ISO/IEC JTC1/SC2/WG2 N4534
Date: 2014-02-04

Proposal to Encode a Slavonic Punctuation Mark in Unicode

Aleksandr Andreev* Yuri Shardt Nikita Simmons

PONOMAR PROJECT

1 Introduction

This document is a proposal to encode an additional character in the Supplemental Punctuation block of the Unicode Standard. This character is widely used in Church Slavonic manuscripts, written both in the Cyrillic and Glagolitic scripts, as well as in some early Slavonic incunabula for punctuation. It may also be encountered in academic texts that study or reproduce manuscripts or historical printed texts and in published editions that mimic medieval manuscripts. The encoding of this character is thus required for two purposes: first, for the correct digital storage of the contents of manuscripts and their search and display in electronic media; second, for use by palæographers, linguists, and scholars of liturgics in academic literature. The proposed character is summarized below.

2 Description of Character

We propose to name the character "Dash with Left Upturn". In previous versions of this proposal, we had proposed the names "Slavonic Spear" or "Slavonic Paragraphos". The first is a translation of the Slavonic term "Kopie", used for this character by Cleminson et al. (2010). Consultation with the authors of that paper revealed that the term "Kopie" was created as a translation of the Greek term "obelus." In modern typography, the term obelus most commonly refers to the division sign and related symbols used for obelism. In our opinion, the present character is more properly associated with the paragraphos and not the obelus, based both on its appearance and function. However, since the paragraphos typically occurs in the margin and this character commonly occurs in line at the end of a block of text, we propose instead the generic name "Dash with Left Upturn." We should note that while existing texts on Slavonic palæography (e.g., Karsky (1979)) do identify this character, they do not provide it a name (see Figure 1).

The proposed character is used in certain Ustav (Uncial) manuscripts to indicate either a full stop, a medium stop, or the conclusion of a section. For example, in the Ostromir Gospel, an important 11th Century manuscript, this symbol is used to indicate the end of a pericope or the end of a heading preceding a pericope; in this usage, it commonly occurs next to a Middle

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

Dot (period) as can be seen in Figure 2. The character plays a similar function in the Sava's Book,[1] another 11th Century manuscript, as can be seen from Figure 3. Both of these texts – the Ostromir Gospel and the Sava's Book – have been digitized and made available on-line by the Manuscripts.ru project; however, that project uses a custom font and an *ad hoc* codepage which encodes all Slavonic characters (including this one) in the Private Use Area of Unicode. The Ponomar Project is presently engaged in the process of converting these texts to Unicode.

The usage of this character is not limited to the Cyrillic script; it may also occur in Glagolitic manuscripts. Figure 4 presents an example of this character in the *Psalterium Sinaiticum*, an 11th Century Glagolitic Psalter, where the character is used to indicate the end of a Psalm verse. In Figure 5, the same character is seen in an academic reprint of the text of the *Psalterium Sinaiticum* (though given in Cyrillic transcription, as is common for studies in Glagolitic palæography). Note that in the critical edition of this manuscript published by Sever′yanov (1922), the character has been reprinted both in the text and in various comments of the critical apparatus.

While it is primarily used in the manuscript tradition, this character may also be encountered in printed texts, though its usage in printed matter is admittedly more rare. Nonetheless, examples of this character used in Slavonic incunabula – early books printed in the Church Slavonic language – can be seen in Figures 7 and 8. Note that the shape of this character in the printed editions has become somewhat stylized. Finally, in Figure 6 we present an example of the "Dash with Left Upturn" used in a modern setting to separate blocks of lyrics in Znamenny musical notation. Here, the character is used out of a desire to imitate the manuscript tradition.

Note that in most (though not all) instances, this character is preceded by some variant of the period. This may be the "Three Dot Punctuation" (Figures 1 and 6); the "Middle Dot" (Figure 2); or the "Colon" (Figures 4, 7 and 8). In Figure 3, the character occurs all by itself. Other configurations may be observed, including the use of Tricolon, Four Dot Punctuation, Five Dot Punctuation, or other ornamental punctuation characters. Since the character preceding the "Dash with Left Upturn" can vary, the characters should properly be encoded separately and the correct graphical representation should be left to the font and rendering system, where it can be achieved with the appropriate use of kerning. Thus, for example, we believe that it would be incorrect to encode the "Three Dot Punctuation" and the "Dash with Left Upturn" as a standalone ligature.

3 Similar Characters

The "Dash with Left Upturn" character bears a visual similarity with several dash or hyphen characters already available in Unicode. As well, the Unicode Standard includes already the Greek Paragraphos (U+2E0F) and two related characters, the Forked Paragraphos and the Reversed Forked Paragraphos. We summarize the appearance and function of all of these characters in Table 1.

As can be clearly seen from Table 1, the characters already encoded in Unicode cannot be used to represent the Slavonic character in question because they have a specific required appearance or function that differs from those of the proposed character. The various dash characters – En Dash, Em Dash, and Figure Dash – have required width because of their ty-

[1] So-named after the priest Sava (Sabbas), who inscribed his name on the original folios.

Table 1: Proposed Character and Similar Characters

Name	**Codepoint**	**Appearance**	**Function**
Figure Dash	U+2012	‒	Must be same width as digits in a font.
En Dash	U+2013	–	Must be half the width of the Em dash, below.
Em Dash	U+2014	—	Must be the width of one em, used for demarcation and interpolation
Horizontal Bar	U+2015	―	Quotation dash used to introduce quoted text.
Swung Dash	U+2053	⁓	A tilde character; for example, as used in dictionaries and other linguistic work.
Hyphen-Minus	U+002D	-	Used for Hyphen or Minus Sign.
Soft Hyphen	U+00AD		An invisible character used to indicate discretionary hyphenation.
Armenian Hyphen	U+058A	֊	Used in Armenian.
Hyphen	U+2010	‐	Unambiguous representation of the Hyphen.
Non-breaking Hyphen	U+2011	‑	Hard hyphen used when line breaking is not allowed.
Hyphen Bullet	U+2043	⁃	A short horizontal bullet.
Paragraphos	U+2E0F	⸏	Lies at the baseline.
Forked Paragraphos	U+2E10	⸐	Lies at the baseline and has a notch on left side.
Rev. Forked Paragraphos	U+2E11	⸑	Reversed version of U+2E10.
Dash with Left Upturn	(U+2E43)	⸃	Proposed character

pographical functions whereas the character in question can be of any arbitrary width. The Horizontal Bar (U+2053) must not be used to represent this character because, in addition to the graphical differences, it is widely used in Russian and other languages to indicate quotation. Mapping the proposed Slavonic character to the Horizontal Bar would produce undesirable problems for working with Church Slavonic text in an otherwise Russian-language document. The various hyphen characters have specific appearances and functions related to hyphenation which must not be compromised, not the least because Church Slavonic typography (at least of the Synodal recension) may also use hyphenation. Finally, the three paragraphoi characters used in Greek text all lie along the baseline whereas the proposed character is typically raised from the baseline.

We should also note specifically that it is not appropriate to use U+2053 SWUNG DASH to encode the proposed character, even though this has been done by some authors, in part because an adequate codepoint for the proposed character does not exist in Unicode or in most existing *ad hoc* codepages. For example, in the reproductions of manuscripts given by Yelkina (1960), the Swung Dash is regularly used in place of the "Dash with Left Upturn." Besides the fact that the Swung Dash should not be stylized as a "Dash with Left Upturn" in fonts because it has legitimate functional purposes for linguists and palæographers, we note that both the Swung Dash and the "Dash with Left Upturn" may occur in Slavonic text (as can be seen in Figure 9) and merging the two characters – even though their purpose is perhaps similar – seems entirely inappropriate. While it is not the purpose of typography to reproduce *every* nuance of the manuscript tradition, it is also incorrect to disregard the manuscript tradition to such an extent that apparently different characters become convoluted.

4 Character Properties

In light of the fact that no existing character in Unicode can be properly used to represent the Slavonic character in question, we propose to encode this character at a new codepoint; because the character can occur in both Cyrillic and Glagolitic texts, it is most appropriately encoded in the Supplemental Punctuation block of Unicode, as proposed in Table 2 and the subsequent listing of character properties.[2] We propose to add the note "Used in Church Slavonic" as an annotation in order to facilitate the correct usage of this character by the Church Slavonic language community.

Table 2: Proposed Character

Character	Codepoint	Name
⸗	U+2E43	DASH WITH LEFT UPTURN

5 Character Properties

The following entry is proposed for addition to UnicodeData.txt:

[2]Anecdotal evidence provided by Stephen Emmel suggests that this character may also occur in Coptic and Greek. Not being experts in the field of Coptic or Greek palæography, we cannot comment on this matter. It suffices to say that encoding this character in the Supplemental Punctuation block and without the term "Slavonic" in its name makes it readily available for use with Coptic or Greek text should the need arise.

```
2E43;DASH WITH LEFT UPTURN;Po;0;ON;;;;;N;;Used in Church
   Slavonic;;;
```

Above, we described the primary function of the "Dash with Left Upturn" as indicating the end of a sentence, phrase or pericope. In this usage, line breaking commonly occurs following the character. In addition, as we discussed above, the "Dash with Left Upturn" is usually preceded by one or more decorative punctuation characters, such as a colon, middle dot, or three dot punctuation. These characters are never separated from the "Dash with Left Upturn" by a line break, as can be seen in Figures 2, 4, 6, 7 and 8. It is thus proposed that the line breaking property of the "Dash with Left Upturn" be set to Break After (BA). This is also consistent with the line breaking property of the various Greek Paragraphoi already in Unicode. Therefore, the following entry is proposed for addition to LineBreak.txt:

```
2E43;BA # DASH WITH LEFT UPTURN
```

Finally, because this character is used both in the Cyrillic and Glagolitic scripts, it is proposed that the character's script property be set to Common, by adding the following entry to Scripts.txt:

```
2E43          ; Common # Po        DASH WITH LEFT UPTURN
```

This would agree with the script property of other characters in the Supplemental Punctuation block.

References

Cleminson, R., V. Baranov, A. Rabus, D. Birnbaum, and H. Miklas (2010). Proposal for a unified encoding of Early Cyrillic glyphs in the Unicode Private Use Area. *Scripta & e-Scripta* (8-9), 9–26.

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

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

Sever′yanov, S. N. (1922). *Синайская псалтырь: глаголическій памятник XI вѣка.* Памятники старославянского языка. Изд. Отдѣленія русскаго языка и словесности Россійской академіи наук.

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

Figure 1: Example of the Dash with Left Upturn (boxed in red, together with U+2056, THREE DOT PUNCTUATION) used in a modern publication. Source: Karsky (1979, p. 224).

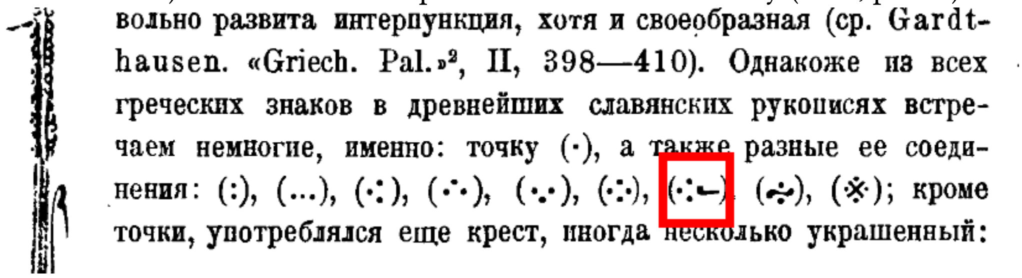

вольно развита интерпункция, хотя и своеобразная (ср. Gardthausen. «Griech. Pal.»², II, 398—410). Однакоже из всех греческих знаков в древнейших славянских рукописях встречаем немногие, именно: точку (·), а также разные ее соединения: (:), (...), (·:), (·˙·), (·.·), (·:·), (·:–), (·÷·), (※); кроме точки, употреблялся еще крест, иногда несколько украшенный:

Figure 2: Examples of the Dash with Left Upturn (boxed in red, together with U+00B7, MIDDLE DOT) used at the end of a pericope and at the end of a heading. Source: Ostromir Gospel.

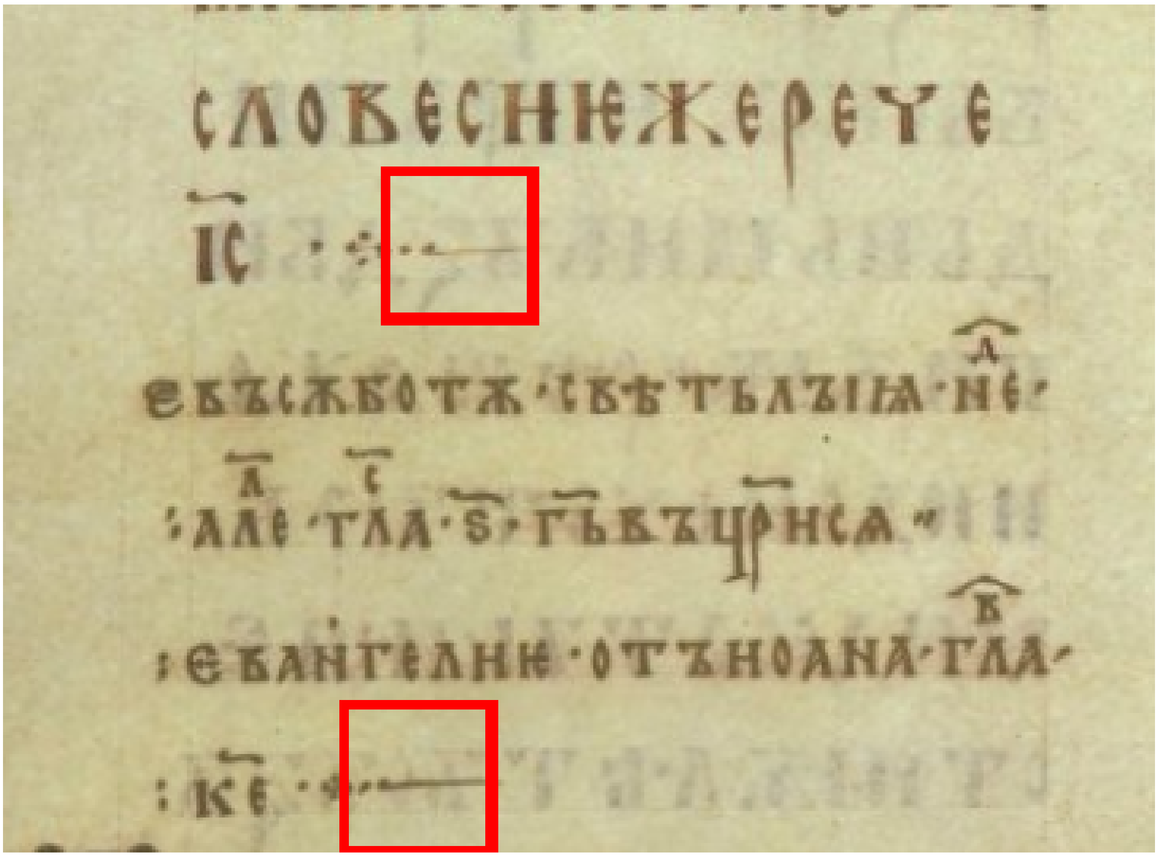


Figure 3: Examples of the Dash with Left Upturn (boxed in red) used at the end of a pericope and at the end of a heading. Source: Sava's Book as reproduced by Schepkin (1903).

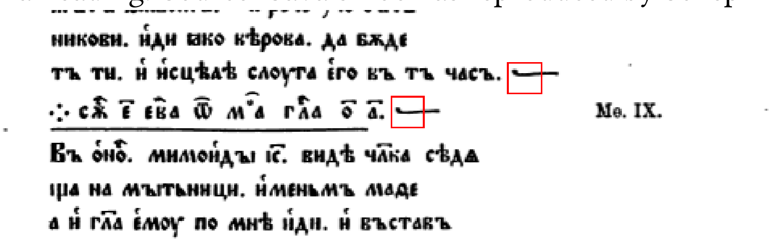

никови. иди ꙗко кѣрова. да бѫде
тъ ти. и исцѣлѣ слоуга его въ тъ часъ.
·:· сл҃ ѕ҃ ева ѿ м҃ѳа гл҃а о҃ а҃. Мѳ. IX.
Въ оно. мимоидꙑ іс҃. видѣ чл҃ка сѣдѧ
ща на мꙑтьници. именьмъ маде
а и гл҃а емоу по мнѣ иди. и въставъ

Figure 4: Examples of the Dash with Left Upturn (boxed in red) used at the end of psalm verses written in the Glagolitic script. Note that the character is preceeded by U+003A COLON. Source: Psalterium Sinaiticum.

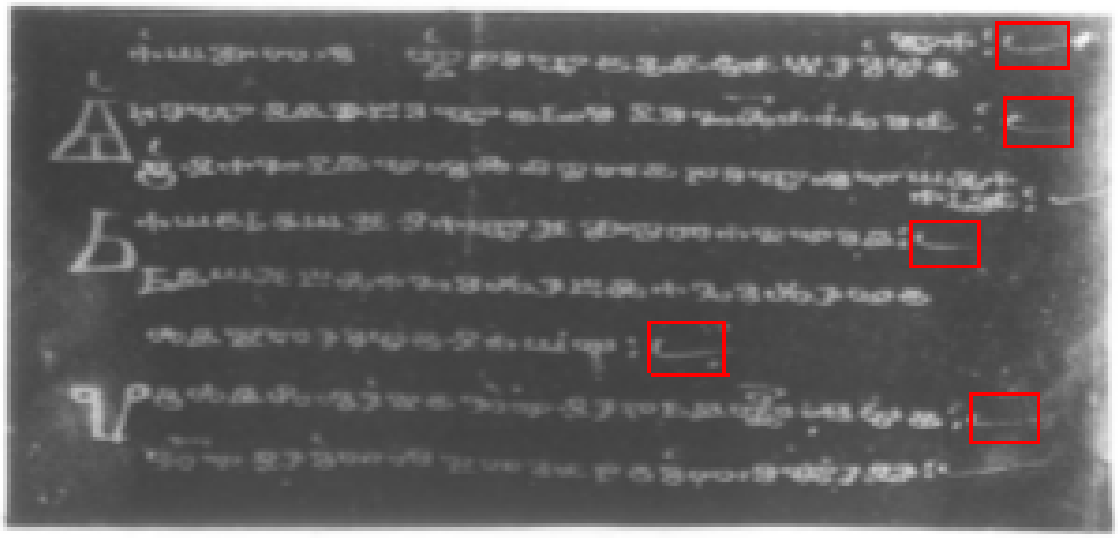


Figure 5: Examples of the Dash with Left Upturn (boxed in red) used at the end of a psalm verse. Note that the character is preceeded by U+003A COLON. Source: Psalterium Sinaiticum as reproduced by Sever'yanov (1922).

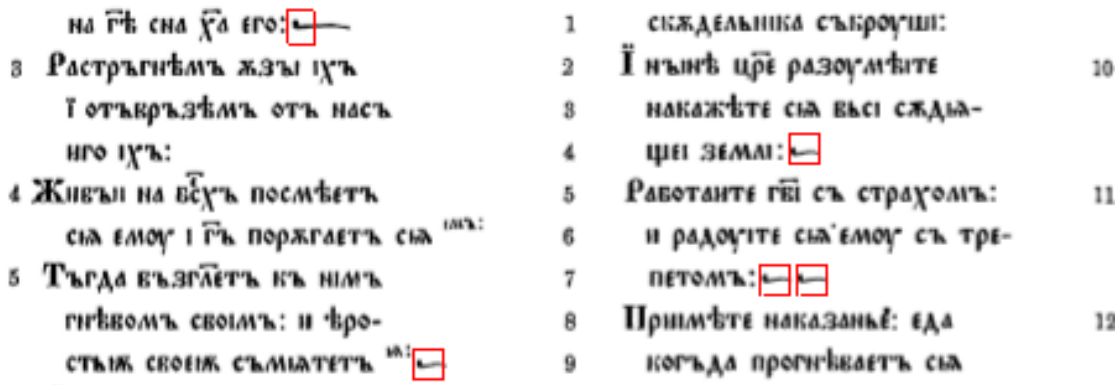


Figure 6: Examples of the Dash with Left Upturn (boxed in red) used to separate lyrics in Znamenny musical notation. Note that the character is preceeded by U+2056 THREE DOT PUNCTUATION. Source: Materials of the 2nd Conference of Practitioners of Znamenny Chant, Moscow, 2004.

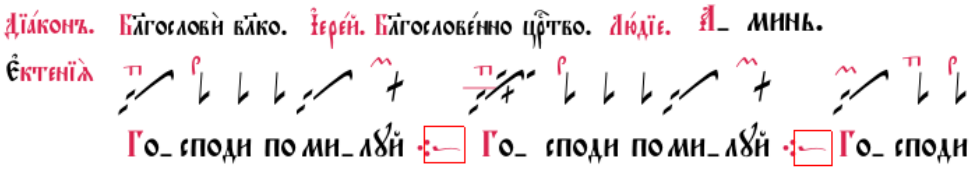


Figure 7: Examples of the Dash with Left Upturn (boxed in red, together with U+003A COLON) used in Slavonic Incunabula. Source: *Apostolos*, Press of Yakov Babich, Vilnius, 1525.

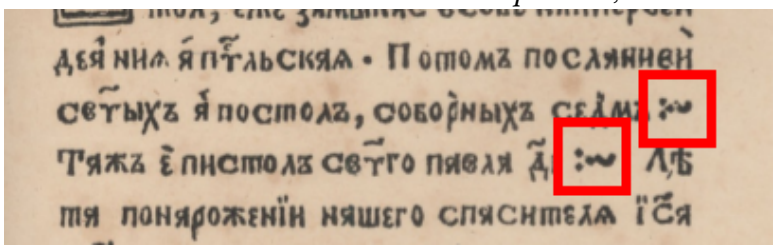


Figure 8: Example of the Dash with Left Upturn (boxed in red, together with U+003A COLON) used in Slavonic Incunabula. Source: *Octoechos*, Press of the Mamonich family, Vilnius, *c.* 1574.

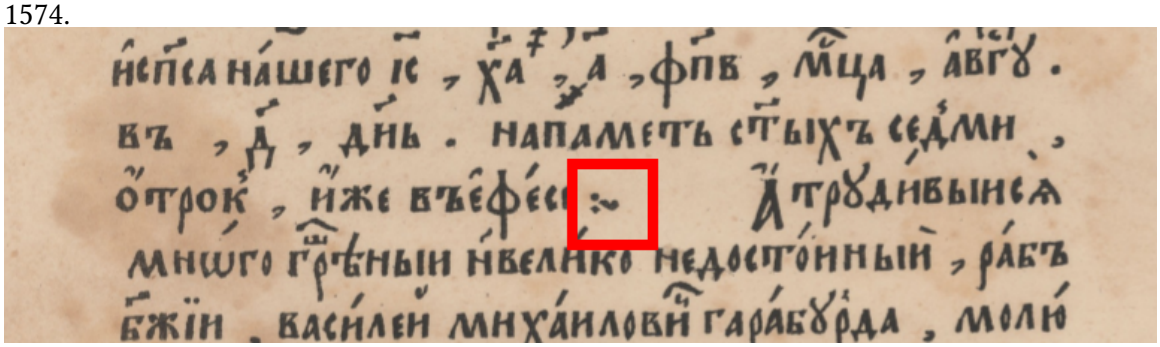
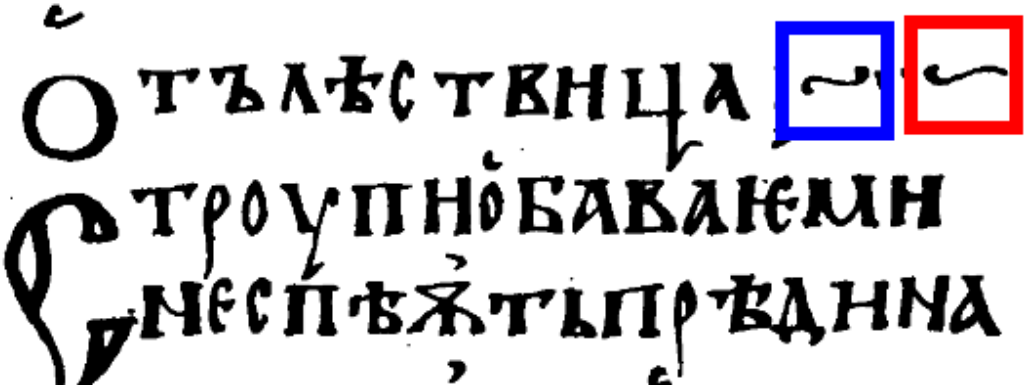


Figure 9: Note the use of both the Dash with Left Upturn (boxed in red) and U+2053 SWUNG DASH (boxed in blue). Source: *Izbornik* of Svyatoslav, near Kiev, 1073, as presented by Karsky (1979).



ISO/IEC JTC 1/SC 2/WG 2
PROPOSAL SUMMARY FORM TO ACCOMPANY SUBMISSIONS
FOR ADDITIONS TO THE REPERTOIRE OF ISO/IEC 10646[1]
Please fill all the sections A, B and C below.
Please read Principles and Procedures Document (P & P) from http://std.dkuug.dk/JTC1/SC2/WG2/docs/principles.html **for guidelines and details before filling this form.**
Please ensure you are using the latest Form from http://std.dkuug.dk/JTC1/SC2/WG2/docs/summaryform.html.
See also http://std.dkuug.dk/JTC1/SC2/WG2/docs/roadmaps.html **for latest *Roadmaps*.**

A. Administrative

1. **Title:** *Proposal to Encode a Slavonic Punctuation Mark in Unicode, Revised*
2. Requester's name: *Aleksandr Andreev, Yuri Shardt, and Nikita Simmons*
3. Requester type (Member body/Liaison/Individual contribution): *Individual contribution*
4. Submission date: *12/13/2013*
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: *Supplemental Punctuation*
2. Number of characters in proposal: *1*
3. Proposed category (select one from below - see section 2.2 of P&P document):
   A-Contemporary ( ) B.1-Specialized (small collection) (X) B.2-Specialized (large collection) ( )
   C-Major extinct ( ) D-Attested extinct ( ) E-Minor extinct ( )
   F-Archaic Hieroglyphic or Ideographic ( ) G-Obscure or questionable usage symbols ( )
4. Is a repertoire including character names provided? *YES*
   a. If YES, are the names in accordance with the "character naming guidelines" in Annex L of P&P document? *YES*
   b. Are the character shapes attached in a legible form suitable for review? *YES*
5. Fonts related:
   a. Who will provide the appropriate computerized font to the Project Editor of 10646 for publishing the standard? *Aleksandr Andreev*
   b. Identify the party granting a license for use of the font by the editors (include address, e-mail, ftp-site, etc.): *Hirmos Ponomar font distributed by Aleksandr Andreev, Yuri Shardt, Nikita Simmons under GNU GPL http://www.ponomar.net/ or aleksandr.andreev@gmail.com*
6. References:
   a. Are references (to other character sets, dictionaries, descriptive texts etc.) provided? *YES*
   b. Are published examples of use (such as samples from newspapers, magazines, or other sources) of proposed characters attached? *YES*
7. Special encoding issues:
   Does the proposal address other aspects of character data processing (if applicable) such as input, presentation, sorting, searching, indexing, transliteration etc. (if yes please enclose information)? *NO*

8. Additional Information:

Submitters are invited to provide any additional information about Properties of the proposed Character(s) or Script that will assist in correct understanding of and correct linguistic processing of the proposed character(s) or script. Examples of such properties are: Casing information, Numeric information, Currency information, Display behaviour information such as line breaks, widths etc., Combining behaviour, Spacing behaviour, Directional behaviour, Default Collation behaviour, relevance in Mark Up contexts, Compatibility equivalence and other Unicode normalization related information. See the Unicode standard at http://www.unicode.org for such information on other scripts. Also see Unicode Character Database (http://www.unicode.org/reports/tr44/) and associated Unicode Technical Reports for information needed for consideration by the Unicode Technical Committee for inclusion in the Unicode Standard.

[1] Form number: N4102-F (Original 1994-10-14; Revised 1995-01, 1995-04, 1996-04, 1996-08, 1999-03, 2001-05, 2001-09, 2003-11, 2005-01, 2005-09, 2005-10, 2007-03, 2008-05, 2009-11, 2011-03, 2012-01)

C. Technical - Justification

1. Has this proposal for addition of character(s) been submitted before? YES
   If YES explain: This is a further revision of L2/13-140, based on feedback from UTC
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; Victor A. Baranov; Ralph Cleminson
   If YES, available relevant documents: E-mail correspondence
3. Information on the user community for the proposed characters (for example: size, demographics, information technology use, or publishing use) is included? YES
   Reference: See Section 2, Description of Character
4. The context of use for the proposed characters (type of use; common or rare) Rare
   Reference: See examples in Proposal
5. Are the proposed characters in current use by the user community? YES
   If YES, where? Reference: Academic literature on Slavonic paleography; Znamenny sheet music
6. After giving due considerations to the principles in the P&P document must the proposed characters be entirely in the BMP? NO
   If YES, is a rationale provided?
   If YES, reference:
7. Should the proposed characters be kept together in a contiguous range (rather than being scattered)? N/A
8. Can any of the proposed characters be considered a presentation form of an existing character or character sequence? NO
   If YES, is a rationale for its inclusion provided?
   If YES, reference:
9. Can any of the proposed characters be encoded using a composed character sequence of either existing characters or other proposed characters? NO
   If YES, is a rationale for its inclusion provided?
   If YES, reference:
10. Can any of the proposed character(s) be considered to be similar (in appearance or function) to, or could be confused with, an existing character? YES
    If YES, is a rationale for its inclusion provided? YES
    If YES, reference: See Section 3, Similar Characters
11. Does the proposal include use of combining characters and/or use of composite sequences? NO
    If YES, is a rationale for such use provided?
    If YES, reference:
    Is a list of composite sequences and their corresponding glyph images (graphic symbols) provided? N/A
    If YES, reference:
12. Does the proposal contain characters with any special properties such as control function or similar semantics? NO
    If YES, describe in detail (include attachment if necessary)
13. Does the proposal contain any Ideographic compatibility characters? NO
    If YES, are the equivalent corresponding unified ideographic characters identified?
    If YES, reference: