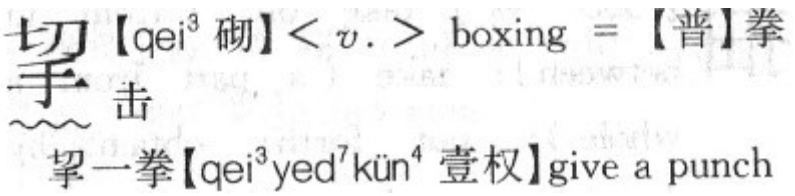
Further feedback on IRGN2551

by John Knightley

for discussion by IRG.

(2022-10-22)

1) Case study 3 uses UTC-00636/UK-20317 but this example does not match the others and should be replaced by a more suitable character like UTC-00119 鬮



Case study 3 refers to the Cantonese specific characters from *Concise Cantonese-English Dictionary* (Yáng Míngxīn 杨明新, 1999). Yang Mingxin did not create these Cantonese specific characters but collected the characters from elsewhere. The vast majority are already encoded, the fact the several dozen of them are not encoded is nothing to shout about and entirely consistent with his claim to have collected them. On page 597 of *Concise Cantonese-English Dictionary*, in Chinese, Yang Mingxin explicitly states that the characters marked with a wavy line underneath are only (仅) those collected from a selection of real(民间) sources, such as Cantonese opera scripts, local chronicles (地方志), etc and even some found in Cantonese publications of other provinces and countries. “本书反映粤方言特殊词汇的文字符号﹏，仅据部分民间资料，如粤剧脚本，地方志等，甚至连境外，国外出版物上的一些写法也吸收过来了。”

Whilst it is correct to say the character was discussed that is about all it has in common with the other 3. Something that would be similar to the others is say UTC-00119 鬮 [illegible]門龍 aka the Monroe character, a character created in modern times by Mr Monroe to represent his own name based on the Japanese pronunciation of it. At the time it was also noted that even if Mr Monroe published books with the character in it this would still not be sufficient to get it encoded. The result would be then four case studies three of characters that never have been submitted to the IRG, and one that may be submitted in the future. It would be good to note whether or not either of the other two have ever been submitted to a member body or simply hypothetical examples that to date no one has ever seriously suggested they be consider for submission to the IRG.

2) Answering the original question and adding IRG prospective.

The original question was if someone publishes a book with characters they have made up, does that mean that the IRG has automatically to accept it? To this the simple answer is, “No it does not.” Furthermore it should be noted that this is also already covered in PnP. If only used by the person who made the character it would be a character with only evidence from a original and questionable source. PnP 2.2 (d) 2 (a) states, “Original Source (證據源限制): The source of evidence must be considered authoritative by IRG, as validated by past literature and IRG experts. IRG has the right to reject characters from questionable sources.”

The next PnP item also gives an solution and reminds submitters of the dangers of submitting characters with only a single source. PnP 2.2 (d) 2 (b) “Multiple Sources (多 源 證 據): Supply character use evidence from multiple independence sources. IRG has the right to reject characters with evidence of use from only a single source, especially if the source is not considered authoritative by IRG.”

In passing it should also be noted, that since the IRG does not accept submissions from the public but members and the members are all groups. These are often easily dealt with. This may not be apparent to others as the deliberations of most member bodies is not a matter of public record.

Putting more IRG prospective would be an improvement, but since the final version would be available on the IRG website making this does not mean that it should necessarily be added to PnP.

3) Analysis of several meanings of 生造字(shengzaozi) etc

Here some of the several meanings of 生造字(shengzaozi) etc are presented to help discussion.

What do 生造字(shengzaozi) and 自造字(zìzàozì) literally ‘self-created characters’ mean? In Chinese linguistics they synonyms, which though sometimes used as a translation of the English ‘coined characters’ or visa versa, because of semantic shift has a taken on a different meaning. The latter also is used as a term in computer science. Let us look at these 3 different usages starting with the meaning usually found in Chinese linguistics since the term was introduced in Chinese.

The first written response was from Toby Tso who says he has concerns about the expression ‘avoid encoding Shēngzàozì Characters (生造字)’, and further states “I agree to avoid[ing] the use of the term ‘Shēngzàozì Characters’ (生 造 字 , literally means ‘made-up characters’) in IRG PnP.” His concern being that if strictly applied then it would prevent the encoding of characters used for ‘different Sinitic languages’ (that is characters used languages other than Putonghua in China). He concludes with a question, “Moreover, as the recent IRG working sets have been dealing with

Sawndip characters (古壯字), if the principle is adopted, will the submission of Sawndip characters be affected and become almost impossible to be encoded?". (see https://appsrv.cse.cuhk.edu.hk/~irg/irg/irg57/IRGN2482FeedbackToby.pdf)

Why did Toby say this? It is because the term 生造字 (Shengzaozi) when used in China by linguists talking about the Han script it usually refers to (1) characters not part of the standard Mandarin Chinese character set or (2) characters used with a meaning different to that they have in standard Mandarin Chinese. Please note here, and below 'standard Mandarin Chinese' refers to those Mandarin characters usually found in dictionaries and so also does not include most characters only used in names of places and people. As a result all characters used for writing Cantonese, Japanese, Korean, Vietnamese, Zhuang and any other language or dialect you can think of that uses the Han script but that are not also used for Mandarin itself are automatically Shengzaozi characters. Not to mention almost all those used only for names of places and people. One established principal of the IRG is that characters are treated the same regardless of the language they are use for, as such I would argue that it would be wrong to add to IRG PnP anything that discriminates based on whether or not a character is a 生造字 (shengzaozi) character in this sense of the word. In this sense it is correct to say a Mandarin dictionary should avoid including 生造字 (shengzaozi) characters, but wrong to say that the IRG should avaoid encoding 生造字 (shengzaozi) characters.

In linguistics the standard translation of 生造字 (shengzaozi) into English would be 'coined characters'. A definition based on the English translation of 生造字 (shengzaozi), and 自造字(zìzàozì), literally 'self-created characters' would be (1) recently created characters or (2) existing characters recently given a new meaning. Of course from the IRG prospective only new characters are of interest not old ones. As Ken points out the word 'modern' is superfluous. What this means in one very real sense when fully understood that when a character was created is irrelevant, since all characters the IRG are either (1) not already encoded characters, or (2) already encoded characters presented in a new way. The main job of the IRG is to decide which is which. Those agreed to be of type (2) it unifies, and those agreed to be of type (1) it encodes. Simply put the job of the IRG is more clearly put as to encode new characters, that is coined characters.

If whether or not something is modern is a critical factor in saying a character is not suitable to encode, then this means that the characters in Yú Shǎoléi's book not suitable at present because they were published in 2011, but would be suitable if they had been published in 1911. Does that not then mean that the characters would become suitable for encoding when it ceases to be modern in say 2061, or if you prefer 2111. There are criteria which make a character suitable for encoding, and it is fair to say the older characters are more likely to reach these criteria than newer ones, but not that all older ones meet them. There are somethings the make a character unsuitable for recommendation by the IRG and these are already mentioned in PnP, namely (1) it is not a Han script character (2) it is a logo or image and (3) it is

unifiable with an already encoded character, that it is not really a new character. I agree with Andrew West who uses the term in this sense and says in his feedback, “IRG has managed for almost 30 years without rules for self-created characters, and I am not convinced that such rules need to be added to the PnP now …”.

In computing 自造字(zìzàozì) are new glyphs that are made on a computer. The most common being user-defined glyphs, that whilst they use the PUA, the glyph data itself is stored on the computer concerned, documents can be printed, but the data itself expires with the physical computer. In theory one could make disc image of such a computer and run it on a virtual machine. There is also commercial software used by publishers that uses the PUA and stores the glyph data in a data file, and documents created this way can be viewed and edited on other machines using the same software. New glyphs can also be stored in a ttf file in the PUA, and this is a requirement for all characters submitted to the IRG. IRGN2551 has the true but somewhat tautological sentence: “To avoid confusion, IRG refers to these newly created characters in computer systems that are not suited for encoding as **private characters**.” Tautological in that these **private characters** not only identified by how the they are stored but also by being not suitable for encoding, hence **private characters** are by definition characters not suitable for encoding.

This section of IRGN2551 is all very interesting, but by the nature of computing would need updating from time to time if added to PnP. Furthermore it seems to have little direct relevance to the work of the IRG, as IRG only considers characters in the public domain.

4) Conclusion

The IRG PnP is sufficient as it stands to solve the original question raised. Just simply because something has been discussed by the IRG does not of itself make it suitable for inclusion in PnP. With the completion and acceptance of ws2021 the IRG will have been instrumental in the encoding of over 100,000 characters and should be congratulated on that. The current road map already has allocated a further 50,000 characters specifically for use by the IRG. With so much space already road mapped, and no obvious candidates for the remaining empty planes, surely we should be answering the question of how best to process the many characters that still need encoding.