Doc Type:	**Ideographic Rapporteur Group Document**
Title:	**Proposal to improve IRG process**
Source:	**Wang Xieyang (王谢杨)**
Status:	**Individual Contribution**
Action:	**For consideration by IRG**
Date:	**2020-09-02**

Here are four parts of the document discussing the following issues:

1.Ways to deal with the characters wrongly put in D-set.

2.Avoid encoding Shengzaozi Characters(生造字).

3.Concretize the process of defining non-congnate characters.

4.The possibility of unifying some rarely used ideographs regardless of their meanings.

If the suggestions are confirmed by IRG, the IRG PnP should be changed accordingly.

1. Ways to deal with the characters wrongly put in D-set

In IRG#56, we found two Vietnam characters wrongly put in the D-set due to editorial error. After IRG#56, three China characters were found wrongly postponed:

<div align="center">

https://hc.jsecs.org/irg/ws2017/app/?find=GXM-00267

https://hc.jsecs.org/irg/ws2017/app/?find=GXM-00303

https://hc.jsecs.org/irg/ws2017/app/?find=GXM-00370

</div>

New evidences of these three characters are posted and accepted in July, 2020(after the time of discussion record), but none of them are moved from D-set to M-set. I don't think we should blame this to anyone while we can absolutely make this happen less.

Here are my suggestions to deal with the problem:

In order to avoid this happening, I suggest we:

1.Skip no comment tagged *unification*, *evidence*, *glyph design & normalization* and *other* in our meeting.

2.In meeting mode, IRG ORT should give a special color to D-set characters who have new comments before or during the meeting.

3.We should get clear agreement from experts concerned or other experts who represent them that their new comments about the D-set character are well discussed in our meeting.

However, we can't assure that this or other kinds of errors won't happen again. So the messures to deal with it should be added to IRG PnP.

I suggest that we create a document for this kind of editorial errors every time we find it. If the M-set is not frozen, the characters wrongly put in D-set should be added back to M-set immediately; if the M-set is frozen already, the characters should be added to the next IRG working set WITHOUT being counted into the quota. The document work can be done by the editors of each source or someone designated.

2. Avoid encoding Shengzaozi Characters(生造字).

Encoding characters created by oneself in modern times can be dangerous. **Provided this kind of characters can be encoded freely, everyone who can afford to publish a book is able to creat a character himself and get it encoded.** That's also why, in practice, we have long been trying not to encode this kind of characters. Usually, we call these as Shengzaozi Characters(生造字).The defination of Shengzaozi Characters(生造字), however, can't be found in IRG PnP.

Considering some Shengzaozi Characters can be accepted by more and more people and then become Suzi Characters(俗字, also called Common Characters), the scope of Shengzaozi Characters is hard to defined. I suggest that we not define Shengzaozi Characters in IRG PnP but add one item to IRG PnP 2.1.1 said

> "**e. Scope of use**（使用范围限制）: The character should be used in running text by someone except the creator. It will be rejected if a character is considered created by one of the authors of the materials or reference which the proposed evidence is derived.

Since place or people name used characters can be used daily, the encoding of them won't be excluded.

In IRG WS2021, UK and UTC submit some dialect used characters which are aparently created by the author. These characters are from《简明粤英词典》(杨明新 著, 广东高等教育出版社, 1999年) or *The Representation of Cantonese with Chinese Characters (Journal of Chinese Linguistics, Monograph Series Number 18, 2002)* without other use in running text. What's more, the actual source of UTC's evidences is also《简明粤英词典》.For example, UTC-00742:

Fig.1 The evidence of UTC-00742 from IRG ORT

There is a book named 《造文字的反 一个草民的造字运动》(余少镭 著, 广州: 花城出版社, 2011 年 1 月) which contains many characters that the author creates. Every character has its own pronunciation, meaning and even running text.

Fig.2 余少镭: 造文字的反 一个草民的造字运动, 广州: 花城出版社, 2011 年 1 月, P7

Basically, the characters in 余少镭's book have no difference with 杨明新's in some way. Thus, characters from the two books shouldn't be encoded unless other reliable evidences can be found. We will be able to reject such characters according to IRG PnP directly if we add an item as above.

3. Concretize the process of defining non-congnate characters.

Here is a part of a chart in the book named《韩国汉文古文献异形字研究》（吕浩 著，上海：上海人民出版社，2013 年 12 月）

主形字	韓國漢文古文獻異形字	敦煌文獻異形字
備	備、俻、備、脩、脩、俻、備、備、備	俻、備、備
邊	邉、劲、劲、邊、邊、邊	逯、邊、迶、邊
稱	稱、稱、稱、稱、稱、稱、稱、稱、稱、稱、稱、稱	稱、稱、稱、秤
處	處、處、處、處、處、處、霙、処、寬	處、處、凱、處
辭	辤、辭、辭、辭、舜、辭、辭、辭	辭、辝、辝、辥
帶	帶、带、帶、帶、帶、帶	帶
殿	殿、殿、駸、殿、㲋、殿、殿、㲋、㲋、殿	墅、殿、殿、墅
發	桊、㪔、㪔、發、發、㪔、㪔、發、発、發、發、發、發、桊、發、發、發、㪔、桊	㪔、㪔、發、㪔、㪔、㪔、㪔、苃、㪔
歸	埽、歸、歸、歸、歸、歸、歸、歸、歸、帰、歸、歸、歸、歸、歸、埽、埽、歸、歸	埽、歸、歸、歸、歸、歸、隖、隖、歸、歸、歸、歸、埽、飯
龜	龜、龜、龜、龜、龜、龜、龜、龜、電、龜、龜、龜	龜、龜、龜、龜
號	號、號、號、號、號、號、號、号、號、號、號	號、號、號、號、嘸、號、麘、号、号、号、号

Fig.3 吕浩: 韩国汉文古文献异形字研究, 上海: 上海人民出版社, 2013 年 12 月, P27

This chart is just a tip of the iceberg. Many Han characters have a wide range of glyphs. It is very likely that one of the glyphs inherit one rare meaning of the character while other glyphs don't. Then even if the glyph has little difference from others, it is of great possibility that we disunify them. Actually, however, the users even don't care. Some characters are usually written as very similar glyphs in the practice and most people even don't think there is difference between them. In this case, the disunification is not very necessary.

In IRG PnP 2.1.3, it says

> Ideographs with different glyph shapes that are unrelated in historical derivation (non-cognate characters) are not unified no matter how similar their glyph shapes may be.
>
> **……**
>
> Because shape analysis alone may not tell non-cognateness or semantic differences, it is the submitter's responsibility to provide information and supporting evidence in order to invoke the non-cognate rule.

It doesn't say, however, which kind of information and evidence is needed in spercific. What's more, IRG PnP doesn't clarify how to deal with the mixing used simlar glyphs. Considering the large quantity of Han characters' glyphs and the conclusion of our discussion of GDM-00241(⿰木匋) in IRG WS2021 and UK-10757(⿸门丁) in IRG WS2017, I suggest

> 1) Add one sentence to the quoted paragraph 1 that "Ideographs with unifiable glyph shapes should be considered congnate if they can be used without distinction in fact".
>
> 2) Add one sentence to the quoted paragraph 2 that "For unifiable ideographs, the information and supporting evidence provied by submitters should be able to clearly explain a) the pronunciation of the two ideographs have no historical derivation; b) the meaning of the two ideographs have no relationship.

4.The possibility of unifying some rarely used ideographs regardless of their meanings.

After trying to concretize the process of defining non-congnate characters, I think there is a possibility to unify some rarely used **unifiable** ideographs regardless of their meanings. We can take it as two or more meanings of one character but two non-congnate characters if their meanings are unrelated. If the user does care for the shape, fixing the problem on the level of fonts or using IVS should be preferred then.

This is just a thought after brain storming, it will be very dangerous if we actually do it.

(End of Doc)

Feedback on the Suggestion of IRGN2482 to Avoid Encoding Shēngzàozì Characters or Made-up Characters
　　　　　　　　　　　　　　　　　　　　　　　　　　　2021-09-16

Thanks to Wᴀɴɢ 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 (古壮字), if the principle is adopted, will the submission of Sawndip characters be affected and become almost impossible to be encoded?

(End of Document)

Subject: Comments regarding Section 3 in IRG N2482: Proposal to improve IRG process
(3. Concretize the process of defining non-cognate characters)
Date: 2021.09.15.; 2021.09.16. (b)
Authors: KIM Kyongsok (KR)
Status: Individual contribution
Relevant docs: IRG N2482

1. Background

- Section 3 of IRG N2482 is shown below:

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

2020-09-02

Doc Type:	Ideographic Rapporteur Group Document
Title:	Proposal to improve IRG process
Source:	Wang Xieyang (王谢杨)
Status:	Individual Contribution
Action:	For consideration by IRG
Date:	2020-09-02

3. Concretize the process of defining non-congnate characters.

Here is a part of a chart in the book named 《韩国汉文古文献异形字研究》(吕浩 著, 上海: 上海人民出版社, 2013 年 12 月)

主形字	韓國漢文古文獻異形字	敦煌文獻異形字
備	備、俻、僃、俻、俻、俻、俻、偹、備	俻、備、備
邊	遻、込、込、邊、邊、邊	邉、邉、邉、邊
稱	稱、稱、穪、稱、桶、穪、稱、稱、稱、穪、穪、稱	稱、穪、稱、秤
處	處、處、處、處、歴、處、處、処、處	處、處、凱、處
辭	辭、辭、辭、辭、舜、辭、辭、辭	辭、辞、辝、辤
帶	帶、帶、帶、帶、帶、帶	帶
殿	殿、殿、殿、殿、殿、殿、殿、殿、殿、殿	壂、殿、殿、壂
發	癹、癹、癹、癹、癹、癹、癹、發、発、發、發、薆、發、發、發、發、癹	癹、發、發、癹、發、發、發、茇、發
歸	埽、歸、歸、歸、歸、歸、歸、歸、歸、埽、歸、歸、歸、歸、埽、埽、歸、歸	埽、歸、歸、歸、歸、歸、陽、陽、歸、歸、歸、歸、埽、皈
龜	龜、龜、電、龜、龜、龜、龜、龜、電、龜、龜、龜	龜、龜、龜、龜
號	號、鯱、鯱、號、號、鯱、鯱、号、鯱、鯱、號	騙、號、號、豿、嘷、唬、麐、號、号、号、号

Fig.3 吕浩: 韩国汉文古文献异形字研究, 上海: 上海人民出版社, 2013 年 12 月, P27

This chart is just a tip of the iceberg. Many Han characters have a wide range of glyphs. It is very likely that one of the glyphs inherit one rare meaning of the character while other glyphs don't. Then even if the glyph has little difference from others, it is of great possibility that we disunify them. Actually, however, the users even don't care. Some characters are usually written as very similar glyphs in the practice and most people even don't think there is difference between them. In this case, the disunification is not very necessary.

In IRG PnP 2.1.3, it says

> Ideographs with different glyph shapes that are unrelated in historical derivation (non-cognate characters) are not unified no matter how similar their glyph shapes may be.
> ......
>
> Because shape analysis alone may not tell non-cognateness or semantic differences, it is the submitter's responsibility to provide information and supporting evidence in order to invoke the non-cognate rule.

It doesn't say, however, which kind of information and evidence is needed in spercific. What's more, IRG PnP doesn't clarify how to deal with the mixing used simlar glyphs. Consideing the large quantity of Han characters' glyphs and the conclusion of our discussion of GDM-00241(囗木勹) in IRG WS2021 and UK-10757(囗冂冂) in IRG WS2017, I suggest

1) Add one sentence to the quoted paragraph 1 that "Ideographs with unifiable glyph shapes should be considered congnate if they can be used without distinction in fact".

2) Add one sentence to the quoted paragraph 2 that "For unifiable ideographs, the information and supporting evidence provied by submitters should be able to clearly explain a) the pronunciation of the two ideographs have no historical derivation; b) the meaning of the two ideographs have no relationship.

2. The author of IRG N2482 suggests as follows:

2) Add one sentence to the quoted paragraph 2 that "For unifiable ideographs, the information and supporting evidence provided by submitters should be able to clearly explain
a) the pronunciation of the two ideographs have no historical derivation;
b) the meaning of the two ideographs have no relationship.

3. Comments regarding the above suggestions:

3.1 pronunciations of characters

information and supporting evidence provided by submitters should be able to clearly explain
   a) the pronunciation of the two ideographs have no historical derivation;
==>

   1) In some cases, the pronunciation of characters are helpful to decide whether or not to unify, but not always.

   2) Sometimes one may not know the pronunciation of characters (for example, characters in old documents).

   3) Some characters have several pronunciations. Japanese Kanji characters usually have several (sometimes more than 10) pronunciations. Even when the pronunciations of two slightly different glyphs are different, actually the two can be one and the same character.

   4) When two glyphs are used in different countries/regions, it is likely that pronunciations are different.
   - Furthermore, a submitter may have to contact an expert in other countries to know the pronunciations there.

   5) I understand what the author of N2482 is trying to incorporate in the IRG PnP. However, I am not sure if this is a resonable approach.

3.2 meanings of characters

information and supporting evidence provided by submitters should be able to clearly explain
   b) the meaning of the two ideographs have no relationship.
==>

   1) In some cases, the meaning of characters are helpful to decide whether or not to unify, but not always.

   2) Sometimes, one may not know the meaning of characters (for example, proper names such as personal names or place names).

   3) Some characters have several meanings. Even when the meanings of two slightly different glyphs are different, actually the two can be one and the same character.

   4) When two glyphs are used in different countries/regions, sometimes meanings may be different.
   - Furthermore, a submitter may have to contact an expert in other countries to know the meanings there.

   5) I understand what the author of N2482 is trying to incorporate in the IRG PnP. However, I am not sure if this is a resonable approach.

3.3 A conclusion

  – Even when the pronunciations and meanings of two similar glyphs are different, the two glyphs may or may not be unifiable.

  – Based on the discussions in 3.1 and 3.2 above, it may be desirable not to add the following sentence which seems too restrictive.

For unifiable ideographs, the information and supporting evidence provided by submitters should be able to clearly explain
  a) the pronunciation of the two ideographs have no historical derivation;
  b) the meaning of the two ideographs have no relationship.

* * *