Universal Multiple-Octet Coded Character Set
International Organization for Standardization
Organisation Internationale de Normalisation
Международная организация по стандартизации

Doc Type: Working Group Document
Title: Preliminary Proposal to enable the use of Combining Triple Diacritics in Plain Text (two possible solutions)
Author: Karl Pentzlin
Status: Individual Contribution
Action: For consideration by JTC1/SC2/WG2 and UTC
Date: 2010-09-23

Foreword for this preliminary proposal

This proposal is presented to the JTC1/SC2/WG2 October 2010 meeting in Busan, with the kind request to take a decision which of the two solutions presented here shall be followed in principle in the final proposal.
(While there are hints that the solution B is already preferred by the UTC, this could not be confirmed from the minutes of the UTC meetings. This preliminary proposal intends to document the possible solutions publicly; a public documentation of the decision about the way to proceed, by a SC2/WG2 resolution, is welcomed.)

1. Introduction

In several phonetic and dialectological transcription systems, diacritical marks are found which span over three base characters. However, there are only four forms found so far: conjoining bows above or below the base characters, a tilde below the base characters, and a roof (i.e. a circumflex) above the base characters.

Two of them (the bows) already were proposed (thus, the examples on which the proposals were based are not repeated here *[in this preliminary proposal; in the final version they are intended to be included]*).

- The "bow above" (here shown as U+1AFD COMBINING TRIPLE INVERTED BREVE ABOVE) was proposed in L2/09-028 (N3571) "Proposal to encode additional characters for the Uralic Phonetic Alphabet" by Finnish and Irish NB as U+1DFB COMBINING TRIPLE INVERTED BREVE.
- The "bow below" (here shown as U+1AFD COMBINING TRIPLE BREVE BELOW) was proposed in L2/08-392 "Proposal to encode a combining diacritical mark for Low German dialect writing" as U+1DFD COMBINING TRIPLE BREVE BELOW.
- Regarding the third mark, an example is shown in Fig. 1.

As there many projects to encode characters for the dialectology of several countries are ongoing, such triple marks are urgently needed. Fig. 3 shows an example from Slovenia, proving that the triple marks are really productive.

To use such triple marks with the Latin script provides a challenge for rendering engines anyway.

Here, two solutions are proposed:

- **Solution A** proposes to encode the combining triple marks as single characters, like it is done for combining double marks like U+035C...U+0362, U+1DCD, and U+1DFC.
- **Solution B** proposes to use building blocks, which can be used to create bows spanning over/below an arbitrary long sequence of base characters, using the mechanism provided by U+FE24...FE26 to accomplish continuous macrons over sequences of Coptic letters.

In principle, both solutions can be accepted together, to use the triple diacritics for units with defined semantics (like in dialect writing/transcription), while to use the building blocks for prosody and other kinds of marking arbitrary sequences of characters.
This has precedence in encoding U+FE22/FE23 COMBINING DOUBLE TILDE LEFT/RIGHT HALG in parallel to U+0360 COMBINING DOUBLE TILDE.

In fact, accepting Solution A, together with U+FE27...FE2C from Solution B, provides the maximum of flexibility and usability together with a high degree of consistency.

2. Discussion of Solution A

To have the triple diacritical marks encoded as characters of their own, has the advantage that the character identity is clearer from the beginning on. Semantic units, as the elements of dialectological writing/transcribing, are preserved.

The triple marks, as proposed here, are to be combined with (i.e. to be input after) the first of the three base characters that they are to be applied to. This is in conformance with the use of the already encoded double marks, which are combined with the first of the two base characters that they are applied to.

To the rendering engine, the new demands are marginal. Even now, with double diacritics spanning over two base characters, a rendering system of quality has to take into account the width of both base characters, and the placement of all combining marks which are attached to any of these two base characters. Thus, the rendering engine can generate its graphical output only when the second base character (which in fact follows the double mark) with its entire single combining marks is read. The introduction of triple combining characters extends this for one base character more, without introducing something totally new.

To use the system of Canonical Combining Class Values in an optimal way, two new such values should be introduced (235 "Triple below" and 236 "Triple above", following the existing 233 "Double below" and 234 "Double above").
This would have the effect to get the triple marks "outermost", beyond the double marks which are the "outermost" diacritics until now (ignoring the special case of iota subscriptum).
If any stability policy prevents this, the values 233 and 234 can be used for triple marks also, as no example of stacking double and triple marks together is known until now.

3. Discussion of Solution B

This solution follows the principle outlined in the UTC Action Item 120-A87 (see L2/09/225, re L2/09-281 COMBINING TRIPLE INVERTED BREVE and other triple-length combining marks" by Deborah Anderson), where is stated:

" Action Item for Peter Constable (UTC Liaison to WG2): Triple marks should be handled by the mechanism associated with two part diacritics in the Combining Half Marks block at U+FE20."

First, this solution requires the scope of the already encoded characters:
U+FE20 COMBINING LIGATURE LEFT HALF
U+FE26 COMBINING CONJOINING MACRON
U+FE20 COMBINING LIGATURE LEFT HALF
to be expanded:

- Any sequence of one base character with U+FE20 applied, one or more base characters each with U+FE26 applied, and one base character with U+FE21 applied, shall yield a ligature bow over the complete sequence of base characters, starting with the one to which U+FE20 is applied, and ending with the one to which U+FE21 is applied.
- Then, the same applies to sequences regarding the proposed U+FE27, U+FE2C, U+FE29, except that the ligature bow goes under the character sequence.
- Likewise, the same applies to sequences regarding the proposed U+FE28, U+FE2C, U+FE29, except that this yields a tilde under the character sequence.
  As both the ligature bow and the tilde end in a turn up, and as the sequences are to be rendered in a special way anyway determined by the starting part, the end parts for bow and tilde are unified in U+FE29.
  (Otherwise, either the obvious reservation for the left and right “macron below” halves had to be dropped, or the characters could not have been placed into the same block.)
- Also, this applies to sequences regarding the proposed U+FE2D, U+FE2F, U+FE2E, except that a circumflex instead a bow is applied over the sequence.
  The application of more than one U+FE2F achieves only one peak displayed regardless how much U+FE2F are contained in the sequence.

  (This shows one disadvantage of Solution B, compared with Solution A: Three different characters are needed to get the same graphical representation, which is achieved with Solution A with only one character.)

The demands on the rendering engine are more extensive than for Solution A.
Until now, the "building block principle" only duplicates the way two special characters may be represented (U+FE20/U+FE21 resemble U+0361, and U+FE22/U+FE23 resemble U+0360), or renders at straight lines which cause no real trouble if not applied in the correct order (i.e. a base character with U+FE24 is to be followed by one with U+FE26, and one with U+FE26 is to be followed by one either with U+FE26 again or with U+FE25).

Solution B would extend this to arbitrary long bows that have to be computed after processing the whole character sequence until the correct end part has to be found, including doing all error processing if the expected parts of the sequence are not found, or if the required beginning part is missed.

However, there are applications at least for the bows where it is desirable to apply bows of arbitrary length in plain text (see fig. 2).

Also, copying and pasting base characters within such sequences may be more transparent to the user, as (at least when the base characters with all its combining diacritical marks are affected) carry their "bow part" with them. However, this may be a drawback when the triple marks are part of units of a dialectological writing/transliterating system.

4. Proposed Characters — Solution A

The code points are based on the assumption that the new block "Combining Diacritical Marks Extended-A at U+1AB0…U+1AFF, as contained in other proposals, is accepted.

U+1AFC COMBINING TRIPLE CIRCUMFLEX ABOVE

U+1AFD COMBINING TRIPLE INVERTED BREVE ABOVE

U+1AFE COMBINING TRIPLE BREVE BELOW

U+1AFF COMBINING TRIPLE TILDE BELOW

Properties:

```
1AFC;COMBINING TRIPLE CIRCUMFLEX ABOVE;Mn;236;NSM;;;;;N;;;;;
1AFD;COMBINING TRIPLE INVERTED BREVE ABOVE;Mn;236;NSM;;;;;N;;;;;
1AFE;COMBINING TRIPLE BREVE BELOW;Mn;235;NSM;;;;;N;;;;;
1AFF;COMBINING TRIPLE TILDE BELOW;Mn;235;NSM;;;;;N;;;;;
```

New Canonical Combining Class Values:

235 – Triple below
236 – Triple above

5. Proposed Characters — Solution B

U+FE27 COMBINING LIGATURE BELOW LEFT HALF

U+FE28 COMBINING TILDE BELOW LEFT HALF

U+FE29 COMBINING LIGATURE OR TILDE BELOW RIGHT HALF

U+FE2A, U+FE2B: reserved (for "combining conjoining macron below" halves, if needed)

U+FE2C COMBINING CONJOINING MACRON BELOW

U+FE2D COMBINING CONJOINING CIRCUMFLEX ASCENDER

U+FE2E COMBINING CONJOINING CIRCIMFLEX DESCENDER

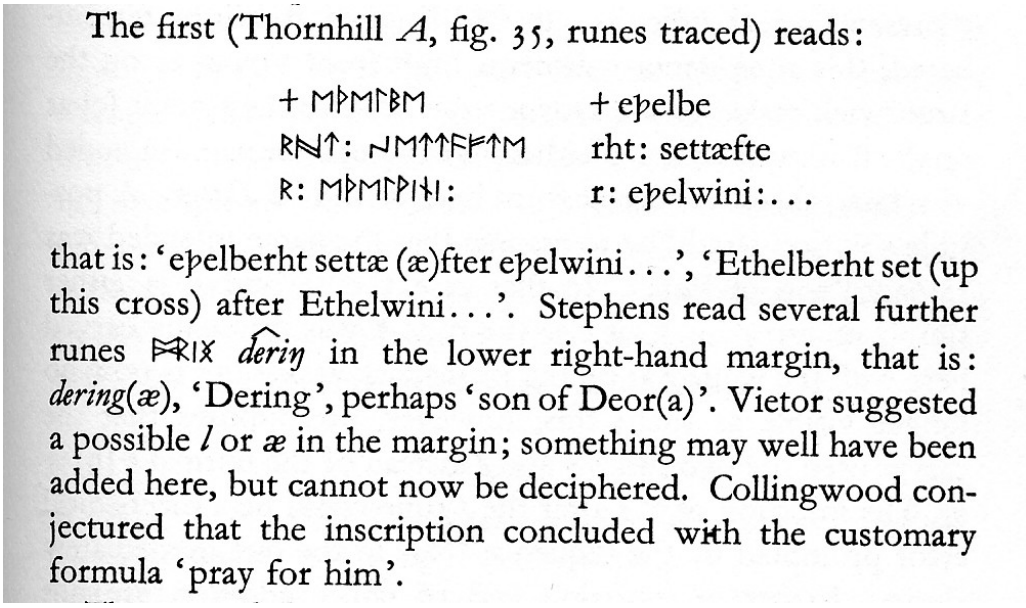
U+FE2F COMBINING CONJOINING CIRCUMFLEX PEAK
- the conjoining of more than one of U+FE2F render in a larger single peak on the affected base characters

Properties:

```
FE27;COMBINING LIGATURE BELOW LEFT HALF;Mn;230;NSM;;;;;N;;;;;
FE28;COMBINING TILDE BELOW LEFT HALF;Mn;230;NSM;;;;;N;;;;;
FE29;COMBINING LIGATURE OR TILDE BELOW RIGHT HALF;Mn;230;NSM;;;;;N;;;;;
FE2C;COMBINING CONJOINING MACRON;Mn;230;NSM;;;;;N;;;;;
FE2D;COMBINING CONJOINING CIRCUMFLEX ASCENDER;Mn;230;NSM;;;;;N;;;;;
FE2E;COMBINING CONJOINING CIRCUMFLEX DESCENDER;Mn;230;NSM;;;;;N;;;;;
FE2F;COMBINING CONJOINING CIRCUMFLEX PEAK;Mn;230;NSM;;;;;N;;;;;
```

6. Examples and Figures

Fig. 1: Elliot's "Runes" (Manchester University Press, 1987), showing an example of a three-letter bind rune transliterated with a circumflex over the three letters "der".
Thanks to Andrew West for providing this example.

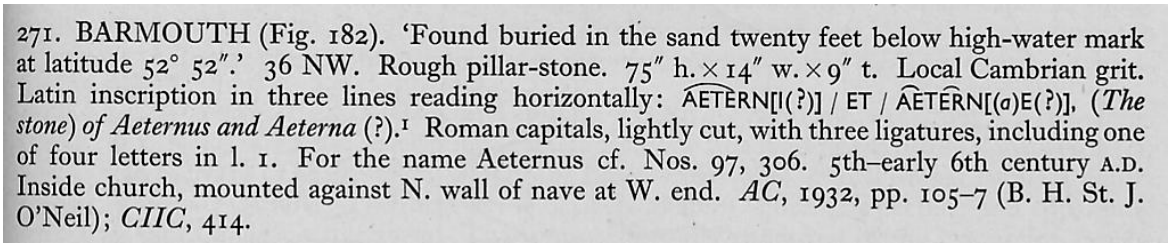


The first (Thornhill *A*, fig. 35, runes traced) reads:

+ ᛗᚦᛗᛚᛒᛗ	+ eþelbe
ᚱᚻᛏ: ᚿᛗᛏᛏᚠᚠᛏᛗ	rht: settæfte
ᚱ: ᛗᚦᛗᛚᚦᛁᚻᛁ:	r: eþelwini:...

that is: 'eþelberht settæ (æ)fter eþelwini...', 'Ethelberht set (up this cross) after Ethelwini...'. Stephens read several further runes ᚹᚱᛁᚷ *dêriŋ* in the lower right-hand margin, that is: *dering*(æ), 'Dering', perhaps 'son of Deor(a)'. Vietor suggested a possible *l* or æ in the margin; something may well have been added here, but cannot now be deciphered. Collingwood conjectured that the inscription concluded with the customary formula 'pray for him'.

Fig. 2: V. E. Nash-Williams, "The Early Christian Monuments of Wales". Cardiff, 1950: showing bows upon letter sequences of different length, which denote some features of the text, rather than marking special phonetic units with distinctive semantics.
Thanks to Andrew West for providing this example.



271. BARMOUTH (Fig. 182). 'Found buried in the sand twenty feet below high-water mark at latitude 52° 52″.' 36 NW. Rough pillar-stone. 75″ h. × 14″ w. × 9″ t. Local Cambrian grit. Latin inscription in three lines reading horizontally: AETERN[I(?)] / ET / AETERN[(a)E(?)], (*The stone*) *of Aeternus and Aeterna* (?).[1] Roman capitals, lightly cut, with three ligatures, including one of four letters in l. 1. For the name Aeternus cf. Nos. 97, 306. 5th–early 6th century A.D. Inside church, mounted against N. wall of nave at W. end. *AC*, 1932, pp. 105–7 (B. H. St. J. O'Neil); *CIIC*, 414.

Fig. 3**:** Excerpt from the PUA of the Font ZRCola.ttf used for the Obščeslavjanskij ingvističeskij atlas, showing a lot of transcription units using bows on three base letters (at EE0C, EE25...EE28, EE3D).
Thanks to Peter Weiss for providing that font.

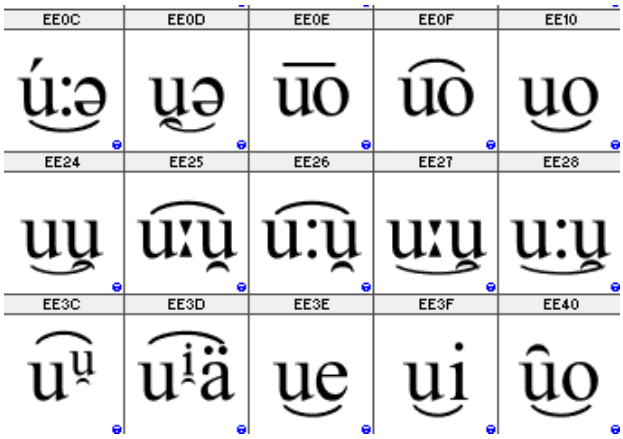


Fig. 4**:** A specimen for the triple tilde below from:
Wenker, Georg, et al.: Deutscher Sprachatlas auf der Grundlage des Sprachatlas des Deutschen Reichs, Marburg (Lahn) 1927-1956; introduction, p. 18.