IRG

ISO/IEC JTC1/SC2/WG2/**IRGN2482 Feedback**

SOURCE:	Toby TSO
STATUS:	Individual Contribution
ACTION REQUIRED:	To be considered by IRG
№ OF PAGES:	2

Feedback on the Suggestion of IRGN2482 to Avoid Encoding Shēngzàozì Characters or Made-up Characters

2021-09-16

Thanks to WANG Xieyang, IRGN2482 is an excellent contribution.

Regarding the second issue, 'avoid encoding *Shēngzàozì* Characters (生造字)', I have some concerns that if the principle is strictly implemented, the possibility to encode vernacular characters in different Sinitic languages (Chinese languages), especially minority Sinitic languages, will become nearly impossible in the future.

I agree to avoid the use of the term '*Shēngzàozì* Characters' (生造字, literally means 'made-up characters') in IRG PnP, but the requirement that 'the character should be used in running text by someone except the creator', in my opinion, still too harsh if the principle is implemented aggressively.

I believe that there are two main practical reasons for the lack of running text (for vernacular characters in different Sinitic languages, submitted to IRG):

1. **We have to admit that Chinese folk preferred to write in the language that can be understood across the whole of China** — formerly 文言 (Classical Chinese) and now Mandarin. This has resulted in vernacular characters appearing only when necessary, usually in traditional songs or opera scripts that emphasise the local speaking. It is difficult to cover all the vernacular morphemes in these materials, and a language can change very quickly before it is standardised in a written form. I do not have a reference on this, but judging from my current study of the vernacular characters in 陽江, 廣東, a decade is enough time for quite a lot of morphemes to emerge. This is why dictionaries that emphasise Sinographs (CJK ideographs) have to 'make up' some characters.

2. **Nowadays, computers are used to process documents.** Previously, the folk could 'make up'

characters for unique morphemes in their vernacular, and some of the characters 'can be accepted by more and more people and become *Súzì* Characters (俗字)', as mentioned in IRGN2482. But today, a person who wants to write a minority Sinitic language in full Sinographs hopes that the IME to be able to type the characters they want, and in most cases, the only characters available to refer to is in a dictionary full of 'made-up characters'.

If all these 'made-up characters' are rejected, it will be impossible for downstream IME and font vendors to include them, it will be very difficult for minority Sinitic language speakers to produce 'running texts', those only Sinographs (even if they are 'made-up') for unique morphemes will never become common characters (俗字), and a large number of these morphemes will remain without any Sinograph written form. The issue has become a 'chicken or the egg' dilemma.

In my opinion, if a morpheme in a minority Sinitic language have only one case of Sinograph written form, even if it is 'made-up', to encode them is not an abuse of the current mechanism. The second issue in IRGN2482 requires that the submitted character 'should be used in running text by someone except the creator', but even so, it is still possible (for the two reasons mentioned above) that the submitted character may not have been widely used and become a common character. The fact that different books refer to each other and the same 'made-up character' was chosen suggests that in some minority Sinitic language, some morphemes did not develop any common character, and the 'made-up characters' are the only Sinograph written case.

But I admit that it is very difficult to establish the case — the best way is to ask a native speaker to testify, but even this would only lead to a rough conclusion, and would require a great deal of effort on IRG contributors to reach such a rough conclusion.

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?

(End of Document)