

PONOMAR PROJECT

Proposal to Encode Some Outstanding Early Cyrillic Characters in Unicode



Yuri Shardt, Nikita Simmons, Aleksandr Andreev

In old, Slavic documents that come from Eastern Europe in the centuries between A.D. 1400 and 1700, it is possible to come across various unusual characters, whose forms have not yet entered the Unicode repertoire. As well, these symbols can occasionally be found in more recent publications by those Orthodox Christians (Old Believers) that do not accept the typographic changes introduced by Patriarch Nikon in the middle 1600. The exact forms of these symbols often vary between different places and times. In the interest of creating a standard for faithfully typesetting Slavonic manuscripts, there is a need to include these symbols in Unicode. Since these symbols are often found in Slavonic Church books, these symbols will be encoded in the “Extended Cyrillic Block B” of the Unicode standard. The suggested new characters can be divided into three categories: characters, diacritical marks, and character combinations.

Characters

Table 1 presents a summary of the proposed characters for encoding in Unicode: Double O and Crossed O, which are found in many early Slavonic manuscripts. The double o is used in the words for both two (**двое**), both (**обо**), and twelve (**обана**десять, **двою**надесять), where the bold letters denote the usual placement of the proposed double o. This letter would complement the other forms of o that are already present in the Unicode standard, including the monocular o, the binocular o, the double monocular o, and the multiocular o. The crossed o is primarily used in the word for neighbourhood (**окрест**) in early Slavonic manuscripts. This character’s usage is more common than that of the multiocular o and hence should also be added. Examples of both characters can be found in Figure 1 and Figure 2.



Table 1: Summary of the Proposed Characters for Encoding

Proposed Character	Proposed Name	Location	Comments
	CYRILLIC [CAPITAL, SMALL] LETTER DOUBLE O	U + A698 U+A699	This letter is used in the words for two (двое), both (обо), twelve (обана десять and двою надесять).
	CYRILLIC [CAPITAL, SMALL] LETTER CROSSED O	U + A69A U+A66B	This letter is used in the word for neighbourhood (окрест).

Diacritical Marks

Table 2 presents a summary of the proposed diacritical marks for inclusion in Unicode: the Combining Cyrillic De-I and the Combining Cyrillic Zhe-E. Both of the double superscripts are similar in nature to the already encoded Combining Cyrillic Es-Te (U+2DF5) and are not divisible ligatures. This is especially true of the combining Cyrillic zhe-e as used in the Ostrog Bible and early Bulgarian manuscripts, where it is an almost unidentifiable combination of zhe and e to represent an abbreviation for the letters zhe (ѣѣ). Both characters should be included in the Unicode standard. Examples are shown in Figure 4, Figure 5, Figure 6, and Figure 9(right).

Table 2: Proposed Diacritical Mark




Proposed Diacritical Mark	Proposed Name	Proposed Location	Comments
	COMBINING CYRILLIC DE-I	[U+ A66C]	This is not the most common, or even attested form for the diacritical marks. However, to be consistent with the other combining marks this is the proposed shape.
	COMBINING CYRILLIC ZHE-E	[U+ A66D]	

Ligatures

Table 3 lists some commonly encountered undivisible ligatures that occur consistently in Slavonic documents: Cyrillic Ligature A-Uk, Cyrillic Ligature El-Uk, and Cyrillic Ligature Te-Ve. The Cyrillic ligature a-uk is composed of an a (а) and an uk (ѣ). It is commonly used in words for joy (радѣѣ-) and its derivatives. This word is commonly written with the given ligature with a superscript Cyrillic de (see, for example, Figure 7(right) and Figure 8). In Figure 8, it can be seen that this ligature need not occur every time the word is written. In fact, in the given passage, there are 2 instances of the same word (shown in a blue box) that are written without the ligature. The Cyrillic ligature el-uk is used as an abbreviation for denoting the name Luke in

Gospel lectionaries (see, for example Figure 7(left)). Finally, the Cyrillic ligature te-ve is used in early Slavonic documents to represent a combined te-ve at the whim of the scribe (see, for example, Figure 9).

Table 3: Cyrillic Ligatures

Typical Proposed Ligature	Proposed Name	Proposed Location	Comments
	CYRILLIC LIGATURE A-UK	[U+A66E]	
	CYRILLIC LIGATURE EL-UK	[U+ A??]	
	CYRILLIC LIGATURE TE-VE	[U+A??]	

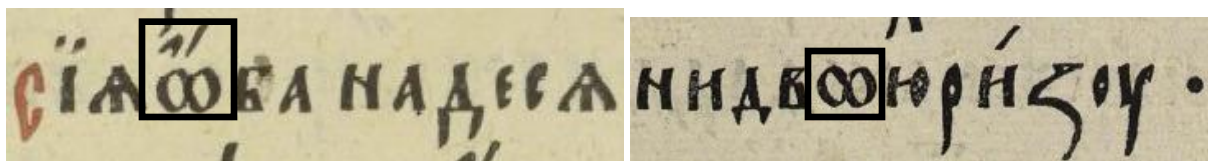


Figure 1: Extract from the 1553/4 Gospel showing the use of the Cyrillic double on (boxed).

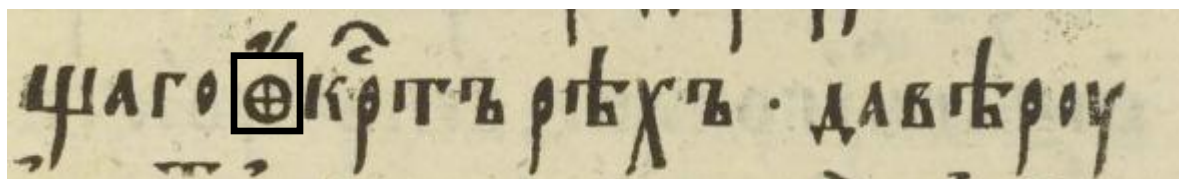


Figure 2: Extract from the 1553/4 Gospel showing the use of the Cyrillic crossed on (boxed).

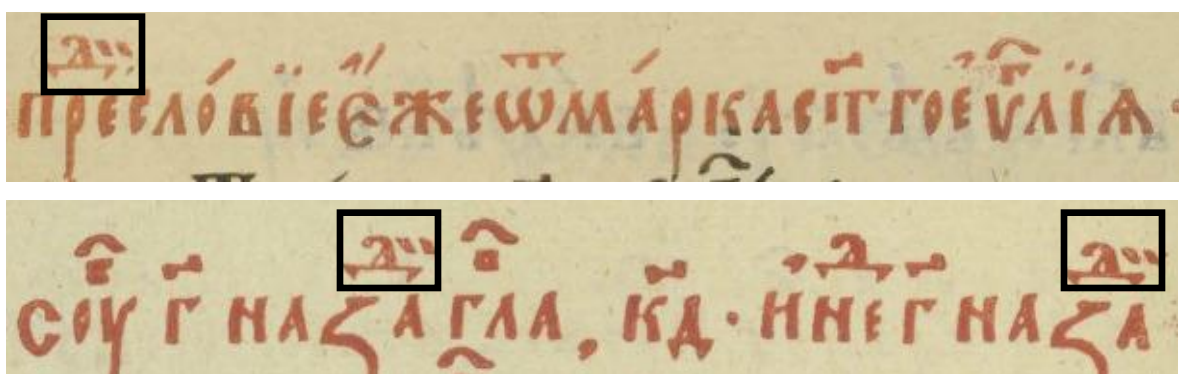


Figure 3: Extracts from the 1553/4 Gospel showing the use of the combining Cyrillic de-i (boxed).

Ѣ, ѢДИНЪ ГРѢХЪ Ѣ И ѢДИНЪ БЪ ВСЕСЪЖЖЕНІЕ. И ДА
ПРИНЕСЕПЪ А КЪ ЖЕРЦУ, И ВЪЗМЕПЪ ЖРЕЦЪ ѢЖЕ ЗА

Figure 4: Extract from the Book of Leviticus in the Ostrog Bible of the combining Cyrillic de-i (boxed) (Turkonjak, 2003).

БОТЪ НОСЛЫША • ПРИСТУПІВШИ

Figure 5: Extract from the Ostrog Bible of the combining Cyrillic zhe-e (boxed).

РЕЧЕЖЕ АБРАМЪ ЛОГПЪ, ДАНЕ БУДЕПЪ
СВАРЪ МЕЖДЪ МНОЮ И ПЛОБОЮ, И МЕДЪ

Figure 6: Extract from the Ostrog Bible of the combining Cyrillic zhe-e (boxed).

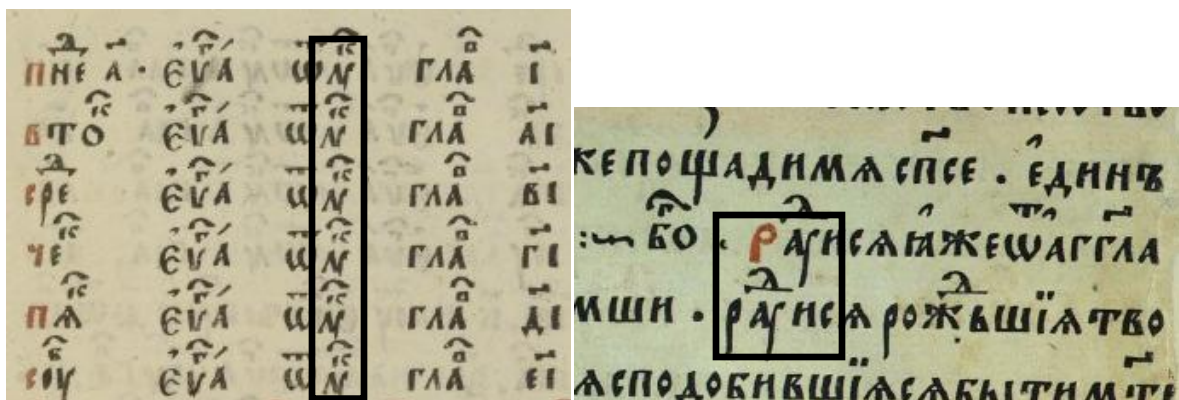


Figure 7: (left) Lectionary from the 1575 Vilnius Gospel showing the use of the Cyrillic ligature el-uk (boxed examples) (Евангелие (The Gospel Books), 1575). (right) Extract from the 1553/4 Gospel containing the Cyrillic ligature a-uk (boxed example).

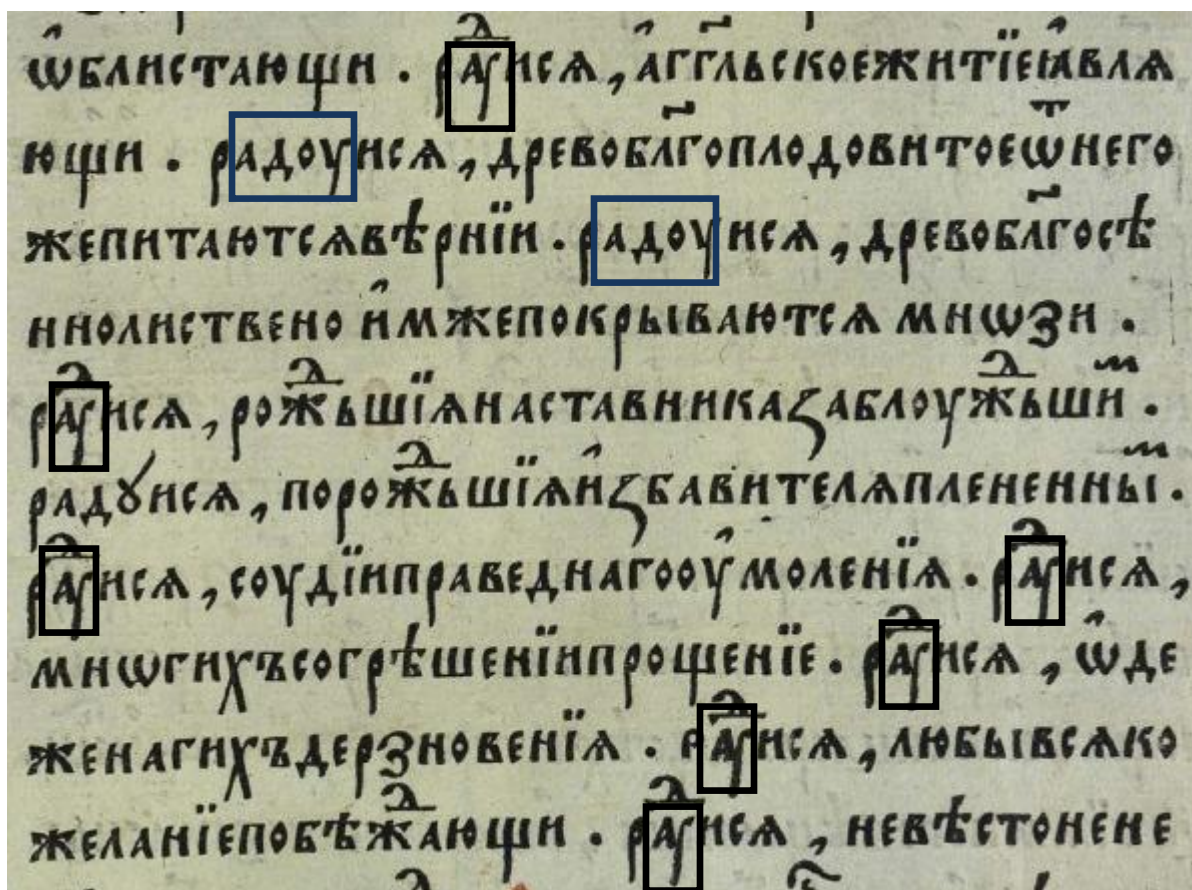


Figure 8: Another example of the Cyrillic ligature a-uk from the 1553/4 Gospel. The examples boxed in black contain the proposed ligature, while the blue box contains the same word not written using a ligature. In all cases, the boxed word is the imperative active reflexive form of the verb rejoice.



Figure 9: Examples of the Cyrillic ligature te-ve (first two from the left) and the combining Cyrillic zhe (right) from a Bulgarian manuscript of the early Slavic style. All examples have been boxed.

References

Turkonjak, R. (. (Ed.). (2003). *Книги Виходу і Книга Левіт (The Books of the Exodus and Leviticus in the Ostrog Bible)*. L'viv, Ukraine: Українське Біблійне Товариство (Ukrainian Bible Society).

Евангеліє (The Gospel Books). (1575). Vilnius, Polish-Lithuanian Commonwealth: Печ. Петр Тимофеев Мстиславец (Publishing House of Petr Timofeev Mstislavec).

ISO/IEC JTC 1/SC 2/WG 2
PROPOSAL SUMMARY FORM TO ACCOMPANY SUBMISSIONS
FOR ADDITIONS TO THE REPERTOIRE OF ISO/IEC 10646¹

Please fill all the sections A, B and C below.

Please read Principles and Procedures Document (P & P) from <http://www.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://www.dkuug.dk/JTC1/SC2/WG2/docs/summaryform.html>.

See also <http://www.dkuug.dk/JTC1/SC2/WG2/docs/roadmaps.html> for latest Roadmaps.

A. Administrative

1. Title:	Proposal to Encode Some Outstanding Early Cyrillic Characters in Unicode
2. Requester's name:	<i>Yuri Shardt, Nikita Simmons, Aleksandr Andreev</i>
3. Requester type (Member body/Liaison/Individual contribution):	<i>Individual Contribution</i>
4. Submission date:	<i>October 3, 2010</i>
5. Requester's reference (if applicable):	
6. Choose one of the following:	
This is a complete proposal:	<i>YES</i>
(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):	<i>NO</i>
Proposed name of script:	
b. The proposal is for addition of character(s) to an existing block:	<i>YES</i>
Name of the existing block:	<i>U+ A66x</i>
2. Number of characters in proposal:	<i>7</i>
3. Proposed category (select one from below - see section 2.2 of P&P document):	
A-Contemporary	<i>Yes</i>
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	
4. Is a repertoire including character names provided?	<i>YES</i>
a. If YES, are the names in accordance with the "character naming guidelines" in Annex L of P&P document?	<i>YES</i>
b. Are the character shapes attached in a legible form suitable for review?	<i>YES</i>
5. Who will provide the appropriate computerized font (ordered preference: True Type, or PostScript format) for publishing the standard?	<i>Yuri Shardt</i>
If available now, identify source(s) for the font (include address, e-mail, ftp-site, etc.) and indicate the tools used:	<i>Hirmos Ponomar v.6 (contact Yuri Shardt at yuri.shardt@ualberta.ca for the font)</i>
6. References:	
a. Are references (to other character sets, dictionaries, descriptive texts etc.) provided?	<i>YES</i>
b. Are published examples of use (such as samples from newspapers, magazines, or other sources) of proposed characters attached?	<i>YES</i>
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)?	<i>NO</i>

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 <http://www.unicode.org/Public/UNIDATA/UCD.html> and associated Unicode Technical Reports for information needed for consideration by the Unicode Technical Committee for inclusion in the Unicode Standard.

¹ Form number: N3152-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)

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? academics; Old Rite Believers	
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?	small, but active
Reference:	
4. The context of use for the proposed characters (type of use; common or rare)	common
Reference:	
5. Are the proposed characters in current use by the user community?	YES
If YES, where? Reference: In typesetting older documents authentically	
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?	YES
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 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?	
If YES, reference:	
Is a list of composite sequences and their corresponding glyph images (graphic symbols) provided?	
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 character(s)?	NO
If YES, is the equivalent corresponding unified ideographic character(s) identified?	
If YES, reference:	