ISO/IEC JTC1/SC2/WG2/IRG **IRGN 1406**

Title: ***Han Unification Issues***

Authors: Richard Cook, Thomas Bishop, Ken Lunde

Date: 6 June 2008

"文 U+6587 / 攵 U+6535"

Note the form of the 文 U+6587 component in some characters under that radical in *Kang Xi* :

http://linguistics.berkeley.edu/~rscook/images/KangXi_1997/0477.gif

The 文 U+6587 component has a combining form identical to 攵 U+6535.

This fits in with discussion at IRG #29 of the fact that unification rules are different for components vs. independent characters. That is, depending on the size/position/context of the form when it is used as a component, some unifications are evident that might not otherwise be applied.

In this case, when 文 appears as rightside component, it sometimes has the form indistinguishable from 攵. Now, 文 and 攵 components are typically not unified. But here, since *Kang Xi* explicitly identifies 文 as the radical, there is no question that 攵 is 文. Note also that *Kang X*i mentions 文/攵 conflation under the entry for 攵 under 攴 radical.

http://linguistics.berkeley.edu/~rscook/images/KangXi_1997/0468.gif

A related odd/interesting thing is the position of 𣁋 [U+2304B] in the *Extension B* code charts. It derives from the *Kang Xi* radical assignment, which as far as we can tell is an error on the part of the *Kang Xi* editors. It seems they had finished the 攵 radical, and found 𣁋 when they were working on 文, and so just put it there under 文, when in fact it really should be classified under 攵 instead.

"叠 U+53e0 / 疊 U+758a"

Consider 㲲 U+3cb2, which seems to have simplified 叠 component in some 5/12 of available CJK fonts, but traditional 疊 component in the others. This is evident in these screenshots of Apple's OSX *Character Palette* :

http://linguistics.berkeley.edu/~rscook/images/3CB2.tiff

http://linguistics.berkeley.edu/~rscook/images/4D11.tiff

The simplified form is a "traditional simplification" (like 无/無), but the 叠/疊 forms are not explicitly unified (in an official multi-column codechart, or in SuperCJK 14), though PRC orthographic standards assert the variant relation and the PRC preference for the 叠 form.

The 叠/疊 pair seems to be unified, at least according to GB standards. For example, 䴑 U+4D11 is rendered differently in GB 18030-2000 and GB 18030-2005.

Below is the full encoded (Unicode 5.1) set with 叠/疊 comp:

BMP: 叠/疊: 㜼㩹氎鸈
SIP: 疊: 𠠝𣡭𣰼𡬶𡬷𨇼𨏦
SIP: 叠: 𢶣𣜖𣿬

Note that the BMP forms exhibit 叠/疊 variation in available fonts (though the above glyphs all happen to have 叠 in this typeface), but the (rare) SIP forms do not (reflecting instead the precise form of the component seen in ISO/IEC 10646 Ext. B representative glyphs).

It is interesting that although 疊/叠 is an “official” traditional/simplified pair (meaning, according to the *Wenlin* and *UniHan* definition, that one is Big5 and the other GB2312), there are no such “official” pairs with these as *components*.

The 叠/疊 pair is another case in which simplification of the stand-alone character does not imply simplification of the same character as component in another character. Other examples are as follows (after 《简化字总表》, Table 1; see the *Notes* below**):

么(麼)	*but*	嬷(嬤)
只(隻)	*but*	愯
报(報)	*but*	蕔
办(辦)	*but*	𥷕
术(術)	*but*	𦺔, 𣾷
为(為)	*but*	寪?

(And contrast with these the case of 言 U+8a00, which is simplified to “讠” only as a *component*.)

All of this suggests that 疊/叠 were indeed originally unified in Unicode/10646 (though not explicitly, but only in implementations or in legacy standards), and later disunified in Ext. B (intentionally or not), with encoding of the 3 SIP code points in the following list (note that the members of the first pair appear here to be duplicates because of the specific fonts used, though the official BMP representative glyph uses 疊):

㩹 U+3A79 / 𢶣 U+22DA3
𣡭 U+2386D / 𣜖 U+23716

These are the only encoded pairs we know of, besides 疊/叠. But the situation becomes especially complicated when one considers that the following are all encoded variants of 疊/叠:

㬪曡畳疂疉𡀉𣆹𣈍𤴁

[U+3b2a][U+66e1][U+7573][U+7582][U+7589][U+21009][U+231b9][U+2320d][U+24d01]

END NOTES:

**The book 《语文文字规范手册》 (3rd edition, 1997 语文出版社, ISBN 7-80126-131-3/H.34) contains the updated (as of 1997) versions of the following PRC standards (among others): 《简化字总表》 (1986) and 《第一批异体字整理表》 (1955). 叠/疊 is in the latter as standard/variant (not simplified/full form).

《简化字总表》 has 3 tables, plus indexes. The 1st table has 350 不作简化偏旁用的简化字 which as a rule are simplified only as whole characters, not as components (碍, 肮, 袄, ...). The 2nd table has 132 可作简化偏旁用的简化字 characters to be simplified both as whole characters and wherever they occur as components (爱, 罢, 备, ...); and 14 可作简化偏旁用的简化偏旁 components to be simplified wherever they occur as components. (Six of those components — five simplified and one full form — are unencoded, but Wenlin has assigned them PUA codepoints; they should be encoded if only so we can communicate about these standards; but also for use in CDL and IDS.) The 3rd table gives a non-exhaustive list of 1753 examples of compound characters that contain the components listed in the 2nd table.

Relevant PRC orthographic standards such as those mentioned above should be translated into English and taken into account by IRG for development of Han unification principles, and for development of a standard variant mechanism to automatically convert rare full forms into rare simplified forms.