From: François Yergeau, Patrick Andries
Re: N'Ko Proposal in Amendment 2

The purpose of this document is to clarify some of the issues regarding the Canadian comments on the encoding proposal for the N'ko script in Amendment 2, with a focus on the arguments put forward by Mamady Doumbouya in WG2 N2898.

We first want to make clear that we welcome the encoding of N'ko and that this response should be viewed in the light of our attempt to improve the excellent work that lead to the N'ko proposal. We do not believe that any of the suggestions we have made – the removal of a few characters – would endanger an actual Unicode conformant implementation of N'ko.

1) Regarding the encoding of certain old forms of letters (07E8, 07E9 and 07EA)

N2898 first refers to the historical aspect of these characters offered by the delegate from Ireland. It would seem that these aspects would militate against the separate encoding of these characters. All scripts have undergone shape changes through their histories, and the usual manner to deal with such changes is not to encode old shapes separately, but to consider them as shape variants of the same letters.

N2898 stresses the necessity of faithfully reproducing the older writings of Solomana Kanté, in their original form. This necessity is not contested, but must be put in perspective with the same necessity facing any other script where letter shapes have changed. The solution is to use an appropriate font, not to encode the old forms. The Latin script has known many now obsolete forms such as Uncial and Fraktur. These forms are not and will not be encoded. When an old document needs to be reproduced in its original form (or when old character shapes have to be displayed in a modern document, as in Dr. Mamadi Baba Diané's N'ko version of the Qur'an) a font for this purpose is used, showing the old shapes for the letters. The same situation obtains for other scripts, with the same solution.

We have been told repeatedly that N'ko users require a plain text distinction of these old characters. It is noteworthy that N2898 does not mention this. And this begs the question of why such a plain text distinction would be required in the case of N'ko, while other scripts with older forms are handled with a font change. Why is N'ko special in this regard? There could be a hint in the fact that N'ko orthography has changed (two OLD RA's are now written RRA), but then again this is not unique, all orthographies vary, as well as the supporting script. And even if this orthographic change were to justify OLD RA, what about OLD JA and OLD CHA?

In addition to the paucity of arguments in favour of encoding the old shapes as characters, we also wish to bring attention to an argument that militates against this encoding: having a single code for, say, OLD RA and RA means that many kinds of text processing, most importantly search, will treat them identically. This means that someone searching for a word in a database or on the Internet does not have to remember to search it in both old and new forms (except of course in the case of orthographic change). This is a real benefit of *not* encoding the old shapes, which should not be underestimated.

2) Regarding N'ko diacritics

N2898 seems to indicate a belief that using the common diacritics (03xx) for N'ko would somehow

make them "secondary characters or not part of N'Ko character set but merely borrowed from other systems." This belief is misguided, since the 03xx diacritics are common, meant for use by all scripts and not reserved for any particular script. The Unicode standard is crystal-clear on this (section 7.7, http://www.unicode.org/versions/Unicode4.0.0/ch07.pdf): "The combining diacritical marks in this block are intended for general use with any script." In fact, those common diacritics are already used by many scripts – regardless of origin –without making them secondary or foreign in any of those scripts.

It has been suggested that the origin of diacritics should be a factor in deciding whether to use the common diacritics for marks of similar shapes. This is at best impractical: having to prove a common historical link between a new script's generic-looking diacritics and the ones in the common block would mean, if applied systematically in the future, that one would have to embark on a protracted study which could always be subject to review or worse, proven wrong (do we change the assignments then?). But to what avail would one need to do this laborious study? For which compelling technical reason? We believe it is best to consider the similarities in shape and behaviour of diacritics, not origin, when determining whether those from the common block are appropriate. Current practice seems to support this view, with U+0308 being used for the Greek dialytika, the German umlaut, the French tréma, the double derivative, the Cyrillic diaeresis, as a mark for numerals in Hebrew [1] and as a Syriac diacritic, without any common origin (or meaning, for that matter) being obvious at all.

N2898 also argues that all N'ko characters should be contiguous ("require ... to hunt for it characters in the other parts of the Unicode ... will be ... an error..."). But having characters in various blocks is already the common situation for most if not all scripts! Here is another excerpt from the Unicode standard (section 2.8, http://www.unicode.org/versions/Unicode4.0.0/ch02.pdf):

> "Characters used in a single writing system may be found in several different blocks. For example, characters used for letters for Latin-based writing systems are found in at least nine different blocks: Basic Latin, Latin-1 Supplement, Latin Extended-A, Latin Extended-B, IPA Extensions, Phonetic Extensions, Latin Extended Additional, Spacing Modifier Letters, and Combining Diacritical Marks."

This last block is precisely the 03xx that we are concerned with here. We would like to understand why N'ko would differ from the common case, which is to have characters from several blocks. There may be a reason but it still needs to be put forward.

It may be that *some* diacritics used in N'ko are truly unique to it and that no combining marks from 03xx are suitable (for instance their combining classes need to be different). These would of course need to be encoded specifically for N'ko (and presumably in the N'ko block), but they need to be argued as such, otherwise the designation of the 03xx block as generic stops making any sense.

In conclusion, in the absence of better evidence of the need for the old variant shapes as characters and for diacritics outside the generic block, we believe the most prudent approach is to remove these characters. Such characters could be added at a later date if evidence is provided.

---

1 Documentation accompanying the fonts Ezra SIL and SBL Hebrew recommends the use of U+0307 and U+0308 with Hebrew base characters for the single and double dots used in a rare style (neither biblical nor modern) of marking Hebrew numerals.