are swapped with each other very often. ISO10646 should not encode every single permutation. Only the normalized form should be encoded.

Example 2:



U+2EAF2 should not be encoded, because it is a strict transcription of a handwritten form of 魯:



(from http://coe21.zinbun.kyoto-u.ac.jp/djvuchar?query=%E9%AD%AF)

Sometimes, the first two strokes of 魯 become separated too far and the direction of the first stroke is changed such that it becomes “overturned 八”.

The corruption between ク and overturned 八 is a common systematic occurrence. Thus, U+2EAF2 is corrupted form of a variant Kaishu form of 魯, and thus should not be encoded as a CJK Unified Ideograph, but via IVD to 魯 (U+9B6F).

If the normalized form cannot be determined, then the form can be encoded as-is as a new character. However, submitters should not try to encode the pre-normalized form as new character if the normalized form can be easily inferred from the context, based on semantic analysis or other versions. They should be encoded via IVD instead.

These rules should not be treated as a traditional “unification” rules because they are not variation across “C/J/K/V/T/H/KP/V/M” regions, but generally variations across time.

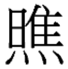
3) Errors: One-off corruptions found on tombstone carvings

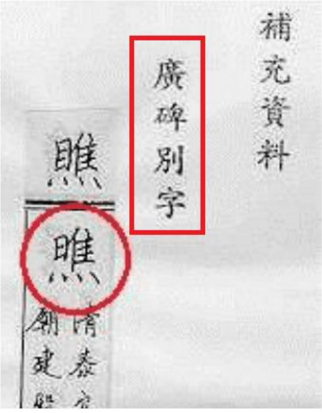
One-off corruptions found on the tombstone carvings should not be coded into ISO10646. ISO10646 is an industrial standard to encode meaningful characters. It should not a dumping ground for glyphs which appear once or twice in dictionaries intended to collect all sorts of obscure and ill-written “characters”. It is very common for tombstone carvings to contain unique-but-wrong characters since it is impossible to correct the mistakes made by the tombstone carver.

To reflect this rule, tombstone glyphs (primary sources) and dictionaries which draw from tombstones (secondary sources), such as 《廣碑別字》、《碑別字新編》、《偏類碑別字》etc should be deemed insufficient proof of “new character” if no evidence of the exact same glyph from other sources can be presented. Evidences from dictionaries which quote these sources (tertiary sources such as 《中華字海》) should not be deemed as sufficient proof either.

Assuming we regard the source as authoritative, we should regard the written relationship between the corrupted glyph and the “normalized” glyph as authoritative evidence that the two glyphs are the same character, and thus eligible for IVD (given that number of components is the same).

Example 1 - Missing horizontal stroke in a basic radical:

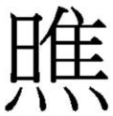
02270	火 86.12			 TE-6F6B				
	⿱睢灬							
	竪	N/A						



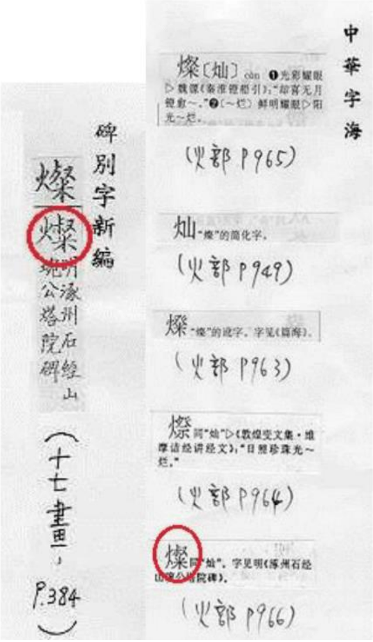
補充資料

廣碑別字

TE-6F6B



Example 2 – Swapped out component: 㕛 (phonetic of phonetic component) → 叔



碑別字新編

燦

燦

涿州石經山琬公塔院碑

（十七畫，P.384）

中華字海

燦〔灿〕càn ❶光彩耀眼▷魏徵《[illegible]》："[illegible]无月[illegible]～。"❷〔～烂〕鲜明耀眼▷阳光～烂。

（火部 P965）

灿 "燦"的简化字。

（火部 P949）

燦 "燦"的讹字。字见《篇海》。

（火部 P963）

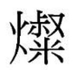
燦 同"灿"▷《敦煌变文集·维摩诘经讲经文》："日照珍珠光～烂。"

（火部 P964）

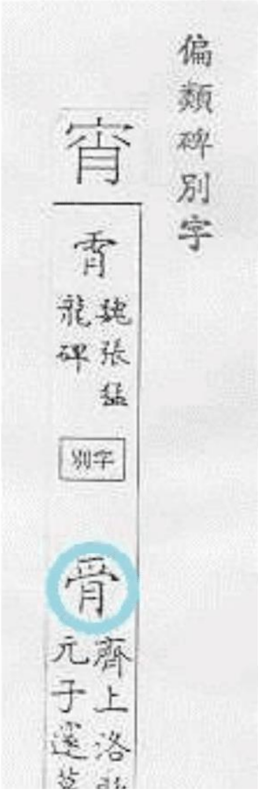
燦 同"灿"。字见明《涿州石经山[illegible]公[illegible]院碑》。

（火部 P966）

T13-2D55



Example 3 – Various forms of severe corruption:



偏類碑別字

胥

胥

魏張猛龍碑

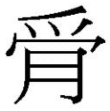
別字

胥

齊上洛

元子邃

T13-2D5A



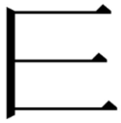
C. Rejection of Non-Han characters

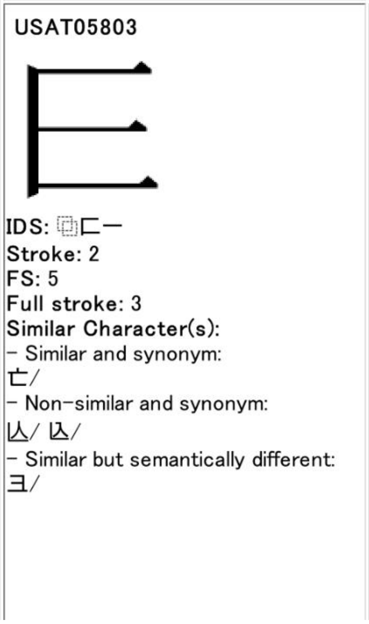
Non-characters should be rejected from encoding as CJK Unified Ideographs. They should be coded in the proper ISO10646 blocks.

1) Character Components

Geometrically decomposed components should be encoded as components in a block such as CJK Radicals Supplement or other Extension. They should not be encoded as normal ideographs as they are semantically not ideographs, and since they generally do not have any pronunciation they will not be typeable in conventional IMEs in China or Japan.

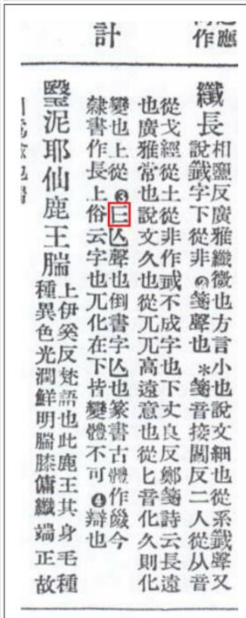
Examples:

00005	一 1.2 ⿻匚一 折 N/A				 USAT05803		disunified to 00005, UTC provides more evidence of 00003, irg 48. pending unified to 00003, irg46.



USAT05803

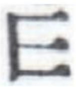
IDS: ⿻匚一
Stroke: 2
FS: 5
Full stroke: 3
Similar Character(s):
– Similar and synonym:
亡/
– Non-similar and synonym:
亾/ 亾/
– Similar but semantically different:
ヨ/



計

纖長相鹽反廣雅纖微也方言小也說文細也從糸韱聲又說韱字下從非の韱聲也＊韱音接閻反二人從从音從戈經從土從非作戜不成字也下丈良反鄭箋詩云長遠也廣雅常也說文久也從兀兀高遠意也從匕音化久則化變也上從③⿻匚一亾聲也倒書字亾也篆書古體作镾今隸書作長上俗云字也兀化在下皆變體不可④辯也

毉泥耶仙鹿王腨上伊奚反梵語也此鹿王其身毛種種異色光潤鮮明腨膝傭纖端正故



This character is a purely geometrical decomposition of 長 (the top part) and has no meaning or pronunciation whatsoever. It should not be eligible to be encoded as CJK Unified Ideograph.

There are past examples of such components being encoded. I believe these past examples are doing the wrong thing and such practice should not be continued.

Side note: the synonym information of USAT05803 provided by the submitter is not correct.

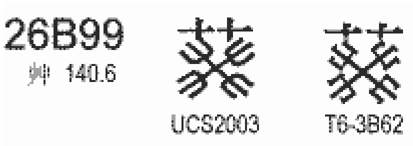
2) Kana ligatures

Kana ligatures such as tomo 𪜈 (Ligature of ト (to) and モ (mo)) should not be encoded as CJK Unified Ideographs, as they are of unrelated script (Kana-script vs Han-script).

However, I have no opinion to the encoding Han-Kana / Han-Latin Ligatures as CJK Unified Ideographs.

3) Kai-Seal ligatures

Kai-Seal ligatures should not be encoded. (Actually, this can also fit into the case of “Unorthodox Transcription”.) Either the Seal script character is encoded in the Shuowen Block (and/or via IVD), or only the Kai form is encoded, or the character be encoded via IVD. Example which should not be encoded as CJK Unified Ideograph:


26B99
艸 140.6
UCS2003
T6-3B62

There are already multiple existing cases of encoding such characters into CJK Unified Ideographs. In retrospect, I would question the validity of encoding those such characters. The nature of these characters should be well analyzed before their inclusion into ISO10646.