Sebastian Kempgen

Unicode 4.1 and Slavic Philology
Problems and Perspectives (II)

0. Introduction

In the first paper devoted to the current state of encoding Slavic characters into Unicode (KEMPGEN 2006) [1] we gave an introduction to the topic, outlined various achievements in the current version of Unicode (v. 4.1, 2005), mainly because not all of them are common knowledge or available in standard fonts like *Times* or *Times New Roman*, and then we discussed many of the historic or special characters of the Slavic languages that have not yet been encoded in Unicode. Space did not permit us to cover everything in the first article, so this second article will take up where the first paper left, and will also present more illustrative material for some of the topics from the first article. The reader might find it useful to be familiar with the first article before reading the present one, but knowledge of the first article is not a prerequisite.

As for the first paper, it should be noted that the purpose of this paper is *not* to claim that all characters mentioned in this article *should* be encoded in Unicode, but merely 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.

One important thing that is also worth repeating 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*, *Y* 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 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 (although of course only now under OS X with an extended Unicode character set).[2]

---

1 Available electronically from the 'Kodeks' server: http://kodeks.uni-bamberg.de/

2 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.

To make up for the deficiencies in the standard fonts mentioned above, the author 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 the above-mentioned reference fonts. 'Kliment Std' 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 offer support for a character in question.

1. Unicode Blocks

For practical purposes, Unicode, a table consisting of 65.536 cells in which each character has its own unique number and also a name, is normally organized into "blocks". Thus, for example, all Cyrillic characters for the Slavic languages form one block, Soviet additions to the Cyrillic alphabet form another block etc.

Fig. 1: Cyrillic Unicode Block ('Lucida Grande' font)

This pertains to the Latin accented characters used by today's Slavic languages, and to the contemporary orthography of Cyrillic. The differences – and problems – however begin to start immediately beyond that point.

2. Notes on available characters and solutions

2.1. The Uk ligature

Among the historical additions to the Slavic block there is the 'Uk' digraph, i.e. Oy / oy. Sometimes this character is rendered as a digraph, and sometimes as a vertical ligature, i.e. ᪸ ᪹ (see Fig. 1). However, one can either have the digraph or the ligature in the Slavic Cyrillic block, not both.

Just browsing through all characters available in a given font will, however, reveal the presence of ᪸ ᪹ at another location: They are available as *Latin* letters in the Latin Extended-B block, character codes [0222] and [0223]. This is somewhat unexpected because it is well known to Slavists that these characters were introduced into the Cyrillic alphabet under the influence of the Greek alphabet and its writing conventions. In the Greek section of Unicode, however, these vertical ligatures aren't available. The reason why the vertical ligature is not present in Unicode as a Greek character is because Greek official bodies vetoed its introduction on the basis that it's use is 'a mistake', and that it will be corrected by teachers. However, the vertical ligature is being used today even in printed form, see Fig. 2 for an example in the name of a famous Greek biscuit company.

Fig. 2: Greek vertical OY ligature in use today

Anyway, from the Greek alphabet the ligature found its way into the alphabets of the Algonquin and Huron languages. Thus, the Greek vertical ᪸ ᪹ became a *Latin* letter present in two Indian languages but the language it was borrowed from, Greek, does not have that character in Unicode because officials did not like it there, and the need of philologists wasn't considered relevant enough to overturn the official vote.

Now from a user's standpoint it wouldn't make much of a difference if the ᪸ ᪹ is a Greek, a Cyrillic or a Latin letter – if it's there, it can be used. That's indeed true, but it is also true that one will then have a mixture of Greek and Latin letters within one word, or of Cyrillic and Latin. This has negative side effects on sorting, spell checking, hyphenation etc. and such a mixture of characters should be avoided. For the very same reasons, Unicode has a Cyrillic 'A' even if already has a Latin 'A'!

The same observation is also valid for other characters: Cyrillic characters not yet available in Unicode may look identical to Latin characters elsewhere in other Unicode blocks, and thus it is tempting to use them instead. However, as

we wanted to point out here, that is not the ideal approach, although certainly practical if no other solution is available at a given time.

2.2. Long and short vowels

Available in Unicode are precomposed characters for most long and short vowels:

Long vowels: ā ē ī ō ū ȳ
Short vowels: ă ĕ ĭ ŏ ŭ _

What's missing from this chart is a short y, which, however, can be composed from its parts: ў. One may be tempted to simply use the Byelorussian character instead, which looks identical. While for a printed text or an on-screen representation it makes no difference whether the ў is Cyrillic or Latin, this distinction is important for settings concerning the language of the text in a word-processor file, for hyphenation, for sorting etc., i.e. for text processing and text encoding – see above.

3. Problems in Unicode 4.1 – continued

3.1. Štokavian Accents – the missing double grave accent

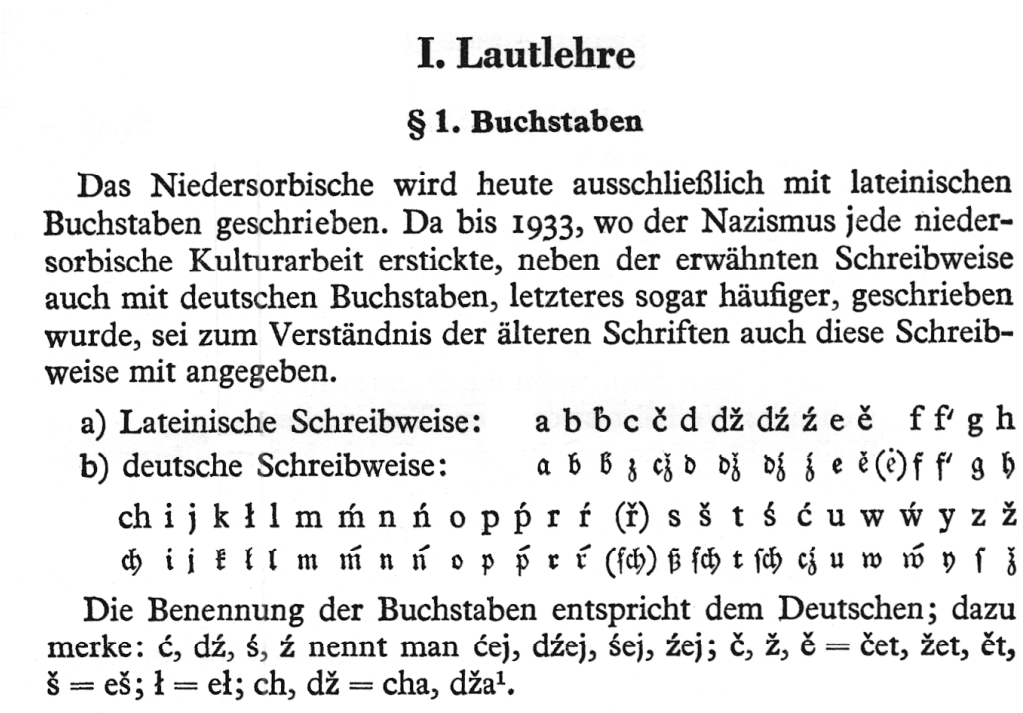
Serbian and Croatian use four accent marks to denote the four possible combinations of long vs. short duration with rising vs. falling tone. The accepted norm uses acute (´) for long-rising, grave (`) for short-rising, an inverted bow (̑) for long-falling and a double grave accent (̏) for short-falling. 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. Of these 24 characters, 23 are covered as precomposed characters by Unicode 4.1, while one has been forgotten: *r with grave accent*: ȓ. This omission was treated more fully in the first article. As one can see from the preceding sentences, it is sometimes necessary, when one writes about accents, to typeset them by themselves. This is why Unicode has among its blocks a section devoted to 'spacing modifiers', i.e. accents with their own positive width, and a section devoted to 'combining diacritics', i.e. accents which have zero-width, i.e. which will not move the insertion point in a word-processor, and which will be displayed instead above or under the preceding character. Such 'combining diacritics' are also known as 'flying accents'.

While most accents required for Slavic phonetics are available as spacing accents as well as non-spacing accents, the one accent missing is the double grave accent. It is available in the 'Combining Diacritical Marks' block at [030F], but not in the 'Spacing Modifiers' section. There, however, we find a "modifier letter middle double grave accent" at number [02F5] among several other "middle" modifiers. They all differ from normal modifiers in their height:

they do not sit *above* characters but to the side of them: x‟x. In other words: a normal *spacing modifier double grave accent* is missing from Unicode at present and should be added as a 'spacing clone' of the corresponding combining diacritic.

3.2. Sorbian and Polish orthography

In the first article on this subject, we already mentioned that for *Sorbian* several characters are missing. We will be presenting some additional illustrations here.



I. Lautlehre

§ 1. Buchstaben

Das Niedersorbische wird heute ausschließlich mit lateinischen Buchstaben geschrieben. Da bis 1933, wo der Nazismus jede niedersorbische Kulturarbeit erstickte, neben der erwähnten Schreibweise auch mit deutschen Buchstaben, letzteres sogar häufiger, geschrieben wurde, sei zum Verständnis der älteren Schriften auch diese Schreibweise mit angegeben.

a) Lateinische Schreibweise: a b b̓ c č d dž dź ź e ě f f‘ g h

b) deutsche Schreibweise: a b b̓ ʒ cz̓ d dz̓ dź̓ ź e ě(ė) f f‘ g h̓

ch i j k ł l m ḿ n ń o p ṕ r ŕ (ř) s š t ś ć u w ẃ y z ž

ch i j k ł l m ḿ n ń o p ṕ r ŕ (ſch) ß ſch t ſch cz̓ u w ẃ y ſ ʒ̓

Die Benennung der Buchstaben entspricht dem Deutschen; dazu merke: ć, dź, ś, ź nennt man ćej, dźej, śej, źej; č, ž, ě = čet, žet, ět, š = eš; ł = eł; ch, dž = cha, dža[1].

Fig. 3: Character table from Šwela (1952, 1).

The characters that are not available in Unicode are precomposed b' and f', which may not really be necessary given that they can be composed from their parts quite satisfactorily, and the *long s* (ſ) *with stroke*. See Fig. 5 for some more samples. The *long s with stroke* is an addition to Unicode that would be needed. Because this character was always and only printed using broken script fonts ('Fraktur'), the lowercase character is not the 'normal' s, but the long s (ſ). Implemented for a modern typeface, it would look like this:

Fig. 4: Serifed design of S and long s with stroke

z, ſ weiches s (lesen) za – ſa für (**Sa**che), zły – ſły böse.
s, ß hartes s, ß (Grüße, miß): se – ße sich (Erbse), gus – guß die Gans (der **Gu**ß).
c, ʒ wie dt. z: cas – ʒaß die Zeit, cera – ʒera Linie (**zerr**en).
ź, ʑ ganz weiches dt. sch mit j verbunden: źiwy – ʑiwy wild.
ś, ſch dt. sch mit j verbunden: śele – ſchele Kalb, śpa – ſchpa Stube.
ć, cʑ dt. tsch mit j verbunden = tś: ścina – ſcʑina Rohr, ścaniś – ſcʑaniſch schillern.
dź, dʑ = d + ź: zdźaržaś – ſdʒarʒaſch erhalten.
ž, ʒ volles sanftes sch, wie im Worte **j**ournal: žaba – ʒaba der Frosch, žywy – ʒywy lebendig.
š, ſch härter als dt. sch (Pa**sch**a), šesć – ſcheſcʑ sechs, šnarl – ſchnarl der Goldammer (**sch**narren), běšćo – běſchcʑo ihr waret.
č, cʒ volles tsch (Kut**sch**er) = tš: lažčej – laʒcʒej leichter.
dž, dʒ = d + ž: ldžej – ldʒej leichter.
ř, ſch = sch, steht in manchen älteren Schriften nach k, p, t statt heutigem š: křej – kšej, třawa – tšawa.

Fig. 5: Character samples from Šwela (1952, 3).

In the history of *Polish* orthography, the work of Jakub PARKOSZ (ca. 1440) stands out as the first attempt to distinguish hard and soft characters. In the first part of the article on Slavic philology and Unicode, we already presented a scan from modern printed editions of his works. We will present here the handwritten original:

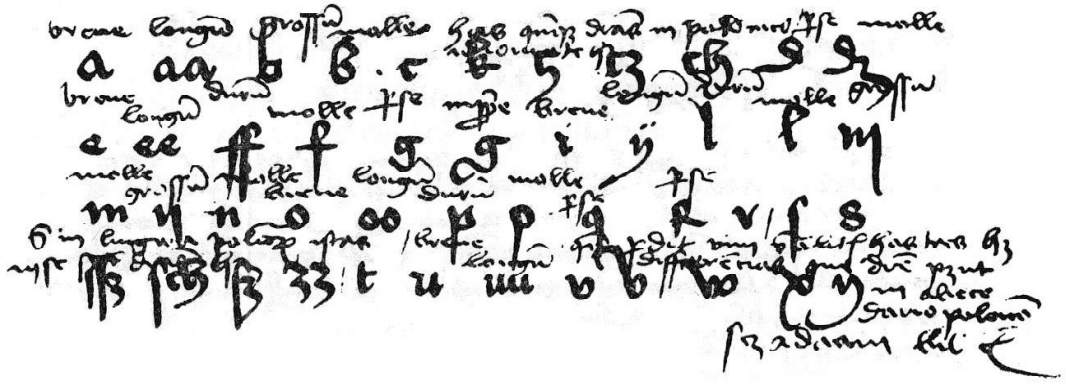


Fig. 6: Character samples from J. Parkosz (URBAŃCZYK/OLESCH 1985, 53)

Here, one can se the *square and round b* in the first line under the Latin descriptions ‘grossum’ (hard) and ‘molle’ (soft), and the *square and round p* in the third line, again with ‘grossum’ and ‘molle’ above them. As was already stated in the first article, Parkosz’s suggestions never became popular even

with his contemporaries because these distinctions were hard to realize and more a question of writing styles or font design than true graphemes.

We will present additional here material concerning the so-called *r rotunda*, the 'round r'. Just as the Latin *s* is available in Unicode in two forms, the standard form and the *long s*, the *r rotunda* needs to be encoded as the second form of the *r*:

Fig. 7: *r, r rotunda, s*, and *long s* ('Breitkopf' font)

In contrast to German, in Polish the *r rotunda* carried a functional load: it helped to distinguish -rz- [ž] from -r-z- [rz] (see URBAŃCZYK/ OLESCH 1985, 36 on Murzynowski 1551). The *r rotunda* was also in use for Sorbian, see the second line of the table in Fig 8:

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	—	—	*ŕ*	—	*ŕ*	—
ß (ſſ)	ß	ss	ß	*s*	*s*	*s*	s
ſch	ßch	sch	ſch	*š*	*š*	*š*	š
ſch	ſch	schj	ſch	*ś*	*ś*	—	—
t	t	t	t	*t*	*t*	*t*	t

Fig. 8: Alphabet chart from MUCKE (1891; repr. 1965, 18)

3.3. Unified Jers and Nasals

In Cyrillic texts, we find a so-called ‘unified jer’, which is written as a middle form between the two standard jers, and we also find so-called ‘unified nasals’:

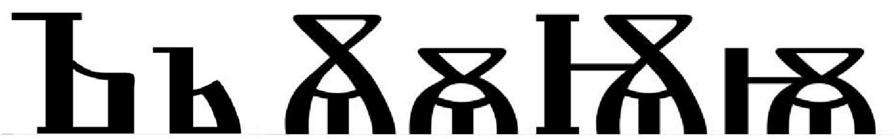


Fig. 9: Unified Jer and unified nasals (‘Method’ font)

All these character pairs are not yet available in Unicode. Their very essence lies in the fact the scribes could not decide or did not know or were not aware of which one of each pair to write, and therefore they went for a middle form. This, of course, means that these middle forms cannot be identified with either one as being a stylistic variant. At this point, it should be noted that these ‘unified’ letters are not identical to the characters used in the so-called ‘one-jer-texts’ which exhibit only one of the jers, but where the form of this letter is clearly one of the two.

3.4. Transliteration of Glagolitic Nasals

Now that the Glagolica has been included in Unicode in its version 4.1, it is only logical to check if all characters needed for the *transliteration* of Glagolitic into Cyrillic are available. First problems in this area were noted in the first article – the case of the missing *Cyrillic Iota*.

The transliteration of the nasal vowels from the Glagolitic alphabet into Cyrillic letters presents another currently unsolved problem. Let’s have a look at the transliteration:

Ⱔ	↔	Ѧ	✓
Ⱕ	↔	Ѧ	¬
Ⱗ	↔	Ѩ	✓
Ⱘ	↔	Ѫ	✓
Ⱙ	↔	Ѭ	✓
Ⱖ	↔	‘YO’	¬

The Ⱔ – nasal E (first line) – can be transliterated into Cyrillic. In the second line, there is special character that occurs in Zographensis and in Marianus: Ⱕ; it is encoded in Unicode as a separate glyph even if it is recognized to be a variant of the glyph in the first line.

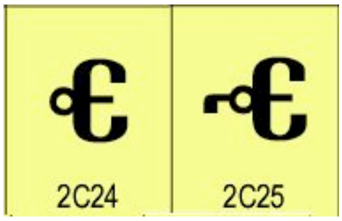
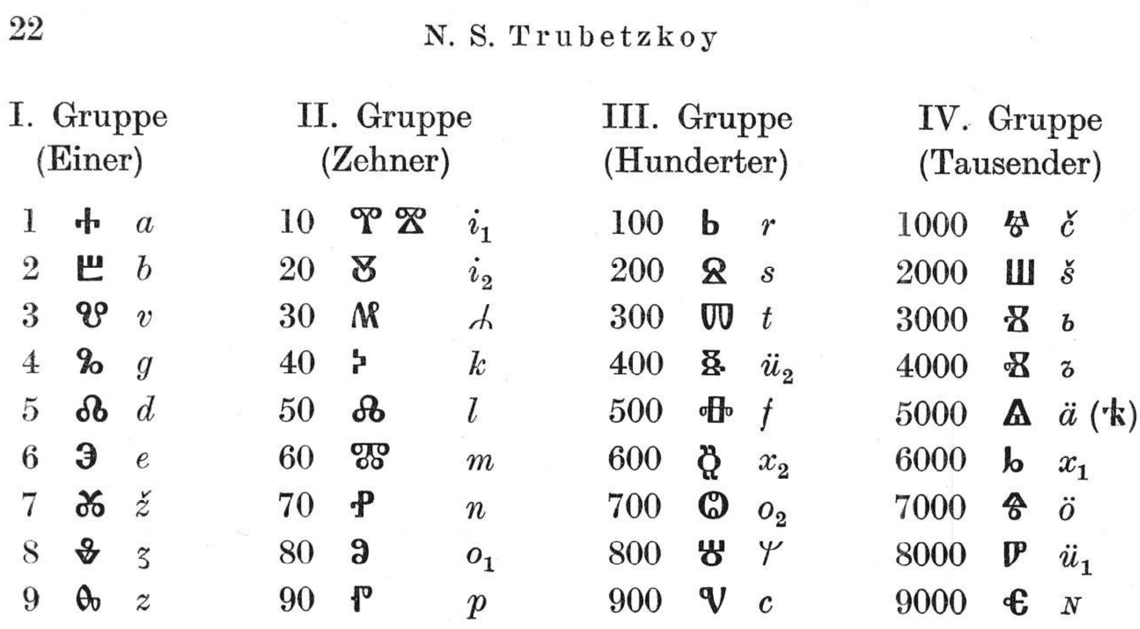


Fig. 10: Variants of EN in Unicode

According to TRUBETZKOY, the glyph Ⱕ denotes a nasal O; however, because its shape is derived form the nasal E (Ⱔ < Ⱕ), it is usually transliterated using a modified Cyrillic nasal E (Ѧ < Ѧ) whose form actually occurs in Suprasliensis (but is not encoded as a separate letter in Unicode). Consequently, the Ѧ should be added to the Cyrillic block not only in its own right, but primarily because it is needed to transliterate Ⱕ.

The jotated nasal E (Ⱗ), the nasal O (Ⱘ) and the jotated nasal O (Ⱙ) can all be transliterated, as lines three to five show. However, sometimes the need arises to single out the first part of the jotated nasal O, i.e. Ⱖ, for example, to write about it, and it also seems to occur in a non-connected form in actual manuscripts. Here, the situation becomes even less clear: it is unclear how to transliterate this part into Cyrillic; sometimes we simply find the Latin digraph ‘YO’ although it clearly does not represent the phonetic value correctly.

TRUBETZKOY’s OCS Grammar (1968, 22) presents two additional interesting transliteration characters:



22 N. S. Trubetzkoy

I. Gruppe (Einer)			II. Gruppe (Zehner)			III. Gruppe (Hunderter)			IV. Gruppe (Tausender)		
1	Ⰰ	*a*	10	Ⰹ Ⰺ	i_1	100	Ⱃ	*r*	1000	Ⱍ	*č*
2	Ⰱ	*b*	20	Ⰻ	i_2	200	Ⱄ	*s*	2000	Ⱎ	*š*
3	Ⰲ	*v*	30	Ⰼ	*ʎ*	300	Ⱅ	*t*	3000	Ⱏ	*ъ*
4	Ⰳ	*g*	40	Ⰽ	*k*	400	Ⱆ	$ü_2$	4000	Ⱐ	*ь*
5	Ⰴ	*d*	50	Ⰾ	*l*	500	Ⱇ	*f*	5000	Ⱑ	*ä* (ѣ)
6	Ⰵ	*e*	60	Ⰿ	*m*	600	Ⱈ	x_2	6000	Ⱒ	x_1
7	Ⰶ	*ž*	70	Ⱀ	*n*	700	Ⱉ	o_2	7000	Ⱖ	*ö*
8	Ⰷ	*ʒ*	80	Ⱁ	o_1	800	Ⱛ	*ϒ*	8000	Ⱓ	$ü_1$
9	Ⰸ	*z*	90	Ⱂ	*p*	900	Ⱌ	*c*	9000	Ⱔ	*N*

Fig. 11: TRUBETZKOY’s transliteration of Glagolitic

The transliteration for numbers “30” and “800” are unusual additions, derived from the shapes of the Glagolitic letters, it seems, and mixed into the set of Latin letters used for the rest of table. This seems to be a singular use of these nonstandard characters, so they may not merit inclusion into Unicode.

This, however, brings us to the next interesting area:

3.5. Transliteration of Glagolitic into Croatian (Latin)

While Glagolitic has been and usually is being transliterated into Cyrillic, we should not forget that in Croatia this is not the case: there, the Croatian or Square Glagoljica is transliterated using Latin characters.[3] If we take a look at Fig. 12, we see one character that is not present in Unicode:

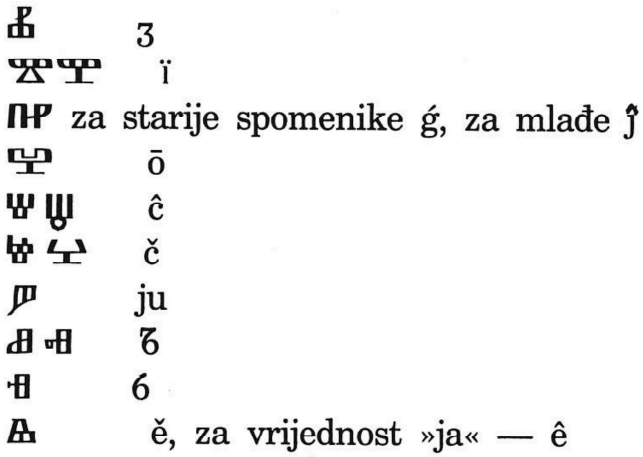


Fig. 12: Croatian transliteration of Glagolitic (BRATULIĆ 1995, 98)

A not-so-common character is the c with circumflex (ĉ) which is available in Unicode at [0108] and [0109]; it can, of course, also be composed from its parts. The j with diacritic needs to be looked at more closely. While Fig. 12 seems to show a *j with circumflex* (which is available in Unicode at [0134] and [0135]), another figure from the same source clearly shows shows an inverted breve:

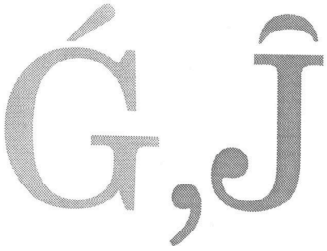


Fig. 13: Croatian transliteration of Glagolitic (BRATULIĆ 1995, 146)

3 Bulgarian or Round Glagolica and Croatian or Square Glagoljica are considered to be the same script in Unicode, and the difference between them to be stylistical.

A *j with an inverted breve* is not yet available in Unicode. Also, such a character cannot be made up from its parts because a 'dotless j' is also not yet available so that would be another character that needs consideration.

3.6 Croatian Glagoljica: variants or separate letters?

The encoding of the Glagolica presents another interesting case: in the first drafts of the original submission, the two variants representing the back jer (first line) were both present individually, just as the two variants representing the front jer (second line).

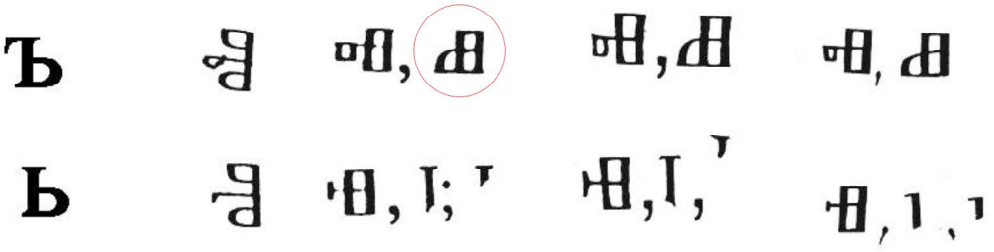


Fig. 14: Glagoljica variants of the Jers

In the final proposal, however, only the variants for the front jer survived, while the second variant for the back jer (encircled in Fig. 14) has not been encoded separately. While the two characters are surely related, one could still have preferred to have them both available.

3.7 Bosančica

This brings us to the one Slavic script that seems to have not yet been considered as such at all: The Bosančica.

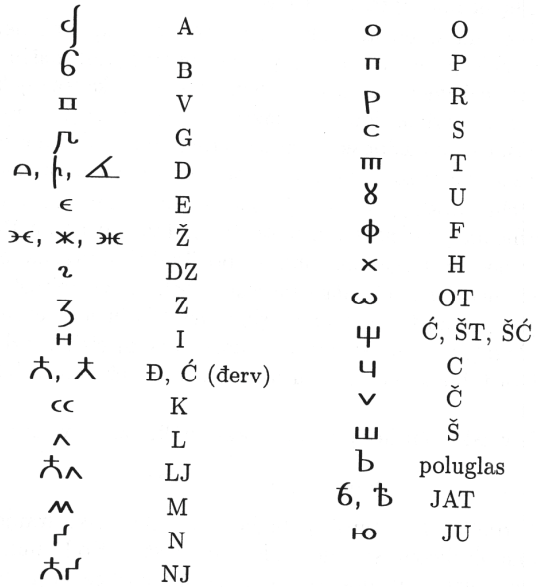


Fig. 15: Bosančica alphabet (from Žubrinić 1996,70)

The question which of these characters are different enough to warrant separate encoding is not easy to answer. Several candidates could be singled out, primarily ‘V’, ‘D’, Ć/Đ, and ‘N’, with Ć/Đ being the favorite.

3.8. Superscripts

Superscripts are another example where a solution has already been implemented for medievalists working with the Latin alphabet, while no comparable solution exists for Cyrillic. Nearly the complete Latin alphabet is currently already as superscripts; Fig. 16 shows a sub-set only:

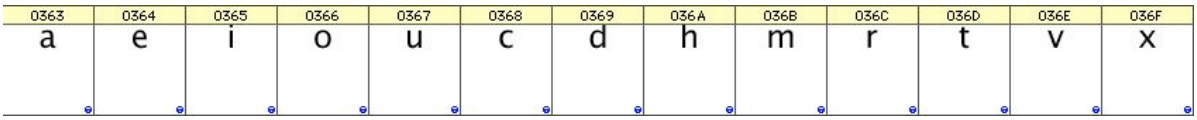


Fig. 16: Combining Latin superscripts in Unicode

These superscripts allow the writing of u͛ (= ü), a͛ (= ä), o͛ (= ö) etc., which is very important for medieval German.[4] However, as we said, no superscripts are available for Cyrillic, even if some have very distinct shapes, see Fig. 17 for the ‘d’ and the ‘z’, maybe less so for the ‘x’.

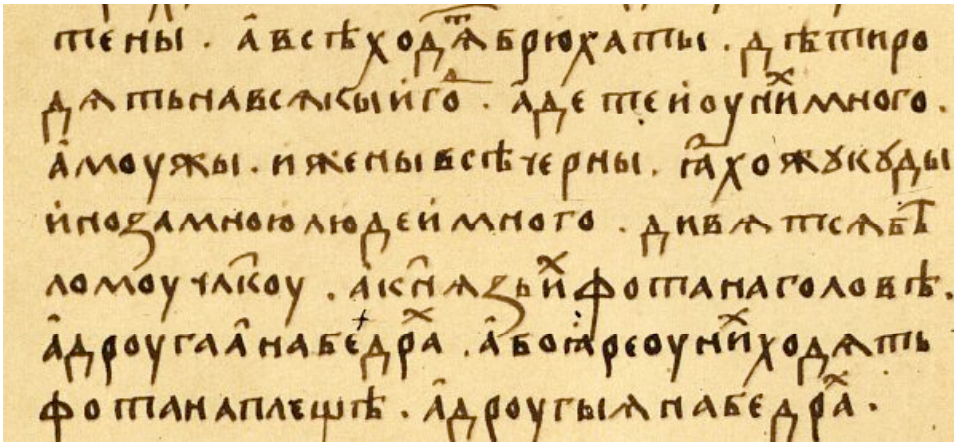


Fig. 17: Cyrillic superscripts in Afanasij Nikitin’s ‘Voyage Beyond the Three Seas’ (*Troickij spisok*)

4 German orthography is remarkable in that it always had clear rules of how to replace non-ASCII-characters by sequences of ASCII characters: ä > ae, ö > oe, ü > ue, ß > ss. These same replacements are also being used today in e-mail addresses (i.e. a person called ‘Müller’ will always choose the sequence ‘mueller’, never ‘muller’, for his/her e-mail address).
These rules do not seem to be well known abroad where the umlaut characters are usually simply replaced by the base character instead (i.e. ä > a, ö > o, ü > u). Very strange-looking is, however, the practice to subsitute a Greek beta (β) if no German *double s* (ß) is available – this can sometimes be observed in Russian printing.

3.9. Ligatures

Ligatures are, by their very nature, a most pleasing subject, aesthetically speaking. The history of Latin typesetting knows much more ligatures than a normal user today will be aware of. Fig. 18 gives an impression of an extended set of ligatures in a modern 'expert' digital font.

Fig. 18: Expanded set of Latin ligatures in a modern typeface

In Unicode, ligatures are considered "presentation forms". Quite a number of such forms are required for Arabic where the shape of a character depends on its position in the word (beginning, middle, end). For reasons of compatibility with legacy code pages, Unicode preserves some ligatures for Latin (see Fig. 20, left part). However, for Cyrillic and Glagolitic, there are no ligatures at all available in Unicode. Both script systems do know quite a number of ligatures. In Cyrillic, there are at least a dozen common ligatures (see Fig. 19), but there are hundreds of actual ligatures to be found in texts.

Fig. 19: Sample OCS Cyrillic ligatures ('Method' font)

The same is true for Glagolitic, and Fig. 20 (right half) shows a well-known table listing ligatures consisting of two and even three characters. It may be worth mentioning here that Greek manuscripts also show a large number of ligatures, which are also not present in Unicode.

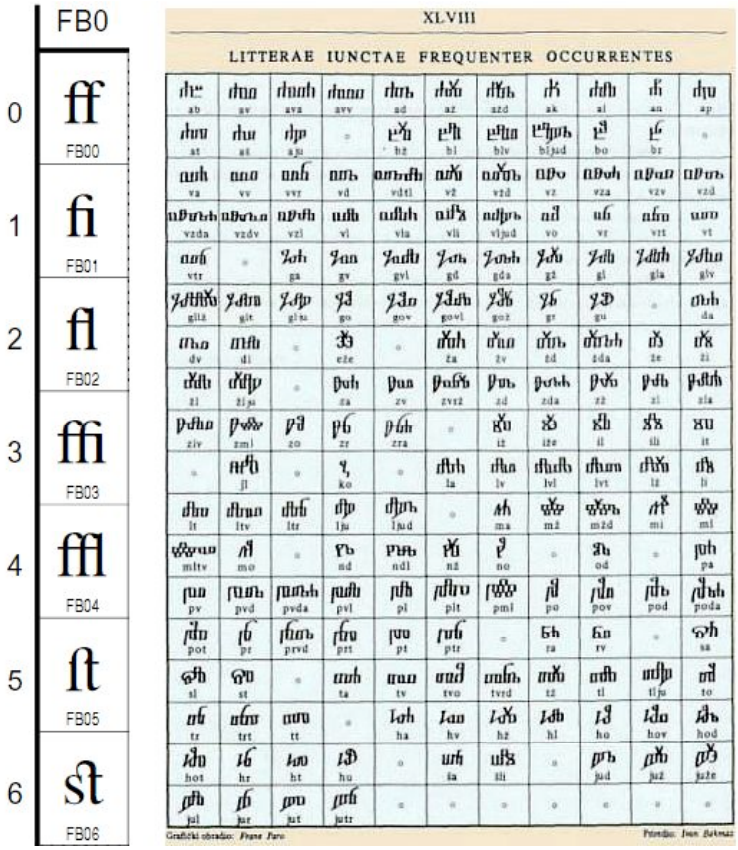


Fig. 20: Latin ligatures in Unicode, and a table of Glagolitic ligatures

3.10. Numbers

Fig. 21 shows the "Numbers" block in Unicode, consisting of various fractions and then Roman numerals in uppercase and lowercase:

2150	2151	2152	2153	2154	2155	2156	2157	2158	2159	215A	215B	215C	215D	215E	215F
			⅓	⅔	⅕	⅖	⅗	⅘	⅙	⅚	⅛	⅜	⅝	⅞	⅟
2160	2161	2162	2163	2164	2165	2166	2167	2168	2169	216A	216B	216C	216D	216E	216F
Ⅰ	Ⅱ	Ⅲ	Ⅳ	Ⅴ	Ⅵ	Ⅶ	Ⅷ	Ⅸ	Ⅹ	Ⅺ	Ⅻ	Ⅼ	Ⅽ	Ⅾ	Ⅿ
2170	2171	2172	2173	2174	2175	2176	2177	2178	2179	217A	217B	217C	217D	217E	217F
ⅰ	ⅱ	ⅲ	ⅳ	ⅴ	ⅵ	ⅶ	ⅷ	ⅸ	ⅹ	ⅺ	ⅻ	ⅼ	ⅽ	ⅾ	ⅿ

Fig. 21: Fractions and Latin (Roman) numbers in Unicode

In contrast to this, for Slavic no precomposed numbers are available. As a visual reminder of the Slavic writing conventions for numbers might serve Fig. 22. For Slavic, we have the 'thousand' symbol encoded, but nothing else.

Fig. 22: Slavic numbers (source: web, from an unnamed 1977 almanac)

That is, for each number, the user is expected to type at least four parts: a half-high dot, then the letter, a titlo over the letter, and again a closing half-high dot. The problem here is that to successfully compose the Slavic numbers from their parts, we need a titlo that goes above *two letters* – see numbers 11 to 19, and such a titlo is not yet available, only one which goes directly above a letter. The best implementation for the longer titlo would probably be a nonspacing diacritic that is to be typed between the two letters, very similar in principle to the 'combining double macron' that is available in Unicode at [035E]. So, a '*combining double titlo*' is what should be added to Unicode.

3.11. Balkan Philology

A broader perspective on Slavic writing systems is in order if we want to fully cover all their uses. First, there is the use of *OCS Cyrillic in Romania*, where it was used until the 19th century. From a manual of Romanian paleography, there is clear evidence of an additional accepted OCS letter (see Fig. 23).

LITERA Ѧ

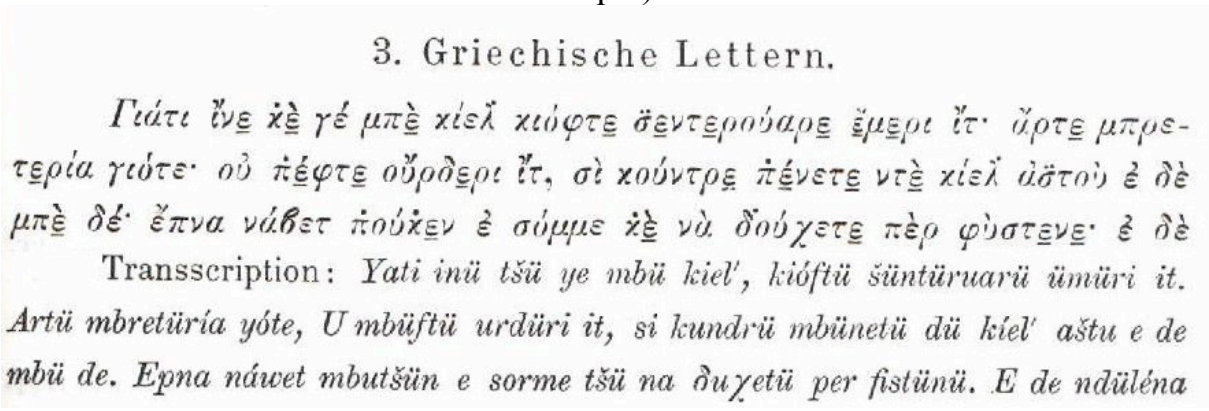
Litera Ѧ, obținută prin modificarea grafică a literei Ѫ[78], în documentele muntenești ale secolului al XV-lea apare o singură dată, într-un act de întărire dat pentru mănăstirea Govora[79] de voievodul Radu cel Mare: **ѡт глав поıанє долє по ѦН[80] долинѹ Чоржчєв** „din capul poienei, în jos, prin (pe în) valea Ciorîcei" (1499 iul. 13, Arh. St. Buc., S. I., nr. 118).

Deși atestată într-un cuvînt românesc, totuși unicul exemplu în care găsim această literă nu ne dă posibilitatea de a ne referi la valoarea sau valorile acestei litere în general[81]. În acest unic caz, grafia ne determină să afirmăm că Ѧ notează nazalitatea vocalei care precedă pe Н[82].

Fig. 23: OCS letter 'IN' in Romanian (DJAMO-DIACONIȚĂ 1971, 49)

This same character is also known from Slavic texts, representing a similar, but not identical sound.

It is well-known that the *Greek* script has been used on the Balkans for neighbouring languages: It was in use in Bulgaria during the First Kingdom before OCS was to become the official language with its own alphabet, and it was used for Macedonian in the 19th century before the language acquired an official status (after WWII). The Greek script was also used for Albanian, and in the sample shown in Fig. 24 we see some interesting character-diacritic combinations not known in Greek itself (epsilon with underline, kappa with over-dot, epsilon with gravis and underline, sigma with dieresis, pi with over-dot – all in the first two lines of the sample).



3. Griechische Lettern.

Γιάτι ἴνε̱ κὲ̱ γέ μπὲ̱ κιελ̓ κιώφτε̱ ἄε̱ντερούαρε̱ ἔμε̱ρι ἴτ· ἄρτε̱ μπρε-
τερία γιότε· οὺ πέ̱φτε̱ οὔρδε̱ρι ἴτ, σὶ κούντρε̱ πέ̱νετε̱ ντὲ̱ κιελ̓ ἀάτοὐ ἐ δὲ
μπὲ̱ δέ· ἔπνα νάβετ πούκε̱ν ἐ σύμμε κὲ̱ νὰ δούχετε̱ πὲρ φῦστε̱νε̱· ἐ δὲ

Transscription: *Yati inü tšü ye mbü kiel̓, kióftü šüntüruarü ümüri it.*
Artü mbretüría yóte, U mbüftü urdüri it, si kundrü mbünetü dü kiel̓ aštu e de
mbü de. Epna náwet mbutšün e sorme tšü na duχetü per fistünü. E de ndüléna

Fig. 24: Albanian written using Greek letters (after FAULMANN 1880)

At present there is no evidence that the Arabic script when applied to Macedonian yielded any additional characters, but its application to Byelorussian and Bosnian would have to be carefully researched to see whether this adaptation resulted in any characters or character plus diacritic combinations that are not available yet in Unicode or cannot be produce from available parts.

3.12. General Phonetics

Concluding our overview of areas that should be considered from a Slavicist's point of view, general phonetics should no be completely forgotten. While most phonetic symbols are already available in Unicode, there are some symbols that were either not sanctioned by the IPA but still are in use or have been suggested at some point in time. Fig. 25 shows a list of various such additions not yet implemented in Unicode at present.

Fig. 25: Some phonetic symbols missing from Unicode

Many of these symbols are present in PULLUM/LADUSAW (1986), an excellent source on phonetic symbols, their history and meaning. It might be worth noting that one these symbols is being used by the Bavarian dialect atlas:

Fig. 26: open o-e ligature (PULLUM/LADUSAW 1986, 120)

4. Encoding Strategies and Questionable Characters

4.1 Ukrainian Ghe-upturn

Not every special character, however, that researchers have noted needs inclusion into Unicode. Let us look at such a case here.

TRUNTE (2001, 324–327) remarks that SMOTRYC'KYJ uses a special character in his 1619 grammar and he even typesets it. The character in question can be seen in Fig. 27 in the Greek loanwords *Grammátiki, Orfográfia, Etymologïa*. TRUNTE calls this character an 'allographe' or 'variant' (2001, 325) of г, and he says that it "stands for" Greek gamma (327).

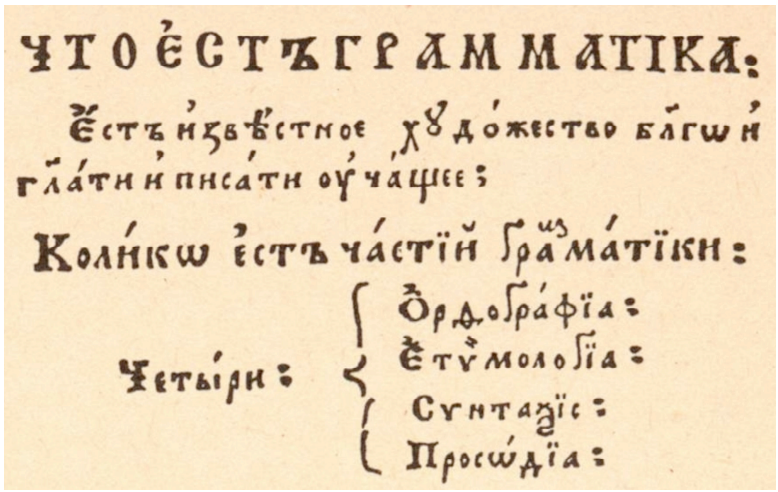


Fig. 27: Use of Ghe-upturn in SMOTRYC'KYJ (1619, a v)

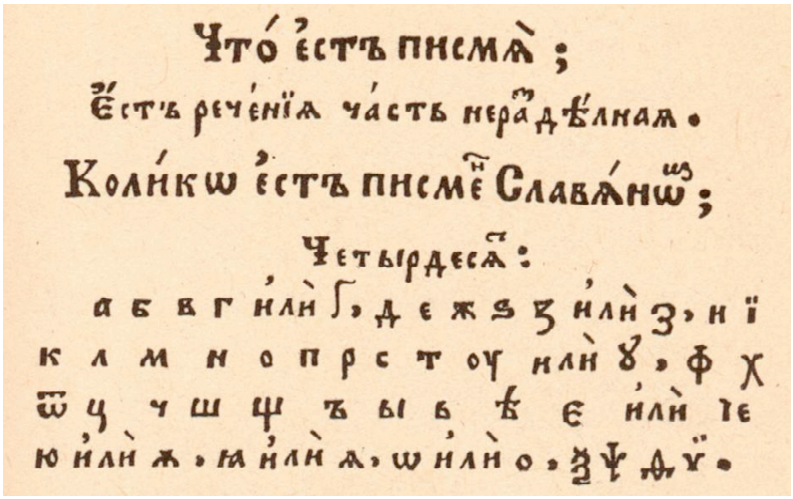


Fig. 28: Alphabet from SMOTRYC'KYJ (1619, a g)

Now what do these pictures really show? We see a "g" that does not fit the rest of Church Slavonic letters very well, and although it is reminiscent of the *capital* Cyrillic letter Г or Ґ, it is clearly a *lowercase* letter in both figures. Note that SMOTRYC'KYJ uses the 'normal' Г in the uppercase-only title question 'What is Grammar?' and, in the same word, but printed in lowercase letters, the Ghe-upturn in the fourth line of Fig. 27.

By pure chances, the author happened to have the label of a Greek wine bottle which allowed for some interesting observations: This label has the same letter in it which SMOTRYC'KYJ uses: again, it clearly is a lowercase letter (see the first word in the first line, where it is used twice, and the first word in the second line of the text). Now it becomes clearer that what SMOTRYC'KYJ uses is nothing but a *Greek lowercase Gamma* in the context of OCS letters. This explains the difference in design (it is lighter than the true OCS characters) and the special form.[5]

Fig. 29: Modern Greek wine label with script gamma

5 The same Greek gamma is also present in the table of Greek letters in Ivan Fedorov's Ostrog 'Azbuka' from 1578 (see 1983, 2).

As the wine label shows, this form of the Greek lowercase gamma is still in use today when a font is required that should look a bit more traditional and handwritten.

If we now take a look at a Ukrainian grammar (RUDNYC'KYJ 1943, 2), we again see the same character in his alphabet table where he shows printed and handwritten forms of all characters:

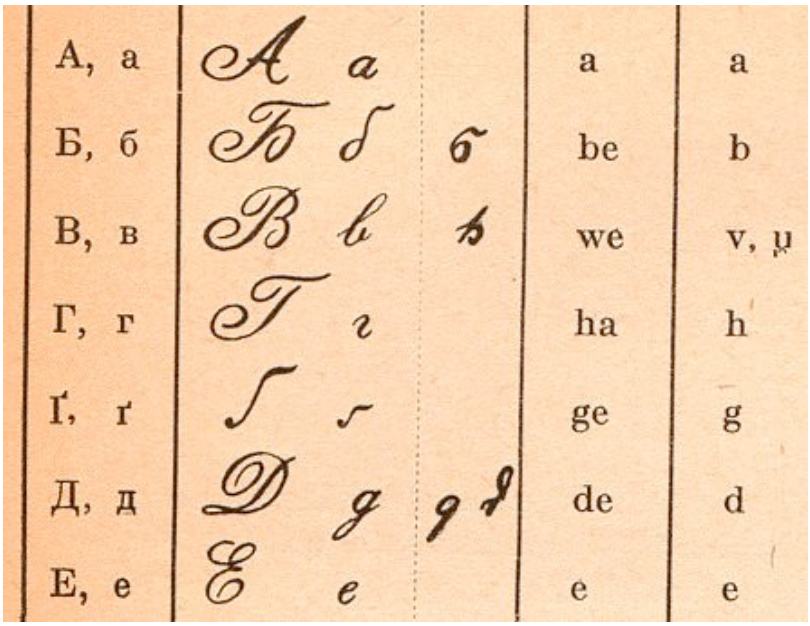


Fig. 30 Ukrainian G and Ghe-upturn (RUDNYC'KYJ 1943, 2)

As we can see from these figures, the Greek Gamma has been adapted into the construction principles underlying the Cyrillic alphabet which features many lowercase characters that are "small versions" of the respective uppercase letter where the Latin alphabet uses special forms for the lowercase letters (which derive from handwriting). Just compare Cyrillic T-т, M-м to Latin T-t, M-m etc.[6] That is, the special tall Greek lowercase Gamma which SMOTRYC'KYJ uses was later reinterpreted as being an uppercase character, and a small version was derived from it to form the lowercase character. This also explains why in hand-written form, the normal Cyrillic G and the Ghe upturn look different, although the printed forms are very similar to each other (see Fig. 30).

So, the remarks by TRUNTE could and probably should be worded a bit differently: In SMOTRIC'KIJ's Grammar we do see a new character, and because it stands for the sound [g], and not [h], it is a *phoneme* and not an *allophone*: if it is already considered to be a *Cyrillic* character, it isn't a tall variant of г, but of ґ ! The very fact, however, that the tall Greek version of the Gamma is being

6 See KEMPGEN (1993) for some quantitative measurements of alphabet systems with regard to their internal structure.

used can be considered as a proof that this is still a borrowed Greek character and not yet a Cyrillic character. In contrast to this, one can indeed say, as TRUNTE does, that ѳ ‘stands for’ Greek theta (θ) in certain manuscripts, because both characters are different to each other and clearly belong to different script systems. But in our case we simply observe the origin of the Ukrainian ‘Ghe upturn’ character which began its life as a borrowed tall lowercase Greek script gamma.

If one wants to typeset SMOTRYC’KYJ today, there would be several solutions: one could simply use the modern Cyrillic *ghe upturn* character (i.e. ґ) or, if more fidelity with the look of the original is required, an alternate tall form of the lowercase *ghe upturn* could be used. Finally, another obvious solution would be to use a *Greek* font with a tall lowercase gamma, just as SMOTRYC’KYJ did for his ‘Grammar’. In any case, this is not a new character that would be missing from Unicode, and it seems a somewhat strange omission that TRUNTE does not identify the special character that he sees in SMOTRYC’KYJ’S ‘Grammar’ with the Ukrainian *ghe upturn* of today.

4.2. Encoding variants

This brings us to a more general consideration of how variants can be encoded within the Unicode framework. Let us take the Jery as an example (see Fig. 31). First, we have an old form (with the ‘hard sign’ as its first part) and a newer form (which has the ‘soft sign’ as its first part. This newer version also has two basic connected or ligature variants (see middle section): the connecting line either is somewhere at half height or at the top the character. And then we also have instances where the second part already has the dotted i as its second part.

Fig. 31: Variants of the ‘jery’

Basically, there would be four ways to encode variants:

1) use of the ‘private area’ for everything but the standard form;
2) separate fonts – one for the standard form, another one having variants;
3) use of font and software technology;
4) have Unicode, Inc. add them to the standard.

These solutions have their advantages and disadvantages:

The use of the ‘private are’ has the advantage of having all variants in a single font and the disadvantage of lack of compatibility between fonts and their users’ documents.

The use of different fonts for variants has the advantage that all variants have the same Unicode number as the basic glyph which means all software treats them as the same character with only different 'looks'. An advantage is also that such a solution is immediately possibly. The disadvantage here is that for each new variant of a letter, one needs a new font – possibly a dozen for some characters. As only certain characters do have variants at all, this means that many slots in these additional fonts would remain empty because they are not needed. This clearly is not the most economical or practical way to handle variants either.

Solution 3 can be demonstrated by taking a look at the Mac OS X version of 'Lucida Grande', the system font. This font has some Czech alternate characters built-in, but not in the private area, see Fig. 32.

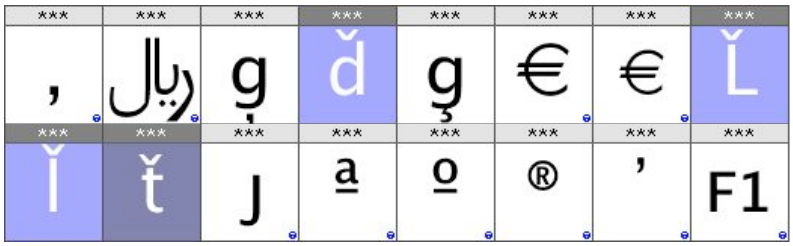


Fig. 32: Built-in variants in 'Lucida Grande' font

The clue to the non-use of the private area are the three asterisks above each of the character: characters in the private area have their own number which is not the case here. A user accesses these alternate shapes through functions built into his word processor. 'Text Edit', for example, OS X's standard text editor, has a menu entry 'Character Shapes' > 'Traditional form'. See also Fig. 33 for a possible user interface solution – it shows a part of the 'character palette' window where a user can simply point and click on a variant to select it.

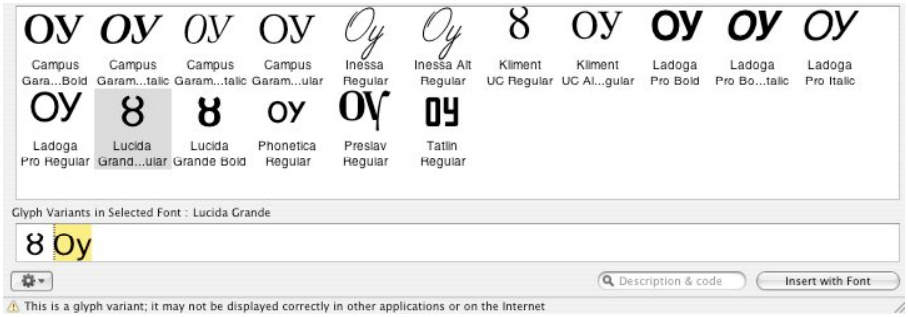


Fig. 33: User interface for selecting variants in OS X

The advantage of such a solution is that all variants are built into the basic font, no redundancy of empty slots in additional fonts is present, but there are also some important disadvantages: in this case the user must rely on software

vendors of word processors to implement a function to select 'traditional forms' in their software. Also, the correct display of these alternate shapes cannot be guaranteed with fonts other the one that has originally been used, and cross-platform compatibility is even less clear. This, therefore, is the solution that requires the most coordinated effort by various software vendors (font vendors, word-processor vendors, and system software vendors) and as such is the least practical solution.

Solution 4 would be ideal if we could rely upon Unicode, Inc. to add all recognized variants to the standard in a timely manner and font vendors to follow up with expanded character sets in their fonts. However, the disadvantage here is that no immediate solution is possible if a solution is needed now and not in several years time. Also, more importantly, the success of such a strategy is unclear, even if some variants are already encoded as separate glyphs in Unicode:

Fig. 34: Sample variants available in Unicode

Another problem is the important question which shapes to treat as different characters and which ones to treat as variants, stylistic or other.

5. Conclusion and Outlook

As both papers have shown, there are quite a few areas in Slavic philology that need careful consideration which characters to submit to Unicode, Inc., for inclusion in a next revision of the standard. These amount to several dozen characters at the very least. If we compare the treatment that Latin has been given with the treatment of Cyrillic and Glagolitic, we clearly see that the Latin alphabet is at present much better supported with respect to the needs of medievalists, paleography etc., although even there many wishes remain unfulfilled at the moment. Much of the support an alphabet receives from Unicode and within Unicode depends on the initiative of individuals working in the field who are at the same time interested in computer technology and therefore willing to submit proposals to the Unicode consortium.

As of this writing (April 2006), Ralph CLEMINSON has submitted a proposal to Unicode, Inc. for inclusion of some 40 characters (most of them pairs). The author of the present article contributed various illustrations and suggestions apart from his oral presentation at the Sofia conference (which formed the basis for both articles). All characters included in the proposal are *Cyrillic* – Old Church Slavonic and its transliteration as well as Old Russian are the main

focus. The proposal does not include *all* Cyrillic characters mentioned in our two articles, but many of them, and *no Latin* characters. Consequently, more work needs to be done, in the area of Cyrillic characters as well as in the area of Latin characters, and even Glagolitic.

Abstract

The present paper is the second in a series presenting an overview of Slavic philology with respect to version 4.1 of the Unicode standard. Reviewing East, West and South Slavonic languages, their alphabets and writing systems, the first paper already revealed at least a dozen characters that need to be encoded in Unicode, among them several (soft) Cyrillic characters with a tail at the right side, the Cyrillic old-style Ѣ, a Cyrillic Iota (uppercase and lowercase), a Cyrillic dotless lowercase ı, the Cyrillic Paerok (no distinction between uppercase and lowercase), number signs, accents etc. This second article presents even more characters, among them a Latin S with stroke, the 'r rotunda', problems in the transliteration of Glagolitic into either Cyrillic or Latin, a broader perspective on Balkan philology etc.

References

Bratulić, J.:
1995 *Leksikon hrvatske glagoljice*. Zagreb.
Cleminson, R.:
2006 *Proposal for additional cyrillic characters*. pdf file available from: http://www.unicode.org/~dwanders/06042-cleminson-cyrillic.pdf
Djamo-Diaconiţă, L.:
1971 *Limba documentelor slavo-române emise în Ţara Românească în sec. XIV şi XV*. Bucureşti.
Faulmann, K.:
1880 *Illustrirte Geschichte der Schrift*. Wien–Pest–Leipzig.
Kempgen, S.:
1993 Spezifika slawischer Schriften. In: ders. (ed.), *Slavistische Linguistik 1992*, München, 111–143.
2006 Slavic Philology and Unicode (I). To appear in: A. Miltenova et al. (eds.), *Proceedings of the 'Azbuky net' Conference, Sofia, Oct., Oct 24-27, 2005*. Sofia. 25 pages.
Mucke, K.E.:
1891 *Historische und vergleichende Laut- und Formenlehre der niedersorbischen (niederlausitzisch-wendischen) Sprache*. Leipzig (Reprint 1965).

Parkoszowicz, J.:
1985 *Traktat o ortografii Polskiej Jakuba Parkosza*. Opracował Marian Kucała. Warszawa: PAN.
Pullum, G.K., Ladusaw, W.A.:
1986 Phonetic Symbol Guide. Chicago–London.
Šwela, B.:
1952 *Grammatik der niedersorbischen Sprache*. Leipzig (2nd ed.). First edition Cottbus 1906.
Trubetzkoy, N.S.:
1968 *Altkirchenslawische Grammatik. Schrift-, Laut- und Formensystem*. Hrsg. v. R. Jagoditsch. 2. Auflage. Graz–Wien–Köln.
Trunte, N.:
2001 *Slavenskij jazyk. Ein praktisches Lehrbuch des Kirchenslavischen in 30 Lektionen. Zugleich eine Einführung in die slavische Philologie. Band 2: Mittel- und Neukirchenslavisch* (Slavistische Beiträge, Bd. 370, Studienhilfen, Bd 9). München.
The Unicode Consortium
1991 *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* (Slavistische Forschungen, Bd. 37). Köln–Wien: Böhlau.
Žubrinić, D.:
1996 *Hrvatska Glagoljica*. Zagreb.

Федоров, И.:
1983 *Азбука Ивана Федорова 1578*. Москва.

Appendix: Some fonts and their features for slavists

	Kliment Std v. 1.7 (Win/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	some	most	some
Croatian Digraphs	✓	✓	– –
Maced. Translit. (ǵ)	✓	✓	– –
Štokavian Accents	✓	✓	– –
Nasal o	✓	✓	– –
Latin Extended Additional (256)	some	✓	ca. 1/3
Maced. Translit. (ḱ)	✓	✓	– –
Russ. Hist. Translit. (f̀, ỳ)	✓	✓	– –
Sorbian (ḿ, ṕ)	✓	✓	– –
IPA – Phonetic	some	2/96	1/96
Spacing Modifiers	✓ **80/80**	11/80	9/80
Translit. of Jers	✓	– –	– –
Combining Diacritics (= „Flying Accents“)	✓ **112/112**	40/112	5/112
Greek			
Modern Greek	✓	✓	✓
Archaic Letters (Koppa, Stigma, Sampi…)	✓	– –	– –
Classical Greek	– –	✓	– –
Cyrillic			
Std. Russian & Slavic	✓	✓	✓
Macedonian Add. (è, ѝ)	✓	✓	– –
Hist. Add. (ѣ ѧ ѫ …)	✓	– –	– –
Ukrainian Ghe (Ґ ґ)	✓	✓	✓
Non-Slavic Cyrillic (ex GUS-Countries)	– –	ca 1/2	ca. 1/10
Glagolitic	– –	– –	– –
Transliteration into Cyrillic	✓	– –	– –

	Kliment Std v. 1.7 (Win/OS X)	Times/ Helvetica v. 5.0d10e1 (OS X)	Times NR/ Arial v. 3.05 (Win/OS X)
Armenian, Georgian, Hebrew, Arabic, Ethiopian	– –	– – (supported by other fonts)	Hebrew, Arabic
General Punctuation	**70/112**	18/112	27/112
Superscripts/Subscripts (0…9)	**30/46**	– –	– –
Currency (Euro…)	1/48	3/48	6/48
Comb. Diacr. for Symbols (O)	✓	– –	– –
Number Forms	**49/64**		
Add. Fractions (2/3…)	✓	– –	6/13
Roman Numerals	✓	✓	– –
Arrows	6/112	– – (complete in Apple Symbols)	7/112 (complete in Wingdings)
Mathematical Operators (∏, ∫, ≠ …)	45/256	12/256 (complete in Apple Symbols font)	15/256 (complete in other fonts)

Bibliographische Angaben / Bibliographical Entry:

Sebastian Kempgen: Unicode 4.1 and Slavic Philology – Problems and Perspectives (I). In: A. Miltenova, D. Radoslavova, E. Pancheva (eds.), *Computer Applications in Slavic Studies. Proceedings of Azbuky.net. International Conference and Workshop. 24–27 October 2005*, Sofia, Bulgaria. Sofia 2006, 131–159.

The original presentation given at the conference is available separately.

Bibliographische Angaben / Bibliographical Entry:

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

The original presentation given at the conference is available separately.