L2/05-223

Comments on L2/05-172 (On Criteria for disunifying Diacritics)

From : P. Andries and F. Yergeau

The recently published L2/05-172 makes some arguments and proposes some criteria for disunifying combining diacritics. We believe some of the arguments deserve examination and criticism, and that the proposed criteria need substantial improvement.

The Unicode Standard states about the Combining Diacritical Marks block: “The combining diacritical marks in this block are intended for general use with any script.” It may be true — or not — that this text is sometimes misunderstood as if it was intended to be a normative directive that these diacritical marks should be used with all scripts. Is it written black on white, however, that the U+03xx block *may* be used with, and is even intended for, any script, and is not limited to “primarily for use with the European scripts derived from Greek” as L2/05-172 argues. We don't quite see how it is possible to interpret otherwise the phrase “intended for general use with *any* script” [our emphasis].

L2/05-072 argues that applications would not reasonably support use of common diacritics in typographic traditions other than Greek-based, such as CJK or Sumero-Akkadian syllables. This is in direct contradiction with Tifinagh, Hebrew and Syriac usage and the recent implementation of N'ko by P. Andries [see L2/05-169], which demonstrated exactly zero additional difficulty and zero additional limitations from using the common diacritics instead of the proposed N'ko-specific diacritics. It is just as easy to write 0x0308 in an OpenType table as it is to write 0x07F3. This argument of support by applications being more difficult or less likely to be obtained with common diacritics appears completely bogus and should not influence the criteria for disunifying diacritics.

L2/05-072 correctly points out that Unicode already has numerous script-specific diacritics and lists a number of them. It is fairly short, however, on the reasons why this was done. In some cases this is known, in others not, but we can easily start a list of possible motivations:

- Some diacritics need properties different from the common ones in order to behave correctly in the given script.
- Some diacritics have no match among the common ones.
- Some diacritics may have been encoded for no good reason at all, other perhaps than compatibility, a common diacritic could have been recommended instead.

We don't see any of these as providing a motivation or creating a precedent for encoding more script-specific diacritics whenever a common one is adequate. Of course diacritics that need different properties or that have no adequate match among the common diacritics should be encoded. But, perhaps, consideration should then be given to encoding them among the common diacritics, if there is any likelihood that they might of use in other scripts. The goal of Unicode is not to encode as many distinct characters as

possible, but to support writing systems. We believe we must strive for simplicity and generality.

In light of this, let us now look at the proposed disunification criteria:

a. *the mark has been borrowed from another script, but has been significantly modified to fit with the ductus of the borrowing script*

The first half does not belong here. There is nothing in L2/05-172 or elsewhere that would support basing encoding decisions on a character being borrowed or not. Unicode encodes characters, not their history. As for the second half, it is clear that if a diacritic has no match within the common ones (either as a result of adjustment after being borrowed or otherwise), it should be encoded separately. This criterion needs to be modified and clarified.

b. *the mark forms part of a set of marks in the script (for example a set of tone marks), but only some members of the set can be unified with existing marks*

This is not supported by any argumentation in L2/05-172 or elsewhere. It seems to be designed solely to justify *a posteriori* the misguided encoding of some N'ko-specific diacritics. This criterion needs to be removed entirely.

c. *the mark has a specific function unrelated to the generic diacritical mark (e.g. use of the mark as a vowel sign as opposed to the use of a similar-shaped mark as a diacritic). In such case the two uses might also require explicit differences in their character properties.*

Function should have no bearing on disunification. Dot-above is already used in very different functions in natural languages, old IPA and in mathematics. The function of diæresis in German is very different from that in French or in mathematics. The criterion needs modification, the part about different Unicode character properties only should be retained.

d. *the range of glyphic appearance is markedly different from the generic diacritical mark*

As written, this criterion appears too vague to be of much use. The common diacritics already have very ranging appearances in various typographic traditions, even within the Greek-based scripts which L2/05-172 singles out. We do not know how to improve the criterion to make it useful, other than stating the obvious: if no existing diacritic matches the needed one, then a new one needs to be encoded.

e. *the layout behavior is different and requires different support*

Layout behaviour is dependent on the properties, so this is redundant with *c*. Diacritic placement is generally supported in one of two ways in modern rendering engines:

substitution to a precomposed form or relative placement to an anchor point defined on the corresponding base letter (or diacritical mark when stacking diacritics). Either method can be used to accommodate high-end script-specific diacritic placement (for example setting an identical diacritic mark slightly higher in one script than in another, this behaviour is simply determined by the base letter, see L2/05-169 for placement of N'ko marks). It is thus unclear for us what is meant by "requiring different support" (from what?), this part needs to be fleshed out to become a useful criterion.

L2/05-172 also discusses a concern that has been raised (in L2/05-169, which L2/05-172 fails to reference) about script-specific diacritics increasing the potential for spoofing. It correctly points out that Unicode already offers a very large spoofing potential. We do not think this provides any motivation for worsening the case.

L2/05-172 also correctly states that a blanket prohibition (which nobody is asking for, as far as we know) would not be very useful. We concur: there are cases where encoding confusable diacritics is necessary, such as when the properties must be different. The security downside must then be accepted and properly dealt with.

We differ strongly, however, when L2/05-172 argues that script-specific diacritics can be helpful for security. While it is true that a security-conscious implementation can deal quite effectively with the risk, we are convinced that not increasing the risk in the first place can only be better than increasing it and then relying on mostly as yet non-existent or undeployed, and in any case fallible mechanisms to counter it.

In conclusion, we believe generic combining marks are a powerful feature of Unicode which sets it apart: all scripts may benefit from these marks right now, even for cases undreamt of by ISO standardizers, without waiting years until ISO approves the addition of identically looking and behaving but script-specific signs. Any disunification has to be grounded, as much as possible, on compelling technical grounds (e.g. differing stacking order, different placement of potentially concurrent similar signs on the same base letters) or obvious visual differences and not on subjective appreciations (e.g. the intricate genealogy of a diacritic, a preference to group all functionally similar or historically related signs under one block). Introducing additional subjective criteria could only lead to sterile and lengthy discussions with no practical benefits.