Feedback to ISO/IEC JTC1/SC2/WG2 IRG N2223

Date: 2017/10/17

Source: suzuki toshiya

Title: Comments on UCV#363 Handling

Action: Consideration in IRG#49 meeting

1. UTC-02625 and U+588F, and UCV#363

In IRG#49, Japanese comment on UTC-02625 is being discussed. Although the upper right componet of UTC-02625 is slightly different from that in U+588F, this pair is clearly covered by the source code separations listed in ISO/IEC 10646 Annex S.

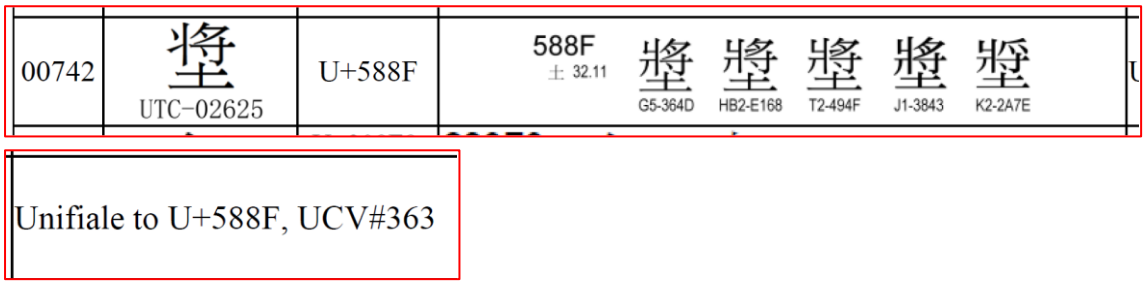

Figure 1: Japanese comment on UTC-02625

Figure 2: UCV #363

Figure 3: Source Separation Example in Annex S and the situation in the code chart.

This shape difference was discussed in Ext F development. In the development of Ext F, UK had once suggested to disunify this difference, possibly because of the overlooking of the source separations.

2E484 and 2E5AB

There is no precedence for unifying 壯 and 壮, which in China are treated as traditional and simplified forms rather than glyph variants. Therefore 2E484 should not be unified with 2E5AB.

Figure 4:IRGN2146 UK response

As recorded in WG2 N4727, IRG decided to unify them.

2E484 and 2E5AB

There is no precedence for unifying 壯 and 壮, which in China are treated as traditional and

According to SCS, they should be unified

Figure 5:IRG N2146 suzuki comment 2 (the red note is a record of IRG dicision, not my personal opinion)

Code point	Unified with	Deleted source reference (Font code)
2E40B	8521	JMJ-060130 (F1-991C)
2E484	2E5AB	USAT-03743 (F2-6355)
2E5E8	2789B	JMJ-058726 (F2-6788)

Figure 6: WG2 N4727 record

Figure 7:Unification of KC-06763 and USAT-03743 in UCS 5ed (DIS text)

For UTC-02625, UK again suggests to disunify them by quoting 3 disunifications in Ext E. However, Ext E disunifications might be regarded as the mistakenly disunifications, because the reconsideration of the source separation examples should be discussed in WG2, but there was no such discussion in the recent WG2.

2. Expected Impacts by the Obsoleting of UCV#363

In IRG#49, some experts discussed the possibility to obsolete UCV#363. However, considering the number of the "unpaired" characters, obsoleting of UCV#363 may induce so many derived simplified characters. Estimating the number of the characters by IDS, the results are:

- the characters including 爿: 352 (URO 61, ExtA: 21, ExtB: 198, ExtC: 9, ExtE: 36, ExtF: 27)
- the characters including 丬: 47

The existing separations listed in UCV#363 is just 12, therefore, the number of affected characters could exceed 200, if UCV#363 were obsoleted. Such impact conflicting with Annex S should not be justified only by the existence of several disunification examples. I suggest to use IVS to minimize the impact.

3. Expected Impacts by New NUC separating 将 and 將 (added for revision 3 of this document)

In IRG#49, some experts suggested to keep UCV#363 but add new NUC separating 将 and 將. This pair is clearly conflict with 2 source separation examples.

However, the number of the characters affected by this new separation rule might be useful for further discussion.

- The number of the characters including 將: 47
- The number of the characters including 将: 16

As a result, there might be about 30 separate encodings of simplified characters. It is smaller impact in comparison with that by obsoleting UCV#363, but it could be still uneasy.

If the view of the subtle difference in shape is important, we already have the technology standard of Ideographic Variation Sequence (IVS), which is suitable to clarify the appearance of abstract character shape. We should move on to unify, not to dis-unify, the ideographs, for the sake of ease of use. Once Thomas Carlyle said, “the Fraction of Life can be increased in value not so much by increasing your Numerator as by lessening your Denominator”, so is the number of encoded characters. We must consider what is the best solution for the common users, not the pedants. Unlike 30 years ago when the Unicode was first appeared, we already have an enough storage and resolutions for storing images of documents rather than encoding them. The needs to subdividing the encoded character for subtle difference is lessening day by day, and it just increases the confusion on searching and collating the documents.

(end of document)