L2/13-153

Proposal to Use Standardized Variation Sequences to Encode Church Slavonic Glyph Variants in Unicode

Aleksandr Andreev[*] Yuri Shardt Nikita Simmons

PONOMAR PROJECT

Abstract

The authors propose an approach to encoding variants of letters in the Church Slavonic language within the Unicode standard via the use of variation sequences. Eight Cyrillic characters are considered, because they have been identified as having variant glyph forms that have an orthographic or grammatical significance in Church Slavonic incunabula, early printed texts in Russia, or texts of the Synodal period of a Kievan provenance. It is proposed to select variants implicitly via the use, where possible, of contextual substitution rules in OpenType or SIL Graphite, which, where necessary, can be overwritten explicitly by the use of variation selectors. For these purposes, it is also proposed to encode the base form of each character as a variation sequence. The list of proposed variation sequences is provided in Table 1.

1 Variation Sequences

The Unicode standard encodes characters, not glyphs. However, as the Unicode standard admits (p. 556), sometimes "the need arises in text processing to restrict or change the set of glyphs that are to be used to represent a character." In formatted text, such changes are indicated by choice of font or style. However, in some situations, the need arises to select glyph variants in a plain text environment where resorting to font-level techniques is impossible or inconvenient. This is especially the case when the displayed glyph form has a particular grammatical or typographical significance.

For this purpose, Unicode provides *variation selectors* – combining marks that provide a "mechanism for specifying a restriction on the set of glyphs ... [or] specifying variants ... that have essentially the same semantics but substantially different ranges" (ibid). A *variation sequence* consists of a base character and a variation selector and is interpreted by the rendering system as affecting the appearance of the base character. Variation sequences are *not* a general code extension mechanism; rather, only those sequences are allowed that are "defined in the file StandardizedVariants.txt in the Unicode Character Database" (ibid).

Standardized variation sequences are currently implemented in the Unicode Standard (version 6.2) for the display of variant glyph forms of some mathematical characters; glyph variants used for Japanese emoji; and for glyph variant forms used in Mongolian and Phags-Pa, where the appearance of the glyph is dictated by complex rules of orthography. Chapter 13 of the Unicode standard describes the use of Variation Sequences in Mongolian:

> Some letters have additional variant forms that do not depend on their position within a word, but instead reflect differences between modern versus traditional orthographic practice or lexical considerations – for example, special forms used for writing foreign

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

> words. On occasion, other contextual rules may condition a variant form selection. For example, a certain variant of a letter may be required when it occurs in the first syllable of a word or when it occurs immediately after a particular letter …
>
> When a glyph form that cannot be predicted algorithmically is required (for example, when writing a foreign word), the user needs to append an appropriate variation selector to the letter to indicate to the rendering system which glyph form is required.

On the basis of this discussion, it can be seen that variation sequences in Unicode can serve one of two purposes in displaying a text in a given writing system.

Contextual glyph forms used outside of context. In a writing system, certain characters may need to be represented by a specific glyph form based on context. For example, characters may take different glyph forms when they occur in the initial, medial, or final position; when they occur in combination with or next to other characters; or when they occur with the presence of diacritical marks. In these instances, the appropriate glyph form is typically selected by specifying the relevant rules to the rendering system via a font technology such as OpenType or SIL Graphite. However, in some instances, it may be necessary to invoke the contextual glyph form *outside* of its usual context *or* to override the default contextual substitution behavior. In these instances, the use of variation sequences is appropriate.

Grammatical or orthographic variants. Some characters may have different glyph forms depending on complex rules of grammar and spelling in the underlying writing system and language. The glyph forms are still treated as the same character – they have the same semantics and for purposes of listing the "alphabet" of the writing system are the same character. However, their use indicates that a word may have a different meaning, a different function, or a different etymology. In these instances, contextual prediction of the glyph form is impossible without resorting to the use of dictionaries and artificial intelligence. Thus, variation sequences provide an appropriate mechanism for specifying which glyph form should be used.

2 Application of Variation Sequences

Pentzlin (2011) has identified two ways in which a variant glyph may be selected:

- explicitly, when the user specifically inserts the relevant *variation selector* to define a *variation sequence* that the rendering system interprets as a glyph variant;
- implicitly, by a higher-level protocol (typically appropriate rules in OpenType, SIL Graphite, or similar technologies) which instructs the rendering system to interpret a specific occurrence of the base character as if it were be followed by the variation selector which constitutes the variation sequence.

Implicit selection occurs whenever a locale-specific glyph selection or a contextual glyph substitution takes place. Explicit selection is the approach to be taken whenever the user needs to select a grammatical or orthographic variant, to select a contextual glyph form outside of context, or to override implicit selection. In order to facilitate the overriding action of explicit selection, it is necessary to also provide explicit variation sequences for the standard (base) glyph form to be used whenever an implicit selection is to be avoided. The examples below demonstrate this functionality.

3 Variation Sequences vs. Other Mechanisms

Other mechanisms can be contemplated for selecting glyph variant forms. In particular, the following come to mind:

3.1 Use of Advanced Typographic Features

Advanced typographic features are presently available in two technologies – OpenType and SIL Graphite. When a glyph variant needs to be specified *explicitly*, that is outside of context, contextual glyph substitution in either technology cannot be used. Instead of using variation sequences, one could contemplate resorting to alternative glyph selection mechanisms, namely, in OpenType, the Stylistic Alternatives feature and the Stylistic Sets, and in SIL Graphite, developer-defined "optional features." In both technologies, these features were designed to allow various typographical effects, for example, alternate glyphs that would give words an informal appearance or add some graphical effect to the typeface. The user could select a glyph variant at the font level by "turning on" the particular font feature.

However, there are a number of problems with this approach. First, it is contingent on the presence of fonts and OpenType and Graphite aware applications. While fonts could be embedded into webpages – and this technology is supported in most modern browsers – and into PDF documents, there is no way to "embed" fonts into text documents. Nor is there any standardized methodology at the moment to turn on advanced typographic features in HTML (clearly an oversight on the part of W3C). The storage and exchange of texts on the Internet, on the other hand, is done in a text-only format (or, at most, using a markup language). Reliance on these features would require the bundling of fonts together with texts and the correct display of the texts would be contingent on the user's ability and desire to install and correctly use the font and relevant OpenType or Graphite aware software. Second, and more importantly, while the Unicode standard is a universal standard for encoding, there is no universal standard for font features. Font developers may choose to implement OpenType features, Graphite features, or both. There is no requirement that developers implement the same features and there is no naming conventions for these features (in fact, the name of the Graphite feature is entirely up to the developer). There is no mechanism for requiring developers to map a specific variant form to a specific feature. Thus, in practice, this approach would make the distribution of texts outside of a closed platform-software system impossible because the given text would only be properly displayed when bundled with a specific font that uses a specifically-named Graphite or OpenType feature, specifically selected in the text via some markup language.

3.2 Ad-hoc Markup Languages

The second option is to supplement the encoding of these variant glyphs at the font level with the use of an ad-hoc markup language. For example, an XML-like approach to encoding the word моє́ю would look something like this:

м<variant num=”1”>о</variant>е́ю

A rendering system could be used to select the necessary glyph form by "turning on" a specific font feature or replacing the base glyph with another glyph mapped in the Private Use Area. Such an approach would supplement any markup used to store non-textual material in text-only data – for example, data about color, alignment, pagination, font face, font size, and so forth. However, there is a fundamental difference between the storage of color and font face information and the encoding of glyph variants; namely, the color or font information is an attribute of layout while the glyph variants are a part of the text itself. Color, outside of a limited number of writing systems (for example, Znamenny and Byzantine Music notations), while significant in the layout, does not provide any grammatical or orthographic information.

Essentially, the use of markup language for encoding glyph variants constitutes a return to 8-bit codepages and *ad hoc* markup languages such as the HIP standard described in passing below. It suffers from a number of problems. Namely, it requires the use of a processor (interpreter or renderer) to preprocess a text and covert it from computer code into a human-readable appearance. Coupled with the fact that the display of these glyph variants in the human-readable form would still be contingent on the use of OpenType or Graphite aware fonts and applications (or on the encoding of glyphs in the Private Use Area), this approach would again restrict the exchange of text data by limiting the end user to a limited set of applications that both support OpenType and can be scripted (for example, XƎTEX or a web browser) or that can allow the installation of add-ons that can perform conversion between the markup language and some private-use encoding scheme (for example, Microsoft Office or LibreOffice). Indeed, this would be a step backwards from the existing HIP standard, because at least the HIP implementation already provides a set of tools for converting HIP-encoded texts into an ad-hoc 8-bit encoding whereas in this standard, such tools would have to be created once again.

3.3 Encoding additional glyphs

Another approach is to encode the glyph variants as standalone characters in Unicode. Despite the fact that "Unicode encodes characters, not glyphs", the standard does include certain characters that arguably are glyph variants and not standalone characters. One such example from the Cyrillic block is the glyph encoded at U+A641 and called Cyrillic Small Letter Zemlya; it could be argued that this is not a letter at all, since it consistently appears in both printed and manuscript texts as a variant of U+0437, Cyrillic Small Letter Ze.

However, encoding glyph variants as separate characters at their own codepoints raises a number of practical problems. First, none of the proposed variant glyphs has a capitalized analog. This means that the capitalized analog for each of these glyph variants is the capitalized analog of the base glyph. Yet any capitalization function must be bijective (invertible); this means that we must not only encode the glyph variants but also encode capitalized forms that do not differ visually from the capitalized form of the base glyph. Clearly, such an approach would lead to confusion.

Encoding glyph variants as separate characters leads to a second implementability issue. Namely, the contextual rules that govern the selection of a glyph variant need not be the same in different locales or typefaces. That is to say, the same text may be rendered in different ways depending on the specified contextual rules. If instead of relying on contextual rules that can be overwritten by the use of variation sequences, we allow variant forms to be "hard-coded" at the codepoint level, the interoperability of a text is lost. Put differently, the same text occurring in two different locales not only has a different visual appearance but also has a different underlying encoding. This leads to problems for search and matching, string comparison, and similar operations, although it is foreseeable that these problems could be alleviated by an appropriate collation tailoring. Nonetheless, the approach creates more problems for the end user than is desirable. This would doubtlessly lead to incorrect results among those developers not aware of the intricacies of the Slavonic writing system; at a minimum, these issues would need to be adequately addressed in the annotations to the characters and / or Unicode documentation.

4 Church Slavonic

Church Slavonic is a highly codified, living, literary language used by the Slavs. Presently, various recensions of Church Slavonic are used by Slavic Orthodox Churches, such as the Russian Orthodox Church, and by Slavic Byzantine-Rite Catholic Churches. Historically, the language was used not only for liturgical texts and religious literature but also for secular academic literature, such as grammars, lexicons, and even astronomical treatises. This proposal will only focus on Church Slavonic texts printed in the Cyrillic alphabet. The issue of properly encoding the texts printed in Glagolitic will be considered in a separate proposal.

4.1 Recensions of Church Slavonic

With the advent of the printing press, Church Slavonic developed a number of printing traditions that we term "recensions." These recensions have somewhat different orthographic and grammatical rules, especially when they apply to glyph variants. We identify three recensions: that of South and West Slavic Incunabula, early Polu-ustav (semi-uncial) type; and Synodal Polu-ustav type. The last of these, Synodal type, is the recension used since the early 1700's and up to today in the liturgical books of the Russian Orthodox Church. Early Polu-ustav reflects the printing traditions of Moscow and Lithuania in the 17th Century; the Muscovite version of Polu-ustav continues to be imitated in the printed books of the Old Ritualists – those Russian Orthodox who did not accept the reforms of Patriarch Nikon. Incunabula are the first set of Slavonic books printed in the West Slavic and South Slavic lands, particularly the editions of Schweipolt Fiol in Cracow (c. 1500); the editions of Božidar Vuković, whose 1517 *Sluzhebnik* opened the work of a Serbian press in Venice; and those by Francysk Skaryna, who between 1517 and 1519 printed the Bible in 22 volumes in Prague. Each printing tradition has somewhat different rules that govern the use of glyph variants.

4.2 Existing Encoding Models

In addition to other, architectural, considerations, Cyrillic glyph variants need to be encoded in Unicode to maintain backward-compatibility with existing encoding standards and thus to assure that text can be correctly converted from these standards to Unicode. Two such legacy encoding models are widely used – the Hyperinvariant Presentation (HIP) and the so-called Belgrade standard. The HIP standard uses an 8-bit encoding and an ad-hoc markup language to represent Church Slavonic text using an expansion of the CP1251 Cyrillic codepage. Since the HIP standard includes most of the variant forms discussed in this proposal, the absence of a standardized approach to encoding these variants in Unicode has made the full conversion of text from HIP to Unicode impossible. This has led to adverse affects for the community of Church Slavonic scholars and users, because no universal standard for text and software applications currently exists.

If some characters or variants are not available in the Unicode standard, users or developers may choose to encode them in the Private Use Area (PUA). One such approach is the Belgrade standard discussed by Kostić et al. (2009), which uses the PUA to encode Slavonic text. The problem with the use of the PUA is that it is, by its very definition, private; thus, any encoding scheme that places characters in the PUA is not suitable for working with text outside of a closed software-platform environment. The PUA should only be used to encode entities used internally by software, but not to encode text meant for distribution to a wide audience. In summary, there exists a need to standardize the encoding of these glyph variants in Unicode in order to achieve compatibility with the HIP and various private-use schemes. A full transition of the user community to Unicode will not be possible until these variants are added to the repertoire of Unicode. In listing the glyph variants in the Table below we also provide their analogs in HIP and the Belgrade Standard for reference.

5 Church Slavonic Glyph Variants

This section lists the glyph variant forms that have been identified based on an exhaustive study of modern and early Church Slavonic printed texts; that is, of printed texts of the Synodal, Poluustav, and Incunabula recensions. We note that for purposes of this proposal, we are not concerned with manuscripts, except as a point of reference to support the origins of various typographical conventions. All of the proposed glyph variants are summarized in Table 1.

Table 1: Table of Proposed Variation Sequences

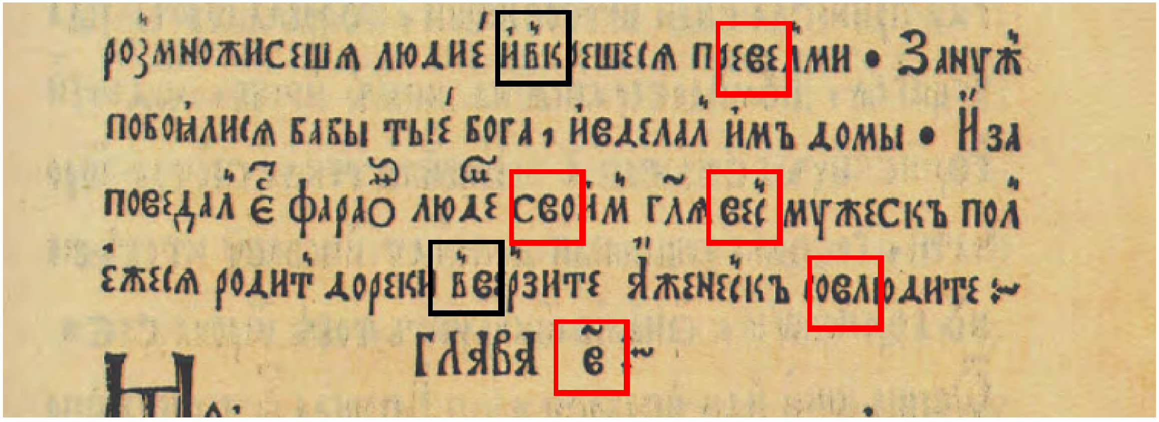
Sequence	Glyph	Name	HIP	Belgrade
U+0432 U+FE00	в︀	CYRILLIC SMALL LETTER VE VARIANT-1 ROUNDED VEDI	-	-
U+0432 U+FE0F	в️	CYRILLIC SMALL LETTER VE BASE FORM	в	E053
U+0434 U+FE00	д︀	CYRILLIC SMALL LETTER DE VARIANT-1 LONG-LEGGED DOBRO	<д>	-
U+0434 U+FE0F	д️	CYRILLIC SMALL LETTER DE BASE FORM	д	E055
U+043E U+FE00	о︀	CYRILLIC SMALL LETTER O VARIANT-1 NARROW ON	<о_>	E069
U+043E U+FE0F	о️	CYRILLIC SMALL LETTER O BASE FORM	о	E06A
U+0441 U+FE00	с︀	CYRILLIC SMALL LETTER ES VARIANT-1 WIDE SLOVO	<с>	E167
U+0441 U+FE0F	с️	CYRILLIC SMALL LETTER ES BASE FORM	с	E06F
U+0442 U+FE00	т︀	CYRILLIC SMALL LETTER TE VARIANT-1 TALL TVERDO	<т>	-
U+0442 U+FE01	т︁	CYRILLIC SMALL LETTER TE VARIANT-2 OLD-STYLE TVERDO	<|т|>	-
U+0442 U+FE0F	т️	CYRILLIC SMALL LETTER TE BASE FORM	т	E070
U+044A U+FE00	ъ︀	CYRILLIC SMALL LETTER HARD SIGN VARIANT-1 TALL HARD SIGN	<ъ>	-
U+044A U+FE0F	ъ️	CYRILLIC SMALL LETTER HARD SIGN BASE FORM	ъ	E080
U+0463 U+FE00	ѣ︀	CYRILLIC SMALL LETTER YAT VARIANT-1 TALL YAT	<jь>	-
U+0463 U+FE0F	ѣ️	CYRILLIC SMALL LETTER YAT BASE FORM	jь	E086
U+A64B U+FE00	ꙋ︀	CYRILLIC SMALL LETTER MONOGRAPH UK VARIANT-1 CHECKMARK-SHAPED UK	<ov>	-
U+A64B U+FE0F	ꙋ️	CYRILLIC SMALL LETTER MONOGRAPH UK BASE FORM	у	E072

5.1 Cyrillic Letter Ve Variant

This is a variant form of the Cyrillic Small Letter Ve (0432), known as the "round form" for its characteristic shape. This form appears in incunabula of a West Slavic provenance as well as in later Poluustav printed texts of a Lithuanian or Kievan provenance. In Figure 1 we present an example from the Bible of Francysk Skaryna, printed in Prague circa 1519. In this particular example, the variant form is used whenever the letter Ve does not take a diacritical mark (combining letter, titlo, or payerok) and the base form is used whenever the letter Ve occurs with a combining mark. However, observe that when the letter Ve occurs under a titlo not in a contraction but as the numeral 2, the variant form is used. Thus, the variant form is a contextual glyph variant that sometimes is used outside of context.

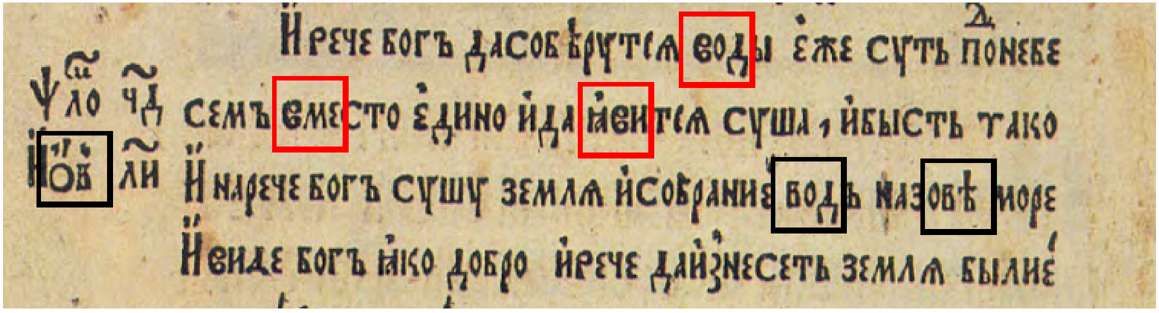
While this usage is the norm in Skaryna's edition of Exodus, it does not hold elsewhere in his Bible. In some places, Skaryna's usage of the two glyph variants appears to be haphazard and even whimsical, as can be seen from Figure 2. There has yet to be a critical study of Skaryna's text and so the extent to which he followed any given rules and the extent to which his usage influenced others has yet to be established. In addition to its use in the Bible of Skaryna, the variant form of Ve also occurs in the *Ostrog Bible* published by Ivan Fedorov in Lithuania in 1581 and other sources. Based on our study of the sources, however, it is clear that the usage of this glyph cannot be predicted algorithmically and thus must be handled via a variation sequence.

Figure 1: Typical Cyrillic Small Letter Ve (boxed in black) and variant form (boxed in red). Source: Bible printed by Francysk Skaryna, Prague, circa 1519.



розмножисеша людие и вкреѣшесѧ превелми • Занꙋж
побоѧлисѧ бабы тые бога, и сделал имъ домы • И за
повелал е фараѡ люде своим глѧ всѣ мꙋжескъ пол
ежесѧ родит до реки вверзите ѧженескъ совлюдите
главѧ е

Figure 2: Typical Cyrillic Small Letter Ve (boxed in black) and variant form (boxed in red). Source: Bible printed by Francysk Skaryna, Prague, circa 1519.

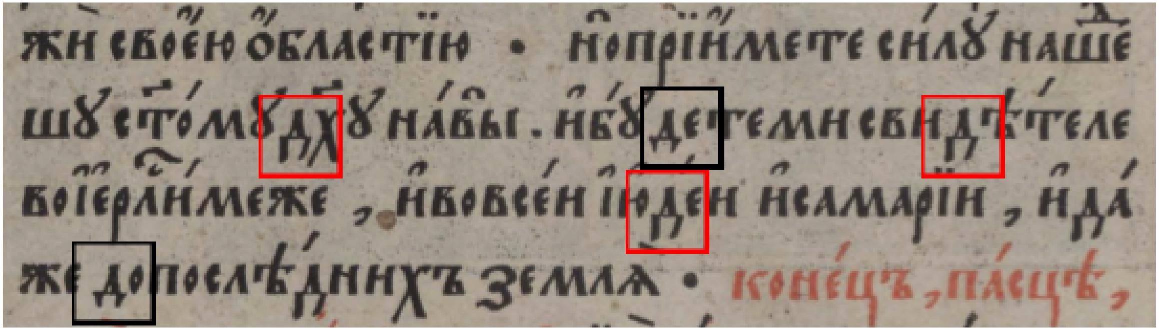


Ѱло чд
Йов ли
И рече богъ да соберꙋтсѧ воды еже сꙋть понебе
семъ вместо едино и да ꙗвитсѧ сꙋша, и бысть тако
И нарече богъ сꙋшꙋ землѧ и собрание вод назовѣ море
И виде богъ ѧко добро и рече да изнесеть землѧ былие

5.2 Cyrillic Letter De Variant

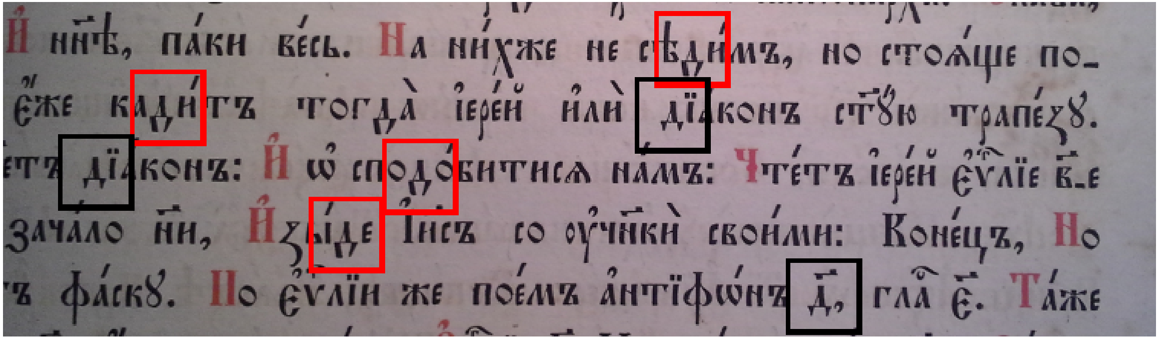
This is a variant of the Cyrillic Small Letter De (0434), known as the "Long-legged Dobro". In the

Figure 3: Typical Cyrillic Small Letter De (boxed in black) and variant form (boxed in red). Note that both forms can occur in initial or medial positions. Source: *Apostol*, Moscow: Tipografia of Ivan Fedorov, 1564.



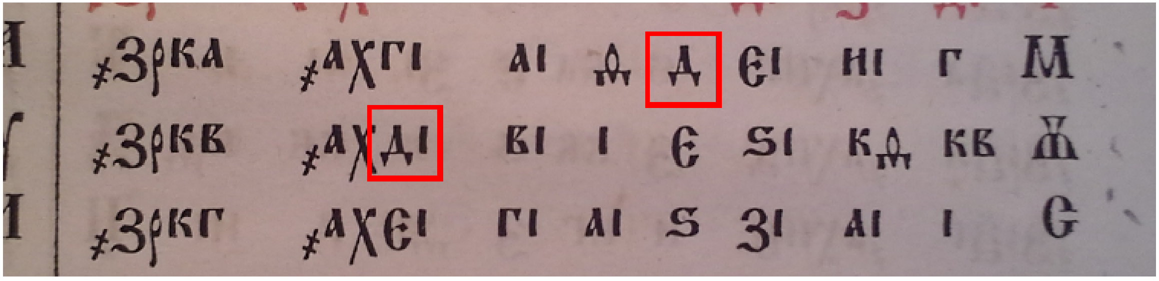
жи своєю ѡбластїю • но прїимете силꙋ нашє
шꙋ стомꙋ дꙋхꙋ на вы. и бꙋдете ми свидѣтеле
во ієрлимѣ же, и во всеи іюдеи и самарїи, и да
же до послѣднихъ земля • конецъ, пасцѣ,

Figure 4: Typical Cyrillic Small Letter De (boxed in black) and variant form (boxed in red). Note that the variant occurs in medial position while the base form in initial position but only the base form is used for numerals. Source: *Typicon*, Kiev: Lavra of the Kiev Caves, 1893.



нинѣ, паки весь. На нихже не сѣдимъ, но стоѧще по-
ѥже кадитъ тогда ієрей или дїаконъ стꙋю трапезꙋ.
етъ дїаконъ: И ѡ сподобитисѧ намъ: Чтетъ ієрей єѵлїе б҃е
зачало ни, Изыде Інсъ со ѹчнки своими: Конецъ, По
ъ фаскꙋ. По єѵлїи же поемъ антїфѡнъ д҃, гла є҃. Таже

Figure 5: Typical Cyrillic Small Letter De (boxed in black) used for numerals. Source: *Typicon*, Kiev: Lavra of the Kiev Caves, 1893.



І	҂зрка	҂ахгі	аі	ѧ	д	єі	иі	г	М
Г	҂зркв	҂ахді	ві	і	є	ѕі	кѧ	кв	Ѫ
І	҂зркг	҂ахєі	гі	аі	ѕ	зі	аі	і	Є

manuscript tradition of the Poluustav era, the base form was written in the initial position while this variant was written in the medial or final positions. This convention was carried over to the Lithuanian print tradition, and can be observed in the printed text of the Statutes of Lithuania (Karsky, 1979, p. 186). However, unlike scribes, typesetters began to use both forms indiscriminately; this was particularly the case in Moscow, where the base form was used when the amount of vertical space between lines of text was limited or a collision needed to be avoided with diacritical marks on the line below. Figure 3 presents the typical usage of both the base form and the variant form in a text of Muscovite provenance. The usage of the two glyphs is completely haphazard and cannot be predicted algorithmically. This usage can still be observed today in texts published by Old Ritualists, who have maintained a print tradition that mimics the older Muscovite type forms. Eventually, the variant form completely fell out of use in Muscovite typography and, with rare exceptions, it does not occur in Synodal texts of a Muscovite origin.

However, the variant form continued to be used extensively in Synodal editions of a Kievan provenance. Figure 4 presents an example from a Synodal recension book published in Kiev where it can be clearly observed that the base form is used in initial position and the variant glyph form is used in medial position, in keeping with earlier rules of usage. However, whenever the letter De occurs as part of a numeral, only the base glyph form is used. This become even clearer from looking at Figure 5, which presents a calendrical chart out of the same book. It can be observed that, when it occurs as part of a numeral, the letter De is encountered only in the base glyph form, regardless of position. Since in charts of numbers, the titlo used to indicate that the letters form a numeral is often omitted (as is the case in this example), it is impossible to predict algorithmically that the group of letters constitutes a numeral and thus impossible to use contextual rules to select the base glyph form.

In summary, the letter De variant is a contextual glyph form that often occurs outside of context.

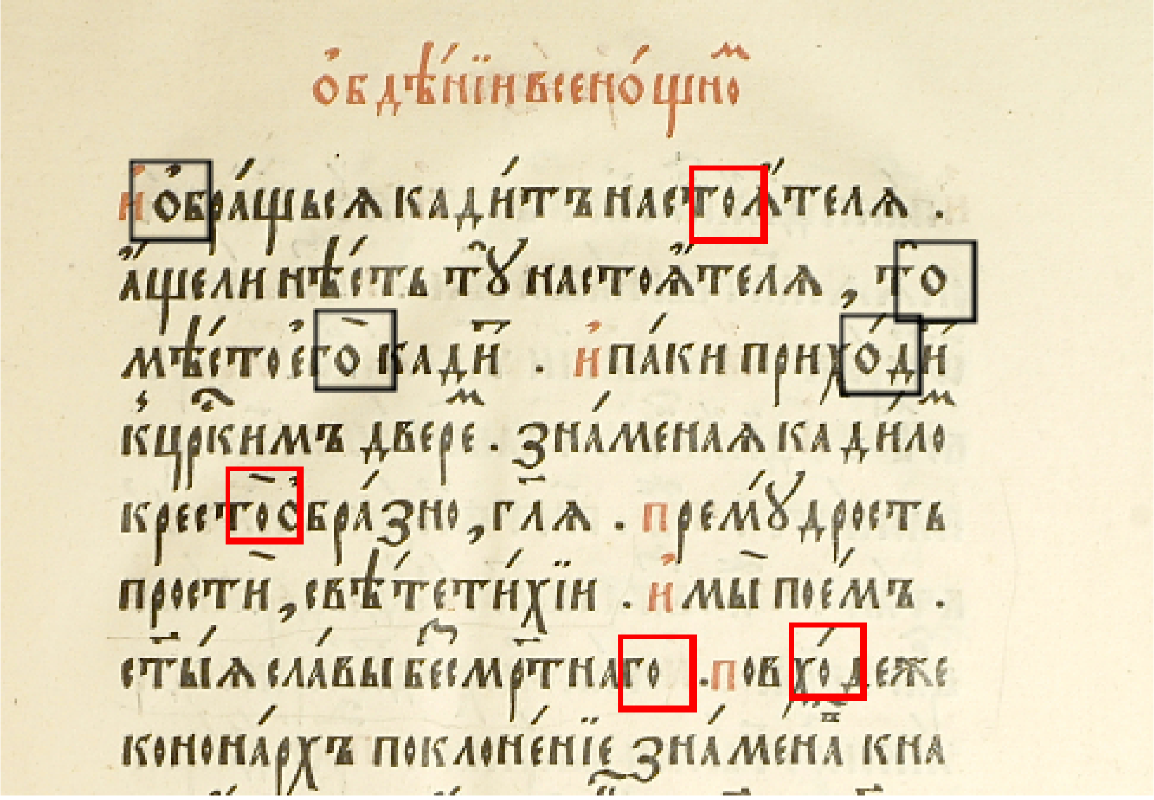
5.3 Small Letter O Variant

This is a variant form of the Cyrillic Small Letter O (043E), known as the "Narrow On". This form is widely used in Slavonic typography of all recensions. In the earliest Poluustav printed texts, rules governing the usage of this glyph variant were not fixed, and so this glyph form may be found both in the medial and the final positions and may be either accented or unaccented. This can be observed in Figure 6. In later printed editions, the usage stabilized and the narrow form came to be used whenever the letter O does not take an accent, while the base form was used in the accented position. This is true of modern texts printed by the Russian Old Ritualists. However, this usage was not always adhered to strictly, as can be observed from Figure 7. Thus, this glyph is a contextual variant that often occurs outside of context.

In addition to the base form, Unicode includes the wide form of the Letter O, called "Round Omega" (047B). This wide form is used in Church Slavonic orthography in a very specific circumstance: only in the initial position, for example, in the word ѻ҆ц҃ъ (*father*), or, when in the medial position, as the initial letter of a stem in a compound, as in the word пра́ѻц҃ъ (*forefather*, *ancestor*). Since it has a specific grammatical function – to indicate the first letter of a root that starts with о – this form should not be used to encode the base form of the Letter O (043E). We can observe from Figure 7 that all three forms of the Letter O (043E, 047B and the Variant Form) may occur in a typeface and all may be either accented or unaccented. It follows that it would not be correct to use the "Round Omega" to encode the base form of the Cyrillic Letter O and to use the codepoint of the Cyrillic Letter O to encode the narrow variant.

In modern Synodal typography, the variant form of the letter O is encountered extremely rarely, only as an apparent space-saving device; here it would need to be accessed every time explicitly via a variation selector.

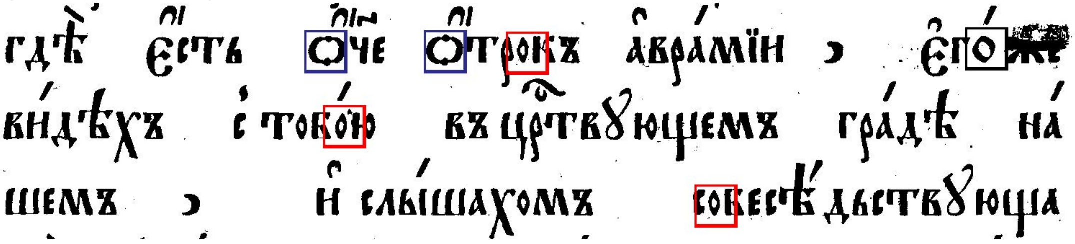
Figure 6: Typical Cyrillic Small Letter O (boxed in black) and Variant Form (boxed in red). Note that both forms can occur with or without an accent. Source: *Oko Tserkovnoye*, Moscow: Pechatnyi Dvor, 1610.



о бдѣнїи всенощно

И обращьсѧ кадитъ настоѧтелѧ.
аще ли нѣсть тꙋ настоѧтелѧ, то
мѣсто его кади. И паки приходи
к црскимъ двере. знаменаѧ кадило
крестообразно, глѧ. Премꙋдрость
прости, свѣтетихїи. И мы поемъ.
стыѧ славы безмртнаго. По обходе же
кононархъ поклоненїе знамена кна

Figure 7: Note the use of three forms of the letter O – the typical Cyrillic Small Letter O (boxed in black), the Cyrillic Letter Round Omega (boxed in indigo) and the Variant Form (boxed in red). Source: *Prolog*, Moscow: Tipografia of the Moscow Old Believers, 1915.

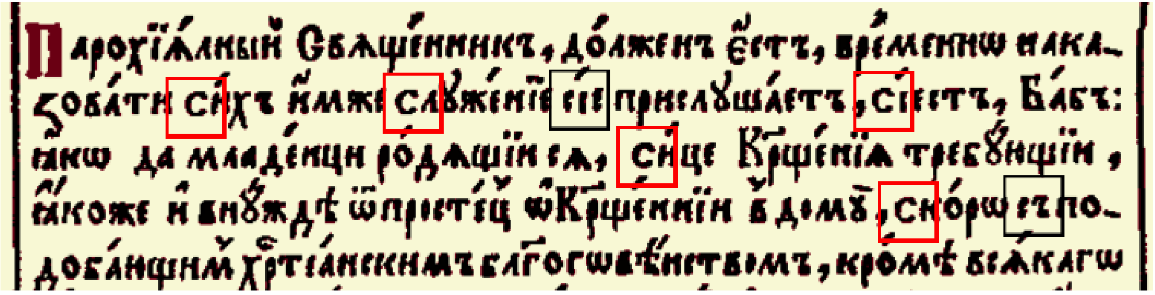


гдѣ есть ѻче ѻтрокъ авраамїи : его же
видѣхъ с тобою въ црствꙋющемъ градѣ на
шемъ : и слышахомъ бобесѣдствꙋюща

Finally, both in Poluustav and in Synodal recension texts, the narrow form of the glyph occurs as the first glyph of the digraph letter ᲂу. In fact, writing оу instead of ᲂу is generally not correct.[1] Unicode had initially encoded the digraph Uk as a standalone character (0479). However, the typographic tradition strongly suggests that it is properly treated as two glyphs; for example, when in a text the initial letter of a paragraph is set in red type, it is typical for only the ᲂ glyph to be set in red and not the entire ᲂу digraph. Likewise, the capitalized form of the digraph may be either Оу or ОУ, depending on the context. Thus, the codepoints 0478 and 0479 should be considered deprecated and the digraph ᲂу is most properly encoded as 043E 0443. This scheme demonstrates clearly that the narrow On is in fact a contextual variant: in the context of 043E 0443, the letter O is always displayed as the narrow variant, unless in very rare instances (such as in this paragraph, see above) it is necessary to override the contextual behavior. The overriding action takes place by explicitly selecting the base glyph.

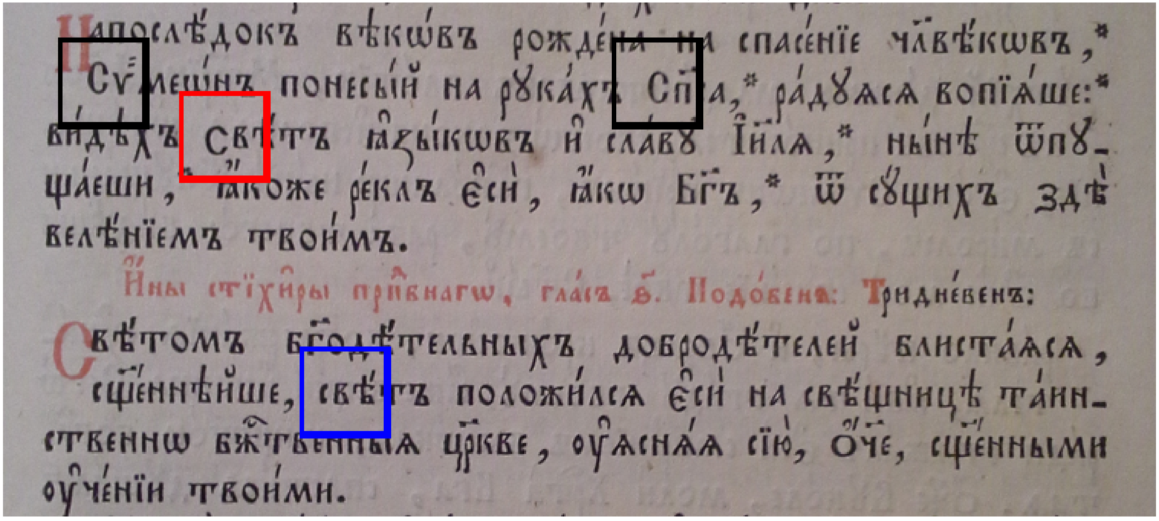
5.4 Small Letter Es Variant

Figure 8: Typical Cyrillic Small Letter Es (boxed in black) and variant form (boxed in red). Source: *Trebnik* of Metropolitan Peter (Mohyla), Kiev: Lavra of the Kiev Caves, 1646.



Парохїа́лный Сщ҃енникъ, до́лженъ є҆́стъ, вре́меннѡ нака-
зова́ти сн҃хъ и҆́мже слꙋже́нїе є҆гѡ̀ прислꙋша́етъ, сїе є҆́стъ, Ба́бъ:
ꙗ҆кѡ да младе́нцы роди́вшїисѧ, сн҃це крще́нїѧ тре́бꙋющїи,
ꙗ҆́коже и҆ бꙋ́дꙋщїе ѿ при́чтїѧ ѿ крще́нїи в дом̾ꙋ, скорѡ є҆гѡ по-
доба́ющимъ хрс҇тїанскимъ бл҃гоче́стїемъ, кромѣ̀ всѧ́кагѡ

Figure 9: Typical Cyrillic Small Letter Es (boxed in blue) and variant form (boxed in red). The capital form has been boxed in black. Source: *Menaion* for February, Kiev: Lavra of the Kiev Caves, 1893.



И҆ на послѣ́докъ вѣ́кѡвъ рожде́нна на спасе́нїе чл҃вѣ́кѡвъ,*
Сѵ́меѡнъ понесы́й на рꙋка́хъ Сп҃са,* ра́дꙋѧсѧ вопїѧ́ше:*
ви́дѣхъ свѣ́тъ ѧ҆зы́кѡвъ и҆ сла́вꙋ І҆и҃лѧ,* нн҃ѣ ѿпꙋ-
ща́еши, ꙗ҆́коже ре́клъ є҆сѝ, ꙗ҆́кѡ Бг҃ъ,* ѿ сꙋ́щихъ здѣ̀
велѣ́нїемъ твои́мъ.
И҆ны̑ стїхи́ры прп҃бнагѡ, гла́съ в҃. Подо́бенъ: Тридне́венъ:
Свѣ́томъ бг҃одѣ́тельныхъ добродѣ́телей блиста́ѧсѧ,
сщ҃еннѣ́йше, свѣ́тъ положи́лсѧ є҆сѝ на свѣ́щницѣ та́ин-
ственнѡ бж҃ственныѧ цр҃кве, ᲂу҆ѧсна́ѧ сїю̀, ѻ҆́ч҃е, сщ҃енными
ᲂу҆че́нїи твои́ми.

This is a variant form of the Cyrillic Small Letter Es (0441), known as the "Wide Slovo". This form is only encountered in initial position and only in texts of a Kievan provenance. In Kievan texts of the Synodal recension – that is, modern liturgical texts of the Russian Orthodox Church – this variant

[1] But it does occur in some publications, notably in the 1619 *Grammar* of Meletius Smotrytsky. This existence of such exceptions bolsters the need to encode variation sequences in order to override contextual substitution rules.

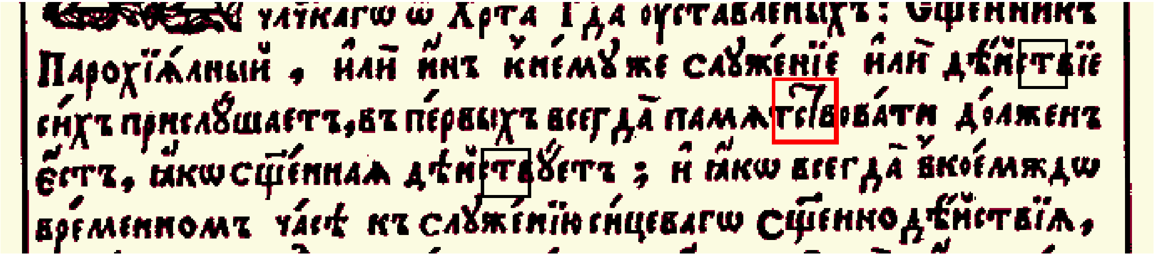
form is used in words that refer to the Divinity but are not divine names (*nomina sacra*). This can be clearly seen from Figure 9. Observe that the variant form (boxed in red) is used as the initial letter of the word свѣ́тъ (light) when it refers to Christ ("light of the Gentiles", an allusion to Luke 2:32). On the same page, we observe the base form of the letter used in the same word свѣ́тъ (light) when it refers to a saint ("light upon a candlestick", an allusion to Matthew 5:15). Thus, the variant is used in the first example simply to distinguish that the word light in this context refers to Christ. Observe also that the variant form **is not** a capital form of the letter Cyrillic Es, since the capital form may also be seen on this page in the word Сѷмеѡ́нъ (Symeon), a proper name, and in the word Сп҃съ (Savior), a *nomen sacrum*, both boxed in black.

In earlier printed texts of the Poluustav era, the typographical and orthographic rules were less rigid, but the same general pattern of usage may be observed. Figure 8 presents an example from the Trebnik (Euchologion) of Metropolitan Peter (Mohyla), a monumental 17th Century text that is still important both as a practical reference for clergy and as a fundamental primary source for the study of the development of Eastern Orthodox ritual. In this text, both forms of the letter Es are encountered (as well as the capitalized form), though the pattern of usage is less clear. The base form appears to be used in conjunctions and other less important words while the variant form is used for nouns. With regular frequency, the demonstrative pronoun се́й (this one) and its declined forms are written with the base form when they refer to an object or concept, and written with the variant form when they refer to a person.

It can be surmised from this discussion that the variant form of the letter Es is a grammatical variant and is used to communicate grammatical or syntactical information out of a desire to make theological texts as precise as possible. As such, the usage of this variant cannot be predicted algorithmically and it must be specified explicitly by use of a variation selector.

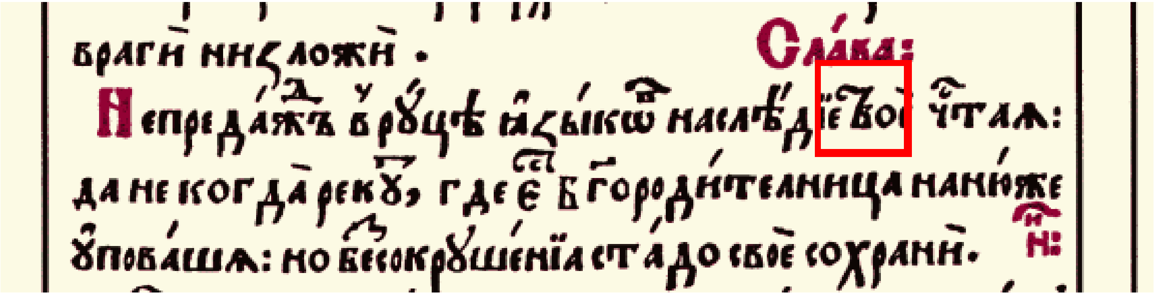
5.5 Small Letter Te Tall Variant

Figure 10: Typical Cyrillic Small Letter Te (boxed in black) and variant form (boxed in red). The variant form appears to be used as a space-saving device. Source: *Trebnik* of Metropolitan Peter (Mohyla), op. cit.



члчкагѡ ѿ Хрта Гда оуставленыхъ: Сщенникъ
Парохїѧльный, или инъ кнемꙋже слꙋженїе или дѣйствїе
сихъ прислꙋшаетъ, въ первыхъ всегда памѧтствовати долженъ
єстъ, ꙗкѡ сщенная дѣйствꙋетъ; и ꙗкѡ всегда вкоемждо
времененомъ часѣ къ слꙋженїю сицевагѡ сщеннодѣйствїѧ,

Figure 11: Cyrillic Ligature Te Ve (boxed in red). The variant Tall Tverdo occurs as the first component of this ligature. Source: *Trebnik* of Metropolitan Peter (Mohyla), op. cit.



враги низложи.
Слава:
Непредаждъ врꙋцѣ ꙗзыкѡмъ наслѣдїе твое чтꙋщаѧ:
да некогда рекꙋ, гдѣ є Бгъ городительница наших же
ꙋповаша: но безсокрꙋшенїа стадо свое сохрани.
н:

This is a variant form of the Cyrillic Small Letter Te (0442), known as the "Tall Tverdo". This glyph commonly occurs in Poluustav printed editions, where it is used as a space saving device. The typical usage is demonstrated in Figure 10. Liturgical texts are commonly laid out in justified alignment. In the earliest printed Church Slavonic texts, hyphenation (the transfer of a portion of a word to a new line) was avoided out of a desire for theologically precise language. When texts are typeset without any hyphenation, if the amount of inter-word spacing cannot be further reduced, the letter Te can be represented using this variant form to obtain additional kerning with the preceding letter.

In storing digital versions of these texts, it is important to preserve the use of this variant for two reasons. On the one hand, the need exists to use digital methods to study and analyze the typographic conventions used by early typographers. On the other hand, in producing either reprints of older texts (such as the *Trebnik* of Peter Mohyla) or new texts, there is often a need to imitate early typographic conventions; for example, many Old Ritualist texts are still printed without hyphenation, and thus use space-saving glyph variants.

In modern technologies, some functionality is provided via the *JSTF* table in OpenType, which can allow font developers to specify rules for selecting and positioning various glyphs in justified text. Some selection of space-saving glyph variants can be handled by specifying the appropriate substitution rules in the JSTF table. However, the need then immediately arises to override the automated substitution rules or to present space-saving glyph variants outside of context (for example, in documentation accompanying fonts and computer software). In addition, in a plain text setting, the text is stored independently of layout information, and so the glyph variant must be specified explicitly via the use of variation selectors.

In some instances, where even more space-saving compression is desired, adjoining letters Te and Ve are written in Slavonic texts as the ligatureт‍в; in these cases, the Tall Tverdo variant forms the first component of the Te-Ve ligature, as can be seen in Figure 11. The ligature components are joined by the use of the Zero Width Joiner (200D). In instances where the two components need to be displayed independently (as inт в), the variant form must again be specified explicitly via the use of variation selectors. (For the display of ligature components independently, Unicode recommends the use of the Zero Width Non-Joiner, but in this instance, it would provide the undesirable result т в, as the Te component would be displayed using the base glyph form).

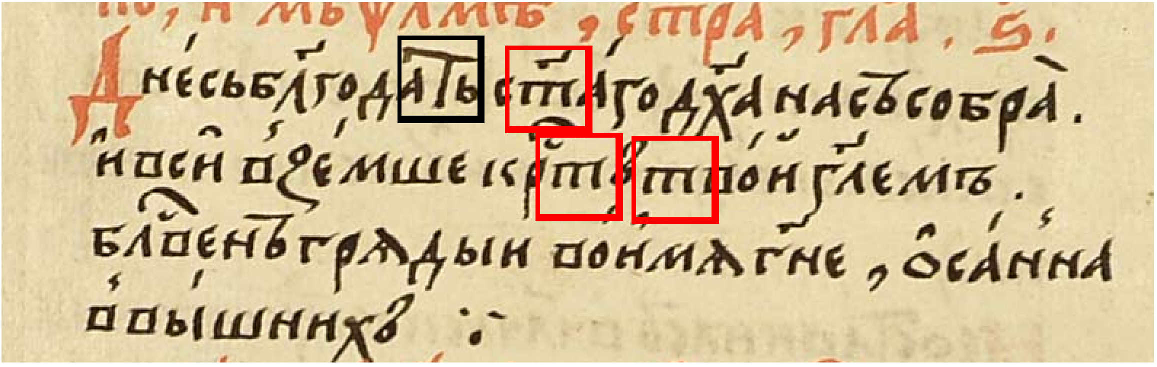
5.6 Small Letter Te Old-Style Variant

This variant of the Cyrillic Letter Te (0442), also called (in the HIP standard) the "three-masted tverdo" and written with all three vertical strokes touching the baseline (т), has a complex history. According to Karsky (1979, p. 198), in the 15th Century, this becomes the most prevalent form of the letter Te in Church Slavonic manuscripts. What later becomes the base form (or perhaps the tall form) is used in the manuscript tradition as a space-saving device. This is demonstrated in Figure 12. In printed editions of Church Slavonic texts, the base form (т) begins to dominate, and the old style form gradually drops out of usage. Some editions, particularly of Lithuanian or Ukrainian provenance, however, use both forms interchangeably, perhaps to imitate the manuscript tradition. We demonstrate an example of this usage in Figure 13.

On the other hand, the Old Style variant of the letter Te becomes the dominant form of the letter in printed Russian (vernacular) texts through the 19th Century, until it is replaced by the base form in more recent publications (see Figure 14 for examples of this usage).

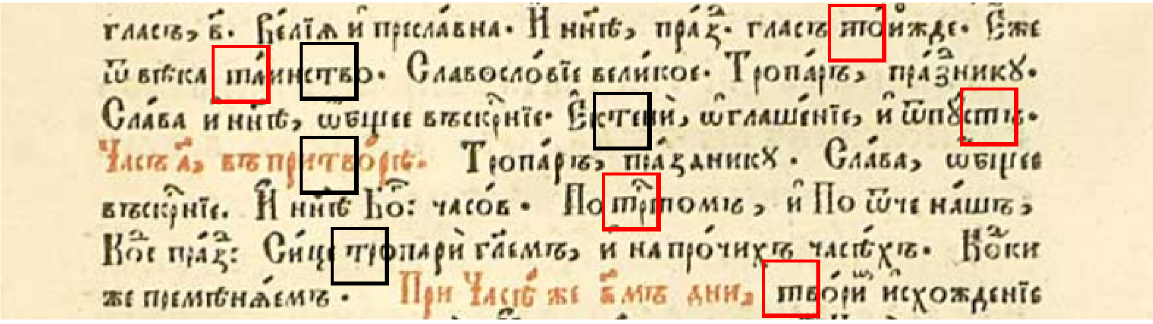
When Church Slavonic or vernacular Russian texts are encoded in Unicode, the default form of the letter Te should be provided at the font level. Thus, a Church Slavonic font imitating the orthographic traditions of older manuscripts should use the old style variant by default. In some instances, the base form needs to be selected explicitly, and this is accomplished via the use of a variation selector. Likewise, fonts designed to imitate the orthographic traditions of 19th Century Russian should use the

Figure 12: Typical Cyrillic Small Letter Te (boxed in black) and Old Style variant form (boxed in red). The variant form is the more prevalent form in this manuscript, and is used by default. Source: *Kanonnik*, a Poluustav manuscript written in 1616.



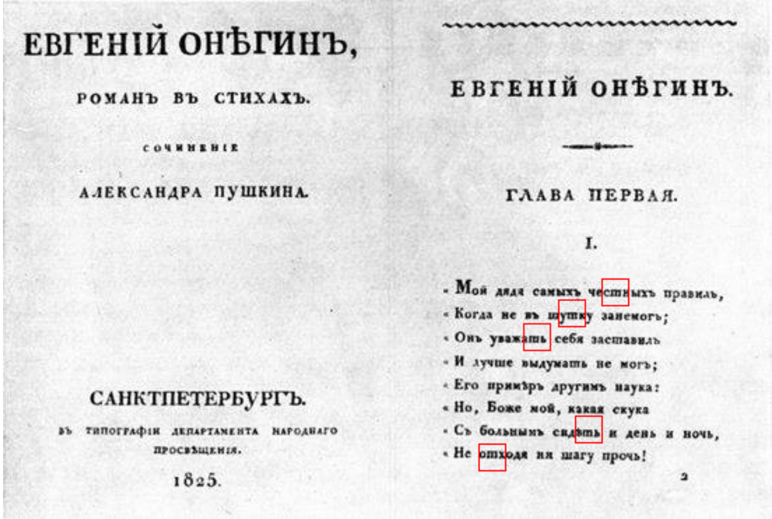
Днесь бл҃годать стаго дх҃а нас собра.
и вси вземше крⷭ҇т твой гл҃емъ.
бл҃нъ грѧдыи во имѧ гн҃е, ѡсанна
в вышних ⁛

Figure 13: Typical Cyrillic Small Letter Te (boxed in black) and Old Style variant form (boxed in red). Both forms are used interchangeably, and may occur either in initial or medial position. Source: *Flowery Triodion*, Lvov, 1642.



гласъ, в҃. Велїа и преславна. И нн҃ѣ, праз. гласъ тойже. Еже
ѿ вѣка таинство. Славословїе великое. Тропарь, празднику.
Слава и нн҃ѣ, общее воскресенїе. Ектенїи, ѡглашенїе, и ѿпустъ.
Часъ а҃, въ притворѣ. Тропарь, праздникꙋ. Слава, общее
воскресенїе. И нн҃ѣ Бо: часов. По тр҃помъ, и По Ѿче нашъ,
Кѡⷣ праз: Сиже тропари глемъ, и на прочихъ часѣхъ. Боки
же премѣнѧемъ. При Часѣ же а҃ мъ дни, твори исхожденїе

Figure 14: The Old-Style variant form of the letter Te (boxed in red) used in vernacular Russian. Source: a printing of Aleksandr Pushkin's *Eugene Onegin*, St Petersburg, 1825.



ЕВГЕНІЙ ОНѢГИНЪ,
РОМАНЪ ВЪ СТИХАХЪ.
СОЧИНЕНІЕ
АЛЕКСАНДРА ПУШКИНА.

САНКТПЕТЕРБУРГЪ.
ВЪ ТИПОГРАФІИ ДЕПАРТАМЕНТА НАРОДНАГО ПРОСВѢЩЕНІЯ.
1825.

ЕВГЕНІЙ ОНѢГИНЪ.

ГЛАВА ПЕРВАЯ.

I.

« Мой дядя самыхъ честныхъ правилъ,
« Когда не въ шутку занемогъ;
« Онъ уважать себя заставилъ
« И лучше выдумать не могъ;
« Его примѣръ другимъ наука:
« Но, Боже мой, какая скука
« Съ больнымъ сидѣть и день и ночь,
« Не отходя ни шагу прочь!

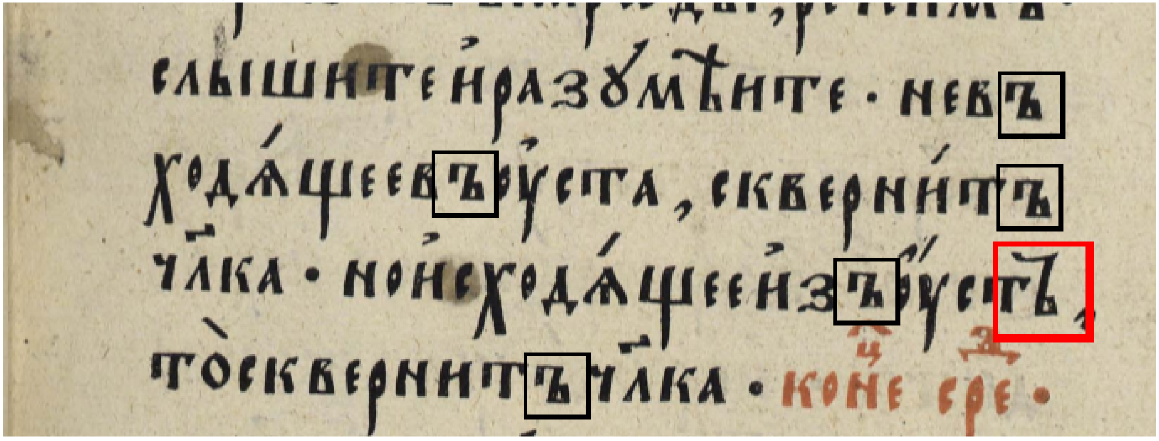
2

old style form by default, with the possibility of selecting the base form via the use of a variation selector. On the other hand, Church Slavonic fonts designed for Synodal-era texts and Cyrillic fonts designed for modern Russian should use the base form by default, with the possibility of selecting the variant form via the use of variation selectors. We propose to name this form the Old-Style form of Tverdo in the Unicode standard, since it is the form used in “old” Russian type. Alternatively, the name “Three-masted Tverdo” may also be used.

5.7 Small Hard Sign Variant

A number of early Church Slavonic publications (particularly, those by the Moscow Anonymous Typografia), use a variant of the Cyrillic Small Letter Hard Sign (044A). This variant form is used as a space saving device in texts with right-justified alignment because the left tip of the hard sign kerns above and over the preceding letter. This usage can be seen in Figure 15. The typographic tradition featuring this variant of the hard sign reflects the earlier manuscript tradition. In later publications, with the advent of hyphenation under the influence of Western typographical norms, this usage was largely forgotten. Nonetheless, for the reasons outlined above in our discussion of the Tall Te variant, this space-saving character needs to be preserved in the digital storage and reproduction of those texts where hyphenation is undesirable or inappropriate.

Figure 15: Typical Cyrillic Small Letter Hard Sign (boxed in black) and variant form (boxed in red). The variant form appears to be used as a space-saving device. Source: *Gospel Book* published by the Anonymous Typografia, 1553.

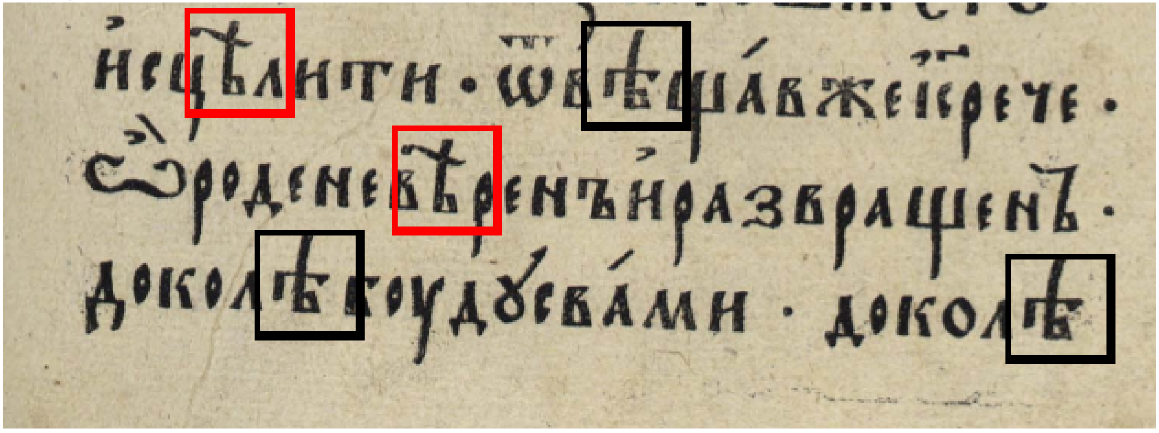


слꙑшите и҆ разꙋмѣите · не въ
хо́дѧщее въ ѹ́ста, скверни́тъ
чл҃ка · но и҆сходѧ́щее и҆зъ ѹ́стъ,
то̀ ескверни́тъ чл҃ка · коне ере ·

5.8 Small Yat Variant

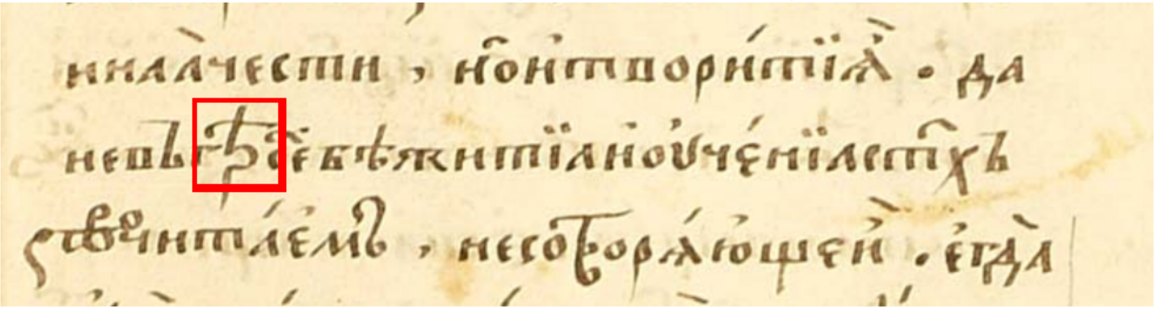
The texts that use the variant form of the Hard Sign also use a variant form of the Cyrillic Small Letter Yat (0463). This variant form is likewise used as a space-saving device, as can be seen in Figure 16. In the manuscript tradition of Slavonic Poluustav, the letters Re and Yat are sometimes written together as a ligature (ᲈ). This is demonstrated in Figure 17. When creating a digital version of these texts encoded in Unicode, the characters are combined into a ligature via the use of the Zero Width Joiner. As in the example above with the Te-Ve ligature, when the elements of the Re-Yat ligature need to be displayed independently (as in рѣ), the correct variant of the Yat needs to be selected explicitly via the use of a variation sequence.

Figure 16: Typical Cyrillic Small Letter Yat (boxed in black) and variant form (boxed in red). The variant form appears to be used as a space-saving device. Source: *Gospel Book* published by the Anonymous Typografia, 1553.



исцѣлити · ѿвѣщавже і҆с рече ·
ѡ родене вѣренъ и развращенъ ·
доколѣ боудоу свами · доколѣ

Figure 17: Example of the Re-Yat ligature. Source: *Izmaragd*, Moscow, 16th Century.



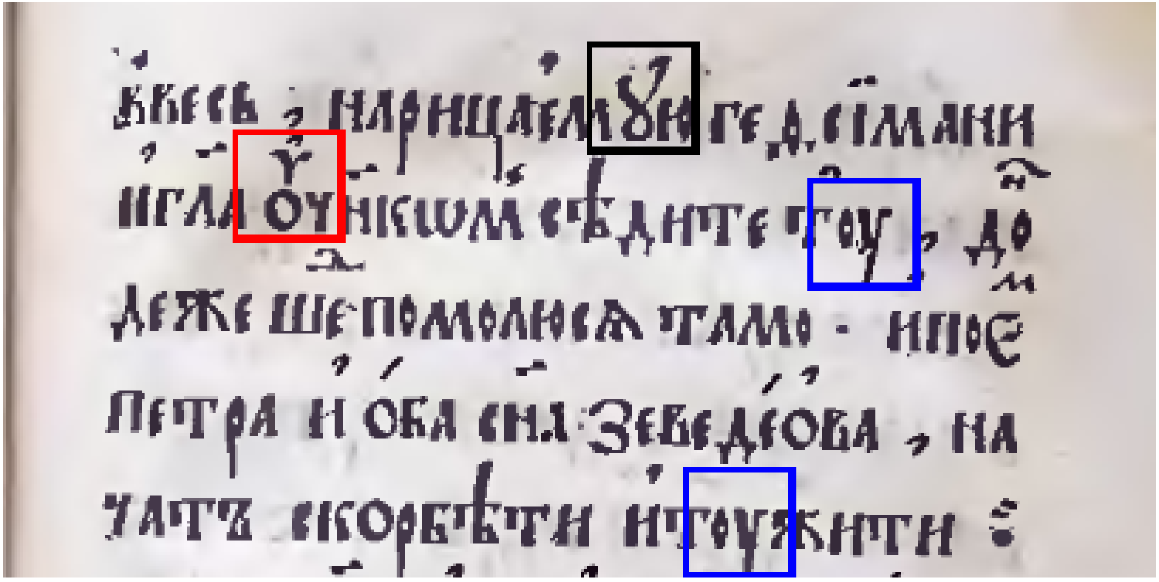
и на лчести , но и непоритиѧ · да
не ѡбрѣве бѣжнитиа и оучениа лестиѵхъ
своихъ тѣмъ , не собораꙗющеи · егда

5.9 Small Letter Monograph Uk Variant

In Synodal Church Slavonic, the orthography has been standardized and the digraph ѹ always occurs in the beginning of words while the monograph form ꙋ occurs in medial or final position (and never in initial position). However, such standardization is not the case in earlier recensions, including Poluustav Church Slavonic texts, where the monograph form and the digraph form of the letter are used more or less interchangeably. In some Poluustav editions, we find a variant of the Cyrillic Small Letter Monograph Uk (U+A64B) used when the monograph Uk takes a Psili (breathing mark). This variant form is called the “checkmark-shaped form” because of its characteristic appearance. Its evident purpose is to avoid complexities of positioning the breathing mark within the branches of the base form of the Uk character (as in ꙋ̓). In Figure 18 we show an example of the variant form used in initial position and in Figure 19 we show the same variant form used in medial and final positions. Note that in all of these examples, the base form and the digraph form ѹ occur in addition to the variant form.

While the selection of the variant form of the Uk character in these instances can take place at the font level via the use of advanced typographic features, instances may arise when the variant form needs to be used without a breathing mark. In such instances, the variant form must be selected explicitly via the use of a variation selector. In other instances, it may be required to turn off the default contextual substitution behavior of the font, in which case the base form of the letter needs to be selected explicitly with the use of a variation selector.

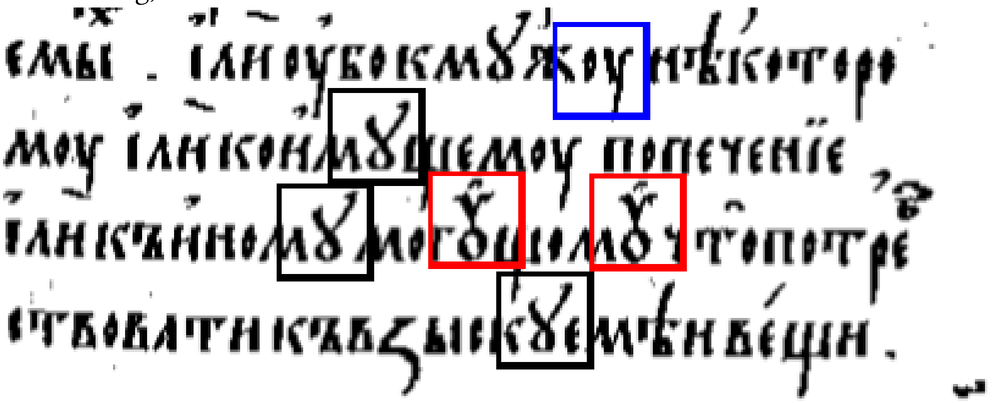
Figure 18: Typical Cyrillic Small Letter Monograph Uk (boxed in black) and variant form (boxed in red). Note also the use of the Cyrillic Small Letter U (as part of the digraph ѹ). Source: *Gospel Book* published in Vilnius, 1575.



6 Implementation

This section discusses the proposed implementation of Cyrillic variation sequences in Unicode. As was discussed above, in Poluustav typography, the Cyrillic Letter O usually occurs in the base form when it takes a diacritical mark and as the narrow variant when it occurs without any diacritical marks. Quite frequently, this default behavior must be overridden explicitly. We have then the following examples:

Figure 19: Typical Cyrillic Small Letter Monograph Uk (boxed in black) and variant form (boxed in red). Note also the use of the Cyrillic Small Letter U (as part of the digraph оу). Source: *Book on Fasting* published in Ostrog, 1594.



Render	Encoding	Comments
мо́й	U+043E U+0301	Base form occurs in context
мо︀́й	U+043E U+FE00 U+0301	Variant form selected explicitly
мой	U+043E	Variant form occurs in context
мо️й	U+043E U+FE0F	Base form selected explicitly (overriding contextual substitution)

These examples demonstrate that in most cases, the selection of the correct glyph form is relegated to the rendering system, which chooses the necessary glyph on the basis of contextual rules. In those instances when the contextual rules must be overwritten explicitly, the presence of the variation selector specifically forces the rendering system to select the desired glyph. This is a much more flexible approach, since the alternative approach – not relying on contextual substitution but explicitly selecting the narrow glyph whenever it occurs – would require the end user to enter the variation selector explicitly every time a narrow glyph variant is required. Given that the narrow glyph form of the On can account for up to 50% of the occurrences of the Letter O in Poluustav text, such an approach would be tedious.

Allowing the glyph selection to take place implicitly with occasional explicit overriding has one additional benefit. Consider the same example, but now in the Synodal recension of Church Slavonic, where the narrow On is used only on very rare occasions:

Render	Encoding	Comments
мо́й	U+043E U+0301	Base form displayed as expected
мо︀́й	U+043E U+FE00 U+0301	Variant form selected explicitly
мой	U+043E	Base form displayed as expected
мо️й	U+043E U+FE0F	Explicit selection of base form has no effect (no contextual rule to override)

Under this implementation, the underlying text in both examples is encoded identically. At the font level, the contextual substitution rules determine the selection of the desired glyphs for each locale – Synodal recension or Poluustav recension. If the user needs to select the appropriate glyph outside of context, this is done by explicitly invoking the variation selector.

6.1 Rationale for Encoding Base Form as Sequence

Relying on contextual rules to select the correct glyph form most of the time is contingent on the existence of a mechanism to override contextual substitution. We propose to define such a mechanism by explicitly encoding the base glyph form as a variation sequence.

Alternatively, one could contemplate the use of the Zero Width Non-Joiner (ZWNJ, U+200C) as a control character that overrides contextual substitution. This is somewhat problematic because of the currently defined behavior of the ZWNJ. The ZWNJ is presently used to “obstruct the normal ligature/cursive connection behavior” and is to “have the desired effect naturally for most fonts”, meaning that no special rules involving ZWNJ need to be written at the font level (Allen et al., 2012, p. 551). In other words, the sequence of characters [base char], ZWNJ is not treated as a ligature itself; rather, in the sequence of characters [base char], ZWNJ, [second char], the contextual rules between [base char] and [second char] are ignored. In the above examples, where the contextual substitution rules are defined on the basis of the presence of diacritical marks, the presence of ZWNJ would not only obstruct contextual substitution rules but also obstruct the correct positioning of the diacritical mark over the base letter. Thus, to use the ZWNJ to override contextual substitution in these examples is impossible.

If the Unicode Consortium chooses not to encode a variation sequence for the base variant form, the functionality of ZWNJ would need to redefined so that in fonts, the sequence [base char], ZWNJ is always treated as a ligature substitution, resulting in the base glyph form of the character. However, this may have undesirable adverse effects for other writing systems – such as Arabic or Devanagari – that rely on the ZWNJ for shaping behavior.

7 Summary

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

```
U+0432 U+FE00; VARIANT-1 ROUNDED VEDI; # CYRILLIC SMALL LETTER VE
U+0432 U+FE0F; BASE FORM; # CYRILLIC SMALL LETTER VE
U+0434 U+FE00; VARIANT-1 LONG-LEGGED DOBRO; # CYRILLIC SMALL LETTER
   DE
U+0434 U+FE0F; BASE FORM; #  CYRILLIC SMALL LETTER DE
U+043E U+FE00; VARIANT-1 NARROW ON; # CYRILLIC SMALL LETTER O
U+043E U+FE0F; BASE FORM; #  CYRILLIC SMALL LETTER O
U+0441 U+FE00; VARIANT-1 WIDE SLOVO; #  CYRILLIC SMALL LETTER ES
U+0441 U+FE0F; BASE FORM; #  CYRILLIC SMALL LETTER ES
U+0442 U+FE00; VARIANT-1 TALL TVERDO; #  CYRILLIC SMALL LETTER TE
U+0442 U+FE01; VARIANT-2 OLD-STYLE TVERDO; # CYRILLIC SMALL LETTER
   TE
U+0442 U+FE0F; BASE FORM; #  CYRILLIC SMALL LETTER TE
U+044A U+FE00; VARIANT-1 TALL HARD SIGN; # CYRILLIC SMALL LETTER
   HARD SIGN
U+044A U+FE0F; BASE FORM; # CYRILLIC SMALL LETTER HARD SIGN
U+0463 U+FE00; VARIANT-1 TALL YAT; # CYRILLIC SMALL LETTER YAT
U+0463 U+FE0F; BASE FORM; # CYRILLIC SMALL LETTER YAT
U+A64B U+FE00; VARIANT-1 CHECKMARK-SHAPED UK; # CYRILLIC LETTER
   MONOGRAPH UK
U+A64B U+FE0F; BASE FORM; # CYRILLIC SMALL LETTER MONOGRAPH UK
```

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.

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

Kostić, Z., V. Savić, et al. (2009). Standard of the old slavonic cyrillic script. In G. Jovanović, J. Grković-Major, Z. Kostić, and V. Savić (Eds.), *Standardization of the Old Church Slavonic Cyrillic Script and Its Registration in Unicode.* Serbian Academy of Sciences and Arts.

Pentzlin, K. (2011). Proposal to add variation sequences for latin and cyrillic letters. Document number: L2/10-280. Available online: `http://www.pentzlin.com/Variation-Sequences-Latin-Cyrillic3.pdf`.

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 use Standardized Variation Sequences to Encode Church Slavonic Glyph Variants***
2. Requester's name: *Aleksandr Andreev, Yuri Shardt and Nikita Simmons*
3. Requester type (Member body/Liaison/Individual contribution): *Individual contribution*
4. Submission date: *07/19/2013*
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: *Standardized Variants*
2. Number of characters in proposal: *17*
3. Proposed category (select one from below - see section 2.2 of P&P document):

A-Contemporary		B.1-Specialized (small collection)		B.2-Specialized (large collection)	*X*
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, aleksandr.andreev@gmail.com
   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 developed by Aleksandr Andreev, Yuri Shardt and Nikita Simmons. Licensed under GNU GPL and available from http://www.ponomar.net/
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*
   The proper use of variation sequences to encode these characters is discussed. See §6, Implementation

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? *YES*

Reference: *See Section 4*

4. The context of use for the proposed characters (type of use; common or rare) *common*

Reference: *See Section 5 for details on each character*

5. Are the proposed characters in current use by the user community? *YES*

If YES, where? Reference: *Used in the HIP Standard to encode digital versions of books & manuscripts*

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? *YES*

If YES, reference: *See Section 3 for rationale of encoding as Variation Sequences*

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? *N/A*

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? *N/A*

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 and Section 6*

Is a list of composite sequences and their corresponding glyph images (graphic symbols) provided? *YES*

If YES, reference: *See 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: