Unicode 4.1 and Slavic Philology
Problems and Perspectives (I)

Sebastian Kempgen, University of Bamberg, Germany

The present papers aims to give an overview of the current state of encoding the characters used to write Slavic languages, past and present, in Unicode. We will have a look at the Latin tradition as well as Cyrillic writing, and will also cover Glagolitic. The purpose of this paper is *not* to claim that all characters mentioned in this article *should* be encoded in Unicode, but mereley to point out those areas where further investigation is needed, where a common understanding of the principles and practice of treating characters should be developed among Slavicists. The paper is also meant to be a contribution to formal proposals which will be submitted to Unicode, Inc. to have more characters encoded.

The present paper is divided into three main sections: After a short introduction, we will present some solutions that are already available within Unicode – not all of them very well known; the main part of the paper will then be devoted to the “missing pieces”, i.e. characters not yet encoded in Unicode. The article ends with a short outlook of the topics to be covered in the second article on the same subject (Kempgen 2006).

Many different Slavic languages will be covered in this overview, not all of them in the same detail, but we will try to focus on the most important aspects nevertheless. The discussion will mainly touch Russian, Polish, Sorbian, Croatian, Bulgarian, Old Russian and Old Church Slavonic in its various incarnations. The discussion, as is to be expected, focuses on letters or characters as its subject, so orthography, palaeography and phonetics, in their synchronic and diachronic aspects, are the linguistic disciplines this paper fits into.

1. Introduction

The history of encoding characters and scripts for the use by computers has been written about many times and from very different perspectives. For the purpose of the present paper it will be sufficient to sum it up as follows:

- ASCII is a standard with 2^7 = 128 characters. The ASCII standard is a sub-set of all further encoding standards. It covers the English (Latin) alphabet, uppercase and lowercase, numbers, and punctuation.
- “Code Pages” were the next step in the development of operating systems; they consist of 2^8 = 256 characters. Code pages have been defined by operating system software vendors for their platforms, notably Microsoft for Windows, and

© Sebasatian Kempgen 2006. Orig. publ. in: *Computer Applications in Slavic Studies. Proceedings of "Azbuki.Net" International Conference and Workshop, 24–27 October 2005, Sofia, Bulgaria*. Sofija 2006, 131-159.

Apple for the Macintosh. Code Pages are tables which have the 128 ASCII characters in their first (upper) half, and additional 128 characters in their second (lower) half. These additional characters can, for example, comprise a set of accented Latin characters as required by a certain language or by a group of languages, but can also contain the Cyrillic alphabet, for example. However, one cannot have both accented Latin *and* Cyrillic in a single code page.

Code pages differ from computing platform to computing platform, so tools were necessary to translate characters in a text file from one platform to another, say a text containing East European characters from a PC for use on a Macintosh. Not all Slavic characters, languages or scripts were ever covered by code pages: Cassubian is an example of a Latin-writing language, and Old Church Slavonic in its Cyrillic and Glagolitic tradition are also examples of scripts that were never defined as code pages, so only ad hoc or custom solutions were possible here, with some becoming more widespread than others. Code pages characterize the computing situation during the ’90s.

• In 1991, several hard- and software vendors founded Unicode, Inc. to define a common encoding standard that would eventually cover all languages of the world with all their characters. This standard came to be known simply as ‘Unicode’. It has 2^16 = 65.536 slots for characters which were thought to present ample space to fit each and every character of all the world’s scripts into. The Unicode standard gradually evolved over the next years, and operating system software vendors implemented more and more support for Unicode in their software. The code pages which were in existence before the conception of Unicode were all integrated into the new standard to guarantee compatibility with legacy data.

• Version 4.1 of the Unicode standard was finalized and published in 2005, and it featured an important addition for Slavicists, namely the inclusion of the Glagolitic script. Thus, it seems appropriate today to present an overview of the current state of encoding the Slavic languages and scripts in Unicode 4.1. As we said before, Unicode is an evolving standard, and hints of an upcoming version 5.0 are already to be found.[1]

Unicode as an encoding standard can be thought of as a gigantic table with 65.536 cells; each cell has a numerical value and contains a specific character. In addition, this character can also have its own unique name, for example “Tse-cyrillic”. Character are defined by their internal numbers, not by their names, and in principle this should indeed allow for every defined character to be transported from platform to another unaltered. What should be kept in mind, how-

[1] The website for Unicode, Inc. is at http://www.unicode.org. It covers all aspects of the Unicode standard and of its evolution. While in the beginning printed editions served as the standard reference, downloadable pdf files serve the same purpose today. When the present article refers to ‘Unicode docs’, we refer to these pdf files (together about 50 smaller files that cover a specific section of the standard or one global file that contains everything).

ever, is that characters are represented on-screen by using fonts, and fonts may cover only certain subsets of Unicode. Also, one should be aware that the basic fonts used on PCs and Macs fonts evolve over time to incorporate more characters, which means, that as a piece of software they carry version numbers. If one refers to a font having or not having a certain feature, one should carefully note the version of the font that serves as a reference. The Appendix to the present article contains a list of features from Unicode that are important to Slavic philology, and it notes the availability of that feature in current versions of standard fonts like *Times* and *Times New Roman.* As can be seen from that table, these fonts differ markedly in what they – currently – have to offer to a slavicist, and especially a medievalist.

For practical purposes, the aforementioned gigantic Unicode table can be thought of as consisting of “blocks”. The most common blocks are various blocks containing Latin characters, one block contains phonetics characters, another Cyrillic characters for the Slavic languages, and the next one those for the non-Slavic additions, another block various accents and diacritics and so on. For common Western languages, these blocks are not unlike the former code pages in scope, while they do differ in size – some may be small, others may be large.

Unicode terminology differs a bit from linguistic usage: linguists usually talk about characters, graphemes, allographs and letters, whereas Unicode knows characters, glyphs, and character codes. These different terminologies can largely be neglected for the purpose of the present paper.

Apart from the blocks touched upon above, Unicode also provides a so-called ‘private area’ consisting of 6.400 slots. Seeing that there might be the need to use non-standard characters, company logos etc. in texts, this private area was meant to house such elements. The ‘risk’, so to speak, of using it lies in the fact that compatibility between fonts and documents cannot be guaranteed any longer. This may be not be problem for texts which are used by a small user community. For example, a company may decide to put its logo into a slot of the private area so all workers can produce texts using that logo. However, it is easy to foresee a situation where several groups of users (say the medievalists of various philologies, like Germanic, Romance, Slavic) start to use the same slots for their different needs. This can easily lead to problems in publishing papers, in printing etc. At present, this private area is already in use, and its use is not yet coordinated between philologies.[2]

[2] Philologies seem to differ in their efforts to coordinate the use of this private area within their discipline or between them. There seem to be more efforts for coordination in Germanic, English and Romance philologies, as evidenced by MUFI, the *Medieval Unicode Font Initiative* (http://gandalf.aksis.uib.no/mufi/).

The process of evolving the Unicode standard further is through written submissions and proposals. These proposals need to follow a given structure, and they have technical and philological aspects. It is the burden of the persons submitting proposals to clearly demonstrate why a character is needed, how and when it was used, what its relevance for today's computing needs is etc. Unicode.org publishes a 'pipeline' of characters and scripts that will eventually be included into the next revision of the standard so double submissions and the work connected with putting together such a proposal can be avoided. This is very important because the process of getting a proposal accepted is lenghty and includes several steps – it can in principle take anything from months to years as evidenced by the efforts of having Glagolitic being accepted as part of Unicode. Unicode.org also publishes a lists of rejected scripts and characters so interested parties are spared the work of proposing them again. The *Klingon* language/ script may serve as an example of a rejected script. Lastly, the documentation of the Unicode standard will naturally contain some errors, mistakes, though every effort is being made to minimize them. All known errata are also published on the Unicode web-site. All material that will be presented in the third section of this paper has, of course, been checked against the already published errata and the pipeline of forthcoming characters.

2. The Status Quo – A Few Appetizers

In this section we will have a look at what is already available to Slavic philology, what has already been accomplished, what is already covered by standard solutions. This section will of course not expand on trivial accomplishments. Modern Cyrillic, for example, has been a part of the Unicode standard from version 1 on. But nevertheless, Unicode presents some solutions for Slavicists not everybody will be aware of. Thus, it is certainly worth to have a look at some of them, the more so because some of these solutions will serve later as basis to outline deficiencies in the Unicode 4.1.

One important thing that is worth repeating here is that if a character, may it be a letter, an accent or some other sign is defined in Unicode, that does not mean that a certain font contains an image of that character. Thus, it would be a misconception to think that because font *X* does not have character *Y*, it is not available in Unicode. Most people today will, for example, use *Times New Roman* (from Monotype) on the PC (and now on the Macintosh, too), and *Times* (from Linotype) on the Macintosh. A comparison of these two common fonts (see Appendix) will show that, unfortunately, *Times New Roman* has much less to offer to a Slavicist than *Times*, but *Times* is not as widespread on the PC whereas on the Macintosh it is a standard since the invention of desktop publishing in the 80’s.[3]

As an appetizer may also serve the fact that the autor of the present article has produced a font named *Kliment Std* that is available for free. This font is aimed especially at Slavic medievalists, and it features a lot of those characters not present in either *Times* or *Times New Roman*. The font is available for download from http://kodeks.uni-bamberg.de/AKSL/Schrift/KlimentStd.htm and also from the ‘Repertorium’ website. It is being used in the present paper where the standard fonts do not offer support for a character in question.

2.1. Transliteration of pre-revolutionary Russian orthography

Θө: Ḟḟ Unicode: 1E1E, 1E1F (Latin Extended Additional)
Ѵѵ: Ẏẏ Unicode: 1E8E, 1E8F (Latin Extended Additional)

For *F with overdot*, the Unicode documentation mentions “Irish Gaelic (Old Orthography)” as an example for the use of this character; for *Y with overdot*, no such information is given. It is not important whether the documentation for Unicode has any reference to Russian here or not; it is not even important if Unicode, Inc. is aware of this specific use of these two character pairs – what’s important to the user is only that these characters do exist and can be used.

As for the presence in standard fonts, it should be noted that *Times New Roman* does not have these characters at present, whereas *Times* has both (see Appendix).

2.2. Transliteration of Soft Sign and Hard Sign

ЬЪ: ʹ ʺ Unicode: 02B9, 02BA (Spacing Modifiers)
“Modifier Letter Prime”, “Modifier Letter Double Prime”

A bit unexpected to many Slavists might be the fact that Unicode has special characters that should be used for the transliteration of the soft and hard sign. In other words: the curly quote (i.e. ’) which is so common today in printed material regarding Slavic is, strictly speaking, incorrect from a Unicode perspective.[4]

[3] It might be worth pointing out that version 3.05 of both *Times New Roman* and *Arial* for the Macintosh have the same character set as their PC counterparts. Thus, any data and file exchange that is based on these fonts will be completely without problems. This is true for the Latin accented characters used by today’s Slavic languages, and for the contemporary orthography of Cyrillic. The differences – and problems – however begin to start immediately beyond that point.

[4] This brings up another interesting point: Even if a specific Unicode character is available for a given purpose, nobody forces a user to use it – one can use an incorrect character accidentally or if one prefers it to the right one or if the right one is not available in the basic font. Therefore the requirement, often heard today, to “use only Unicode fonts” does not guarantee

That Unicode should have separate characters for the transliteration of the soft and the hard sign is understandable, though: for purposes such as sorting, searching etc., one will want to be able to distinguish between punctuation and characters. In this case, the Unicode docs explicitly mention the use of these two symbols, which also denote “primary stress” and “exaggerated stress” in a phonetic context. The shape of these two glyphs may be acceptable in phonetics, but it seems questionable if they are the best solution aesthetically for the soft and hard sign, and it remains to be seen if these two characters will really become widespread for this use.
In any case, neither *Times* nor *Times New Roman* has these two symbols at present.

2.3. Transliteration of Macedonian

Ѓѓ	Ǵǵ	Unicode: 01F4, 01F5 (Latin Extended-B)
Ќќ	Ḱḱ	Unicode: 1E30, 1E31 (Latin Extended Additional)

The Unicode documentation correctly mentions that these two character pairs are needed for the transliteration of *Macedonian*; however, it also mentions *Serbian* for the first character pair, which is not correct. Interesting in this case is also the fact that these two character pairs are distributed between two blocks which are, so to speak, “miles apart”. This is clear evidence of the fact that proper support for the transliteration of Macedonian has not originally been a systematic consideration when these characters were introduced, or otherwise one would expect to see both character pairs close together in one block.

As for support of these character by our reference fonts, *Times* has both pairs, *Times New Roman* has neither of the two.

2.4. Support for the Cassubian alphabet

Cassubian, a minority language in Poland with a very lively Internet-using community, never had its own code-page on any computing platform, so users had to switch between the standard Western character set and the CE (Central European) code page if they wanted to write their language; however, all characters were in principle available. Now with Unicode, the disctinctions between code pages are largely irrelevant and it is now possible to have one keyboard driver to allow input of all Cassubian characters.

Cassubian characters from the Western character set:	Ãã Òò Ùù Ôô Éé Ëë
Cassubian characters from the CE code page:	Łł Óó Éé Ńń Ôô Żż

full Unicode-compatibility in *encoding* the text – users may be more interested in in the *presentation* form of their text: if the output matches their intentions, they don’t care whether they have used the correct character or not.

Only those characters with diacritics are given here, and one can see that Cassubian uses quite a few characters not used by Polish (all pairs from the Western character set listed here).

2.5. Nasal vowels

In the area of slavic phonetics and phonology, it may be worth noting that Unicode has nasal variants of all five basic vowels *a, e, i, o, u*. Of these, four, namely *Ąą Ęę Įį Ųų*, are needed for Polish and Lithuanian orthography and therefore are part of the Unicode standard, but not *Ǫǫ*. So it is a very welcome addition to find *Ǫǫ*, which Unicode documentation says is needed for Sami and Old Icelandic. Of course, it is also possible to put a breve or macron over these nasal vowels to indicate 'short' or 'long' duration.[5]

From our two reference fonts, *Times* has a glyph for nasal o, *Times New Roman* doesn't. Codepoints are [01EA] and [01EB], respectively.

2.6. Long and short vowels

For Slavic phonetics, long and short vowels are very important. Here, the situation is as follows (only lowercase letters are given):

Basic vowels:	long and short: ā ă ē ĕ ī ĭ ō ŏ ū ŭ
Additional vowels:	only long: ȳ

Of course, *y with breve above* can be composed from its parts: y̆.[6] Thus, for Latin writing, all essential glyphs are there. For Cyrillic, one will have to resort to composing anything besides standard orthography by using the base character and then putting a diacritic on top.

2.7. Vowels with underdot (indicating accent)

Sometimes, especially in typesetting poetry, vowels with underdots are used to indicate stress position. Among Latin precomposed characters, six vowels are available with such a diacritic below:

[5] 'macron' is the name of the symbol used to denote long duration. – In addition, it might be noted that a *long nasal o* is available as a precomposed character (01EC: Ǭ, 01ED: ǭ), because it is also needed for Old Icelandic. However, no other nasal vowel is available as a precomposed character in its long or short variant so we won't stress the presence of this single one too much here. As for our reference fonts, *Times* has this character, but not *Times New Roman.*

[6] Actually, the *Times* font has a bug in the "Combining Diacritical Marks" block: its diacritics do not really combine but have their own positive width, just as the corresponding diacritics in the "Spacing Modifiers" block. Other fonts, like *Lucida Grande* or my own *Kliment Std.* have combining diacritics which do in fact allow composition. *Times New Roman* does not have the breve or the macron in this block.

Latin precomposed: ạ ẹ ị ọ ụ ỵ

For Cyrillic, no such precomposed characters are available, so here one has to use combining diacritics to achieve the same effect:

Cyrillic composed: а̣ е̣ н̣ о̣ у̣ ы̣ э̣ ю̣ ѩ̣
The Cyrillic sample uses the medieval characters shapes of the *Kliment Std* font.

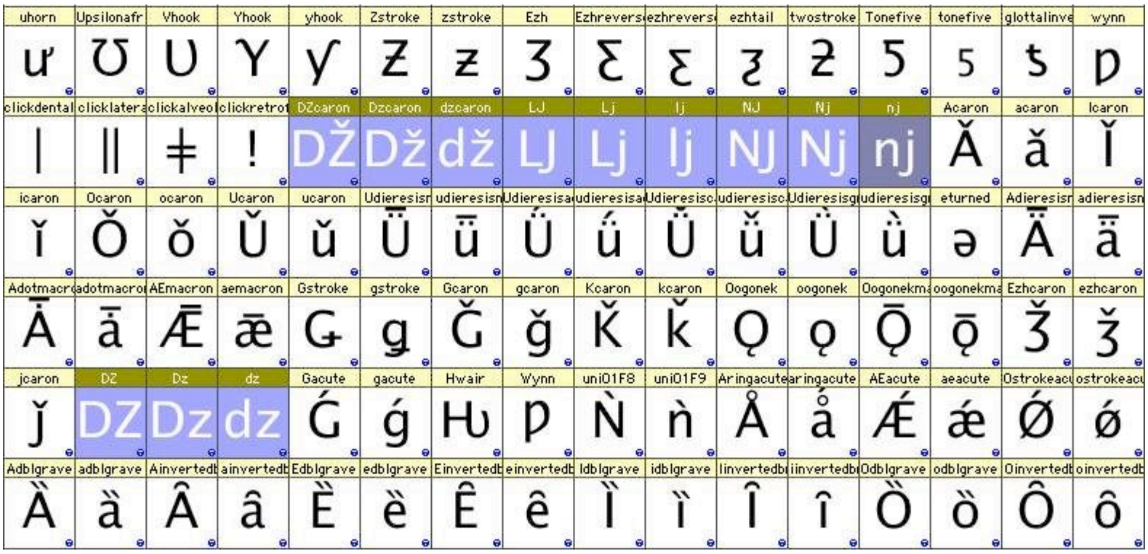
2.8. Croatian Digraphs matching Serbian
An unusual feature of the Unicode standard is that it contains (in the Latin Extended-B block) three groups of three characters each, the so-called "Croatian digraphs matching Serbian Cyrillic letters":

LJ Lj lj - NJ Nj nj - DŽ Dž dž ≈ Љљ Њњ Џџ

Nearby in the same Unicode block, there is another similar group:

DZ Dz dz ≈ Ѕѕ

See the following on-screen representation of the corresponding section from the Latin Extended-B block, with the characters in question highlighted:[7]



uhorn	Upsilonafr	Vhook	Yhook	yhook	Zstroke	zstroke	Ezh	Ezhrevers	ezhrevers	ezhtail	twostroke	Tonefive	tonefive	glottalinve	wynn
ư	Ʊ	Ʋ	Ƴ	ƴ	Ƶ	ƶ	Ʒ	Ƹ	ƹ	ƺ	ƻ	Ƽ	ƽ	ƾ	ƿ
clickdental	clicklatera	clickalveo	clickretrot	DZcaron	Dzcaron	dzcaron	LJ	Lj	lj	NJ	Nj	nj	Acaron	acaron	Icaron
ǀ	ǁ	ǂ	ǃ	Ǆ	ǅ	ǆ	Ǉ	ǈ	ǉ	Ǌ	ǋ	ǌ	Ǎ	ǎ	Ǐ
icaron	Ocaron	ocaron	Ucaron	ucaron	Udieresisr	udieresisn	Udieresisa	udieresisa	Udieresisc	udieresisc	Udieresisg	udieresisg	eturned	Adieresisr	adieresisn
ǐ	Ǒ	ǒ	Ǔ	ǔ	Ǖ	ǖ	Ǘ	ǘ	Ǚ	ǚ	Ǜ	ǜ	ǝ	Ǟ	ǟ
Adotmacr	adotmacro	AEmacron	aemacron	Gstroke	gstroke	Gcaron	gcaron	Kcaron	kcaron	Oogonek	oogonek	Oogonekma	oogonekma	Ezhcaron	ezhcaron
Ǡ	ǡ	Ǣ	ǣ	Ǥ	ǥ	Ǧ	ǧ	Ǩ	ǩ	Ǫ	ǫ	Ǭ	ǭ	Ǯ	ǯ
jcaron	DZ	Dz	dz	Gacute	gacute	Hwair	Wynn	uni01F8	uni01F9	Aringacute	aringacute	AEacute	aeacute	Ostrokeac	ostrokeac
ǰ	Ǳ	ǲ	ǳ	Ǵ	ǵ	Ƕ	Ƿ	Ǹ	ǹ	Ǻ	ǻ	Ǽ	ǽ	Ǿ	ǿ
Adblgrave	adblgrave	Ainvertedb	ainvertedb	Edblgrave	edblgrave	Einvertedb	einvertedb	Idblgrave	idblgrave	Iinvertedb	iinvertedb	Odblgrave	odblgrave	Oinvertedb	oinvertedb
Ȁ	ȁ	Ȃ	ȃ	Ȅ	ȅ	Ȇ	ȇ	Ȉ	ȉ	Ȋ	ȋ	Ȍ	ȍ	Ȏ	ȏ

Fig. 1: Croatian digraphs matching Serbian and Macedonian Cyrillic letters

For the last group, the Unicode documentation does not give a specific use, but is clear that these are "Croatian digraphs matching Macedonian Cyrillic letters".

[7] The table also shows several characters we have already talked about – the nasal o, the g with acute for Macedonian transliteration, and many of the characters important for the discussion in section 3.1 (see below).

Of course, it could also be indicated that we need *DŽ Dž dž* ≈ *Џџ* for Macedonian, too. The reasoning behind the introduction of these characters was that 'Serbocroatian is one language which can be written with two alphabets, and therefore it should be possible to convert any text written in one alphabet into its counterpart from the other – and back again'.

As one can see from the samples above, each group consists of three digraphs: one uppercase only, one mixed uppercase (first character) and lowercase (second character), and one lowercase only. Although a normal user might not be aware of this, the typical uses would be a headline in all capitals (*DŽAS - JAZZ)*, the beginning of a sentence, where only the first letter is capitalized (as in *Džas*, for example), and never the first two (**DŽas*, for example, looks horrible, even if it sounds the same), and a 'normal' form (like *džas*).

As for our reference fonts, *Times* has all these digraphs, *Times New Roman* doesn't have any of them.

These observations may conclude the discoveries that can be made among the already available characters in the current Unicode standard, either with or without support from standard fonts.

3. A Case of Blues – Missing Pieces in Unicode 4.1

This next section will be devoted to problems and missing features of the current Unicode standard from a Slavicists point of view. Available space does not permit us to touch upon everything that would be worth mentioning, so only a selection from various Slavic languages, language areas and times will be presented here (our second article on the same subject will present additional material).[8]

3.1. Štokavian Accents – the missing 'r grave accent'

Serbian and Croatian use four accent marks to denote the four possible combinations of long vs. short duration with rising vs. falling tone. The accepted norm uses acute (´) for long-rising, grave (`) for short-rising, an inverted bow (̑) for long-falling and a double grave accent (̏) for short-falling.[9] These four diacritics go over any of the following six vowels: *a e i o u r*, totalling 6 x 4 = 24 different character combinations.

One of the pleasant surprises of Unicode is that it contains "Additions for Slovenian and Croatian" in its Latin Extended-B block – see the last line of Fig.

[8] Also, it should be duely noted that others, notably Birnbaum (1996 & 2002), have already covered several of the topics that will follow in the next section, although sometimes from a different perspective and in a different context. See also Berdnikov/Lapko (1999).

[9] The inverted bow is sometimes misinterpreted by people writing about Serbocroatian without being really familiar with the language: the circumflex (ˆ) can sometimes be found instead, or the macron (¯). Both are not correct here.

1 above. However, a closer look reveals a strange omission. Here is an overview of available characters (only lowercase characters are given here):

á é í ó ú ŕ	–	Latin 1
à è ì ò ù _	–	Latin 1
ȁ ȅ ȉ ȍ ȕ ȑ	–	Latin Extended-B
ȃ ȇ ȋ ȏ ȗ ȓ	–	Latin Extended-B

Fig. 2: Štokavian accents in Unicode – one is missing

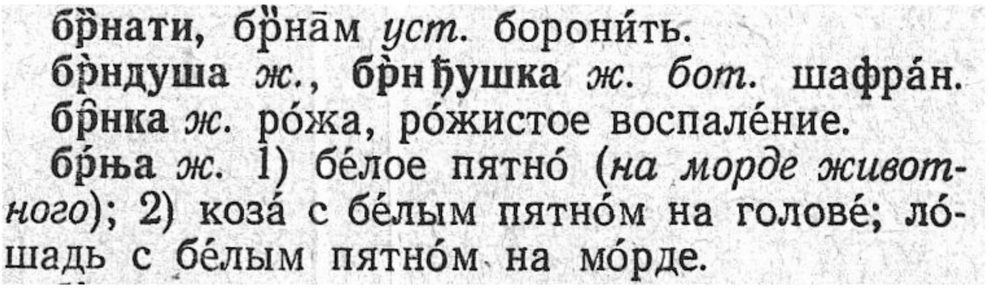
Strange as it may seem, one character has obviously been forgotten, namely *r with grave accent*: r̀. Now the fact itself that all of these vowel-diacritic combinations have been added to Unicode may be astonishing by itself, but there is some logic to it: The letters in the first two rows (Latin 1) are part of standard orthographies, so they were already present in Unicode, which means half of the whole system was already implemented. The remaining characters were added to the Latin Extended-B block to complete the Štokavian accent system. However, it was overlooked that *r with grave accent* (r̀) isn't part of any alphabet and not present in the Latin 1 block. Thus, we have 23 of 24 characters that are available as preaccented characters, and one has to made up from its parts. A strange omission indeed, and one which should surely be corrected from a user's perspective. However, Unicode, Inc. has changed its attitude in recent years with respect to precomposed characters: they won't be accepted if they can equally well be composed from available parts. The omission is especially strange if we look at some – not completely random – sample words that need the missing glyph: *Sr̀bija*, and *hr̀vatski*.

As for our reference fonts – *Times* has the Štokavian accents while *Times New Roman* doesn't. The *r with grave accent* cannot be composed using the current version of the *Times* font, however, because of the bug mentioned above (the combining grave accent behaves as a spacing character, not as a combining character), so the sample again uses the *Kliment Std* font.

The Unicode documentation may mention both Slovenian and Croatian, but it's really the Croatian system that was implemented here. One the one hand, Slovenenian does not use all of the 24 characters mentioned above, and, on the other hand, requires additional characters (cf. Comrie/Corbett 1993: 390ff.) in a phonological transcription:[10] ə̀ (Schwa with grave accent), ə̏ (Schwa with double grave accent), and ẹ́ (e with dot below and acute above). These samples again use the *Kliment Std* font, as neither *Times* nor *Times New Roman* support these character combinations.

[10] For Croatian, there is no distinction here between orthography with added accents and tones and a phonological transcription – they are identical. For Slovenian, this is not true.

It should also be mentioned that Serbian uses the same four accents in its Cyrillic writing, as can be seen from the following sample (from Tolstoj 1970):



бр̏нати, бр̏на̄м *уст.* борони́ть.

бр̀ндуша *ж.*, **бр̀нђушка** *ж.* *бот.* шафра́н.

бр̑нка *ж.* ро́жа, ро́жистое воспале́ние.

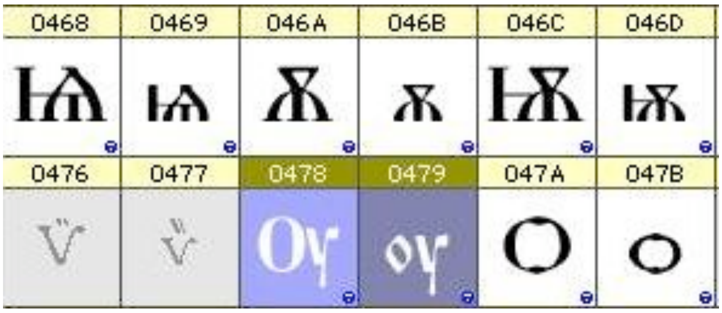
бр́ња *ж.* 1) бе́лое пятно́ (*на морде животного*); 2) коза́ с бе́лым пятно́м на голове́; ло́шадь с бе́лым пятно́м на мо́рде.

Fig. 3: Serbian Cyrillic use of accents

The scan – which is interesting insofar as it shows all four accents in a row over *r* – also shows the use of the macron (¯) denoting a long non-accented vowel. It is worth pointing out here that the Cyrillic portion of Unicode does not have its own diacritics, not even speaking of precomposed Štokavian accents.

3.2. Cyrillic Uk

The Cyrillic block presents a problem which can be treated with the same arguments that led to the inclusion of three variants each for the Croatian digraphs (see above, section 2.8.):



0468	0469	046A	046B	046C	046D
Ѩ	ѩ	Ѫ	ѫ	Ѭ	ѭ
0476	0477	0478	0479	047A	047B
Ѷ	ѷ	Оу	оу	Ѻ	ѻ

Fig. 4: Two slots for 'Uk' – mixed implementation

It is well known that the horizontal digraphs *Оу – оу* are variants of the vertical ligatures *ȣ ȣ*. However, to represent the two vertical ligatures with two separate letters, we really need three pairs: *ОУ - Оу – оу*, not two, as the Unicode table shows. Further, it is not completely clear which character pair should go into the first slot (i.e. 0478). Some fonts realize this position with uppercase-only letters, others with the mixed variant (see sample). Clearly, some action is required here to introduce the third pair into Unicode and to clarify which slot to assign the uppercase-only variant to.

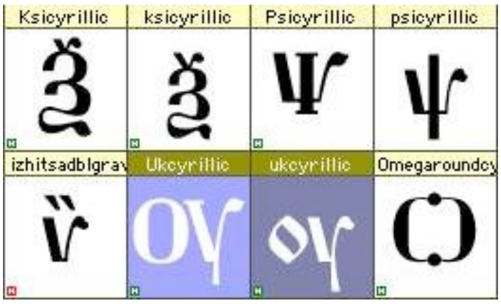
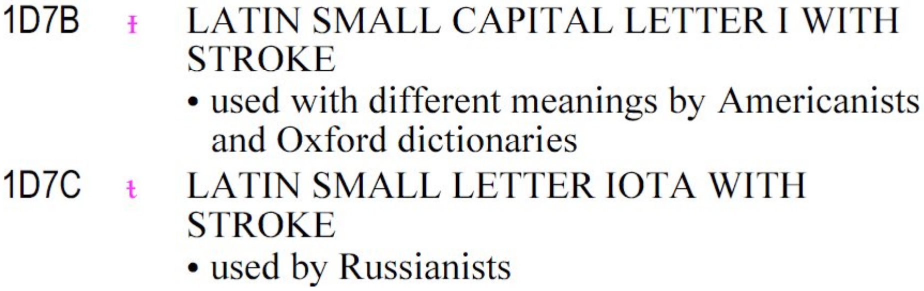


Fig. 5: Uppercase-only and lowercase-only implementation

It could be argued that as long as the solution is not clear, it is better to have the uppercase-only variant in [0478], because word-processors usually have a function to capitalize words written in lowercase letters and vice versa. But a lowercase word-form like ѹгоднти would look like ѸГОДНТИ after applying this function if we were to have the mixed uppercase-lowercase variant in [0478] – not to everyone's typographical taste, one would assume.

3.2. Slavic Phonetics

Just added to Unicode version 4.1. have been two characters that are supposed to be useful for Russian phonetics:

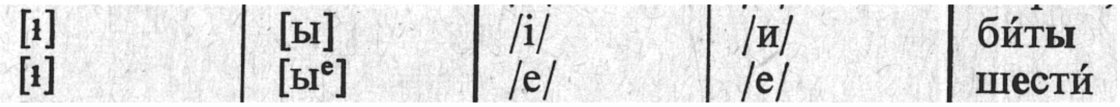


1D7B ᵻ LATIN SMALL CAPITAL LETTER I WITH STROKE
- used with different meanings by Americanists and Oxford dictionaries

1D7C ᵼ LATIN SMALL LETTER IOTA WITH STROKE
- used by Russianists

Fig. 6: Latin small capital ɪ with stroke and iota with stroke in UC

Having consulted all phonetic manuals on Russian that were available to me, I have not seen a single instance where the small iota with stroke has been used. However, there are many instances, where another, similar letter is being used:

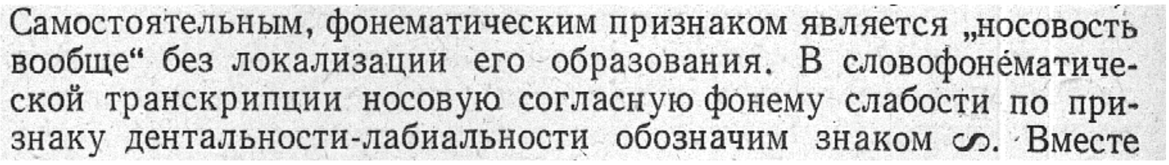


[ɨ]	[ы]	/i/	/и/	би́ты
[ɨ]	[ы^{е}]	/e/	/е/	шести́

Fig. 7: Dotless ı with stroke (Gabka 1975: 34)

It is a *dotless ı with stroke*, and it is commonly used to denote the unaccented realization of *ы*. It is not the lowercase capital ɪ that is being used here in the first row of this table, it is indeed a dotless i!

If we take a closer look at the transcription system used by Avanesov in his various books on Russian phonetics, we see that he mixes Cyrillic with some Latin and/or IPA characters (h j ʌ ə), throwing in some Greek characters (γ α), and making extensive use of accents, too.
However, he also introduces a new phonetic symbol, essentially an “S” turned sideways, to denote “nasality as such”:

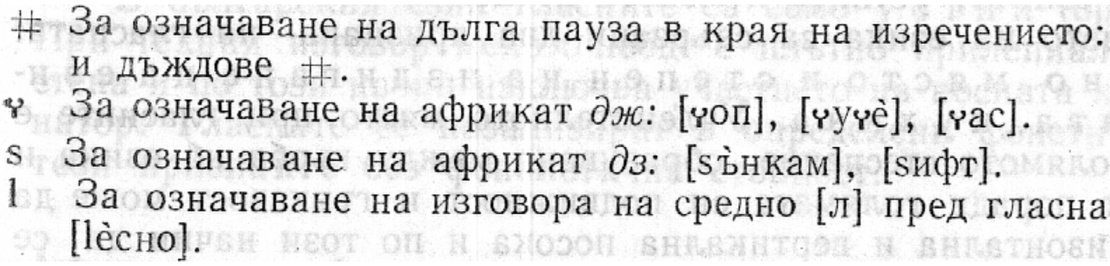


Самостоятельным, фонематическим признаком является „носовость вообще“ без локализации его образования. В словофонéматической транскрипции носовую согласную фонему слабости по признаку дентальности-лабиальности обозначим знаком ∽. Вместе

Fig. 8: Phonetic symbol introduced by Avanesov (1974: 50)

The symbol is available as a technical or mathematical symbol in Unicode [223D], but not as a phonetic symbol.

In Bulgarian phonetics, we have another interesting phenomenon: the revival of an Old Church Slavonic character, used to denote the affricate [dž]:



За означаване на дълга пауза в края на изречението: и дъждове #.
ɤ За означаване на африкат *дж:* [ɤоп], [ɤуɤѐ], [ɤас].
ѕ За означаване на африкат *дз:* [ѕъ̀нкам], [ѕифт].
l За означаване на изговора на средно [л] пред гласна [lѐсно].

Fig. 9: OCS ɤ used in Bulgarian phonetics (Gramatika 1982, T. 1: 29)

While the transcription is in principle Cyrillic, it also makes use of the Latin ‘l’ to write the ‘middle’ or ‘European’ l-sound (not as soft as the Slavic soft l, and not as hard as the hard Slavic l).

The two phonetic symbols mentioned in this section may, however, be regarded as individual, non-standard uses, which do not merit an expansion of the Unicode standard.

3.3. Polish and Sorbian Orthography

The history of Polish orthography exhibits many interesting developments, not all of which can be discussed here at equal length.
Jakub Parkosz (or Parkoszowicz) in his work on Polish orthography (written ca. 1440) was looking for a way to distinguish between hard and soft consonants. The solution he arrived at was to draw characters like ‘b’ and ‘p’ with a round belly to denote soft consonants, and with a square part to denote a hard consonant. His suggestions never really caught on with is contemporaries and actually

were never used in the then-emerging typesetting. However, he is a prominent figure in the history of Polish orthography, and modern editions of his work try to mimic his handwritten original by using character forms created especially for this purpose (cf. 1985, 78). The square vs. round distinction is more a stylistic variation comparable to modern type vs. broken script, but the need to reproduce these letters today is one of the criteria that Unicode, Inc. has formulated for any successful submissions.

More successful proved the distinctions used by St. Murzynowski in his “Orthographia Polska” (Königsberg 1551). Among the characters not present in Unicode today are a *soft b* (b’), the *m with a combining overline*, and the so-called “r rotunda”, which is also well-known from Germanic philology.

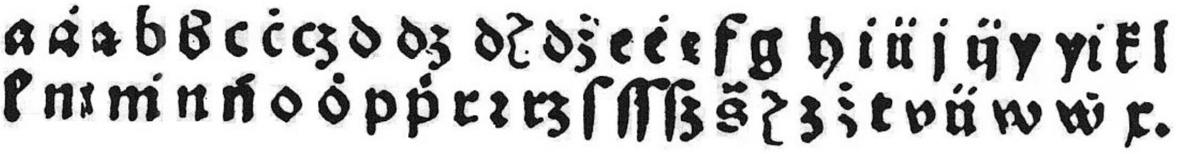


Fig. 10: Alphabet by St. Murzynowski (1551)

While the “r rotunda” was only a stylistic variation in medieval German (it was used when the neighbouring letter was round), Murzynowski proposed another use: he wanted to distinguish those cases where the combination *rz* denotes one sound, i.e. [ž] from those where it denotes two sounds, i.e. [r-z]. Thus, the difference for him was *functional*, not only aesthetic.

Being linguistically closely related to Polish (and Czech), Sorbian is in need of very similar additions as those just mentioned. While several Sorbian characters are included in Unicode (in blocks Latin Extended-A and Latin Extended Additional), others still await a solution (see Fig. 11). They include a *soft b* (b’) and a *soft f* (f’) which can in principle be composed from their parts. These two characters were removed from the official orthography of Sorbian only in 1948 (Upper Sorbian) resp. 1952 (Lower Sorbian; see Comrie/Corbett 1993: 601, 606). A completely new character, however, is the “stroked S”. Because Sorbian has always been printed using German broken script, the lowercase counterpart to the “S with stroke” is not the standard *s* but the “long s” (ſ) already available in Unicode (without stroke) in slot [017F], block Latin Extended-A.

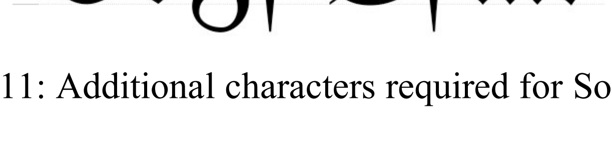
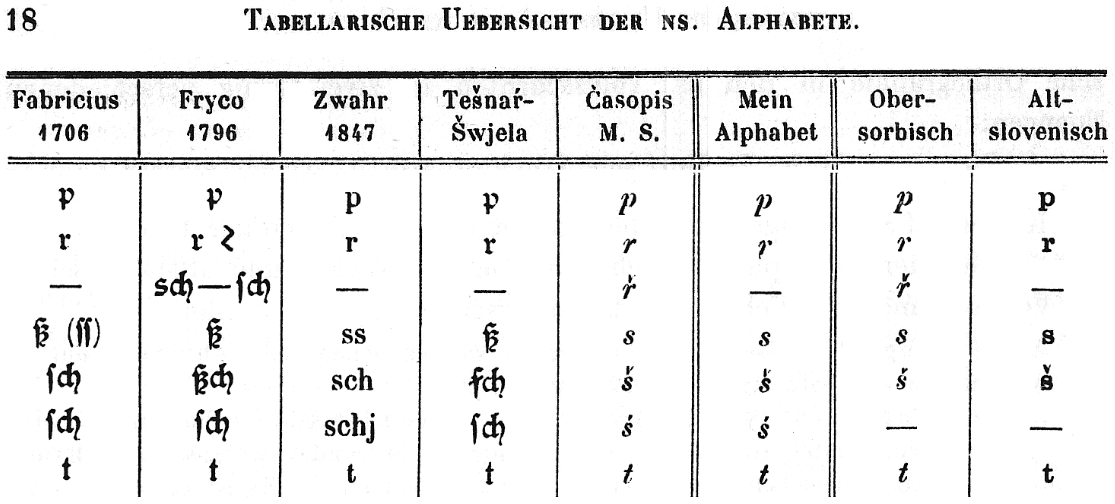


Fig. 11: Additional characters required for Sorbian

The *s with stroke* seems to have been introduced for the printing of a bible in 1868; it was limited (?) to the trigraph *sch* which denotes the sound [š] in German. This corresponds to its funtion: *sch* with stroked long s denotes [š], while *sch* with a normal long s denotes [ś]. See the following table from Mucke (1965: 18) which also includes the “r rotunda”:



18 TABELLARISCHE UEBERSICHT DER NS. ALPHABETE.

Fabricius 1706	Fryco 1796	Zwahr 1847	Tešnaŕ-Šwjela	Časopis M. S.	Mein Alphabet	Ober-sorbisch	Alt-slovenisch
p	p	p	p	*p*	*p*	*p*	p
r	r ꝛ	r	r	*r*	*r*	*r*	r
—	sch—ſch	—	—	*ř*	—	*ř*	—
ẜ (ſſ)	ẜ	ss	ẜ	*s*	*s*	*s*	s
ſch	ẜch	sch	ẜch	*š*	*š*	*š*	š
ſch	ſch	schj	ſch	*ś*	*ś*	—	—
t	t	t	t	*t*	*t*	*t*	t

Fig. 12: Distribution and use of normal vs. stroked long s

The *stroked s* and the *r rotunda* are very interesting challenges typographically because they have never been designed for modern printing types, only for historic broken scripts. In a modern serifed font, they would look like this:

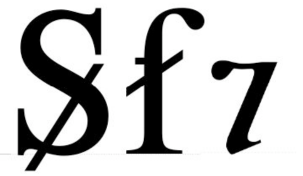


Fig. 13: Design of stroked S/long s and r rotunda

3.4. Czech Orthography

It is well-known from Roman writing onwards that ‘v’ stood for the sound [u]; this was also true for medieval German, and from there it migrated to Czech. In his edition of Johannes’ Hus “Orthographia Bohemica”, Schröpfer (1968: 82, Fn. 25) says that until 1849 initial [u] was written *v* in Czech, and [v] was written *w*. As this initial vowel could also be long, the combination *v́* is required for the historical orthography of Czech, for example, for the word *v́dolj* ‘valley’. Similarly, historical German needs *v̈,* i.e. a *v* with a dieresis above. Old tombstones, by the way, are a good source for samples for this specific character.

3.4. Historical Cyrillic letters

The basic principle that governs the first editions of the Unicode standard as far as Cyrillic is concerned, has been clearly stated in the printed edition of Unicode v. 1:

> "The early form of the Cyrillic alphabet is regarded as a *font change* from modern Cyrillic, because the early Cyrillic forms are relatively close to the modern appearance" (The Unicode Standard 1991: vol. 1, 44).

The application of this principle, however, creates as many problems as it solves, as we'll see below. However, there is a glimmer of hope:

> "If, at some future date, the old letterforms are adequately documented and the need for them demonstrated, then they can be added to this [the Cyrillic-Extended] block" (The Unicode Standard 1991: vol. 1, 45).

That's exactly the status quo today, where several slavists join forces to do exactly this: to demonstrate the use and importance of older letters missing so far from the Unicode standard, with the aim to finalize a submission to Unicode, Inc. The present paper fulfills its purpose if it lines out some of the deficiencies in Unicode 4.1 and proposes some good arguments for the inclusion into Unicode of some of them.

The application of the principle cited above is more or less trivial in some cases, for example **А ~ А**, **Б ~ Б**, **Щ ~ Щ**. In other cases, we have a functional identity, if not an identity in the history of the glyphs themselves: **Л ~ Љ**, **Н ~ Њ** vs. **Ѓ ~ Ѓ**, **Ќ ~ Ќ**. However, Sobolevskij (1908: 52) also lists a *soft d*, and Ščepkin (1967: 111) shows a *soft m*, too:

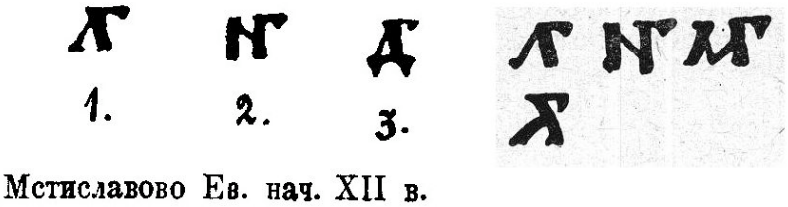


Fig. 14: Old Russian soft consonants (with tail)

The *soft d*, i.e. Д, could be said to correspond funtionally to a Ђ, although both forms are completely unrelated historically, and no such correspondence exists for the *soft m.* So, while for some characters it would be possible to think of a solution consisting of creating a separate font containing old or alternate character shapes with identical function, no uniform solution is possible for all of these soft consonants.[11]

[11] We will not touch upon another set of soft consonants here, those written with the corresponding OCS diacritic, and inverted bow at the right side of the character.

The yotated jat (ꙓ) is also missing from Unicode so far. It seems to occur in Old Russian only, and not very frequently. However, it is important for the history of the Russian form of the Cyrillic script because it evolved into the cursive form of the jat that looks like a ligature of an *n* with a soft sign, ь. Because in Slavic we have the well-known correspondence of yotated and non-yotated vowel letters (а – ꙗ, є – ѥ, ѧ – ѩ, ѫ – ѭ, о – ю), the yotated jat would fit into this row rather nicely, too.

The best-known problem is surely the case of the Old-Russian **Ꙗ** where we have two competing correspondences, **Я** and **Ѧ**. Unicode 4.1 doesn't have a separate entry for **Ꙗ**, but it is very much needed to faithfully transcribe old texts. Therefore, it is one of the characters that are being added to the private area, but while this is certainly a stop-gap solution, this character is really much too important and too regular for the private area. It absolutely should be a regular citizen of the Unicode standard.

It is common knowledge that Peter the Great replaced **Ꙗ** with **Я**. Less attention is usually given, however, to another replacement that occurred in Early Modern Russian. In the 18th century, the [jo] sound was written, very logically, with the digraph ıô. To faithfully reproduce this character, the circumflex should sit *between* the two characters, but we will limit ourselves here to this approximation.[12] We find this ıô in the Academic Dictionary of 1783-94, and although Karamzin came up with ë as a replacement in 1797, the Academic Grammar of 1802 still uses the old form.

3.4. Accents and numbers, Paerok

A separate, but closely related subject is the encoding of Old Russian accents, breathing marks, numbers etc. Of the two dozen or so elements (not counting superscript letters), only a few can be considered to be encoded in Unicode 4.1. Further, it is well-known that the system of these accents – if it even can be called a system at all – underwent some changes over time, and not all scribes knew the rules when and where to write these diacritics. From the overview that Čerepnin gives (1956: 374-376), we will limit ourselves here to the *erok, ertica,* or *paerok*, as the no. 26 from Fig. 15 was called. The standard Russian name today is *paerok.*

[12] Actually, the ıo was printed in several ways: the first part could be a Latin or a Cyrillic i, and the Latin i could be with dot or dotless.

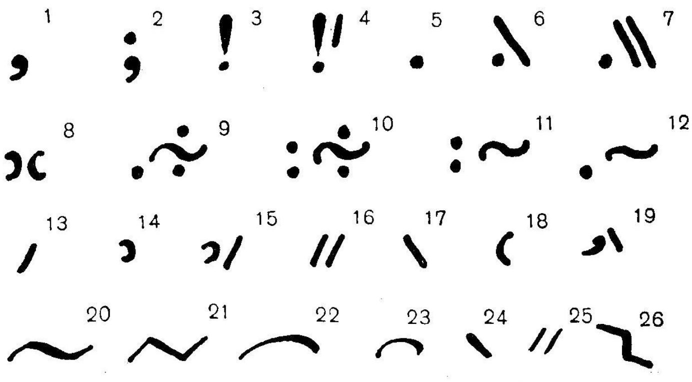
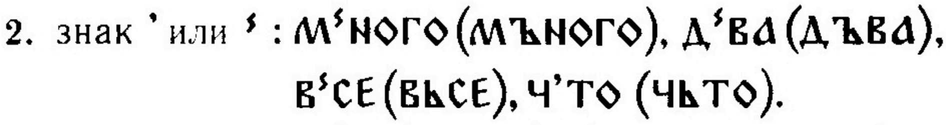


Fig. 15: Old Russian accents and breathing marks (Čerepnin 1956: 375)

Manuscripts actually know two different basic forms of this glyph: one is very similar or identical to a single quote or apostrophe (’), the other, older, form is a distinctive ‘S’-like shape:



2. знак ʼ или ꙿ : мꙿного (мъного), дꙿва (дъва), вꙿсе (вьсе), чʼто (чьто).

Fig. 16: The basic form of the *Paerok* (Samsonov 1973: 57)

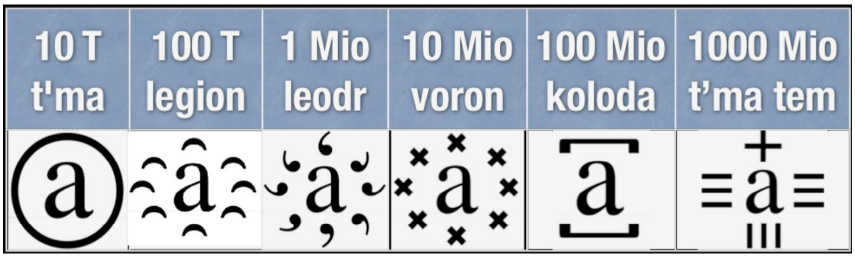
The *paerok* stands mostly between characters, but can also appear above characters. However, it is not a diacritic in the sense that it *modifies* the preceding character. Rather, it is a special glyph that appears *instead* of a character, namely the *jer*.[13]

The first form of the *paerok* can be considered to be available in Unicode: at position [02BC] there is a “modifier letter apostrophe”, and the documentation says that “many languages use this as a letter of their alphabets”. This is also true for some contemporary Slavic languages (for example, Macedonian, and, for a short period after the revolution of 1917, Russian), and it correctly describes the use of the *paerok*, too. Actually, Leskien calls the *paerok* an ‘apostrophe’.

Not so clear, however, is the second form of the *paerok.* In Unicode, there is a “combining vertical tilde” [033E], which looks similar to the *paerok,* so some font designers or authors have identified both, a solution we wouldn’t think is correct. So the *paerok* is a character that needs to be encoded separately.

As for number signs, Vostokov (1863: 9) describes the following system:

[13] And because the paerok appears *instead* of a normal character, its superscript, diacritic-like use has to be solved in the same way all other superscript letters are treated: either all of them have to be encoded separately in Unicode or they all have to be treated as presentation forms (using mark-up to denote their position).



10 T t'ma	100 T legion	1 Mio leodr	10 Mio voron	100 Mio koloda	1000 Mio t'ma tem

Fig. 17: Old Russian number signs according to Vostokov

From the six signs, only two are encoded in the Cyrillic block (*legion* and *leodr*), and the first one, a combining circle, could possibly be identified with slot no. [20DD], block “combining diacritical marks for symbols”, but that doesn’t appear to be the ideal solution.
It should be noted, however, that the glyphs that mark large numbers are, to a large extend, theoretical entities – there was hardly any actual need to count to such large numbers in early Rus’.

3.4. Transliteration of Glagolitic

Now that the Glagolitic script has been accepted by the Unicode consortium as being a separate script, it is only natural to check if all the necessary characters are available which are commonly used for the transliteration of Glagolitic into Cyrillic (just as Latin counterparts to certain Serbian / Macedonian characters have been added to have a 1:1–correspondence). In fact, the common usage by librarians to transliterate Glagolitic into Cyrillic served as an argument *not* to encode certain Glagolitic variants separately.

If we take, for example, the Glagolitic *i*, we’ll see that there is certain problem here. The Glagolica has three different forms for the *i* (see Fig. 18), and these three characters – all encoded in Unicode 4.1 – are usually transliterated with the four Cyrillic characters in the second row.

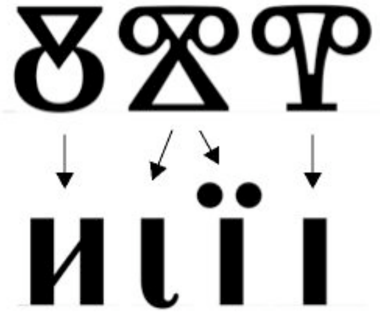
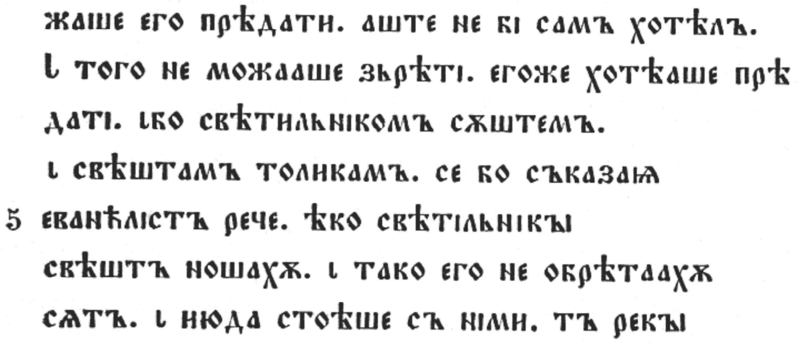


Fig. 18: Transliteration of Glagolitic ‘i’

The first and the third Glagolitic *i* are consistently transliterated using the indicated characters, but for the middle Glagolitic *i*, the transliteration differs between authors, with the Iota being more common (see below).
Of the four Cyrillic characters in Fig. 18, only two are present in Unicode, and two are missing: the Cyrillic Iota and the Cyrillic dotless lowercase ı.
The Cyrillic Iota was introduced by Jagić expressly for the purpose of transliterating Glagolitic. Samples of both an uppercase and a lowercase variant are present in the following scan (Jagić 1911: 232) and also frequent in his edition of the Codex Marianus:



жаше его прѣдати. аште не бі самъ хотѣлъ.
Ꙇ того не можааше зърѣті. егоже хотѣаше прѣ
даті. ꙇбо свѣтильнікомъ сѫштемъ.
ꙇ свѣштамъ толикамъ. се бо съказаꙗ
еванѣлістъ рече. ѣко свѣтільнікꙑ
свѣштъ ношахѫ. ꙇ тако его не обрѣтаахѫ
сѫтъ. ꙇ нюда стоѣше съ німи. тъ рекꙑ

Fig. 19: Jagićs use of Cyrillic Iota (uppercase and lowercase)

Of the two characters, Cyrillic Iota and Cyrillic I with dieresis, the former is clearly the standard most editions use for the transliteration of the middle Glagolitic character from Fig. 18. The second one can be considered a minority use, and quite rightly so: the use of diacritics leads, in this case, to a confusion as to which diacritics are present in the original, and which ones are only part of the transliteration.

With every proposal that is submitted to Unicode, Inc., one has to include a font that shows the character in question. This requirement poses an interesting typographical challenge: Until now, a Cyrillic Iota has only existed in the blackletter designs used in printing OCS texts, but not in modern typefaces. Therefore the question arises what a Cyrillic Iota should look like in serifed typefaces like *Times* (or *Times New Roman*). Fig. 20 presents the answer.

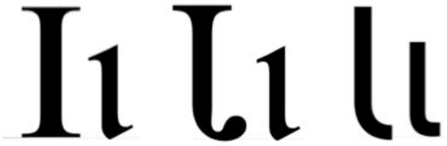
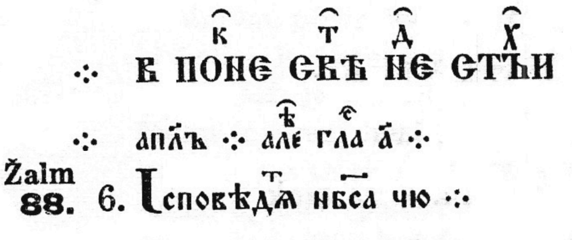


Fig. 20: Design of the Greek Iota (left) and Cyrillic Iota (Roman and Sans)

If we first take a look at the lowercase and uppercase Greek Iota (left), we see that it has an individual character shape for the lowercase letter, while the uppercase letter is identical to the Latin uppercase I. Consequently, the Cyrillic uppercase Iota cannot be identical to the Greek uppercase Iota, because we need distinct lowercase *and* uppercase forms: the uppercase I is needed for the third Glagolitic letter from Fig. 18. The lowercase Cyrillic iota can be borrowed from and be identical to the Greek lowercase iota, but the uppercase Cyrillic Iota has to be designed differently. Following the design principles of a serifed typeface, the uppercase Cyrilic Iota must look like a flipped Latin J. Thus, Fig. 20 shows a serifed Cyrillic Iota in the middle, and a sans-serif Cyrillic Iota at the right side.[14]

Such a serifed Cyrillic Iota has been, as it turned out after the author arrived at his own solution, actually been used before. It may be a black-letter design, but the design principles are the same, and it serves as a confirmation of the solution demonstrated in Fig. 20:



⁘ В ПОНЕ СВѢ НЕ СТЪИ

⁘ апл҃ъ ⁘ але гл҃а а҃ ⁘

Žalm 88. 6. Ɩсповѣдѧ нбса чю ⁘

Fig. 21: Serifed (black-letter) uppercase Cyrillic Iota (Kurz 1955: 3)

The transliteration of the three Glagolitic I's actually show that we need another Cyrillic character, namey a lowercase only addition: a dotless small variant of the uppercase I with the same design: ɪ. In the Cyrillic block, we have two pairs of I's so far that look like Latin characters: I – i and Ï – ï. As is usual with Latin alphabets, in the first pair the lowercase i has an overdot, but the uppercase doesn't have one. In contrast to Turkish, the dot is not a distinctive element. For the transliteration of Glagolitic, however, we need a third lowercase letter: ɪ. This is the same glyph that survived in the digraph (ligature) ъɪ, and as the first part of ю, ѥ, ѩ, ѭ. As can be seen from Fig. 18, and as every edition of a OCS text will prove, it is impossible to use a standard i with overdot in the transliteration of a Glagolitic text. This *must* be a dotless i. However, it cannot be the Latin dotless i which is already available in Unicode, i.e. ı, because that is a Latin character, and it is not the smaller variant of the capital letter I that we need for the Cyrillic transliteration. The Cyrillic lowercase dotless ɪ must be the same glyph that is available as [026A] in the 'IPA Extensions' block, only as a Cyrillic character in the Cyrillic block. Further, the Cyrillic lowercase dotless ɪ cannot be said to be an earlier variant of the modern character i: The modern i could be said to be a dotless ı with a dot above (and the modern dotless ı could be said

[14] With the sans-serif uppercase Iota, the design actually isn't simply a flipped J as that does not produce an optimal optical effect. The design of the uppercase variant is more 'a long version' of the lowercase letter, just as in traditional OCS printing types (see Fig. 19).

to have a precursor looking like ı), but once an i with a dot has been encoded as a Cyrillic character, the dotless form ı must be a separate entry. Lastly, there is ample evidence from Old Russian texts that an ı and and an ı with an overdot exist side by side in the same text, from the hands of the same scribe.

4. To be continued…

Within the space of the present article, only a handful of missing characters and areas could be presented and covered. Therefore, a second article will cover additional topics such as the Cyrillic transliteration of Glagolitic nasals, the transliteration of Glagolitic into Croatian Latin, Croatian Glagoljica characters, Bosančica, Cyrillic superscripts and ligatures, additional considerations from a broader perspective of Balkan philology (OCS-Cyrillic for Romanian, Greek for Albanian and Macedonian), and General Phonology to name the most important areas.

References

Berdnikov, A., Lapko, O. 1999. *Old Slavonic and Church Slavonic in TEX and Unicode.* pdf file available for download from: http://www.uni-giessen.de/partosch/eurotex99/berdnikov2.pdf

Birnbaum, D. 1996. Standardizing characters, glyphs, and SGML entities for encoding early Cyrillic writing. *Computer Standards and Interfaces* 18, 201-252.

Birnbaum, D. 2002. *Unicode for Slavic Medievalists*. Presentation for Sofia conference. Available at http://clover.slavic.pitt.edu/~repertorium/resources/unicode_sofia_1_post.pdf

Comrie, B., Corbett, G.G. (eds.) 1993. *The Slavonic Languages*. London—New York.

Dobreva, M. 2003. *Mediæval Slavonic Written Cultural Heritage in the E-World: The Bulgarian Expertise*. pdf paper, publication forthcoming.

Gabka, K. (ed.) 1975. *Einführung in das Studium der russischen Sprache. Phonetik und Phonologie*. Düsseldorf.

Kempgen, S. 2005. *Kliment Std – A Free Font for Slavic Medievalists*. http://kodeks.uni-bamberg.de/AKSL/Schrift/KlimentStd.htm

Kempgen, S. 2006. Unicode 4.1 and Slavic Philology – Problems and Perspectives (II). In: T. Berger, J. Raecke, T. Reuther (eds.), *Slavistische Linguistik 2004/2005*, München, 223–248.

Kurz, J. (ed.) 1955. *Evangeliarum Assemani. Codex Assemani 3. slavicus glagoliticus*. Ediderunt Josef Vajs & Josef Kurz. Tomus II. Edidit Josef Kurz. Pragae.

Mucke, K.E. 1891. *Historische und vergleichende Laut- und Formenlehre der niedersorbischen (niederlausitzisch-wendischen) Sprache*. Leipzig (Reprint 1965).

Parkosz, J. 1985. *Traktat o ortografii Polskiej Jakuba Parkosza*. Opracował Marian Kucała. Warszawa: PAN.

Schröpfer, J. 1968. *Hussens Traktat „Orthographia Bohemica“* […]. Wiesbaden.

Unicode 1991. The Unicode Consortium (ed.), *The Unicode Standard. Worldwide Character Encoding. Version 1.0.* Vols. 1-2. Reading, Mass.: Addison-Wesley Publ. Co.

Unicode v. 4.1 2005. Complete code tables are available in a single pdf file for download: http://www.unicode.org/Public/4.1.0/charts/Codecharts.pdf

Urbańczyk. S., Olesch, R. (eds.) 1983. *Die altpolnischen Orthographien des 16. Jahrhunderts. Stanisław Zaborowski, Jan Seklucjan, Stanisław Murzynowski, Jan Januszowski* (Slavistische Forschungen, Bd. 37). Köln–Wien: Böhlau.

Аванесов, Р. И. 1974. *Русская литературная и диалектная фонетика*. Москва.

Востоков, А. Х. 1863. *Грамматика церковно-словенскаго языка изложенная по дрейвнейшимъ онаго письменнымъ памятникамъ*. СПб. (Reprint Köln 1980).

Граматика 1982. *Граматика на съвремения български книжовен език*. Том I. Фонетика. София.

Самсонов, Н. Г. 1973. *Древнерусский язык*. Москва: Высшая школа.

Соболевский, А. И. 1908. *Славяно-русская палеография*. Изд. 2-е. СПб.

Толстой, И. И. (сост.) 1970. *Сербскохорватско-русский словарь*. Изд. 3-е, исправл. и доп. Москва.

Черепнин, Л. В. 1956. *Русская палеография*. Москва.

Щепкин, В. Н. 1967. *Русская палеография*. Москва.

Ягичъ, И. В. 1911. “Глаголическое письмо.” В: *Энциклопедія славянской филологіи*, вып. 3, “Графика у славян”, СПб, 51–230.

Abstract

The paper presents an overview of Slavic philology with respect to version 4.1 of the Unicode standard. A review of East, West and South Slavonic languages, their alphabets and writing systems reveals at least a dozen characters that need to be encoded in Unicode, among them a Latin S with stroke, the ‘r rotunda’, several (soft) Cyrillic charactes with tail, the Cyrillic old-style Ꙗ, a Cyrillic Iota (uppercase and lowercase), a Cyrillic dotless lowercase ı, the Cyrillic Paerok (no distinction between uppercase and lowercase), number signs, accents etc.
A second article (to appear) on the same subject will present more areas which need attention and careful consideration, and will feature more material in support of the missing characters covered in this first article

Appendix: Standard fonts and their features for Slavists

	Lucida Grande v. 5.0d8e1 (OS X)	Times / Helvetica v. 5.0d10e1 (OS X)	Times New Roman / Arial v. 3.05 (Win/OS X)
Basic **Latin**	✓	✓	✓
Latin-1 Supplement (= Western Europe)	✓	✓	✓
Latin Extended-A (= Eastern Europe & more)	✓	✓	✓
Latin Extended-B	✓	most	some
Croatian Digraphs	✓	✓	– –
Maced. Translit. (ǵ)	✓	✓	– –
Štokavian Accents	✓	✓	– –
Nasal o	✓	✓	– –
Latin Extended Additional (256)	✓	✓	ca. 1/3
Maced. Translit. (ḱ)	✓	✓	– –
Russ. Hist. Translit. (ḟ, ẏ)	✓	✓	– –
Sorbian (ḿ, ṕ)	✓	✓	– –
IPA – Phonetic	✓	2/96	1/96
Spacing Modifiers	✓	11/80	9/80
Translit. of Jers	✓	– –	– –
Combining Diacritics (= „Flying Accents“)	✓	40/112	5/112
Greek			
Modern Greek	✓	✓	✓
Archaic Letters (Koppa, Stigma, Sampi…)	✓	– –	– –
Classical Greek	✓	✓	– –
Cyrillic			
Std. Russian & Slavic	✓	✓	✓
Macedonian Add. (ѐ, ѝ)	✓	✓	– –
Historical Add. (ѣ ѧ ѫ …)	✓	– –	– –
Ukrainian Ghe (Ґ ґ)	✓	✓	✓
Non-Slavic Cyrillic (ex GUS-Countries)	✓	ca 1/2	ca. 1/10

	Lucida Grande v. 5.0d8e1 (Mac OS X)	Times/ Helvetica v. 5.0d10e1 (OS X)	Times NR/ Arial v. 3.05 (Win/OS X)
Armenian, Georgian, Hebrew, Arabic, Ethiopian	Hebrew	– – (supported by other fonts)	Hebrew, Arabic
General Punctuation	✓	18/112	27/112
Superscripts/Subscripts (0…9)	✓	– –	– –
Currency (Euro…)	✓	3/48	6/48
Comb. Diacr. for Symbols (O)	✓	– –	– –
Number Forms	✓		
Add. Fractions (2/3…)	✓	– –	6/13
Roman Numerals	✓	✓	– –
Arrows	20/112	– – (complete in Apple Symbols)	7/112 (complete in Wingdings)
Mathematical Operators (∏, ∫, ≠ …)	18/256	12/256 (complete in Apple Symbols font)	15/256 (complete in other fonts)