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

Doc Type: Working Group Document
Title: Proposal to enable the use of Combining Triple Diacritics in Plain Text
Source: German NB
Status: National Body Contribution
Action: For consideration by JTC1/SC2/WG2 and UTC
Date: 2011-05-22

1. Introduction

In several phonetic and dialectological transcription systems, diacritical marks are found which span over three base characters. Such marks are: conjoining bows above or below the base characters, and a tilde below the base characters.

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.

The solution presented here 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:

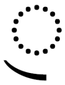
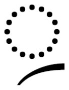
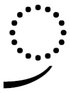
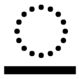
" Action Item for Peter Constable (UTC Liaison to WG2): Triple marks should be handled by the mechanism associated with two part diacritics in the Combining Half Marks block at U+FE20."

> Note: WG2 N3915 = L2/10-353 "Preliminary Proposal to enable the use of Combining Triple Diacritics in Plain Text" had proposed to encode four triple diacritics as units rather than being composed of blocks:
> U+1AFC COMBINING TRIPLE CIRCUMFLEX ABOVE (not addressed in this proposal)
> U+1AFD COMBINING TRIPLE INVERTED BREVE ABOVE (also proposed in WG2 N3571)
> U+1AFE COMBINING TRIPLE BREVE BELOW
> U+1AFF COMBINING TRIPLE TILDE BELOW
> It is to be noted that MUFI 3.0 contains a U+F1FC COMBINING TRIPLE BREVE BELOW in its PUA.

First, this solution expands the scope of the already encoded characters:
U+FE20 COMBINING LIGATURE LEFT HALF
U+FE26 COMBINING CONJOINING MACRON
U+FE21 COMBINING LIGATURE RIGHT HALF:

- 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+FE2A, U+FE29, except that the ligature bow goes under the character sequence.
- Likewise, the same applies to sequences regarding the proposed U+FE28, U+FE2A, 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 are unified in U+FE29.

2. Proposed Characters

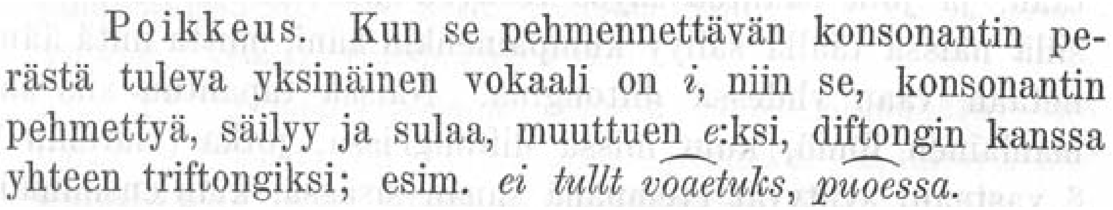
	U+FE27	COMBINING LIGATURE BELOW LEFT HALF
	U+FE28	COMBINING TILDE BELOW LEFT HALF
	U+FE29	COMBINING LIGATURE OR TILDE BELOW RIGHT HALF
	U+FE2A	COMBINING CONJOINING MACRON BELOW

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;;;;;
FE2A;COMBINING CONJOINING MACRON;Mn;230;NSM;;;;;N;;;;;
```

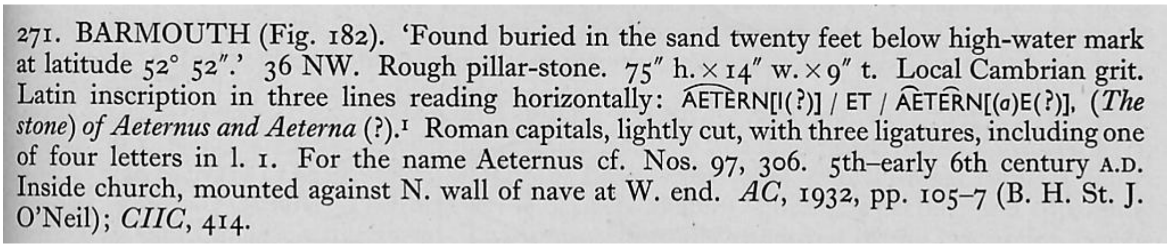
3. Examples and Figures

Fig. 1: The diacritic "combining triple inverted breve" indicates a triphthong (a group of three vowels in one syllable) in the UPA. Example from Ossian Grotenfelt, "Pohjois-Hämeen kielimurteesta", in Suomi II:12 (page 321, rows 22–25).
This figure and its legend is copied from fig. 1 of WG2 N3571 = L2/09-028 "Proposal to encode additional characters for the Uralic Phonetic Alphabet" by Klaas Ruppel, Tero Aalto, Michael Everson, 2009-01-27.



Poikkeus. Kun se pehmennettävän konsonantin perästä tuleva yksinäinen vokaali on *i*, niin se, konsonantin pehmettyä, säilyy ja sulaa, muuttuen *e*:ksi, diftongin kanssa yhteen triftongiksi; esim. *ei tullt voaetuks, puoessa.*

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 / A͡ETE͡RN[(*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 in its PUA several of transcription units using bows on three base letters (at EE0C, EE25...EE28, EE3D).
Thanks to Peter Weiss for providing that font.

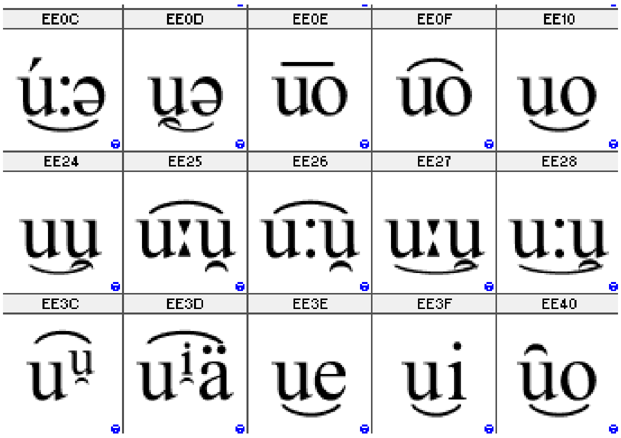


Fig. 4: A specimen for the triple tilde below from:
Wenker, Georg, et al.: Deutscher Sprachatlas auf der Grundlage des Sprachatlas des Deutschen Reichs, Marburg (Lahn) 1927-1956; introduction, p. 18.

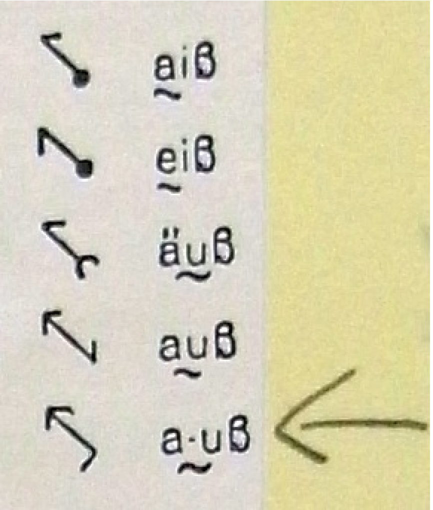
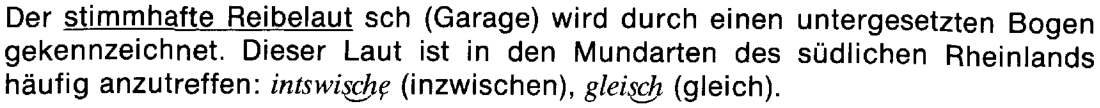
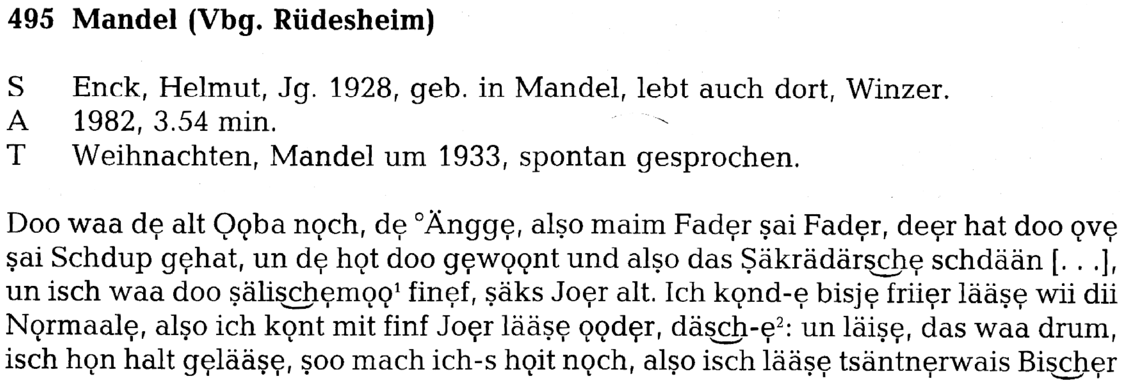


Fig. 5: Introduction a triple breve below for a German dialectal orthography, from:
Honnen, Peter: Rheinische Dokumenta. Lautschrift für rheinische Mundarten.
Pulheim 1987, ISBN 3-7927-0947-3; p. 16.



Der stimmhafte Reibelaut sch (Garage) wird durch einen untergesetzten Bogen gekennzeichnet. Dieser Laut ist in den Mundarten des südlichen Rheinlands häufig anzutreffen: *intswisc͜hẹ* (inzwischen), *gleisc͜h* (gleich).

Fig. 6: Sample of the German dialectal orthography introduced in the source of fig. 5, from:
Cornelissen, Georg, Honnen, Peter, Langensiepen, Fritz (ed.):
Das rheinische Platt • Eine Bestandsaufnahme (Handbuch der rheinischen Mundarten, Teil 1: Texte) – Köln 1989, ISBN 3-7927-0689-X; p. 495.



495 Mandel (Vbg. Rüdesheim)

S Enck, Helmut, Jg. 1928, geb. in Mandel, lebt auch dort, Winzer.
A 1982, 3.54 min.
T Weihnachten, Mandel um 1933, spontan gesprochen.

Doo waa dẹ alt Ọọba nọch, dẹ °Änggẹ, alṣo maim Fadẹr ṣai Fadẹr, deẹr hat doo ọvẹ ṣai Schdup gẹhat, un dẹ họt doo gẹwọọnt und alṣo das Ṣäkrädärsc͜hẹ schdään [. . .], un isch waa doo ṣälisc͜hẹmọọ[1] finẹf, ṣäks Joẹr alt. Ich kọnd-ẹ bisjẹ friiẹr lääṣẹ wii dii Nọrmaalẹ, alṣo ich kọnt mit finf Joẹr lääṣẹ ọọdẹr, däsc͜h-ẹ[2]: un läiṣẹ, das waa drum, isch họn halt gẹlääṣẹ, ṣoo mach ich-s họit nọch, alṣo isch lääṣẹ tsäntnẹrwais Bisc͜hẹr

ISO/IEC JTC 1/SC 2/WG 2
PROPOSAL SUMMARY FORM TO ACCOMPANY SUBMISSIONS
FOR ADDITIONS TO THE REPERTOIRE OF ISO/IEC 10646[1]
Please fill all the sections A, B and C below.
Please read Principles and Procedures Document (P & P) from http://www.dkuug.dk/JTC1/SC2/WG2/docs/principles.html **for guidelines and details before filling this form.**
Please ensure you are using the latest Form from http://www.dkuug.dk/JTC1/SC2/WG2/docs/summaryform.html**.**
See also http://www.dkuug.dk/JTC1/SC2/WG2/docs/roadmaps.html **for latest *Roadmaps*.**

A. Administrative

1. **Title:** ***Proposal to enable the use of Combining Triple Diacritics in Plain Text***
2. Requester's name: *German NB*
3. Requester type (Member body/Liaison/Individual contribution): *Member Body Contribution*
4. Submission date: *2011-05-22*
5. Requester's reference (if applicable):
6. Choose one of the following:
    This is a complete proposal: *Yes*
    (or) More information will be provided later: *No*

B. Technical – General

1. Choose one of the following:
    a. This proposal is for a new script (set of characters): *No*
        Proposed name of script:
    b. The proposal is for addition of character(s) to an existing block: *Yes*
        Name of the existing block: *Combining Half Marks*
2. Number of characters in proposal: *4*
3. Proposed category (select one from below - see section 2.2 of P&P document):

A-Contemporary	*X*	B.1-Specialized (small collection)		B.2-Specialized (large collection)	
C-Major extinct		D-Attested extinct		E-Minor extinct	
F-Archaic Hieroglyphic or Ideographic				G-Obscure or questionable usage symbols	

4. Is a repertoire including character names provided? *Yes*
    a. If YES, are the names in accordance with the “character naming guidelines” in Annex L of P&P document? *Yes*
    b. Are the character shapes attached in a legible form suitable for review? *Yes*
5. Fonts related:
    a. Who will provide the appropriate computerized font to the Project Editor of 10646 for publishing the standard?
    To be announced
    b. Identify the party granting a license for use of the font by the editors (include address, e-mail, ftp-site, etc.):
6. References:
    a. Are references (to other character sets, dictionaries, descriptive texts etc.) provided? *Yes*
    b. Are published examples of use (such as samples from newspapers, magazines, or other sources) of proposed characters attached? *Yes*
7. Special encoding issues:
    Does the proposal address other aspects of character data processing (if applicable) such as input, presentation, sorting, searching, indexing, transliteration etc. (if yes please enclose information)? *Yes*
    Reference to the mechanism of "combining marks" (see text)
8. Additional Information:

Submitters are invited to provide any additional information about Properties of the proposed Character(s) or Script that will assist in correct understanding of and correct linguistic processing of the proposed character(s) or script. Examples of such properties are: Casing information, Numeric information, Currency information, Display behaviour information such as line breaks, widths etc., Combining behaviour, Spacing behaviour, Directional behaviour, Default Collation behaviour, relevance in Mark Up contexts, Compatibility equivalence and other Unicode normalization related information. See the Unicode standard at http://www.unicode.org for such information on other scripts. Also see http://www.unicode.org/Public/UNIDATA/UCD.html and associated Unicode Technical Reports for information needed for consideration by the Unicode Technical Committee for inclusion in the Unicode Standard.

[1] Form number: N3702-F (Original 1994-10-14; Revised 1995-01, 1995-04, 1996-04, 1996-08, 1999-03, 2001-05, 2001-09, 2003-11, 2005-01, 2005-09, 2005-10, 2007-03, 2008-05, 2009-11)

C. Technical - Justification

1. Has this proposal for addition of character(s) been submitted before? *No*
   If YES explain *(There was a preliminary proposal: WG2 N3915 = L2/10-353)*
2. Has contact been made to members of the user community (for example: National Body, user groups of the script or characters, other experts, etc.)? *Yes*
   If YES, with whom? *See text*
   If YES, available relevant documents:
3. Information on the user community for the proposed characters (for example: size, demographics, information technology use, or publishing use) is included? *Yes*
   Reference: *Scientific use (see text)*
4. The context of use for the proposed characters (type of use; common or rare) *Common*
   Reference: *See text*
5. Are the proposed characters in current use by the user community? *Yes*
   If YES, where? Reference: *See text*
6. After giving due considerations to the principles in the P&P document must the proposed characters be entirely in the BMP? *Yes*
   If YES, is a rationale provided? *Yes*
   If YES, reference: *To keep them in line with related characters*
7. Should the proposed characters be kept together in a contiguous range (rather than being scattered)? *Yes*
8. Can any of the proposed characters be considered a presentation form of an existing character or character sequence? *No*
   If YES, is a rationale for its inclusion provided?
   If YES, reference:
9. Can any of the proposed characters be encoded using a composed character sequence of either existing characters or other proposed characters? *No*
   If YES, is a rationale for its inclusion provided?
   If YES, reference:
10. Can any of the proposed character(s) be considered to be similar (in appearance or function) to an existing character? *No*
    If YES, is a rationale for its inclusion provided?
    If YES, reference:
11. Does the proposal include use of combining characters and/or use of composite sequences? *Yes*
    If YES, is a rationale for such use provided? *Yes*
    If YES, reference: *See text*
    Is a list of composite sequences and their corresponding glyph images (graphic symbols) provided? *Yes*
    If YES, reference: *See text*
12. Does the proposal contain characters with any special properties such as control function or similar semantics? *No*
    If YES, describe in detail (include attachment if necessary)
13. Does the proposal contain any Ideographic compatibility character(s)? *No*
    If YES, is the equivalent corresponding unified ideographic character(s) identified?
    If YES, reference: