L2/05-247

ISO/IEC JTC1/SC2/WG2 N2981
2005-09-05

Organisation internationale de normalisation
Международная организация по стандартизации
International Organization for Standardization

Jeu universel de caractères codés sur plusieurs octets
Universal Multiple-Octet Coded Character Set

Doc Type: Working Group Document
Title: Comments on N2976 (On Criteria for disunifying Diacritics)
Source: P. Andries and F. Yergeau
Status: Individual Contribution
Date: 2005-09-05

The recently published N2976 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 to make them acceptable.

Purpose of common diacritics

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 as N2976 purports — 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. It is 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 N2976 argues. We don't quite see how it is possible to interpret otherwise the phrase “intended for general use with *any* script” [our emphasis].

N2976 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, which do use some common combining diacritics, and the recent implementation of N'ko by P. Andries [see N2949], 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.

Long on analogy, short on justifications and explanations

N2976 correctly points out that the UCS 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 (such as the Arabic ones listed in N2976) need properties different from the common ones in order to behave correctly in the given script (e.g. be presented in proper stacking order after normalisation).
- Some diacritics have no match among the common ones.
- Some diacritics may have been encoded for no good reason at all, other perhaps sometimes 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 separately. 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 the UCS is not to encode as many distinct characters as possible, but to support writing systems. Unexplained analogy to past encoded diacritics seems to be the prime justification for the criteria suggested. We believe we must rather strive for simplicity and generality justified by compelling technical reasons and not a proposer's preference to group diacritics in a neat contiguous block when this could create duplicates and increase the security risks.

Review of the proposed disunification criteria

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

> *The main criterion is that of the shape of the glyph, since that is the chief identifier of a diacritical mark. When the range of glyphic appearance of a diacritical mark may be markedly different from the range typical of the generic diacritical mark, disunification may be preferred.*

As written, this criterion appears too vague to be of much use. The common diacritics already have very wide ranging appearances in various typographic traditions, even within the Greek-based scripts which N2976 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.

> a. *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 could be considered candidates for unification with existing marks.*

This is not supported by any argumentation in N2976 or elsewhere. Incidentally, most, if not all, scripts already use symbols from different blocks, part of a set or not. Why this particular exception? For which compelling technical reason? This criterion seems to be designed solely to justify *a posteriori* the misguided encoding of some N'ko-specific diacritics. It needs to be removed entirely.

> b. *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, only the part about different character properties should be retained.

> c. *the display behaviour is fundamentally different and requires different support. For example, U+A806 SYLOTI NAGRI SIGN HASANTA looks like a combining circumflex, but requires different display support.*

Display behaviour is dependent on the properties, so this is mostly redundant with *b*. Diacritic placement is generally supported in one of two ways in modern rendering engines: substitution to a precomposed glyph 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 placement of diacritics (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 N2949 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 (as was already requested in a previous document submitted to L2/UTC). A single unexplained analogy is no explanation or fleshing out.

As for U+A806 SYLOTI NAGRI SIGN HASANTA, it simply required different properties from U+0302 CIRCUMFLEX ACCENT and therefore required distinct encoding. There is no need to invoke a "display behaviour" criterion to justify its encoding.

> d. *the mark has been borrowed from another script, but has been significantly modified to fit with the ductus of the borrowing script, disunification may be preferred.*

The borrowed part does not belong here. There is nothing in N2976 or elsewhere that would support basing encoding decisions on a character being borrowed or not. ISO 10646 encodes characters, not their history. As for the second part, 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, as already requested to L2/UTC.

Security issues swept under the carpet

N2976 also states that there are security issues involved with disunification, without providing any detail, simply pointing to UTR #36. This is insufficient. There is a significant increase in the potential for spoofing when encoding script-specific diacritics that resemble common ones (see N2949). If WG2 is to have criteria for the disunification of combining diacritical marks included in its *Principles and Procedures* document, then this important security concern must be part of the set of criteria. This one would generally oppose disunification, unless trumped by other factors such as different properties (the security downside must then be accepted and properly dealt with) or sufficiently distinct appearance (which reduces the potential for spoofing increasingly as the shapes diverge).

Conclusion

In conclusion, we believe generic combining marks are a powerful feature of the UCS which sets it apart: all scripts may benefit from these marks right now, even for cases undreamt of by ISO standardizers, without waiting for years while ISO considers the addition of identically looking and behaving but script-specific signs. Any disunification has to be grounded, as much as possible, in 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 but a potentially increased security risk.