Old Cyrillic in Unicode*

Ivan A Derzhanski
Institute for Mathematics and Computer Science, Bulgarian Academy of Sciences
iad@math.bas.bg

The current version of the Unicode Standard acknowledges the existence of a pre-modern version of the Cyrillic script, but its support thereof is limited to assigning code points to several obsolete letters. Meanwhile mediæval Cyrillic manuscripts and some early printed books feature a plethora of letter shapes, ligatures, diacritic and punctuation marks that want proper representation. (In addition, contemporary editions of mediæval texts employ a variety of annotation signs.) As generally with scripts that predate printing, an obvious problem is the abundance of functional, chronological, regional and decorative variant shapes, the precise details of whose distribution are often unknown. The present contents of the block will need to be interpreted with Old Cyrillic in mind, and decisions to be made as to which remaining characters should be implemented via Unicode's mechanism of variation selection, as ligatures in the typeface, or as code points in the Private space or the standard Cyrillic block. I discuss the initial stage of this work.

The Unicode Standard (Unicode 4.0.1) makes a controversial statement:

> The historical form of the Cyrillic alphabet is treated as a font style variation of modern Cyrillic because the historical forms are relatively close to the modern appearance, and because some of them are still in modern use in languages other than Russian (for example, U+0406 "I" CYRILLIC CAPITAL LETTER I is used in modern Ukrainian and Byelorussian). Some of the letters in this range were used in modern typefaces in Russian and Bulgarian. Prior to 1917, Russian made use of *yat*, *fita*, and *izhitsa*; prior to 1945, Bulgaria made use of these three as well as the *big yus*.

(The last statement is incorrect, incidentally: Ѳ and Ѵ went out of use in Bulgarian a century earlier.) In fact, the historical forms that are in modern use (or have been used in modern typefaces) are negligibly few, compared to the ones that aren't (and haven't); likewise, there is a plethora of modern forms with no historical prototypes (the ones invented expressly for non-Slavic languages). None the less, there are other benefits from treating the two forms of the script as style variations. There is the fact that oldstyle typefaces are often used for setting text in Modern Bulgarian, Russian and Serbian in an 'Olde Cyrillick' style, which may be an argument for designers of Old Cyrillic typefaces for cutting oldstyle forms of letters that are used now in these languages, though not in mediæval manuscripts and even in books such as the *Fish Primer* (Beron 1824), a book printed in an oldstyle script but with modern-style capitalisation. On the other hand, Gerov's *Dictionary*, published at the turn of the 20th century (Gerov 1895–1908), is set in a modern typeface, but with two 'historical' letters that haven't made it into Unicode's Cyrillic plane and one incorrectly analysed there.

This implies a complex relationship between the two forms of the script. For example, mediæval manuscripts feature several shapes of the little jus (Ѧ ѧ ѧ ѧ ҿ), as well as the letter Ѩ (with its variant Ӭ). In the *Fish Primer* the variety is reduced to one form of each letter. Gerov's *Dictionary* gives the letters the shapes я Я and ıa IA (with connecting bars), respectively. In modern usage only the former letter survives; it is not readily recognised as a style variation of ѧ Ѧ, though, so oldstyle Cyrillic typefaces can usefully contain both.

Several kinds of characters need to be accounted for:

---

* Work done in Prague in February–April 2005; especially discussions with Zdenka Ribarová. Most information on the historical usage of letters and variant forms is from Karskij 1928 and Andrejčin 1977; the inventory of letters is partly based on the appendix to Camuglia & Ribarov (ms.).

- letters (majuscule and minuscule), including variant letter shapes;
- superscript letters;
- ligatures;
- diacritics;
- punctuation;
- annotation signs.

Of these, only the first category is relevant to both historical and modern Cyrillic; contrariwise, the last one (or perhaps the last three) can be presumed to be very similar to the needs of mediævalists working with other scripts (especially Greek and Roman).

It is a large part of the problem that variation comes in many kinds. It may be phonetic (if the letters stood for different sounds), functional (if they stood for the same sound but were associated with different positions in the word, or with individual words), chronological, regional or merely decorative. Many pairs of related shapes are obtained by reversal (ѕ ƨ), by the use of diacritics (і ї ї, ѡ ѡ̈, ѵ ѵ̋), by variation of width (ҁ ⊂), by enlisting Glagolitic letters (ⰵ ⱘ, ⰰ ⰴ); there is also much unclassifiable variation (т м 7). It stands to reason that similar cases should be treated in similar ways if possible.

For each shape, the following options are available:

- Find a matching shape to unify it with
  - in the Cyrillic plane,
  - in another public plane,
  - in the part of the private area used by the Medieval Unicode Font Initiative (MUFI).
- Relay it to Unicode's mechanisms for handling ligatures and diacritics, if it can be viewed as a compound character.
- Place it in a vacant part of the private area.
- Negotiate its inclusion into the primary Cyrillic space.

It is a general policy of Unicode to unify diacritics and punctuation across scripts but to split letters, so a Cyrillic letter shouldn't be unified with a Greek or Roman one (either from the public planes or from MUFI's part of the private area), but diacritics, punctuation and annotation signs may be.

For each pair of historical shapes which kind of are the same, but kind of aren't (e.g., ҙ and ӡ, both of which correspond – in function as well as history – to modern 3, U+0437), these options exist:

- Call them the same thing, and postulate that a font will contain one or the other in position U+0437, but never both. (No two shapes are ever identical in handwriting; the line needs to be drawn somewhere.) For this, it is a necessary condition that the shapes are not found in the same text (meaning that this is the obvious strategy for the treatment of chronological or regional variants); it may not be a sufficient condition, however (Unicode does separate some variant Greek letters, such as U+03B8 θ and U+03D1 ϑ).
- Call them different glyphs for the same character, and postulate that if a font contains both (which it needn't do), then one will be generated by the character code alone (U+0437 3) and the other by the same code with a variation selector, a little-used part of the Unicode toolbox (3+VSn). This is advisable if the two are in free variation in documents of the same tradition, but it is desirable to preserve the difference.
- Call them different characters, unify the first (which won't always be easy to identify) with the character code in the public plane (ӡ=U+0437 3), and

- find another code in the same public plane to unify the second with (the obvious choice for ҙ would be U+04E1 ӡ, a letter used in Abkhaz for the affricate /ʒ/, in view of the shape), or
- use a code from a private area, and perhaps aim for inclusion into the primary space.

A tentative Old Cyrillic-minded interpretation of the Cyrillic Unicode plane

§1. Modern Cyrillic letters that never had Old Cyrillic forms, nor need any:

048A–048F	Ҋҋ Ҍҍ Ҏҏ
0492–04CE	Ғғ Ҕҕ Җҗ Ҙҙ Ққ Ҝҝ Ҟҟ Ҡҡ Ңң Ҥҥ Ҧҧ Ҩҩ Ҫҫ Ҭҭ Үү Ұұ Ҳҳ Ҵҵ Ҷҷ Ҹҹ Ӄӄ Ӆӆ Ӈӈ Ӊӊ Ӌӌ Ӎӎ
04D2–04DF	Ӓӓ Ӕӕ Ӗӗ Әә Ӛӛ Ӝӝ Ӟӟ
04E2–04F5	Ӣӣ Ӥӥ Ӧӧ Өө Ӫӫ Ӭӭ Ӯӯ Ӱӱ Ӳӳ Ӵӵ

The letters ө Ө and ү Ү (used respectively for *ö* and *ü* in Altaic languages), though they look like mere graphic modifications of о О and у У, in fact go back to ѳ Ѳ and ѵ Ѵ, which were abolished in 1918 (Musaev 1982:22–23), but the former pair is separated by the sounds, and the latter by the shapes.

§2. Modern Cyrillic letters that weren't used in pre-Modern times, but can use oldstyle forms for the purpose of writing text in Slavic languages in an 'Olde Cyrillick' style:

0400	0450	Ѐ	ѐ
0401	0451	Ё	ё
0403	0453	Ѓ	ѓ
0408	0458	Ј	ј

0409	0459	Љ	љ
040A	045A	Њ	њ
040C	045C	Ќ	ќ
040D	045D	Ѝ	ѝ

042F	044F	Я	я

A ј-like letter is found, in fact, in Russian cursive, but as a variant of і.

The letter я superseded ѧ (as Gerov 1895–1908 explicitly states), though graphically it is a descendant of both ѧ and ѩ, which fell together in speedy (Russian) writing. Their separation (an artefact of Unicode) is perhaps advantageous, as oldstyle Я я letters have been cut, for the purpose of setting 'Olde Cyrillick' text in contemporary Bulgarian and Russian; one might also note that Dal' 1880–1882 lists 'Ѧ' as a separate entry in his *Dictionary*, as a 'nasalised [vowel] still heard in Polish and of interest only to the linguist'.

§3. Letters with regional or chronological variants, of which presumably one form should be chosen in a typeface. Codes of the form Kyr# and Kys# are assigned to letters not currently supported by Unicode:

0410	0430	А	а	ѧ	or 'alpha' (similar to MUFI's U+F200 LATIN SMALL LETTER INSULAR A), a chronological semi-uncial variant.
0411	0431	Б	б	ѣ	or 'tilted Б' (90° clockwise), a Bosnian variant.
0412	0432	В	в	ѣ	or 'loopy B' (a variant used in some texts), or one tilted (90° anticlockwise), frequent in 17c texts.
Kyr6	Kyr7			ҙ	or ditto with the crossbar going through the entire width of the letter.

041C	043C	М	м	ᴍ	or ditto with a deeper middle loop.
0420	0440	Р	р	ᴩ/ᵨ	
0422	0442	Т	т	ᴛ/ᵀ	Other variants include a 7-like shape.
Kys0	Kys1			ᵾ/ᵿ	Non-initial letter (as opposed to the digraph оу). Has a variety of shapes (ᵾ ȣ; ᵿ ᵾ). The Latin Extended-B letters Ȣ U+0222 LATIN CAPITAL LETTER OU and ȣ U+0223 LATIN SMALL LETTER OU are available in Code2000, but not in Arial Unicode MS, and at any rate they pertain to a different script.
0424	0444	Ф	ф	ᵩ/ᵠ	
0426	0446	Ц	ц	ᶇ	or ditto with a shorter and more distant tail, or reversed.
0427	0447	Ч	ч	ʏ/ч	
0460	0461	Ѡ	ѡ	ᴡ/ω	
0472	0473	Ѳ	ѳ	ᵩ/ᵠ	

§4. Letters which tend not to appear in oldstyle Cyrillic typefaces:

040F	045F	Ꙗ	ꙗ	Old Cyrillic form used in Beron 1824.
0490	0491	Ґ	ґ	'hard Г', in South Russian/Galician/Ukrainian since 17c.
Kys4	Kys5			The Cyrillic letter Yn, used for *î*, cooccurs in Roumanian with ѫ, used for *ă*, and so must be treated as a separate letter, not a local variant.

§5. More difficult cases. In the tables **0** indicates the primary variant (for which there may be more than one possible shape, the choice then being left to the designer of the typeface), and **1**, **2** and **3** are optional variants, to be obtained with variation selectors:

				0	**1**	**2**	
0402	0452	Ђ	ђ	ћ?			These two modern Serbian letters are descendants of the same Old Cyrillic letter, which is closer to ћ in shape, but was originally closer to ђ in sound, being voiced; its use for both voiced and voiceless sounds dates from the 14c.
040B	045B	Ћ	ћ	ћ?			
0404	0454	Є	є	∈			A broader shape was mostly used word-initially and a narrower one word-medially/finally. Allowing for a matching difference in sound (indeed, ∈ replaced ѥ by the 14c), the unification of these two shapes with є (Ukrainian /je/) and е respectively seems appropriate, although the modern uppercase shapes differ in roundness, and the lowercase ones in openness, rather than width.
0415	0435	Е	е	є	ϵ		
042D	044D	Э	э	ɜ		϶	A mediæval Bulgarian shape, originally interchangeable with ∈ or є (of which it was a mere graphic variation), but a separate letter in Modern Russian and most alphabets derived from it.

041E	043E	O	o	ѻ	ѻ		A broader shape was mostly used word-initially and a narrower one word-medially/finally.
047A	047B	ⵔ	ⵔ	ⵔ			
KyrA	KyrB			⊙			Used mostly, though not exclusively, instead of ⵔ in forms and derivatives of the word ⊙ѻ.
KyrC	KyrD			⊙			Used word-initially/postvocalically from the 14c onwards: ⊙ѻ. (Perhaps a graphic variant of ⊙?)
KyrE	KyrF			ꙮ	ꙮ		Used in дꙮє, ѻбꙮє etc.; also in ꙮкоу. (Even this is not the whole story: ꙮ appears in one place in the word ꙮчи, and there is a 'multiocular' shape – appropriately in a word with this very meaning – made of 10 ⊙s in a 15c psalter, but that is by no means a regular part of the script.)
0460	0461	Ꙍ	ѡ	ѡ	ӧ		Both ѡ and ∈ were used word-medially instead of ѻ and є for disambiguating between certain pairs of homonymous word forms, which is to say that both ѡ and ⵔ were to ѻ as ∈ was to є, though in different ways (an argument in support of considering ⵔ a form of ѡ, as the Unicode description ROUND OMEGA suggests). The diæresis was decorative.
047C	047D	Ꙍ̂	ѡ̂	ѡ̅			The same thing as ѡ with titlo (and in fact no different from any other letter with this diacritic), as currently defined.
047E	047F	Ꙍ̄	ѡ̄	ѡ̄			The same thing as ѡ with superscript т.
0421	0441	C	c	с		⊂	The wide variant is found very seldom, always word-initially. Thus the distribution of the variants of с parallels that of the variants of е and о, obviously because of the similarity of the forms of the letters, a phonetic difference being out of the question here.
0405	0455	S	s	ѕ			ѕ is sometimes an episemon (6), but ꙅ is much more common in this function.
Kyr4	Kyr5			ꙅ			
Kyr6	Kyr7			ҙ			Has the same sound value (/ζ/) as ѕ/ꙅ, but isn't graphically related to them, and is never used as an episemon.
04E0	04E1	Ʒ	ʒ	ҙ?			The modern letter ʒ is used in Abkhaz for /ζ/, and has been proposed for /ð/ in some Samoyedic (Tereščenko 1986) and Tungusic languages. If it is considered too exotic as a counterpart of ҙ, the latter should be treated as a graphic variant of з.
0417	0437	З	з	з	ҙ?		
0406	0456	І	і	ı	і?		Mostly an episemon (10). As a letter, used primarily as a prevocalic or line-final form; also in the name іс 'Jesus'. The variant with one dot is very rare historically. In Beron 1824 the capital version has no dots, whereas the lowercase one has *two* unless it bears an accent mark.

0407	0457	Ї	ї	ї	і?		Rare form, used word-initially or after и. Has a variant with a double acute. (Modern version used in Ukrainian for /ji/.)
Kyr8	Kyr9			ι			Only used in transcriptions of Glagolitic text.
0418	0438	И	и	н/и	н?	и?	It could be postulated that if both shapes exist in a typeface, then и+VS1 must generate a letter with a horizontal bar (the original one) and и+VS2 one with a tilted bar (the modern one, used in Beron 1824), and what happens without a variant selector is up to the typeface.
041D	043D	Н	н	N/н	N?	н?	Similarly, if both shapes exist in a typeface, н+VS1 must generate a letter with a tilted bar (the original one, used in Beron 1824) and н+VS2 one with a horizontal bar (the modern one), and what happens without a variant selector is up to the typeface.
0423	0443	У	у	ү			The historical letter ү was a graphic variant of ѵ; the modern letter у was introduced in the place of ꙋ.
0474	0475	Ѵ	ѵ	ѵ			
0476	0477	Ѷ	ѷ	ѷ			The double grave was decorative, as it appears.
Kys2	Kys3			ꙋ			A variant Bulgarian Glagolitic letter (not found in Croatian Glagolitic), only used in one word: ꙋлъми ‘βουνοί’. It is unknown whether this reflects an actual (i.e., phonetic) difference.
0466	0467	Ѧ	ѧ	ѧ/ѧ			Word-initial/postvocalic letter in some traditions (as opposed to ꙙ). Postconsonantal in others (as opposed to ꙙ or ѩ). Postconsonantal letter in Church Slavonic and in Beron 1824 (as opposed to ѩ).
KysC	KysD			ꙙ/ꙙ			Word-initial/postvocalic letter in some traditions (as opposed to ѧ). Postconsonantal in others (as opposed to ѧ or ꙝ).
KysA	KysB			ⰵ			A Glagolitic letter, only used in Cyrillic in one word: ѧⰵг҃лъ ‘ἄγγελος’. (The word is also spelt аггелъ, анћелъ and аньѥлъ.)
046A	046B	Ѫ	ѫ	ѫ/ѫ			The latter seems to be a palæographical variant.

§**6.** Letters with breves. They can be obtained by using U+0306 COMBINING BREVE; three of them, however, are available as precomposed characters:

04D0	04D1	Ӑ	ӑ	ѧ̆	Used in Beron 1824. (The modern version is used in Chuvash.)
0419	0439	Й	й	й	Old Cyrillic form used in Beron 1824.
				ῐ	Found in the Sinai Psalter.
040E	045E	Ў	ў	ү̆?	
				ѣ̆	

Kys8	Kys9			ѧ̆	Used in Beron 1824.

In most modern typefaces the diacritic over ѝ and ў is slightly different from a regular breve, though originally it wasn't.

§7. Letters derived from digraphs:

				0	1	2	3	
0478	0479	Оу	оу	оѵ	оѵ	ѳѵ		The digraphs оѵ and оѵ were freely interchangeable (predictably, since so were ѵ and ѵ at first); ѳѵ is found in ѳѵѣѳ. (Theoretically оу could have as many forms as о has *times* as many as у has. Fortunately, this was not so in practice.) Word-initial (as opposed to ѹ).
Kyr0	Kyr1	Ъі	ъі	ѣі	ѣ̈і	ѣі		Modern version (without a connecting bar) used in Gerov 1895–1908.
042B	044B	Ы	ы	ьі		ыі	ьї	The primary Old Cyrillic variant should have a connecting bar.
04F8	04F9	Ӹ	ӹ	ьї				Cf. one of the Old Cyrillic variants of ы. It seems, however, that ыі and ьї were in free variation, whereas in the distribution of і and ї there were at least tendencies, if not rules. (The modern version, with the dots centred horizontally, is used in Mountain Mari.)

§8. Iotated letters. These seem to call for a unified strategy; e.g., VS1 for a diæresis, VS2 for reversal, VS3 for reversal plus diæresis:

				0	1	2	3	
042E	044E	Ю	ю	ю	ӭю	оі	ої	The reversed forms were used before the 13c. It is possible that reversal indicated a difference in sound (which would call for a separate character).
0464	0465	Ѥ	ѥ	ѥ	ӭѥ			Modern version used in Gerov 1895–1908. (It is a IE/ie, *not* a Ѥ/ѥ ligature, with connecting bars; typeface designers should be instructed to cut it with E/e rather than Є/є as its right-hand part.)
0468	0469	Ѩ	ѩ	ѧ	ӭѧ			Modern version never actually used.
KysE	KysF			ѣ̀				Found only in the 13c Caransebeş Octoich.
046C	046D	Ѭ	ѭ	ѫ	ӭѫ	ѫі	ѫї	
Kyr2	Kyr3			ѧ	ӭѧ			Modern version (IA/ia ligatures with connecting bars) used in Gerov 1895–1908. Word-initial/postvocalic letter in Church Slavonic and in Beron 1824 (the postconsonantal one is ѧ).
Kys6	Kys7			ѣ̀ѣ				Used once in the *Izbornik* of 1076, otherwise only in Svjatoslav's *Izbornik*.

References

Camuglia, M. & K. Ribarov, ms. ‘Old Church Slavonic in Codes’.

Андрейчин, Л. Д. 1977. «Черковнославянска и гражданска азбука през възраждането». / *Из историята на нашето езиково строителство*, София: «Народна просвета», 151–156.

Берон, П. (Петръ Х. Беровичъ) 1824. **Ѣукварь съ различнꙑ поученїѧ** ‘A Primer with Diverse Instructions’ (popularly known as Рибен буквар ‘Fish Primer’).

Геров, Н. 1895–1908. *Рѣчникъ на блъгарскъй языкъ / съ тлъкувание рѣчи-ты на блъгарскы и на русскы* (‘Dictionary of the Bulgarian Language, with intepretation of the words in Bulgarian and Russian’).

Даль, В. И. 1880–1882. *Толковый словарь живаго великорускаго языка* (‘Explanatory Dictionary of the Living Great Russian Language’).

Карский, Е. Ф. 1928. *Славянская кирилловская палеография*. Ленинград.

Мусаев, К. М. (отв. ред.) **1982.** *Опыт совершенствования алфавитов и орфографий языков народов СССР*. Москва: «Наука».

Терещенко, Н. М. 1986. «Алфавит нганасанского языка; Алфавит энецкого языка». / П. Я. Скорик (отв. ред.), *Палеоазиатские языки*, Ленинград: «Наука», 45–47, 50–52.