L2/19-346

Title: On Encoding Policy of Gongche Notations and Upcoming Para-ideographs (tentative)
Author: Wang Yifan

[Due to time constraints, the document below contains the unrevised contents which I provided to Script Ad Hoc. I would like to ask experts in UTC to examine the possible impact on the traditional CJK Unified Ideographs framework by accommodating characters described hereinafter in its blocks, in particular:

- **whether those seven Gongche notation characters (and subsequent ones) should be regarded, in the CJKUI's concept, as of no difference than the existing;**
- **whether CJKUI would be still equivalent to Han characters, as described in the current Unicode standard, after the planned inclusion of those Gongche characters;**
- **whether the inclusion of those Gongche characters to CJKUI indicates that the inclusion of those in Section 3.2 is also openly allowed in future, unlike previous characters (Section 3.3) that derived from each finite set of respective standards**

]

1. Introduction

This document discusses the problem newly posed by Gongche characters (see WG2 N4967) to the paradigm of CJK Unified Ideographs (thereafter CJKUI), upon its inclusion to the block being discussed on WG2 and IRG. I find that this problem is something has not previously been discussed systematically, but having been, it will benefit future addition of a non-definite number of East Asian characters.

2. Summary

The rest of this document argues that:

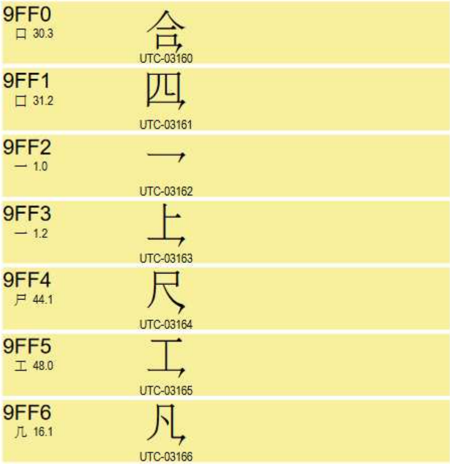
- The new Gongche characters can be placed among a bigger set of characters which are marginal in relation to CJKUI (or usual Han)
- The marginal characters have issues on compatibility with CJKUI, or its encoding process
- We need measures to make some distinction between the marginal characters and CJKUI

The document does **NOT** argue for the following, in case of possible misunderstanding:

- Gongche notations are not ideographs because they are musical notations
- Move/deprecate existing characters or duplicate them in some new block

3. Problems

3.1. Gongche and Unified Ideographs



9FF0	口 30.3	UTC-03160
9FF1	囗 31.2	UTC-03161
9FF2	一 1.0	UTC-03162
9FF3	一 1.2	UTC-03163
9FF4	尸 44.1	UTC-03164
9FF5	工 48.0	UTC-03165
9FF6	几 16.1	UTC-03166

There is an issue regarding seven Gongche notation characters, which have been being assigned code points in the CJK Unified Ideographs block as of the latest CD of ISO/IEC 10646:2019 (WG2 N5032), and then in the Ext. B block for the next CD (WG2 N5100; M68.01), which is contrary to the WG2 recommendation M67.10 for encoding them in a new block named "CJK Unified Ideographs Supplement." Currently recorded discussions are seen in WG2 N5020, WG2 N5066, and WG2 N5106 as far as I have found. Excerpts below:

> *Mr. Andrew West: There are other already encoded characters similar to these. 6 dot above and 2 dot above are examples of music symbols, from HKSAR, are already encoded in the URO.*[1]
> *Mr. Peter Constable: Their use being exclusively for Musical notation, there should not be any issues for unification.*
> *Dr. Lu Qin: No one in IRG is against encoding these – but the opinion is that they are unlike the other ideographs used for text.* [...]
> *Prof. Kyongsok Kim: I think these are symbols derived from regular ideographs.* [...]
> *Mr. Peter Constable: Thinking of analogy to Latin script – we have symbols derived from Latin script in a separate block. These are there only for legacy compatibility reasons.* [...]
> *Dr. Lu Qin: The hooks at the bottom make these symbols totally unrelated to the meaning of ideographs above the hook.*

> *These "Gongche" characters used for musical notation as the symbol have different characteristic and different usage from CJK unified ideographs. Further, the differences of shape from base characters cannot be considered in the framework of existing unification process.*

> *This proved unpractical for code chart production reason. After all, these characters have radical and stroke count like any other CJK ideograph, information which cannot be easily conveyed in a symbol block. After further feedback, it was considered easier to add them to the more convenient CJK addition area (i.e. end of the URO block starting at U+9FF0)*

> *The block was called* ***CJK Unified Ideographs Supplement*** *(emphasis added). Therefore, they were*

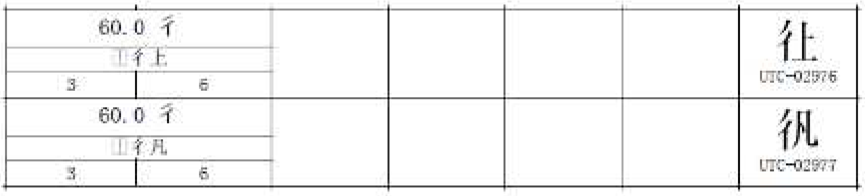
[1] From the context, it seems to refer to 鿈 (U+9FC8) and 鿉 (U+9FC9) in Table 1 of WG2 N4967.

already categorized as CJK unified ideographs. [original emphasis]

Multiple points are involved in the discussion above. Here I would like to reorganize main issues and state my observations as follows.

3.1.1. Are the 7 same with regular CJKUI?

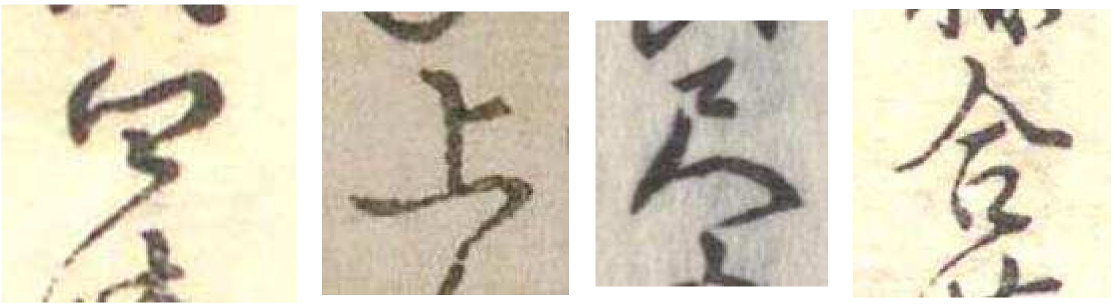
In the light of original proposal (WG2 N4967), no. It also mentions other two unencoded Gongche notation characters that are included in IRG Working Set 2017, which currently in review. That means IRG has accepted them and their undisputed identity as CJKUI candidates unlike those seven.



60.0 彳 ⿰彳上 3 6					UTC-02976
60.0 彳 ⿰彳凡 3 6					UTC-02977

Due to the inherently composite nature of Han script, adding "diacritic" components on an existing character does not immediately make them deviate from the framework of Han, so that a large number of transcription characters, mathematical characters, and regional "annotated" characters (cf. Vietnamese reading marks: WG2 N4915) have place in the CJKUI blocks. This is because the resulted characters still do not violate Han's structure or stroke patterns. Here is the major difference between them and the seven characters: the latter contain an exotic element as the printed style of Han.

Table 1: Handwriting character with swash (from CODH Dataset[2])



四	上	尺	合

The final swash as what is seen in those characters is a common trait in vertical (cursive) handwriting but not distinctive, thus is generally discarded in the printed style and subject to unification. The swash-like element could otherwise not appear in the character-final position of a full CJKUI (which should use printed serifed typeface as the representative glyph). This is what is meant by "some characters' appearances do not conform to the general principles of writing Chinese ideographs" in the original proposal.

[2] http://codh.rois.ac.jp/dataset/index.html.en

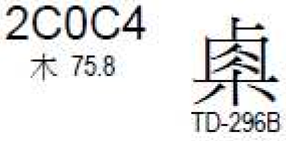
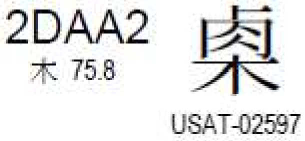
Therefore conceptually, they are more analogous to Letterlike Symbols and Arabic Mathematical Alphabetic Symbols by fixating otherwise unmeaningful features.

3.1.2. Do the 7 affect other CJKUI?

It can be said that there is a certain impact on CJKUI if the seven character are included in the blocks. While it may seem that adding a character has nothing to do with other characters in general, CJKUI has a mechanism of IDS (Ideographic Description Sequence) that serves both public and IRG use. IDS is formalized to allow CJKUI[3], Radicals, and CJK Strokes as its components (as in ISO/IEC 10646:2017 Annex I; Unicode 12.0, p. 725), which means a new character in the CJKUI is available for description of other characters' structures.

As IRG heavily relies on mechanical IDS checking, accepting ambiguous shapes as input could be dangerous. Some of the Gongche characters have simple and versatile-looking shapes that can be useful in the middle of character for users who want to describe a character as faithfully as possible. However, such usage often results in non-confluent decomposition that makes machine checking difficult. IRG has not been free from duplicate characters induced by nonstandard IDS.

Table 2: An actual duplicate error by incompatible IDS

Code Chart	 2C0C4 木 75.8 TD-296B	 2DAA2 木 75.8 USAT-02597
IDS for IRG	⿳⺊冈木	⿱⿰肉一木[4]

As those seven characters evidently only make sense en bloc and have no ability to derive, it is safer to exclude them from possible IDS components.

Another concern is that they may disturb the unification rules of IRG. IRG reviews if a newly submitted character has significant difference with existing characters following the general rules of UCS and its own precedent list. Both, however, assume that characters have valid structures as Han. If we accept characters with undefined ambiguous structure, we might not able to make rule-based unification judgement on that part.

Especially some of those Gongche characters have simpler constructions that is very possible to have confusable characters in future submissions. In this case, we could only conduct unification on a case-by-case basis which takes great resources[5] while IRG usually processes 100+ of characters each meeting. It can undermine the accuracy of IRG checking, the consistency of the unification policy, and the

[3] Any Ideographic characters can be used, but the Unicode standard discourages from using non-CJKUI characters for CJKUI.

[4] Sic; should still be ⿵肉一.

[5] It should be noted that what meaning the character currently has (being a musical notation or not) is *not* a decisive factor when discussing unification (or the non-cognate rule).

reliability of the standards.

3.1.3. Are the 7 worth a distinct status?

As expressed in cited comments, opening a new block or making some special treatment only to accommodate those seven exceptions may be too inefficient. However, there are some CJK characters from other fields, in a relatively small (compared to the entire CJKUI set) but unignorable number, that just behave like the seven Gongche characters.

They share common characteristics with said Gongche characters, namely:

- Not having a valid structure as normal Han characters
- Nevertheless usually intermixed with Han characters seamlessly

I can provisionally define them as **para-ideographs** in this document. They have long existed in the actual world but rarely been brought into standardization, presumably postponed because their unruliness. While the Gongche notation form its own system, I believe that it is more helpful and effective to deem those remaining Gongche characters forerunners of this bigger group. The actual examples of them are given in the next section.

3.2. Upcoming problematic characters

Currently we are not able to estimate the upper bound of the number of para-ideographs, but we know at least three separate, stable origins. The characters from these origins are expected to be an open class as much as CJKUI itself, except being considerably smaller. They are respectively:

a. **Mystical characters in religious documents**

   Unusually shaped characters that have fixed pronunciation or meaning appear in Buddhist and Daoist (Taoist) chants, spells, or other contents. I can provide basic information below for some Buddhist characters.

b. **Regional characters incorporating non-Han element**

   There are domain-specific proper name or shorthand characters partially comprise other letters as their components. I can provide basic information below for some of such characters widely recognized in Japan.

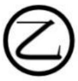
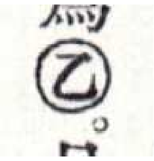
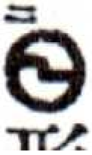
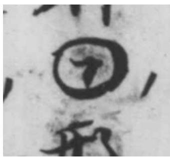
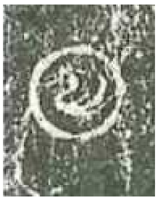
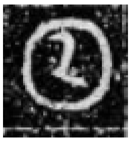
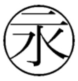
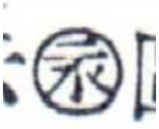
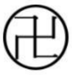
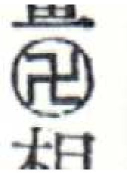
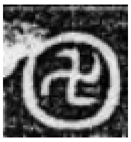
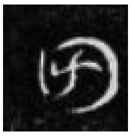
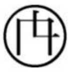
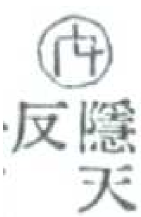
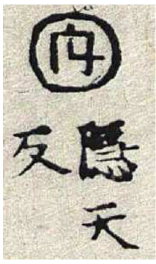
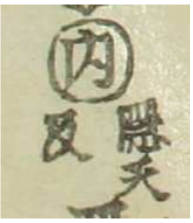
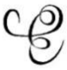
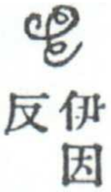
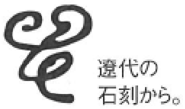
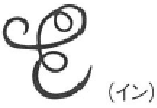
c. **Indigenous characters in non-Chinese, essentially Han-script writing systems**

   Some writing systems that has adopted Han system may integrate a small number of non-siniform components. I can provide basic information below for so-called “pictograms” (象形字) in Sawndip (Zhuang characters).

3.2.1. Buddhist characters

Buddhist scriptures also contain a number of “original” Zetian characters which have not undergone later Kai-like stylization (楷化). Since the inventor, Empress Wu (武則天 Wu Zetian), was an ardent patron of Buddhism, many religious texts written and compiled during her reign, which have been circulated widely in East Asia, reflect those glyph shapes.

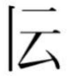
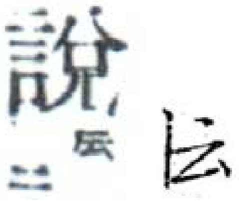
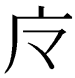
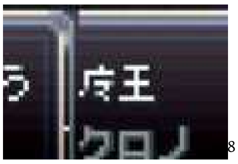
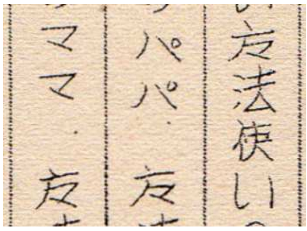
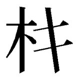
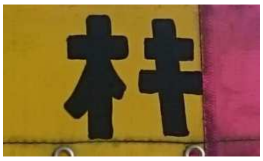
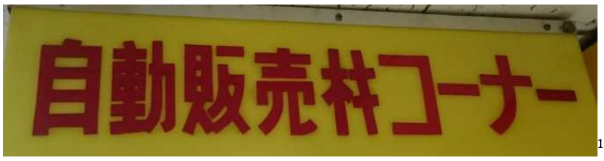
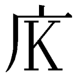
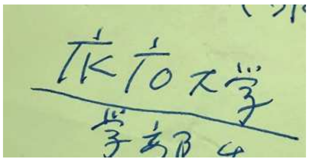
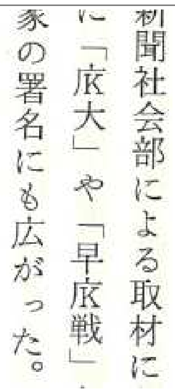
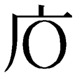
Table 3: Examples of Buddhist para-ideographs

Glyph	Description	Images
	Reading: *rì* (Mandarin) Meaning: "sun, day" Zetian character for 日 that later transcribed as 囸. Not an enclosed ideograph. The component in the middle usually looks like 乙, but not very stable.	   [6] 
	Reading: *yuè* (Mandarin) Meaning: "moon, month" Zetian character for 月. Not an enclosed ideograph.	
	Reading: *yuè*(?) (Mandarin) Zetian character for 月 that later transcribed as 匨, but also used as a Zen symbol for obscure meaning. The inner element is often wavy and sometimes reversed.	  
	Reading: *yān* (Mandarin) An incantation character in 釋摩訶衍論 with unknown meaning. Some editions seem to have 内 inside.	  
	Reading: *yīn* (Mandarin) An incantation character in 釋摩訶衍論 with unknown meaning. Said to be attested in epigraphy as well.	  

[6] http://www.robundo.com/robundo/blog/?p=4341

3.2.2. Japanese characters

Table 4: Examples of Japanese regional para-ideographs

Glyph	Description	Images
	Reading: *to-iu* (Japanese) Meaning: “named, quoth” An old katakana-kanji ligature (ト云) for the phrase. The shape could be either that ト and 云 share the center bar, or 云 with a leftward slash.	 [7]
	Reading: *ma* (Japanese) An abbreviation for 魔 or 摩 replaced with homophonous katakana マ. Heavily used in handwriting and lo-res display.	 [8]  [9]
	Reading: *ki* (Japanese) Meaning: “machine” Abbreviated 機 with homophonous katakana キ, mostly for the meaning “machine”, widely used in sign lettering.	 [10]  [11]
	Reading: *kei* (Japanese) Abbreviated 慶 in handwriting with homophonous Latin *K*, mostly for proper name Keio (a renowned private school).	 [12] 
	Reading: ō (Japanese) Same as above, *O* for 應(応).	

聞社会部による取材に
に「大」や「早戦」
家の署名にも広がった。

[7] http://komonjo.rokumeibunko.com/binran/goji.html

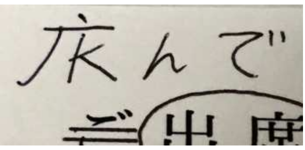
[8] https://twitter.com/sc2rm/status/814273830579408896

[9] http://mahoroba3.cocolog-nifty.com/blog/2018/12/post-b2a3.html

[10] https://twitter.com/inchorin/status/890157700125933568

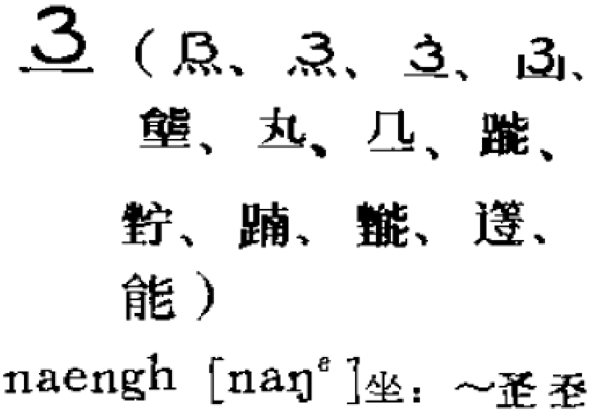
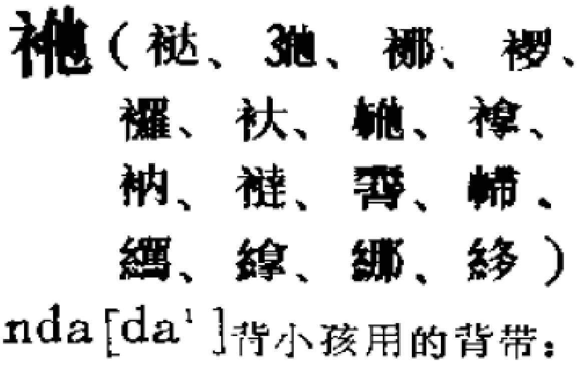
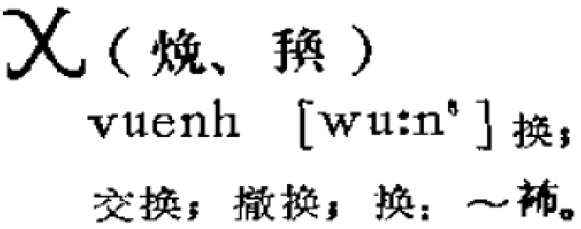
[11] https://twitter.com/inchorin/status/890987250640502784

[12] https://twitter.com/ketokate/status/951049551393177600

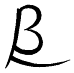
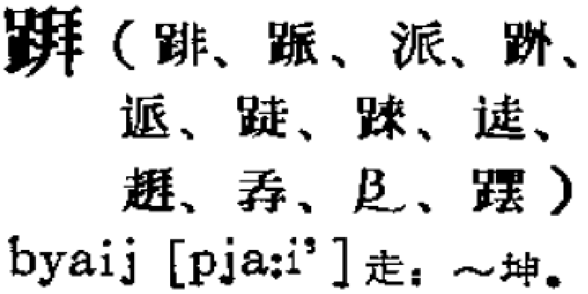
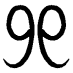
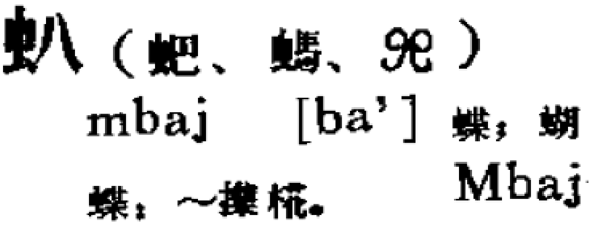
		[13]

3.2.3. Sawndip characters

Table 5: Examples of Sawndip para-ideographs

Glyph	Description	Images
3	Reading: *naengh* (Zhuang) Meaning: “sit down” Some Sawndip characters contain a shape resembles the Arabic numeral 3, that represents a person. This character has multiple variants with the person shape combined with different abstract strokes.	 naengh [naŋ^6]坐：～歪歪
3他	Reading: *nda* (Zhuang) Meaning: “baby carrier” This character integrates the person shape, as the semantic component, with a normal Han element 他 as phonetic.	 nda [da^1]背小孩用的背带：
乂	Reading: *vuenh* (Zhuang) Meaning: “exchange” The character is composed of two crossed wavy lines, as if the script *x*. Note that another character with two crossed straight lines (*ca*, *cah*, or *cax*) is also found in Sawndip.	 vuenh [wu:n^6] 换； 交换；撤换；换：～裤。

[13] https://twitter.com/fdrbdr/status/1021665154050609153

	Reading: *byaij* (Zhuang) Meaning: “walk” A shape looking as if the person shape and a component like ⻌ has merged together.	 byaij [pja:i'] 走：～坤。
	Reading: *mbaj* (Zhuang) Meaning: “butterfly” A pictogram with symmetrical strokes that depicts the symbolic large wings of a butterfly.	 mbaj [ba'] 蝶；蝴蝶：～撲梹。 Mbaj

3.3. Characters with similar trait in CJK Unified Ideographs blocks

I would also like to account for characters already accepted as Unified Ideographs with unusual representative glyphs which may have been deemed or resemble para-ideographs in any way.

Table 6: Existing CJKUIs with unusual shapes

Code	**4E44**	**3403**	**3514**	**3AB3**	**3AC8**	**20137**	**201AD**	**201C7**	**2034B**
Glyph	乄	㐃	㔔	㪳	㫈	𠄷	𠆭	𠇇	𠍋
Code	211A2	219B9	219D1	22013	26B99	26E57	2A708	2CEFF	2CF02
Glyph	𡆢	𡦹	𡧑	𢀓	𦮙	𦹗	𪜈	𬻿	𬼂
Code	2D047	2D37B							
Glyph	𭁇	𭍻							

Not all of them, however, should be classified as para-ideographs. Many of them merely retain cursive loops faithfully as administrative proper name characters, and Han users can easily equate them with shapes whose loop element disconnected.

What rightfully regarded as para-ideographs, that cannot be reduced to valid Han strokes, are half of them: U+3403, U+3514, U+3AB3, U+3AC8, U+20137, U+201AD, U+219B9, U+22013, U+26B99, U+2CF02. Two characters, U+4E44 and U+2A708, are even highly questionable as ideographs if not errors, but the former may be counted in as it is inherited from the original standard. Although U+2CEFF is a variant of U+2CF02, its structure is not totally invalid as CJKUI.

Table 7: Details of existing para-ideograph equivalents in CJKUI

Glyph	**Code**	**Description**

乄	(4E44)	Contains an unusual stroke PG (only for this character). Grandfathered in from JIS X 0212. Actually identical to 〆 (U+3006; *shime*).
㐃	3403	Contains illegal straight diagonal lines. Grandfathered in from PKS C 5700-2 1994.
㔔	3514	Contains an unusual stroke Q (only for them), which is actually hangul jamo *ieung* ㅇ here. Grandfathered in from PKS C 5700-2 1994.
㪳	3AB3	
㫈	3AC8	
𠄷	20137	Seal script glyph that corresponds to 𠫼 (U+20AFC). Contains illegal straight diagonal lines. Seemingly grandfathered in from Dai Kanwa Jiten.
𠆭	201AD	A variant Seal glyph that corresponds to 侌 (U+4F8C). Contains illegal complex curves. Grandfathered in from important source dictionaries.
𡦹	219B9	Variant of 𠕍 (U+2054D) which contains unusual complex curves. Grandfathered in from important source dictionaries.
𢀓	22013	A variant Seal glyph that corresponds to 巨 (U+5DE8). Contains illegal complex curves. Seemingly grandfathered in from Dai Kanwa Jiten.
𦮙	26B99	An archaistic variant of 葵 (U+8475) which contains unusual curves and illegal straight diagonal lines. Seemingly grandfathered in from Dai Kanwa Jiten.
𬼂	2CF02	Cursive variants of 也 (U+4E5F) mostly for Japanese particle *nari*. Contains unusual handwriting curves. There was criticism against them in the ballot for not having normal glyph shapes (WG2 N4656).

They were originally included in the blocks as finite sets in each standard, with no other remedies available. Even though they already have the status as CJKUI, the same problem as that of new para-ideographs equally applies, thus we need some means to correctly handle them.

4. Possible Solutions

4.1. New block

The most straightforward solution to new para-ideographs is to put them in separate block(s) from CJKUI, so that they can bypass regular CJKUI-internal restrictions and discussions. We can also potentially divert the idea of Supplementary Ideographs proposed by Japan (WG2 N4948). However, certain attention must be paid lest it be a “garbage dump” of East Asian characters.

Characters in the block should have following properties:

```
sc=Hani
```

```
Ideo=Yes
UIdeo=No
```

The currently estimated maximum number of additions from each field in Section 3.2 is respectively, around 10 for Buddhist; under 20 for Japanese; and under 20 for Sawndip. Of course, there would be extra calls from other sources.

4.2. Special property or annotations

Para-ideographs can be marked with special property(ies) or some flag(s) in order to signal that they should be handled separately from other CJKUI. The semantics of the property will be, for example, that the character contains illegal shapes, that it cannot be used as a general CJKUI component, that it has a specially assigned identity, or that it has not undergone normal IRG processing. Consequently, other parts of the standards, such as the IDS grammar, may have additional changes.

This is desirable anyways for existing non-conformant characters listed in Section 3.3, including those whose identity as ideograph is also doubted.

4.3. Additional CJK Strokes or components

It is also possible to accept all characters as having right shapes within CJKUI and record their peculiar strokes as independent constituents of CJKUI structure. By doing this, we can at least quantize those features to correctly describe the glyph structures and put them on the same table where we discuss the unification rules. In this case, we must amend new entries in the CJK Strokes block, and possibly register some components equivalent to multiple strokes somewhere, should we encounter shapes which we cannot decide how to break down.

Note that CJK STROKE PG (U+31E2) and CJK STROKE Q (U+31E3) are standing examples of such effort to incorporate exotic strokes (see Table 7).

References

Chan, Eiso, et al. (2018). “Updated proposal of Gongche characters for Kunqu Opera” [WG2 N4967].

Collins, Lee, and Ngô Thanh Nhàn (2017). “Proposal to Encode Two Vietnamese Alternate Reading Marks” [WG2 N4915].

Guangxi Zhuangzu Zizhiqu Shaoshu Minzu Guji Zhengli Chuban Guihua Lingdao Xiaozu 广西壮族自治区少数民族古籍整理出版规划领导小组办公室 (ed.) (2012). *Sawndip sawloih* 古壮字字典. Nanning: Guangxi Minzu Chubanshe.

Japan (2018). “Proposal to start the development of CJK Regional Supplementary Ideographs and terminate the development of CJK Unified Ideographs” [WG2 N4948R].

Sasahara, Hiroyuki 笹原宏之 (2006). *Nihon no kanji* 日本の漢字. Tōkyō: Iwanami Shoten.

Sasahara, Hiroyuki (2017). *Nazo no kanji: yurai to hensen o shirabete mireba* 謎の漢字：由来と変遷を調べてみれば. Tōkyō: Chūō Kōron Shinsha.

SAT Daizōkyō Text Database. http://21dzk.l.u-tokyo.ac.jp/SAT/index.html

Suignard, Michel (2015). “Disposition of comments on PDAM2.2 to ISO/IEC 10646 4th edition” [WG2 N4656].

Suignard, Michel (2019a). “Additional repertoire for ISO/IEC 10646:2019 (6th ed.) CD-2” [WG2 N5032].

Suignard, Michel (2019b). “Draft disposition of comments on ISO/IEC CD.2 10646 6th edition” [WG2 N5066].

Suignard, Michel (2019c). “Additional repertoire for ISO/IEC 10646:2019 (6th ed.) CD-3, Draft” [WG2 N5100R3].

Suignard, Michel (2019d). “Disposition of comments on ISO/IEC CD.2 10646 6th edition” [WG2 N5106].

Takuhon Moji Database 拓本文字データベース. http://coe21.zinbun.kyoto-u.ac.jp/djvuchar

Umamaheswaran, V.S., and Michel Suignard (2019). “Unconfirmed minutes of WG 2 meeting 67” [WG2 N5020].

(End of Document)