Old Slavonic and Church Slavonic in TEX and Unicode

Alexander Berdnikov, Olga Lapko

The characteristic features of Cyrillic (Old Slavonic and Church Slavonic) writing systems are analyzed and compared. The old numbering rules and the difference between the canonical orthodox Church Slavonic and 'old believer' Church Slavonic are considered as well. It is shown that Old Slavonic and Church Slavonic differ strongly, and should at the very least be considered as two well distinguished dialects of the same writing system. An analysis of the current state of the Unicode 04xx encoding page shows that it is not sufficient to represent the Old Slavonic and Orthodox Church Slavonic writings adequately. The project of T2D encoding which enables the representation in TEX of out-of-date Bulgarian texts (from the middle of the 19th century till 1945), Russian texts (1703–1918 and emigrant literature) and Church Slavonic/Old Slavonic texts, is described.

Introduction

When in December 1998[1] the encodings T2A/T2B/T2C became a standard part of LATEX 2ε, a significant break between Latin and Cyrillic alphabets as supported by LATEX 2ε was eliminated. But there are still a lot of symbols present in Cyrillic and absent from LATEX (see [1], for example)—namely, the Old Slavonic and Church Slavonic letters—and it is necessary to add them to the set of LATEX encodings as well.

But before we can do that, we should investigate the Old Slavonic and Church Slavonic writing systems a little bit more.

A brief history of Cyrillic

Cyrillic is a relatively young writing system, and we know (or at least we think that we know) its authors. Slavonic writing was invented by St. Cyrill (Constantine) and St. Method in 863 (or in 855 as some historians state). It is more or less proved that at first it appeared in a form currently called *Glagolitic* although ancient historical books call it *Cyrillic* for the name of its inventor

[1] Ok, *this* time 'December 1998' appeared to be in April 1999 ☺.

St. Cyrill. A little bit later (between 893 and 927) the writing system which we call Cyrillic appeared, and there is a hypothesis that it was introduced not by St. Cyrill and St. Method but by their successor and disciple Climent Ochrydsky. (It is also not clear whether some letters were added later to the original Cyrillic alphabet or not.)

Although Cyrillic letters are quite different from Glagolitic, there is a nearly one-to-one correspondence between the glyphs of these writing systems. (Cyrillic in general is more rich—it has a longer history and many glyphs were added to it after Glagolitic was already dead. Just one symbol *gherv* exists in Glagolitic but not in Cyrillic—it corresponds to a sound which disappeared soon after 863.) Since the phonetic analysis and decomposition of the Old Slavonic language was more important work than the assignment of graphical shapes to these sounds, it is acceptable to call St. Cyrill and St. Method the authors of Cyrillic even if Climent Ochrydsky or somebody else is the actual author of its graphics.

After its appearance the new writing system became popular and there is an enormous quantity of manuscripts based on Cyrillic (Old Bulgaria was a growing and cultural kingdom). Cyrillic became the writing system for the significant part of the Slavonic world (at least the part that adhered to the eastern branch of the Christian church). Due to slow divergence of the united Slavonic people into nations, different writing traditions became specific for different regions. In parallel, the Slavonic language itself evolved as well—for example, some sounds became obsolete even for the next generation.

When writing with thin reeds was superseded by writing with pens (goose feathers), the quantity of manuscripts increased rapidly,[2] and writing rules became much less rigorous and more dependent on the writer. The other effect specific to this period (due to the expense of material—i. e., parchment—used for writing) is that abbreviations, the abbreviation symbols, and the trend to compress letters and to create vertical ligatures appeared widely in Cyrillic.

The Slavonic writing system was influenced also by the fact that many texts were copied from the original Greek sources—so, the aspiration symbols (hard and soft) are placed arbitrarily and mean nothing in Slavonic texts;[3] letters ξ, ψ, θ, ω in words taken from Greek represent just the same sounds as letters

[2] But there may be no causal relation between these two facts: both are provoked by the same permanent process of cultural growth [16].

[3] Some sources insist that aspirations are conserved in Old Slavonic texts for calligraphic reasons only and are copied directly from the Greek texts; some sources [16] state that placing of aspiration symbols is not arbitrary—although we don't know exactly what they mean, they are

кс, пс, ф, о in Slavonic words (and quite often were substituted by them), and the numbering system (numbers are represented by letters—see section "Numbering system" on Page 15) follows the order of the Greek alphabet, etc.

So, there are many variant letter forms and writing rules for the manuscripts created during this period. Quite literally, ancient Cyrillic writing was characterized by some anarchy ☺ instead of well-defined rules, and Cyrillic manuscripts display an impressive variety of glyphs, styles and traditions of writing. The canonical Old Slavonic alphabet is represented in [2, 3], but it does not cover the whole variety of Old Cyrillic (see [18], for example).

In the middle of the 14th century Balkan Slavonic countries were seriously assaulted by Osmans, and at the end of the 14th century they were conquered and almost totally destroyed (although the remainder of the Byzantine Imperium fell only in 1453). It was a great loss for Slavonic culture, and since that time the centre of Slavonic writing was moved to the East. The process of spontaneous orthographic and phonetic evolution of the Slavonic writing system continued there as well.

But while the main purpose of the early Slavonic manuscripts was to reproduce the *meaning* of the text, the exact reproduction of the *form* and *pronunciation* of the sacred texts became more important now (at least with respect to Church writings) since the original language is not alive any more. As a result, artificial grammar rules and special diacritical signs appeared (which helps in pronouncing the Church texts *exactly* as they were pronounced several centuries ago). By comparison with the former period, writing rules are more or less formalized and it is strictly prohibited to change them.

This stage in the development of Old Slavonic and Church Slavonic writings was fixed in the middle of the 17th century when patriarch Nikon initiated the 'correction' (or, more correctly, new translation) of the sacred Church books. Starting from that moment the Church Slavonic writing system has been fixed up through the present, and the result is shown in [4, 5, 6].

While this is true for Orthodox Church writings, there was a small group of people who did not accept Nikon's reformations (so-called 'old believers') and continued to follow the former traditions. The main disagreement between these groups was in understanding the sacred texts and the ways in which the sacred ceremonies should be performed, but there is also some difference

somehow connected with the pronunciation, as in every writing system apart from Hebrew. Nobody knows the truth . . . ☺

in the Church writing system conserved up to now[4]—see section "Church Slavonic writing" on page 9 and also [6, 23].

Church Slavonic writing was definitely not suitable for civil purposes. For practical applications the ordinary (script) writing system was created step by step (middle 14th–15th centuries)—with simplified rules, useful abbreviations, round letter shapes far from those in Church books, etc. Although it originated in the Church Slavonic script, by the middle of the 17th century it was definitely a separate calligraphic art, and there are special textbooks dating from this period showing how to write correctly (although such writing was not used in typography). A well-known reformation of the alphabet by Peter the Great in 1707–1708 was based mainly on this *de facto* writing system.

In 1707–1708 the new official civil alphabet was introduced by the Russian tsar Peter I ([25, 26]). It differs strongly from the Church Slavonic writing and its appearance was affected by practical requirements: the reformation of the state required typographically printed textbooks (mathematics, mechanics, ballistics, engineering, geography, etc.), and the Church Slavonic system was definitely not suitable for that purpose. Peter the Great simplified the letter shapes making it closer to Latin, cancelled non-necessary and doubling letters, deleted artificial stresses and phonetical symbols, included the new letters Э/э and Я/я necessary for new sounds (and used *de facto* in handwritten scripts), and introduced arabic notation for numbers. It seems that the first variant (1707–1708) was much more radical with respect to the obsolete letters and only later, under pressure from the Orthodox Church, most of these letters were reinstated (1708–1710).

Slow evolution of the new Russian writing system[5] continued up to 1917–1918 when the next significant reform took place. Although performed in the early days of the new communist regime (the state laws introducing the new Russian alphabet were issued and signed by the officials on December 23, 1917, and October 10, 1918), this reformation was based on long-term work performed in 1904–1917 by the Academy of Sciences, and its main purpose was to simplify the orthography and to delete obsolete and unnecessary letters inherited from Church Slavonic writing. The present state of Russian grammar

[4] It is possible that other branches of the Orthodox Church also use some sub-dialects of the canonical writing—if you have some information about it please let me know.

[5] In about 1735, Й/й was accepted officially as a separate letter and the letters *ksi* (ξ), *psi* (ψ) and *izhitsa* (υ) were thrown away. In 1738 the letter *i* was substituted for *ï*. In 1758 *izhitsa* (υ) was reinstated. In 1797 N. M. Karamzin introduced the letter Ё/ё for the digraph *IO/io* used previously. Subsequent changes were negligible.

and its alphabet was fixed in 1956 (it did not introduce any changes to the alphabet, only improved the grammar rules). This alphabet is shown in figure 1, and it is necessary to emphasize that the letter Ё/ё is still there as a separate symbol (regardless of the fact that in printing it is often substituted by Е/е).

This does not mean that the history of Cyrillic is restricted to the Russian language only. After the wars with Turks at the end of the 19th century, Bulgaria became an independent state and reintroduced Cyrillic as its official writing system (some of its features were borrowed from the Russian one as it was at that moment, and in 1945 the Bulgarian writing system was updated by deleting 'big yus' and 'yat' and modifying the grammar). The same is true for the Serbian and Macedonian alphabets based on Cyrillic. Ukrainian, Byelorussian and Moldavian alphabets (before the latter was changed into the Romanian one) inherited most features from the civil Russian writing system, but now they are developing independently. The same is true for the Mongolian writing system based on Cyrillic and numerous languages of the national minorities of the Russian Federation and Former Soviet Union with the alphabets *based on* civil Cyrillic but with their own rules and specific features (some of these are reviewed in [11]).

Old Slavonic writing

Old Slavonic writing did not distinguish between uppercase and lowercase letters although the first letter in a chapter was usually drawn artistically and colored. The canonical alphabet is shown in [2, 3] although it is necessary to note that there are many variant graphical shapes and ligatures (for example, reversed *ie* and S-shaped *zelo*, the ligatures Н+Г and Л+Г, etc.) not shown there. (For historical reasons briefly explained in section "A brief history of Cyrillic" on page 1 there is some flexibility and violation of canonical standards in Old Slavonic writing.)

Here is a brief description of the characteristic features of Old Slavonic writing:

- Cyrillic originates in Byzantine writing and inherits its features and rules to a great extent. For example, Cyrillic sometimes keeps the Greek notations for corresponding sounds—for example, the sound [*u*] in Greek is written like *o*υ while single υ represents the sound [*ü*]. Although in Slavonic there was no sound like [*ü*], the sound [*u*] was represented as ѹ in Cyrillic.
- Some letters were conserved in Cyrillic to keep numerical notation in agreement with the Byzantine one. So, the letters ξ, ψ have no meaning in Cyrillic because there were no such sounds in any Slavonic language before

the 10th century, but they are kept in the Cyrillic alphabet to represent the numbers 60 and 700. (These letters were also used in Greek words inserted in native Cyrillic writings, but in many cases they were substituted by the pairs пс and кс even in words taken from Greek.)

- Some sounds were represented by two different letters, which also reflects the fact that Cyrillic was derived from the Byzantine (Greek) writing system. For example, the sound [*o*] may be written as ѡ (*omega* = *long o* in Greek) or as о (*omicron* = *short o* in Greek) while there was only one sound [*o*] in the Slavonic language.[6] In *ustav* writings (the most ancient Cyrillic manuscripts) *omega* (ѡ) was used mainly for numerical notation, and even in words taken from Greek the sound [*o*] was written as *o*. (But sometimes the usage of *omega* reflects the origin of the word.) Later, in *semi-ustav* writings *omega* is used more frequently, but it appears to be for decorative reasons only. Similarly, in *semi-ustav* writings the letter *round o* or *wide-o* appears for the same reason—sometimes it was used where the sound [*o*] requires the stess, but in most cases its usage is more or less arbitrary.

- Like the sound [*o*], the sound [*f*] was also represented by two letters: *fita* (ѳ) and *fert* (ф). *Fita* (ѳ) was used primarily for words taken from Greek where this letter was used following the rules of the Byzantine grammar. These two letters also have different numerical meanings—9 and 500 (see section “Numbering system” on Page 15).

- The sound [*i*] was also represented by two letters: *izhei* = Н and *izhe* = І. They have different numerical meanings: Н = 8, І = 10 (because of this feature they are sometimes called *octal-i* and *decimal-i*). The letter *I* was used relatively rarely, and mainly in cases where space is critical (for example, if there are two н one after another, the second one is written as і). In *semi-ustav* writings the letter І was used more frequently, and it became the tradition to put it after vowels. Sometimes the two-dot form of І can be seen in *semi-ustav* writings, but in ancient manuscripts this letter is used without any dot exclusively.

- Since Cyrillic inherits many features of the Byzantine writing system, the letter *az* is always written using the round shape (i. e., closer to lowercase Greek *alpha* than to capital latin A).

- The letter *shta* (now transformed into *shscha*) was written with the tail or descender below the middle stem, not as it is written currently (Щ). The

[6] Some authors [16] state that the difference in letters reflects the difference in pronunciation, but it seems to be rather questionable.

reason is that this letter is the ligature between the letter *sh* (Ш) and the letter *t* (Т).

- The letter *izhei* which corresponds to modern Cyrillic И (vowel [*i*]) was written as modern Cyrillic Н (consonant [*n*]), while the letter *nash* corresponding to the consonant [*n*] was written similar to the latin (Greek) N.
- In the same manner the letter *cy* representing the consonant [*c*] (letter Ц in modern Cyrillic) was written as Ч (in modern Cyrillic it represents the consonant [*ch*]) while the letter *cherv* representing in Old Slavonic the consonant [*ch*] was written in a manner similar to latin Y.
- The letter *short i* (Й, modern form Й/й) appeared in the 14th century, but was fixed as the canonical form only in the 17th century. (The 'old believer' Church Slavonic writings still do not use it in some positions, where it is required by the orthodox Church rules.)
- Some letters have more than one graphical shape in Old Slavonic manuscripts. The letter *izhitsa* corresponding to Greek υ (upsilon) has two graphical shapes: v-shaped and y-shaped letters. Similarly, the letter *uk* has two shapes: the ligature 'оу' and the γ-shaped letter (which is actually the vertical ligature constructed from the same letters). The letter *zelo* in Old Slavonic has two variant forms: *S-shaped* (sometimes with a tick in the middle) and *Z-with-tail*. We can see *wide-o*, *narrow-o* and *omega* in Old Slavonic writings representing the sound [*o*] where *wide-o* and *narrow-o* can be exchanged freely in writing. More recently in *semi-ustav* manuscripts the *wide-ie* and *narrow-ie* shapes appear corresponding to the same sound [*e*]. Such alternative shapes played mostly a decorative role although in some cases (especially in ancient manuscripts) they were used to economize space where it was critical. Later, the variant shapes got some orthographic meaning (which was fixed strictly in the 17th century after Nikon's Church reformation when the orthodox Church Slavonic writing formally appeared).
- As has already been mentioned, the letter *zelo* has two variant forms: *S-shaped* and *Z-with-tail*. Formerly this letter represented the phoneme [*dz´*], which evolved into soft [*z´*] and then disappeared by transforming into ordinary [*z*]. The letter *zemlya* similar to that in pronunciation is also written as *Z-with-tail*; the only difference between these two letters is in the size of the tail and in an optional tick for *zelo*. Since the letter *zemlya* slowly evolved to its modern shape (З) by increasing the tail and making the z-component smaller and higher, sometimes the difference between *zelo*

and *zemlya* can be established only through historical or grammatical context. Moreover, since the original meaning of *zelo* was lost and this sound became indistinguishable from *zemlya*, quite often Cyrillic manuscripts, except the ancient ones, use these letters in the wrong way.

- Originally the letters *er* (Ъ) and *erj* (Ь) represented short vowels (semi-vowels) similar to [*o*] and [*e*], respectively. But in time, these sounds disappeared from Slavonic language. As a result, in some cases (under stress) О substitutes Ъ and Е substitutes Ь, and in some other cases (without the stress) they just disappeared. (As a result, in modern Cyrillic these letters change their meaning totally—they are used to mark hard and soft pronunciation, and in Bulgarian the letter Ъ is used to represent a specific Bulgarian vowel which has no relation with the original Slavonic sound.)
- Similar to *er* and *erj*, the letters *small yus* and *big yus* represented specific Slavonic vowels close to [o^{η}] and [e^{η}]. It seems that the proper pronunciation of these sounds was lost by the 10th century since quite often these letters are used in the wrong way even in ancient manuscripts. In spite of this fact, grammatically correct usage of the letters *big yus* and *small yus* was kept until the 16th century. (In modern Church Slavonic *big yus* is substituted by У, *iotified big yus*—by Ю, *small yus* and *iotified small yus*—by *small yus* or *iotified az* defining the sound [*ya*], or by *az* after sibilants.)
- The sound [*e*] has two different pronunciations and in *ustav* manuscripts it was represented by two different symbols: the letter *ie* (Є) was pronounced as modern Russian Э, and the letter *iotified ie* was pronounced as modern Russian Е. *Iotified ie* was used after vowels, at the beginning of words, and in a few exceptions. In *semi-ustav* manuscripts there was no difference between hard [*e*] and soft [*e*]. Although we can see *wide-ie* and *narrow-ie* in these manuscripts, this is mainly decoration, not the requirements of grammar. (In spite of this, correct pronunciation of hard and soft [*e*] was conserved up to the 18th century, and the 'old believers' keep it even now.)
- Many Cyrillic letters were created as ligatures. These are: *uk* which is the combination of О and У, *shta* = the combination of Ш and Т, *ery* combining Ъ and І or Н. A special case of ligatures is the *iotified* letters which are the combination of the letter *izhe* (І) connected by a horizontal line with the following vowel. *Iotified az*, *iotified ie*, *iotified small yus*, *iotified big yus* are created in such a way. Surprisingly, the letter *yu* (Ю) is also the iotified form derived from the letter *uk* (the ligature ОЮ) by throwing away the second component У.

- The letter *ery* (Ы) was the ligature between the letter *er* (Ъ) and *izhe* (І) or *izhei* (Н = modern И). So, in ancient manuscripts it is written as ЪІ or as ЪН with a few exceptions, and only later in *semi-ustav* writings is it transformed into its modern form Ы (i. e., when the letters Ъ and Ь lost their original meanings).
- There is one more letter, *gherv*, in the alphabet shown in [2]. It is used only in modern scientific texts, although it was absent in native Old Slavonic writings. The reason it was introduced is very simple: it corresponds to the only letter in Glagolitic which has no analog in Cyrillic. (In Glagolitic it represents the sound which disappeared when Cyrillic appeared.) So it is used to represent the original Glagolitic writing by Cyrillic transcripts in *scientific* literature, and for nothing more.

Church Slavonic writing

The Orthodox Church Slavonic alphabet is shown in [4, 5]. There are the following differences between Church Slavonic and Old Slavonic (*ustav*) writings:[7]

- The order of the alphabet is changed, some letters changed their names, some letters changed their shape. Some letters became obsolete and are excluded from the alphabet, but as a compensation the new letter *ot* appears which is the ligature between O (Greek omega) and T with three stems. (In Old Slavonic the name *ot* was reserved for Greek ω which is called *omega* in Church Slavonic.)
- Uppercase and lowercase letters appear.
- The letter *ie* exists in two graphical forms (although it is included as a single letter in the alphabet): *wide-ie* is used at the beginning of words, *narrow-ie* is used in the middle and at the end of words. Additionally, letters *wide-ie* and *narrow-ie* are used to distinguish grammatical forms in foreign words (‘фарісей’ and ‘фарісэй’, for example). Letter *yat* is pronounced as [*e*] as well now and in this respect is undistinguishable from other *ie*. (Formerly

[7] As already mentioned in section “A brief history of Cyrillic” on page 1, there is no sharp boundary between Old Slavonic and Church Slavonic, but rather the smooth and continuous evolution of the common writing system. Since the rules and requirements of Church Slavonic are strictly specified, we can, under some (rather weak) assumptions, call all the features of Slavonic writing outside the canonical rules of Church Slavonic *Old Slavonic* even if such a classification is rather artificial and incorrect.

it was read as [ие] or [э] but nowadays this sound is lost even in Church Slavonic.)

- The letter *zelo* exists only as a single variant (S-shaped). In Old Slavonic this letter represented the phoneme [дз] which has now disappeared. In Church Slavonic it is pronounced similar to [з] and is used only for a limited and well-defined set of words. It also represents the digit '6' (see section "Numbering system" on page 15).
- The letter *zemlya* is modified—it is now mostly written as modern Cyrillic З.
- The letter named *izhei* is absent—the letter *izhe* (И/и) is used instead. The letters И/и are used before consonants. In all other cases the letter ї is used to represent the sound [*i*] (see below).
- The latin letter I/ї, which was called *izhe* in Old Slavonic, is now called *ii* and is written with two dots in the lowercase form (when it is used without diacritical accents, of course). Sometimes capital I is drawn with a bold dot in the middle of the stem, or with a calligraphic ring (hole). Letters I/ї are used before vowels (while И/и are used before consonants and before consonants in foreign (greek) words where they substitute for the greek letter ι and diphthongs *e*ι, *o*ι. The lowercase letter is used with two dots where this sound is pronounced without the stress, and without dots when the stress is placed explicitly or only implied (rare case).

  The lowercase letter і with one dot was used in civil Russian texts before 1918 (see section "A brief history of Cyrillic" on page 1). All these letters are pronounced like [и].
- Letter *on* is transformed into *wide-o* and *ordinary-o*. *Ordinary-o* is used in the middle and at the end of a word. *Wide-o* is used at the beginning of the word, inside some specific words and at the boundary inside a compound word. In addition to these two forms of *o*, there is *narrow-o* which is used only for the ligature 'оу' representing the *uk* letter at the beginning of words.
- The Greek letter ω is also pronounced [*o*] and exists in two variants: *calligraphic omega* with titlo and aspiration, and *ordinary omega*. The *calligraphic omega* is used to express an exclamation. *Ordinary omega* is used in prefixes and prepositions, to distinguish phonetically equivalent grammatical forms, and for words taken from Greek.

- The separate letter *ot* appears, which is the vertical ligature of letter 'o' and letter 't'. It is used in prefixes and prepositions and pronounced as [*ot*].
- The usage of *uk* letters now obeys strictly defined rules. The form ѹ is used at the beginning of words, the form ѵ—in the middle and at the end. It is necessary to note that *narrow-o*, not *ordinary-o*, is used for the ligature ѹ.
- The letters *c*, *ch* and *ery* changed their graphical shape and are written as in modern Cyrillic.
- The letter *shta* is called *shscha* now (although it is still written in the same manner).
- The letters *small yus* and *iotified az* changed their meaning. Now these letters represent the sound [*ya*]. The letter *iotified az* is used at the beginning of a word, the letter *small yus*—in the middle and at the end of a word.
- The letters *big yus*, *iotified big yus*, *iotified small yus* and *iotified ie* are marked as obsolete and are not included in the alphabet (although, for example, *big yus* is still used in the Church Slavonic calendar for special purposes). The letter *iotified small yus* is substituted by *iotified az* or *small yus*, the letter *iotified big yus*—by Ю/ю, the letter *big yus*—by *uk* (ѵ-shaped or ѹ-ligature), the letter *iotified ie*—by *wide-ie* or *narrow-ie*.
- The letter *izhitsa* is used in words taken from Greek and may be pronounced as [*v*] or [*i*]. When it is pronounced as [и], it has a diacritical sign above it (it may be stress, or aspiration, or reversed hungarian umlaut (double grave), or something else).
- The letters ѳ, ѯ and ѱ are used only for words taken from Greek.
- The numbering system is changed a little—see section "Numbering system" on page 15 for details.

The variant form of Church Slavonic writing is still used by 'old believers' (see [6]). It includes just the same 40 letters in a slightly different order and with a single exception—*big yus* is included while ѡ is superimposed with *o*, —but there are some differences in their graphical shape and usage as well. For example, the letter *zemlya* is written in most cases as *z-with-round-tail*, not as modern Cyrillic з (which is typical for orthodox Church Slavonic writings). The letters *wide-o*, *ordinary-*о and *omega* correspond to the same position of the alphabet (the letter *wide-o* is used as the capital (uppercase) letter, and *omega* is used as the lowercase letter only). Surprisingly, the uppercase form for the sound [я] is *iotified az*, and the lowercase form is *small yus*. Similarly, when

small yus (it conserves its special role in 'old believer' Church Slavonic) is used in text, its uppercase form is written as *small yus*, but its lowercase form is written as *iotified az*.

Diacritics and punctuation symbols

The following diacritics and punctuation symbols are used in Old Slavonic and Church Slavonic:

- Ordinary *titlo* was used for numbers (see section "Numbering system" on page 15) and to represent abbreviated words. (Most typical words have the standard abbreviations which enabled to economize expensive parchment used for writing.) *Semi-ustav* manuscripts use more abbreviations than the *ustav* manuscripts. The variety of graphical shapes used in old manuscripts for *titlo* is great, but logically all these shapes represent just one symbol. It is also necessary to note that the abbreviations used in Old Slavonic writings are quite different from the canonical abbreviation system used in modern Church Slavonic texts.
- In addition to ordinary *titlo*, there were so-called *titlo-in-letters* also used to mark abbreviations. While the ordinary *titlo* is just an empty square bracket placed horizontally over the abbreviated word, *titlo-in-letters* is the small (skipped) letter placed over the abbreviated word—typically it is a consonant—and marked by a curvilinear brace-type symbol placed horizontally. While in Church Slavonic only a limited set of letters can be used to construct the *titlo-in-letters*, in Old Slavonic nearly any letter could be used for this purpose.
- The intermediate variant is the case when the abbreviated letter is placed above the word without the special curved symbol. Such letters form a special system of diacritical signs since in general their shape is quite different from that of the letters used for ordinary text.
- When the letters *er* (Ъ) and *erj* (Ь) are skipped in an abbreviated word, it is substituted by a special diacritical sign *paerok* (*jerok*,*jerik*) placed above the word where the original letter *er* or *erj* is implied. (In Church Slavonic *paerok* is equivalent to *er* or *erj*—it obeys the same grammatical rules and is pronounced similarly. In Old Slavonic [16] *paerok* was also used to indicate the Greek ε.) Although we can find a variety of graphical shapes for that symbol in ancient manuscripts (for example, in the *Ostromirovo Evanghelie* [Ostromir' Gospel], a breve placed between two consonants is used), two

shapes are more or less canonical: *tilde* rotated by 90° and a little straight *integral*-style sign (logically, both symbols are exactly the same).

- In some cases a special diacritical sign (apostrophe or, alternatively, frown) is used to indicate soft consonants.
- To produce the *short i* (й), a special diacritical sign (breve or, alternatively, soft aspiration *dasia*) was placed above *izhei*: ѝ, ѝ.
- Since in most cases Old Slavonic texts are translations from Greek, quite often they inherited the same diacritical signs (although in the Slavonic language these symbols mean nothing). The most frequent are the *aspiration* signs—hard (*dasia* or *Spiritus asper*) and soft (*psili* or Spiritus lenis)—copied from Greek texts. Graphically aspiration symbols are similar to small open and close round braces or apostrophe and reversed apostrophe. (Following the example of the Unicode tables, they could be transformed into a breve-shaped sign as well.)
- Diacritical signs (aspiration, stress) when combined with *titlo* may be placed above or below this sign.
- Although there is *no* palatalization in the Old Slavonic language at all (no soft, nor hard) and although the Slavonic words are pronounced quite differently from those in Greek, palatalization signs of both types are placed (more or less randomly) in Old Slavonic manuscripts from the very beginning, and at the end of the 14th century (*semi-ustav* manuscripts) they started to play an orthographic role. A palatalization sign is placed not over the first vowel (as it is in Greek), but over each vowel without a preceding consonant as well. At the same time aspiration with stress appears (*apostróph*) and *paerok* between two vowels at the boundary between syllables becomes obligatory.
- The Old Slavonic writing system was continuous: words were not emphasized, capital letters were used only at the beginning of chapters but not at the beginning of sentences, and the end of a chapter was usually marked with a special sign (some combination of bars and dots—there is a great variety of these symbols in Slavonic manuscripts).
- There were no punctuation symbols in Old Slavonic (in the ordinary meaning) although some sentences or fragments of sentences may be separated by dots. In such a case the dots were placed vertically at the mid-height of the letters, not at the baseline.

The set of diacritical and punctuation signs in Church Slavonic is much larger. This is explained by the requirement to reproduce exactly not only the meaning, but also the pronunciation of old sacred texts written in a nearly dead language. (When the Old Slavonic language was alive, the correct pronunciation was implied *de facto*.) So, the following new diacritics and punctuation symbols appear in Church Slavonic:

- Three different stresses appear in Church Slavonic:
  - sharp stress (ѡ́)—*oxýa* (latin *acutas*),
  - heavy or blunt stress (ѡ̀)—*várya* (latin *gravis*),
  - clothed stress (ѡ̂)—*kamóra* (latin *circumflexus*?).
- Similar to Greek there is the aspiration sign (hard aspiration *dásia* or, as it is called in Church Slavonic, *zvátelstvo*). Soft aspiration *psili* is not used in Church Slavonic.
- Aspiration may be combined with sharp and blunt stresses. Aspiration with sharp stress (ѽ) is called *íso*, aspiration with blunt stress (ѿ) is called *apostróphe*. In orthodox Church Slavonic aspiration and stress are joined horizontally. In 'old believer' Church Slavonic *apostróphe* may be constructed as ѡ̀̂.
- A special diacritical sign called *okovy* (◌̏ or ◌̋) is placed over *izhitsa* when it should be read as [*i*], not as [*v*], and there is no other diacritical sign above it.
- *Paerok* in modern Church Slavonic substitutes *er* (ъ) only, not *erj* (ь). It is also used to mark the short pause at the boundary between the parts of a compound word or between prefix and root.
- To produce the *short i* (й) only *breve* may be used above the *izhe*: й.
- In Church Slavonic the abbreviation system based on *titlo*, *titlo-in-letters* and tiny letters placed above the word is much better standardized and formalized. In particular, there are only 5 *titlo-in-letter* combinations: with 'с' (*slovo-titlo*), with 'г' (*glagol-titlo*), with 'д' (*dobro-titlo*), with 'о' (*on-titlo*), and with 'р' (*rcy-titlo*). These *titlo-in-letter* symbols are so specifically drawn that they should definitely be considered separate glyphs.
- There is a special footnote symbol called *kavyka*. It is drawn like a *breve* after the end of the word—кавыка◌̆. The footnote is represented as a marginal note or, more conventionally, at the end of the page. (In modern Church

Slavonic the footnote marks are usually represented in a standard way by arabic numbers and the footnotes are at the end of the page or at the end of the whole text.)

- The following punctuation signs appear in Church Slavonic:
  - Ordinary dot—placed above the baseline at the middle of the ordinary letter height; it is heavier than *small dot* (see below).
  - Small dot—it is not so heavy as an ordinary dot, and is used to divide into parts long and compound sentences. The most significant difference is that the sentence after the small dot starts with a lowercase letter.
  - Comma (,).
  - Colon (:).
  - Semicolon—is substituted by small dot or colon.
  - Ellipsis—is substituted by colon.
  - Question mark—is drawn as semicolon (;).
  - Exclamation mark (!).

Numbering system

In Old Slavonic and Church Slavonic, numbers were written as letters with special marks. When some letter or combination of letters represents a number, it is surrounded by dots (centered with respect to its height), and the symbol *titlo* is centered above it ([7, 8]). (In Church Slavonic the dots surrounding the number are not necessary if it is evident from the context that this is a number, and *titlo* is placed above the rightmost letter.)

The number of letters in the alphabet is enough to represent units (1–9), tens (10–90) and hundreds (100–900) (see Table 1). The order of letters used for digital notation follows the Greek alphabet, not the Cyrillic. Some letters changed their numerical meaning:

- In ancient manuscripts Greek *koppa*[8] is used for 90 while the Cyrillic letter *chervj* (Ч or Y) is used for this purpose later.
- In Old Slavonic *izhitsa* is used, and in Church Slavonic *uk* (without preceding *o*) is used to represent 400.
- 800 is represented by *omega* in Old Slavonic and by *ot* (the vertical ligature of *omega* and *t*) in Church Slavonic.
- *Small yus* sometimes is used in Old Slavonic to represent 900 while only *cy* (Ц) is used for this purpose in Church Slavonic.

In addition, it is necessary to take into account that some letters changed their graphical shape: for example, *izhei* = 8 was drawn as H in Old Slavonic while it is drawn as И in Church Slavonic, *cherv* = 90 was drawn as Y in Old Slavonic while it is drawn as Ч in Church Slavonic, *cy* = 900 was drawn as Ч in Old Slavonic while it is drawn as Ц in Church Slavonic.

Thousands are preceded by a special *thousand sign* (for example,[9] ·Ā· = 1 → ҂Ā· = 1000, ·B̄· = 2 → ҂B̄· = 2000, ·Ḡ· = 3 → ҂Ḡ· = 3000, etc.).

Similar to current digital notation, tens are placed to the left of units, and thousands—to the left of tens when more than one digit was necessary (·ЛЄ̄· = 35, ·РКӢ· = 128, ҂АКД̄· = 1024). The exceptions are the numbers from 11 to 19 where units are placed first: ·АЇ· = 11, ·ВЇ· = 12, . . . , ·ѲЇ· = 19 (for example, 1111 = ҂АРАЇ, not 1111 = ҂АРІА·). Sometimes in Old Slavonic units and tens are typed separately: "·М̄· и ·Ḡ·" means 43 (i. e., "40 and 3").

To represent extra large numbers (more than 1.000.000) in Church Slavonic the *thousand sign* is repeated several times:

҂Ā· = 1000, ҂Ī· = 10.000, ҂P̄· = 100.000,

҂҂Ā· = 1.000.000, ҂҂Ī· = 10.000.000, ҂҂P̄· = 100.000.000,

҂҂҂Ā· = 1.000.000.000, ҂҂҂Ī· = 10.000.000.000, etc.

[8] The correct Latin name of this Greek letter is *qoppa* but in Unicode tables it is named as *koppa* which is closer to its 'Russian' name. Sometimes in Old Slavonic and Church Slavonic texts it is mixed up with *stigma*—another obsolete Greek letter used in Old Greek to define the number 6 [14, 15, 17], and sometimes *stigma* is used independently for some Church holidays. Moreover, in Greek script *qoppa* exists in two variant forms [14, 15, 13]. But the discussion of such details and the 'white noise effects' in transferring the typing traditions from Greek to Cyrillic is surely outside the scope of this paper.

[9] Here *titlo* is substituted by *macron* and *thousand sign* is substituted by *tick*.

In Old Slavonic such big numbers were the extremely rare exceptions, and for this reason they are decorated differently and have special names:

- *T'ma* = 10.000—letter А inside a circle,
- *Legion* or *nesved'* =100.000—letter А inside a circle constructed from 8 dots,
- *Leodr* = 1.000.000—letter А inside a circle constructed from 8 commas with tails oriented outside the circle,
- *Vran* = 10.000.000—letter А inside a circle constructed from 8 crosses,
- *Koloda* = 100.000.000—letter А between two arcs: *breve* above and *frown* below,
- *T'ma tem* = 1.000.000.000—letter Ы inside a circle constructed from 7 minuses and one plus placed above the letter.

In modern Church Slavonic such notations are obsolete while the names like *t'ma* and *legion* are still in use.

The Unicode Cyrillic page

Now we can check how well the Unicode Cyrillic range `04xx` (in its current state) suits the purpose of representing the Old Slavonic and Church Slavonic writings.

Taking into account the preceding sections, it can be seen that the Unicode Cyrillic page `04xx` (as it concerns Old Slavonic and Church Slavonic) contains some mixture of glyphs from these writing system but not *all* necessary glyphs. Here is an analysis of what is present and what is absent:

- It contains Greek *koppa*, used in Old Slavonic to represent the number 90 (uppercase and lowercase). There is also a proposal by Michael Everson ([10]) to include the (currently obsolete) symbols for 10.000, 100.000 and 1.000.000 in positions `0487`, `0488` and `0489`.[10] But other old numerical notation symbols (10.000.000, 100.000.000 and 1.000.000.000) are not even considered.

[10] The symbols for 100.000 and 1.000.000 will be included in Unicode version 3, in positions `0488` and `0489`; it may be intended that the combining circle at position `20DD` is to be used as the symbol for 10.000 [14, 1].

- It contains two aspiration symbols *dasia* and *psili* as taken from Greek in Old Slavonic manuscripts. But it does not contain the combinations of *dasia* (the only aspiration symbol used in Church Slavonic) with the stresses ´ and `. (The stresses themselves can be taken from the Unicode page *Combining Diacritical Marks*.)
- There is only one *titlo* (0483) while there are two symbols *titlo* in Old Slavonic and Church Slavonic: *ordinary titlo* and *titlo-in-letters*, and there are precise grammatical rules specifying when each symbol should be used. Moreover, in Church Slavonic there is the well-defined set of letters which can be used together with *titlo-in-letters*, and it seems that all such combinations should be included as separate symbols (graphically they are quite different from the tiny letters placed above the word and under the *titlo-in-letters*).
- It does not contain the diacritical sign *paerok*.
- It does not contain the letter *iotified az* used in Old Slavonic and Church Slavonic. The reason may be that this letter is used in parallel with *small yus* to represent the sound [*ya*]—so, perhaps it could be considered as the variant form for *small yus*? But in Old Slavonic it definitely is a separate letter, and even if in Church Slavonic the letter *ya* has two graphical shapes—*small yus* and *iotified az*—they should *both* be included in the Unicode tables as is done with *narrow-o* and *wide-o* (041E/043E and 047A/047B) or with *omega* and *calligraphic omega* (0460/0461 and 047C/047D).
- It does not contain the Cyrillic analog of the letter *gherv* used only in Glagolitic, but which could be encountered in scientific publications where Glagolitic manuscripts are reproduced in Cyrillic.
- It does not contain the letter *zelo*. With some degree of imagination 'zelo' could be identified with *dze* (0405 and 0455 in Unicode). But even in this case the capital *zelo* should contain the tick or thick dot in its middle part and like other accented letters and letters with modifiers should occupy a separate cell in the Unicode table. (It is necessary to note that such letters as *barred-o* (04E8 and 04E9) and *fita* (0472 and 0473), *izhitsa* (0474 and 0475) and *accented izhitsa* (0476 and 0477) *are* included in Unicode as separate symbols.)
- The variant shapes of letters *zemlya* and *zelo* (Z-with-tail and Z-with-descender) are not included in Unicode since it is the principal proposition of the Unicode Consortium "not to include the variant glyphs" (unfortunately, quite often violated when Latin-based writing systems are considered). The same is true for the letter *nash* which is drawn in Old Slavonic

and Church Slavonic as the intermediate form between 'H' and 'N'. The letter *i without dots* is also not included for the same reason. The letter *shta* (ш with the descender below the middle stem), later transformed into Cyrillic *shscha* (щ, 0429 and 0449), is absent as well. The alternative shapes for *ery* (042B/044B)—ЬН, ЪI, ЪН—are absent.

(This is not a defect of the Unicode tables, which definitely should standardize *symbols*, not *glyphs*. But this feature complicates the creation of standard fonts used to reproduce old texts and the proper encoding of these texts as well.)

- Letter *uk* is included as the ligature 'o+y' only. The alternative graphical shape (γ-shaped *uk*) is not included. (If it is superimposed with the Cyrillic letter 'Ү/y', it seems rather strange.) Similarly, the letter *ie* exists in Church Slavonic in *two* graphical shapes—*wide-ie* (similar to ѣ) and *narrow-ie* (similar to ѣ). With some imagination *wide-ie* can be substituted by Є/є 0404 and 0454), and *narrow-ie*—by Е/е (0415 and 0435), but such substitution does not reflect the encoding markup.

  (The fact that in Church Slavonic the letters *uk* and *ie* have two graphical shapes, and there are strict grammatical rules for when each shape should be used, is not taken into account by the Unicode tables. But exactly the same situation is true for the letters *omega* and *calligraphic omega* (0460/0461 and 047C/047D) and *ordinary-o* and *wide-o* (041F/043E—here the letters have two graphical shapes as well, and *both* shapes are included in the Unicode table!)

The project of T2D encoding

Recent work on standardizing Cyrillic as LaTeX 2ε encodings [12, 11] increases the compatibility between Unicode and LaTeX. Already existing encodings T2A, T2B and T2C (see section "Conclusion" on page 23) cover all existing Cyrillic alphabets, except the accented characters. To achieve full compatibility and to add into LaTeX the Old Slavonic and Church Slavonic letters, T2D encoding is suggested.

First of all, it is necessary to emphasize that T2D *is not* intended for the exact reproduction of Old Slavonic and Church Slavonic texts. Its main aim is:

- to reproduce adequately Russian texts in the orthography used before 1918 and in emigrant literature until the 1970s,

- to reproduce adequately Bulgarian texts as they appeared before 1945 (when the Bulgarian writing system was reformed),
- to include into a main document fragments and citations from Old Slavonic manuscripts and Church Slavonic writings in stylized form—i. e., by keeping their general features but without exact and adequate reproduction of their *graphics*,
- to achieve full compatibility between the Unicode tables and the standard Cyrillic encodings used in LaTeX 2ε.

That is, T2D is intended mainly for scientific texts, and even for popular literature, more than for serious and deep investigations. It cannot be used for exact reproduction of Old Slavonic and Church Slavonic texts, but it should be suited to include into an ordinary text citations and bibliographic references in such a way that they do not disturb the flow of the modern text and simultaneously adhere to the main rules of Old and Church Slavonic writings.

Extraction of the out-of-date Russian and Bulgarian writings into T2D helps to cancel the ambiguity existing in T2C where a single glyph could represent *two* letters—i. e., the letters which are similar graphically but different logically (*semisoft sign* and *yat*, *o-barred* and *fita*). As a result the encoding T2C was modified slightly (see its current state in section "Conclusion" on page 23).

The current variant of the T2D encoding is shown in tables 2 and 3. It was constructed by keeping the common parts of T2A/T2B/T2C with the Russian alphabet (necessary for out-of-date Russian and Bulgarian texts), accents and ASCII letters and symbols, adding the glyphs used in Russian before 1918 and Bulgarian before 1945, adding the Old Slavonic letters and symbols from the Unicode encoding table `04xx`. Since the set of symbols currently included in Unicode is not enough to reproduce Old Slavonic and Church Slavonic texts (see section "The Unicode Cyrillic page" on page 17), the most significant symbols were added.

Some letters were included twice since their graphical shape is quite different in Church texts and civil texts (the variant 'old' and 'new' shapes are essential, at least for the most important letters—otherwise there is some visual discomfort when reading old citations typed in modern-style letters[11]). Some variant glyphs for the same letter (like `\phi` and `\varphi` in mathematics) are included

[11] In general it is more correct to solve this problem through use of special 'old-style' font families. But taking into account the enormous number of fonts required by the EC-font convention, it appears that it is easier to include the most different letters in two shapes into the same font than to create a special set of fonts which differ in only a few letter shapes.

for the most essential variants as well. Since the diacritics in Church Slavonic are richer than those in modern Cyrillic, it was necessary to delete ff-ligatures and to add specific diacritical signs. The most serious disadvantage of T2D is the absence of *titlo-in-letters* symbols, but there is definitely no space for them in T2D which should follow the general LaTeX 2ε rules (*titlo-in-letters* should be constructed from the *round titlo* and a tiny ordinary letter glued with it).

As a result we get a set of glyphs which is sufficient for reproducing the 'visually-logical' structure of Old Slavonic and Church Slavonic texts using a modern font family. It is necessary to emphasize once again, that T2D solves the problem of representation for Old Slavonic and Church Slavonic texts only approximately and under some assumptions about the simplification of their original structure. For example, serious scientific texts on paleographics can contain such enormous numbers of variant glyphs that Omega with its 65.538 symbols may be necessary (although it seems that graphical illustrations may be a better tool for the adequate reproduction of ancient texts in this case). T2D is definitely not suited for such tasks—it just enlarges the set of Unicode characters to the minimal envelope sufficient to type Cyrillic texts following the general Old Slavonic and Church Slavonic rules.

It can be seen that the adequate reproduction of Old Slavonic and Church Slavonic texts requires many more glyphs than could be placed in a single encoding which follows the severe rules of LaTeX 2ε (ASCII latin symbols in 32–127, just the same pairs of uppercase and lowercase letters as in T1, etc.). Some special encoding X⟨*n*⟩ is necessary to solve this problem. Due to the enormous number of variant shapes used in Old Slavonic and Church Slavonic it is reasonable to divide it into three parts of 256 characters each:

- Cyrillic letters, numbers, general punctuation and diacritical signs (aspirations, stresses, etc.).
- Accent-like symbols (first of all—titlos in letters), specialized and exotic diacritical signs, old-style numbering symbols, decorative symbols (asterisks of different type), etc.
- Glagolitics (including the variant Glagolitic symbols).

Such a structure makes it possible to fit all the necessary glyphs and even to leave some space for future upgrades if more exotic symbols/letters/ligatures are discovered in Old Cyrillic.

Although we tried to keep T2D as close as possible to the other T2∗-encodings, some symbols in T2D are different from those in T2A/T2B/T2C:

"0B = ordinary titlo (was: cedilla),
"0C = titlo in letters (was: ogónek),
"0D = paerok (was: palochka),
"17 = clothed stress *kamóra* (was: compound word mark),
"1B = Old-Slavonic aspiration *psíli* (was: ff-ligature),
"1C = Old-Slavonic aspiration *dásia* (was: fi-ligature),
"1D = Church-Slavonic aspiration *zvátelstvo* (was: fl-ligature),
"1E = apostróphe—aspiration with *várya* (was: ffi-ligature),
"1F = íso—aspiration with *oxýa* (was: ffl-ligature),
"9E = thousand sign (was: currency sign).

Latin 'I' (with a single dot) as used in Russian before 1918 is taken from the ASCII part of the table. 'I' with two dots is included as the ordinary letter, and letter 'I' without any dot is included as well. The latter is extremely useful for the variant shape of *ery* (ЪI) and for adequate reproduction of Old Slavonic texts. Similarly, the variant shapes for *ery* ЪН and ЬН should be composed from two letters as well when they are necessary. (Another reason to separate 'I without dot' from Latin 'I' is that quite often it is drawn with a bullet or a circle in the middle of the main stem—i. e., not as latin 'I'.)

Some variant shapes and specific ligatures are not included (*zelo* as Z-stroked-with-tail, *uk* as the bull head 'ȣ', y-shaped *izhitsa*, mirrored S-shaped *zelo*, *t* with three stems, Г-shaped *t* and mirrored Г-shaped *t*, the ligatures ЛГ, МГ, НГ, ЛЮ, МО, etc.) since there is no space for them. Although they may be important for some specific applications, the majority of Old Slavonic and Church Slavonic citations can survive without them. You should wait for X⟨*n*⟩ encoding which supports Old Slavonic and Church Slavonic better ☺.

Some accented letters (for example, *uk*, *yat*, *omega*, *izhitsa* with aspiration and/or stresses) are usually drawn as separate glyphs because it is difficult to compose them beautifully from the standard pieces. Such accented forms are absent in T2D (no space, no space ... ☹), and special macros should be created to compose them more or less artistically.

The letters *yat*, *hard sign*, *soft sign* and *ery* with extremely high stems are considered to be artistic (i. e., font family specific) shapes and are not even considered as candidates for T2D encoding. Nevertheless, it is very promising to make special font families for T2D encoding which simulate the shapes used in the 18th century and *ustav*/*semi-ustav* writings.

Titlo-in-letters diacritical signs should be composed from the *titlo-in-letters* glyph and the corresponding letter set at `\tiny` font size. Similarly, the letters placed as diacritical signs above the abbreviated words are absent in T2D

and are taken from the \tiny font. (As a result, such diacritical signs are unavailable for extremely small font sizes.) There should be special macros for diacritical signs of both types in the macro package supporting the T2D encoding.

Combinations of aspiration with stresses are included as separate accents. In this way they may be placed above letters by the standard TEX tools without artificial hacks and, which is more important, in this case they may be drawn more elegantly than is possible through the 'brute force' composition of two boxes. *Kamora* is included as separate symbol to distinguish it from *frown* used sometimes in Old Slavonic (it is also different graphically). *Kavyka* and *breve* for *short i* (й) are overlapped with the ordinary *breve* inside the T2D encoding. (But they may be separated *logically* by macros with different names inside the macro package.)

The digraphs 'оу' and 'ѿ' are not included into T2D although the separation of *uk* (оу) into its own cell was very attractive. One of the arguments was that we need *three* cells for such digraphs—lowercase (оу), uppercase (ОУ) and title (Оу). Since neither T2 nor T2∗ supports the title forms, these digraphs are not included in T2D either.

Obsolete Old Slavonic digital notation (see section "Numbering system" on page 15) is not supported by T2D. The difference between orthodox Church Slavonic and old-believer Church Slavonic should be maintained (if necessary) by the macro package or, preferably, by the User. Similarly, the historical changes in letter shapes (*cy*: Ч→Ц, *n*: N→Н, etc.) should be supported by macros or by the User. The centered dot is created from the ordinary period by a special macro as well.

Conclusion

Let us summarize the current state of Cyrillic in LATEX. Now LATEX supports the following encodings:

- T2A supports the languages

  Abaza, Avar, Agul, Adyghei, Azerbaidzan, Altai, Balkar, Bashkir, Belorussian, Bulgarian, Buryat, Gagauz, Dargin, Dungan, Ingush, Kabardino-Cherkess, Kazah, Kalmyk, Karakalpak, Karachaevskii, Karelian, Kirgiz, Kumyk, Komi-Zyrian, Komi-Permyak, Lak, Lezgin, Macedonian, Mari-Mountain, Mari-Valley, Moldavian, Mongolian, Mordvin-Moksha, Mordvin-Erzya, Nogai, Oroch, Osetin, Rus-

sian, Rutul, Serbian, Tabasaran, Tadjik, Tatar, Tati, Teleut, Tofalar, Tuva, Turkmen, Udmurt, Uzbek, Ukrainian, Hanty-Obskii, Hanty-Surgut, Gipsi, Chechen, Chuvash, Crimean Tatar

and consists of the symbols shown in the figures 2, 3, 4, 7.

- T2B supports the languages

  Abaza, Avar, Agul, Adyghei, Aleut, Altai, Balkar, Belorussian, Bulgarian, Buryat, Gagauz, Dargin, Dolgan, Dungan, Ingush, Itelmen, Kabardino-Cherkess, Kalmyk, Karakalpak, Karachaevskii, Karelian, Ketskii, Kirgiz, Komi-Zyrian, Komi-Permyak, Koryak, Kumyk, Kurdian, Lak, Lezgin, Mansi, Mari-Valley, Moldavian, Mongolian, Mordvin-Moksha, Mordvin-Erzya, Nanai, Nganasan, Negidal, Nenets, Nivh, Nogai, Oroch, Russian, Rutul, Selkup, Tabasaran, Tadjik, Tatar, Tati, Teleut, Tofalar, Tuva, Turkmen, Udyghei, Uigur, Ulch, Khakass, Hanty-Vahovskii, Hanty-Kazymskii, Hanty-Obskii, Hanty-Surgut, Hanty-Shurysharskii, Gipsi, Chechen, Chukcha, Shor, Evenk, Even, Enets, Eskimo, Yukagir, Crimean Tatar, Yakut

  and consists of the symbols shown in the figures 2, 3, 5, 7.

- T2C supports the languages

  Abkhazian, Bulgarian, Gagauz, Karelian, Komi-Zyrian, Komi-Permyak, Kumyk, Mansi, Moldavian, Mordvin-Moksha, Mordvin-Erzya, Nanai, Orok (Uilta), Negidal, Nogai, Oroch, Russian, Saam (Sàmi, Lappish), Tati, Teleut, Hanty-Obskii, Hanty-Surgut, Evenk, Crimean Tatar

  and consists of the symbols shown in the figures 2, 3, 6, 7.

- T2D supports the languages

  Old Russian (before 1918), Old Bulgarian (before 1945), Old Slavonic and Church Slavonic

  and consists of the symbols shown in the tables 2 and 3, figures 3 and 7.

(For some technical reason the tables in [12] were typed incorrectly, and that's why they are reproduced here again.)

The encoding T2C was modified a little before December 1998 to separate the modern writing systems and the out-of-date writing systems (and to add a new Samí letter as well). The current state of the T2D encoding is β-level, but it seems that it is close to the final state. The encoding X2—'Cyrillic glyph

container'—may need to be revised in the future because its current state is not in agreement with T2A/T2B/T2C. Future projects include the development of the X⟨*n*⟩ encoding to support Old Slavonic and Church Slavonic adequately.

As compared with Unicode, these LaTeX encodings are more complete than the current state of the `04xx` Cyrillic segment (even after the new additions proposed recently). There are some modern languages which are not supported by the Unicode tables, and there are also variant symbols and letters which are not considered by Unicode as separate ones—all these symbols and letters are available in LaTeX now. The current project of the T2D encoding also contains some elements absent from `04xx`—these are some diacritical marks, variant forms for some Cyrillic letters which are *not* the variant ones considered in Old Slavonic and Church Slavonic, and the letter 'Iotified A' which was skipped (or missed) by Unicode for some unknown reason.

It may be a good idea to achieve better agreement between the LaTeX encodings and the Unicode encodings.[12] Unfortunately my own efforts in this direction were not too successful—partly due to my own fault and partly due to the unchangeable belief of the Unicode Team that they know Cyrillic better than the native users ☺. May be somebody more lucky and more vigorous could do it in the future—who knows?

Acknowledgements

The results represented in this paper were discussed and criticized in the `PVT-TEX` mailing list organized by Vladimir Volovich (the list has private and local status, so do not ask me how to subscribe to it), and I would like to express my warmest thanks to Mikhail Grinchuk, Wladislaw Tchernow, Andrew Slepuhin, Dmitry Dmytriev, Vladimir Volovich, Valentin Zaitsev, Dmitry Smirnov, Andrew Janishewsky, Mikhail Kolodin and Alexey Burykin for their valuable help.

I would like to thank also Valeriy Ushakov (`uwe@ptc.spbu.edu`) who introduced our group to Dr. Richard McGowan and Dr. Kenneth Whistler from the Unicode Consortium (special thanks to them as well), and to Dr. Michael Everson, the Unicode expert on Cyrillic, for valuable and friendly email communications, the interesting data represented on his WWW page [9], his un-

[12] There is also an additional big gap in both the LaTeX and Unicode tables—namely, the letters (currently obsolete) which were invented and used for some time for minor peoples (mainly for North Region) in the 1920s in Russia. But it is a separate and difficult project to collect and to classify all these glyphs since the original sources are nearly unavailable.

believable activity in linguistics, fonts and writing systems, fighting for the linguistic rights of the minor nations, etc., and for his vigorous leadership in the ISSS *Alpha Project* [10] as well.

Great thanks to Barbara Beeton, Jörg Knappen and Nikolai Serikoff for their valuable remarks on the draft version of this paper. Special thanks to Barbara Beeton for her patient correction of my 'Penguin English' into the normal one ☺ and to Philip Taylor for the well-organized work on submitting, reviewing and representing onto the Internet the 'paper-less' EuroTEX'99 papers (including this one).

This work was partially supported by a grant from the Dutch Organization for Scientific Research (NWO grant No 07–30–007).

References

[1] Unicode Home Page:
http://www.unicode.org/

[2] Old Slavonic alphabet (Glagolitic and Cyrillic) as scanned from [19]:
http://members.xoom.com/ianin/oldslav/oldcyr1.gif

[3] Old Slavonic alphabet (Cyrillic only) as scanned from [20]:
http://members.xoom.com/ianin/oldslav/oldcyr2.gif

[4] Church Slavonic alphabet as scanned from [20]:
http://members.xoom.com/ianin/oldslav/church1.gif

[5] Church Slavonic alphabet as scanned from [21]:
http://members.xoom.com/ianin/oldslav/church2.gif

[6] Church Slavonic alphabet as scanned from [23]:
http://members.xoom.com/ianin/oldslav/church3a.gif,
http://members.xoom.com/ianin/oldslav/church3b.gif,
http://members.xoom.com/ianin/oldslav/church3c.gif

[7] Church Slavonic numbering system as scanned from [20]:
http://members.xoom.com/ianin/oldslav/number1.gif

[8] Church Slavonic numbering system scanned from [23]:
http://members.xoom.com/ianin/oldslav/number2a.gif,
http://members.xoom.com/ianin/oldslav/number2b.gif

[9] Michael Everson's Home Page:
http://www.indigo.ie/egt/

[10] ISSS *Alpha Project* Home Page:
http://www.stri.is/ISSS-WS/Alpha/

[11] A. Berdnikov, O. Lapko, M. Kolodin, A. Janishevsky, A. Burykin: "*Alphabets necessary for various Cyrillic writing systems (towards X2 and T2 encodings)*", in Proceedings of *EuroTEX-98*, Saint-Malo, 1998.

[12] A. Berdnikov, O. Lapko, M. Kolodin, A. Janishevsky, A. Burykin: "*Cyrillic encodings for LATEX 2ε multilanguage documents*", in Proceedings of the 19th Annual TEX Users Group Meeting. August 17–20, 1998, Toruń, Poland.

[13] Yannis Haralambous: "*From Unicode to Typography, a Case Study: the Greek Script*", 14th International Unicode Conference, Boston, 1999,
http://genepi.louis-jean.com/omega/boston99.pdf

[14] Barbara Beeton: private communication.

[15] Jörg Knappen: private communication.

[16] Nikolai Serikoff: private communication.

[17] Большой энциклопедический словарь "Языкознание". Главный редактор В. Н. Ярцева. Научное издательство "Большая Российская энциклопедия", Москва, 1998.

[18] Е. Ф. Карский: "Славянская кирилловская палеография", Москва, Наука, 1979.

[19] Н. М. Елкина: "Старославянский язык", Москва, Государственное учебно-педагогическое издательство Министерства просвещения РСФСР, 1960.

[20] Иеромонах Алипий (Гаманович): "Грамматика Церковно-Славянского языка", ОАО "Санкт-Петербургская типография N 6", 1997. (репринтное воспроизведение издания 1964 года).

[21] Д. Тихомиров, Е. Тихомирова: "Букварь", Редакция "Воскресение", Москва, 1991.

[22] Азбука церковно-славянского языка. Программа "Обновление гуманитарного образования в России", Москва, Интерпракс, 1995.

[23] Церковно-Славянская азбука. Изд-во "Церковь", Москва, 1994 ("лето от Адама 7502"), Издание Старообрядческой Митрополии Московской и всея Руси.

[24] Н. Л. Сукачев: “Экскурсы в историю письма (знак и значение)”, Санкт-Петербург, 1998, 140 с.

[25] А. Г. Шицгал: “Русский типографский шрифт. Вопросы истории и практика применения”, “Книга”, Москва, 1974.

[26] А. Г. Шицгал: “Русский гражданский шрифт (1708–1958)”, Москва, 1959.

[27] В. Н. Соловьев: Русское правописание. Орфографический справочник: словарь, комментарий, правила, 2-е издание, исправленное и дополненное, Санкт-Петербург, 1997.

Addresses

Alexander Berdnikov
Institute of Analytical Instrumentation
St. Petersburg
Russia
E-Mail: berd@ianin.spb.su

Olga Lapko
Mir Publishers
Moscow
Russia
E-Mail: olga@mir.msk.su

1	А	az	10	І	izhe (Old) i (Ch)	100	Р	rcy
2	В	vedi	20	К	kako	200	С	slovo
3	Г	glagoli	30	Л	lyudi	300	Т	tverdo
4	Д	dobro	40	М	myslite	400	ѵ	izhitsa (Old)
							ѹ	uk (Old)
							у	uk (Ch)
5	Е	jest′	50	N	nash	500	Ф	fert
6	S	zelo	60	ѯ	ksi	600	Х	her
7	З	zemlya	70	О	on	700	ѱ	psi
8	И	izhe (Ch)	80	П	pokoi	800	ѡ	ot (Old)
	Н	izhei (Old)					ѿ	ot (Ch)
9	ѳ	fita	90	Ү	cherv (Old)	900	Ч	cy (OLd)
				Ч	cherv (Ch)		Ц	cy (Ch)

Table 1: Numbers in Old Slavonic and Church Slavonic. (In Old Slavonic uppercase Greek *koppa* for 90 and *small yus* for 900 can be used as well.)

Аа, Бб, Вв, Гг, Дд, Ее, Ёё, Жж, Зз, Ии, Йй, Кк, Лл, Мм, Нн, Оо, Пп, Рр, Сс, Тт, Уу, Фф, Хх, Цц, Чч, Шш, Щщ, Ъъ, Ыы, Ьь, Ээ, Юю, Яя

Figure 1: Modern Russian alphabet

	x0/x8	x1/x9	x2/xA	x3/xB	x4/xC	x5/xD	x6/xE	x7/xF
0x	̀	́	̂	̃	̈	̋	̊	̌
	̑	̄	̇	̡	̧	I	⟨	⟩
1x	“	”	̆	̛	̨	–	—	
	0	1	J	ff	fi	fl	ffi	ffl

Figure 2: Accent part for T2A/T2B/T2C

"00	grave (blunt stress *várya*)	"10	double open quote
"01	acute (sharp stress *oxýa*)	"11	double close quote
"02	circumflex	"12	frown
"03	tilde (variant *okóvy*)	"13	double grave (variant *okóvy*)
"04	umlaut (variant *okóvy*)	"14	Cyrillic breve
"05	double acute (variant *okóvy*)	"15	endash
"06	circle	"16	emdash
"07	hachek	"17	∗ clothed stress *kamóra*
"08	breve (variant Cyrillic breve)	"18	percentage zero
"09	macron	"19	dotless-i
"0A	dot	"1A	dotless-j
"0B	∗ ordinary titlo	"1B	∗ Old-Sl. aspiration *psíli*
"0C	∗ titlo in letters	"1C	∗ Old-Sl. aspiration *dásia*
"0D	∗ paerok	"1D	∗ Ch.-Sl. aspiration (*zvátelstvo*)
"0E	left angle brace	"1E	∗ apostróphe (asp. + *várya*)
"0F	right angle brace	"1F	∗ íso (asp. + *oxýa*)

Table 2: T2D encoding—accent part (characters 0–127). ASCII letters and symbols are placed in "20–"7F as done for all T2∗-encodings.

	x0/x8	x1/x9	x2/xA	x3/xB	x4/xC	x5/xD	x6/xE	x7/xF
2x	␣	!	"	#	$	%	&	’
	(	)	*	+	,	-	.	/
3x	0	1	2	3	4	5	6	7
	8	9	:	;	<	=	>	?
4x	@	A	B	C	D	E	F	G
	H	I	J	K	L	M	N	O
5x	P	Q	R	S	T	U	V	W
	X	Y	Z	[	\	]	^	_
6x	‘	a	b	c	d	e	f	g
	h	i	j	k	l	m	n	o
7x	p	q	r	s	t	u	v	w
	x	y	z	{	|	}	~	-

Figure 3: ASCII part for T2A/T2B/T2C/T2D

"80/"A0	az (α-shaped)	"90/"B0	fita (θ)
"81/"A1	iotified az	"91/"B1	koppa (also called stigma)
"82/"A2	small yus	"92/"B2	uk (γ-shaped)
"83/"A3	iotified small yus	"93/"B3	yat
"84/"A4	ghervj	"94/"B4	cherv (Y-shaped)
"85/"A5	nash (N-shaped)	"95/"B5	psi (ψ)
"86/"A6	zemlya (Z-shaped)	"96/"B6	shta
"87/"A7	ksi (ξ)	"97/"B7	big yus
"88/"A8	i with two dots	"98/"B8	iotified big yus
"89/"A9	i without dots	"99/"B9	wide ie (ε-shaped)
"8A/"AA	izhitsa (υ-shaped)	"9A/"BA	iotified wide ie
"8B/"AB	omega (ω)	"9B/"BB	narrow ie (ε-shaped)
"8C/"AC	wide o	"9C/"BC	Russian yo
"8D/"AD	ot (ѿ)		
"8E/"AE	calligraphic omega		
"8F/"AF	zelo (S-shaped)		
"9D	numero sign	"BD	double basequote
"9E	∗ thousand sign	"BE	double guillemet left
"9F	section sign	"BF	double guillemet right

Table 3: T2D encoding—letters and symbols (characters 128–255). Russian uppercase letters are placed at "C0–"DF, Russian lowercase letters are placed at "E0–"FF as done for all T2∗-encodings.

	x0/x8	x1/x9	x2/xA	x3/xB	x4/xC	x5/xD	x6/xE	x7/xF
8x	Ґ	Ғ	Ҕ	Ҭ	Һ	Ӂ	Ҙ	Ӆ
	Ї	Ҟ	Ҡ	Ӄ	Ӕ	Ҥ	Ҥ	Ѕ
9x	Ө	Ҫ	Ӱ	Ү	Ұ	Ҳ	Ҷ	Ҷ
	Ҷ	Ҽ	Ә	Ӊ	Ё	№	¤	§
Ax	ґ	ғ	ҕ	ҭ	һ	ӂ	ҙ	ӆ
	ї	ҟ	ҡ	ӄ	æ	ҥ	ҥ	ѕ
Bx	ө	ҫ	ӱ	ү	ұ	ҳ	ҷ	ҷ
	ҷ	ҽ	ә	ӊ	ё	„	«	»

Figure 4: Specific Cyrillic letters for T2A

	x0/x8	x1/x9	x2/xA	x3/xB	x4/xC	x5/xD	x6/xE	x7/xF
8x	Ҕ	Ғ	Ґ	Ҕ	Һ	Ӂ	ҕ	Ӡ
	Ҍ	Ҟ	Ҡ	Ҟ	Ҩ	Ҥ	Ҧ	Ҋ
9x	Ө	Ҫ	Ӱ	Ү	Ҳ	Ҳ	Ҳ	Ҷ
	Ҷ	Ҥ	Ә	Ҽ	Ё	№	¤	§
Ax	ҕ	ғ	ґ	ҕ	h	ӂ	ҕ	ӡ
	ҍ	ҟ	ҡ	ҟ	ҩ	ҥ	ҧ	ҋ
Bx	ө	ҫ	ӱ	ү	ҳ	ҳ	ҳ	ҷ
	ҷ	ҥ	ә	ҽ	ё	„	«	»

Figure 5: Specific Cyrillic letters for T2B

	x0/x8	x1/x9	x2/xA	x3/xB	x4/xC	x5/xD	x6/xE	x7/xF
8x	Ҏ	Ҭ	Ҭ	Ҕ	Ԧ	Ҏ	Ҏ	Ӡ
	Ӎ	Ҟ	Ҡ	Ҟ	Ҩ	Ӎ	Ҧ	Ҋ
9x	Ө	Ҿ	Ҿ	Ҍ	Ѵ	Ҳ	Ҵ	Ҵ
	Ҷ	Ԩ	Ә	Ӂ	Ё	№	¤	§
Ax	ҏ	ҭ	ҭ	ҕ	h	ҏ	ҏ	ӡ
	ӎ	ҟ	ҡ	ҟ	ҩ	ҥ	ӎ	ҋ
Bx	ө	ҿ	ҿ	ҍ	ѵ	ҳ	ҵ	ҵ
	ҷ	ԩ	ә	ӂ	ё	„	«	»

Figure 6: Specific Cyrillic letters for T2C

	x0/x8	x1/x9	x2/xA	x3/xB	x4/xC	x5/xD	x6/xE	x7/xF
Cx	А	Б	В	Г	Д	Е	Ж	З
	И	Й	К	Л	М	Н	О	П
Dx	Р	С	Т	У	Ф	Х	Ц	Ч
	Ш	Щ	Ъ	Ы	Ь	Э	Ю	Я
Ex	а	б	в	г	д	е	ж	з
	и	й	к	л	м	н	о	п
Fx	р	с	т	у	ф	х	ц	ч
	ш	щ	ъ	ы	ь	э	ю	я

Figure 7: Russian letters for T2A/T2B/T2C/T2D