L2/10-155

title: *Reservations on L2/09-419 proposal*

authors: *Arno Schmitt & Khaled Hosny*

Koranic text is one of the most widely written Arabic texts and though it is written in Arabic language, it needs special handling because of the its unique orthographic conventions.

Unicode has long supported special Koranic characters, but there still few more needed for proper encoding of Koranic text.

The L2/09-419 proposal takes step towards full Koran coverage in Unicode, but we believe it has several shortcomings that we try to address in this document.

Tanwīn marks

The L2/09-419 proposal proposes encoding a set of *idġām* (sequential) *tanwīn* marks, and though we believe in the necessity of the proposed marks, we think it is not enough and an additional set of *iqlāb tanwīn* still required.

There are millions of copies of the Koran just using one set of *tanwīn* (ً ٌ ٍ), but to the best of our knowledge not a single one uses the proposed sequential *tanwīn* (ً ٌ ٍ) alone without using *iqlāb tanwīn* (ً ٌ ٍ) as well—here is a line first from a typical South-Asian print, then from the King-Fahd edition:

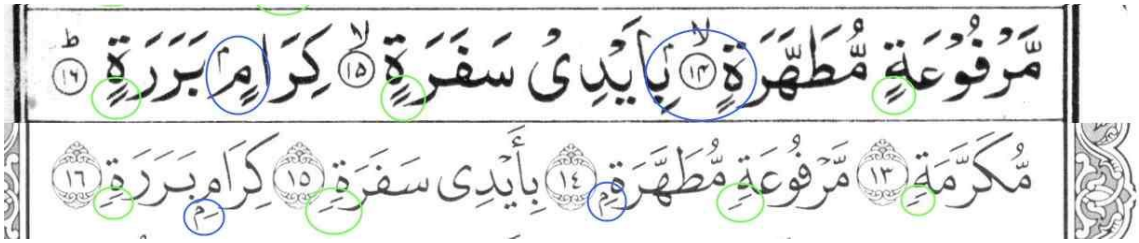


Figure 1 Different *tanwīn* forms, in the first line only standard *tanwīn* is used, in the second both the proposed sequential *tanwīn* (encircled in green) and the not yet proposed *iqlāb tanwīn* (encircled in blue) are used, thus to encode only half of the missing *tanwīn* marks is not sufficient for full representation of Koran.

High small *wāw*

One of the special features of Koranic text is the use of small letters to denote letters that were omitted in the original Koranic orthography. The small letters can be defined into two main groups; non-spacing and spacing. Both groups are transparent as regard to Arabic shaping; that is they neither cause the surrounding letters nor break already joined ones. The main difference between both groups is that the non-spacing group are combining characters in the sense that they set on a base letter, while the spacing group *float* between base letters.

The case against encoding high small *wāw*

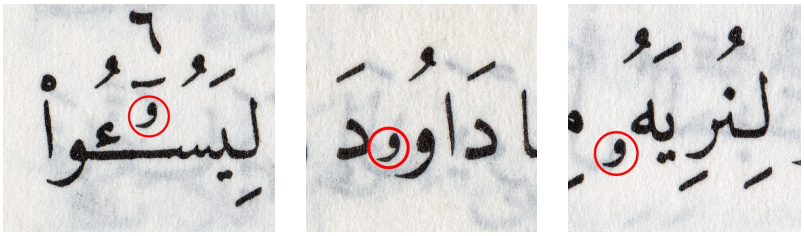
The L2/09-419 proposal proposes encoding a new combing small *wāw* as companion to the existing U+06E5 ARABIC SMALL WAW. The main argument for encoding a separate combining small *wāw* is that the currently encoded small *wāw*, U+06E5 ARABIC SMALL WAW, is a "*non-joining spacing character*", and thus does not fulfil the need of small *wāw* and the middle of the word, and so a new combining non-spacing small *wāw* is required.

The new proposed character is, however, against Unicode's general rule of not encoding positional variants of Arabic letters, since the proposed small *wāw* is essentially a middle of the word variant of U+06E5 ARABIC SMALL WAW as can be evidenced by the fact that it only exists in the middle of the word.

We believe that the proper approach is not to encode a new positional form U+06E5 ARABIC SMALL WAW, but to change the joining propriety of U+06E5 ARABIC SMALL WAW from non-joining to transparent, thus allowing for its occurrence in the middle of the word without breaking the joining of surrounding letters.

It should also be noted that the proprieties of the proposed new small *wāw* doesn't really fit with the middle form of the small *wāw*, since the proposed character is a combining mark while the middle form of small *wāw* is a spacing character as can be evidenced by the fact that it doesn't set on the base characters but between them, some times causing elongation of the inter-character connecting stroke to accommodate it.

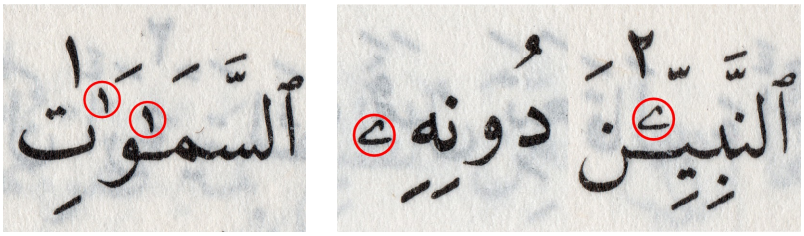
Spacing Arabic small letters have some unique characteristics; they are joining transpar-



middle | middle, between non-joined letters | final

Figure 1 Different positional forms of small *wāw*.

ent and thus doesn't interrupt the joining of the text, yet they are not combining marks, so instead of setting above a base letter they *float* between them. As can be seen in the following examples:



small *ʾalef* | small *yāʾ*

Figure 2 Different Arabic small letters, note how the left *ʾalef* is higher though in between two unjoined letters.

Though Unicode has both U+06E6 ARABIC SMALL YEH and U+06E7 ARABIC SMALL HIGH YEH, we believe that the later is a mere positional form of the former, the same way the proposed small *wāw* is a positional form of U+06E5 ARABIC SMALL WAW, and thus it was a mistake to encode it separately.

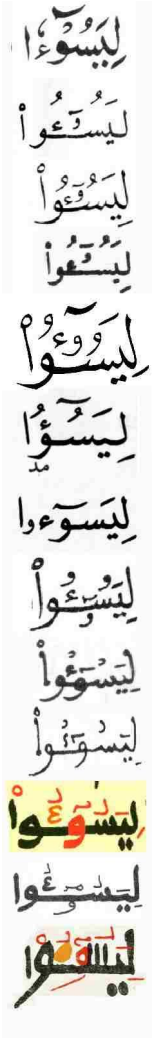


Figure 3 Different representations of small *wāw* showing that is not a combining mark but rather a spacing transparent character. Note in the last examples it is written in the size of regular *wāw*, sometimes in a different colour, attached to the rest of the word but still transparent in its joining behaviour.