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

In memoriam Alan M. Kent.

0. Summary. This proposal requests the encoding of one combining mark. If this proposal is accepted, the following character will exist:

1ACF   COMBINING OVERCURL

　　　　　• used in medieval Cornish, English, Latin

　　　　　• often fuses typographically with its base letter

1. Previous discussion. A variety of objections to this character have been raised since it was proposed in 2017. None of them responds to the unique behaviour and use of this diacritical abbreviation mark, however. It is noteworthy that WG2 and the UTC agreed that the abbreviation mark existed.

1.1. The combining character accepted. At first the diacritical mark was accepted at a WG2 meeting and put on a ballot. The US asked to have the character removed because of discussion at the UTC.

1.2. Atomic characters proposed. The US suggested that atomic characters be encoded. This was objectionable because the character really is a diacritical mark. There is no "letter n̂" in any language in the same way that there is a "letter ŋ".

It's simply wrong to consider these like this. Moreover the sign is found everywhere in Britain, from Middle Cornish to Middle English and Middle Scots. Because there is no character for it, some writers have indicated it (decades back) with an apostrophe. But the apostrophe is a different character. In any case, the argument for "atomic characters" was nothing more than "font developers won't know how to draw it", which is pretty poor argument given the simplicity of this feature and the complexity of many scripts in the standard. It's a loop that attaches to the base character in a convenient way. I proposed a Unicode Technical Note on the typography of the character but never received a response to that.

1.3. Unify with INVERTED BREVE ABOVE. I think it was Peter Constable who suggested unifying it with U+0311 But the INVERTED BREVE ABOVE is used in traditional Slavicist notation of Serbo-Croatian phonology to indicate long falling accent. It is placed above the syllable nucleus, which can be one of five vowels (â ê î ô û) or syllabic r̂.

It cannot be suggested that South Slavic linguists would accept the overcurl as a valid glyph variant.

1

2. Progress. An impasse exists with regard to the correct encoding of the COMBINING OVERCURL proposed in N4907. Experts in SC2 the UTC and in SC2 have agreed that the evidence indicates the existence of the polyvalent diacritical mark, but there has not been consensus on how it should be encoded. The issue is that some in the UTC have suggested that it might be too difficult for font developers to support attaching the overcurl to a base letter. I really don't believe that is true. I authored Irish ballot comments which stated clearly that the attachment of the mark to base letters in the manuscripts differs from letter to letter and hand to hand, and that in type attaching it "conveniently" was sufficient. Moreover, the UCS contains scripts which have *far more* complicated glyphs. Consider that every Cyrillic letter can be enclosed by U+0488 and U+0489. That's complex enough even if the ink doesn't touch. Or consider Lanna or Telugu or Sinhala.

But there's a solution. The wrong thing to do is to try to encode these characters atomically, as was suggested in 2017 or 2018. There are no orthographies which use these as *letters* per se—they are always genuine diacritical marks, which are polyvalent, meaning *m* or *n* or other things, or nothing. In terms of the user community (medievalists with a specific interest in palaeography), the usage scenario is one in which there are real advantages to treat the OVERCURL in the same way that U+0306 COMBINING OVERLINE, U+0311 COMBINING INVERTED BREVE, and U+0352 COMBINING FERMATA and other palaeographic marks are. To edit texts with these as precomposed elements would not be convenient—I have myself had extensive experience editing two Cornish texts and one Middle Scots text and correcting and counting and doing other things with characters like this would definitely be hindered with atomic characters.

Belgian palaeographer Albert Derolez told me that the OVERCURL is not really known in continental Europe. It is definitely a development in Britain, and I have seen it in wide use in Middle Cornish, Middle English, and Middle Scots. There is a Scottish Gaelic manuscript where Gaelic is written in a Middle-Scots-based orthography (*The Book of the Dean of Lismore*) and I suspect it will turn up there as well. It may not occur in Middle Welsh texts.

2.1. COMBINING ZIGZAG ABOVE. We have at present a character which can both stand free above a letter, or be fused with it: U+035B COMBINING ZIGZAG ABOVE. In European manuscripts it tends to have a lightning-bolt or s-shape, and it sits above consonants or vowels. But in Britain it is drawn with a hairline that attaches to the base letter (not in the same place that the OVERCURL attaches!) both in manuscripts and in lead-type fonts. These are the ones I used in my edition of the Cornish Passion Poem:

$$\text{č d̂ ĝ m̂ n̂ p̂ r̂ ʒ̂ ŝ ft̂ t̂ þ̂ û v̂ ŵ x̂}$$

Every one of these can also be drawn without this glyph. It is the same character, and many British manuscripts use the simpler form (particularly those in Latin). In fact before I decided to support traditional British typography for this character, I had this subset of glyphs in my font:

$$\text{d̂ ĝ m̂ n̂ p̂ r̂ ft̂ t̂}$$

And I was quite satisfied with these. They are legible and they are reasonable representations of the ZIGZAG. And with no Opentype features at all, it's still legible (though *d* and *t* aren't great):

$$\text{č d ĝ m̂ n̂ p̂ r̂ ʒ̂ ŝ t̂ þ̂ û v̂ ŵ x̂}$$

This is how the ᴢɪɢᴢᴀɢ is represented in the monowidth font which I use for plain-text editing, including the generation of indexes. The combination with *t* could use some tweaking, but this is completely legible at small sizes.

<p style="text-align:center; font-size:1.5em">č̓ d̓ g̓ m̓ n̓ p̓ r̓ ʒ̓ s̓ t̓ þ̓ u̓ v̓ w̓ x̓</p>

And in Continental Europe, the ꝓ-type ligature is unknown. Only the ꞅ-type ligature is used.

The same solution can be accepted for the ᴏᴠᴇʀᴄᴜʀʟ.

2.2. ʟᴀᴛɪɴ sᴍᴀʟʟ ʟᴇᴛᴛᴇʀ ɪs. We have a similar situation with the character U+A7CD ʟᴀᴛɪɴ sᴍᴀʟʟ ʟᴇᴛᴛᴇʀ ɪs which, ideally, should *always* combine typographically with letters which precede it:

<p style="text-align:center; font-size:1.5em">cẜ dẜ fẜ gẜ kẜ rẜ ʒẜ ſtẜ</p>

These too are legible when proper ligation is not supported:

<p style="text-align:center; font-size:1.5em">cẜ dẜ fẜ gẜ kẜ rẜ ʒẜ ſtẜ</p>

And they are legible in a monowidth font:

<p style="text-align:center; font-size:1.5em">cẜ dẜ fẜ gẜ kẜ rẜ ʒẜ ſtẜ</p>

2.3. The ᴏᴠᴇʀᴄᴜʀʟ. The same holds for the ᴏᴠᴇʀᴄᴜʀʟ. These are the letters with overcurl that I used in my printed edition of the Cornish Passion Poem:

<p style="text-align:center; font-size:1.5em">a⃝ e⃝ g⃝ i⃝ m⃝ n⃝ r⃝ s⃝ t⃝ u⃝ y⃝</p>

Here is the same set opentype features turned off. Once again, it is still legible:

<p style="text-align:center; font-size:1.5em">a⃝ e⃝ g⃝ i⃝ m⃝ n⃝ r⃝ s⃝ t⃝ u⃝ y⃝</p>

Here are the same ones in Everson Mono; I used this working with Andrew West, who assisted in drawing up the Index Verborum:

<p style="text-align:center; font-size:1.5em">a⃝ e⃝ g⃝ i⃝ m⃝ n⃝ r⃝ s⃝ t⃝ u⃝ y⃝</p>

This is a simple half-circle at a 45° angle. It is a legible and distinctive representation. I would have no problem whatsoever having the code charts display something like this, with whatever useful informative note we might have in the names list, and would be happy that a Unicode Technical Note can give further guidance for this and indeed for the ᴢɪɢᴢᴀɢ, which has this behaviour (whether we recognized that or not back when it was encoded).

<p style="text-align:center; font-size:1.5em">◌⃝</p>

The ᴏᴠᴇʀᴄᴜʀʟ is productive, and is not a part of any formal orthography; it has a specialist use for medieval palaeography The best way to support it is by the addition of a single character to the UCS. At the WG2 meeting in Hohhot on 2017-09-27, the Medieval Ad Hoc recommended to encode the character. It has since been published (Williams, Everson, & Kent 2020),

3. Linebreaking. Line-breaking properties for these are suggested as follows.

1ACF: CM (Combining Mark)

4. Unicode Character Properties. Character properties are proposed here.

`1ACF;COMBINING OVERCURL;Mc;210;L;;;;;;N;;;;;`

5. Draft Unicode Technical Report.

DRAFT Unicode Technical Note No. XXX: -Rendering the COMBINING OVERCURL

Michael Everson

2023-06-01

U+1ACF COMBINING OVERCURL is a medieval abbreviation character with a similar function to U+0306 COMBINING OVERLINE, U+0311 COMBINING INVERTED BREVE, and U+0352 COMBINING FERMATA. It differs from them in that it fuses with the base consonant as an upward- and leftward-swinging swash. This abbreviation mark was only recently discovered as a palaeographically relevant feature of some medieval handwriting, and there has been no typographic tradition for representing it. Some older transcriptions of Middle English texts indicated it with an apostrophe ', but these are not palaeographic readings, and the apostrophe is a workaround in these texts, and of course is used as a character in its own right for other things. Note that in Middle English the COMBINING OVERCURL may either be decorative, or indicate a final -e, the status of which is problematic in Middle English, which was undergoing a process of loss of that vowel in that position. In Middle Cornish, the COMBINING OVERCURL may either be decorative, or indicate a following -m or -n. To date the COMBINING OVERCURL has only been observed with a subset of the Latin script (with the letters a e i m n r s t u y), but this technical note presents reasonable typographic representations for combination with the set of basic Latin letters, to guide and inspire implementors who wish to include support for it in their fonts.

Some background on the relation of this abbreviation mark with others may be beneficial. It appears that the original "nasal replacement" character was simply a horizontal bar. Later this came to be written with a curve, which sometimes had a dot inside it. The OVERCURL originated as a decorative swash, probably mimicking the INVERTED BREVE, but it differs from that in a very important way. The OVERLINE, INVERTED BREVE, and FERMATA always mean that the letter they are above is to be followed by an m or an n. But the OVERCURL is ambiguous: the letter it is attached to may be followed by an m or an n, or the OVERCURL may be nothing more but a decorative swash. Sometimes the dot of the FERMATA is also given by the scribe. This never means anything, but it can be represented by following the OVERCURL with U+0307 COMBINING DOT ABOVE, and drawing it underneath the OVERCURL.

$$\bar{a} \ \hat{a} \ \hat{a} \ \widehat{a} \ \widehat{a} \ \bar{a} \ \hat{a} \ \hat{a} \ \widehat{a} \ \widehat{a}$$
$$\overline{m} \ \hat{m} \ \hat{m} \ \widehat{m} \ \widehat{m} \ \overline{m} \ \hat{m} \ \hat{m} \ \widehat{m} \ \widehat{m}$$

The rule for type design is simple. Any reasonable fusion that respects a font's style should be acceptable; a ring in the lowercase y, for instance, is not obligatory. It may look good in a serif font. Here we have Times plain on the left, and Baskerville plan and italic in the centre and on the right.

yỹ yỹ yỹ

But the original hand in the medieval Cornish poem Pascon agan Arluth had a different ductus for y, and so the OVERCURL had no loop. The loop itself is of no consequence. The OVERCURL simply has to attach at a convenient point, and swing over towards the left.

ɤ ɤ̃

Here is an example of the attested forms found in the Cornish Pascon agan Arluth. These are the letters a e i m n r s t u and y:

ᴀ ꜳ ℴ ℴ̃ , Ꝺ m m̃ n n̅

γ γ̃ ꝛ ꝛ̃ t ꝑ n̅ ɤ ɤ̃

Clearly the base characters can differ considerably in shape from modern typeset forms, which is why the advice about how to attach the OVERCURL is properly a recommendation to attach it "conveniently".

Here is the Latin alphabet where each of the letters has been followed by the COMBINING OVERCURL (attested letters coloured in blue):

a b c d e f g h i j k l m n

o p q r s t u v w x y z

Opentype features may not always be available. The standalone combining mark U+1DFA should be designed in fonts so that it fits reasonably and legibly over the base characters where Opentype features cannot be activated. This preserves legibility; it is not perfect, but it does the job. A good font should do better. Here is an example:

a b c d e f g h i j k l m n

o p q r s t u v w x y z

	1AB	1AC	1AD	1AE	1AF
0	◌ 1AB0	◌ 1AC0			
1	◌ 1AB1	◌ 1AC1			
2	◌ 1AB2	◌ 1AC2			
3	◌ 1AB3	◌ 1AC3			
4	◌ 1AB4	◌ 1AC4			
5	◌ 1AB5	◌ 1AC5			
6	◌ 1AB6	◌ 1AC6			
7	◌ 1AB7	◌ 1AC7			
8	◌ 1AB8	◌ 1AC8			
9	◌ 1AB9	◌ 1AC9			
A	◌ 1ABA	◌ 1ACA			
B	◌ 1ABB	◌ 1ACB			
C	◌ 1ABC	◌ 1ACC			
D	◌ 1ABD	◌ 1ACD			
E	◌ 1ABE	◌ 1ACE			
F	◌ 1ABF	◌ 1ACF			

This block contains many specialized combining marks, often derived from European dialectology transcriptional conventions.

Used for German dialectology

1AB0	̈̂	COMBINING DOUBLED CIRCUMFLEX ACCENT
1AB1	̈̊	COMBINING DIAERESIS-RING
1AB2	̊̊	COMBINING INFINITY
1AB3	̣↓	COMBINING DOWNWARDS ARROW
1AB4	̣⋮	COMBINING TRIPLE DOT
1AB5	̲x	COMBINING X-X BELOW
1AB6	̰	COMBINING WIGGLY LINE BELOW
1AB7	̹	COMBINING OPEN MARK BELOW

→ 0328 ̨ combining ogonek

→ 0345 ͅ combining greek ypogegrammeni

1AB8	̹̹	COMBINING DOUBLE OPEN MARK BELOW
1AB9	ϱ	COMBINING LIGHT CENTRALIZATION STROKE BELOW
1ABA	ϱ	COMBINING STRONG CENTRALIZATION STROKE BELOW

Marks surrounding other diacritics or letters

These marks are often intended to indicate partial or uncertain applicability of another diacritic.

1ABB	̆̆	COMBINING PARENTHESES ABOVE
		• intended to surround a diacritic above
1ABC	̆̆	COMBINING DOUBLE PARENTHESES ABOVE
1ABD	̖̗	COMBINING PARENTHESES BELOW
		• intended to surround a diacritic below
1ABE	⟨◌⟩	COMBINING PARENTHESES OVERLAY
		• intended to surround a base letter
		• exact placement is font dependent

Used in Scots dialectology

1ABF	̫w	COMBINING LATIN SMALL LETTER W BELOW
		• indicates voiced labialization

→ 02B7 ʷ modifier letter small w

1AC0	̫ʍ	COMBINING LATIN SMALL LETTER TURNED W BELOW
		• indicates voiceless labialization

Marks next to or surrounding other diacritics

The left or right parentheses marks are used in extended IPA in combination with voicing or devoicing diacritics to indicate initial or final partial applicability.

1AC1	͗◌	COMBINING LEFT PARENTHESIS ABOVE LEFT
1AC2	◌͗	COMBINING RIGHT PARENTHESIS ABOVE RIGHT
1AC3	͐◌	COMBINING LEFT PARENTHESIS BELOW LEFT
1AC4	◌͑	COMBINING RIGHT PARENTHESIS BELOW RIGHT
1AC5	̣̣	COMBINING SQUARE BRACKETS ABOVE

Phonetic sign

1AC6	̣#	COMBINING NUMBER SIGN
		= superscript octothorp
		• used extensively in J.P. Harrington€™s transcriptional notation

Used in extended iPA

1AC7	̑̑	COMBINING INVERTED DOUBLE ARCH ABOVE
		• used over letters with a descender

→ 032B ̫ combining inverted double arch below

→ 1DF1 ̫ combining latin small letter w

1AC8	̟+	COMBINING PLUS SIGN ABOVE

→ 031F ̟ combining plus sign below

1AC9	̟++	COMBINING DOUBLE PLUS SIGN ABOVE
1ACA	̟++	COMBINING DOUBLE PLUS SIGN BELOW

Used in the Middle English Ormulum

1ACB	̋̋	COMBINING TRIPLE ACUTE ACCENT

→ 0308 ̋ combining double acute accent

1ACC	ᵹ̇	COMBINING LATIN SMALL LETTER INSULAR G

→ 1D79 ᵹ latin small letter insular g

1ACD	ꞃ̇	COMBINING LATIN SMALL LETTER INSULAR R

→ A783 ꞃ latin small letter insular r

1ACE	ꞇ̇	COMBINING LATIN SMALL LETTER INSULAR T

→ A787 ꞇ latin small letter insular t

Used throughout medieval Britain

1ACF	ᷚ	COMBINING OVERCURL
		• used in medieval Cornish, English, and Scots
		• used either as an abbreviation or an otiose mark

→ 0305 ̅ combining overline

→ 031F ̑ combining inverted breve

A. Administrative

1. Title
Revised proposal to add the combining overcurl to the UCS
2. Requester's name
Michael Everson, Nicholas Williams, †Alan M. Kent
3. Requester type (Member body/Liaison/Individual contribution)
Individual contribution.
4. Submission date
2023-06-16
5. Requester's reference (if applicable)
6. Choose one of the following:
6a. This is a complete proposal
Yes.
6b. More information will be provided later
No.

B. Technical – General

1. Choose one of the following:
1a. This proposal is for a new script (set of characters)
No.
1b. Proposed name of script
1c. The proposal is for addition of character(s) to an existing block
Yes
1d. Name of the existing block
Combining Diacritical Marks Extended
2. Number of characters in proposal
1.
3. Proposed category (A-Contemporary; 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)
Category A.
4a. Is a repertoire including character names provided?
Yes.
4b. If YES, are the names in accordance with the "character naming guidelines" in Annex L of P&P document?
Yes.
4c. Are the character shapes attached in a legible form suitable for review?
Yes.
5a. Who will provide the appropriate computerized font (ordered preference: True Type, or PostScript format) for publishing the standard?
Michael Everson.
5b. If available now, identify source(s) for the font (include address, e-mail, ftp-site, etc.) and indicate the tools used:
Michael Everson, Fontographer.
6a. Are references (to other character sets, dictionaries, descriptive texts etc.) provided?
Yes.
6b. Are published examples of use (such as samples from newspapers, magazines, or other sources) of proposed characters attached?
Yes.
7. 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.
8. 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.
See above.

C. Technical – Justification

1. Has this proposal for addition of character(s) been submitted before? If YES, explain.
No.
2a. 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.
2b. If YES, with whom?
The authors are members of the user community, preparing new editions of the complete Cornish corpus with palaeographic readings.
2c. 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?
Medievalists, Celticists, Germanicists, and other scholars.
4a. The context of use for the proposed characters (type of use; common or rare)
Used historically and in modern editions.
4b. Reference
5a. Are the proposed characters in current use by the user community?
Yes.

5b. If YES, where?

Scholarly publications.

6a. After giving due considerations to the principles in the P&P document must the proposed characters be entirely in the BMP?

Yes.

6b. If YES, is a rationale provided?

Yes.

6c. If YES, reference

Accordance with the Roadmap. Keep with other punctuation and combining characters.

7. Should the proposed characters be kept together in a contiguous range (rather than being scattered)?

No.

8a. Can any of the proposed characters be considered a presentation form of an existing character or character sequence?

No.

8b. If YES, is a rationale for its inclusion provided?

8c. If YES, reference

9a. Can any of the proposed characters be encoded using a composed character sequence of either existing characters or other proposed characters?

No.

9b. If YES, is a rationale for its inclusion provided?

9c. If YES, reference

10a. Can any of the proposed character(s) be considered to be similar (in appearance or function) to an existing character?

No.

10b. If YES, is a rationale for its inclusion provided?

10c. If YES, reference

11a. Does the proposal include use of combining characters and/or use of composite sequences (see clauses 4.12 and 4.14 in ISO/IEC 10646-1: 2000)?

Yes.

11b. If YES, is a rationale for such use provided?

Yes.

11c. If YES, reference

It is a proposal for a combining character.

11d. Is a list of composite sequences and their corresponding glyph images (graphic symbols) provided?

Yes.

11e. If YES, reference

12a. Does the proposal contain characters with any special properties such as control function or similar semantics?

No.

12b. If YES, describe in detail (include attachment if necessary)

13a. Does the proposal contain any Ideographic compatibility character(s)?

No.

13b. If YES, is the equivalent corresponding unified ideographic character(s) identified?