

Title: Response to Feedback to IRGN2612  
Source: Henry Chan  
Date: 2023-10-17  
Status: Individual Contribution to IRG #61  
Action: For consideration by IRG  
Pages: 3

### Response to John Knightley's comments:

John suggests that 責 should be removed from UCV #190d, claiming that there are many existing disunification examples in Extension B. However, it has been agreed in IRG meetings that Extension B disunification examples are not considered as existing unification examples.

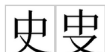
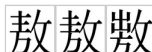
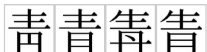
The quoting of the sources from MOE Dictionary is merely to prove that they are variants, and to prove the variation is systematic. It is not used as a direct basis for unification.

In recent IRG meetings, variants arising from using a different component form for the same radical have been changed from NUCV to UCV, since these differences are more appropriately unified and coded using IVS. For example NUCV #404, #405 has been changed to UCV, and new UCV #417, #430 has been added:

404 · unifiable 艸++ 	405 · unifiable 攴支 	417 · unifiable 月肉 	430 · unifiable 彳水 
---	--	--	--

In essence these are calligraphic simplifications which are greatly generalizable and happened as part of the transformation between the Seal script, Clerical script, and Regular script.

These unifications are not limited to different forms of radicals. Besides 責, we also have other UCV examples where the “complex” or “full” form which corresponds more closely to the form in Seal script is now unified with the “common” form which is predominantly used in Regular script:

469 · unifiable 史叟 	202 · unifiable 敖 	319 · unifiable 青青青青 
---	--	---

My suggestion is that we codify this general class of unification so we can reduce the number of variants we need to code.

This is indeed a departure from the encoding model in Extension B, but also consider that the initial encoding model in URO does not include any of these “complex” or “full” forms, and none of the original source standards in the URO contain these forms.

Note, the “complex” form which is found in dictionaries is often not even consistent. Consider the following encoded forms:

勝 - 勝 勝 - □ 滕 - 滕

vs

朕 - 𨾏

vs

膳 - 𨾏 膳 - 𨾏

In the first set, the 月 has been expanded to the original form 舟, while in the second set, the 关 has also been expanded to 𨾏, while in the third set, even the 𨾏 has been expanded to the full form 𨾏火収.

By encoding these variant forms as separate characters, it means users either need to already know the exact form used by a particular text, or the digitization system needs to maintain huge mapping tables for the variant characters back to the common characters. Both of which are unwieldy for use.

The fact is the general public do not even recognize these characters, so most digitization systems choose to convert the characters to the common ones, which results in an irreversible loss of information.

This is no longer an issue if the characters are unified and the variant forms are encoded as variants in an IVD collection. Any Unicode compliant text processing software should automatically handle variants encoded this way correctly, without the need for additional mapping tables and custom logic.

## Response to Tao Yang and Chen Zhixiang's comments:

The goal of updating the IRG PnP to address unification regarding differences in transliteration is mainly to target the issue around one or more components taking a different form due to calligraphic simplification or component merging.

Similar examples are like 更 and 𠄎, 曹 and 𦉳, 晉 and 𦉳 etc. Note these “complex” or “full” forms have already been encoded as separate characters, so any new UCV only applies to characters which include them as components.

The goal of suggesting the update was not to block the encoding or suggest unifying characters like 𠄎止它 and 𠄎止虫, or 𠄎止矢 to be unified to 族.

Obviously the two types of transliteration are not the same.

The first one is the transliteration of one or more components into a single joined component, which arises due to the stroke simplification that happened between the development of the Clerical script and Regular script.

The second one is the transliteration of ancient characters where the character composition is not the same.

---

The feedback from Tao Yang and Chen Zhixiang also mentions “right or wrong should be judged by philologists”.

However, IRG should not be involved in the academic viewpoint of “right or wrong”. As long as a character has a required use, whether it is right or wrong, as long as it is not a printing error in limited distribution, it should be coded – either as a new character, or unified with an existing character.

If a character is unified it does not mean it cannot be coded. It only means it will not be coded as a new character under IRG, and it automatically means it can be coded as a variant in an IVD collection.

If philologists deem a particular “complex” or “full” form to be particularly noteworthy and correct according to some orthographic standard or academic theory, it does not mean that it needs to be coded as a separate character. Unification only needs to consider that two forms are interchangeable semantically, and submitters are only required to provide one representative glyph.

The matter of which form is “correct” or which form is “more correct” should be defined by the user community of these variants. If a given user community considers some forms are “more correct” than others, it can be represented by placing them into different IVD collections. This should be kept outside the scope of IRG.