L2/09-036

In Re Public Review Issue #128
John H. Jenkins
25 January 2009

I'm afraid I disagree with the proposed update to UTS #37. I have two main objections.

1) Strengthening the criteria for propriety of using a variation sequence for a particular form (and the use of the word "must" in the proposed wording) implies that somebody will be doing either enforcement or, on behalf of Unicode and WG2, vetting proposals to see that the criteria are met. It's not at all clear to me who would be responsible for that. Presumably, it would be the IRG, but the IRG already has a very full plate and is seriously backlogged. I don't know that the IRG could process variation sequence proposals at any reasonable rate, and I don't think it's prudent to overload the IRG with yet more responsibility.

2) More seriously, I don't understand why we are continuing to insist that y-variants be separately encoded. By definition, y-variants (two forms which mean the same thing but have different abstract shapes as defined by Annex S) would not typically occur in the same text. What, therefore, is the value of continuing to separate them?

In particular, I don't think that there is value in continuing to separately encode simplified Chinese characters from their traditional counterparts. There are, in fact, several thousand traditional forms with no currently encoded simplified forms which could in theory be produced by applying existing rules. When such are attested, do we really want to go through the process to encode them?

Nor are simplified characters the only source of unencoded y-variants. Another example would be UTC00344 ( ). This form comes from Lau's *Practical Cantonese-English Dictionary*, which defines it as a variant form of U+69AB (榫). The two characters have different overall geometries and are therefore non-unifiable under Annex S rules. This is the kind of character which I would very much like to be able to represent using a variation selector, as there seems to be no practical benefit to keeping them separate.

At the moment, the variant situation with regard to Han is frankly a shambles. We all know that there are numerous variant characters which have been separately encoded, but there is no authoritative variant data available and so processes such as fuzzy matching for Han are largely impossible. Continuing to add forms which are unifiable under a strict interpretation of the character-glyph model does not make this problem any easier or hasten the time when it's dealt with properly.

I think that the main objections to allowing y- and z-variants to be represented via variation sequences are that it makes the definition of allowable variant too vague, and that it might end as a pseudo-encoding.

Part of the problem is that, while we have a formal definition of z-variant courtesy of Annex S, we don't have one for y-variant. We may want to address this. Alternatively, we may simply want to leave the definition of "acceptable y-variant" in the hands of those developing the VS proposals.

As for the pseudo-encoding issue, there are a number of sub-concerns. One is that a future candidate for encoding as a separate ideograph might be rejected on the basis of its already being representable via a variation sequence. Similarly, we would have the potential for two visually identical forms being representable via different character sequences. This second is a problem which already exists, not only in non-Han text, but even in Han, given the z-variants, compatibility ideographs, and accidental duplications we already have.

I should point out that in any event, the former potential problem already exists. Given the IRG's use of the non-cognate rule, it is entirely possible that a variant form currently defined as a z-variant (and therefore unifiable) would be discovered to be non-cognate with the base form. In such a case, it would be a potential candidate for encoding, even if a variation sequence representing it is defined. (Granted, this scenario is unlikely.)

In any event, I really think that we need to seriously do what we can to reduce the overhead for extending the set of distinguishable forms in Han. The IRG is the bottleneck, and working to reduce its workload I think in the long run is in everybody's best interest.