L2/15-306

Feedback to L2/15-266

John H. Jenkins

1 November 2015

The following email was received on the Unihan mailing list from Henry Chan regarding L2/15-266:

To sum up, here's some suggestions I have to the current list of kZVariants.

A. Exclusions

A1. Exclude Characters with Different "abstract shape" i.e. sound & meaning (=non-cognates in IRG terms)

-- 幺 (U+5E7A) / 么 (U+4E48) (two completely different characters in PRC) = kSpecializedSemanticVariant?

-- 本 (U+672C) / 夲 (U+5932) (definitely non-cognate, the latter sometimes as a corrupted form of the former) = kSpecializedSemanticVariant?

-- 刊 (U+520A) / 刋 (U+520B) (definitely non-cognate, the latter sometimes as a corrupted form of the former) = kSpecializedSemanticVariant?

A2. Exclude Official PRC Simplifications

-- 狀 U+72C0 / 状 U+72B6

-- 妝 U+599D / 妆 U+5986

-- 壯 U+58EF / 壮 U+58EE

-- 莊 U+838A / 庄 U+8358

-- 將 U+5C07 / 将 U+5C06

Note 1: despite what Annex S may suggest, no unifying precedence of official traditional/simplified character exists

Note 2: 奨 5968 / 奬 596C / 獎 734E need not be excluded because the official simplification is 奖 U+5956

B. Inclusions

B1. Include Exact Duplicates in Extension A & B

-- e.g. 炮 (U+3DB7) / 炮 (U+2420E)

-- e.g. 堅 (U+249BC) / 堅 (U+249E9)

-- e.g. 鞲 (U+2a415) / 鞲 (U+24bd2)

-- Other IRG duplicates identified at: http://appsrv.cse.cuhk.edu.hk/~irg/irg/irg24/IRGN1132_DefectReportForSuperCJK.pdf

B2. Include unifyable characters under the UCV, and disunified by error:
-- e.g. 敞 (U+2304B) / 敞 (U+22F38): one has a Kangxi source and the other a Hanyu Dazidian source; the sources refer to the same 切音 and Shuowenjiezi entry.

B3. Include other potentially unifiable cases not addressed in UCV:
-- e.g. 复 (U+590D) / 夏 (U+3686) & 縠 (U+7DEE) / 縠 (U+2608A)
-- Many are the classical case of newer characters stealing kangxi codes from older characters, because one of the four dictionaries suck at normalizing variants.

If these principles are fine, I can start clawing out characters and putting them into these categories :)

-- Henry

I responded to Henry as follows:

I'm fine with B1 and B2, but not with A1 and A2, and I'm iffy with B3.

Ultimately, this is something the UTC needs to weigh in on. Unless you'd rather write up a formal document, I'll forward your email to the UTC, and we can discuss it at the next meeting with my document.

Lee Collins further responded:

I don’t have a problem with B3, but I agree with John on A1 and A2. Even if characters are non-cognate, some of the examples given are often interchanged in practice. When searching, I want to be able to normalize to one and find the mapping useful.

Lee