L2/13-139

Proposal to Encode Combining Half Marks Used for Cyrillic Supralineation in Unicode

Aleksandr Andreev* Yuri Shardt Nikita Simmons

PONOMAR PROJECT

Abstract

A Proposal to add two additional characters to the Combining Half Marks block of Unicode to be used for the correct presentation of supralineation in Church Slavonic texts.

1 Introduction

The Cyrillic writing system used to record the Church Slavonic language uses the character *titlo* (Slavonic: тѝтло) as a combining mark, most often placed over a single character. The placement of this mark has several uses. First, it may be used to indicate that the letter or group of letters are to be interpreted as a numeral (*e.g.*, а҃ = 1, ка҃ = 21, рка҃ = 121). Second, it may be used to indicate that a letter or group of letters are missing from a word, which is thus an abbreviation (*e.g.*, цр҃ь = цáрь, *king*). Finally, it may be used to indicate a *nomen sacrum*, an abbreviation for writing divine names (*e.g.*, бг҃ъ = Бóгъ, *God* vs. бóг҃ъ, a false deity). In the Unicode standard, the titlo has been encoded as U+0483, Combining Cyrillic Titlo.

In several instances, a titlo may occur over two or more letters. First, this usage is evident in iconographic inscriptions. For example, in iconography, one finds the inscription МР҃ ѲУ҃ (a Slavonic rendition of the Greek Μήτηρ του Θεού, *Mother of God*) or ІС҃ ХС҃ (Їисоу҃ Хрїстóс҃, *Jesus Christ*). One may also find inscriptions where the titlo balances over more than two letters, for example цр҃ь дв҃дъ (Цáрь Дави́дъ, *King David*), as can be seen in Figure 4. A correct econding mechanism for the titlo in these cases is necessary for the use of iconographers and students of iconography.

Second, in early Ustav (Uncial) manuscripts of Church Slavonic, the titlo is commonly found to balance over two or more letters, both when indicating a numeral and an abbreviation / *nomen sacrum*. Figure 1 and 2 reveal examples from the Sava's book (Саввина книга), an eleventh century Cyrillic Church Slavonic evangeliary; and from the Codex Suprasliensis (Супрасльский сборник), an eleventh century Church Slavonic Menaion. Currently, several online projects are undertaking the task of presenting such early manuscripts in a digital format; while no encoding scheme can be sufficient to transmit all elements of a manuscript, an encoding scheme should provide for a way of encoding the essential elements of the writing system. The use of the titlo over several characters is one such element, given, especially, that in some instances, a semantic difference may exist.

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

Third, the titlo may balance over two or more letters in academic publications that study early manuscripts. The examples in Figure 3 are taken from Yelkina (1960) and reveal a usage of the titlo over two or three letters. This is done to emphasize the particular features of the manuscripts being considered.

2 Proposed Characters

The correct method in Unicode for encoding a titlo over multiple characters is via the use of Combining Half Marks (U+FE20 – U+FE2F). These codepoints are used to encode "combining marks that apply to multiple base letterforms" (Allen et al., 2012, p. 243). These marks are implemented in a way such that "a discontiguous sequence of the combining half marks corresponds to a single combining mark" (*ibid.*). One common use for these marks is for a particular type of supralineation used in Coptic (Allen et al., 2012, p. 228).

As of version 6.2, the Unicode standard provides combining half marks for an inverted breve (U+FE20 and U+FE21). Since the inverted breve is used to encode a Cyrillic *kamora* (circumflex accent), these codepoints cannot be used for the *titlo*. The Unicode standard also provides combining half marks for a tilde (U+FE22 and U+FE23). However, since a tilde differs from a *titlo* both in visual appearance and in function, these codepoints also should not be used to encode a *titlo*. Finally, the standard provides three marks (U+FE24, U+FE25 and U+FE26) used for Coptic supralineation. Of these, the Combining Macron Left Half (U+FE24) and Combining Macron Right Half (U+FE25) cannot be used for the *titlo* because, first, a *titlo* has a distinct visual appearance from a macron and, second, according to the Unicode documentation, the Combining Macron halves are designed to "extend from the middle of the first character in the sequence" of the supralineation "to the middle of the last character in the sequence" (Allen et al., 2012, p. 228). In contrast, the *titlo* commonly balances over the entire character.

We therefore propose for encoding two additional characters, the Combining Titlo Left Half and the Combining Titlo Right Half, to be encoded at U+FE2E and U+FE2F, respectively. When a *titlo* is to balance over three or more characters in a Church Slavonic letter sequence, we propose that the existing Combining Conjoining Macron (U+FE26) be used over the middle elements of the sequence, as the Conjoining Macron is visually identical to the middle part of *titlo*. This usage of the Combining Conjoining Macron is in keeping with the recommendations set forth in Irish NB and German NB (2011). Thus, the abbreviation ц҆р҃ѣ҆ would be encoded ц҆ followed by U+FE2E, р҃ followed by U+FE26, and ѣ҆ followed by U+FE2F. The two proposed characters are summarised in the Table 1.

Table 1: Table of Proposed Characters

Glyph	Codepoint	Name
◌͡	U+FE2E	COMBINING CYRILLIC TITLO LEFT HALF
◌͜	U+FE2F	COMBINING CYRILLIC TITLO RIGHT HALF

3 Implementation

Any sequence of one base character with U+FE2E applied, zero or more base characters each with U+FE26 applied, and one base character with U+FE2F applied, shall yield a *titlo* over the complete sequence of base characters, starting with the one to which U+FE2E is applied, and ending with the one to which U+FE2F is applied. This follows exactly the recommendations set forth in Irish NB and German NB (2011).

At the font level, one of two implementations is possible. The first implementation relies on glyph substitution. The sequence of combining marks beginning with U+FE2E and ending with U+FE2F can be replaced with a single glyph for a double, triple, quadruple (or longer) *titlo*. This substitution can take place via the *ccmp* feature in OpenType (in the subsitution table, the "Ignore base glyphs" flag needs to be set) or via an appropriate substitution rule in SIL Graphite. Under this approach, problems occur with the positioning of the composed *titlo* glyph, as correct positioning needs to take into account both the different width and height of the base glyphs. In SIL Graphite, it is possible to write positioning rules that take into account the horizontal and vertical glyph metrics of the base glyphs and would thus correctly position the composed combining glyph. In OpenType, to our knowledge, this is not possible. Rather, it would be necessary to write contextual positioning rules that would determine the horizontal and vertical position of the composed glyph on the basis of the sequence of base glyphs. In practice, this becomes quite tedious as the number of glyph classes becomes large.

The second implementation approach is to create precomposed glyphs of the base characters with titlo halves of the appropriate height and width for each of the possible combinations of base characters. The correct precomposed glyph is then selected via the use of contextual substitution rules. This approach has the advantage that the order of glyphs is preserved. Since the precomposed glyph of the base letter and half mark can be given an appropriate glyph name in the font in accordance with the Adobe Glyph Naming convention, the correct codepoints in the correct order will be preserved under such operations as copying from a PDF document. Under the first approach, the correct order is not preserved, since all of the half marks are eliminated and replaced with a single mark that combines with the first base character. On the other hand, this approach is by far even more tedious than the first approach, as it requires the creation of precomposed glyphs for at least every single letter of the Church Slavonic Cyrillic alphabet.

The authors feel that in the future an extension to OpenType should be considered that allows the use of existing technologies for the correct "joining" of combining marks, for example, via the use of the *curs* (cursive attachment) feature. Presently, this is not possible. However, if this feature were extended to combining marks over different base glyphs, it would allow half marks to be joined together visually without resorting to glyph substitution and complex contextual rules. This would greatly simplify the implementation of half marks used in Church Slavonic or Coptic supralineation, as well as in other settings.

4 Character Properties

The following entries are proposed for addition to UnicodeData.txt:

```
FE2E;COMBINING CYRILLIC TITLO LEFT HALF;Mn;230;NSM;;;;;N;;;;;
FE2F;COMBINING CYRILLIC TITLO RIGHT HALF;Mn;230;NSM;;;;;N;;;;;
```

Figure 1: Examples of the *titlo* used over multiple letters (highlited in red) and over a single letter (highlited in blue). Source: Sava's book, as reproduced by (Schepkin, 1903) .

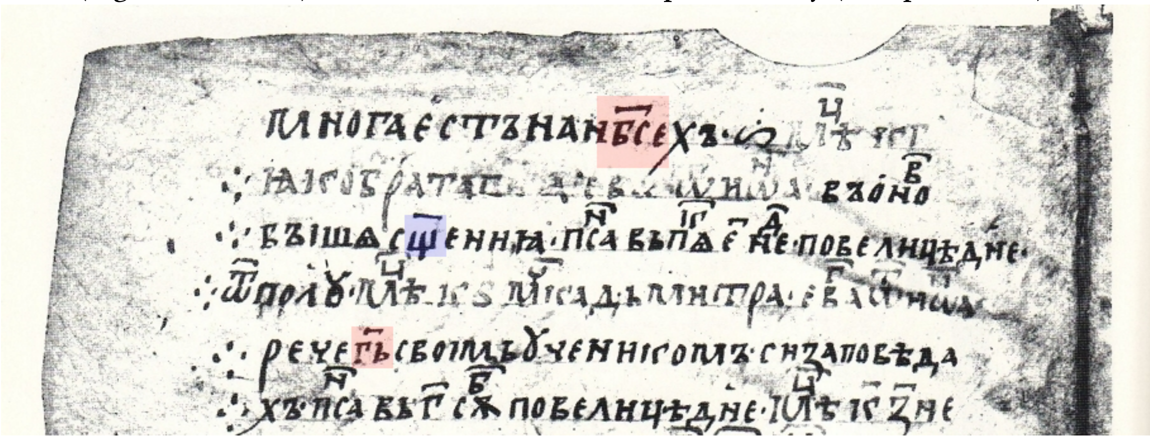


Figure 2: Examples of the *titlo* used over multiple letters. Source: Codex Suprasliensis, as reproduced by (Sever'yanov, 1904) .

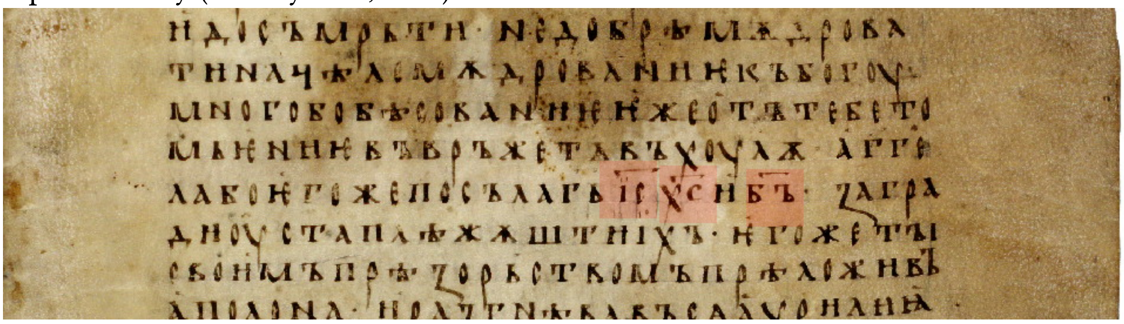
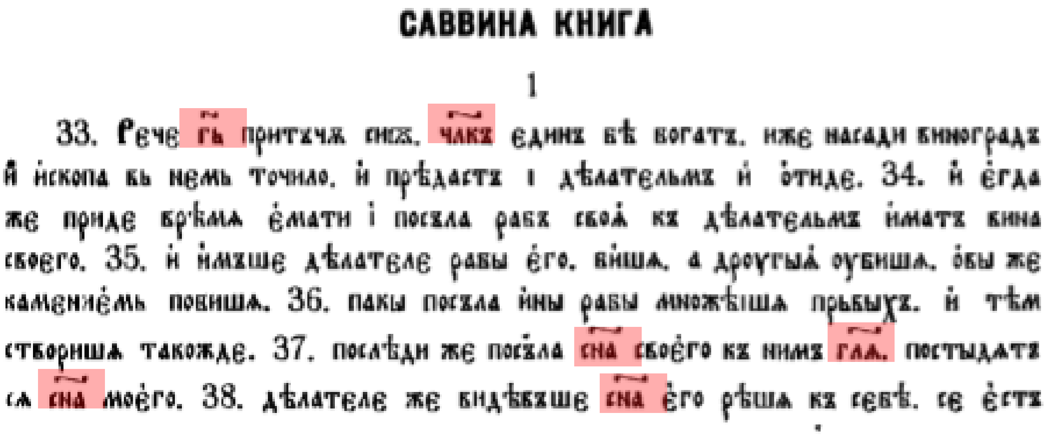


Figure 3: Examples of the *titlo* used over multiple letters in an academic setting. Source: (Yelkina, 1960).

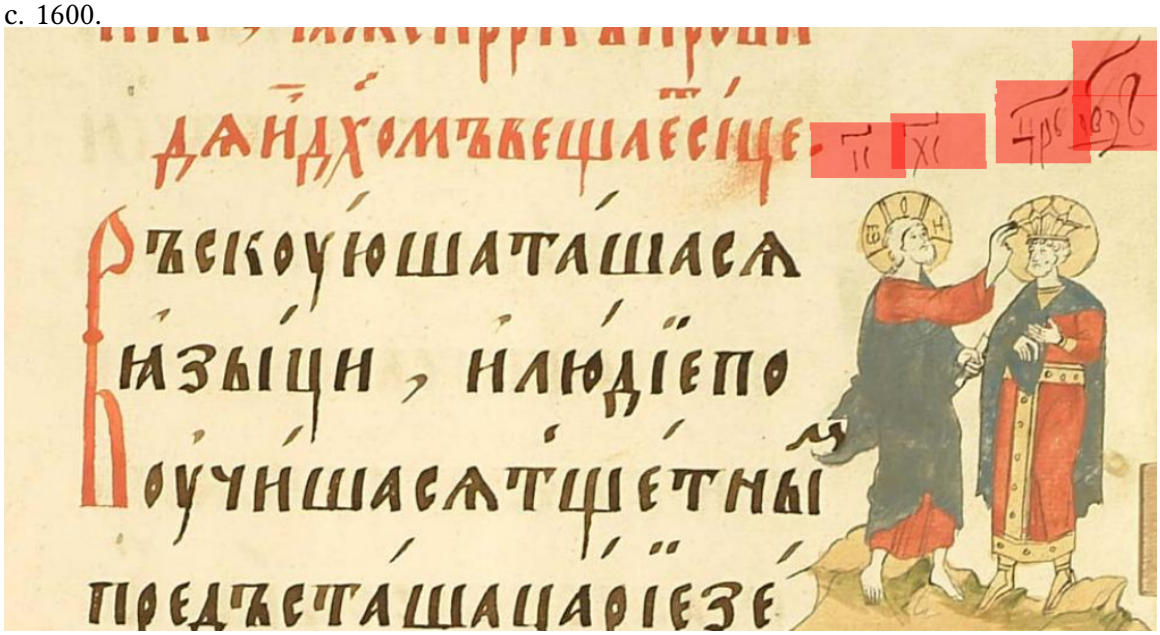


САВВИНА КНИГА

1

33. Рече гь притъчѫ сиѭ. члкъ единъ бѣ богатъ. иже насади виноградъ и ископа въ немь точило. и прѣдастъ і дѣлателемъ и отиде. 34. и егда же приде врѣмѧ ѥмати і посъла рабъ своѩ къ дѣлателемъ имать вина своего. 35. и имъше дѣлателе рабы его. биша. а дроугыѩ оубиша. овы же каменіемь побиша. 36. пакы посъла ины рабы множѣиша прьвыхъ. и тѣм створиша такожде. 37. послѣди же посъла сна своего къ нимъ гла. постыдатъ сѧ сна моего. 38. дѣлателе же видѣвъше сна его рѣша къ себѣ. се естъ

Figure 4: Examples of the *titlo* used in iconographic inscriptions. Source: illustrated *Psalter*, c. 1600.



References

Allen, J. D., D. Anderson, J. Becker, R. Cook, M. Davis, P. Edberg, M. Everson, A. Freytag, J. H. Jenkins, R. McGowan, L. Moore, E. Muller, A. Phillips, M. Suignard, and K. Whistler (2012, September). *The Unicode Standard Version 6.2 – Core Specification.* Mountain View, CA: The Unicode Consortium.

Irish NB and German NB (2011). *Revised Proposal to enable the use of Combining Triple Diacritics in Plain Text.* Working Group Document N4078.

Schepkin, V. N. (1903). *Саввина книга.* Number v. 1, pt. 2 in Памятники старославянского языка. Изд. Отд-нія русскаго языка и словесности Императорской академіи наук.

Sever'yanov, S. N. (1904). *Супрасльская рукопись.* Number v. 1 in Памятники старославянского языка. Изд. Отделенія русскаго языка и словесности Имп. Академіи наук.

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

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://std.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://std.dkuug.dk/JTC1/SC2/WG2/docs/summaryform.html **.**
See also http://std.dkuug.dk/JTC1/SC2/WG2/docs/roadmaps.html **for latest *Roadmaps*.**

A. Administrative

1. **Title:** *Proposal to Encode Combining Half Marks used for Cyrillic Supralineation*
2. Requester's name: *Aleksandr Andreev, Yuri Shardt, Nikita Simmons*
3. Requester type (Member body/Liaison/Individual contribution): *Individual contribution*
4. Submission date: *07/08/2013*
5. Requester's reference (if applicable): *N/A*
6. Choose one of the following:
   This is a complete proposal: *YES*
   (or) More information will be provided later:

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: *2*
3. Proposed category (select one from below - see section 2.2 of P&P document):
   A-Contemporary ( ) B.1-Specialized (small collection) (X) 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?
   Aleksandr Andreev
   b. Identify the party granting a license for use of the font by the editors (include address, e-mail, ftp-site, etc.):
   Hirmos Ponomar font distributed by Aleksandr Andreev, Yuri Shardt, Nikita Simmons under GNU GPL
   http://www.ponomar.net/ or aleksandr.andreev@gmail.com
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*
   Implementation of half marks using OpenType and SIL Graphite is briefly discussed

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 Unicode Character Database (http://www.unicode.org/reports/tr44/) and associated Unicode Technical Reports for information needed for consideration by the Unicode Technical Committee for inclusion in the Unicode Standard.

[1] Form number: N4102-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, 2011-03, 2012-01)

C. Technical - Justification

1. Has this proposal for addition of character(s) been submitted before? NO
   If YES explain
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? *Slavonic Typography Society*
   If YES, available relevant documents: *Online discussion at http://cslav.orthonet.ru/*
3. Information on the user community for the proposed characters (for example: size, demographics, information technology use, or publishing use) is included? NO
   Reference:
4. The context of use for the proposed characters (type of use; common or rare) *Rare*
   Reference: *See Section 1, Introduction*
5. Are the proposed characters in current use by the user community? YES
   If YES, where? Reference: *See references to academic literature in Proposal*
6. After giving due considerations to the principles in the P&P document must the proposed characters be entirely in the BMP? NO
   If YES, is a rationale provided?
   If YES, reference:
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, or could be confused with, an existing character? YES
    If YES, is a rationale for its inclusion provided? YES
    If YES, reference: *Unicode distinguishes between Titlo and Tilde or Macron*
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 Section 2, Proposed Characters*
    Is a list of composite sequences and their corresponding glyph images (graphic symbols) provided? *N/A*
    If YES, reference:
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 characters? NO
    If YES, are the equivalent corresponding unified ideographic characters identified?
    If YES, reference: