IRG N2176 (v2 modified 2016/10/20)

Title: New UCV Suggestions

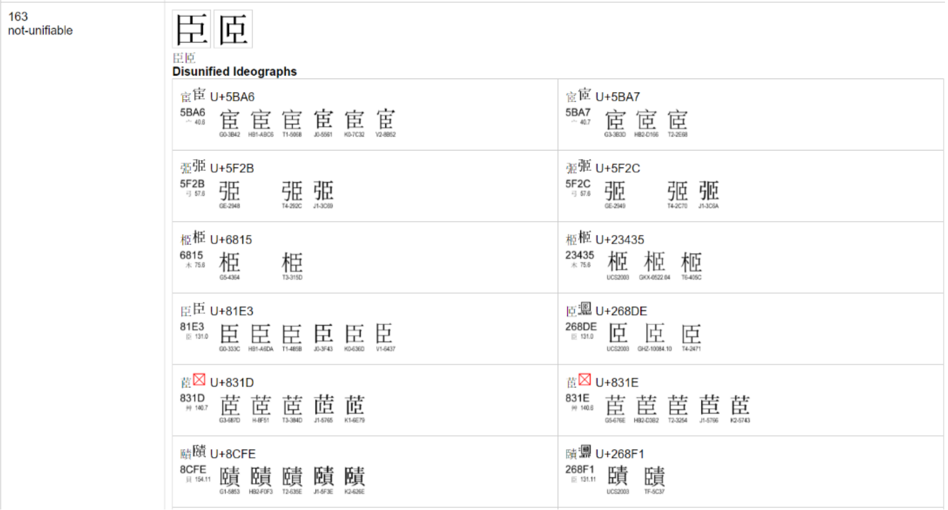
Author: Henry Chan

Type: Individual Contribution to IRG #47

Issue 1A

𦣞 vs 臣

According the current NUCV, this pair cannot be unified:

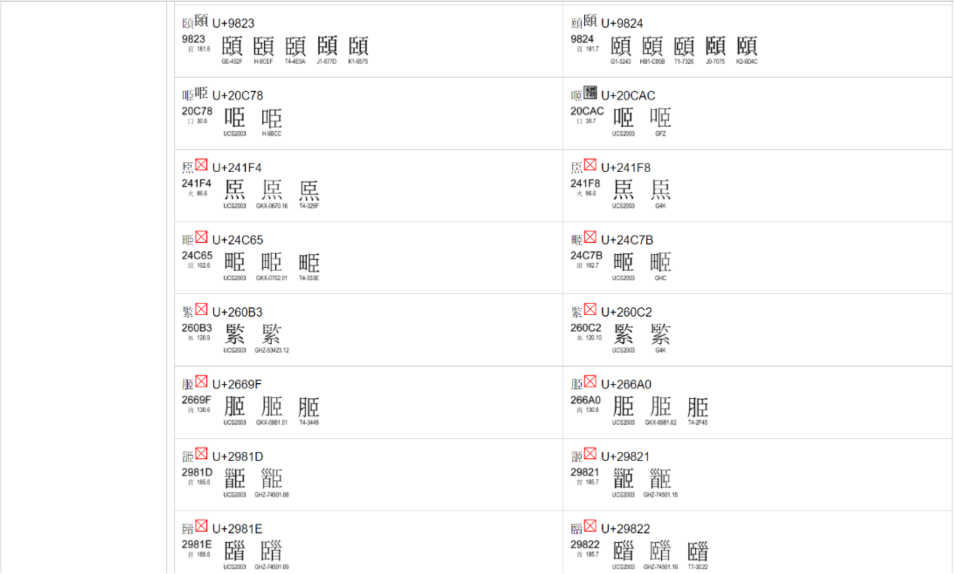


163
not-unifiable

臣 𦣞

Disunified Ideographs

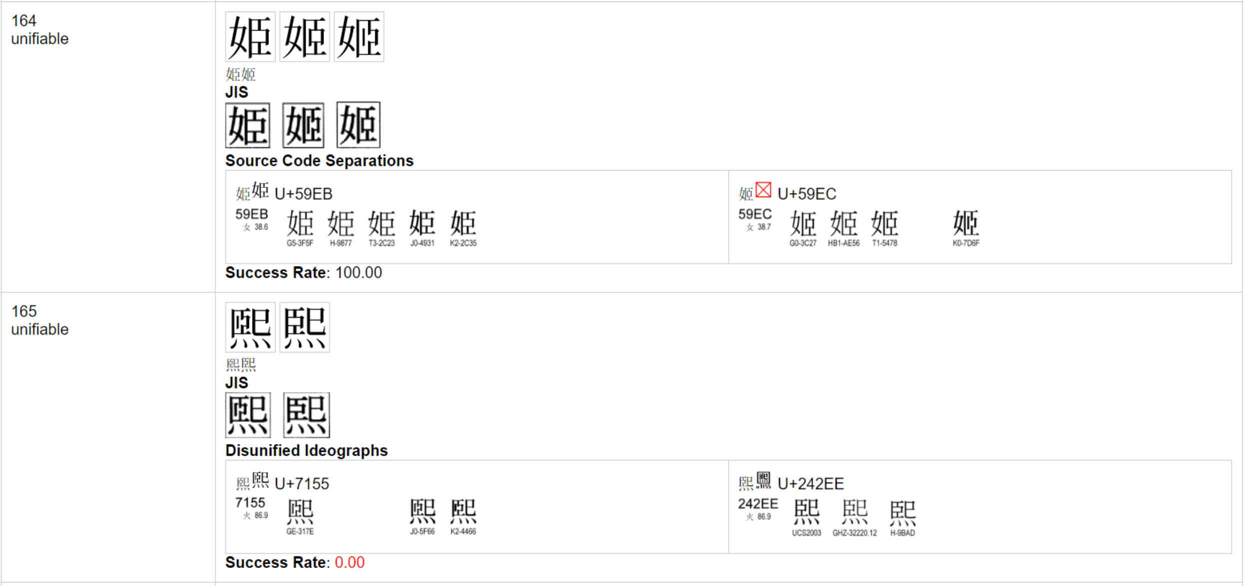
宦 U+5BA6 5BA6 宀 40.6 宦 宦 宦 宦 宦 宦 G0-3B42 HB1-ABC6 T1-5068 J0-5561 K0-7C32 V2-8B52	宧 U+5BA7 5BA7 宀 40.7 宧 宧 宧 G3-3B3D HB2-D166 T2-2E68
弫 U+5F2B 5F2B 弓 57.6 弫 弫 弫 GE-2948 T4-292C J1-3C69	弬 U+5F2C 5F2C 弓 57.6 弬 弬 弬 GE-2949 T4-2C70 J1-3C6A
栕 U+6815 6815 木 75.6 栕 栕 G5-4364 T3-315D	𣐵 U+23435 23435 木 75.6 𣐵 𣐵 𣐵 UCS2003 GKX-0522.04 T6-405C
臣 U+81E3 81E3 臣 131.0 臣 臣 臣 臣 臣 臣 G0-333C HB1-A6DA T1-485B J0-3F43 K0-636D V1-6437	𦣞 U+268DE 268DE 臣 131.0 𦣞 𦣞 𦣞 UCS2003 GHZ-10084.10 T4-2471
茝 ☒ U+831D 831D 艸 140.7 茝 茝 茝 茝 茝 G3-687D H-8F51 T3-384D J1-5765 K1-6E79	茞 ☒ U+831E 831E 艸 140.6 茞 茞 茞 茞 茞 G5-676E HB2-D3B2 T2-3254 J1-5766 K2-5743
賾 U+8CFE 8CFE 貝 154.11 賾 賾 賾 賾 賾 G1-5853 HB2-F0F3 T2-635E J1-5F3E K2-626E	𦣱 U+268F1 268F1 臣 131.11 𦣱 𦣱 UCS2003 TF-5C37



頣 U+9823 9823 頁 181.6 頣 頣 頣 頣 頣 GE-452F H-9CEF T4-4E5A J1-677D K1-6575	頤 U+9824 9824 頁 181.7 頤 頤 頤 頤 頤 G1-5243 HB1-C05B T1-7326 J0-7075 K2-6D4C
𠱸 U+20C78 20C78 口 30.6 𠱸 𠱸 UCS2003 H-9BCC	𠲬 U+20CAC 20CAC 口 30.7 𠲬 𠲬 UCS2003 GFZ
𤇴 ☒ U+241F4 241F4 火 86.6 𤇴 𤇴 𤇴 UCS2003 GKX-0670.16 T4-326F	𤇸 ☒ U+241F8 241F8 火 86.6 𤇸 𤇸 UCS2003 G4K
𤱥 ☒ U+24C65 24C65 田 102.6 𤱥 𤱥 𤱥 UCS2003 GKX-0762.01 T4-333E	𤱻 ☒ U+24C7B 24C7B 田 102.7 𤱻 𤱻 UCS2003 GHC
𦂳 ☒ U+260B3 260B3 糸 120.9 𦂳 𦂳 UCS2003 GHZ-53423.12	𦃂 ☒ U+260C2 260C2 糸 120.10 𦃂 𦃂 UCS2003 G4K
𦚟 ☒ U+2669F 2669F 肉 130.6 𦚟 𦚟 𦚟 UCS2003 GKX-0981.01 T4-3445	𦚠 ☒ U+266A0 266A0 肉 130.6 𦚠 𦚠 𦚠 UCS2003 GKX-0981.02 T4-2F45
𩠝 ☒ U+2981D 2981D 首 185.6 𩠝 𩠝 UCS2003 GHZ-74501.08	𩠡 ☒ U+29821 29821 首 185.7 𩠡 𩠡 UCS2003 GHZ-74501.15
𩠞 ☒ U+2981E 2981E 首 185.6 𩠞 𩠞 UCS2003 GHZ-74501.09	𩠢 ☒ U+29822 29822 首 185.7 𩠢 𩠢 𩠢 UCS2003 GHZ-74501.16 T7-3E22

In reality, only U+268F1 is cognate with U+8CFE. (U+24C7B, U+2981D, U+29821, U+2981E and U+29822 cannot be verified). The other characters are non-unifiable.

Also, following this rule, there are also rules 164 and 165 which indicate when 𦣝 or 臣 are in ［⿰女〓］ or ［⿱⿰〓巳灬］, they are unifiable:



164
unifiable

姬姬

JIS

Source Code Separations

U+59EB
59EB
女 38.6
G5-3F5F H-9877 T3-2C23 J0-4931 K2-2C35

U+59EC
59EC
女 38.7
G0-3C27 HB1-AE56 T1-5478 K0-7D6F

Success Rate: 100.00

165
unifiable

熈煕

JIS

Disunified Ideographs

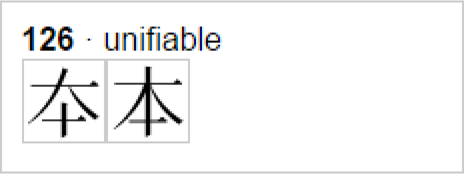
U+7155
7155
火 86.9
GE-317E J0-5F66 K2-4466

U+242EE
242EE
火 86.9
UCS2003 GHZ-32220.12 H-9BAD

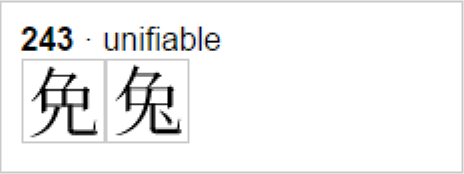
Success Rate: 0.00

𦣝 and 臣 are completely different in pronunciation, so if two characters containing these component, it should be easy to identify if they are non-cognate. In handwriting, the two components are somewhat frequently messed up. Such a rule may unnecessarily hinder the unification of “improper” variants.

This is similar to these other UCV rules in nature:



126 · unifiable



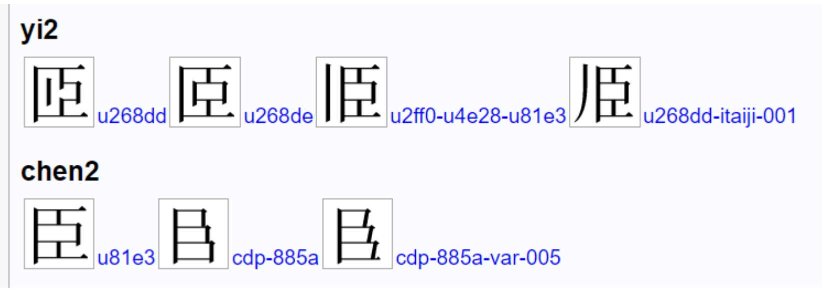
243 · unifiable

Their use in characters is often as a phonetic component, but it is easy to distinguish.

It is suggested that this rule be removed from the NUCV.

Issue 1B

Per handwriting conventions, it is suggested that these variants be (independently) added to the UCV:



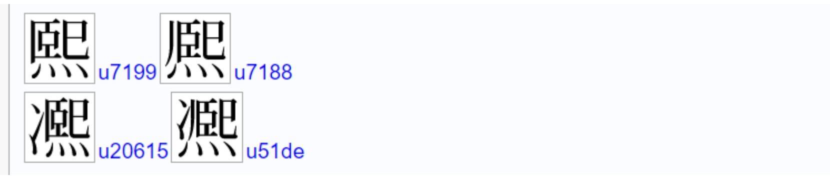
yi2

𦣝 u268dd 𦣞 u268de ⿰丨臣 u2ff0-u4e28-u81e3 u268dd-itaiji-001

chen2

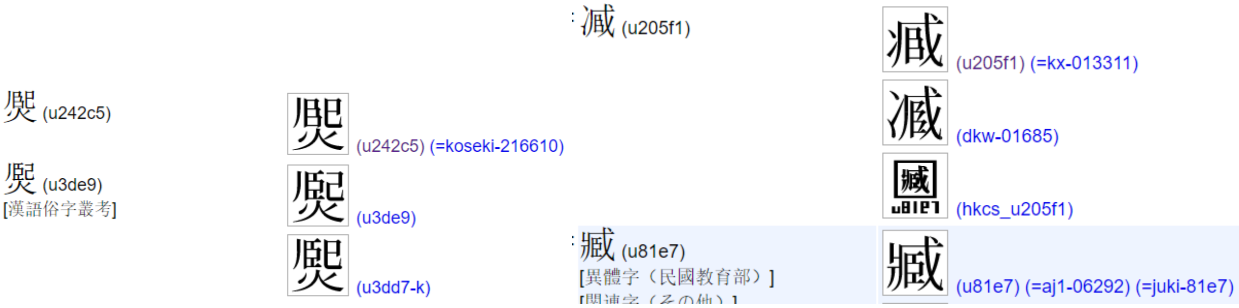
臣 u81e3 cdp-885a cdp-885a-var-005

Existing disunified characters are as follows:



熙 u7199 熈 u7188

𠘕 u20615 凞 u51de



𤋅 (u242c5)

㷩 (u3de9)
[漢語俗字叢考]

𤋅 (u242c5) (=koseki-216610)

㷩 (u3de9)

㷗 (u3dd7-k)

𠗱 (u205f1)

臧 (u81e7)
[異體字（民國教育部）]
[関連字（その他）]

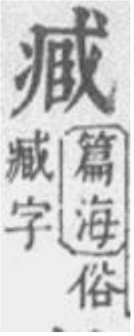
𠗱 (u205f1) (=kx-013311)

(dkw-01685)

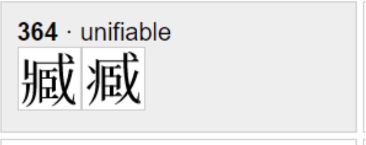
(hkcs_u205f1)

臧 (u81e7) (=aj1-06292) (=juki-81e7)

The Kangxi entry for U+205f1 is:



which indicates they are cognate, and could have been unified, when complemented by rule 364:

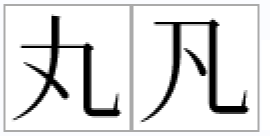
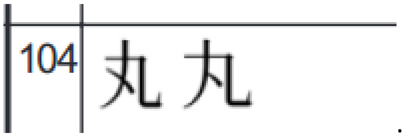


364 · unifiable

臧 𠗱

This rule substitutes UCV rule 163, 164, 165, and 167, and complements rule 364.

Issue 2

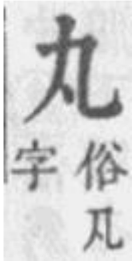
Suggest merging  into rule #104: 

104 丸 丸

.

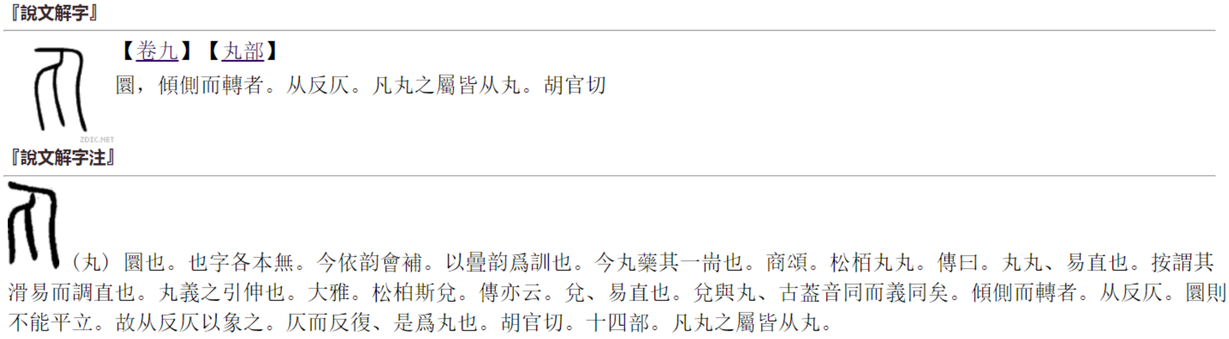
Rationale

U+2007D is the ancient form of U+4E38. According to Kangxi dictionary, Kangxi thinks that 丸 is a 俗體 (corrupted form) of 𠁽:



丸
俗𠁽
字

The origin of the ancient shape is from seal script, where the “人” does not protrude out of the container:



『說文解字』

【卷九】【丸部】

圜，傾側而轉者。从反仄。凡丸之屬皆从丸。胡官切

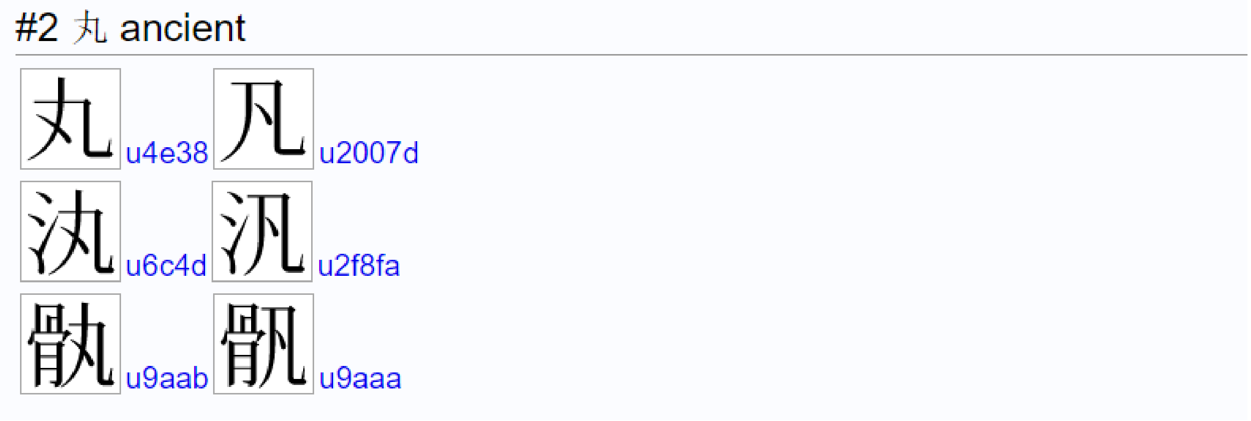
『說文解字注』

（丸）圜也。也字各本無。今依韵會補。以疊韵爲訓也。今丸藥其一耑也。商頌。松桷丸丸。傳曰。丸丸、易直也。按謂其滑易而調直也。丸義之引伸也。大雅。松柏斯兌。傳亦云。兌、易直也。兌與丸、古蓋音同而義同矣。傾側而轉者。从反仄。圜則不能平立。故从反仄以象之。仄而反復、是爲丸也。胡官切。十四部。凡丸之屬皆从丸。

The shape of U+2007D 𠁽 is easily confused with U+51E1 凡. As such, as mentioned in IRGN2174 Part 1, U+2F8FA has been mistakenly unified to U+6C4E 汎 instead of U+6C4D 汍.

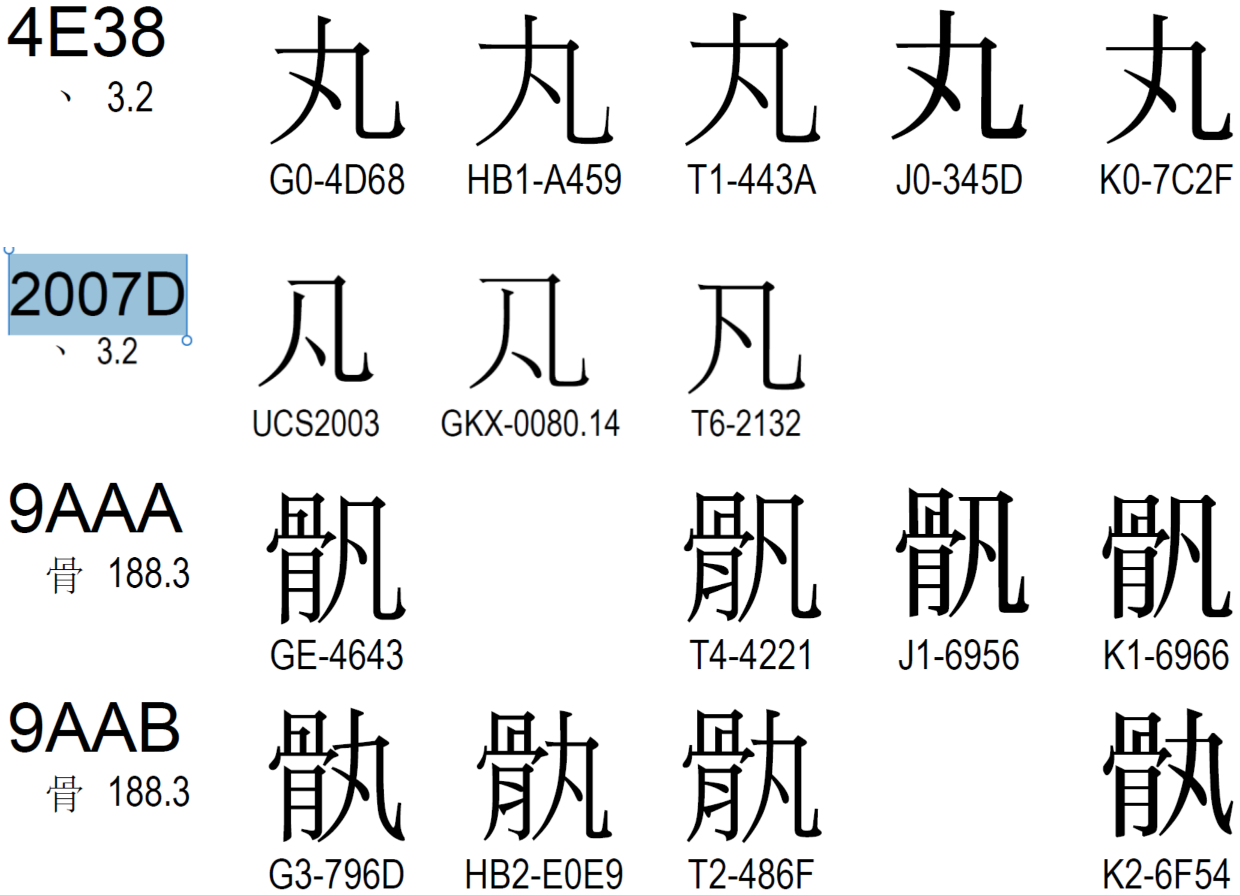
The small top protrusion is significant in distinguishing the etymology of the character in the Kangxi Dictionary. In modern life, the shape of U+2007D 𠁽 is rare. The shape of U+4E38 丸 is often used, as a single character or as a character component. It is recommended that U+2007D and U+4E38 be unifiable components to reduce any confusion.

Existing Disunified Examples:



#2 丸 ancient

丸 u4e38	𠁽 u2007d
汍 u6c4d	汎 u2f8fa
骫 u9aab	骪 u9aaa

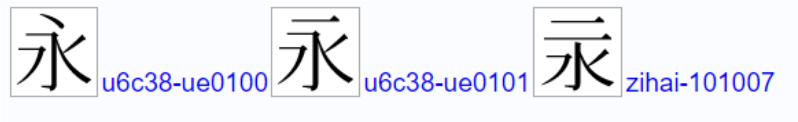


4E38 、 3.2	丸 G0-4D68	丸 HB1-A459	丸 T1-443A	丸 J0-345D	丸 K0-7C2F
2007D 、 3.2	𠁽 UCS2003	𠁽 GKX-0080.14	𠁽 T6-2132		
9AAA 骨 188.3	骪 GE-4643		骪 T4-4221	骪 J1-6956	骪 K1-6966
9AAB 骨 188.3	骫 G3-796D	骫 HB2-E0E9	骫 T2-486F		骫 K2-6F54

The glyphs for U+9AAA from G-source, T-source and K-source may considered for changing to better reflect the etymology.

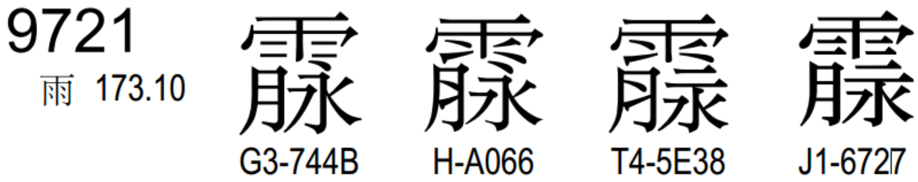
Issue 3

Suggest to add this pair:

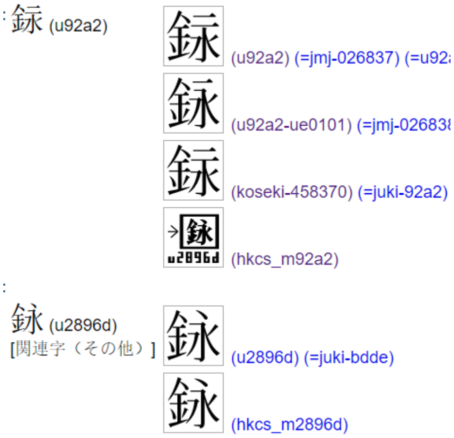


Characters with the 永 phonetic are frequently written as the shape “zihai-101007” in Kangxi dictionary.

Existing Unification example:

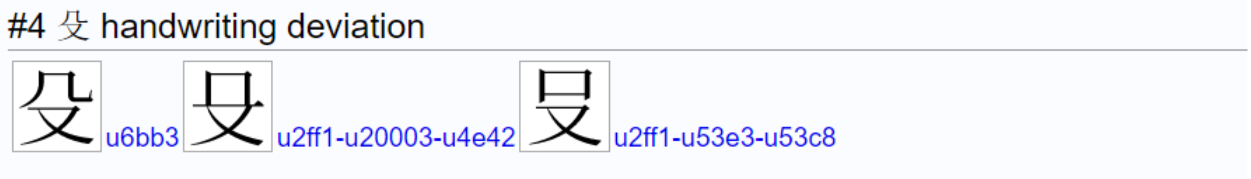


Existing Disunification (mis-disunification) example from Extension-B:



Per the discussion in IRG#47, this difference is usually normalized away by ROK. We should make them unifiable as well, as not all regions can carry out normalization to their glyphs.

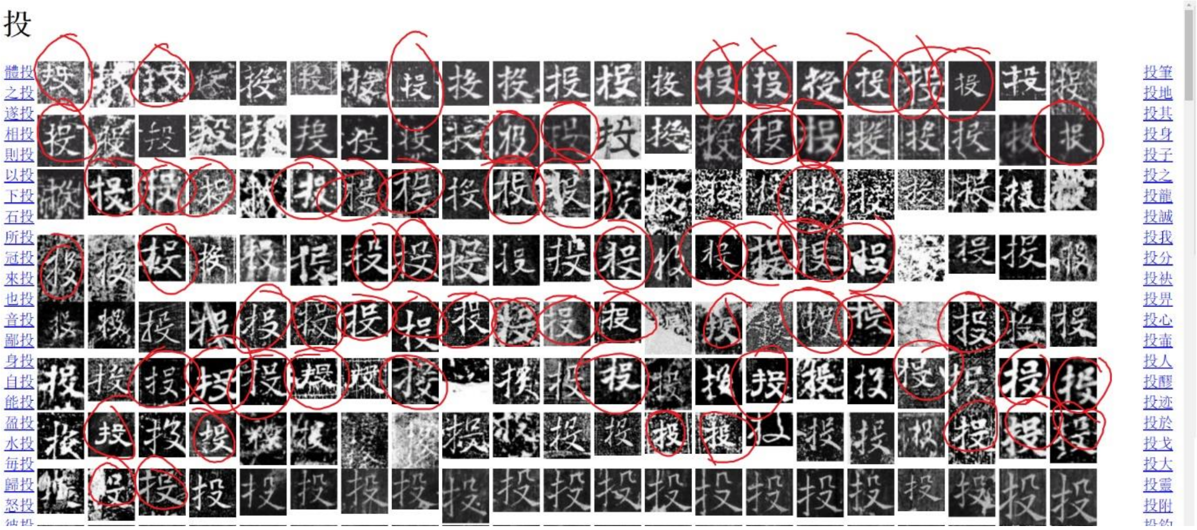
Issue 4

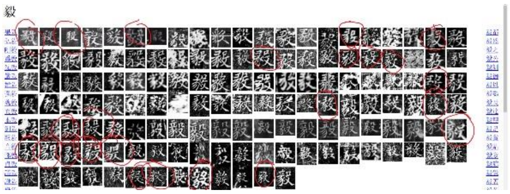


#4 殳 handwriting deviation

殳 u6bb3 u2ff1-u20003-u4e42 u2ff1-u53e3-u53c8

These pairs are handwritten deviation forms frequently seen in calligraphy:





Allowing the disunification of could lead to disastrous results.

However, it may be too late to do anything with the existing characters in Extension F. The disunification of the JMJ characters in Extension F should not be regarded as a “general rule”.

Existing Disunified cases:

U+2A832 𪠲: variant of U+20B1B 𠬛. 殳 is unifiable with 𠬛.

U+23CDA 𣳚: variant of U+6C92 沒. 𠬛 component is unifiable with 殳.

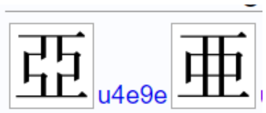
U+22919 𢤙: variant of U+2289E 𢢞.

U+253CF 𥏏: variant of U+6BB9 殹

U+22936 𢤶: Variant of U+6128 愨.

Issue 5

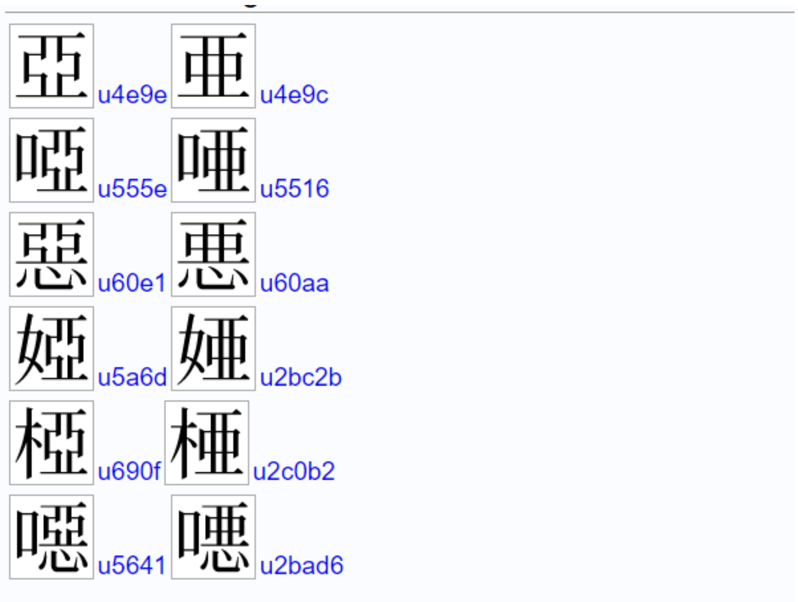
Propose adding the following unification:



亞 u4e9e 亜

The shape difference is minor and could be handled via IVS. The two shapes a purely calligraphic stroke ordering differences.

Current disunified characters:



亞 u4e9e 亜 u4e9c

啞 u555e 唖 u5516

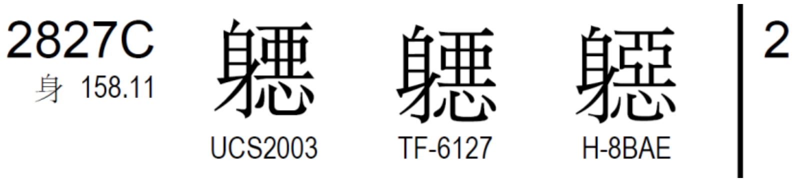
惡 u60e1 悪 u60aa

婭 u5a6d 𫰫 u2bc2b

椏 u690f 𬂲 u2c0b2

噁 u5641 𫫖 u2bad6

A unification example:



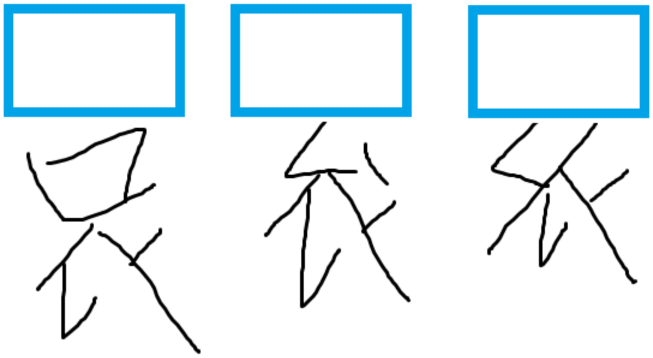
2827C
身 158.11
UCS2003
TF-6127
H-8BAE
2

Previous discussion about disunifying this pair remarked that they should not be disunified, but this could not be regarded as a general unification example.

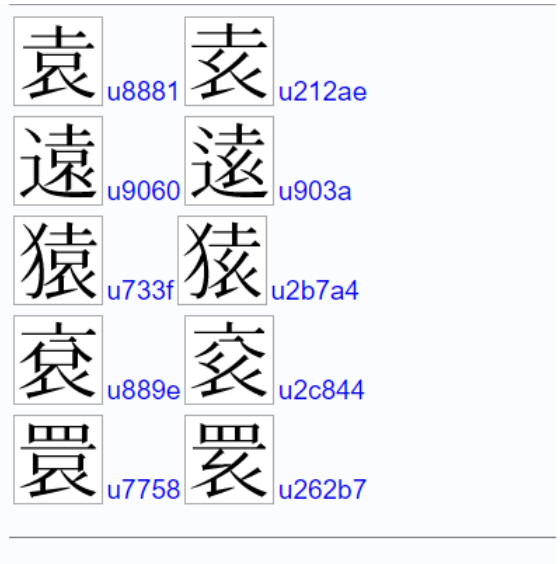
The proposal is to regard make it the preferred unification, given the comments by Whistler to reduce “unnecessary variants”.

Issue 6

Add new entry into UCV:

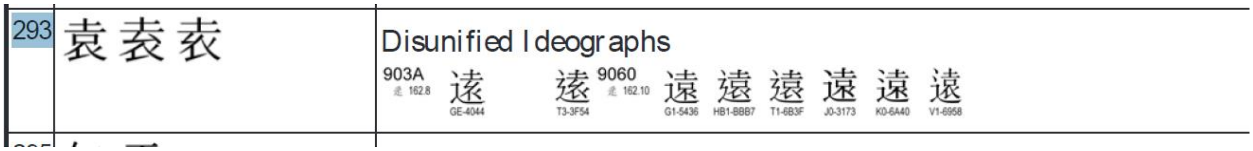


Disunified examples (that are cognate):



袁	u8881	𡊮	u212ae
遠	u9060	逺	u903a
猿	u733f	𫞤	u2b7a4
袞	u889e	𬡄	u2c844
睘	u7758	𦊷	u262b7

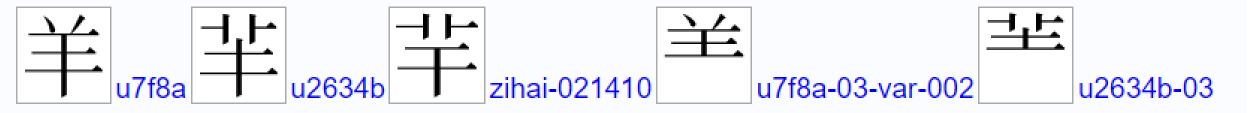
This proposal supersedes #293:



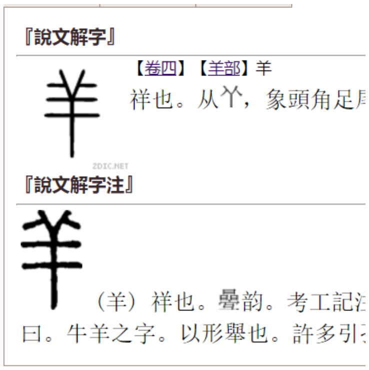
293	袁 𡊮 𡊮	Disunified Ideographs 903A 辵 162.8 逺 GE-4044 逺 T3-3F54 9060 辵 162.10 遠 G1-5436 遠 HB1-BBB7 遠 T1-6B3F 遠 J0-3173 遠 K0-6A40 遠 V1-6958

Issue 7

Add this rule:



These shapes arise from the different Kaishu-fication of the seal shape:



『說文解字』

【卷四】【羊部】羊

祥也。从丫，象頭角足

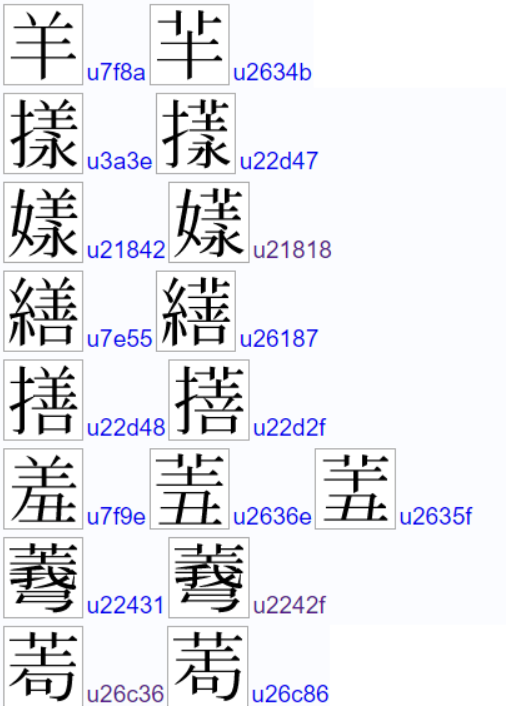
『說文解字注』

（羊）祥也。疊韵。考工記

曰。牛羊之字。以形舉也。許多引

(Screenshot from zdic.net)

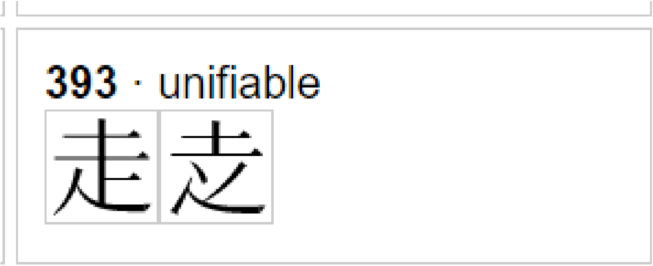
Existing Disunified (mis-disunified) cognates:



Issue 8

We should expand existing UCV rule #393 to cover more scenarios.

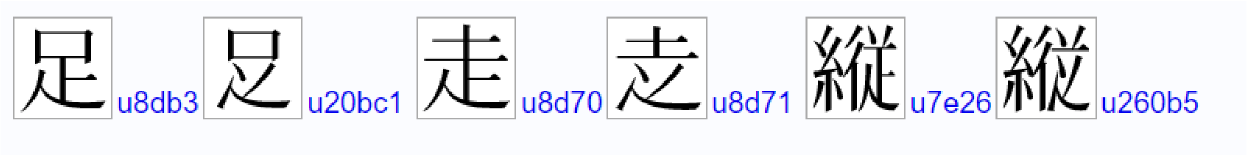
UCV Rule #393:



393 · unifiable

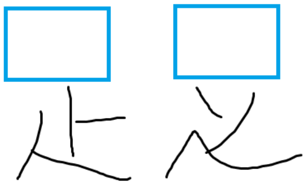
走 赱

Rule 393 should be expanded to cover the following scenarios:



足 u8db3 𠯁 u20bc1 走 u8d70 赱 u8d71 縱 u7e26 𦂵 u260b5

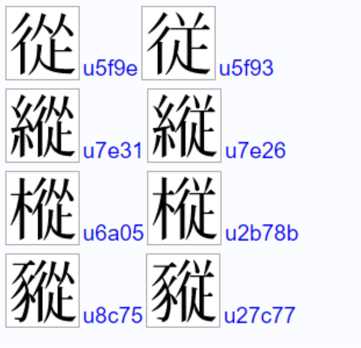
Proposed modified UCV Rule:



#393:

Issue 9

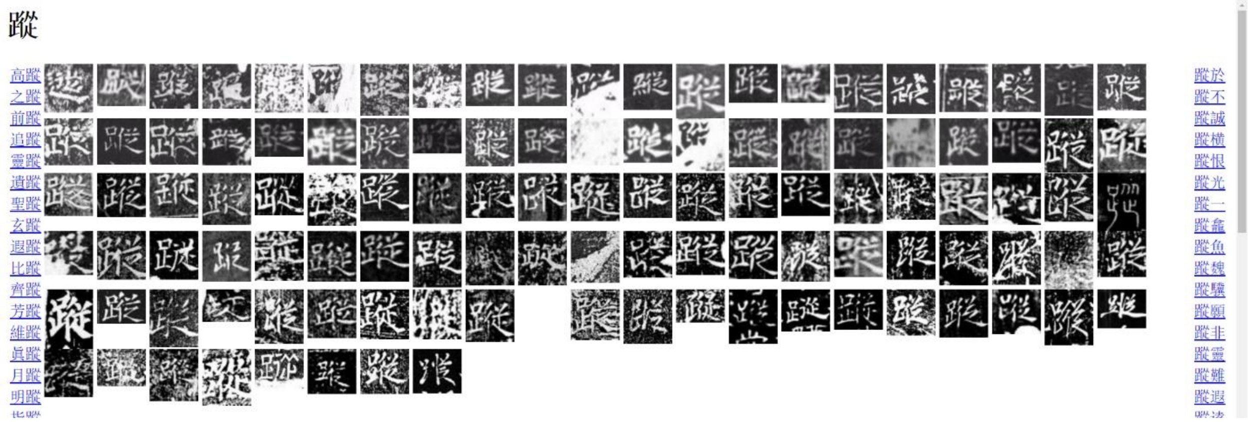
Consider adding rule to UCV for these variants:



從 u5f9e	従 u5f93
縱 u7e31	縦 u7e26
樅 u6a05	𫞋 u2b78b
豵 u8c75	𧱷 u27c77

These variants are cognates.

This variation is quite common. In fact, it is often combined with modified #393 UCV:



蹤

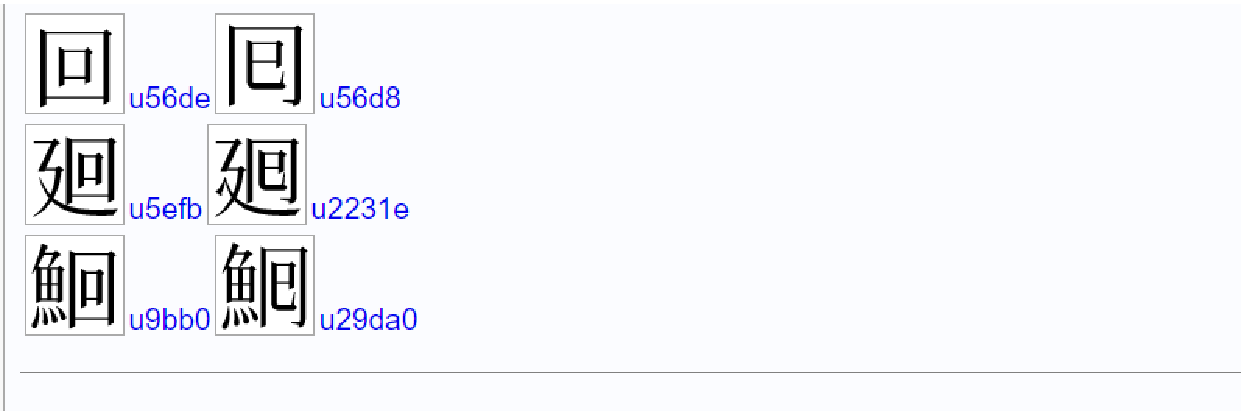
高蹤
之蹤
前蹤
追蹤
靈蹤
遺蹤
聖蹤
玄蹤
遐蹤
比蹤
齊蹤
芳蹤
維蹤
眞蹤
月蹤
明蹤

蹤於
蹤不
蹤誠
蹤横
蹤恨
蹤光
蹤一
蹤龕
蹤魚
蹤魏
蹤驥
蹤願
蹤非
蹤靈
蹤難
蹤遐

However, the shape difference might be “too significant” to unify?

Issue 10

Consider adding rule to UCV for these variants:



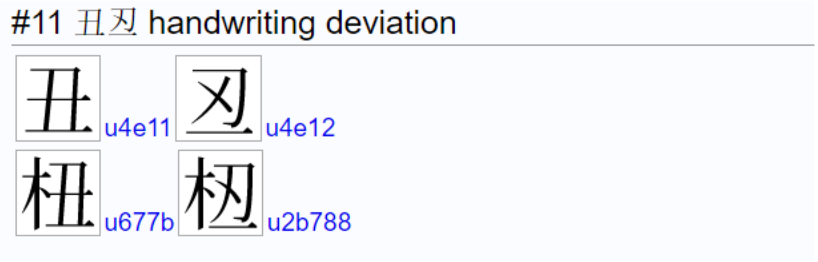
回 u56de 囘 u56d8

廻 u5efb 𢌞 u2231e

鮰 u9bb0 𩶠 u29da0

Issue 11A

Consider adding rule to UCV for these variants:

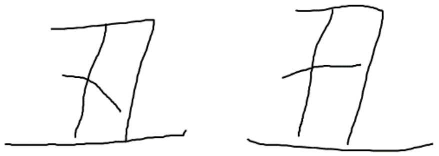


#11 丑丒 handwriting deviation

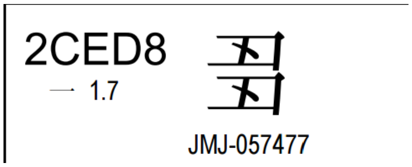
丑 u4e11 丒 u4e12

杻 u677b 𫞈 u2b788

The Zhuang characters submission by PRC has also normalized these shapes:



Issue 11B



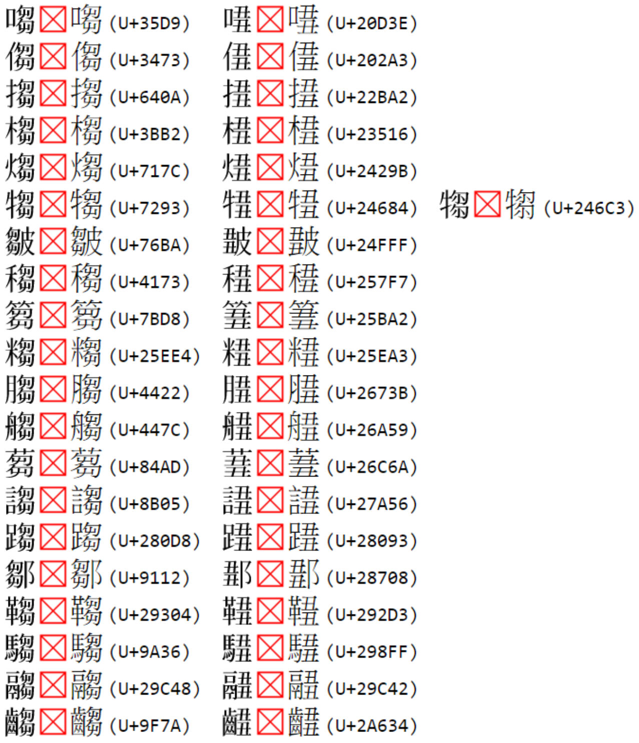
2CED8

一 1.7

JMJ-057477

is a variant of 芻.

In fact, all of these are cognates according to Hanyu Dazidian and Zhonghua Zihai:

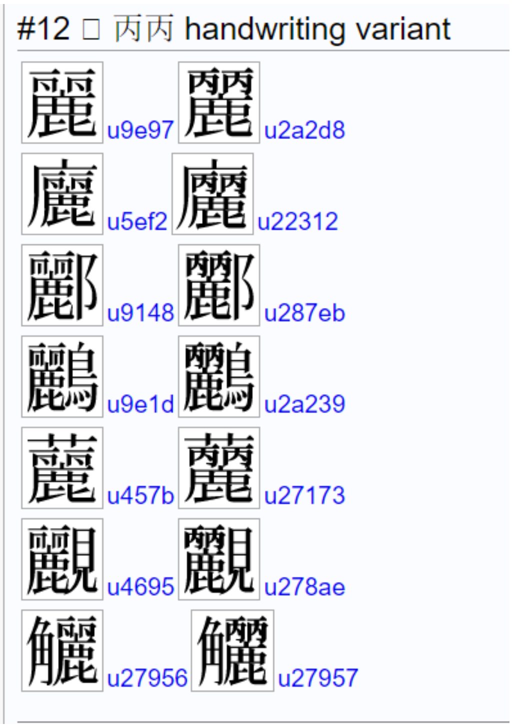


㗙☒㗙 (U+35D9)	𠴾☒𠴾 (U+20D3E)	
㑳☒㑳 (U+3473)	𠊣☒𠊣 (U+202A3)	
搊☒搊 (U+640A)	𢮢☒𢮢 (U+22BA2)	
㮲☒㮲 (U+3BB2)	𣔖☒𣔖 (U+23516)	
煼☒煼 (U+717C)	𤊛☒𤊛 (U+2429B)	
犓☒犓 (U+7293)	𤚄☒𤚄 (U+24684)	𤛃☒𤛃 (U+246C3)
皺☒皺 (U+76BA)	𤿿☒𤿿 (U+24FFF)	
䅳☒䅳 (U+4173)	𥟷☒𥟷 (U+257F7)	
篘☒篘 (U+7BD8)	𥮢☒𥮢 (U+25BA2)	
𥻤☒𥻤 (U+25EE4)	𥺣☒𥺣 (U+25EA3)	
䐢☒䐢 (U+4422)	𦜻☒𦜻 (U+2673B)	
䑼☒䑼 (U+447C)	𦩙☒𦩙 (U+26A59)	
蒭☒蒭 (U+84AD)	𦱪☒𦱪 (U+26C6A)	
謅☒謅 (U+8B05)	𧩖☒𧩖 (U+27A56)	
𨃘☒𨃘 (U+280D8)	𨂓☒𨂓 (U+28093)	
鄒☒鄒 (U+9112)	𨜈☒𨜈 (U+28708)	
𩌄☒𩌄 (U+29304)	𩋓☒𩋓 (U+292D3)	
騶☒騶 (U+9A36)	𩣿☒𩣿 (U+298FF)	
𩱈☒𩱈 (U+29C48)	𩱂☒𩱂 (U+29C42)	
齺☒齺 (U+9F7A)	𪘴☒𪘴 (U+2A634)	

However, the difference between 芻 and JMJ-057477 is too big in my opinion. Consider add to NUCV.

Issue 12

Consider adding a UCV rule for these variants:



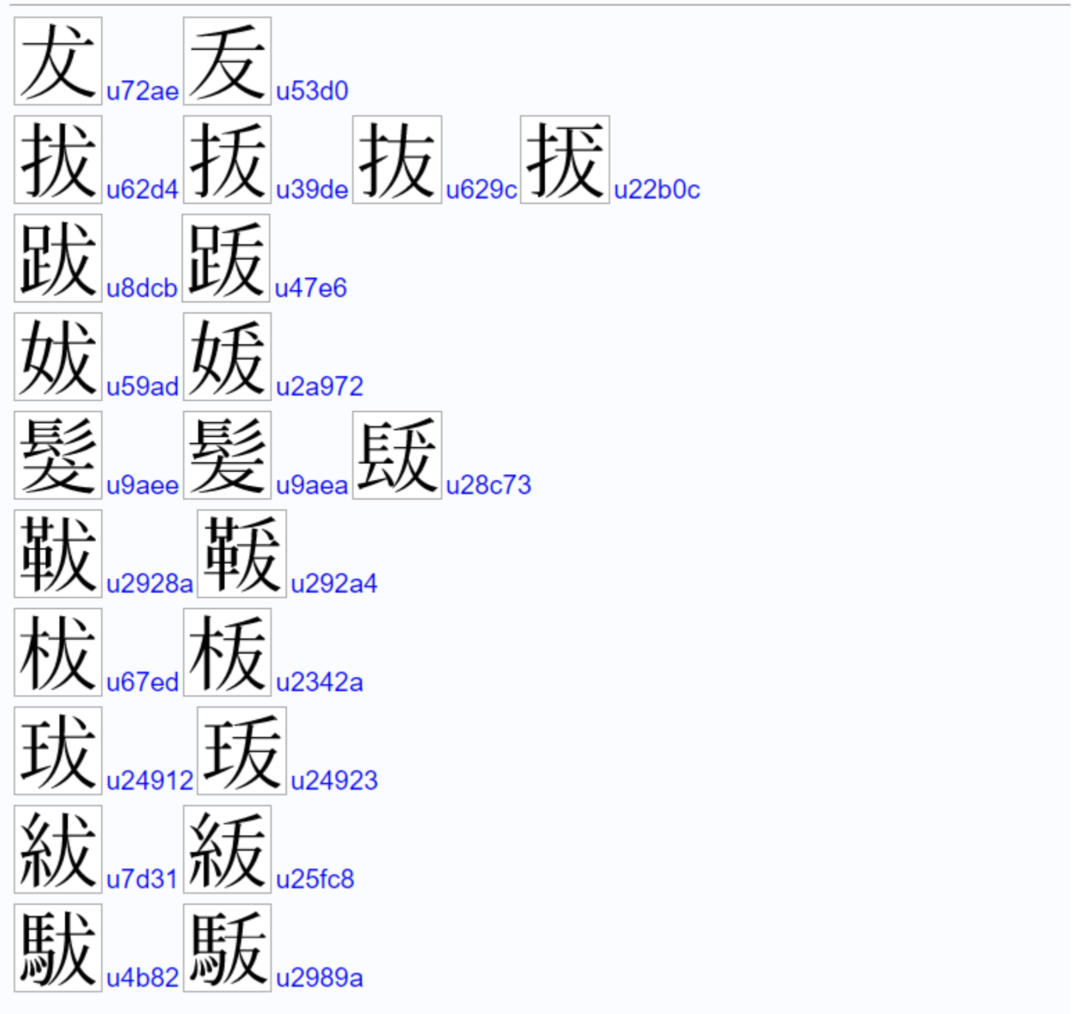
#12 □ 丙丙 handwriting variant

麗	u9e97	𪋘	u2a2d8
廲	u5ef2	𢌒	u22312
酈	u9148	𨟫	u287eb
鸝	u9e1d	𪈹	u2a239
䕻	u457b	𧅳	u27173
䚕	u4695	𧢮	u278ae
𧥖	u27956	𧥗	u27957

They are cognate.

Issue 13:

Consider adding UCV for these variants:



犮 u72ae 叐 u53d0

拔 u62d4 㧞 u39de 抜 u629c 𢬌 u22b0c

跋 u8dcb 䟦 u47e6

妭 u59ad 𪥲 u2a972

髮 u9aee 髪 u9aea 𨱳 u28c73

𩊊 u2928a 𩊤 u292a4

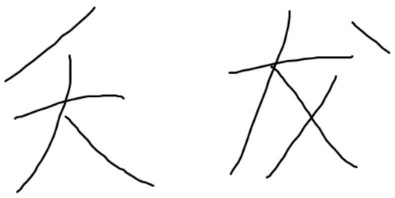
柭 u67ed 𣐪 u2342a

𤤒 u24912 𤤣 u24923

紱 u7d31 𥿈 u25fc8

䮂 u4b82 𩢚 u2989a

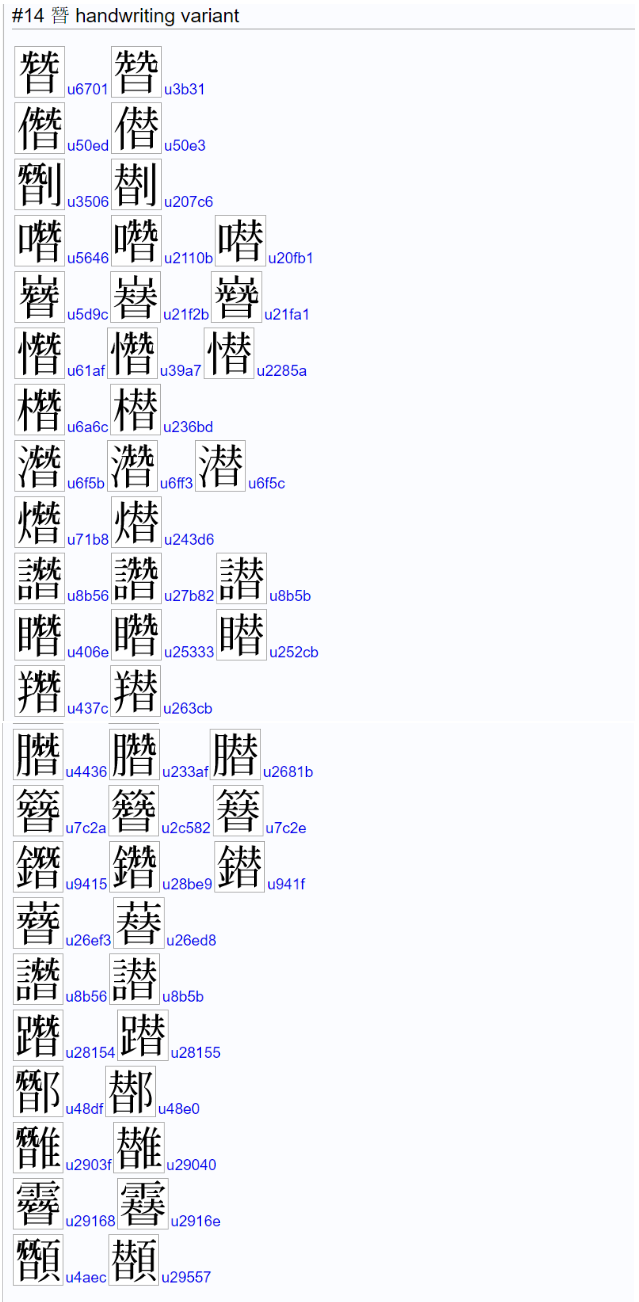
These examples are consistent with ROK's normalization.



ROK normalization also includes this pair. However, such normalization generally requires semantic decomposition to determine the correct glyph shape. Consider adding to UCV with a note.

Issue 14:

Consider adding UCV for these variants:



#14 朁 handwriting variant

朁 u6701 㬱 u3b31

僭 u50ed 僣 u50e3

㔆 u3506 𠟆 u207c6

噆 u5646 𡄋 u2110b 𠾱 u20fb1

嶜 u5d9c 𡼫 u21f2b 𡾡 u21fa1

憯 u61af 㦧 u39a7 𢡚 u2285a

橬 u6a6c 𣚽 u236bd

潛 u6f5b 濳 u6ff3 潜 u6f5c

熸 u71b8 𤏖 u243d6

譖 u8b56 𧮂 u27b82 譛 u8b5b

䁮 u406e 𥌳 u25333 𥋋 u252cb

䍼 u437c 𦏋 u263cb

䐶 u4436 𣎯 u233af 𦠛 u2681b

簪 u7c2a 𬖂 u2c582 簮 u7c2e

鐕 u9415 𨯩 u28be9 鐟 u941f

𦻳 u26ef3 𦻘 u26ed8

譖 u8b56 譛 u8b5b

𨅔 u28154 𨅕 u28155

䣟 u48df 䣠 u48e0

𩀿 u2903f 𩁀 u29040

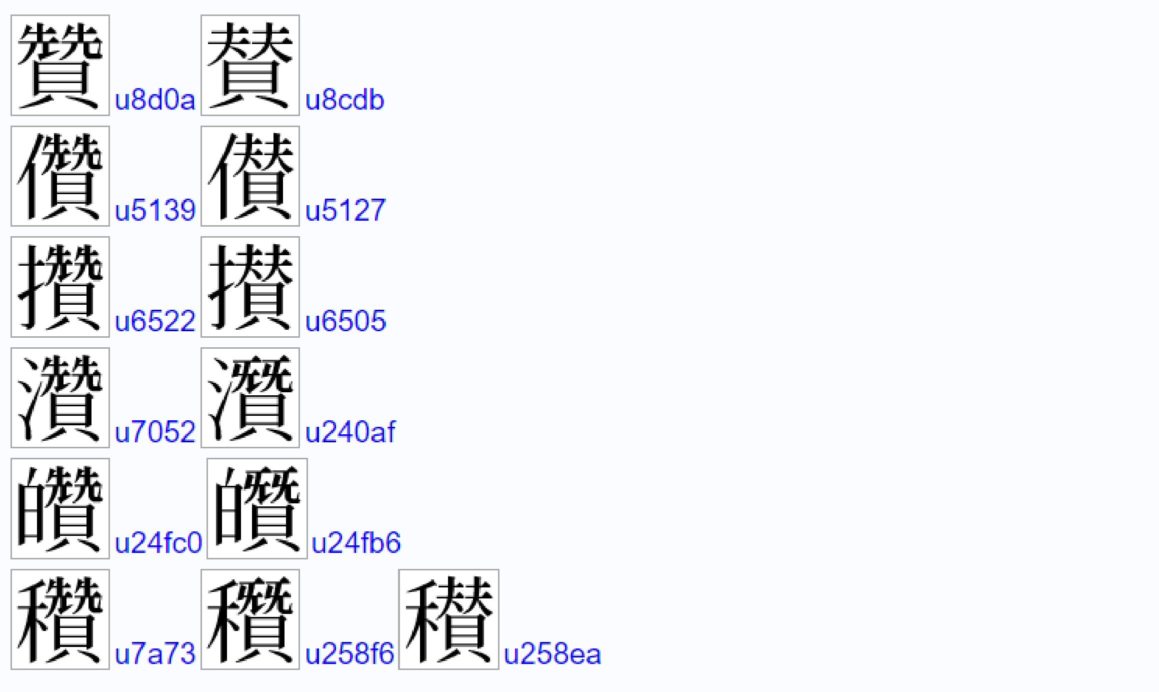
𩅨 u29168 𩅮 u2916e

䫬 u4aec 𩕗 u29557

The more etymologically correct glyph is on the left.

Issue 15:

Consider adding UCV rule for these characters:



u8d0a u8cdb

u5139 u5127

u6522 u6505

u7052 u240af

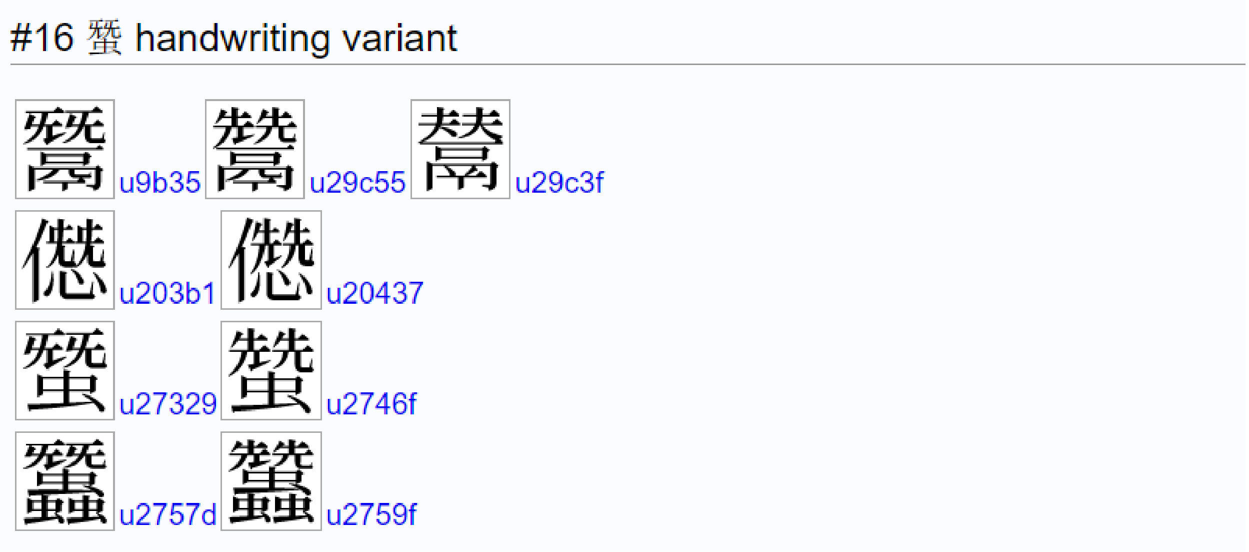
u24fc0 u24fb6

u7a73 u258f6 u258ea

The more etymologically correct form is on the left.

Issue 16

Consider adding UCV rule for these characters:



#16 贊 handwriting variant

u9b35 u29c55 u29c3f

u203b1 u20437

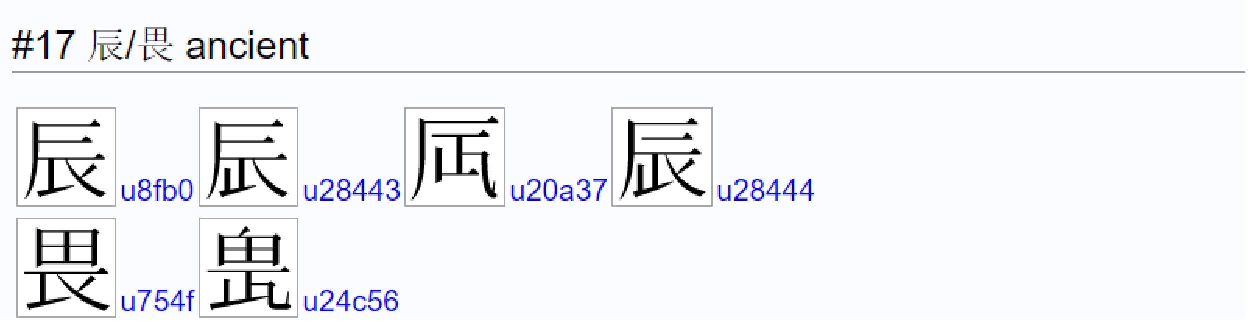
u27329 u2746f

u2757d u2759f

The more etymologically correct form is on the left.

Issue 17:

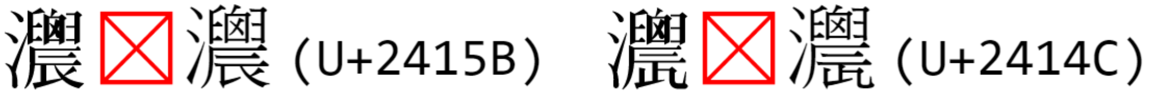
Consider adding these glyphs to a UCV rule:



#17 辰/畏 ancient

辰 u8fb0 𨑃 u28443 𠨷 u20a37 𨑄 u28444

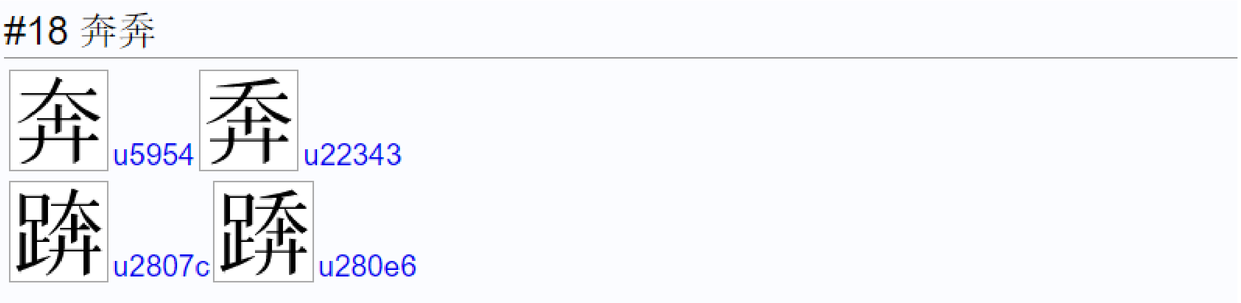
畏 u754f 𤱖 u24c56



These glyphs are derived from different Song/Ming typeface interpretations of the similar in Shuowenjiezi.

Issue 18:

Consider adding these glyphs to a UCV rule:



#18 奔𢍃

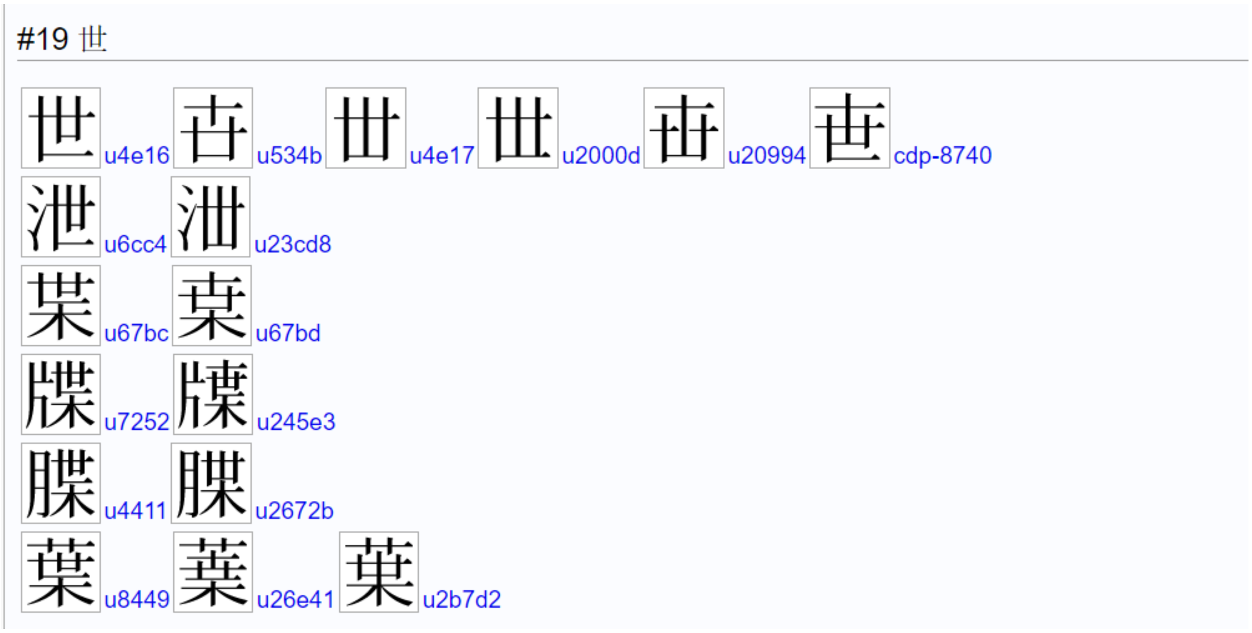
奔 u5954 𢍃 u22343

𨁼 u2807c 𨃦 u280e6

A “running person” is represented in oracle script as a person throwing his arms up and down with his head slanted. It is modernized as 大 (big) (two arms downwards) or 夭 (devil) (head slanted and two arms downwards), which are both characters of unrelated origin. The presence of an additional slant is not contributive to the meaning of the character. (Similar to issue 13)

Issue 19:

Consider adding UCV to cover the following cognates:



#19 世

世 u4e16 卋 u534b 丗 u4e17 𠀍 u2000d 𠦔 u20994 cdp-8740

泄 u6cc4 𣳘 u23cd8

枼 u67bc 枽 u67bd

牒 u7252 𤗣 u245e3

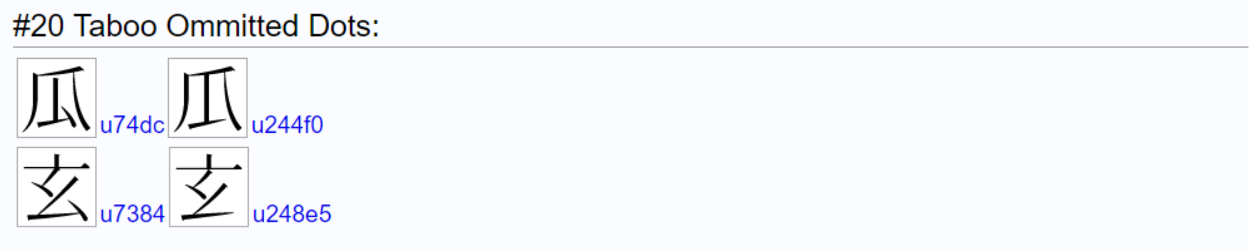
䐑 u4411 𦜫 u2672b

葉 u8449 𦹁 u26e41 𫟒 u2b7d2

The included variations are cognate.

Issue 20:

Consider noting these two cases as mis-disunification of “missing of minor dot”.



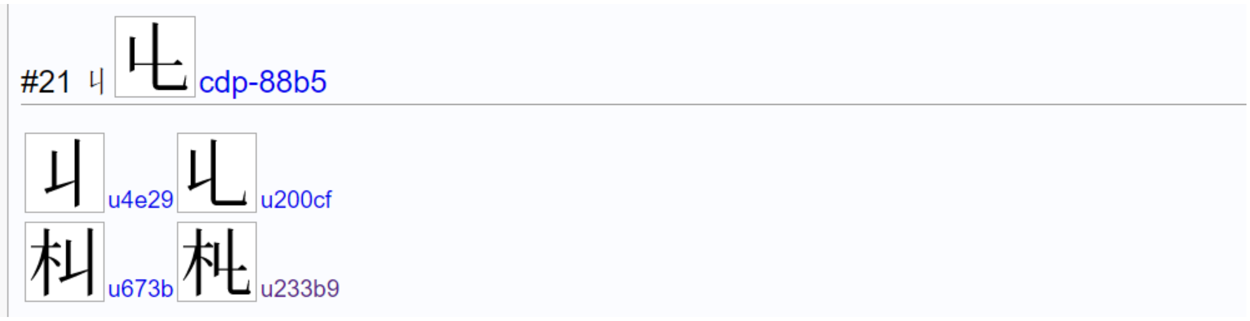
#20 Taboo Ommitted Dots:

瓜 u74dc 𤓰 u244f0

玄 u7384 𤣥 u248e5

Issue 21:

Consider covering the following cognates:



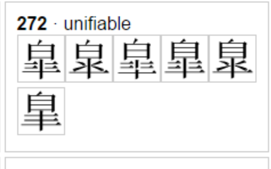
#21 丩 cdp-88b5

丩 u4e29 𠃏 u200cf

朻 u673b 𣎹 u233b9

Issue 22A:

Existing UCV:

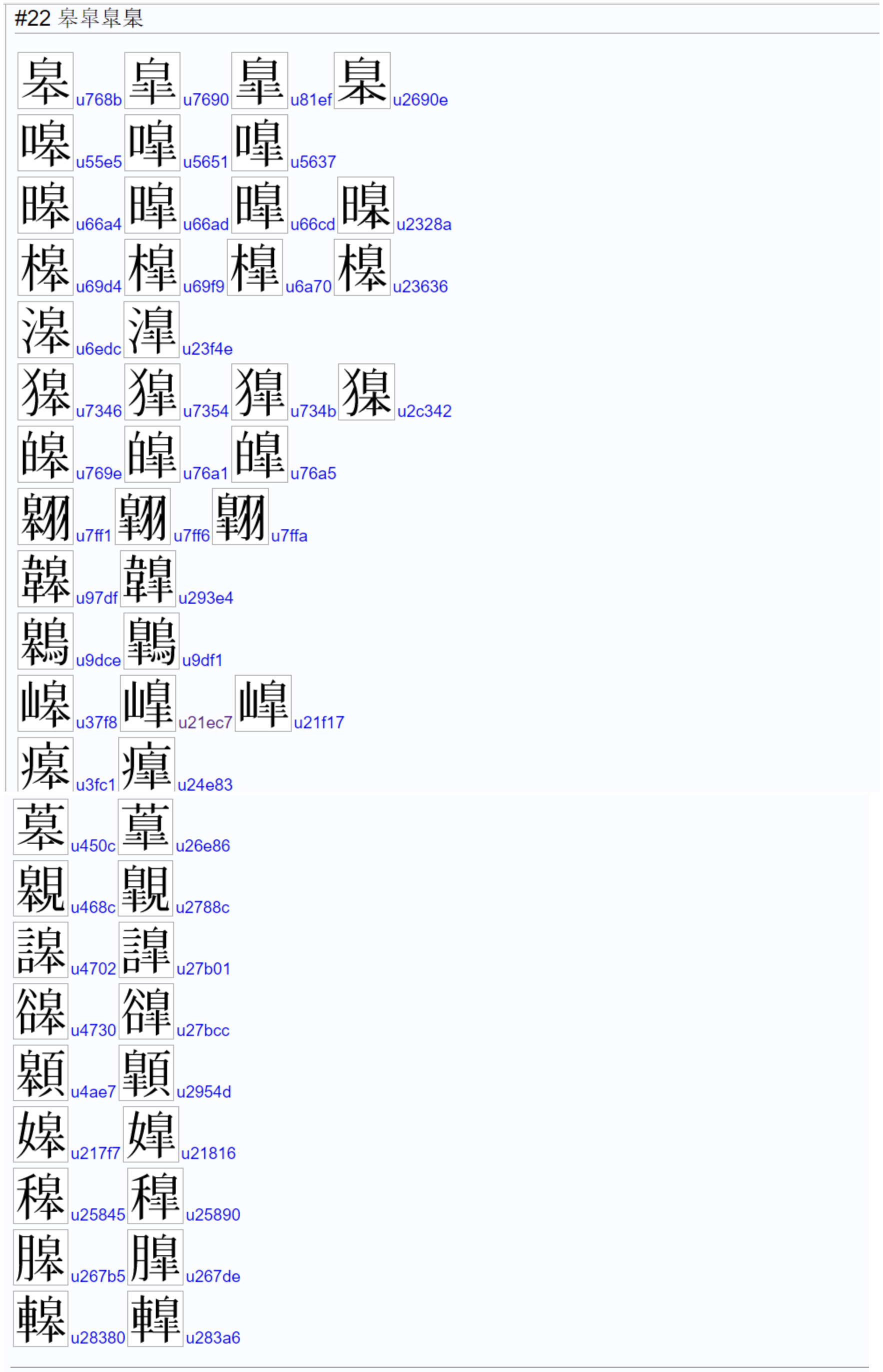


272 · unifiable

皐 皋 皐 臯 𦤎

臯

Consider covering these cognates:



#22 皋皐臯𦤎

皋 u768b 皐 u7690 臯 u81ef 𦤎 u2690e

嗥 u55e5 噑 u5651 嘷 u5637

暤 u66a4 暭 u66ad 曍 u66cd 𣊊 u2328a

槔 u69d4 槹 u69f9 橰 u6a70 𣘶 u23636

滜 u6edc 𣽎 u23f4e

獆 u7346 獔 u7354 獋 u734b 𬍂 u2c342

皞 u769e 皡 u76a1 皥 u76a5

翱 u7ff1 翶 u7ff6 翺 u7ffa

韟 u97df 𩏤 u293e4

鷎 u9dce 鷱 u9df1

㟸 u37f8 𡻇 u21ec7 𡼗 u21f17

㿁 u3fc1 𤺃 u24e83

䔌 u450c 𦺆 u26e86

䚌 u468c 𧢌 u2788c

䜂 u4702 𧬁 u27b01

䜰 u4730 𧯌 u27bcc

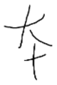
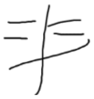
䫧 u4ae7 𩕍 u2954d

𡟷 u217f7 𡠖 u21816

𥡅 u25845 𥢐 u25890

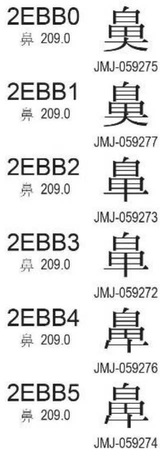
𦞵 u267b5 𦟞 u267de

𨎀 u28380 𨎦 u283a6

The shape difference between  and  are too significant. However, the top part difference of 白 and 自 should be unifiable, when the bottom part is consistent.

Issue 22B:

Considering unifying 白 and 自 in this case as well:



2EBB0
鼻 209.0
JMJ-059275

2EBB1
鼻 209.0
JMJ-059277

2EBB2
鼻 209.0
JMJ-059273

2EBB3
鼻 209.0
JMJ-059272

2EBB4
鼻 209.0
JMJ-059276

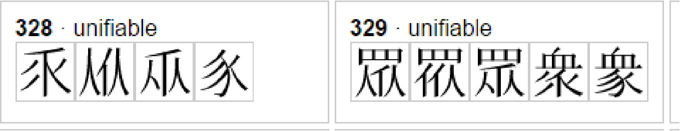
2EBB5
鼻 209.0
JMJ-059274

Issue 23:

#withdrawn

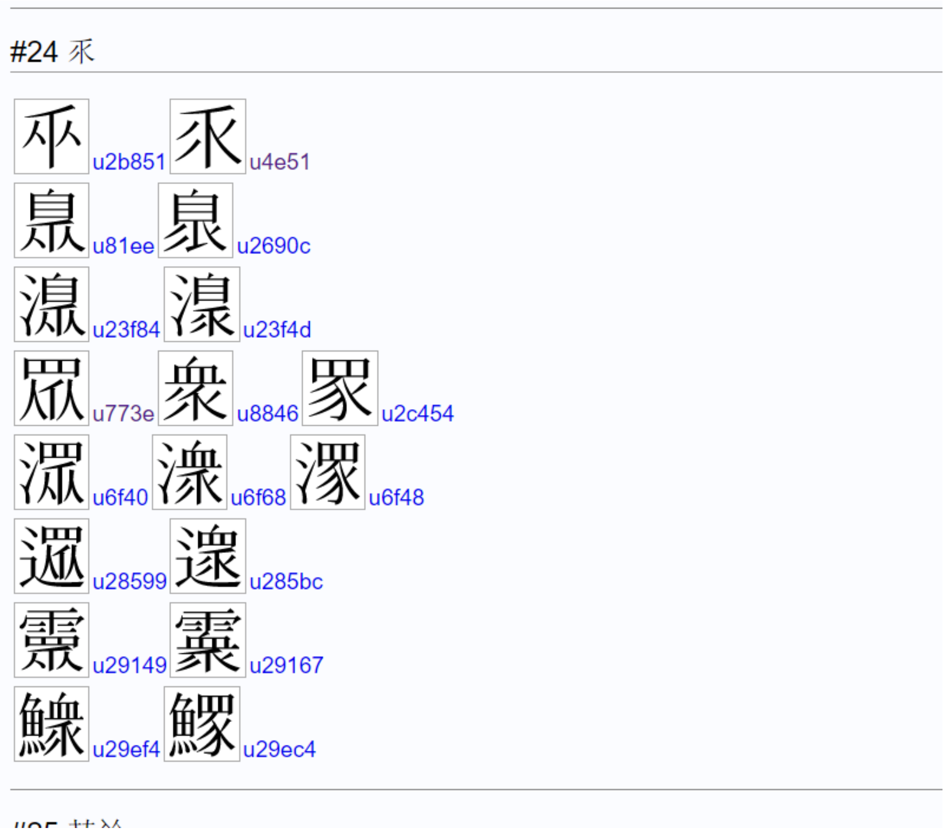
Issue 24:

These glyphs are the same character. Consider marking them as mis-disunifications due to rule #328 and #329:



328 · unifiable

329 · unifiable



#24 乑

𫡑 u2b851 乑 u4e51

臮 u81ee 𦤌 u2690c

𣾄 u23f84 𣽍 u23f4d

眾 u773e 衆 u8846 𬑔 u2c454

潀 u6f40 潨 u6f68 潈 u6f48

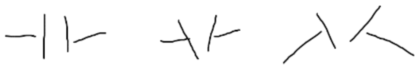
𨖙 u28599 𨖼 u285bc

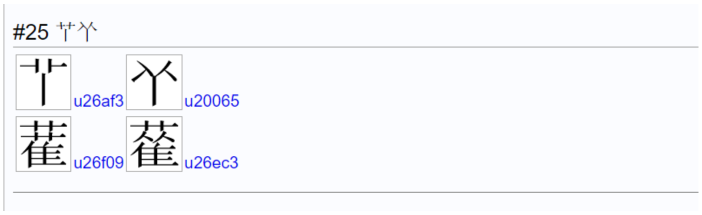
𩅉 u29149 𩅧 u29167

𩻴 u29ef4 𩻄 u29ec4

Issue 25:

Consider adding UCV rule for these variants:

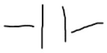
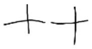




#25 𦫳𠁥

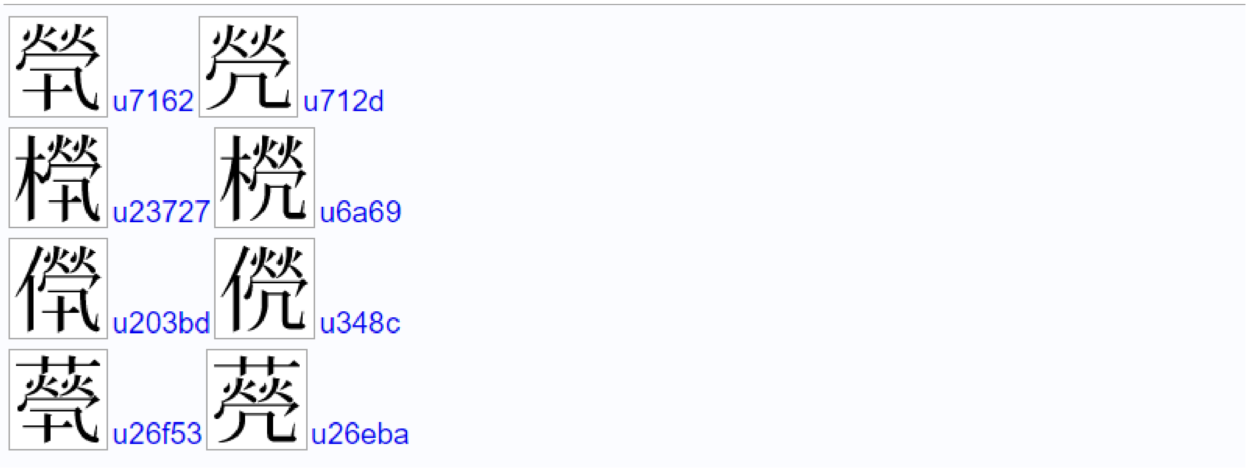
𦫳 u26af3 𠁥 u20065

𦼉 u26f09 𦻃 u26ec3

The glyph design of u26f09 should be  in the middle but the PRC normalization conventions convert it into grass ()

Issue 26:

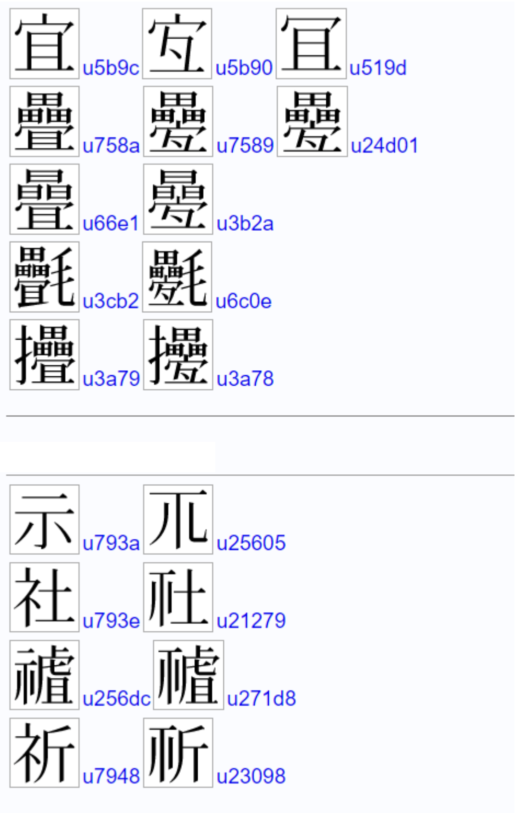
Consider new UCV rule for these cognates:



煢	u7162	焭	u712d
𣜧	u23727	橩	u6a69
𠎽	u203bd	㒌	u348c
𦽓	u26f53	𦺺	u26eba

Issue 27 & 28:

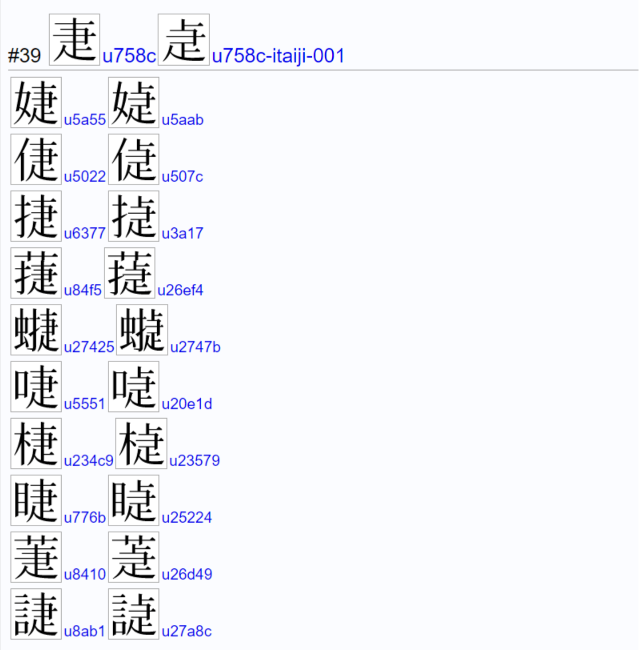
Ditto as above.



宜	u5b9c	宐	u5b90	冝	u519d
疊	u758a	疉	u7589	𤴁	u24d01
曡	u66e1	㬪	u3b2a		
㲲	u3cb2	氎	u6c0e		
㩹	u3a79	㩸	u3a78		

示	u793a	𥘅	u25605
社	u793e	𡉹	u21279
𥛜	u256dc	𧇘	u271d8
祈	u7948	𣂘	u23098

Issue 29: Ditto.

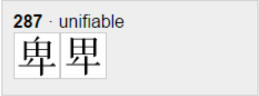


#39 疌 u758c 疌 u758c-itaiji-001

婕	u5a55	媫	u5aab
倢	u5022	偼	u507c
捷	u6377	㨗	u3a17
蓵	u84f5	𦻴	u26ef4
𧐥	u27425	𧑻	u2747b
啑	u5551	𠸝	u20e1d
𣓉	u234c9	𣕹	u23579
睫	u776b	𥈤	u25224
萐	u8410	𦵉	u26d49
誱	u8ab1	𧪌	u27a8c

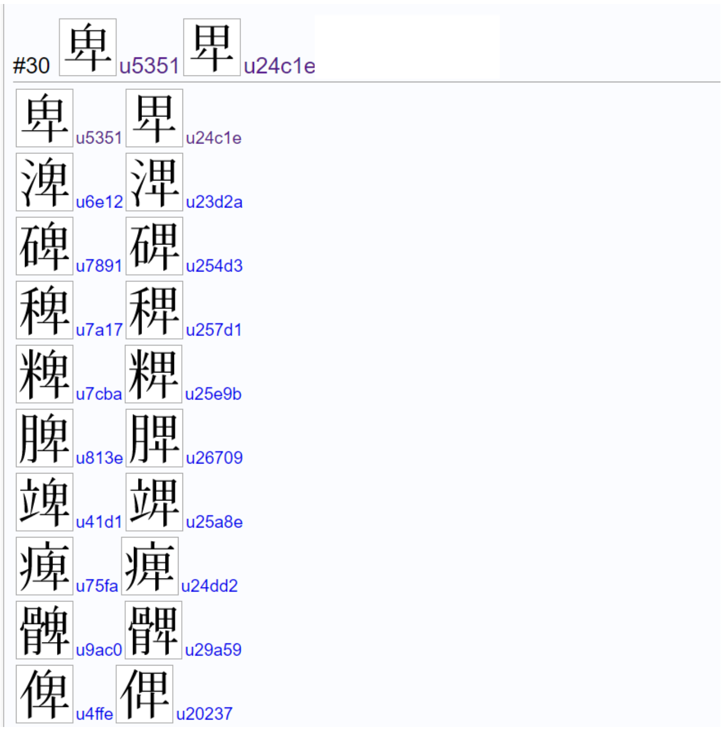
Issue 30:

Update #287 to cover all following characters and mark as mis-disunification:



287 · unifiable

卑 𤰞

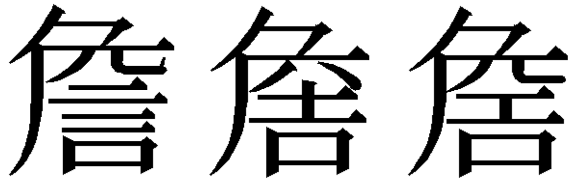


#30 卑 u5351 𤰞 u24c1e

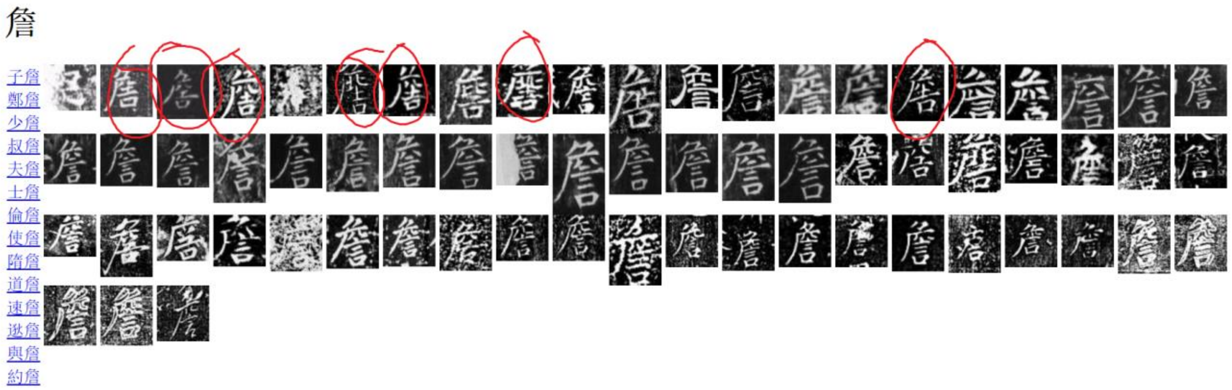
卑	u5351	𤰞	u24c1e
渒	u6e12	𣴪	u23d2a
碑	u7891	𥓓	u254d3
稗	u7a17	𥟑	u257d1
粺	u7cba	𥺛	u25e9b
脾	u813e	𦜉	u26709
䇑	u41d1	𥪎	u25a8e
痺	u75fa	𤷒	u24dd2
髀	u9ac0	𩩙	u29a59
俾	u4ffe	𠈷	u20237

Issue 31 (added per discussion in IRG#47 1st day)

Add to UCV:

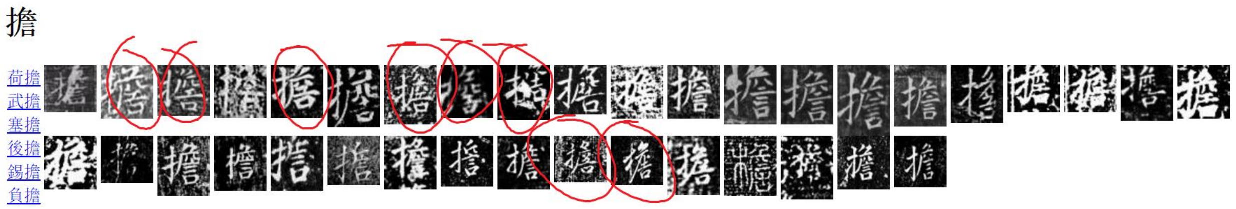


This variation is very frequently seen:



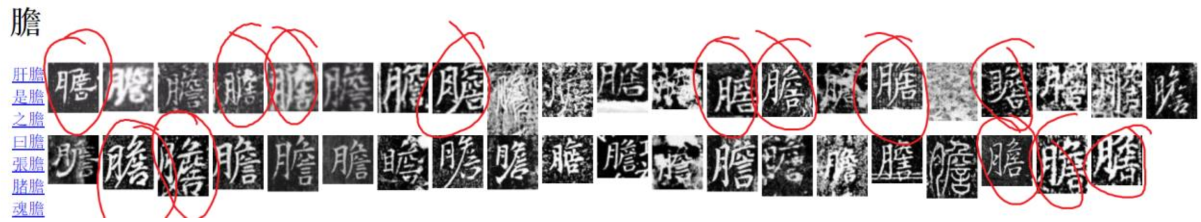
詹

子詹
鄭詹
少詹
叔詹
夫詹
士詹
倫詹
使詹
隋詹
道詹
速詹
逖詹
輿詹
約詹



擔

荷擔
武擔
塞擔
後擔
錫擔
負擔



膽

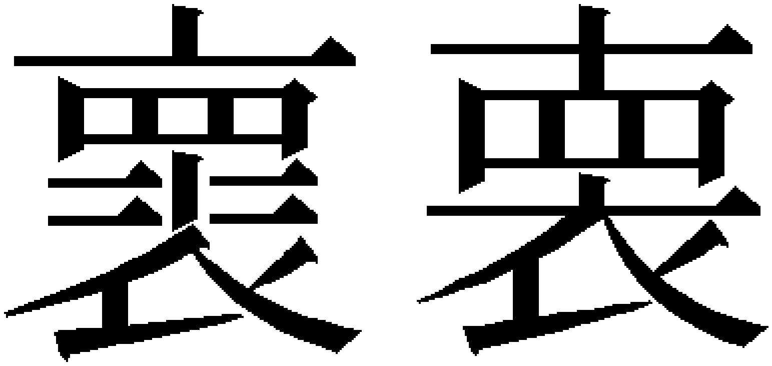
肝膽
是膽
之膽
曰膽
張膽
賭膽
魂膽

As discussed, the allowed disunification could be catastrophic given the high number of characters containing the 詹 phonetic.

No disunification example exist.

Issue 32 (added per discussion of IRG #47 1st day)

Add new NUCV:



As discussed, shape difference is significant: number of components may be counted differently.

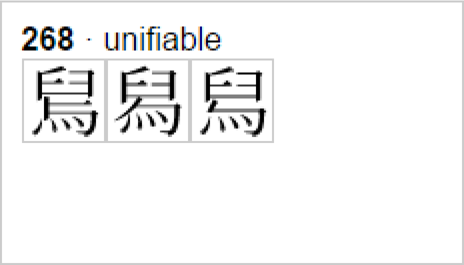
Disunification in URO:

U+61F7 懷 vs U+61D0 懐

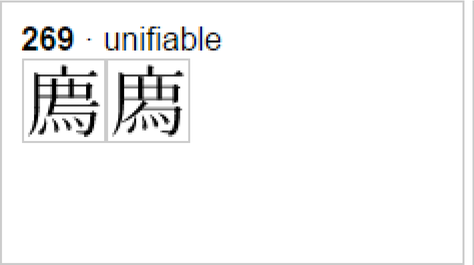
U+58DE 壞 vs U+58CA 壊

Issue 33 (added per discussion on IRG #47 1st day)

Modify #268 and #269:

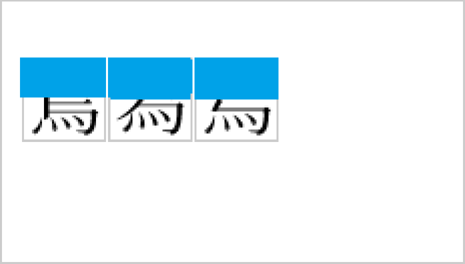


268 · unifiable



269 · unifiable

To



To cover a unification case discussed by IRG regarding U+24261

Discussion:

Traditionally, in certain East Asian regions, a heavy emphasis is placed on “correct” or “proper” form in official contexts. Variants frequently seen in calligraphy is because calligraphy is “an art”. Handwritten characters are often “normalized” to a certain glyph shape before they are added into national encoded character sets.

Each region may have different preferred forms. The ISO/IEC10646 is a standard that encodes via a character basis, not a glyph basis. Therefore, similar forms are usually unified to the same codepoint. This unification across all regions (GHTJKV) is most significant in URO. The current Annex S and hence UCV rules are based on these inter-region unifications precedent.

However, due to legal reasons, some regions may need to assign a new code-point in their national character set for every glyph variant that occurs, in handwriting or print, no matter how small the variance. This is usually particular to family register computer systems.

Unlike the normal character sets submitted to IRG, these family register character sets may contain many variants of the same character. Now, the IRG has decided that IVS is the better solution. This “unification” (intra-region unification) by encoding variants via IVS is also currently called “unification”, but it is different in nature to the unification (inter-region unification) carried out previously. This new type of unification is in nature a kind of “normalization” because the variant shapes are discarded from UCS. The only difference is where it is carried out: at the level of IRG (e.g. TCA’s postponed unifications in IRG #47), or before the submission to IRG (e.g. ROK’s normalization in IRG #47).

In the intra-region unification, often semantic analysis is required, and the definition of “minor difference” is rather arbitrary. More often, it depends on the evidence supplied from the submitter to determine the correct base character it should unify to. In inter-region unification, if the glyph shape difference is rather large, it has been simply disunified (e.g. such as the separate coding of 壞 and 壊 in URO).

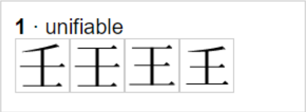
The current UCV is not constructed to handle these two situations well. Also, the current UCV is very long and quite hard to use. Many characters that are mis-disunifications are currently categorized as “Disunified”, because analyzing them is hard.

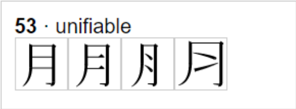
Proposal

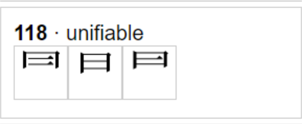
I propose that we split the UCV into two main subgroups: Assimilation and Variance.

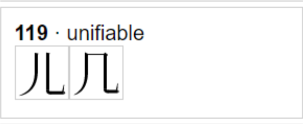
Assimilation should concern where two similar shape but phonetically distinct characters are often mixed up when used inside a character.

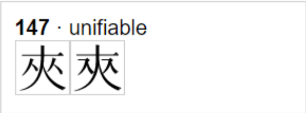
Examples include:

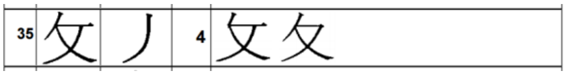


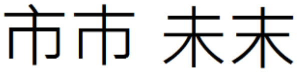




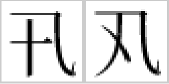






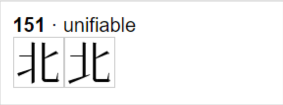


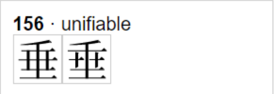
Whether two glyphs containing these components can be unified (same character) or cannot be unified (non-cognate, different character) heavily depends on semantic analysis. A lot of attention should be placed – these are cases where previous unifications have gone (horribly) wrong.

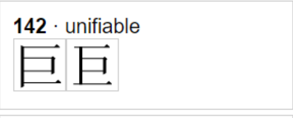
My additional proposal is, IRG should specify a virtual "proper" shape for the separate etymologies, for determining a SC and FS for multi-source and single-source characters. The "proper" shape need not follow Kangxi closely, but should use forms traditionally used to denote different etymology. An example is, in Taiwan and Hong Kong,  are distinguished components which reflect their etymology, which is generally not distinguished in Kangxi. Having a distinguished virtual shape will more intuitively identify any non-cognates.

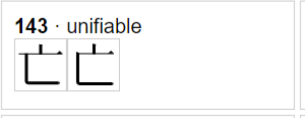
Variance would concern variances in glyph that do not incur phonetic differences for characters than contain them. Unifications in this category should be nearly 100%. Complex components which swap out part of their component with another component with a completely different semantic meaning, but have no added effect on glyphs containing this complex component also belong to this category.

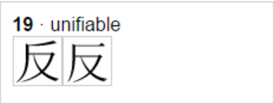
Variances that do not concern any etymological differences:



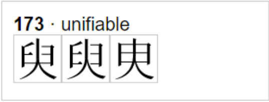


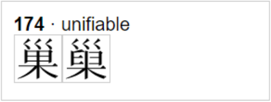


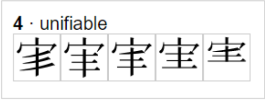


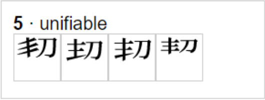


Variant components with components which differ in etymology, but have no added effect:





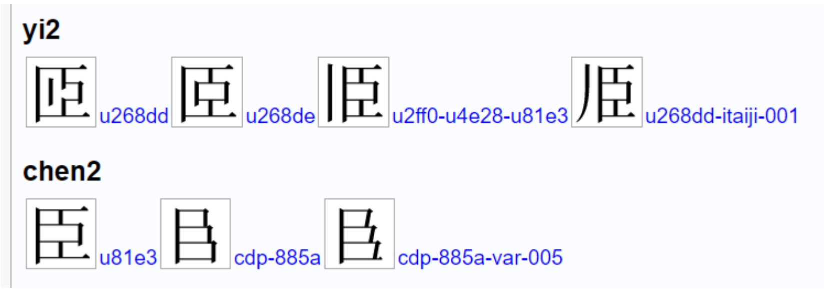




A single "proper" form should be specified for SC and FS counting across all regions.

For this proposal, some rules would need to be split into two or more rules.

Assimilation rules may also include variance rules. (Refer to Issue 1B:)



For the top four glyphs, they are a "variance set 1". For the bottom three glyphs, they are a "variance set 2". The compound of these two variance sets is another "assimilation" rule.

Discussion Item 2:

To aid in duplicate removal and identification of similar shape non-cognates, a virtual normalized glyph can be generated for each glyph submitted according to the proper forms specified in the Assimilation rules and the proper form for the Variation rules. By specifying a normalized virtual glyph, SC can also be easily calculated easily.

The SC does not have to follow anyone's convention closely; it only needs to be consistent. In my opinion, IRGN954AR is a preliminary specification of "proper" forms because the first stroke and stroke count is generally from the glyph of the head character.

The generated virtual glyphs, if necessary, can be included directly in the ISO/IEC10646 standard as well. This is similar to UCS2003 glyphs in Extension B, which effectively act as the glyph that other characters are unified to. The UCS2003 glyphs in the past suffer from the problem that the glyph normalization was not consistent. With "proper" forms specified, we can also avoid this problem.