Ponomar Project
Slavonic Computing Initiative

Proposal to Encode Combining Glagolitic Letters in Unicode

Aleksandr Andreev*, Heinz Miklas, Yuri Shardt

Section 1. Introduction

Glagolitic, also known as "Glagolitsa", is an alphabetic writing system used to record Church Slavonic and other Slavic languages. Originating in the 9th century, it is the earliest known Slavonic alphabet. The creation of the alphabet is attributed to the younger of the teachers of the Slavs, St. Cyril.

Glagolitic writing may be found in mediæval manuscripts and in printed liturgical books, mostly of a Croatian origin. In Bulgaria, Glagolitic was gradually replaced by the Cyrillic alphabet, and this Cyrillic alphabet was subsequently used also by other Slavs. For its part, the Glagolitic script has been preserved by some communities in Croatia even up to the present. Extant Glagolitic texts are of enormous value to linguists, palæographers, and scholars of liturgy.

Support for Glagolitic in the Unicode standard is required for two purposes. First, contemporary specialists need to be able to typographically represent mediæval texts written in the Glagolitic script, both in printed matter (such as academic publications) and in an electronic format (for use with computer analysis, such as string comparison, wordlist generation and searching). To this end, computer fonts that contain the repertoire of Glagolitic characters must be created. Second, owing to the close relationship between the Cyrillic and Glagolitic writing systems, scholars have traditionally represented Glagolitic texts also in Cyrillic transcription. To facilitate the transliteration process, an encoding model that parallels the model for the Cyrillic script needs to be available for Glagolitic.

The base repertoire of Glagolitic characters has been included in the Unicode standard since version 4.0. Nonetheless, this repertoire is incomplete because it lacks combining Glagolitic letters. Such combining letters exist in the Glagolitic script and play a function that is analogous to their role in Cyrillic – that is, they are used in abbreviations that are either space saving devices (for example, commonly written words are often abbreviated) or in *nomina sacra*. For full support of the Glagolitic writing system in Unicode, as well as for proper interoperability between the implementations of the Glagolitic and Cyrillic scripts, we propose the encoding of these combining characters in an additional block entitled *Glagolitic Extended*.

Section 2. Proposed Characters

The following table contains examples of combining Glagolitic letters that occur in various Glagolitic manuscripts and in printed literature. We propose to encode the characters as one block, in the same codepoint order as the base Glagolitic letters encoded at U+2C00 and following. This allows for simple computer manipulation of Glagolitic characters, as well as leaving some encoding positions empty to be used in the unlikely instance that additional combining characters are discovered by researchers and need to be encoded. Note that since a Glagolitic Extension is not in the Roadmaps to Unicode, all of the indicated codepoints in this proposal are provisional codepoints in the Private Use Area (PUA).

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

Name	Codepoint	Appearance	Location in Sources
Combining Glagolitic Letter Azu	U+E000		*Srez.*, p. 42; *MissSin,* 13v18, 44r16-17, 45v20; *PsDem*, 27v6, 42r7, 118r5
Combining Glagolitic Letter Buki	U+E001		*Srez.*, p. 254; *MissSin,* 38r13, 49v20
Combining Glagolitic Letter Vede	U+E002		*Srez.*, p. 59; *MissSin*, 7r2, 22r10, 40v10, 52r17; *PsDem*, 104r8
Combining Glagolitic Letter Glagoli	U+E003		*Srez.*, p. 254; *PsDem,* 117r1
Combining Glagolitic Letter Dobro	U+E004		*Srez.*, p. 228; *MissSin*, 54v14; *PsDem*, 21v1, 42r8, 78r19, 131r9
Combining Glagolitic Letter Yestu	U+E005		*Srez.*, p. 59; *MissSin*, 29r20, 46r12
Combining Glagolitic Letter Zhivete	U+E006		*Srez.*, p. 84
Combining Glagolitic Letter Zemlja	U+E008		*EuchSinV*, 103r16m
Combining Glagolitic Letter Izhe	U+E009		Srez., p. 82
Combining Glagolitic Letter Initial Izhe	U+E00A		*PsSinV*, 177r18
Combining Glagolitic Letter I	U+E00B		Srezn., p. 82
Combining Glagolitic Letter Djervi	U+E00C		Srezn., p. 84; *MissSin*, 18v6
Combining Glagolitic Letter Kako	U+E00D		Srezn., p. 224; *PsDem*, 126r2
Combining Glagolitic Letter Ljudie	U+E00E		Srezn., p. 40, p. 42; *PsDem*, 5v1; *MissSin.*, 30v22, 51r15
Combinign Glagolitic Letter Myslite	U+E00F		Srezn., p. 224; *PsDem*, 50v15, 105r19, 113v3; *MissSin*, 13r18
Combining Glagolitic Letter Nashi	U+E010		Srezn., p. 59; *PsDem*, 21v3; *MissSin*, 20v23
Combining Glagolitic Letter Onu	U+E011		*PsDem*, 126r2.; *MissSin*, 43v14, 43v21-2
Combinign Glagolitic Letter Pokoji	U+E012		Srezn., p. 42, p. 248; *MissSin*, 13r18, 13v10
Combining Glagolitic Letter Ritsi	U+E013		Srezn., p. 42; *MissSin*, 35v15
Combining Glagolitic Letter Slovo	U+E014		Srezn., p. 228; *PsDem*, 126r24, 128v7, 126r24, 126v18
Combinign Glagolitic Letter Tvrido	U+E015		Srezn., p. 36, p. 42, 59; *PsDem*, 118r4; *MissSin*, 18v3, 17v12, 45v20, 46r19

Name	Codepoint	Appearance	Location in Sources
Combining Glagolitic Letter Uku	U+E016	◌	*MissSin*, 43v14
Combining Glagolitic Letter Fritu	U+E017	◌	*MissSin*, 25r13, 29r14, 33r23, 40r15, 22r10
Combining Glagolitic Letter Heru	U+E018	◌	*MissSin*, 19r12, 13v15, 17v14
Combining Glagolitic Letter Shta	U+E01B	◌	*MissSin*, 58r6, 72r/v8
Combining Glagolitic Letter Tsi	U+E01C	◌	Srezn., p. 59; *MissSin*, 19r12
Combining Glagolitic Letter Chrivi	U+E01D	◌	*MissSin*, 53r5, 45r9-10, 39v13
Combining Glagolitic Letter Sha	U+E01E	◌	Srezn., p. 224; *MissSin*, 21r12, 13r1, 17r18, 22v18, 24r23, 24(15)v8
Combining Glagolitic Letter Yeru	U+E01F	◌	*MissSin*, 18v6
Combinign Glagolitic Letter Yeri	U+E020	◌	Srezn., p. 42
Combining Glagolitic Letter Yati	U+E021	◌	Srezn., p. 248; *PsDem*, 83r16; *MissSin*, 40r15
Combining Glagolitic Letter Yu	U+E023	◌	Srezn., p. 254
Combining Glagolitic Letter Small Yus	U+E024	◌	*EuchSinV*, 32v17, 51r11
Combining Glagolitic Letter Yo	U+E026	◌	Does not exist as a single character, but is a component of U+E029.
Combining Glagolitic Letter Iotated Small Yus	U+E027	◌	Srezn., p. 248; *MissSin*, 19r12
Combining Glagolitic Letter Big Yus	U+E028	◌	Mansvetov, p. 362 (given by Mansvetov in Cyrillic transcription only)
Combining Glagolitic Letter Iotated Big Yus	U+E029	◌	*EuchSinV*, 5r10, 17r1, 17v18, 28v3, 29r18 etc.
Combining Glagolitic Letter Fita	U+E02A	◌	*Assem*, 125v5, 149v25; (*EuchSinV*, 83v14, 84v15, 85v4 for /f/)

The following entries are proposed for addition to UnicodeData.txt (note that all codepoints are provisional):

```
E000;COMBINING GLAGOLITIC LETTER AZU;Mn;230;NSM;;;;;N;;;;;
E001;COMBINING GLAGOLITIC LETTER BUKI;Mn;230;NSM;;;;;N;;;;;
E002;COMBINING GLAGOLITIC LETTER VEDE;Mn;230;NSM;;;;;N;;;;;
E003;COMBINING GLAGOLITIC LETTER GLAGOLI;Mn;230;NSM;;;;;N;;;;;
```

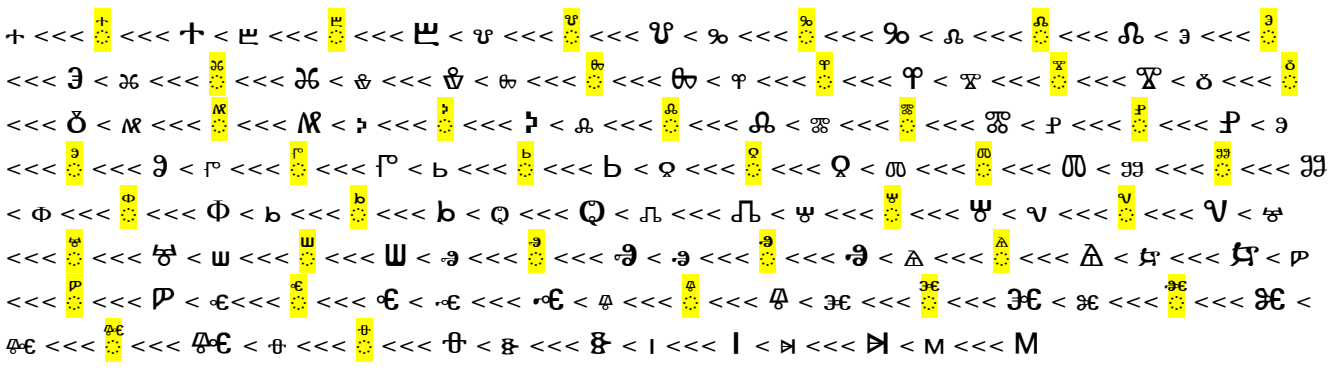
```
E004;COMBINING GLAGOLITIC LETTER DOBRO;Mn;230;NSM;;;;N;;;;;
E005;COMBINING GLAGOLITIC LETTER YESTU;Mn;230;NSM;;;;N;;;;;
E006;COMBINING GLAGOLITIC LETTER ZHIVETE;Mn;230;NSM;;;;N;;;;;
E008;COMBINING GLAGOLITIC LETTER ZEMLJA;Mn;230;NSM;;;;N;;;;;
E009;COMBINING GLAGOLITIC LETTER IZHE;Mn;230;NSM;;;;N;;;;;
E00A;COMBINING GLAGOLITIC LETTER INITIAL IZHE;Mn;230;NSM;;;;N;;;;;
E00B;COMBINIGN GLAGOLITIC LETTER I;Mn;230;NSM;;;;N;;;;;
E00C;COMBINING GLAGOLITIC LETTER DJERVI;Mn;230;NSM;;;;N;;;;;
E00D;COMBINING GLAGOLITIC LETTER KAKO;Mn;230;NSM;;;;N;;;;;
E00E;COMBINING GLAGOLITIC LETTER LJUDIE;Mn;230;NSM;;;;N;;;;;
E00F;COMBINING GLAGOLITIC LETTER MYSLITE;Mn;230;NSM;;;;N;;;;;
E010;COMBINING GLAGOLITIC LETTER NASHI;Mn;230;NSM;;;;N;;;;;
E011;COMBINING GLAGOLITIC LETTER ONU;Mn;230;NSM;;;;N;;;;;
E012;COMBINING GLAGOLITIC LETTER POKOJI;Mn;230;NSM;;;;N;;;;;
E013;COMBINING GLAGOLITIC LETTER RITSI;Mn;230;NSM;;;;N;;;;;
E014;COMBINING GLAGOLITIC LETTER SLOVO;Mn;230;NSM;;;;N;;;;;
E015;COMBINING GLAGOLITIC LETTER TVRIDO;Mn;230;NSM;;;;N;;;;;
E016;COMBINING GLAGOLITIC LETTER UKU;Mn;230;NSM;;;;N;;;;;
E017;COMBINING GLAGOLITIC LETTER FRITU;Mn;230;NSM;;;;N;;;;;
E018;COMBINING GLAGOLITIC LETTER HERU;Mn;230;NSM;;;;N;;;;;
E01B;COMBINING GLAGOLITIC LETTER SHTA;Mn;230;NSM;;;;N;;;;;
E01C;COMBINING GLAGOLITIC LETTER TSI;Mn;230;NSM;;;;N;;;;;
E01D;COMBINING GLAGOLITIC LETTER CHRIVI;Mn;230;NSM;;;;N;;;;;
E01E;COMBINING GLAGOLITIC LETTER SHA;Mn;230;NSM;;;;N;;;;;
E01F;COMBINING GLAGOLITIC LETTER YERU;Mn;230;NSM;;;;N;;;;;
E020;COMBINING GLAGOLITIC LETTER YERI;Mn;230;NSM;;;;N;;;;;
E021;COMBINING GLAGOLITIC LETTER YATI;Mn;230;NSM;;;;N;;;;;
E023;COMBINING GLAGOLITIC LETTER YU;Mn;230;NSM;;;;N;;;;;
E024;COMBINING GLAGOLITIC LETTER SMALL YUS;Mn;230;NSM;;;;N;;;;;
E026;COMBINING GLAGOLITIC LETTER YO;Mn;230;NSM;;;;N;;;;;
E027;COMBINING GLAGOLITIC LETTER IOTATED SMALL YUS;Mn;230;NSM;;;;N;;;;;
E028;COMBINING GLAGOLITIC LETTER BIG YUS;Mn;230;NSM;;;;N;;;;;
E029;COMBINING GLAGOLITIC LETTER IOTATED BIG YUS;Mn;230;NSM;;;;N;;;;;
E02A;COMBINING GLAGOLITIC LETTER FITA;Mn;230;NSM;;;;N;;;;;
```

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

```
E000..E02A   ; Glagolitic # Mn  [43] COMBINING GLAGOLITIC LETTER AZU..COMBINING
GLAGOLITIC LETTER FITA
```

Section 3. Collation

It is proposed that the default collation order given by the DUCET for Glagolitic characters mimic the default collation order for Cyrillic characters as follows:



Section 4. Justification for Encoding

In this section, we explain the rationale behind encoding combining Glagolitic letters in the Unicode standard. Our rationale can be summarized as follows. First, we show that combining Glagolitic characters are in fact distinct from – and in the presentation of text should be handled separately from – their respective base forms. Second, we demonstrate that other possible approaches for handling combining letters in the Glagolitic script – the use of *ad hoc* markup, encoding in the PUA, and the use of advanced font features – are insufficient or overly complex. Finally, we argue that because combining Cyrillic characters have already been encoded in Unicode, correct interoperability between the two scripts demands that combining Glagolitic characters be also encoded.

4.1 Distinction and Use of Combining Characters in Church Slavonic

As we stated in the Introduction section above, superscription in Church Slavonic is used in two instances: in abbreviations (for example, the word **ми́лость** (mercy) is often written as **мл҃ть**) and in *nomina sacra* (for example, the spelling **гд҃ь** (Lord) is used when it refers to God and the spelling **господ́ь** is used when it refers to a secular ruler (a lord), much the same way that capitalization is used in many modern languages).[1] Thus, superscription is a required feature of writing Church Slavonic. Unlike in modern English and other languages where superscription is a stylistic embellishment (e.g., in writing "2nd" as opposed to "2nd"), the superscripted characters in Church Slavonic are combining characters that act like true diacritical marks. In particular, these characters (both in the Cyrillic and Glagolitic scripts) are non-spacing characters, while the "n" and "d" in writing "2nd" in English are spacing characters. Handling such non-spacing characters by positioning spacing characters over a base character would not be correct from the standpoint of text processing in the Unicode standard. In fact, as far as text processing is concerned, the combining characters in Church Slavonic are in no way different from any of the other diacritical marks already encoded.

In his review of L2/14-103, David Birnbaum concedes that "standard modern ChSl [Church Slavonic] orthography does require superscript letters in some words." But then he goes on to write, "I would have regarded the use of the "wrong" letter as culturally incorrect but nonetheless informationally adequate." We strongly disagree with this premise. In many instances, using the inline letter instead of the combining letter is not only "culturally incorrect" but also "informationally inadequate". For example, the sequence **бог**[**ꙗ**] means "god" (a pagan deity) while the sequence **бо҆** (with the combining letter Ge) is an abbreviation for **богоро́дичен҃ъ** ("theotokion" – a type of liturgical hymn). While one could write **бо**ᵍ by using markup-level superscription, this "г" is still a spacing character. One could use kerning at the font level to force the "г" to position over the "о", but such an approach is not correct from the standpoint of text processing. Moreover, under such an approach, the correct meaning of a text stream would not only be determined by its characters but also by markup (or formatting) and by font-level attributes, which cannot be exchanged between users in a text-only format.

In addition, it is also incorrect to write all Church Slavonic words in their full, unabbreviated form (resolving all abbreviations). As we have seen above, **гд҃ь** and **господ́ь** have two different meanings; thus, writing **гд҃ь** as **господ́ь** is "informationally inadequate," not just "culturally incorrect".

1 Throughout we present examples in the Cyrillic script, since it is more familiar to the reader. We will then demonstrate that all of the arguments also hold true for the Glagolitic script because of the relationship that exists between the two scripts.

It is true that strict orthographic conventions did not take shape in Church Slavonic until after 1700 (with the publication of the Elizabeth Bible in 1751) and that Church Slavonic writing of earlier recensions (especially before the advent of the printing press) demonstrates a greater degree of leeway in spelling. Nonetheless, wherever combining letters occur in Church Slavonic, they are always treated as non-spacing marks. Treating them differently in computer-encoded text would be an unreasonable limitation.

In fact, combining characters are already encoded in Unicode for a variety of writing systems. In addition to the combining characters used for Cyrillic, Unicode includes a variety of combining characters used in writing classical Arabic (for example, Honorifics and Koranic annotation signs; for a discussion of these, see L2/01-425); and the various combining Latin characters used for the representation of mediæval texts (see ISO/IEC JTC 1/SC 2/WG 2 N2266 and L2/06-027). In their proposal for combining Latin characters, Everson *et. al.* argue that these characters are necessary to make possible a representation of mediæval text that "does not entail the replacement or the distortion of the original character set." Similarly, a distortion of the Glagolitic character set is undesirable and so combining Glagolitic characters are needed.

4.2 Alternative Approaches

In addition to encoding, three methods to handle combining Glagolitic characters could be contemplated: the use of an *ad hoc* markup language, encoding the characters in the Private Use Area (PUA), and the use of advanced typographic features available in some font technologies.

Use of *Ad Hoc* Markup

The use of *ad hoc* markup languages is the approach proposed by Ralph Cleminson in his response to L2/14-103. Cleminson writes: "[i]t was recommended ... that in those texts (the majority) where superscription is largely a matter of scribal whim, it should be encoded using markup." Of course, given the subsequent encoding of a large number of combining Cyrillic characters in the Unicode standard, this recommendation – which is nowhere articulated in the technical documentation to the Unicode standard – has not been followed. Nonetheless, let us briefly consider this approach. Under this scheme, an *ad hoc* markup language is used to mark up text and to indicate that a certain letter is a superscript. In fact, this approach is not new; it was first developed for Church Slavonic with the creation of the HIP (Hyperinvariant Presentation) technology, a markup language used to represent Church Slavonic with an 8-bit codepage that claims to be a "platform-independent representation of Church Slavonic texts ... designed to record the texts in a readable form." In HIP, the backslash character is used to indicate a superscript; thus, the above example is encoded as:

бо\г

This approach has a number of problems. Though HIP claims to be "readable" (in the sense that it is a set of mnemonic conventions), it is in fact *not* "readable" in the sense that it is not a final presentation form. For example, the text бо\г cannot be presented to an end user on a webpage or in a printed edition. The HIP format – and any markup approach – can only be used for storing text in a plain text format and requires the use of a processor program that converts the markup-encoded text into a format that can be finally presented. The reliance on a processor or converter means that the user of the stored text is limited to a specific set of software tools (those that support add-ons, like Microsoft

Office or those that can be scripted, like the various flavors of TeX).[2] It also means that in addition to the use of the markup language we must use one of the other two approaches (encoding in the PUA or the use of stylistic alternatives) to present the final representation. For example, the processor can convert the HIP-encoded text to a Unicode presentation where the superscript characters are encoded in the PUA. In practice, since a way to work with the final representation of the text without relying on markup now becomes available (the post-processor output), the markup language and the processor itself become obsolete. In other words, as soon as post-processed text free of markup becomes usable, no one will use the text with markup (this is in fact what happened to HIP).

Use of the Private Use Area

The second approach is to encode the combining characters in the Private Use Area of Unicode. At the font level, the positioning of the combining character over the base character is handled via the use of the mark-to-base positioning feature (*mark*) in OpenType. While this approach is attractive because of its simplicity, it presents several problems. First, if the combining characters warrant their own codepoints in the PUA, why do they not warrant codepoints in the body of the Unicode standard? While various attempts to standardize the PUA have been undertaken (for example, the Medieval Unicode Font Initiative), most of these attempts see the use of the PUA as a temporary solution until the necessary characters can be formally proposed for encoding into the standard. Second, one must keep in mind that the Unicode standard is more than encoding: it also defines, for example, character properties, line breaking, and collation. None of these is well-defined for codepoints in the PUA. Characters in the PUA have a `General_Category` property of `Co` (other, private use). In principle, vendors can agree to override the `General_Category` property; to do this, they must "exchange privately defined data which describes [*sic.*] how each private-use character is to be interpreted" (Unicode Standard, p. 558). But since "the Unicode Standard provides no predefined format for such data" (*ibid.*), in practice this means that such a scheme cannot be contemplated beyond a small, tightly-nit user community using a limited set of software. Certainly, this would not allow for easy exchange of Church Slavonic texts over the Internet in a way that is supported by all browsers and mobile devices. Furthermore, the lack of a specified collation table would preclude proper indexation of the text by major search engines and hamper other string manipulation operations. Finally, many software applications do not correctly support OpenType features (including mark-to-base positioning) for characters in the PUA, so the correct rendering of the text would fail. All in all, the use of the PUA can only be contemplated as a temporary measure until such a time as the characters can be formally accepted into the Unicode standard.

Use of Advanced Font Features

The other approach is the use of advanced typographic features, such as the use of stylistic alternatives (*salt*) or stylistic sets in OpenType or the use of custom features in SIL Graphite. For example, the main glyph r can be stored as the base character in the font and the superscript glyph  can be stored as *stylistic alternative 1*. The user can select the superscripted glyph by selecting *stylistic alternative 1* in a text rendering application. The positioning of the superscripted glyph is again handled via mark-to-base positioning.

2 A further problem becomes that markup used for character presentation must be distinguished from internal markup used for text style. For example, how does one store information about font color together with the markup? Would it even be possible to have a universal format that can be used in all software – Notepad, TeX, the various Offices, web browsers, mobile devices, and so forth? (For example, the backslash character is a control character in TeX and its use in HIP makes the implementation of HIP in a TeX context a logistical nightmare).

While this approach looks powerful at first glance, it in fact fails spectacularly. First of all, not all software supports advanced typographic features (how does one select stylistic alternatives in Notepad or even in LibreOffice?) Second, even in software where stylistic alternatives can be selected, indicating that a certain character in a character stream has a certain property (such as being a stylistic alternative) usually interrupts the character stream, preventing the mark-to-base positioning feature from working properly, so the practical rendering of combining characters becomes impossible. Now it could be argued that implementation bugs in software need not concern the Unicode Technical Committee. But why design an encoding model that cannot be used in practice given the existing software?

The third issue with this approach is that it does not provide for any way to store information about the superscript letters in a software- and platform-independent setting. While information about stylistic alternatives can be stored in a specific file format (for example, within a Microsoft Office DOCX file), it can be shared across platforms and applications only via the use of a markup language. For example, one could record the sequence in question in an XML-like format as:

```
бо<salt alt="1">г</salt>
```

In addition to all of the problems involving the use of markup languages discussed above, this approach suffers from further interoperability limitations. No standard for stylistic alternatives exists: font developers can choose to support the combining character as *stylistic alternative 1*; or, they can choose to support it as *alternative 2* and support something else (for example, the italic Russian *г*) as *alternative 1*; or, they can choose not to support stylistic alternatives at all. At least the PUA provides a limited set of codepoints so that some informal agreement between vendors could be reached; under the stylistic alternatives approach, if we also allow for SIL Graphite features (which have vendor-defined names), the number of ways to render the combining letters becomes non-countable and exchange of data across applications and systems in a standard way becomes impossible. Moreover, since stylistic alternatives are designed to be optional, the end user would not even see an error message when a given alternative is not available in a font.

4.3 Interoperability of Cyrillic and Glagolitic

In addition to the reasons presented above, there is one further reason why combining Glagolitic letters should be encoded in the Unicode standard. Namely, the Cyrillic and Glagolitic scripts are closely related; combining characters have already been encoded for Cyrillic; a methodology needs to exist for the simple conversion of texts between Cyrillic and Glagolitic; thus, combining Glagolitic letters should also be encoded.

As we have pointed out in the Introduction section, scholars believe that Glagolitic was the first script used to record Church Slavonic. After the introduction of the Cyrillic script, it gradually replaced Glagolitic in most Slavic cultures. Nonetheless, Cyrillic and Glagolitic scripts are often used interchangeably in the academic community, and it is quite common to publish Glagolitic textual sources in Cyrillic transcription (see Figure 14). Because of this interoperability of Cyrillic and Glagolitic, computer software needs to be able to unambiguously convert (transliterate) between the two writing systems. In the Table below, we present the transliteration scheme commonly used and due to Jagic (1879):

Glagolitic			Cyrillic		
Name	**Codept.**	**Disp.**	**Name**	**Codept.**	**Disp.**
Glagolitic Letter Azu	U+2C30	ⰰ	Cyrillic Letter A	U+0430	а
Glagolitic Letter Buki	U+2C31	ⰱ	Cyrillic Letter Be	U+0431	б
Glagolitic Letter Vede	U+2C32	ⰲ	Cyrillic Letter Ve	U+0432	в
Glagolitic Letter Glagoli	U+2C33	ⰳ	Cyrillic Letter Ge	U+0433	г
Glagolitic Letter Dobro	U+2C34	ⰴ	Cyrillic Letter De	U+0434	д
Glagolitic Letter Yestu	U+2C35	ⰵ	Cyrillic Letter Ie	U+0435	е
Glagolitic Letter Zhivete	U+2C36	ⰶ	Cyrillic Letter Zhe	U+0436	ж
Glagolitic Letter Dzelo	U+2C37	ⰷ	Cyrillic Letter Dze	U+0455	ѕ
Glagolitic Letter Zemlja	U+2C38	ⰸ	Cyrillic Letter Ze	U+0437	з
Glagolitic Letter Izhe	U+2C39	ⰹ	Cyrillic Letter Ukrainian I	U+0456	і
Glagolitic Letter Initial Izhe	U+2C3A	ⰺ	Cyrillic Letter Iota	U+A647	ꙇ
Glagolitic Letter I	U+2C3B	ⰻ	Cyrillic Letter I	U+0438	и
Glagolitic Letter Djervi	U+2C3C	ⰼ	Cyrillic Letter Djerv	U+A649	ꙉ
Glagolitic Letter Kako	U+2C3D	ⰽ	Cyrillic Letter Ka	U+043A	к
Glagolitic Letter Ljudie	U+2C3E	ⰾ	Cyrillic Letter El	U+043B	л
Glagolitic Letter Myslite	U+2C3F	ⰿ	Cyrillic Letter Em	U+043C	м
Glagolitic Letter Nashi	U+2C40	ⱀ	Cyrillic Letter En	U+043D	н
Glagolitic Letter Onu	U+2C41	ⱁ	Cyrillic Letter O	U+043E	о
Glagolitic Letter Pokoji	U+2C42	ⱂ	Cyrillic Letter Pe	U+043F	п
Glagolitic Letter Ritsi	U+2C43	ⱃ	Cyrillic Letter Er	U+0440	р
Glagolitic Letter Slovo	U+2C44	ⱄ	Cyrillic Letter Es	U+0441	с
Glagolitic Letter Tvrido	U+2C45	ⱅ	Cyrillic Letter Te	U+0442	т
Glagolitic Letter Uku	U+2C46	ⱆ	Cyrillic Letter O Cyrillic Letter U	U+043E U+0443	оу
Glagolitic Letter Fritu	U+2C47	ⱇ	Cyrillic Letter Ef	U+0444	ф
Glagolitic Letter Heru	U+2C48	ⱈ	Cyrillic Letter Ha	U+0445	х
Glagolitic Letter Otu	U+2C49	ⱉ	Cyrillic Letter Omega	U+0461	ѡ
Glagolitic Letter Shta	U+2C4B	ⱋ	Cyrillic Letter Shcha	U+0449	щ
Glagolitic Letter Tsi	U+2C4C	ⱌ	Cyrillic Letter Tse	U+0446	ц
Glagolitic Letter Chrivi	U+2C4D	ⱍ	Cyrillic Letter Che	U+0447	ч
Glagolitic Letter Sha	U+2C4E	ⱎ	Cyrillic Letter Sha	U+0448	ш

Glagolitic			Cyrillic		
Glagolitic Letter Yeru	U+2C4F	Ⱟ	Cyrillic Letter Hard Sign	U+044A	ъ
Glagolitic Letter Yeru, Glagolitic Letter Izhe	U+2C4F U+2C39	Ⱟⰹ	Cyrillic Letter Yeru with Back Yer	U+A651	ꙑ
Glagolitic Letter Yeri	U+2C50	ⰰ	Cyrillic Letter Soft Sign	U+044C	ь
Glagolitic Letter Yati	U+2C51	ⰱ	Cyrillic Letter Yat	U+0463	ѣ
Glagolitic Letter Yu	U+2C53	ⰳ	Cyrillic Letter Yu	U+044E	ю
Glagolitic Letter Small Yus	U+2C54	ⰴ	Cyrillic Letter Little Yus	U+0467	ѧ
Glagolitic Letter Iotated Small Yus	U+2C57	ⰷ	Cyrillic Letter Iotified Little Yus	U+0469	ѩ
Glagolitic Letter Big Yus	U+2C58	ⰸ	Cyrillic Letter Big Yus	U+046B	ѫ
Glagolitic Letter Iotated Big Yus	U+2C59	ⰹ	Cyrillic Letter Iotified Big Yus	U+046D	ѭ
Glagolitic Letter Fita	U+2C5A	ⰺ	Cyrillic Letter Fita	U+0473	ѳ
Glagolitic Letter Izhitsa	U+2C5B	ⰻ	Cyrillic Letter Izhitsa	U+0475	ѵ

The Table presents the standard transliteration scheme for the main letters of the Cyrillic and Glagolitic scripts. However, we contend that an automated conversion (transliteration) algorithm also needs to correctly handle diacritical marks and combining letters present in the two scripts.

In the encoding model for the Cyrillic script, the combining Cyrillic letters are already available. In their responses to L2/14-103, Cleminson and Birnbaum present an overall negative view of the use of combining characters for Cyrillic. Cleminson believes that only those combining Cyrillic letters that occur in modern Church Slavonic should have been encoded and Birnbaum writes that he “cannot now endorse the inclusion of any additional superscript Cyrillic or Glagolitic characters.”

In hindsight, we agree that the use of combining characters for Cyrillic superscription was not the best approach (although our criticism of this approach is based on different reasons than those presented by Cleminson and Birnbaum). In our view, this approach is too limiting and does not allow for the unambiguous representation of various complexities occurring in mediæval texts, such as, for example, the use of superscription over multiple base letters or the use of multiple combining letters over one base letter. Also, we regret that the standardization of Glagolitic and Cyrillic was not handled simultaneously, leading to discrepancies in the encoding schemes. Nonetheless, given Unicode's stability policy, there is no use now to criticize an existing implementation. Rather, we desire to make the existing implementation more useful. Thus, because combining Cyrillic letters are encoded, we propose that combining Glagolitic letters also need to be encoded so that a meaningful transliteration scheme can be designed. An approach where Cyrillic superscription is handled via combining characters while Glagolitic superscription is rendered using markup or stylistic alternatives would not make such a transliteration scheme possible. But, as we have pointed out, since it is common to present Glagolitic texts in Cyrillic transcription, the availability of such a scheme is necessary. The Table below presents the proposed transliteration scheme for combining letters. Note, again, that all codepoints for Glagolitic characters are provisional.

Glagolitic			Cyrillic		
Name	**Codept.**	**Disp.**	**Name**	**Codept.**	**Disp.**
Glagolitic Combining Letter Azu	U+E000	◌ⰰ	Cyrillic Combining Letter A	U+2DF6	◌ᲀ
Glagolitic Combining Letter Buki	U+E001	◌ⰱ	Cyrillic Combining Letter Be	U+2DE0	◌Д
Glagolitic Combining Letter Vede	U+E002	◌ⰲ	Cyrillic Combining Letter Ve	U+2DE1	◌ᲂ
Glagolitic Combining Letter Glagoli	U+E003	◌ⰳ	Cyrillic Combining Letter Ge	U+2DE2	◌С
Glagolitic Combining Letter Dobro	U+E004	◌ⰴ	Cyrillic Combining Letter De	U+2DE3	◌ᲄ
Glagolitic Combining Letter Yestu	U+E005	◌ⰵ	Cyrillic Combining Letter Ie	U+2DF7	◌ᲅ
Glagolitic Combining Letter Zhivete	U+E006	◌ⰶ	Cyrillic Combining Letter Zhe	U+2DE4	◌Ъ
Glagolitic Combining Letter Zemlja	U+E008	◌ⰸ	Cyrillic Combining Letter Ze	U+2DE5	◌Ѣ
Glagolitic Combining Letter Izhe	U+E009	◌ⰹ	Cyrillic Combining Letter Yi	U+A676	◌ꙶ
Glagolitic Combining Letter Initial Izhe	U+E00A	◌ⰺ	Cyrillic Combining Letter Iota	N/A[3]	
Glagolitic Combining Letter I	U+E00B	◌ⰻ	Cyrillic Combining Letter I	U+A675	◌ꙵ
Glagolitic Combining Letter Djervi	U+E00C	◌ⰼ	Cyrillic Combining Letter Djerv	U+2DF8	◌ᲈ
Glagolitic Combining Letter Kako	U+E00D	◌ⰽ	Cyrillic Combining Letter Ka	U+2DE6	◌Ᲊ
Glagolitic Combining Letter Ljudie	U+E00E	◌ⰾ	Cyrillic Combining Letter El	U+2DE7	◌᲋
Glagolitic Combining Letter Myslite	U+E00F	◌ⰿ	Cyrillic Combining Letter Em	U+2DE8	◌᲌
Glagolitic Combining Letter Nashi	U+E010	◌ⱀ	Cyrillic Combining Letter En	U+2DE9	◌᲍
Glagolitic Combining Letter Onu	U+E011	◌ⱁ	Cyrillic Combining Letter O	U+2DEA	◌᲎
Glagolitic Combining Letter Pokoji	U+E012	◌ⱂ	Cyrillic Combining Letter Pe	U+2DEB	◌᲏
Glagolitic Combining Letter Ritsi	U+E013	◌ⱃ	Cyrillic Combining Letter Er	U+2DEC	◌Ა
Glagolitic Combining Letter Slovo	U+E014	◌ⱄ	Cyrillic Combining Letter Es	U+2DED	◌Ბ
Glagolitic Combining Letter Tvrido	U+E015	◌ⱅ	Cyrillic Combining Letter Te	U+2DEE	◌Გ
Glagolitic Combining Letter Uku	U+E016	◌ⱆ	Cyrillic Combining Letter Monograph Uk	U+2DF9	◌Ᲊ
Glagolitic Combining Letter Fritu	U+E017	◌ⱇ	(Cyrillic Combining Letter Ef)	(U+A69E)	◌ꚟ
Glagolitic Combining Letter Heru	U+E018	◌ⱈ	Cyrillic Combining Letter Ha	U+2DEF	◌Დ

3 This character will be proposed for encoding by the authors in a separate document.

Glagolitic			Cyrillic		
Glagolitic Combining Letter Shta	U+E01B	◌	Cyrillic Combining Letter Shcha	U+2DF3	◌
Glagolitic Combining Letter Tsi	U+E01C	◌	Cyrillic Combining Letter Tse	U+2DF0	◌
Glagolitic Combining Letter Chrivi	U+E01D	◌	Cyrillic Combining Letter Che	U+2DF1	◌
Glagolitic Combining Letter Sha	U+E01E	◌	Cyrillic Combining Letter Sha	U+2DF2	◌
Glagolitic Combining Letter Yeru	U+E01F	◌	Cyrillic Combining Letter Hard Sign	U+A678	◌
Glagolitic Combining Letter Yeri	U+E020	◌	Cyrillic Combining Letter Soft Sign	U+A67A	◌
Glagolitic Combining Letter Yati	U+E021	◌	Cyrillic Combining Letter Yat	U+2DFA	◌
Glagolitic Combining Letter Yu	U+E023	◌	Cyrillic Combining Letter Yu	U+2DFB	◌
Glagolitic Combining Letter Small Yus	U+E024	◌	Cyrillic Combining Letter Little Yus	U+2DFD	◌
Glagolitic Combining Letter Iotated Small Yus	U+E027	◌	Cyrillic Combining Letter Iotified Little Yus	N/A[4]	
Glagolitic Combining Letter Big Yus	U+E028	◌	Cyrillic Combining Letter Big Yus	U+2DFE	◌
Glagolitic Combining Letter Iotated Big Yus	U+E029	◌	Cyrillic Combining Letter Iotified Big Yus	U+2DFF	◌
Glagolitic Combining Letter Fita	U+E02A	◌	Cyrillic Combining Letter Fita	U+2DF4	◌

4.4 Collation

In his response, Ralph Cleminson writes that the approach of using combining characters "has serious drawbacks for the processing of electronic text." He does not specify what these drawbacks are, but we take this to mean the difficulty in comparing strings. Under this encoding methodology, for example, the strings **нашнχ** and **нашн◌ᷯ** are not equivalent at the codepoint level even though under certain circumstances one would like to treat them as the same string. However, the proper way to compare these strings is not via direct (codepoint by codepoint) comparison but rather via the use of the Unicode Collation Algorithm (UCA), which allows for multi-level comparison codes. Under an appropriate UCA collation table, the characters **χ** and ◌ᷯ could be given the same primary weights but different secondary weights (which is in fact what is implemented for Cyrillic and what we propose for Glagolitic). In a context where the above strings need to be treated as identical one can simply perform a comparison ignoring the secondary weights, while in the context where these strings need to be different, one can compare them at the secondary level. In any case, comparing any two Unicode strings codepoint by codepoint is not advisable from the standpoint of text processing.

4 This character will be proposed for encoding by the authors in a separate document.

Section 5. Examples

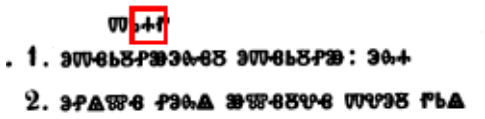


Figure 1: Combining Glagolitic Letter Azu (boxed in red). Source: Srezn.

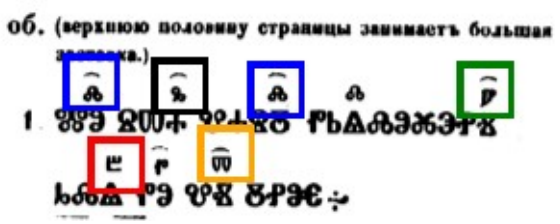


Figure 2: Combining Glagolitic Letters Buky (boxed in red); Glagoli (boxed in black); Ljudije (boxed in blue); Tvrido (boxed in yellow); and Yu (boxed in green). Note the use of the Pokrytie. Source: Srezn.

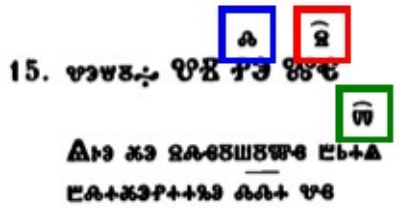


Figure 3: Combining Glagolitic Letters Dobro (boxed in blue); Slovo (boxed in red) and Tvrido (boxed in green). Source: Srezn.

Figure 4: Combining Glagolitic Letter Kako (boxed in red). Note the use of the Pokrytie. Source: Srezn.

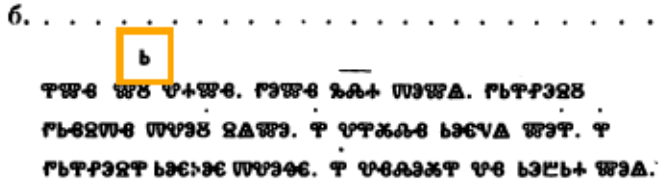


Figure 5: Combining Glagolitic Letter Heru (boxed in orange). Source: Srezn.

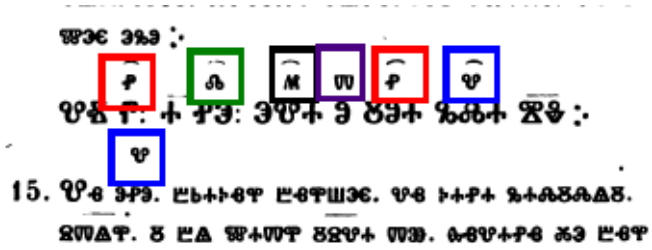


Figure 6: Combining Glagolitic Letters Vede (boxed in blue); Dobro (boxed in green); Djervi (boxed in black); Nashi (boxed in red); and Tvrido (boxed in indigo). Source: Srezn.

Figure 7: Combining Glagolitic Letters Vede (boxed in blue); Zhivete (boxed in green); Myslite (boxed in black); and Sha (boxed in red). Source: Srezn.

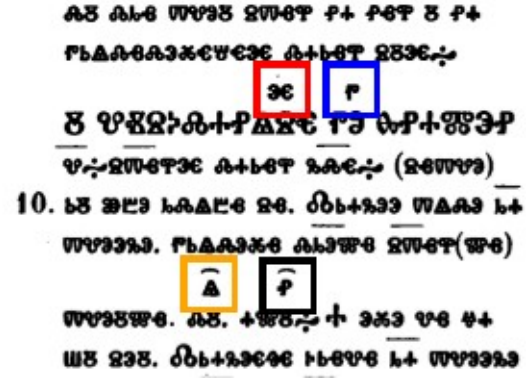


Figure 8: Combining Glagolitic Letters Nashi (boxed in black); Pokoji (boxed in blue); Yati (boxed in orange); and Iotated Small Yus (boxed in red). Source: Srezn.

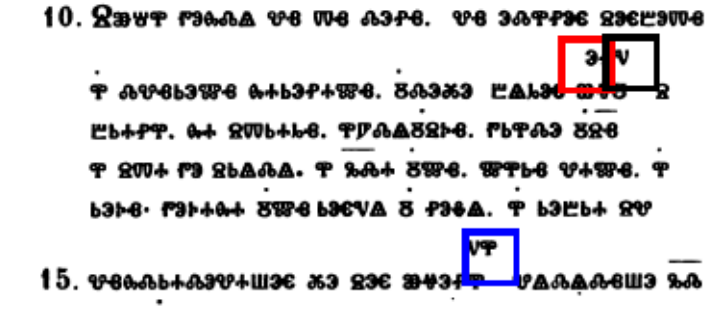


Figure 9: Combining Glagolitic Letters Yestu (boxed in red); Tsi (boxed in black); and Izhe (boxed in blue). Source: Srezn.

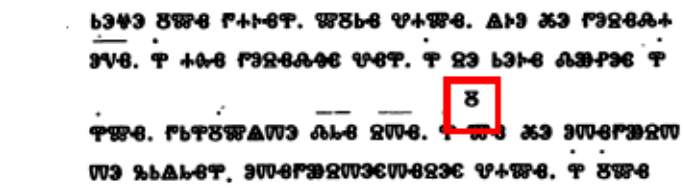


Figure 10: Combining Glagolitic Letter I (boxed in red). Source: Srezn.

Figure 11: Combining Glagolitic Letters Buky (boxed in green); Vedi (boxed in indigo); Djervi (boxed in orange); Kako (boxed in red); Tvrido (boxed in black); and Fita (boxed in blue). Source: Assemaniev Gospel.

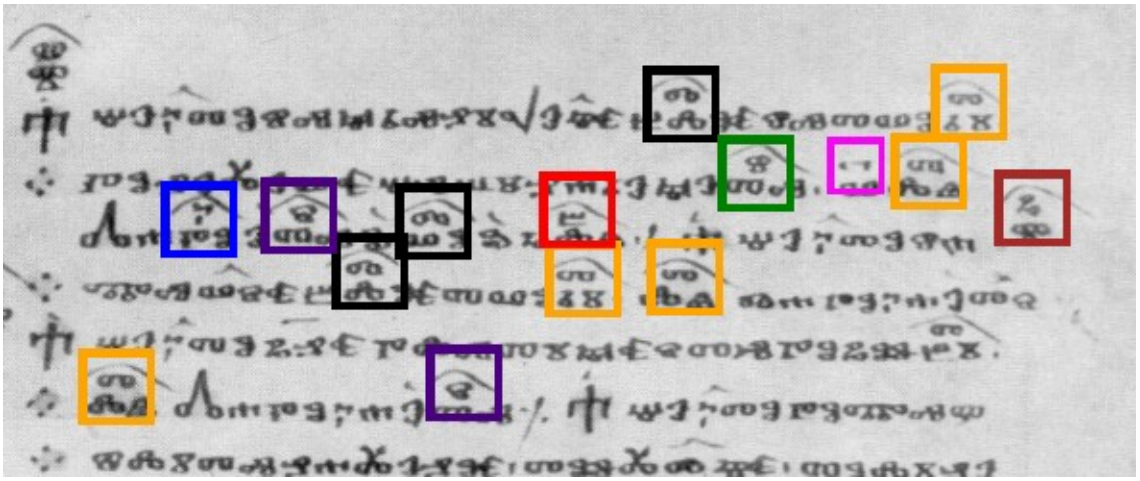


Figure 12: Combining Glagolitic Letters Buky (boxed in red); Glagoli (boxed in brown); Dobro (boxed in black); Initial Izhe (boxed in green); Kako (boxed in blue); Slovo (boxed in indigo); and Tvrido (boxed in orange). Note the use of the Pokrytie with these characters as well as the use of the Combining Vzmet (boxed in magenta). Source: EuchSin.

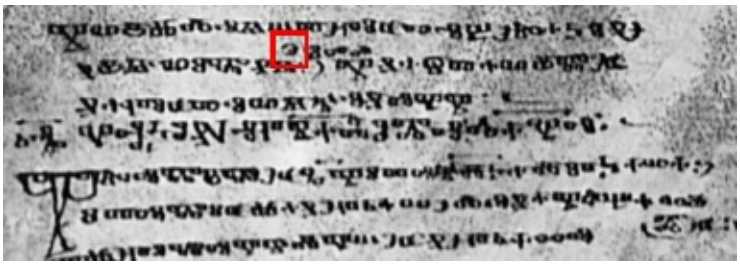


Figure 13: Combining Glagolitic Letter Yestu (boxed in red). Source: PsSin.

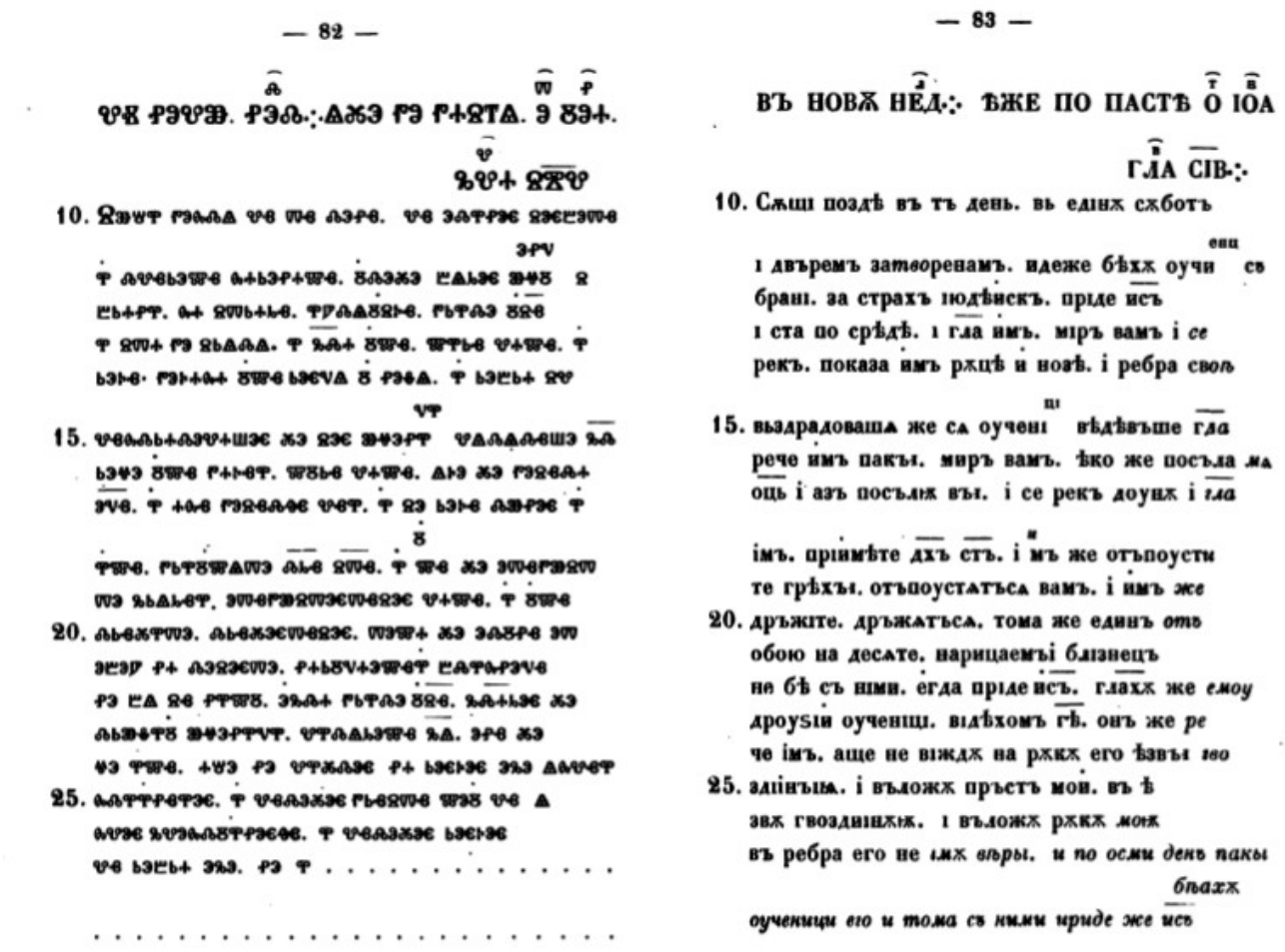


Figure 14: Glagolitic and Cyrillic text presented side-by-side in an academic publication. Source: Srez.

References

Assem: Codex Assemanianus (Vat. slav. 3 glag.), Macedonia, early 11th c. Available online (printed in *Asemanievo evangelie. Faksimilno izdanie*. Sofia: Nauka i izkustvo, 1981).

EuchSinV: Euchologium Sinaiticum, old part (Sin. slav. 37 & SPb. RNB, Glag. 3), Bulgaria, 11th c.; facsimile edition: NAHTIGAL, R. *Euchologium Sinaiticum. Starocerkvenoslovanski glagolski spomenik.* I. *Fotografski posnetek*. Ljubljana: AZU v Ljubljani, Učiteljska tiskarna, 1941.

Everson: EVERSON, Michael *et. al.* "Proposal to add medievalist characters to the UCS". Working Group Document L2/06-027. 2006.

Jagic: JAGIĆ, Vatroslav, *Quattuor Evangeliorum Codex Glagoliticus, olim Zographensis, nunc Petropolitanus*. Berlin: Apud Weidmannos, 1879.

Mansvetov: MANSVETOV, I. D., "Evchologium Glagolski, Spomenik monastira Sinai brda. Izdao D-r Lavoslav Geitler. U Zagrebu, 1882. (Древнейший Славянский требник, изданный г. Гейтлером по рукописи, находящейся в Синайской библиотеке)". In *Прибавления к Твореним св. отцов.* 1883. vol. 32, part 1, pp. 347-390.

MissSin: Sinai Missal (Sin. slav. 5/N), Bulgaria, 11th c.; cf. MIKLAS, H. (ed.): *Glagolitica – Zum Ursprung der slavischen Schriftkultur* (ÖAW, Phil.-hist. Kl., Schriften der Balkan-Kommission, Philologische Abt. 41). Vienna: ÖAW, 2000, pp. 117-129.

PsDem: Psalter of Demetrius (Sin. slav. 3/N), Sinai, 11th c. Facsimile edition: *Psalterium Demetrii Sinaitici (monasterii sanctae Catharinae codex slav. 3/N), adiectis foliis medicinalibus*. Ad editionem phototypicam praeparaverunt M. Gau, D. Hürner, F. Kleber, M. Lettner, H. Miklas sub redactione Henrici Miklas. Cum praefationibus sacri monasterii atque Ioannis Tarnanidae (Glagolitica Sinaitica 1). Vienna: Holzhausen, 2012.

PsSinV: Psalterium Sinaiticum, old part (Sin. slav. 38), Sinai, 11th c., Facsimile edition: ALTBAUER, M.: *Psalterium Sinaiticum. An 11th Century Glagolitic Manuscript from St. Catherine's Monastery, Mt. Sinai.* Skopje: The Macedonian Academy of Sciences and Arts, Goce Delčev Publishing House, 1971.

Srez.: SREZNEVSKIJ, I. I., *Древніе глаголическіе памятники, сравнительно съ памятниками кириллицы*, St. Petersburg: Tipografija Imperatorskoj akademii nauk, 1866.

Glagolitic Extended (Proposed)

Warning: all codepoints are provisional!

0	1	2
◌ U+E000	◌ U+E010	◌ U+E020
◌ U+E001	◌ U+E011	◌ U+E021
◌ U+E002	◌ U+E012	U+E022
◌ U+E003	◌ U+E013	◌ U+E023
◌ U+E004	◌ U+E014	◌ U+E024
◌ U+E005	◌ U+E015	U+E025
◌ U+E006	◌ U+E016	◌ U+E026
U+E007	◌ U+E017	◌ U+E027
◌ U+E008	◌ U+E018	◌ U+E028
◌ U+E009	U+E019	◌ U+E029
◌ U+E00A	U+E01A	◌ U+E02A
◌ U+E00B	◌ U+E01B	U+E02B
◌ U+E00C	◌ U+E01C	U+E02C
◌ U+E00D	◌ U+E01D	U+E02D
◌ U+E00E	◌ U+E01E	U+E02E
◌ U+E00F	◌ U+E01F	U+E02F

U+E000: Combining Glagolitic Letter Azu
U+E001: Combining Glagolitic Letter Buki
U+E002: Combining Glagolitic Letter Vede
U+E003: Combining Glagolitic Letter Glagoli
U+E004: Combining Glagolitic Letter Dobro
U+E005: Combining Glagolitic Letter Yestu
U+E006: Combining Glagolitic Letter Zhivete
U+E007: <not assigned>
U+E008: Combining Glagolitic Letter Zemlja
U+E009: Combining Glagolitic Letter Izhe
U+E00A: Combining Glagolitic Letter Initial Izhe
U+E00B: Combining Glagolitic Letter I
U+E00C: Combining Glagolitic Letter Djervi
U+E00D: Combining Glagolitic Letter Kako
U+E00E: Combining Glagolitic Letter Ljudie
U+E00F: Combining Glagolitic Letter Myslite

U+E010: Combining Glagolitic Letter Nashi
U+E011: Combining Glagolitic Letter Onu
U+E012: Combining Glagolitic Letter Pokoji
U+E013: Combining Glagolitic Letter Ritsi
U+E014: Combining Glagolitic Letter Slovo
U+E015: Combining Glagolitic Letter Tvrido
U+E016: Combining Glagolitic Letter Uku
U+E017: Combining Glagolitic Letter Fritu
U+E018: Combining Glagolitic Letter Heru
U+E019: <not assigned>
U+E01A: <not assigned>
U+E01B: Combining Glagolitic Letter Shta
U+E01C: Combining Glagolitic Letter Tsi
U+E01D: Combining Glagolitic Letter Chrivi
U+E01E: Combining Glagolitic Letter Sha
U+E01F: Combining Glagolitic Letter Yeru

U+E020: Combining Glagolitic Letter Yeri
U+E021: Combining Glagolitic Letter Yati
U+E022: <not assigned>
U+E023: Combining Glagolitic Letter Yu
U+E024: Combining Glagolitic Letter Small Yus
U+E025: <not assigned>
U+E026: Combining Glagolitic Letter Yo
U+E027: Combining Glagolitic Letter Iotated Small Yus
U+E028: Combining Glagolitic Letter Big Yus
U+E029: Combining Glagolitic Letter Iotated Big Yus
U+E02A: Combining Glagolitic Letter Fita
U+E02B: <not assigned>
U+E02C: <not assigned>
U+E02D: <not assigned>
U+E02E: <not assigned>
U+E02F: <not assigned>