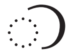
ISO/IEC JTC1/SC2/WG2 N4907
L2/17-358
2017-10-17

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 add one combining character for medieval Cornish to the UCS**
Source: **Michael Everson, Nicholas Williams, Alan M. Kent**
Status: **Liaison Contribution**
Action: **For consideration by JTC1/SC2/WG2 and UTC**
Date: **2017-10-17**
Replaces: **N4902 (L2/17-342)**

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

 1DFA COMBINING OVERCURL

- used in medieval Cornish, English, Latin
- fuses typographically with at least a, e, i, m, n, r, t, u, y

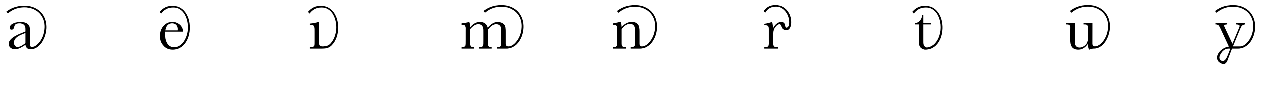
1. COMBINING OVERCURL. Medieval handwriting in Cornish uses a variety of mechanisms for representing an abbreviation of *m* and *n*. These mechanisms occur alongside one another and can all be represented by characters in the UCS (U+0306 COMBINING OVERLINE, U+0311 COMBINING INVERTED BREVE, and U+0352 COMBINING FERMATA).

m̄ m̑ m͒ n̄ n̑ n͒

These are read *mm* and *nn*; *mn* and *nm* are in principle possible but would be extremely rare. These marks also appear on vowels:

ā ȃ a͒ ȳ y̑ y͒

Which can be read *am* and *ym* or *an* and *yn*. But one kind of abbreviation, which seems to be quite productive, is also found:



These forms are polyvalent. This mark may simply be a meaningless swash form, or it may be an abbreviation. Thus the readings for these may be *a*, *e*, *i*, *m*, *n*, *r*, *t*, *u*, *y*, or *am*, *em*, *im*, *mm*, *nm*, *rm*, *tm*, *um*, *ym*, or *an*, *en*, *in*, *mn*, *nn*, *rn*, *tn* (the reading *nt* is attested), *un*, *yn*. There is no way of telling which without interpretating what is in the text. But a palaographic representation of the text is impossible for these last forms without the combining character proposed here.

0.1. On palaeographic readings and character encoding. Eleven years ago in N3027 (L2/06-027), arguments were presented about the nature of palaeographic textual representations. Medievalist editors who make use of such representations do not attempt to achieve *calligraphic* representations of text, which are merely decorative. Rather, they attempt to achieve structural representations of what the scribes have written, in a modern, interchangeable format. In recent times, now that specialists have been working with the characters encoded in the Latin Extended-D and Supplemental Punctuation blocks, gaps in the encoding have been identified. The COMBINING OVERCURL proposed here is one such character.

1.1. Polyvalent signs already encoded. We already have a similar ambiguous situation with U+035B COMBINING ZIGZAG ABOVE, which was encoded as an abbreviation representing *er* and *re*, and with U+0306 COMBINING OVERLINE, as used in medieval Cornish and English. Here are some examples of the former:

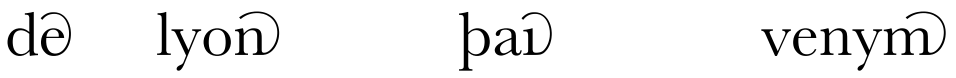
war͛ dr͛ der͛ man͛e

These are forms attested in the manuscripts: the first two words are Middle Cornish *war* 'on' (where the COMBINING ZIGZAG is otiose) and *dre* 'through' (where it represents -*e*), and the second two are Middle English *dere* 'dear' (where the ZIGZAG represents -*e*), *manere* 'manner' (where it represents -*er*-).

Here are some examples of the combining overline used meaningfully and "decoratively".

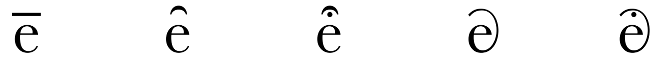
de̅ flogh̅ dragu̅ mygh̅t

Here Cornish *den* 'person' and English *dragun* 'dragon' use the combining overline meaningfully while *flogh* 'girl' and English *myght* 'might' use it without significance. In a palaeographic reading, the scribe's penstrokes are what are being represented. The following examples from Cornish and English materials also occur:



Here Cornish *den* 'person' and English *þaim* 'them' use the combining overline meaningfully while *lyon* 'lion' and English *venym* 'venom' use it without significance.

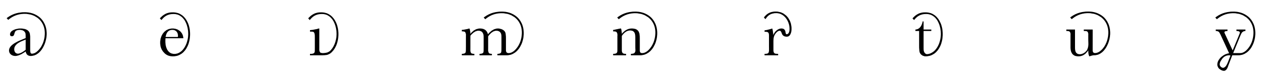
This is not a question of calligraphy, but of palaeography. In the Cornish examples shown below, none can be said to be particularly "calligraphic". The manuscripts we have are examples of *copying* and *writing*, a far cry from the calligraphy Book of Kells. (In *Pascon agan Arluth*, the only genuine attempts at "calligraphy" can be seen in the flowers attached to two of the words in words here and there throughout the text; see Figure 1.) The marks which the scribes make are important to linguistic and orthographic investigations, however—which is why palaeographic editions are important. The scribe writes all of these, and only one of them can't yet be represented in the UCS.



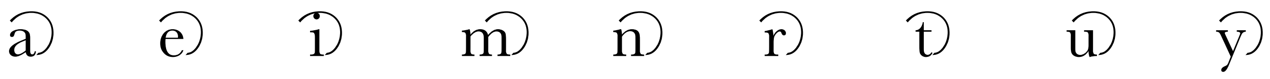
As far as the last one is concerned, the dot there is quite rare rare and can easily be represented (as here) with the existing COMBINING DOT ABOVE. (This does not apply to INVERTED BREVE and FERMATA because those are already encoded.) We do not believe that encoding a *COMBINING OVERCURL WITH DOT would be warranted. We also do not believe that a unification with INVERTED BREVE or FERMATA is possible, since the scribes never write those "meaninglessly". In addition, the ductus of ◌̑ and ◌͒ is toward the right, while the ductus of ◌ is toward the left.

The encoded COMBINING ZIGZAG and COMBINING OVERLINE can both be used to correctly represent these texts, whether the reading of those marks is meaningful or not. Similarly, when we have a word ending in -*e*, we do not know whether it is -*en* or -*em* or -*e*—but with the new COMBINING OVERCURL it is possible to represent the text accurately regardless of the meaning.

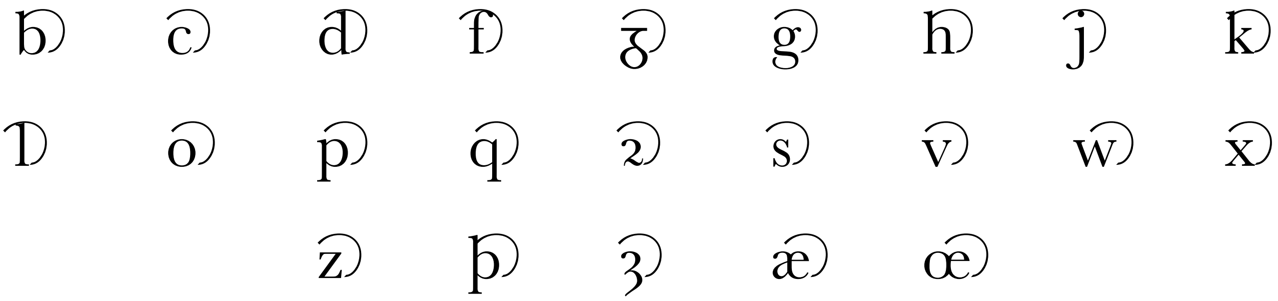
2. Glyph presentation. It is usual in the UCS that diacritics that fuse typographically with base characters are encoded atomically, but since the COMBINING OVERCURL is a productive abbreviation character not used in a standard orthography, we consider it best to encode it as a single combining character. An informative note listing the characters which have been observed making use of it is recommended for the names list. In the event that a font does not fuse the diacritic with its base character, the representation can still be considered legible. A font producer who knows which characters to support can provide optimized glyphs quite easily:



A font producer who does not know which characters to support can still support the overcurl, legibly even if imperfectly:



In fact this works adequately for the rest of the alphabet as well:



Are these the best possible glyphs? No. But they are legible. The alternative to encoding a single combining character as proposed would be to encode the identified characters which conjoin with the OVERCURL atomically, each one as it is discovered. But since the OVERCURL is productive, this would burden the standard and the standardization committees with regular petitions to add characters, with regular arguments over whether they were sufficiently attested, and so on and on. Note that as this document was prepared, the first version showed the diacritic with the seven letters *a, e, m, n, r, u, y*. The next version added the letter *i* (from a Middle English example), and now in this version the letter *t* has been added (from Cornish). This script feature is productive, and others are sure to be found.

There is also some advantage in terms of simplicity for the researcher working on a digital text to be able to search just for instances of the combining character ◌᷺ itself, just as can be done with ◌̄, ◌̂, and ◌̂̇.

It is also the case that many of the existing medievalist characters benefit from expert ligation in proper medievalist fonts. The character U+A7CD LATIN SMALL LETTER IS ꟍ should *always* ideally combine typographically with letters which precede it:

ꟍ cꟍ dꟍ fꟍ gꟍ kꟍ rꟍ ꝛꟍ ſtꟍ

These too are legible when proper ligation is not supported:

cꟍ dꟍ fꟍ gꟍ kꟍ rꟍ ꝛꟍ ſtꟍ

The OVERCURL 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.

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

1DFA: CM (Combining Mark)

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

```
1DFA;COMBINING overcurl;Mc;210;L;;;;;N;;;;;
```

Figures

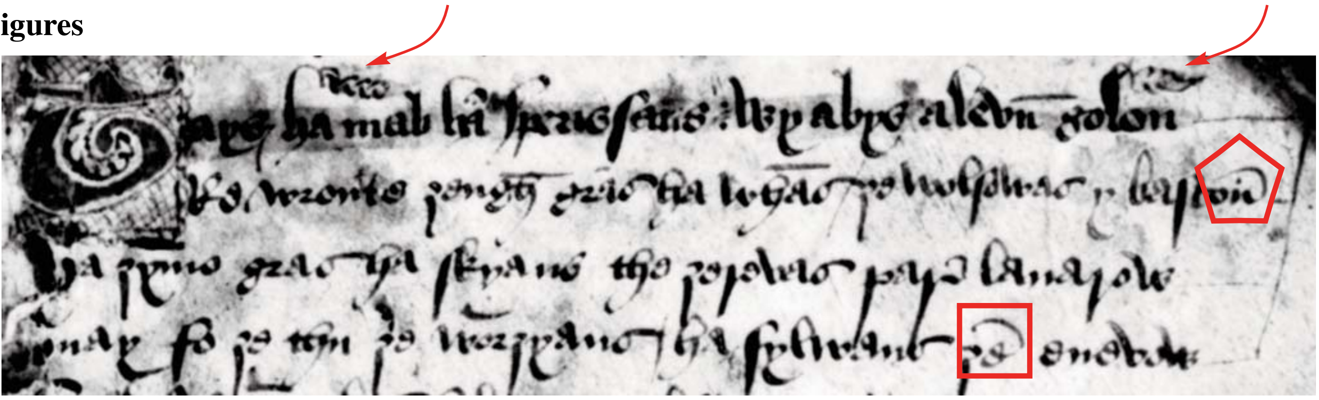


Figure 1. The first line of *Pascon agan Arluth* 'The Passion of our Lord' (BL MS Harley 1782B, fol. 1r), showing COMBINING OVERCURL on *n* in *bascon* 'passion' (where it, along with a dot, is otiose) and *e* ni *ȝen* (where it is meaningful). Arrows point to the "calligraphic" flowers used by the scribe to decorate the first line of the text. The text reads, in palaeographic presentation, normalized text, and translation (from the forthcoming edition in Corpus Textorum Cornicorum):

Tays ha mab hâ ſpeꝛis ſans ⸗ ꝡy abys a levn̄ golon
Re wꝛonte ȝeugh̄ gꝛas ha whās / ȝe wolſowas y baſcoṅ᷑
Ha ȝymo gꝛas ha skyans the ȝerevas parʾ lauaroꝡ
may ſo ȝe thu ȝe woꝛꝛyans / ha ſylwans ȝe᷑ enevoꝡ

Tas ha Mab ha'n Spyrys Sans,
why a bÿs a leun-golon,
re wrauntyo dhywgh grâss ha whans
dhe wolsowes y <u>*Bassyon;*</u>
ha dhymmo grâss ha skians
dhe dherivas pàr dell wòn,
may fo dhe Dhuw dh'y wordhyans
ha selwans <u>*dhe'n*</u> *Gristenyon.*

May Father, Son and the Holy Spirit—
you who pray from the bottom of your heart—
grant you grace and yearning
to listen to his Passion,
and to me grace and wisdom
to recount as well as I can,
that it may be for the glory of God
and the salvation of Christians.

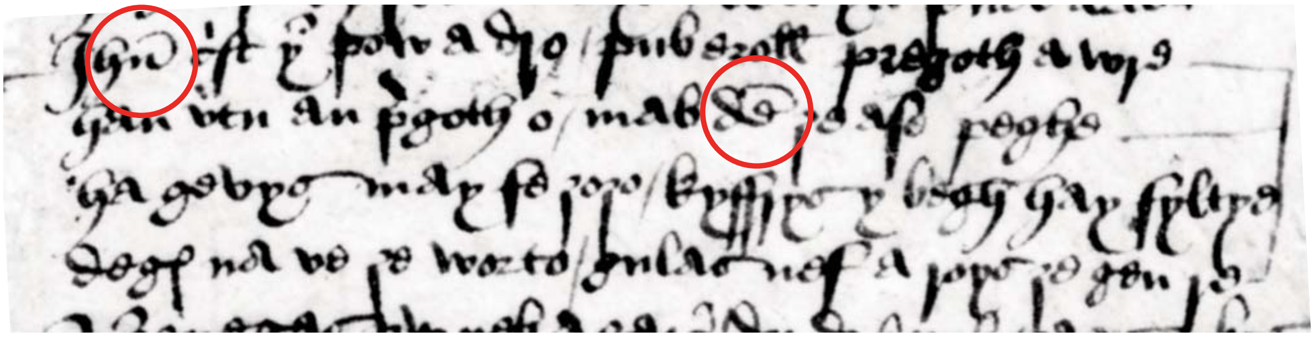


Figure 2. Verse 23 of *Pascon agan Arluth* ‘The Passion of our Lord’ (BL MS Harley 1782B, fol. 3r), showing COMBINING OVERCURL on *u* and *e*. On *u* the swash is not meaningful (*Jesu* is often written with a “meaningless” mark; *Iħu*, *Ihū*, *Ihû*, and *Ihu᷎* all occur in this manuscript) but it is meaningful on the *e* (where it means *n*). The text reads, in palaeographic presentation, normalized text, and translation:

Ihu᷎ cͥſt ŷ pow a dro / pub eꝛoll̄ pꝛegoth a wre
han vͥtu an p᷑goth o / mab de᷎ ʒe aſe peghe
ha gevys may ſe ʒoʒo / kyffrys y beg̅h̅ hay fyltye
degſ na ve ʒe woꝛto / gulas nef a roys ʒe gen re

Jesu Crist i’n pow adro
pùb eur oll pregoth a wre;
ha vertu an pregoth o
mab den dhe asa peha,
ha gevys may fe dhodho
kefrÿs y begh ha’y fylta,
degys na ve dhyworto
gwlas nev ha rës dhe gen re.

Jesus Christ all around the country
used always to preach;
and the essence of the preaching was
that man should give up sinning,
so that there should be forgiven him
both his sin and his corruption,
to the end that the kingdom of heaven should not
be taken from him and given to others.

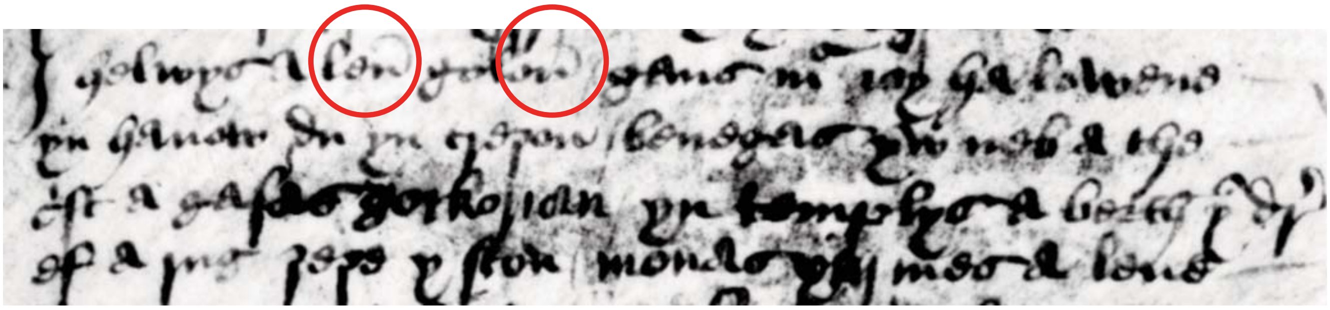


Figure 3. Verse 30 of *Pascon agan Arluth* ‘The Passion of our Lord’ (BL MS Harley 1782B, fol. 3v), showing COMBINING OVERCURL on *u* and *n*. On *n* in *golon* ‘heart’ the overcurl is not meaningful but it is meaningful on the *u* in *leun* ‘full’. The text reads, in palaeographic presentation, normalized text, and translation:

I helwys a leu͗ golon͗ / gans m̃ ioy ha lowene
yn hanow du yn treȝon / benegas yw neb a the
cͥﬅ a gafas goꝛkorian / yn templys a beꝛt̄h̄ ŷ drˢ
ef a rug ȝeȝe y ſcon / monas yn mes a lene

Y helwys a <u>leun</u>-<u>golon</u>
gans meur joy ha lowena
“In hanow Duw intredhon
benegys yw neb a dheu!”
Crist a gafas gwycoryon
i’n templys aberth i’n dre.
Ev a wrug dhedha yn scon
mones in mes alena.

People called out from the bottom of their heart
with great joy and gladness,
“In the name of God among us
is blessed he who comes!”
Christ found traders
in the temples within the city.
Quickly he caused them
to depart away from there

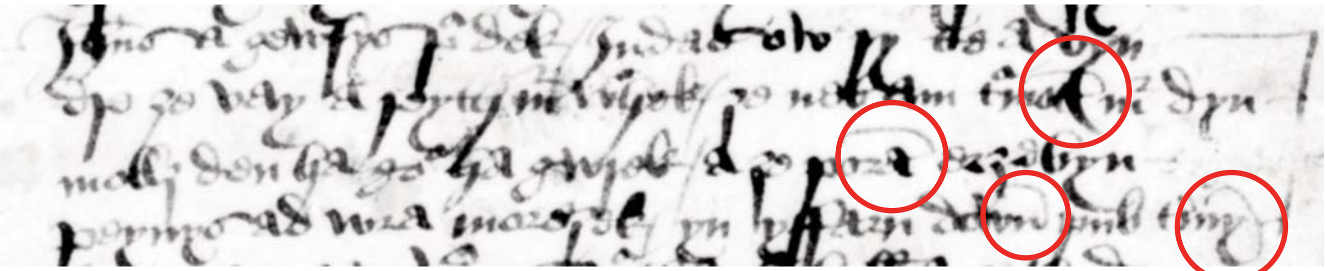


Figure 4. Verse 66 of *Pascon agan Arluth* ‘The Passion of our Lord’ (BL MS Harley 1782B, fol. 6v), showing COMBINING OVERCURL on *t*, *a*, *n* and *y*. On *n* the overcurl is decorative but it is meaningful on the other two (it means *n* on both). The text reads, in palaeographic presentation, normalized text, and translation:

Ih̄us a gewſys p̃ dek / Iudas ow ry te a vyn
dre ȝe vay a reyth mᷣ whek / ȝe neb am t̃moꝺ mᷣ dyn
mollȝ den ha gõ ha gwrek / a ȝe poꝛa᷎ eꝛȝebyn
peynys ad wꝛa moꝛeȝek / yn yffaꝛn down᷎ pub t᷎my᷎

Jesus a gewsys pòr deg,
“Jûdas, ow ry te a vynn,
dre dha vay a reth mar wheg
dhe neb a’m *<u>torment</u>* *pòr dynn.*
Mollath den, ha gour ha gwreg
a dheu *<u>poran</u>* *er dha bynn.*
Painys a’th wra morethek
in iffarn *<u>down</u>* *pùb* *<u>termyn</u>.”*

Jesus spoke very fairly,
“Judas, you will betray me,
by your kiss, which you give me so sweetly,
to him who will torment me very sharply.
The curse of men, both husband and wife,
will come exactly against you,
Pains will render you wretched
in deep hell forever.”

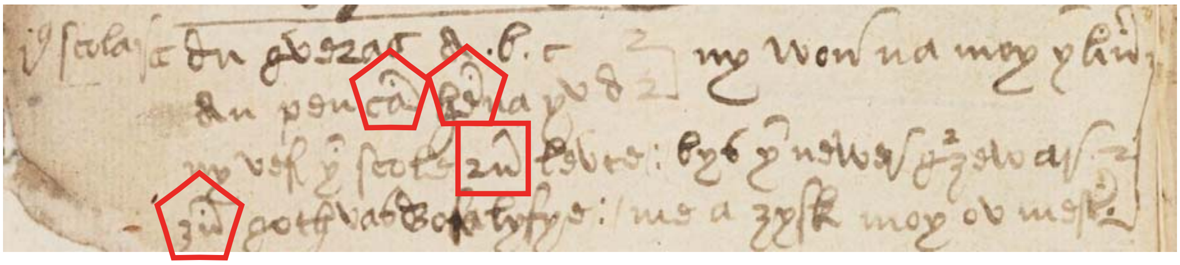


Figure 5. Text from *Beunans Meriasek* (Peniarth MS 105B, fol. 3v), showing an example of the COMBINING OVERCURL in this here (in ꝛu᷎, where it means *m*), alongside the COMBINING INVERTED BREVE (in y̑) and FERMATA (in y͒). There are also three examples of a overcurl with a dot on *a* and *e* (where it means *n*) and *u* (where it means *m*), but in Cornish texts these are rare and it is recommended to use the sequence base + COMBNINING OVERCURL + COMBINING DOT ABOVE; no *COMBINING OVERCURL WITH DOT is proposed. The text reads:

iꝰ ſcolar
du gveꝛas A · b · c / an pen ca̐ he̐na yv d
ny won na moy y̑ liu͛
ny vef y͒ ſcole ꝛu᷎ levte ⁝ bys y̑ newer gᷣʒewar
ʒu̐ gothvas woſa lyfye ⁝ me a ʒyſk moy ov meſt͛

PRIMUS SCOLARIS
Dew gweres A B C, -
an pen can, henna yw D.
Ny won na moy y'n lyver.
Ny vuef yn scol, re'm leouta,
bys yn nyhewer gordhewer.
Dhe'm godhvos, wosa lyvya
me a dhysk moy, ow mester.

FIRST SCHOLAR
God keep A, B, C,
The end of the song, that is D.
I know no more in the book.
I was not at school, by my loyalty,
Until late (?) yesterday evening.
To my knowledge, after dining
I will learn more, my master.

A. Administrative

1. Title
Proposal to add one combining character for medieval Cornish 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
2017-10-17
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 Supplement
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, 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?

Yes.

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

Discussion is given regarding the interaction of the COMBINING OVERCURL and the existing COMBINING DOT ABOVE.

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?