L2/15-014

Proposal to Change Annotations on Some Cyrillic Characters

Aleksandr Andreev[*] Yuri Shardt Nikita Simmons

Ponomar Project
Slavonic Computing Initiative

1 Introduction

Church Slavic (also known as Church Slavonic or Old Slavonic) is a historical literary language of the Slavs. Presently it is used as a liturgical language by the Russian Orthodox Church, various other local Orthodox Churches, and Byzantine-rite Catholic and Old Ritualist communities. As considered in this document, Church Slavic is written in the Cyrillic script. The encoding of Cyrillic characters required for writing Church Slavic has had a long history and a large number of proposals by various authors representing different institutions have been considered by the Unicode Technical Committee. As a result of the piecemeal approach, several inconsistencies in the encoding model have emerged. Understanding that the Unicode Stability Policy prevents us from making changes to the existing implementation, we do, however, request the UTC to update some annotations in order to clarify certain aspects of the encoding model. Our proposed changes to Annotations and the Documentation are presented in this document.

2 U+0479 Digraph Uk

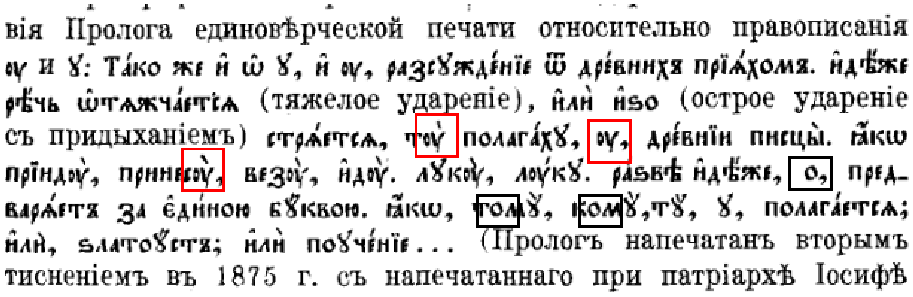
This character has been discussed by the UTC before (see Everson et al. (2006) and Everson et al. (2007)). The authors of Everson et al. (2006, p. 3) state that "What we have regarding CYRILLIC LETTER UK is quite a mess." Since then, the character U+A64B MONOGRAPH UK has been encoded and the annotations on U+0479 and U+0478 have been changed in an attempt to fix the problems with these characters. However, recently a decision was made to encode also U+1C82 CYRILLIC SMALL LETTER NARROW O (see Andreev et al. (2014)) in the Cyrillic Extended-C block. In fact, the digraph character Onik (which is the character that was supposed to be encoded at U+0479), most often appears using the Narrow O as the first character (ᲂу), and not U+043E CYRILLIC SMALL LETTER O (see Figure 1). In the past, the narrow O character was not available in Unicode, but it has now been accepted for encoding in a future version of the standard. What we have now is three different potential spellings for the Onik digraph, which have the representations given in Table 1.

[*] Corresponding author: aleksandr.andreev@gmail.com.

Table 1: Encoding of the Onik digraph

U+1C82 U+0443	ᲂу
U+043E U+0443	оу
U+0479	ambiguous

Figure 1: Orthography of the digraph onik in Church Slavic. Note the usage of Narrow O (encoded at U+1C82) in writing the digraph onik (boxed in red) vs. the usage of Cyrillic Letter O (encoded at U+043E) in writing the letter O (boxed in black). Source: Gamanovitch (1991, p. 13)



віѧ Пролога единовѣрческой печати относительно правописанія
ᲂу и ꙋ: Та́ко же и҆ ѡ҆ ꙋ, и҆ ᲂу, разсꙋжде́ніе ѿ дре́внихъ пріѧ́хомъ. и҆дѣ́же
рѣ́чь ѿтѧжча́етсѧ (тяжелое удареніе), и҆лѝ и҆ѕо (острое удареніе
съ придыханіемъ) стра́етсѧ, то̀ᲂу полага́хꙋ, ᲂу, дре́вніи пи́сцы. ꙗ҆́кѡ
прі́идоу̀, принесоу̀, везоу̀, и҆доу̀. лꙋко́у, лоу́кꙋ. развѣ̀ и҆дѣ́же, о, пред-
вари́етъ за є҆ди́ною бꙋ́квою. ꙗ҆́кѡ, томꙋ̀, комꙋ̀, тꙋ̀, ꙋ, полага́етсѧ;
и҆лѝ, ѕлатоꙋ́стъ; и҆лѝ поꙋче́ніе... (Прологъ напечатанъ вторымъ
тисненіемъ въ 1875 г. съ напечатаннаго при патріархѣ Іосифѣ

Clearly, the character U+0479 has been problematic all along, though the UTC has hesitated to deprecate it completely. Now with the encoding of U+1C82, the problematic nature of this character is compounded. Thus, we propose that the annotation on U+0479 be changed to:

```
this character has ambiguous glyph representation and should not be used
for "digraph onik" use U+1C82 U+0443 or U+043E U+0443
for "monograph uk" use U+A64B
```

Furthermore, the annotation on U+0478, "may be rendered as either monograph or digraph form" seems to reflect the state of the standard before the UTC acted on Everson et al. (2007), where it was proposed to clarify the appearance of U+0478. In light of the past discussion about this character and the proposed changes to U+0479, we propose that the annotations on U+0478 be changed to:

```
this character should not be used
for "digraph onik" use U+041E U+0443  or U+041E U+0423
for "monograph uk" use U+A64A
```

2.1 U+047C and U+047D

These characters were originally encoded in the Unicode standard with an erroneous name and representation. After the UTC ruling on Everson et al. (2006), the representation was corrected and an annotation was added to U+047C, reading "despite its name, this character does not have a titlo, nor is it composed of an omega plus a diacritic". However, no annotation was added to the lowercase form U+047D.

The character that is encoded here is a ligature of the Cyrillic broad (or wide) Omega (encoded at U+A64C and U+A64D) and the 'great apostrof', a stylized diacritical mark consisting of the soft breathing (encoded at U+0486) and the Cyrillic kamora (encoded at U+0311). The broad Omega (U+A64D) can occur by itself, without this diacritical mark, in pre-1700 printed Church Slavic books, though not in modern liturgical texts. Functionally, the character with the diacritical mark is analogous to the Greek character ὦ, which also consists of an Omega, a soft breathing mark and a Perispomene. Both the Greek and Church Slavic characters have identical functions: to record the exclamation 'Oh!' Since U+047C and U+047D were encoded without a canonical decomposition, though they are linguistically decomposable, they should not be decomposed to avoid an encoding ambiguity. However, in our opinion, the annotation as written does not make this clear. We propose to change the annotation on U+047C to read:

```
Alias name: Cyrillic "beautiful omega"
Used for exclamation Oh! in Church Slavic
Despite its name, this character does not have a titlo. It should not be
   decomposed into an omega plus diacritics.
See also: U+A64C CYRILLIC CAPITAL LETTER BROAD OMEGA
```

We further propose to add an analogous annotation for U+047D, reading:

```
Alias name: Cyrillic "beautiful omega"
Used for exclamation Oh! in Church Slavic
Despite its name, this character does not have a titlo. It should not be
   decomposed into an omega plus diacritics.
See also: U+A64D CYRILLIC SMALL LETTER BROAD OMEGA
```

3 U+0484: Combining Palatalization

This character is used in ancient manuscripts and in academic work to indicate that a vowel is softened, a phenomenon called 'palatalization' (Karsky, 1979, p. 230). However, many users incorrectly use this character to encode the Cyrillic Kamora (circumflex accent), which in Unicode has been encoded as U+0311 COMBINING INVERTED BREVE. Furthermore, this character has a cross-reference to U+033E COMBINING VERTICAL TILDE. However, U+033E is not used for palatalization, but is rather used to indicate an omitted yer (soft sign and, in later sources, hard sign) and, very rarely in some manuscripts, an omitted [j] (Karsky, 1979, pp. 228f). The cross-reference probably arises because of a confusion between the function of U+033E and the function of U+02BC MODIFIER LETTER APOSTROPHE, which is used in some manuscripts and in academic publications to indicate palatalization in Cyrillic (see Golyshenko (1987, pp. 52ff) and Yelkina (1960, p. 26)). In any case, the cross-reference is providing further misleading information about the purpose of this character. We propose that the cross-reference to U+033E be removed and that the following annotations be made to U+0484:

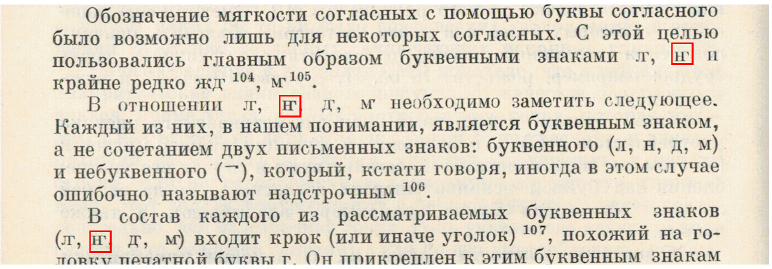
```
Not used for kamora.
See also: U+0311 COMBINING INVERTED BREVE
See also: U+02BC MODIFIER LETTER APOSTROPHE
```

In addition, because this character is used both in Cyrillic and Glagolitic scripts to indicate palatalization (Golyshenko, 1987, p. 44), we propose that the Script property on U+0484 be changed to `Inherited`.

4 U+04A4 and U+04A5 Cyrillic Ligature En Ghe

The Unicode standard includes a number of characters for writing the palatalized (soft) consonants used in some early manuscripts and in academic work: U+A663 CYRILLIC SMALL

Figure 2: Examples of the Cyrillic Letter Soft En (boxed in red) used in academic literature. Source: Golyshenko (1987, p. 42)

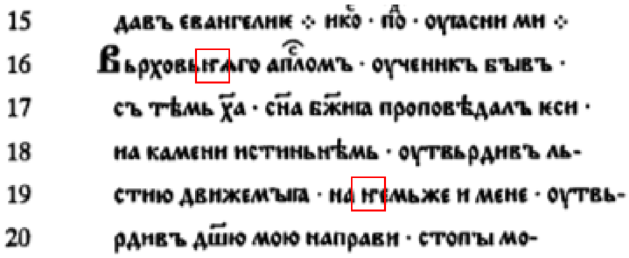


Обозначение мягкости согласных с помощью буквы согласного было возможно лишь для некоторых согласных. С этой целью пользовались главным образом буквенными знаками лг, нг и крайне редко жд[104], мг[105].

В отношении лг, нг, дг, мг необходимо заметить следующее. Каждый из них, в нашем понимании, является буквенным знаком, а не сочетанием двух письменных знаков: буквенного (л, н, д, м) и небуквенного (¬), который, кстати говоря, иногда в этом случае ошибочно называют надстрочным[106].

В состав каждого из рассматриваемых буквенных знаков (лг, нг, дг, мг) входит крюк (или иначе уголок)[107], похожий на головку печатной буквы г. Он прикреплен к этим буквенным знакам

Figure 3: Examples of the Cyrillic Letter Soft En (boxed in red) from Church Slavic texts. Source: Typografsky Ustav, a manuscript written c. 1300. Source: Uspensky (2006, p. 152).



давъ ѥвангелиѥ ⁘ ико · по · оуꙗсни ми ⁘
Вьрховьҥаго ап҃ломъ · оученикъ бывъ ·
съ тѣмь х҃а · с҃на б҃жиꙗ проповѣдалъ иси ·
на камени истиньнѣмь · оутвьрдивъ ль-
стию движемꙑꙗ · на ҥемьже и мене · оутвь-
рдивъ д҃шю мою направи · стопꙑ мо-

LETTER SOFT DE (and its capitalized analog U+A662), U+A665 CYRILLIC SMALL LETTER SOFT EL (and its capitalized analog U+A664), and U+A667 CYRILLIC SMALL LETTER SOFT EM (and its capitalized analog U+A666). The Cyrillic writing system uses one additional soft character: Cyrillic Small Letter Soft En (see examples from academic literature in Figure 2 and examples from reproductions of ancient sources in Figure 3). Instead of encoding this letter as a standalone character, we propose to merge it with U+04A5 (U+04A4 for the capitalized form), since the Ligature En Ghe used in Altay, Mari and Yakut is almost identical visually to the Soft En. Thus, we propose that the following annotation be added to U+04A4:

```
Also used for Cyrillic Capital Letter Soft En
```

and the following annotation be added to U+04A5:

```
Also used for Cyrillic Small Letter Soft En
```

5 U+2DF5 Combining Cyrillic Letter Es-Te

The authors of the present document also submitted to the UTC a proposal to clarify the encoding of Cyrillic composite combining characters (see L2/15-002). Thus, we propose that the appropriate annotations be made to this character. For example:

```
This character should not be used.
Correct spelling is: U+2DED U+200D U+2DEE
```

6 U+0483 Combining Cyrillic Titlo

One finds contradictory information in the Unicode documentation regarding the usage of the characters U+0483 COMBINING CYRILLIC TITLO, U+A66F COMBINING CYRILLIC VZMET and U+0487 COMBINING CYRILLIC POKRYTIE. For example, the Unicode Standard version 7.0 (Section 7.4) reads:

> The [Cyrillic Extended-A] block contains a set of superscripted (written above), or titlo letters, used in manuscript Old Church Slavonic texts, usually to indicate abbreviations of words in the text. These may occur with or without the generic titlo character, U+0483 COMBINING CYRILLIC TITLO, or with U+A66F COMBINING CYRILLIC VZMET.

On the other hand, the annotation on U+0483 reads `not used with letter titlos`.

The problem with the documentation as written is that it is attempting two describe two different practices dating to two different eras. Contrary to the statement in Section 7.4 of the documentation, superscripted (titlo) letters are used not only in manuscript Old Church Slavonic texts, but also in modern Church Slavic. As used in modern Church Slavic, the superscripted letters are usually "covered" by a bow shaped symbol called the "pokrytie" (Slavic for "cover") and encoded in Unicode at U+0487 COMBINING CYRILLIC POKRYTIE. The character "titlo" encoded at U+0483 COMBINING CYRILLIC TITLO is also used in modern Church Slavic to indicate that a letter or letters have been omitted from the spelling of a word; it is also used to indicate numerals, and it is never used to "cover" a superscripted letter. The character U+A66F COMBINING CYRILLIC VZMET is not used in modern Church Slavic.

In Old Church Slavonic manuscripts, the use of titlo, vzmet and pokrytie is more or less interchangeable. Superscript letters usually occur either by themselves or covered by a pokrytie or vzmet (more rarely, titlo) and either the titlo or the vzmet can occur as an indication that letters have been omitted (either in *nomina sacra* or in abbreviations) and in numerals. In fact, the distinction between titlo and vzmet is purely typographical, and the two characters ought to be viewed as two glyphs depicting the same character. For examples, we refer the reader to Karsky (1979, pp. 230ff).

We therefore propose the following changes to the Unicode documentation. Section 7.4 should be amended to read:

> The [Cyrillic Extended-A] block contains a set of superscripted (written above), or titlo letters, used in Church Slavic and Old Church Slavic texts, usually to indicate abbreviations of words in the text. These characters may be followed by U+0487 COMBINING CYRILLIC POKRYTIE, and in Old Church Slavic texts, also by U+0483 COMBINING CYRILLIC TITLO, or by U+A66F COMBINING CYRILLIC VZMET.

We also propose the following changes to Annotations. For U+0483 COMBINING CYRILLIC TITLO, remove the annotation `not used with letter titlos`. Add the annotation:
`used in Cyrillic or Glagolitic to indicate abbreviation or numeral.`
Under "See also", add a reference to U+0487 COMBINING CYRILLIC POKRYTIE. The Script property of this character should be changed to `Inherited`, since it can be used both for Cyrillic and Glagolitic.

For U+0487 COMBINING CYRILLIC POKRYTIE, remove the annotation:

`used only with letter titlos.`

Add the annotation:

`used with combining Cyrillic or Glagolitic letters.`

Under “See also”, remove the reference to U+0311 COMBINING INVERTED BREVE. The character U+0311 is used to encode a Cyrillic Kamora (circumflex accent), not for supralineation. Add a reference to U+0483 COMBINING CYRILLIC TITLO. The Script property of this character should be changed to `Inherited`, since it can be used both for Cyrillic and Glagolitic.

For U+A66F COMBINING CYRILLIC VZMET, remove the annotation:

`used with Cyrillic letters to indicate abbreviation.`

Add the annotation:

`used in Cyrillic or Glagolitic to indicate abbreviation or numeral.`

Again, the Script property of this character should be changed to `Inherited`, since it can be used both for Cyrillic and Glagolitic.

7 Naming Conventions

Because in its listing of the language in the CLDR, the term “Church Slavic” has been adopted, we propose that the Unicode documentation be made to conform with the CLDR terminology and that the words “Church Slavonic” in the documentation be replaced everywhere by “Church Slavic”. The terms “Church Slavic” and “Church Slavonic” are used interchangeably in English, and this naming change is only for the sake of consistency between the two sets of documentation.

References

Andreev, A., Y. Shardt, and N. Simmons (2014). Proposal to Encode Additional Cyrillic Characters Used in Early Church Slavonic Printed Books. ISO/IEC JTC1/SC2/WG2 N4607.

Everson, M., D. Birnbaum, R. Cleminson, I. Derzhanski, V. Dorosh, A. Kryukov, and S. Paliga (2006). On CYRILLIC LETTER OMEGA WITH TITLO and on CYRILLIC LETTER UK. ISO/IEC JTC1 SC2 WG2 N3184.

Everson, M., D. Birnbaum, R. Cleminson, I. Derzhanski, V. Dorosh, A. Kryukov, S. Paliga, and K. Ruppel (2007). Proposal to encode additional cyrillic characters in the BMP of the UCS. ISO/IEC JTC1 SC2 WG2 N3194.

Gamanovitch, A. (1991). *Грамматика Церковно-славянского Языка.* Moscow: Khudozhestvennaya Literatura Press.

Golyshenko, V. S. (1987). *Мягкость согласных в языке восточных славян XI-XII вв.* Moscow: Nauka.

Karsky, E. F. (1979). *Славянская Кирилловская Палеография.* Moscow: Nauka Press.

Uspensky, B. A. (Ed.) (2006). *Типографский устав: Устав с кондакарем конца XI – начала XII века*, Volume 2. Moscow: Yazyki Slavyanskikh Kulʹtur.

Yelkina, N. M. (1960). *Старославянский язык.* Moscow, Russia: Ministry of Education of the RSFSR.