IRG N2744

Date: 2024-10-20

Source:	Gen Kojitani (小路谷 源)
Title:	Reply to N2741
Status:	Individual Contribution
Action:	For consideration by IRG

There has been much discussion about a proposal I made to add some ideographs to Unicode, including some script-hybrid ideographs.

As a general opinion, I personally believe that these Script-Hybrid ideographs should be encoded as CJKUI in principle.

This is because these characters are clearly ideographs that carry both semantic and pronunciation information in a sentence, are used in sentences in languages that use Han ideographs (Chinese, Japanese, Korean, etc.), are not phonetic characters, and are treated as equivalent to Han ideographs (such as **abbreviations** and **ligatures** of specific Han ideographs).

I feel that it is unfair to exclude them because they “look very different,” since characters such as “𠓛”(U+201AD; old alternative form of 霧) and “𦲙”(U+26B99; old alternative form of 葵) are already encode.

I felt that the issue was too complicated to be simply called “Script-Hybrid”. I also regret that there were problems with the content of my previous proposals. Below I will give my views on each case.

1. Han-Katakana, Han-Hangul, Han-Bopomofo ideographs

I personally feel that there is no problem with encoding Han-Katakana, Han-Hangul (and Han-Bopomofo) ideographs as normal CJKUI, since it is easy to define where the brush/pen left off, i.e. the number of strokes. However, I also understand that in these cases, the criteria for merging with similar characters is an issue.

In **N2741**, CheonHyeong Sim mentioned component cognition as a criterion for determining whether or not characters with the same shape but different origins should be merged, but I think that **it is okay to merge** some (the reason why I say "some" will be explained later) of the Han-Katakana, Han-Hangul, and Han-Bopomofo characters **even non-cognitive ones,** i.e. Level 3 ones.

character	similar shape	component cognition	same script
机 (jī & つくえ)	○	○	○
舫 (liz/lǎc & カッター)	○	○	×
軌 (boek/pū & トラック)	○	×	×

Fig. 1: A table explaining cognitive-levels in N2741 by CheonHyeong Sim. The first row defines Level1*, the second row defines* Level2*, and the third row defines* Level3.

The reason is that in these cases, there is no difference in the shape of the character under any circumstances, and the radicals referenced when looking up the dictionary do not change. One of the concerns is that carelessly disunifying existing characters with the same shape may lead to problems such as the **homoglyph problem**.

The example of 月偏/肉月 (肭/朒/腧/腠 & 肭/朒/腧/腠) was given, but in this case, the radical is different, “月” or “肉”, so the index when referring to the dictionary is different, and in Taiwan and Hong Kong, the two are distinguished in terms of their character shapes (“肉” is written in a shape called “提肉旁, i.e. 月”). Therefore, it is necessary to distinguish between them.

On the other hand, as long as the Katakana “ト” and the Han ideograph “卜” are written in the same font, there is no difference in the character shape that distinguishes them. However, in the case of “⿰木丰(alternative form of 材)” / “⿰木キ(Japanese abbreviation of 機)”, the katakana “キ” is usually written with all strokes slightly tilted, and especially the third vertical stroke is rarely written without tilted, so I think that it should be distinguished.

In addition, since the non-Han components of Script-Hybrid ideographs represent pronunciation, when there are non-Script-Hybrid ideographs with the same glyph, if the part that is the same as the non-Han component is not a radical, the radicals are not different between the two, so I think there is no need to distinguish them.

In other words, when the glyphs are the same,

(1) there is no case where they are written differently on the glyph

(2) the radical does not change

I feel that there is no need to distinguish them (at least intuitively) even if they are non-cognate.

However, I am aware of the counterexample case of “叱/吒.” 叱 (U+53F1) is pronounced *chì* in Chinese and means “to scold,” and was originally written as ⿰口七 (U+20B9F 吒) and now written as ⿰口匕 as a result of a misinterpretation of the character. On the other hand, 吒(U+2B738) is pronounced *huà* in Chinese and means “to open the mouth.” The “匕” part of this character is a component that represents the sound, and this character was originally ⿰口匕. In this case, the character is distinguished on the basis that it is clearly **non-cognate**.

Also,

> *I would like to have one more queion for Gen Kojitani. Why did you exclude the four Han-Katakana ideographs mentioned at the top of this document in your newest L2/24-201 document? What is the essential diference between ⿸广マ and ⿰言コ?*

The answer to this question is that there is no essential difference between ⿸广マ and ⿰言コ.

I regret that there was a problem with the way I proposed this. Below I will explain the process of submitting these proposals.

I first found abbreviated characters used in Japan that had not yet been encoded, and applied for addition to Unicode with priority to those for which examples were relatively easy to find. The set of proposals at that time included script-hybrid ideographs (⿸广マ, ⿰木キ, ⿱宀R), but at the time I didn't think they would become such a big source of controversy.

Later, Dr. Ken Lunde contacted me about the problems with these script-hybrid ideographs, and as a compromise, I created a new proposal to encode only non-script-hybrid and Han-Katakana ideographs as regular CJKUI. This is **L2/23-139**.

After further research into script-hybrid ideographs, I found many similar ideographs and examples that were not in L2/23-139, and decided to submit another application for these, but after consulting with Ken Lunde, I came to the conclusion that I should follow the IRG's guidelines at the time, which stated that these script-hybrid ideographs should not be uniformly considered as CJKUI. The proposal at that time was **L2/24-125** and its revised version is **L2/24-201**.

Perhaps I should have thoroughly investigated other Script-Hybrid cases such as those included in L2/24-201 when creating L2/23-139 before creating the proposal document. Or perhaps the fact that I lumped together Script-Hybrid (and non-Script-Hybrid) abbreviations in the same proposal may have caused confusion.

In any case, my personal opinion is that there is no essential difference between ⿸广マ and ⿰言コ, and both should be treated as normal CJKUI.

2. Han-Hiragana ideographs

I deleted it from the current latest proposal L2/24-201 because there were not enough examples (I could only find one example in an advertisement for a company called *Yanmar Diesel*), but I later found another example of the abbreviation 木き for 機 (see images below), so I will also touch on this case.

Fig. 2: Yanmar Diesel's advertisement. This picture is posted on X by @hakkaku_culture on 13 May, 2018, URL : https://x.com/hakkaku_culture/status/995334947077435393/photo/1

Fig. 3: This picture is posted on X by @_iyo1995 on 24 August, 2019, URL : https://x.com/_iyo1995/status/1165040587378057217?s=46&t=fTi4aDUJFwilJX4jEWviTg

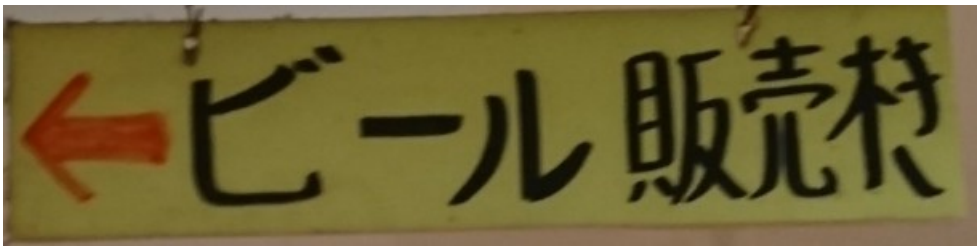


Fig. 4: This picture is posted on X by @1_ichi1 on 26 October, 2019, URL : https://x.com/1_ichi1/status/1188039056824750086?s=46&t=fTi4aDUJFwilJX4jEWviTg

Fig. 5: This picture is posted on X by @gips_apple on 20 August, 2020, URL : https://x.com/gips_apple/status/1296301194063695873/photo/2

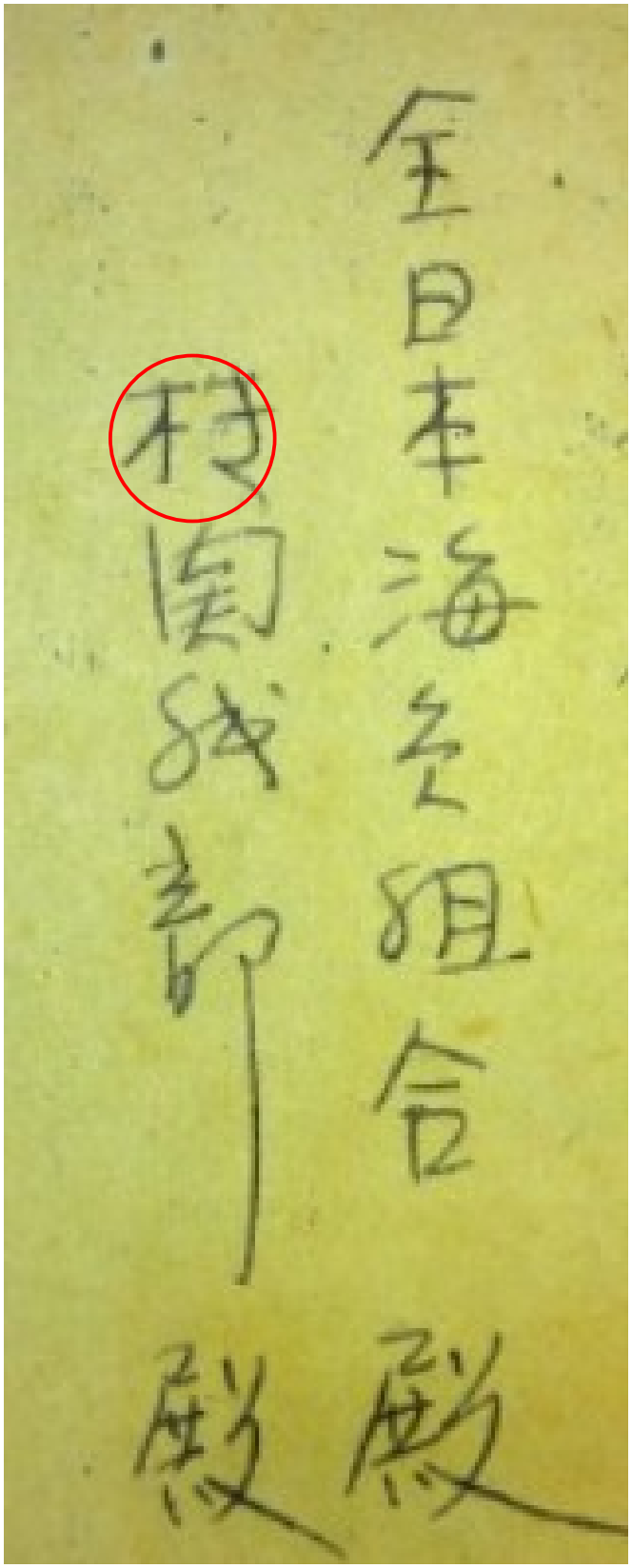


Fig. 6: This picture is posted on X by @rei_akao on 9 December, 2020, URL :
https://x.com/rei_akao/status/1336687000586350592/photo/2

First, my conclusion is that since “𪜊” (*hagata*; U+2D92A) is encoded as CJKUI, it should be encoded as CJKUI. Hiragana and Hentaigana are writing systems used together with Han Ideographs, and I feel there is no reason why characters that have been partially replaced with hiragana as phonetic symbols should not be considered Han Ideographs.

In the case of Han-Hiragana, I don't think there is a big problem in defining the number of strokes because the place where the brush is released is clear for most characters (in fact, it is self-evident that “𛂦” (U+1B0A6; *hentaigana ha*), the left component of 𪜊, is two strokes), but unfortunately, there is ambiguity in the definition of the number of strokes, especially for “**き**”. The number of strokes for “き” changes depending on whether the upper vertical stroke and the lower stroke are written in one stroke or two.

left one is in UD デジタル教科書体 N-R, right one is in Noto Serif JP.

The same problem can occur with other hiragana characters such as “**さ**”(total stroke is 2 or 3), “**そ**”(total stroke is 1 or 2), “**り**”(total stroke is 1 or 2), and “**ゑ**”(total stroke is 1 – 3).

Table 1: Differences in appearance between different hiragana fonts

Noto Serif JP	き	さ	そ	り	ゑ
UD デジタル教科書体 N-R	き	さ	そ	り	ゑ
SimHei	き	さ	そ	り	ゑ
テンドウ EL	き	さ	そ	り	ゑ
Mgen+1m light	き	さ	そ	り	ゑ

My opinion on this is that the number of strokes of hiragana should be defined according to the typeface called “教科書体(***kyōkasho-tai***; textbook typeface)” that is used in education as the “normative” typeface for Japanese. In Japanese education, “き” is usually written with the top and bottom parts separated, and the number of strokes is considered to be 4. In the same way, I think it is correct to consider “さ” as 3 strokes, “そ” as 1 stroke, and “り” as 2 strokes.

As for “ゑ”, it is not currently taught in education, and there is ambiguity between 1 – 3 strokes depending on the typeface, but I think it is correct to set it to the minimum value of 1, as the kTotalStrokes of “草” is 9 (not 10). Also, the fact that “ゑ” is often recorded as a 1-stroke character form in Japanese *kyōkasho-tai* fonts justifies the number of strokes being 1 (see row 2 in Table 1).

3. Han-Latin, Han-Sawndip ideographs

I recognize that the biggest Pandora's box I have opened is the case of **Han-Latin**. Latin characters look very different from Han Ideographs, so there are questions about whether or not to han-ideographize (to normalize letter shape to Han-ideograph-style) their appearance, and the number of strokes is difficult to define. In addition, it is a writing system used in different cultural spheres (however, in recent years, due to the influx of Western culture, it has begun to be used in general texts even in Han Ideographs culture spheres), so I am unsure whether it should be encoded as normal CJKUI or not, and have not yet reached a conclusion. However, I think it is rather correct to encode it as CJKUI.

Sawndip(方块壮字/古壮字), which is used to write **Zhuang** and **Bouyei** languages, also has Script-Hybrid ideographs, and it can be said that it is in exactly the same situation as Han-Latin.

(This is a list of Sawndip characters that I have compiled that are unclear whether they can be integrated with CJKUI, their pronunciations, and their meanings.)

12		aemq [am˦] <古壮字字典, p1-2-1-M> {⿰背今/挭/⿰背包/⿰持⿱人長/蔭/⿰身包/⿰/⿰/⿰身恩/軀} to carry something on one's back	ams [ʔam˦] <布依方塊古文字, p3-c1-r3> ①to carry something on one's back ②to obey eil [ʔɯi˦] <布依方塊古文字, p64-c1-r3> to carry something on one's back
13		-	aml [ʔam˧] <布依方塊古文字, p3-c1-r2> to hit; to throw
14		-	aml [ʔam˧] <布依方塊古文字, p3-c1-r1> a molar

Fig. 7: examples of Sawndip and Han-Sawndip hybrid ideographs

I think that the question of "***Should the appearance be han-ideographized?***" should be considered for each character separately. If there are existing examples of the ideograph that have been han-ideographized (e.g., ⿵门丅 for ⿵门T), then it should be han-ideographized as per the source, whereas if there are no examples (e.g., ⿸广K or Han-Sawndip), then I feel that it should not be han-ideographized. In particular, if there are no examples by Kaishu(楷書), then it should not be han-ideographized. I feel that only a few people can see "宊" and recognize it as "⿱宀R".

In the case of ⿰メ也, it may be difficult to judge because one of the sources has been han-ideographized while other sources have written it as "X 也" (however, the "X 也" notation seems to be related to technical limitations, so it seems correct to han-ideographize it. Further investigation of more sources may be necessary.)

Fig. 9: from https://macaonews.org/features/x%E4%B9%9Fand-ta-the-gradual-rise-of-gender-neutral-pronouns-in-chinese/

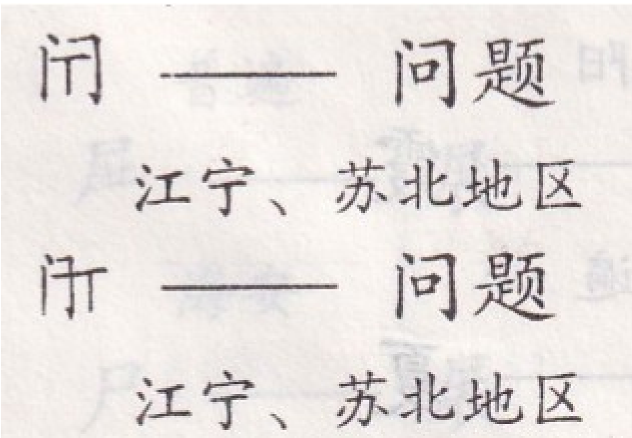


Fig. 8: ⿵门T in 江苏方言总汇

I think the biggest problem with Han-Latin or Han-*Sawndip* is that **the number of strokes cannot be defined**. It is especially difficult to judge the number of strokes that differ depending on whether the font is serif or sans serif (I/l, G/G, etc.) or the part that can be written in a continuous line differently by different people (M, N, etc.). Normally, Latin characters do not have a definition of which part of the character to start writing or where to release the pen, and I am aware that in many Latin writing cultures, the "stroke order" and "numbers of total strokes" is not taught. It is probably the same for *Sawndip*. Personally, I think this problem should be solved in one of the following ways.

1. Following the example of "草", define the **theoretical minimum** number of strokes when writing the Latin character **in normal sans serif fonts** as the number of strokes of the Latin-component. For example, '**l**' has the minimum of one stroke when written in sans serif, and 'M' has the minimum of one stroke when written in a single stroke (see Table 2).
2. The number of strokes in the Latin part is set to **max_value**, and it **is guaranteed to be arranged at the end** of the character with the same radical (姜兆勤's idea)
3. As a revised version of Method 1, in order to eliminate non-intuitiveness, the stroke number can also be the theoretical minimum value for a normal sans serif font when **vertical lines are written only from top to bottom** and **horizontal lines are written only from left to right**. (see table 3).

Table 2: An example of total stroke of Latin letters using Method 1

A	B	C	D	E	F	G	H	I	J	K	L	M
2	1	1	1	2	2	1	3	1	1	2	1	1
N	O	P	Q	R	S	T	U	V	W	X	Y	Z
1	1	1	2	1	1	2	1	1	1	2	2	1
a	b	c	d	e	f	g	h	i	j	k	l	m
2	1	1	1	1	2	1	1	2	2	2	1	1
n	o	p	q	r	s	t	u	v	w	x	y	z
1	1	1	1	1	1	2	1	1	1	2	2	1

Table 3: An example of total stroke of Latin letters using Method 3

A	B	C	D	E	F	G	H	I	J	K	L	M
2	2	1	2	3	3	1	3	1	1	2	1	2
N	O	P	Q	R	S	T	U	V	W	X	Y	Z
3	1	2	2	2	1	2	1	1	1	2	2	1
a	b	c	d	e	f	g	h	i	j	k	l	m
2	2	1	2	1	2	2	1	2	2	2	1	1
n	o	p	q	r	s	t	u	v	w	x	y	z
1	1	2	2	1	1	2	1	1	1	2	2	1

4. Non-Script-Hybrid Ideographs

Finally, I would like to express my thoughts on ideographs that are not strictly script-hybrid.

First, I think that there is no problem in encoding ideographs that are combinations of only Katakana (□ヤ□キハ, □□モタル), which were included in my previous proposal, as normal CJKUI. As you pointed out, since "寀(*utsubo*)" is a precedent, I think that these "Han-component-free" ideographs should also be encoded as CJKUI.

However, for these, the choice of radicals may be a concern, but some radicals (such as "丨" and "丿") are established only for the purpose of classifying the character shape and are unrelated to the meaning of the character, so I don't think it will be a big problem. I think it is appropriate to classify □ヤ□キハ as "丨" and □□モタル as "一" or "儿."

20	兣	• □□モタル • Japanese abbreviation for word "モルタル(mortar)"
21	悀	• □ヤ□キハ • Japanese ligature of "ヤキハ(yakiba)" • often found in sword books of the Edo era

Fig. 10: Idepgraph consisting of only katakana-components in IRG N2637

As CheonHyeong Sim pointed out in N2741, I think that there is no problem if full-width Latin characters are used in cases where Latin characters themselves are treated as Han ideographs (such as **E** as a variant of 医 in Chinese, **Q** as a variant of 級 in Japanese *Geba-ji* letter, and **e** as an exclamation in *Sawndip*). I also agree with the point about adding Hani to script_extension. (In the first place, I personally think that Hani should be added to all script_extension for full-width Latin characters and digits, since full-width Latin characters and digits themselves are used in vertically written documents such as newspapers in Chinese or Japanese.)

A slightly tricky point is whether Sawndip for Zhuang and Bouyei language without any Han-components can be considered CJKUI, and I think this needs further discussion. In particular, there are characters that do not have any Han-components, such as digits used in Bouyei (see image on next page), so I am unsure whether they should be considered CJKUI.

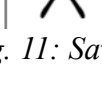
No.	letterform	Zhuang (壯語; zha)	Bouyei (布依語; pcc)
0		-	lingz [liŋ˩] <布依方塊古文字, p152-c2-r2> digit 0
1		-	ndeeul [ɗe:u˧˥] <布依方塊古文字, p206-c2-r6> digit 1
2		-	songl [sɔŋ˧˥] <布依方塊古文字, p247-c1-r1> digit 2
3		-	saaml [sa:m˧˥] <布依方塊古文字, p239-c1-r1> digit 3
4		-	sis [si˧˥] <布依方塊古文字, p255-c1-r4> digit 4
5		-	hah [xa˧˥] <布依方塊古文字, p94-c1-r1> digit 5
6		-	yogt [jɔg˧˥] <布依方塊古文字, p315-c2-r1> digit 6
7		-	sadt [sat˧˥] <布依方塊古文字, p243-c2-r2> digit 7
8		-	beedt [pe:t˧˥] <布依方塊古文字, p21-c2-r7> digit 8
9		-	juh [t͡ɕu˧˥] <布依方塊古文字, p127-c2-r2> digit 9
10		-	sib [sip˧˥] <布依方塊古文字, p258-c1-r4> number 10

Fig. 11: Sawndip digits for Bouyei language

There may be a similar problem with the character [“イン; *in*” that appears in Buddhist scriptures](#).

These are my thoughts on the feedback I received.

(End of document)