Request for Clarification in the Encoding Model for Some Cyrillic Combining 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. This document requests the UTC to clarify the encoding model for Cyrillic as used in typesetting Church Slavic in order to provide for consistency in the encoding of combining Cyrillic letters.

2 Problem Description

Church Slavic uses combining (superscript) Cyrillic letters to indicate that a word has been abbreviated in writing, either as a spelling convention (for example, in *nomina sacra*) or as a space-saving device. Such combining letters have already been encoded in the Unicode standard in the Cyrillic Extended-A and Cyrillic Extended-B blocks. Church Slavic also uses composite combining letters, which consist of either a digraph of two combining letters or a ligature made of two components, both occurring over one base character (in manuscripts, the combining characters may be written quite large and may appear to occur over several characters, but this is a stylistic embellishment). The Unicode standard has already encoded one such composite combining letter, U+2DF5 COMBINING CYRILLIC LETTER ES-TE (◌ⸯ). An example of this character is presented in the first row of Table 1.

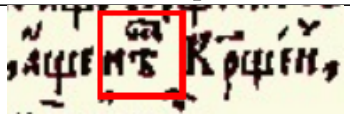
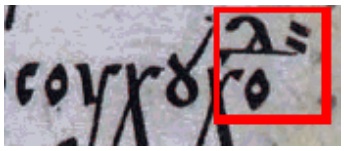
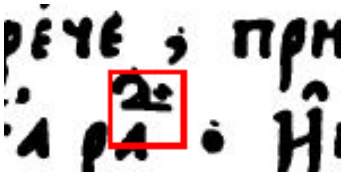
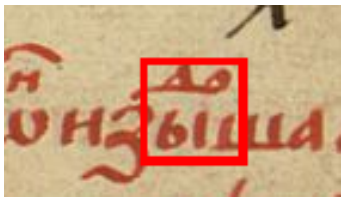
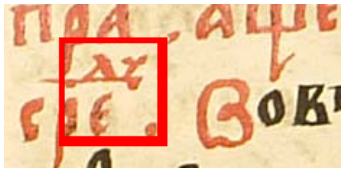
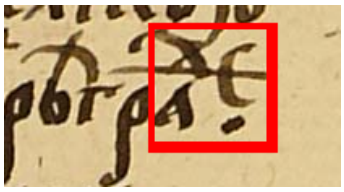
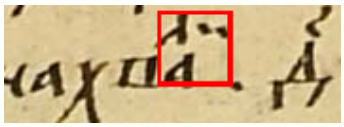
In addition to this character, our research has identified a number of other such composite characters that are used. Some of the examples presented in Table 1 occur in printed editions, which have a stable character repertoire. We also present some additional examples from manuscripts. A cursory review of the manuscripts available to us reveals the presence of the following composite combining titli, in addition to Combining Es-Te: Be-Ie, Be-O, Ve-I, Ve-O, Ge-O, De-Ie, De-I, De-En, De-O, De-U, De-Combining Vertical Tilde, Zhe-Ie, Ze-I, Ka-I, El-I, Em-A, Em-I, Em-U, En-A, En-Ie, Er-I, Te-I, Ha-I, Sha-I, Sha-Ie, Sha-I. We show some of

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

these in Table 1. Undoubtedly additional research in the manuscript tradition would identify many more such forms, since Church Slavic orthography in the manuscript tradition is quite unstable.

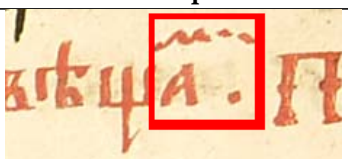
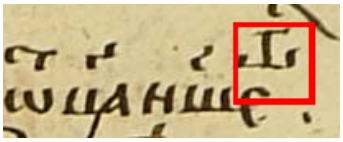
The character Combining De-I (row 3 in Table 1) occurs in the Ostrog Bible, the *editio princeps* of the Church Slavic Bible printed in Ostrog (modern Ukraine) in 1581. Our team is presently actively involved in the digital encoding of this important document and thus there is an urgent need to clarify how this, and other composite combining characters should be encoded. We propose that the UTC rule that the format character U+200D ZERO WIDTH JOINER should be used create combining ligatures or digraphs in Cyrillic.

Table 1: Some Composite Combining Letters Used in Church Slavic

Name	Components	Example	Source
Es-Te	U+2DF5		T
De-I	U+2DE3, U+A675		ATLT
De-I	U+2DE3, U+A675		O
De-O	U+2DE3, U+2DEA		VG
De-Uk	U+2DE3, U+2DF9		Ob. 252
De-Ie	U+2DE3, U+2DF7		Ob. 252
El-I	U+2DE7, U+A675		Ob. 249

Continued on next page

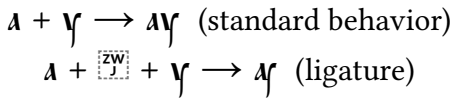
Table 1: Composite Combining Letters (cont'd)

Name	Components	Example	Source
Em-I	U+2DE8, U+A675		Ob. 252
Ge-O	U+2DE2, U+2DEA		Ob. 249

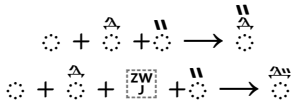
Key to sources:	
ATLT	Lenten Triodion, Moscow: Anonymous Tipografiya, 1555
VG	Gospel Book, Vilnius, 1575
O	Ostrog Bible, Ivan Fedorov, Ostrog, 1581
Ob. 252	Obikhodnik, *ms.* #252, early 17th century in Russia
Ob. 249	Obikhodnik, *ms.* #249, Holy Trinity-St. Sergius Lavra, 1645
T	Trebnik of Metropolitan Peter (Mogila), Kiev, 1646

3 Proposed Solution

The Unicode standard presently provides U+200D ZERO WIDTH JOINER, a format character that requests that two adjoining characters be interpreted by the rendering system as a ligature. The ZWJ may be presently used in Cyrillic to create ligated letters also commonly used in Church Slavic (see Figure 1 for one example) in the following manner:



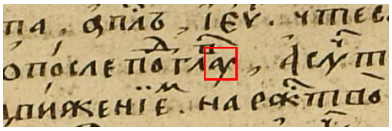
In Indic scripts, ZERO WIDTH JOINER may also be used with combining characters. We propose to allow this behavior for Cyrillic also. Thus, to request that two combining letters form a ligature, the ZWJ should be placed between them, as follows:



With this proposed solution, the default vertical stacking behavior of the characters would be preserved. The existing encoding model for Cyrillic is not altered in any way, and already encoded texts that rely on vertical stacking of combining marks are not affected.

However, the encoding of U+2DF5 COMBINING CYRILLIC LETTER ES-TE in the Unicode standard does mean that an ambiguity is now created regarding the spelling of the ligature Combining Es-Combining Te, which could be encoded as either U+2DF5 or as the sequence

Figure 1: Ligature Cyrillic A-Cyrillic U, which often occurs in Church Slavic documents; Source: Obikhodnik, *ms.* #252 of the collection of Holy Trinity-St. Sergius Lavra, early 17th century in Russia.



U+2DED COMBINING CYRILLIC LETTER ES; U+200D ZERO WIDTH JOINER; U+2DEE COMBINING CYRILLIC LETTER TE. In our view, the encoding of U+2DF5 in Unicode was erroneous since the glyph is not a character, but a ligature, and since other combining ligatures were not encoded also. We recommend that for the sake of consistency and to avoid a spelling ambiguity, the UTC rule that the use of U+2DF5 COMBINING CYRILLIC LETTER ES-TE is now deprecated and that the Es-Te combination should be instead encoded as the sequence U+2DED COMBINING CYRILLIC LETTER ES; U+200D ZERO WIDTH JOINER; U+2DEE COMBINING CYRILLIC LETTER TE.

4 Alternative Solution 1

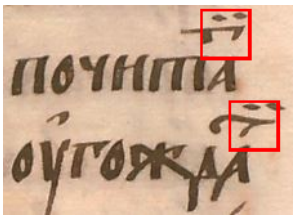
If the deprecation of U+2DF5 COMBINING CYRILLIC LETTER ES-TE is not advisable, then it should be noted that all of the characters that we have identified function in exactly the same way as U+2DF5, and thus, taking the encoding of U+2DF5 as precedent, we propose to encode the seven composite combining letters identified in Table 1 as new characters in the seven remaining slots of the Cyrillic Extended-C block (beginning at the codepoint U+1C89). Further additions to this repertoire would need to be encoded in a new Cyrillic Extended-D block in the Supplemental Multilingual Plane. The advantage of such an approach is that it is simple from an implementation standpoint and entirely consistent with the existing encoding model. The disadvantage of this approach is that it would set a further precedent for the encoding of many more composite combining characters in the Unicode standard, which may or may not be advisable, given that the UTC has taken a stand against encoding any new ligatures in the standard. We do not encourage the UTC to adopt this approach, as the number of combining Cyrillic characters that could then be encoded under this approach is potentially quite large, and this would lead to an unnecessary cluttering of the standard. In addition, the sort order and proper codepoint order for these characters would not be entirely clear.

5 Alternative Solution 2

Under default behavior, multiple combining characters placed above a base character will be stacked vertically. In other words, the currently expected default behavior for multiple combining Cyrillic letters over a base character is as follows:

◌ + ◌ᴀ +◌ᴜ ⟶ ◌ᴜᴀ

Figure 2: Example of Combining Cyrillic Letter Te and Combining Cyrillic Letter I stacked vertically; source: “Nebesa” by John the Exarch of Bulgaria, *ms.* written c. 1500, private collection.



One possible solution to the problem of combining digraphs and ligatures is to define the default behavior for combining Cyrillic letters to be stacking horizontally left-to-right instead of stacking vertically. In rendering systems, where called for by the rules of the writing system and available in the font, the two adjoining characters could form a ligature (via the Glyph Composition / Decomposition feature of OpenType); thus we would have:

◌ + ◌ + ◌ ⟶ ◌

However, in Church Slavic texts, it is also not uncommon to find two combining letters stacked vertically, as can be observed in Figure 2. Redefining the default stacking behavior for combining Cyrillic letters would create problems for encoding such vertically stacked cases. One would need a mechanism to specify that the (now) default horizontal stacking behavior should be avoided. For example, one could then define that placing U+034F COMBINING GRAPHEME JOINER would override the default horizontal stacking. This would extend the function of the CGJ beyond its current usage as a format character used to prevent canonical reordering of combining characters.

We believe that this approach is too complicated. It creates unnecessary complexity in the encoding model, which opens the door for possible misunderstanding and misuse by users, and leads to potential problems for already encoded texts since support for the behavior of CGJ in rendering systems is tenuous. Moreover, it is unclear what the expected behavior should be if three or more combining letters are placed in sequence over a single base character. All of the examples that we have identified in our research only demonstrate two combining letters occurring over a base character. Furthermore, if two combining letters are followed by another diacritical mark, it is unclear with this approach if the diacritical mark should apply to the last character or to both characters, as in the following example:

◌ + ◌ + ◌ + ◌ ⟶ ?

6 Conclusion

We propose that the UTC adopt the first approach and recommend the use of U+200D ZERO WIDTH JOINER for the encoding of combining digraphs and ligatures in Cyrillic. We further propose that mention of this be made in Section 7.4 (Cyrillic) of the Unicode documentation, in the subsection “Cyrillic Extended-A: U+2DE0–U+2DFF” with the addition of text to the first paragraph reading:

Ligatures of two superscripted Cyrillic letters are encoded by placing U+200D ZERO WIDTH JOINER between the two letters. The ZERO WIDTH JOINER is a format character that is used to request that the rendering system connect the adjoining characters to form a ligature. See Section 23.2 Layout Controls for more information. The use of U+2DF5 COMBINING CYRILLIC LETTER ES-TE is deprecated and this ligature should be encoded as the sequence U+2DED COMBINING CYRILLIC LETTER ES; U+200D ZERO WIDTH JOINER; U+2DEE COMBINING CYRILLIC LETTER TE.

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:** *Request for Clarfication in the Encoding Model for Some Cyrillic Characters*
2. Requester's name: *Aleksandr Andreev, Yuri Shardt, and Nikita Simmons*
3. Requester type (Member body/Liaison/Individual contribution): *Individual contribution*
4. Submission date: *01/12/2015*
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: *NO*
      Name of the existing block:
2. Number of characters in proposal: *0*
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)? *NO*

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; Russian Technical Committee 22
   If YES, available relevant documents: E-mail correspondence
3. Information on the user community for the proposed characters (for example: size, demographics, information technology use, or publishing use) is included? YES
   Reference: Section 1 and Section 2
4. The context of use for the proposed characters (type of use; common or rare) Rare
   Reference: Used in the Ostrog Bible and in various mss. See Table 1.
5. Are the proposed characters in current use by the user community? YES
   If YES, where? Reference: In digital reproductions of sources.
6. After giving due considerations to the principles in the P&P document must the proposed characters be entirely in the BMP? N/A
   If YES, is a rationale provided?
   If YES, reference: We propose to treat characters as composites. See Section 3.
7. Should the proposed characters be kept together in a contiguous range (rather than being scattered)? N/A
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? YES
   If YES, is a rationale for its inclusion provided? N/A
   If YES, reference: The one existing character that had been encoded earlier (U+2DF5) should be deprecated.
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? 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 Section 3, Proposed Solution
    Is a list of composite sequences and their corresponding glyph images (graphic symbols) provided? YES
    If YES, reference: Table 1
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: