Character Set Standardization for Early Cyrillic Writing after Unicode 5.1

A White Paper prepared on behalf of the Commission for Computer Processing of Slavic Manuscripts and Early Printed Books to the International Committee of Slavists by

David J. Birnbaum, Pittsburgh, djbpitt@pitt.edu
Ralph Cleminson, Portsmouth, ralph.cleminson@port.ac.uk
Sebastian Kempgen, Bamberg, sebastian.kempgen@uni-bamberg.de
Kiril Ribarov, Prague, ribarov@ufal.mff.cuni.cz

Preface

This White Paper emerged from discussions among the authors at the Slovo conference that took place in Sofia from 2008-02-21 through 2008-02-26. It is partially a response to three documents published by the Serbian Academy of Arts and Sciences: “Standard of the Old Slavic Cyrillic Script” (“Standard”), “Standardisation of the Old Church Slavonic Cyrillic Script and its Registration in Unicode” (“Standardisation”), and “Proposal for Registering the Old Slavic Cyrillic Script in Unicode” (“Proposal”).[1]

Introduction

The purpose of this White Paper is to provide for the benefit of medieval Slavic philologists:

1. A review of the current state of Unicode with respect to encoding early Cyrillic writing.
2. A brief statement of basic Unicode design principles.
3. An overview of the relationship between character set and font technologies.

[1] Bibliographic information about these and other references may be found at the end of this White Paper.

4. A response to “Standard,” “Standardisation,” and “Proposal” that provides a realistic perspective on the compatibility of these documents with modern character set standards.
5. A discussion of the possible need for further expansion of the early Cyrillic character inventory in Unicode.
6. A discussion of strategies for meeting the encoding needs of Slavic medievalists in a standards-conformant way.

This White Paper is contributed for discussion before and during the September 2008 International Congress of Slavists in Ohrid.

The current state of Unicode with respect to encoding early Cyrillic writing

In February 2007 a group of character set specialists and medieval Slavic philologists submitted to the Unicode Consortium (http://www.unicode.org) and ISO JTC1/SC2/WG2 (http://std.dkuug.dk/jtc1/sc2/wg2/) a formal proposal for the modification (correction and expansion) of Unicode resources for encoding medieval Slavic writing (“N3194R”). This proposal was accepted almost in its entirety at a meeting of the Unicode Technical Committee (UTC) at that time, and the characters proposed therein were registered in Unicode 5.1, which became the official current version of Unicode on 2008-04-04. As a result, Unicode 5.1 contains all early Cyrillic characters for which evidence and argumentation has been presented to the UTC.

Unicode design principles

The authors of this White Paper are not members of the Unicode Consortium (UC) or of the UTC and do not speak for those organizations. We nonetheless have practical experience in submitting successful proposals to the UTC for inclusion in Unicode, and one goal of this White Paper is to provide guidelines for future successful proposals to the UTC for the registration of additional characters needed to encode early Cyrillic writing. The design principles below are common knowledge among scholars of character set standardization, but may be unfamiliar to Slavic philologists. A more detailed explanation of these issues is available in UTC TR 17.

Unicode encodes characters, and not glyphs

A *character* is an *informational* unit that has no canonic physical appearance. A *character set*, such as Unicode, encodes an inventory of characters, assigning to each a standardized canonic name, a byte value, and certain other proper-

ties. The glyphs used to illustrate characters in code charts published by the UC are not normative.

A *glyph* is a presentational unit. A *font* encodes an inventory of glyphs. Glyphs are typically used to represent characters, but a single character may be represented by a variety of different glyphs (roman vs italic vs bold; Times Roman vs Helvetica; etc.).

There need not be a one-to-one relationship of character to glyph, either paradigmatically or syntactically. On the paradigmatic level, a single character (e.g., Latin lower-case "a") may be represented equivalently by roman, italic, bold, etc. and Times Roman, Helvetica, etc. glyphs. On the syntactic level, a single unit of writing (e.g., Latin "a" with acute accent) may be represented as one or two characters and as one or two glyphs, where the number of characters is independent of the number of glyphs.

One consequence of the character/glyph distinction is that Unicode is intended to represent text in an informational, but not necessarily presentationally scrupulous manner. Unicode *plain text* (without markup or other additional non-content information) should be legible, but it is not necessarily intended to meet all of the cultural expectations of users.[2] Unicode is thus the character set layer of text representation; culturally satisfactory rendering may require glyph (font) distinctions, as well. More specifically, Unicode is not intended to be entirely adequate for all typographic purposes. For example, if a user wishes to combine early Cyrillic writing rendered in archaic letterforms with modern Cyrillic writing rendered in modern letterforms, it is expected that the user will employ the same Unicode characters for both purposes, and will achieve any required rendering difference by employing different fonts.[3] **For this reason, we are unable to endorse the suggestion in "Proposal" to register early Cyrillic separately from modern Cyrillic so that "we can have both the**

[2] Unicode 2, pp. 18–19.

[3] Much as plain text in modern Cyrillic does not distinguish whether it is in, for example, Russian or Bulgarian (the modern Bulgarian alphabet is a proper subset of the modern Russian alphabet), or whether it is in roman or italic or bold, or whether it is in Times Roman or Helvetica, Cyrillic text also does not indicate whether it is to be rendered in an archaic typeface with "Slavonic" letterforms or in a modern typeface. Both font and linguistic information are intended to be encoded separately (typically through markup). It is plainly not the case that early Cyrillic writing must be rendered with archaic letterforms; counterexamples in a professional Slavistic context range from Horace Lunt's *Old Church Slavonic Grammar* (first published by Mouton in 1958) through Sebastian Kempgen's modern MacCampus "Kliment" family of fonts. It is similarly not the case that modern Cyrillic writing cannot be rendered with archaic letterforms; see, for examples, pictures of commercial uses of old fonts for modern text at http://kodeks.uni-bamberg.de/AKSL/Schrift/AkslHeute.htm .

contemporary and the Old Slavic script in the same font." There is no technical need to include both in the same font, and even if there were, this type of font-based argumentation is unlikely to prove acceptable to the UTC.

The boundaries between characters and glyphs are not always clear, especially across time,[4] but some guidelines for proposing new characters for inclusion in Unicode are:

1. Are two textual items used contrastively to distinguish information? If they do not distinguish information, they are likely to be regarded as glyphic variants of a single character.
2. Are the distinctions primary (e.g., as in minimal pairs) or are they positionally dependent? If the distinctions are positionally dependent, they are likely to be regarded as glyphic variants of a single character.
3. Are the distinctions found consistently and with reasonable frequency in the system, or are they occasional and idiosyncratic? If they are occasional and idiosyncratic, they are unlikely to be regarded as candidates for standardization. **For these three reasons, we are unable to endorse any argument for registration that depends entirely on the mere coexistence in a single document of different letterforms with comparable orthographic and linguistic function.** For example, within an early Cyrillic context, broad and narrow omicron are candidates for independent registration (and both are registered in Unicode 5.1) because they are used in a way that is linguistically contrastive and not orthographically predictable (independently of language) in some written documents. Tall and short jat′ are not candidates for independent registration (and have not been proposed as such) because they have not been shown to be used in a comparably distinctive way.
4. Are textual items independent or composed? In a writing system that, for example, combines "floating" accentual diacritics freely with base alphabetic characters, Unicode prefers to regard an accented letter as a sequence of two characters, rather than as a single precomposed character. Those precomposed characters that exist in Unicode result from the "grandfather" consideration discussed below. This means, for example, that although many combinations of Latin alphabetic letters with accent marks are registered in Unicode as precomposed unitary characters, the UTC is unlikely to agree to register as precomposed unitary characters

[4] For example, Latin "i" and "j" originated as presentational variants of a single informational unit, but are now regarded as different (and, in most writing systems, substantially unrelated) informational units.

combinations of Cyrillic alphabetic letters with accent marks. **For this reason, we are unable to endorse any proposal for the independent registration of textual items that can be represented adequately (on the informational plane) as sequences of Unicode characters.**

5. Were particular textual items present in registered International Organization for Standardization (ISO) or national character set standards during the initial development of Unicode? Textual units that do not meet the requirements for characterhood were nonetheless included in Unicode if they were already present in other standards of this sort. This "grandfather" approach was viewed as a compromise necessary to facilitate the migration of legacy files to Unicode. It does not establish a precedent for adding new non-characters. **For this reason we are unable to endorse any analogical argument for the registration of characters that relies on the presence of structurally comparable "grandfathered" characters in other scripts.**
6. Is there some sort of consensus within a community of users that a particular text element needs to be registered as a separate character, and that it meets the Unicode requirements for characterhood? Because Unicode is a standard, the UTC is unlikely to register new characters that are required by only a single user for idiosyncratic purposes. **For this reason we are unable to endorse any proposal for the registration of characters for which there is no consensus within at least a meaningful subset of specialists.**

Finally, Unicode encodes characters only if they meet the Unicode requirements for characterhood. This restriction emerges from Unicode design principles, and is based on philosophical and structural considerations, and not solely on the desire to keep the number of registered characters low. **For this reason, we are unable to endorse any proposal for registration that relies on the small number of proposed characters (that is, on the relatively low cost of adding a small number of new items) as an argument for their inclusion.**

Unicode encodes scripts, and not alphabets or orthographies

Unicode alphabetic characters typically belong to a particular *script* (e.g., Latin, Cyrillic, Glagolitic, Greek, etc.). The boundaries between scripts are not always clear, especially across time,[5] but a guiding principle has been to observe mod-

[5] For example, Cyrillic originated as an extension and modification of Greek for use in writing a Slavic language, but is now regarded as a different script, while extensions and modifications of

ern cultural conventions. In the medieval Slavic context, Glagolitic and Cyrillic have always been regarded by Slavists as different scripts, while the use of the term "Cyrillic" for both early and modern Cyrillic writing suggests that despite differences in inventory and (in many uses) appearance, such systems nonetheless belong to the same script. **For this reason, we are unable to endorse any argument for the registration of early Cyrillic as a separate script from modern Cyrillic.**

The relationship between character set and font technologies

For the reasons described above, we are unable to endorse any proposal for the registration of additional early Cyrillic characters in Unicode that should more properly be addressed on the font level, rather than the character set level. The assumption underlying "Standardisation" that certain presentational details should be addressed on the character-set level through Unicode registration reflects a misunderstanding of the architecture of modern operating systems and applications.

As was demonstrated by Zoran Kostić at the azbuky.net conference in Sofia from 2005-10-24 through 2005-10-27 and by Sebastian Kempgen at the Slovo conference in Sofia from 2008-02-20 through 2008-02-26, OpenType font technology supports the simultaneous association of multiple glyphs with a single Unicode character. Among major applications support for this feature of Open Type is currently limited to Adobe InDesign, but OpenType is an open standard that is supported on multiple operating systems, and that is likely to gain wider support in applications over time. Meanwhile, it is already possible to encode such complex character/glyph information in XML in an application-independent way using, for example, the "gaiji" (<g>) mechanism of TEI P5.

Response to "Proposal" and "Standard"

Response to "Proposal"

"Proposal" incorrectly states that Unicode does not contain a large number of characters that it does, in fact, contain. Many of those characters were added in Unicode 5.1 and some were registered in earlier versions of Unicode. The presence or absence of specific characters or candidates for registration is discussed in the review of "Standard," below.

Cyrillic in the twentieth century for use in non-Slavic languages of the former Soviet Union and elsewhere are nonetheless regarded as Cyrillic.

“Proposal” asserts that letter shapes in modern and early Cyrillic are different. To the extent that this is true,[6] it is a font-level consideration and therefore not an acceptable argument for Unicode registration. The illustrative glyphs in UC publications are not normative, and Unicode characters are explicitly said to have no normative shape.

“Proposal” asserts that it is not possible to employ modern and early Cyrillic letterforms in a single font. As is discussed above, this is incorrect in the context of OpenType technology. Furthermore, the ability to render different letterforms for the same informational unit in a single font is a not an acceptable argument for Unicode registration.

“Proposal” asserts that letter names in modern and early Cyrillic are different. This is also true of letter names in, for example, some of the modern languages of Europe that use the Latin script, and it is also true for Cyrillic itself—for example, the letter names that “Proposal” uses are Serbian names, which differ from the (more commonly used) Russian names. Characters names also differ between, for example, modern Bulgarian and modern Russian (e.g., “tvërdyj znak” vs. “er goljam”). More important, however, is the fact that Unicode does not try to preserve character names in their native spelling or pronunciation; rather, it uses the English rendition of such names (e.g., “SHTAPIC”) or descriptive English names (e.g., “COMBINING CYRILLIC LETTER TSE” or “IOTIFIED BIG YUS”). Therefore, naming differences are not an acceptable argument for separate Unicode registration.

“Proposal” asserts that some letters (such as djerv) have the same form but different pronunciation in early and modern Cyrillic. This is also true of letters in, for example, some of the modern languages of Europe that use the Latin script, and the same is also true for Cyrillic characters that are already encoded in Unicode for which “Proposal” does not claim the need for a double registration (e.g.m the original Old Church Slavonic [nasal] pronunciation of ѧ and the later non-nasal Russian pronunciation [ja] for the same letter). It is not an acceptable argument for separate Unicode registration.

“Proposal” asserts that some shared letterforms have different sort order properties in modern and early Cyrillic. This is also true of, for example, some of the

[6] See the discussion in footnote 3, above.

modern languages of Europe that use the Latin script.[7] It is not an acceptable argument for separate Unicode registration.

"Proposal" asserts that "[t]he full and user-friendly application of the Old Church Slavic script requires the registering of numerous letters, ligatures, superimposed letters with and without titlos, a large number of diacritical and punctuation marks and all the Old Slavic numerals." It is true that fine typography may require the availability of all of these *glyphs,* but that fact does not constitute an acceptable argument for their registration as separate Unicode *characters*.

In general, "Proposal" seems to be based on the assumptions that 1) a computer character set must be able to support fine typography, 2) plain text files must be encoded so that they can be printed in a culturally expected way, and 3) there is only a single level of representation. Fine typography and printing are unquestionably important concerns, but computer files may also be created for searching and analysis. The division of textual representation into character and glyph levels and the use of markup (whether explicitly, as in XML, or underlyingly, as in word processors that only appear not to use markup) enable computer files to be used for multiple purposes. Attempting to record all fine typographic distinctions at the character level compromises the use of computer files for anything other than simple rendering.[8]

"Proposal" asserts that certain variant letterforms should be registered separately because they are governed by orthographic requirements (citing as examples narrow, broad, and ocular "o"; az and alpha forms of "a"; short and tall "t"; and the "ou" sequence and vertical "uk" ligature). The variants of "o" and "u" are already present in Unicode 5.1 not because they co-occur, but because their use is governed by principles that are neither arbitrary nor fully dictated by position. Registering variants of "a" and "t" would require similar documentation, and not their mere co-occurrence in a single source.

"Proposal" asserts that all superimposed letters with or without titlos should be registered. Unicode 5.1 accepted for registration a large number of early Cyrillic superscript characters. Some of the authors of this white paper opposed that registration (arguing that superscription should be encoded through markup), but now that the UTC has agreed to encode superscript characters because

[7] For example, in Swedish å, ä, and ö are considered separate letters of the alphabet and are sorted after z, which is not the way ä and ö are regarded or treated in German.
[8] The underlying philosophical issue is that textual units may need to be regarded as the same for some purposes (such as sorting or searching) and as different for other purposes (such as fine typographic rendering). Any encoding represents a compromise between these concerns.

they are used in modern Church Slavonic orthography, it would probably be possible to propose the inclusion of additional ones as long as examples could be found of their use. Superscript letters with titlos, however, would be regarded as a sequence of two Unicode characters (that is, the titlo would be treated as a separable diacritic, to be encoded as a separate character).

“Proposal” asserts that all diacritical marks should be registered. Almost all of the diacritical marks listed there already have been registered in Unicode 5.1 (some are individual characters and some can be composed dynamically from multiple characters), about which see the discussion of “Standard,” below.

“Proposal” asserts that all numerals should be registered, citing the existing registration of Ancient Greek numbers. Unicode 5.1 registers all early Cyrillic number signs as combining characters, and the UTC is unlikely to agree to register additional precomposed numerical representations.

“Proposal” asserts that composite characters may be encoded in only two ways: as combinations of base plus diacritic (two characters) or as precomposed accented characters (one character), and that “solution no. 3 does not exist.” It further argues that precomposed characters are more convenient typographically than floating diacritics (which it labels as “typographically incorrect”). In fact, this analysis reflects a confusion of the character and glyph levels. It is certainly awkward to compose an accented *glyph* dynamically (by superimposing two glyphs) because of differences in width, and a *font* may therefore include a full inventory of precomposed *glyphs*. There is, however, no impediment to representing a single precomposed glyph by a sequence of characters. This is supported directly in OpenType, and it may also be represented at a more abstract level (such as through XML markup) that relies on other font technologies for eventual rendering. There *is*, thus, a third solution: the informational level is represented by characters, the presentational level is represented by glyphs, and there need not be a one-to-one correspondence of character to glyph. The principles underlying this third solution have been part of Unicode design considerations since Unicode 1.0; not only is a one-to-one correspondence not a true technical requirement, but it also not likely to be accepted as a relevant argument by the UTC.

Response to "Standard"

The first two pages of "Standard" list 95 proposed characters (49 "basic" and 46 "functional").[9]

Almost all of the proposed characters are already present in Unicode 5.1, and the others are discussed individually below.[10] All registered Cyrillic characters should have upper- and lower-case forms,[11] and, as noted above, a proposal to register separate superscript characters would probably be accepted. A proposal to register superscript characters both with and without titlo would probably not be accepted; superscript letters with titlo should be encoded as sequences of letter plus floating titlo. The inventory can be divided as follows:

Letters already unambiguously present in Unicode 5.1

Belgrade	*Glyph*	*Unicode Code Point*
1	А а	U+0410/U+0430
2	Б б	U+0411/U+0431
3	В в	U+0412/U+0432
4	Г г	U+0413/U+0433
5	Д д	U+0414/U+0434
6	Ѓ ѓ	U+A662/U+A663
7	Е е	U+0415/U+0435
8	Є є	U+0404/U+0454
9	Ж ж	U+0416/U+0436
10	Ꙁ ꙁ	U+A642/U+A643
11	S s	U+0405/U+0455
12	Ԑ ԑ	U+A644/U+A645
13	Ꙁ ꙁ	U+A640/U+A641

[9] The Belgrade proposal has been revised several times since its original publication, and because it has not been possible to "chase" the frequent revisions, updating this White Paper in response to each one, we cite the version distributed at the February 2008 ASO Slovo conference in Sofia. At the time of writing (mid-July 2008) this is the only openly published version, and is available at the Serbian Academy of Sciences Internet address included in the list of Works Cited.

Subsequent revisions of the Belgrade inventory have responded constructively to some (not all) of the concerns raised in this White Paper (which was circulated in draft form to the principal authors of the Belgrade documents), but because the White Paper is intended primarily as a statement of principles and as a guide for future development, rather than specifically as a point-by-point response to the individual details of the Belgrade proposal, the methodological issues addressed here should be understood from that perspective. The present report does take into consideration modifications proposed by Heinz Miklas during discussions at the ASO Slovo conference.

[10] "Standard" was developed according to definitions of character and glyph that are not entirely compatible with those used by ISO and the Unicode Consortium. The "Standard" document also makes no attempt to collate its items with the Unicode early Cyrillic inventory; we provide such a collation as an appendix to this White Paper.

[11] One exception is discussed below.

Belgrade	*Glyph*	*Unicode Code Point*
14	З з	U+0417/U+0437
15	И и	U+0418/U+0438
18	Ї ї	U+0407/U+0457
19	Ꙇ ꙇ	U+A646/U+A647
20	Й й	U+0419/U+0439
22	Ћ ћ	U+A648/U+A649
23	Ћ ћ	U+040B/U+045B
24	К к	U+041A/U+043A
25	Л л	U+041B/U+043B
26	Ꙗ ꙗ	U+A664/U+A665
27	М м	U+041C/U+043C
28	ꙵ ꙵ	U+A666/U+A667
29	Н н	U+041D/U+043D
30	Ҥ ҥ	U+04A4/U+04A5
31	О о	U+041E/U+043E
32	Ѻ ѻ	U+047A/U+047B[12]
33	Ꙩ ꙩ	U+A668/U+A669
34	Ѳ ѳ	U+A66A/U+A66B
35	Ꙍ ꙍ	U+A66C/U+ACCD
36	❀	U+ACCE[13]
38	П п	U+041F/U+043F
39	Р р	U+0420/U+0440
41	С с	U+0421/U+0441
43	Т т	U+0422/U+0442
47	У у	U+0423/U+0443
48	Ф ф	U+0424/U+0444
49	Х х	U+0425/U+0445
51	Ꙍ ꙍ	U+0460/U+0461
52	Ꙍ̂ ꙍ̂	U+047C/U+047D
54	Ꙍ̃ ꙍ̃	U+047E/U+047F
56	Ц ц	U+0426/U+0446
58	Ч ч	U+0427/U+0447
61	Ꙉ ꙉ	U+040F/U+045F
62	Ш ш	U+0428/U+0448
63	Щ щ	U+0429/U+0449
66	Ъ ъ	U+042A/U+044A
67	Ꙑ ꙑ	U+A650/U+A651
70	Ь ь	U+042C/U+044C
71	Ы ы	U+042B/U+044B

[12] The official Unicode name for this character “round omega” is misleading. Slavists refer to it as “broad o,” understanding “o” as Church Slavonic “on” (= Greek “omicron”).
[13] This character does not have separate upper- and lower-case forms.

Belgrade	*Glyph*	*Unicode Code Point*
73	Ъ ъ	U+A64E/U+A64F
74	Ѣ ѣ	U+0462/U+0463
76	Ꙣ ꙣ	U+A652/U+A653
77	Ю ю	U+042E/U+044E
79	Ꙗ ꙗ	U+A656/U+A657
80	Є є	U+0464/U+0465
81	Ѧ ѧ	U+0466/U+0467
82	Ꙙ ꙙ	U+A658/U+A659
83	Я я	U+042F/U+044F
84	Ж ж	U+046A/U+046B
85	Ӂ ӂ	U+A65A/U+A65B
86	Ѫ ѫ	U+A65E/U+A65F
87	Ѩ ѩ	U+0468/U+0469
88	Ꙛ ꙛ	U+A65C/U+A65D
89	Ѭ ѭ	U+046C/U+046D
90	Ѯ ѯ	U+046E/U+046F
91	Ѱ ѱ	U+0470/U+0471
92	Ѳ ѳ	U+0472/U+0473
94	Ѵ ѵ	U+0474/U+0475

Letters already present in Unicode 5.1 but requiring explanation

- 16/17 (І і). Unicode regards early Cyrillic decimal “i” as equivalent to modern Ukrainian and Belarusian dotted “i” (U+0406/U+0456). The normal glyphs for rendering the upper-case version of this item in both early and modern Cyrillic have no dot. The normal glyph for rendering the lower-case version in modern Cyrillic typography has a dot, while the normal lower-case glyph in early Cyrillic writing does not. Early Cyrillic decimal “i” with a dot should be encoded as regular decimal “i” plus a separate superscript dot character [U+0307].[14]
- 44 (ОУ оу). Originally encoded in Unicode with an ambiguous glyphic description (U+0478/U+0479). Unicode 5.1 recommends that the original code points not be used. It creates new code points for vertical “uk” (U+A64A/U+A64B) and recommends that the horizontal “ou” digraph be encoded as a sequence of two characters (“o” is U+041E/U+043E; “u” is U+0423/U+0443).
- 45/46 (ᲈ Ᲊ). An unambiguous vertical “uk” ligature was added to Unicode 5.1 (U+A64A/U+A64B). 45 and 46 should be regarded as glyphic variants

[14] Like any other sequence of this type, this might be mapped to a single glyph for rendering purposes.

of the same underlying character unless they can be shown to satisfy the Unicode requirements for being distinct characters.

- 64 (ШТ шт). Should be encoded as a sequence of "sh" (U+0428/U+0448) plus "t" (U+0422/U+0442).

Letters not present in Unicode 5.1 but possibly candidates for registration

- 21 ("variant djerv"). This should be registered only if it can be demonstrated that it is not merely a presentational variant of djerv (U+A648/U+A649).
- 37 ("double omicron"). This should be registered only if it can be demonstrated that it is not merely a presentational variant of omega (U+0460/U+0461).[15]
- 40 ("soft r"). Soft letters spelled with a physically separate superscript palatalization mark (sometimes offset to the right) should be encoded as sequences of characters. For example, palatal "l" with this spelling should be encoded as regular "l" (U+041B/U+043B) followed by a palatalization hook (U+0484). Soft letters spelled as actual physical combinations are registered separately (thus 6 [U+A662/U+A663], 26 [U+A664/U+A665], 28 [U+A466/U+A467], 30 [U+04A4/U+04A5]). For these reasons, 40 should be registered only if it can be shown to have the latter form (i.e., single physically continuous glyph), rather than base "r" plus physically discrete palatalization hook.[16]
- 42 ("soft s"). According to Heinz Miklas, this item is misrepresented in "Standard" as Cyrillic "s" with an attached palatalization handle, instead

---

[15] Heinz Miklas (p.c.) notes that one might be tempted to explain the instances of closed omega in the *Bdinski sbornik* as the scribe's individual rendering of double omicron, reflecting length according to Serbian orthographic tradition.

[16] Heinz Miklas (p.c.) explains that palatal "r" was represented in some manuscripts (including the *Ostromir Gospel*) by a special form of the following vowel letter (such as "a" with a hook to the left). Whether these hooked shapes should be encoded as independent characters in Unicode or as glyphic variants of jotated vowels (e.g., jotated "a") remains to be determined, but they do not justify encoding a separate palatal "r" character.

During subsequent discussion, Heinz Miklas mentioned a need for soft velars (used to represent palatals). Since these are normally rendered in Cyrillic with a floating palatalization diacritic, they should be encoded as sequences of two characters. Alternatively, it would be possible in a font intended to render early Cyrillic to map glyphic combinations of base "к" or "г" plus palatalization hook to the character cells registered for the modern Macedonian palatals ("Ќ/ќ" = U+040C/U+045C; "Ѓ/ѓ" = U+0403/U+0453). This is marginally legitimate insofar as both spellings employ a velar base plus a diacritic otherwise used to represent softness to designate a palatal. It has, however, two disadvantages: 1) it incorrectly implies that the palatalization hook is a historical variant (paleographically) of an acute accent and 2) its use would result in encoding the voiceless and voiced palatal stops as single characters but the voiceless palatal fricative as a sequence of two characters (U+0425/U+0445) followed by U+0484, which would complicate processing.

of as broad Cyrillic “s”, which is used in the *Codex Suprasliensis* apparently to represent palatal /ś/. This should be registered only if it can be demonstrated that broad and narrow “s” must be distinguished in order to encode the *Codex Suprasliensis* without losing graphemic information.

- 50 (“spidery [or sunny] x”). Heinz Miklas subsequently explained that this was intended to represent soft /x/. As such, it should be encoded in Cyrillic as a sequence of “x” plus separate palatalization diacritic. See the discussion above.
- 53 (“Glagolitic omega”). Should be registered as a Cyrillic character only if it can be shown to be used as a meaningful part of Cyrillic writing. Glagolitic writing shows up in occasional and irregular ways in Cyrillic texts, and should be treated as Cyrillic (and not intrusive Glagolitic) only if it can be shown to satisfy the requirements for *Cyrillic* characterhood when it occurs within a Cyrillic context.
- 55 (“omega with superscript d”). 54 (U+047E/U+047F) is registered because it is viewed in certain contexts as an independent letter of the alphabet.[17] 55 should be registered only if similar arguments can be advanced on its behalf. Otherwise it should be encoded as a sequence of omega (U+0460/U+0461) plus a superscript “d”.[18]
- 57 (“soft ts”). Should be registered only if it can be shown to function as a distinctive part of some Cyrillic writing, and not merely as a glyphic variant of regular “ts” (U+0426/U+0446).
- 59 (“variant ch”). Apparently glyphic variant of regular “ch” (U+0427/U+0447). Should be registered separately only if the glyphs in question can be shown to satisfy the definition of distinct characters.
- 75 (“second jat′”). Should be encoded only if it can be shown not to be a glyphic variant of 74 (U+0462/U+0463) according to the Unicode definition of character.
- 78 (“jotated uk”). Should be encoded only if it can be shown not to be a glyphic variant of 77 (U+042E/U+044E) according to the Unicode definition of character.
- 93 (“long izhica”). Should be encoded only if it can be shown not to be a glyphic variant of 94 (“U+0474/U+0475”) according to the Unicode definition of character.

---

[17] In fact, in some cases “ot” is treated as a separate letter of the Slavonic alphabet even when the omega and “t” are written consecutively and in-line, and not only when the “t” is superscripted over the omega. See Figure 25 in N3194R.

[18] Regular “d” is U+0414/U+0434. Separate superscript “d” is U+2DE3. See the discussion of superscription above.

Letters not candidates for registration

- 60 (“glagolitic ch”). Withdrawn by Heinz Miklas.
- 65 (“glagolitic sht”). Withdrawn by Heinz Miklas.
- 68/69/72 (variants of “jery”). In the early period jery was composed dynamically from back jer plus any of the two Cyrillic or three Glagolitic “i” letters. The dynamic composition means that such spellings should be encoded as sequences of two characters.[19]
- 95 (“izhica with two dots”). This can be composed from regular izhica (U+0474/U+0475) and superscript diaeresis (U+0308).

Diacritics, punctuation, and symbols

“Standard” lists 49 diacritical marks and 26 punctuation marks and symbols. All are present in Unicode 5.1 with a small number of exceptions, which may be proposed for registration if they can be shown to satisfy the Unicode definition of character, and not be to glyphic variants of characters that are already registered. It should be noted that Unicode includes diacritics, punctuation, and symbols that may be used within all scripts (including Cyrillic), as well as those that are script-specific. Cyrillic-specific items in these categories can be proposed only if they do not already exist as either Cyrillic or general characters. Note that some complex punctuation may be already be registered as a unitary character, while other complex punctuation can be constructed as a sequence of characters (e.g., ⁘ may be encoded as two middots [U+00B7] with a colon [U+003A] between them or as a single character).

Items in these categories not present in Unicode 5.1 are:

- Diacritics 31/32/39. 31 and 32 appear to be glyphic variants of 30 (U+1DC3) and 39 appears to be a glyphic variant of 37 (U+0483). Should be registered only if they can be shown be shown to function as distinctive parts of some Cyrillic writing, and not merely as glyphic variants of other registered characters.
- Diacritics 41/42/43/44/45. “Standard” refers to these as titlos located between characters. Unicode already includes a unitary equivalent of 41 (U+035E), and the others can be composed by using a zero width joiner (U+200D) and the appropriate non-spacing diacritic.

---

[19] In subsequent discussion Heinz Miklas also suggested adding both front and back jer followed by “i” with two dots. These should be spelled as sequences of two characters for the same reason. Note that spelling jery with two characters does not imply that it represents two sounds; much as there need not be a one-to-one correspondence between characters and glyphs, there need not be a one-to-one correspondence between characters and sounds or, for that matter, between glyphs and sounds.

- Diacritics 46/47/48/49. “Standard” refers to these as signs (znakovi) between characters. They, too, may be encoded by inserting a zero-width joiner plus a non-spacing diacritic between the two base letters.
- Punctuation/symbol 11. Appears to be a glyphic variant of 10 (U+00F7). Should be registered only if it can be shown be shown to function as a distinctive part of some Cyrillic writing, and not merely as a glyphic variant of other registered characters.
- Punctuation/symbol 19/20. Jotated characters are registered as they are proposed. Accordingly, separate combining jotation bars should not be registered.
- Punctuation/symbol 21. Appears to be a glyphic variant of 22 (U+2E13). Should be registered only if it can be shown be shown to function as a distinctive part of some Cyrillic writing, and not merely as a glyphic variant of other registered characters.

Numbers

“Standard” lists 92 precomposed numbers (combinations of number marks plus letters), 82 “basic” and 10 “functional.” None of these should be registered separately; all should be encoded as combinations of base letters plus combining numerical characters. It also lists the 7 required combining numerical characters, all of which are already present in Unicode 5.1.

Possible need for the expansion of the early Cyrillic inventory in Unicode

Any future expansion of the early Cyrillic inventory in Unicode will require a demonstration that new candidates for inclusion satisfy the Unicode definition of character. This is clearly not the case with most of the items listed in “Standard” that are not currently part of Unicode 5.1, but it may be possible to support the inclusion of a small number of items, discussed individually above, in the future.

Strategies for encoding early Cyrillic writing in a standards-conformant manner

Slavists who wish to encode early Cyrillic writing should do so in a standards-conformant manner so that the resulting documents will not be dependent on specific software or fonts. This means using the Unicode (currently 5.1) character inventory and fonts designed to support a Unicode environment. Users who

require additional characters not registered in Unicode or glyphic variants of standard Unicode 5.1 characters have three options:[20]

1. They can change fonts, representing, for example, the writing of each time and place with a font that represents the writing of that time and place in a culturally acceptable manner.
2. They can represent different glyphs with Unicode Private Use Area (PUA) characters. In this context, it would be helpful if the paleoslavistic community could agree on a sort of microstandardization of a portion of the PUA for that purpose.
3. They can use Open Type technology (which includes glyph selection, substitution, ligation, etc.) to select appropriate glyphic forms. In this case it would be helpful if the paleoslavistic community could agree on a sort of microstandardization of glyph references, which could facilitate coordinated and consistent font development.

Although in general "Standard" is not suitable as a character set inventory or as the basis for an eventual proposal to the UC, it could nonetheless make a meaningful contribution to Slavic electronic text technology both as a rich glyph inventory and as a resource for the microstandardization of the PUA or of Open Type glyphic variants.

Works cited

N3194R

"Proposal to encode additional Cyrillic characters in the BMP of the UCS." ISO/IEC JTC1/SC2/WG2 N3194R. Submitted by the UC Berkeley Script Encoding Initiative (Universal Scripts Project). 2007-03-21. http://std.dkuug.dk/jtc1/sc2/wg2/docs/n3194.pdf .

Proposal

"Proposal for Registering the Old Slavic Cyrillic Script in Unicode." Adopted at the International Conference Held in the Serbian Academy of Arts and Sciences from 15–17 October 2007, Organized by the SASA Language and Literature Department and the SASA Institute for the Ser-

[20] By additional characters not registered in Unicode we mean items that satisfy the Unicode requirements for characterhood, but that have not (yet) been proposed and accepted for inclusion. Their encoding through other means would thus be a temporary solution until they can be registered and assigned an official code point outside the PUA. By glyphic variants of standard Unicode characters we mean items that do not satisfy the Unicode requirements for characterhood, but that some users may nonetheless wish to encode distinctively.

bian Language.
http://www.sanu.ac.yu/Cirilica/Prilozi/Unicode-Explanation.pdf .

Standard

"Standard of the Old Slavic Cyrillic Script." Adapted at the International Conference held in the Serbian Academy of Arts and Sciences from 15–17 October 2007, Organized by the SASA Language and Literature Department and the SASA Institute for the Serbian Language.
http://www.sanu.ac.yu/Cirilica/Prilozi/Standard.pdf .

Standardisation

"Standardisation of the Old Church Slavonic Cyrillic Script and its Registration in Unicode." Conclusions adopted at the international academic conference held at the Serbian Academy of Sciences and Arts (SANU), 15–17 October 2007, organized by the SANU Division for Language and Literature and the SANU Institute for the Serbian Language.
http://www.sanu.ac.yu/Cirilica/Eng.aspx .

TEI P5

TEI: P5 Guidelines. Representation of Non-standard Characters and Glyphs. http://www.tei-c.org/release/doc/tei-p5-doc/en/html/WD.html

Unicode 2

The Unicode Standard 5.0 – Electronic edition. Chapter 2. General Structure. http://www.unicode.org/versions/Unicode5.0.0/ch02.pdf.

UTC TR 17

Whistler, Ken, Mark Davis, and Asmus Freytag. "Unicode Technical Report #17. Character Encoding Model." Version of 2004-09-09.
http://www.unicode.org/reports/tr17/ .