Title: **Feedback to IRGN2340 (and Response to Misc issues)**
Source: Henry Chan
Meeting: IRG #51, Hanoi, Vietnam
Status: Individual Submission
Pages: 5

1. The current unification model is not sufficient.

1.1. The current unification model cannot efficiently handle the large amounts of variants that can be encoded. There are 100,000+ glyphs in the MOE Dictionary, and there are 700,000+ glyphs under collection in the Chinese Characters Repertoire.

1.2. Under the existing unification methodology, only the common differences between the modern standard forms used in China, Japan and Korea locales are unified. The unification methodology does not take into account the historical differences across the many centuries of the evolution of the Han script.

1.3. IRG can only handle 5000 characters in every working set. It will take 236 years ($\frac{670,000-80,000}{5000 \div 2}$) to encode all the glyphs in the Chinese Character Repertoire, assuming that all the characters in the repertoire are not unifiable. Even if a small 10% of the characters are not unifiable according to the current rules, it will take us 24 years.

1.4. Some members think that if IRG did not bother itself so much with the unification, and just stick to the existing rule, then the process could be done more efficiently. Shifting the blame from the problem itself to the people who raise the problem is not a constructive way of solving the problem.

1.5. The unification criteria should be greatly expanded to take into account historical variations of the script.

1.6. For example, the normalization carried out by ROK is a practical way to reduce variants and minimize the number of new characters for encoding.

2. Little has been done by member bodies to improve this situation.

2.1. In one meeting, IRG once suggested TCA to prepare a document for normalizing the glyphs proposed by TCA, as to reduce the number of characters that is necessary to submit. However, TCA had rejected such request, because TCA's characters are encoded verbatim in their sources (MOE Dictionary) and source standards (CNS11643), and that TCA does not commit any normalization in their standardization process.

2.2. The unification criteria of ISO10646 need not be tied closely to the source standards. For example, the unification criteria in the JIS standards is different to that in ISO10646. The unification criteria in the original GB standards can also be described as much looser than ISO10646 as many historical variants are not included, and the encoded glyphs are highly normalized compared with the source glyphs in Kangxi Dictionary or older books such as Guangyun, Jiyun, Leipian, etc.

2.3. Commission of normalization before submission to IRG should be an IRG requirement to reduce the number of characters for encoding.

2.4. TCA requested a clear guideline of what can be unified or what cannot be unified. However, besides ROK, no member body has been able to provide any concrete list or idea of what should be considered unifiable.

3. Independent reviewers have little resources and are working for volunteer.

3.1. I do not agree with the description of review opinion to mass encode characters via IVS as a “personal subjective recognition”.

3.2. As far as I can tell, most of the time in an IRG meeting is spent on the arguing between “It looks too different” and “It looks similar enough”. I believe this way of approaching the unification is “personal subjective recognition”.

3.3. On the other hand, my main argument during the IRG meeting has always been “these kinds of variations are systematic and should be unified consistently”. I have consistently asserted that semantic has priority over difference in glyph shape for deciding unification or disunification.

3.4. In IRGN2309 (IRG Working Set 2017 version 2) Henry Chan Review, I provided example for systematicity and rationale from page 3 – 22. Also refer to pages 32 – 33, 60 – 63, 72, etc., which took information directly from the MOE Dictionary to highlight the systematic and semantic origin as rationale for unification.

3.5. In IRGN2223 (IRG Working Set 2015 Version 4) Henry Chan Review Part 4, I provided detailed rationale for suggested unifications from page 2 – 47, but most were rejected by member bodies as “the variants look too different”.

3.6. Furthermore, it cannot be expected that volunteers provide an “exhaustive” list of possible new unifications. The member body which submit the most glyph variants per character did not even provide *any* suggestions for extending the unification model.

3.7. The review of the huge amounts of trivial variants places an extremely big burden on individual reviewers, and the feedback from member bodies is often hostile.

3.8. Individual reviewers are volunteering based on their free time, but to prepare evidence for the addition of new rules, it takes many hours to collect examples of such variations from multiple sources to showcase the systematic occurrence of such variations. Unfortunately, research is often dismissed by standardization experts from other member bodies on the basis of “it looks too different” or “these characters (sic) need to be used by scholars”. Due to time limit, I have also been unable to supplant all suggestions to use IVD with detailed rationale.

3.9. I believe it is completely misplaced to claim that my review comments to use IVD are “personal subjective recognition”. I think the member bodies which dismiss my evidence as “it looks too different” are the ones being subjective.

3.10. The ad-hoc addition of rules is not scalable. The time and effort put in by reviewers has not translated to meaningful changes inside the encoding process.

3.11. Member bodies themselves hold vast amounts of material which use the variants and have compiled detailed proper form—variant form mappings. Member bodies are in the best position to criticize the existing unification model and propose extensions to the unification model. Pointing a finger at volunteer reviewers, while not proposing any concrete list, nor proposing any methodology to screen away variants, does not solve any problems.

3.12. If member bodies who have the resources do not and cannot draw on the information they have collected for over 30 years to create a clear, exhaustive, objective list to extend the unification model, they cannot reasonably expect volunteers to do an “exhaustive” job for them between an IRG meeting or two.

3.13. Also, in response to unwritten complaint from miscellaneous sources, unrelated to IRGN2340:-

3.13.1. The opinion to not code the variants as new characters is not intended obstruct the encoding of variants needed for scholarly use and cultural preservation. Conspiracy theory is not constructive.

3.13.2. Blaming incomplete implementations to not adopt a technical treatment is not a strong argument.

4. The relationship between unification and IVS should indeed be clarified.

4.1. IRG is currently too conservative in its unification. Since member bodies did not seem to want to modify the unification model. It was proposed to extend IVS to encode variants that are typically not traditionally deemed suitable to unify. Amendments to the IVD text have already taken effect.

4.2. However, TCA expressed concern about designing their own criteria for deciding which variant glyphs could be registered in an IVD, and which variants to be coded as new characters. Also, for glyphs to be coded as variation sequences, the glyphs should technically be unifiable.

4.3. Therefore, for the adoption of IVD, inevitably, reform of the unification model especially with regards to epigraphical variants, calligraphic variants and transliteration variants, is the only remaining option.

5. The concern for over-unification is overrated.

5.1. Submitters have been wary of over-unification as disunification of over-unified characters has a lot of bad consequences for compatibility with existing systems.

5.2. Most over-unified characters are the application of the UCV without consideration of semantics. This situation is less relevant today as submitters need to submit evidence, most of which is dictionary proof.

5.3. Submitters are responsible for ensuring that characters they submit are not unified with non-cognate characters. Submitting glyphs to IVD is also to be compiled by the submitter and is subject to a public review period of not less than 90 days. If the glyph is to be coded via IVD, it will not pass through IRG. Extending the unification model to allow unification of more variants would not lead to higher over-unification error rate for IRG.

5.4. Over-unification of different characters to the same codepoint cause the semantics of characters in existing documents to be ambiguous. However, every variant glyph encoded via IVS has a unique representation. When a variant glyph is encoded as an IVS, and later found out to be separate character in its own regard, the same character can be encoded at a new codepoint with no effect to any existing or future documents.

5.5. In the Adobe Japan 1 IVD collection, many separately encoded characters are also mapped as glyph variants to the more commonly used character.

5.6. Existing characters that are actually minor glyphic variants of more common characters would benefit if they were retrospectively re-encoded as Ideographic Variation Sequences to the common character. Existing datasets could be ported to the new sequences in one-step to take advantage of the Unicode Collation Algorithm (supported natively by languages such as Java) to handle variant-insensitive searching.

6. IVS is mature for adoption.

6.1. Variation Sequences were standardized after the coding of CJK Unified Ideographs Extension B. The technical requirement to distinguish different variants could not be met with loose unification at that time.

6.2. Ideographic Variation Sequences are well supported. Fonts supporting the Adobe Japan 1 IVD are distributed on all major mobile platforms (Android and iOS). The technology is proven.

6.3. The circumstances requiring coding of closely shaped variants separately is no longer valid today. Given the ubiquity of IVS support, and the pressing need to encode the enormous amount of historical glyph variants, the unification model should be updated.

7. If the member bodies do not wish to draft their own list, and do not want to do ad-hoc unifications or create ad-hoc rules, the encoding of all variant characters in Working Set 2017 should be suspended. An ad-hoc group should be formed to discuss extensions to the unification model. **No variants within WS2017 or onwards should be further encoded until the ad-hoc group can reach consensus.**

With regards to new ad-hoc group for reform of the unification model, I believe the principles should be as follows:-

1. **Any expansion of the unification rule shall not violate the non-cognate rule.**

2. **Any expansion of unification rule should not unify any characters with similar shapes but distinguishable semantics with respect to the shape difference**.

   2.1. When any shape difference is reliable distinction of semantics in a certain locale or time frame (typically treated in IRG as part of the non-cognate rule in practice), it shall be disunified.

   2.2. Example: 沉/沈 and 着/著 are not unified even though they are related in historical derivation (cognate), because they are no longer semantically equivalent and cannot be exchanged in general text without loss of semantics.

   2.3. Example: 芸 as a plant, 芸 as simplified for 藝, and 芸 as simplified for 蕓 are unified even though they are not related in historical derivation, because they are structurally equivalent; the semantic difference is not distinguishable by the possible glyph difference of four-stroke grass radical vs three-stroke grass radical.

3. **Unification should be carried out with a holistic view and consider the full spectrum of variants.**

   3.1. The current unification model cannot handle the new variants because it is carried out on a character-comparison basis.

   3.2. Inspecting each character with respect to the “Zhengti” (or any existing coded ideographs) only without considering other variants would lead to unnecessary encoding of uncommon but similar glyphs.

   3.3. In IRGN2211 submitted for IRG #48 I recommended the use of IVD based on character origin, and the unification to the orthodox character. TCA expressed problems with the approach. However, I believe that discussing character origin in an unavoidable part of IRG work. Non-cognate rule is a fundamental part of the current unification model, and non-cognate rule already involves character origin and semantic analysis.

4. **Expansion of unification rule should allow for orthographic changes throughout the maturing of the script.**

   4.1. The regular script is often described as a compromise between Clerical Script and Cursive Script. The script evolved naturally over time and the early forms look markedly different from the modern forms. Sometimes, the ancient forms are identical to forms that have completely different semantics in modern day.

   4.2. For example, 麦 is the form for 夌 instead of 麥 for early texts. If a character has a phonetic component “夌” but has the actual shape of “麦”, the character should be unified towards the corresponding character with shape “夌”. This shall not affect the encoding of character with same shape containing “true 麦” component as a new separate character.

   4.3. For example, the glyphs 夫/天/矢/失 are not well distinguished in Clerical Script and sometimes not in Regular Script. Such components, although recognized to be semantically distinct today, when used as a part of a character, that character may be suitable for unification.

5. **Consideration of encoding via IVD or encoding as new character for variants shall be considered based on submitted evidence to IRG only.**

   5.1. If there is evidence for equivalence, as well as evidence for non-equivalence, the glyph should be processed under the evidence submitted by the member body as first priority.

   5.2. The same source reference may be encoded under different base characters if the evidence suggests so.

   5.3. The encoding of a glyph as an Ideographic Variation Sequence does not forbid the re-encoding of the glyph as a new character, and vice versa.