Romanian Slavonic in Unicode: problems and solutions†

Ion-Mihai Felea⋆

Faculty of Letters, "Alexandru Ioan Cuza" University, Bd. Carol I 11, 700506 Iași, Romania

Abstract

Editors of Slavonic and Slavonic–Romanian text can make use of a large variety of tools (fonts, physical and virtual keyboard layouts, word processors, operating systems) for transcribing and digitizing these texts in a uniform manner. The uniformity of the transcripts is based on Unicode standardization. Our study aims at explaining the place of Slavonic in Unicode and at briefly describing the most accessible tools. To this end, we shall describe the working tools from a historical and functional perspective and then provide examples in which those tools can be or have already been used to obtain a more accurate transcript. The user can choose from the existing methods and tools according to his/her purposes, needs and means. A better understanding of technical data can reduce the working time, improve transcription, accelerate learning times and generally make an editor's work much easier.

1. Slavonic in Unicode. Brief history

The standardization of telecommunication alphabets originates in the invention of the telegraph by Samuel Morse. The need for uniformity grew steadily, so that in the 1960's the American Standards Association created a system called ASCII (American Standard Code for Information Exchange), in which the common characters of the English language were assigned a number from 1 to 127 in the binary system. In the US, the ASCII standard was immediately implemented on computers.

The system has started to show its limitations with the advent of the Internet. On the one hand, ASCII encoding was used on a large scale in the United States and some European countries. On the other hand, some countries, like the Scandinavian ones, used variants of the ASCII code that differed from the American one, while other countries used alphabets that differed entirely from the Latin one. Japan, for example, used four completely different coding systems, which differed from ASCII and were also different from each other. Standardization at the global level became thus a necessity.

This task was entrusted in 1990 to the Unicode consortium, a non-profit organization set up to create, develop and maintain a universal standard for characters. The Unicode standard was based on the same principle. The ASCII characters kept their serial number, and the other characters were gradually assigned a unique and invariable number from 256 to approx. 150 000 (the number of characters existing at the present date). Harmonizing ASCII with a unique system for locating future characters represented an intellectual challenge for which a subtle solution, which was explained in a Computerphile documentary, was provided[1].

The consortium met the highest expectations. Presently UTF8 – Unicode Transformation Format 8-bits, one of the coding systems developed by Unicode, is used by 97% of the web pages.

⋆Email address: *imfelea@gmail.com*.

[1] *youtube.com*.

2. Standardization of Cyrillic and Slavonic characters

Standardization for the Cyrillic alphabet was gradually implemented in various stages, starting with the year 1990. In a first stage, the modern Cyrillic alphabets were coded in a compact, basic block, which was assigned numbers between 1024 and 1279. The Cyrillic character A was assigned the number 1040, Cyrillic Б – 1041, and so on. The specific characters belonging to the Cyrillic alphabet used for Mordvinic, Azeri, Chuvash and other such languages were coded in 2002. This first additional group which was assigned numbers between 1280 and 1327 was called Cyrillic Supplement. The next groups, called Cyrillic Extended A, B and C, Phonetic Extensions, Combining Half Marks, Glagolitic and Glagolitic Supplement, were created mainly for researchers and coded characters such as superscript letters, Iota, as well as Înea and Ge for the Romanian texts. The complete character inventory and their repartition can be consulted on the consortium webpage[2]. We should mention that the list has not been updated by Unicode according to version 9.0, yet the above-mentioned groups can be easily found on the internet. For clarity, we shall refer to character numbers in the decimal system, while Unicode numbering in interfaces is done in hexadecimal.

It is also worth mentioning that the people who documented and instrumented the processes aimed at introducing the additional Slavonic characters in Unicode also contributed to various other issues related to digitization and word processing. Michael Everson and Ralph Cleminson created multi-alphabetic fonts that also cover Slavonic writing (Everson Mono and Dilyana)[3], Vladislav Dorosh is in charge with the maintenance of the Irmologion website, where one can find a wide range of fonts and useful literature on the topic[4], Sorin Paliga designed several keyboard layouts[5], including Romanian and Slavonic ones[6], while Alexandr Andreev, Nikita Simmons and Yuri Shardt are currently the administrators of the Ponomar project[7], which provides excellent fonts, literature and technical support.

The same S. Paliga (2009, p. 179) provided an overview of the solutions available at the time for the transcription of old Slavonic texts into electronic format. Besides the informative content, his study helps us understand the evolution of the tools available today compared with those existing over a decade ago.

3. Tools

3.1. Slavonic Keyboard Layouts

One of the issues that have to be dealt with in order to transcribe a Slavonic text is related to the keyboard. Romanian users normally use a QWERTY layout keyboard or one of its versions adapted for Romanian (diacritic QWERTY), German (QWERTZ), French (AZERTY). A few such layouts are available for Romanian, only two of them—Romanian primary and Romanian secondary—being commonly used[8]. In the Slavic area, a version of ЙЦѸКЕН is commonly used, which displays the respective Cyrillic characters on the first six keys of the bottom row of the keyboard. We are thus entitled to say that quite a few solutions are presently available for rendering Slavonic texts.

---

[2] *unicode.org*.

[3] *evertype.com* and *localfonts.eu*. Ralph Cleminson is also the curator of the Obshtezhitie project, a valuable portal for the study of Cyrillic and Glagolitic texts: *obshtezhitie.net*.

[4] *irmologion.ru*, the website is in Russian.

[5] In the Romanian version of our study we preserved the English term 'layout', largely used in computing. The Romanian equivalent is 'aranjament'.

[6] *unibuc.ro*.

[7] *ponomar.net*, information available both in Russian and English.

[8] Further information and resources are available on Cristian Secară's blog, the IT specialist who elaborated the Romanian keyboard model complying with national standard SR 13392:2004: *secarica.ro*.

3.1.1. Windows

1. The use of layouts existing in the operating system. Starting with the version 10, Windows users are provided two Slavonic keyboards, based on the Russian type ЙЦӲКЕН keyboard, which does not require additional installing.

2. A virtual layout such as the one proposed by Ponomar and some on-line Slavonic dictionaries[9].

3. A QWERTY layout adapted for Slavonic. These keyboards replace most Latin characters with their Slavonic correspondents. Such options are available on the professional webpage of Daniel Bunčić, professor at the Institute of Slavonic Languages of the University of Köln[10]. His webpage provides a wide range of layouts for Modern and Old Cyrillic: Russian (ЙЦУКЕН), Bulgarian (УЕИШЩ), Serbian (ЖАЕРТЗ), Slavonic QWERTY (АВЕРТЫ) and others (Polish, Glagolitic, multilingual, AZERTY).

4. A personalized keyboard. Multiple tools are available for the creation of a personalized keyboard layout. The process is also accessible for users with less expertise in computing, yet it is extremely laborious and unjustified in our opinion, given all the available alternatives. Obviously, there are cases when one could opt for such solution, for instance when the user chooses a DVORAK layout[11] and wants to map the Slavonic characters according to his/her own typing skills/habits.

For a simpler long-term use, two options are available: an ЙЦУКЕН keyboard or a Slavonic QWERTY keyboard, each having certain advantages and disadvantages. For Romanian users, the most significant advantage of the QWERTY keyboard is that they do not need to learn a new arrangement of the characters, which is an important aspect especially when one has to work frequently with a large corpus of Slavonic texts. On the other hand, the ЙЦУКЕН layout comes in turn with a series of advantages. First of all, it allows the transcription of texts in modern Slavonic languages, such as Russian. Secondly, in our opinion, its other keyboard levels are organized more efficiently than in QWERTY. By keyboard level we mean the set of characters that can be obtained in combination with the Shift, AltRight keys, used together or with a deadkey. The most familiar is Level 2, accessed by pressing Shift (or Capslock), which gives access to uppercase letters. For instance, the layouts in Köln (Slavonic QWERTY included) contain uppercase letters on Level 2 (accessed by pressing Shift), while most of the special characters, mainly the superscripts, are obtained by pressing AltRight or a deadkey. However, ЙЦУКЕН places superscript letters on Level 2 (accessed by pressing Shift), whereas uppercase letters are placed on Level 3 (by pressing AltRight). Although Romanian users are definitely more familiar with this type of level organization, in practice superscripts occur more often in Slavonic texts than capital letters and hence the possibility of introducing them by pressing the Shift button seems more ergonomic. In our opinion it is also more efficient to place diacritics on the same line with the numbers. Finally we have to mention that the Köln Slavonic QWERTY layout is of the ISO type, which means that it includes a deadkey between Shift and Z. When ANSI or JIS[12] layouts are used, this key is missing and things get a bit more complicated.

3.1.2. Apple, Linux, Android

Users who work on different operating systems (Apple OS X, GNU/Linux, Android) can find free layouts on the Ponomar website, provided with clear and detailed installing and usage indications[13]. On S. Paliga's professional webpage one can download for free several types of layouts for the Slavic alphabets: Glagolitic, Slavonic, Czech, and others. These are designed for Apple systems users and contain instruc-

---

[9]*gorazd.org*, for instance.

[10]*slavistik.phil-fak.uni-koeln.de*.

[11]Although it represents the de facto state of current keyboards, the QWERTY layout is not the most efficient. Some users prefer the more ergonomic DVORAK layout, which allows a faster collection of the text. A DVORAK version is also available for Romanian, called POPAK, after the name of its creator, Nicolae Popa. Further details are available at: *invatasingur.ro*.

[12]You can determine what type of keyboard you have by following the instructions on this link: *wikipedia.org*.

[13]*sci.ponomar.net*.

tions for use and answers to frequently asked questions[14].

When a user opts for a Slavonic keyboard layout, several aspects might be considered: the operating system (Windows, Mac, Linux or Android), the type of layout (ЙЦУКЕН, DVORAK or QWERTY), the type of keyboard on which the layout will be used (ISO, ANSI or JIS) and the type of texts that are edited. A Romanian Windows user who is not interested in transcribing a text into a Modern Slavic language could successfully use a Slavonic QWERTY layout (АВЕРТЫ) as the Köln one, on an ISO keyboard. An ISO keyboard with blank keys can also be purchased separately and the Slavonic characters can be drawn on it manually to facilitate learning.

3.2. Unicode Fonts

Fonts are essentially an inventory of commonly accepted glyphs / drawings for letters which store a correspondence between a glyph and the Unicode code of a character, along with instructions regarding the interaction between characters, such as placing diacritical marks on letters, kerning (aesthetic spacing), ligatures. The computer receives the information about the pressed key, the font checks the character code in its own inventory (usually displayed for the users as *The quick brown fox jumps over the lazy dog* pangram) and requires the video card to 'draw' the character on the screen. For instance, if we press the key Q while a Slavonic QWERTY layout is activated, the keyboard will inform the computer that the character assigned the Unicode code 467 (in hexadecimal) was pressed, and the Garamond font (used for this article) will require the video card to 'draw' it as: ѧ. The Unicode code of a character is revealed by placing the cursor right after the character and pressing the combination Alt+x. This procedure also works in reverse. Thus, when we discuss of a rare Slavonic character, as closed little Yus (ꙙ), for the sake of clarity we can also mention its Unicode code (A659 in this particular case).

Most fonts currently use Unicode codes, yet 25 years ago a creator of Slavonic fonts could arbitrarily trace correspondences between the glyph and the character. More experienced users probably remember the Method font which displayed Slavonic letters on the screen by actually storing Unicode codes of the Latin alphabet letters. When using exclusively Unicode-compliant fonts, we ensure that our text stores exactly the characters we intended to transcribe in a standardized manner, which is independent of the font, device, software or operating system that our potential readers might use.

3.3. What fonts can be used?

This mainly depends on the purpose. For a strictly textual analysis of a Biblical verse where there is basically no concern regarding the graphic dimension of the text, a font such as Calibri or FreeSerif is more than enough.

In the normalized version, this passage from the Slavonic Tetraevanghelion, *Mt* 4, 3, is perfectly comprehensible: *Послѣди взалка и пристѧпль къ немꙋ искꙋситель, рече а̑ ще сі́й еси бж҃ій, рьци да каменіе се хлѣби бꙋдѧтъ.*

However, if we aim to provide a more exact or even a diplomatic transcription we need to use a dedicated Slavonic font that contains superscripts, diacritic marks, Yotta letters, palatalized consonants, Glagotic letters, alographs, special signs and the like. Fortunately, unlike 20 years ago, multiple, satisfying and often free solutions are readily available:

1. Dilyana: the font conceived by R. Cleminson, which, although quite old, is still successfully used[15].
2. BukyVede: Created by Sebastian Kempgen. We should mention that the Method font, conceived by the same author, was updated to a new version, Unicode compliant.
3. The CyrillicaBulgarian Family: fonts conceived for the CyrilloMethodiana project conducted at "Saint Climent Ohridski" University of Sofia. It was adopted by other projects, such as Gorazd and eRomLex.

---

[14] *dropbox.com*.
[15] *fonts2u.com*.

4. The fonts of the Ponomar.org project. They are destined to various purposes. Ponomar proposes fonts for the synod Slavonic (Ponomar Unicode), pre-Nikonian Slavonic (Feodorovsk Unicode, which imitates the letter of Ivan Feodorov's printing press), uncial manuscripts (Menaion Unicode), fonts destined to academic use (Monomakh Unicode), and decorative fonts (Indiction Unicode). Since Monomakh is destined to researchers, it also contains diacritic marks for the Romanian language, a feature that spares the user of the effort to change layouts when transcribing bilingual texts concomitantly.

Any of the above-mentioned fonts represent a valid solution, although we should admit that we prefer the Ponomar fonts, especially the eponym one and Monomakh. They contain the Cyrillic Extended C set, treat some ligatures better and display the diacritic marks in a more harmonious manner.

4. Problems

4.1. Characters that are not encoded in Unicode

Although the inventory of characters encoded in Unicode is enormous and is permanently updated and enriched, it is possible that a character a certain researcher or amateur user might need is not included in the Unicode inventory, and the screen displays an image such as: ⊠. Two solutions are available:

1. You can ask the Unicode consortium to introduce the character into standard and assign a number to it. You must have your documentation ready, prepare arguments, pass some stages, and if decision-makers regard your request as justified, encoding can be performed. You can find a model for such an accepted proposal on the consortium web page[16].
2. In the meanwhile, you can use Private Use Area (PUA). Unicode intentionally left some empty positions so that users can introduce symbols and glyphs that have not been officially encoded into their own fonts. For instance, when one wants to render faithfully a Romanian text containing multiple ligatures, specific to the chancellery texts, especially those from the late period, one can create a personal font that includes glyphs for the ligatures and assign PUA codes to them. All the known Slavonic fonts have encoded characters or ligatures in PUA, even those especially designed for Unicode (CyrillicaBulgarian10U, Monomakh, Dilyana), as they all try to bridge certain gaps. Generally, the use of PUA may, on the one hand, decrease uniformity and standardization, while on the other hand it can increase the precision of transcription.

A possible candidate for point 1 above is provided by some of the early Slavonic–Romanian texts, where there occurs a form of horizontal рьци. Some researchers, starting with Rosetti (1956, p. 116), claimed that this grapheme belongs to a series of signs that note a vibrant apical *r*. Our intention is not to take this theoretical debate any further, but to emphasize that the researcher who wants to include it in the transcription can submit a scientifically motivated proposal to the Unicode consortium which, once accepted, will allow a more accurate transcription of fragments such as the one extracted from PsVor (Fig. 1). Giuglea (1910, p. 445) also transcribed it using a sigma variant to emphasize it.
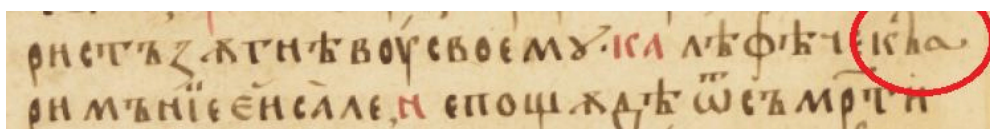

Figure 1: *Voroneț Psalter* (PsVor, *Ps. 77*, 54)

An alternative solution is to include the grapheme in PUA, yet this process should be carried out in the design stage of the respective font.

---

[16]*unicode.org*.

4.2. Morphologic-semantic disambiguation

The Slavonic orthography uses a series of means for antistoichos and the disambiguation of meaning: allographs, diacritics, orientation of accents, vowel alternation. We shall provide just a few examples. The wide variant of *slovo* (1C83, see Fig. 2) is often used in words that refer to divinity: adjectives, theophoric pronouns, the name of Christ. The demonstrative pronoun *céй* is frequently written with the basic letter when referring to objects and with the wide variant when it refers to beings. The wide variant is occasionally used to differentiate homographs in oblique cases from homographs in direct cases. Another example of antistoichos is the pair of normal естъ (U+0435) and wide естъ (U+0454). This alternation can be identified in printed text that distinguish between *жены́* = 'a(l) femeii' [of the woman] (Gen. sg.) and *жены́* = '(pe) femei' [the women] (Nom. pl.). The Slavonic–Romanian texts often use the wide *e* for the initial position of the forms of the verb 'a fi' [to be], as indicated by any brief analysis of Ps1570.

Figure 2: The Cyrillic Extended-C group

The antistoichos and disambiguation marks can also be used concurrently: *нéмощи* = 'a(l) slăbiciunii' [of the weakness] (Gen. sg.) as opposed to *нêмѡщы* = 'slăbiciuni' [weaknesses] (Nom.-Acc. pl.). Some marks and alternations are commonly reproduced by all scientific transcriptions, as in the case of the pair *o/ѡ* in *êгò* = '(pe) el' [him] (Acc. sg.) and *êгѡ̀* = 'a(l) lui' [his] (Gen. sg.). Others are often overlooked. The characters from Cyrillic Extened-C, the most recent Slavonic group in Unicode, often fall into this category. The fonts created prior to 2016 are not the most appropriate for the display of these characters: either they do not include the characters at all, or they only have them encoded in PUA. Other fonts such as Monomakh prove to be useful in this case. Some of these variants are considered formal and their use cannot be predicted by algorithms, as it happens in the case of the pair *з/ꙃ*. More than often, their transcription does not improve digitization or transcription whatsoever, yet there are numerous editions and research paths that could benefit from a faithful transcription of the text. Some of the characters in this group and their localization in Unicode can be examined in Fig. 2.

5. Ligatures

In order for the ligatures to be displayed accurately, the font designer should specify the graphic behaviour of the characters that are part of a ligature. In case the ligatures are coded in PUA, a special type of character called Zero Width Joiner (ZWJ, U+200D) must be introduced between the characters that are to form the ligature. If the font was appropriately created, it will display a pre-drawn ligature. This strategy is largely used by other scripts that use ligatures: Devangari, Arabian; however, for Slavonic, this is still an incipient state.

Ligatures are currently difficult to display in Microsoft Word. In the text processor of the free open-source LibreOffice package, pre-drawn ligatures (such as ҃ѣ ꙋ ꙋ from the Ponomar font) are most often displayed accurately. If we examine the line of Slavonic characters mentioned above with the Alt+x command we can observe that a ZWJ was inserted in between the letters. Fig. 3 exemplifies the manner
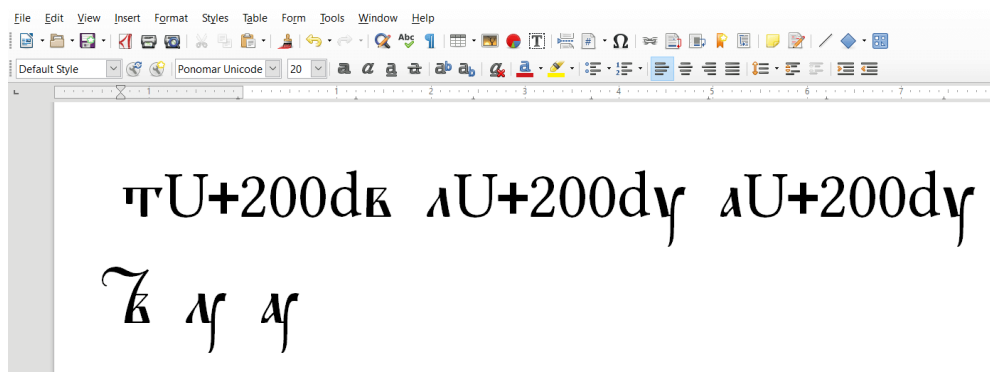
Figure 3: Tverdi-vede, liudi-uk and az-uk ligatures in LibreOffice Writer

in which LibreOffice Word automatically transforms the sequences letter-ZWJ-letter in the top row into the perfect ligatures displayed in the bottom row.

5.1. Spaces

Two types of spaces are specific to Slavonic typography: normal (U+0020), narrow (for clitics, after numerals, certain particles, prepositions). These can be of a break type (common spaces) and non-break type (before comma or full stop). Non-break spaces form a typographic unit together with the characters they precede. This means that in a text processor they will make the transition into the row together with the previous character or word. Fig. 4 displays several types of spaces used for the printing of Coresi's *Apostle*, around 1566.
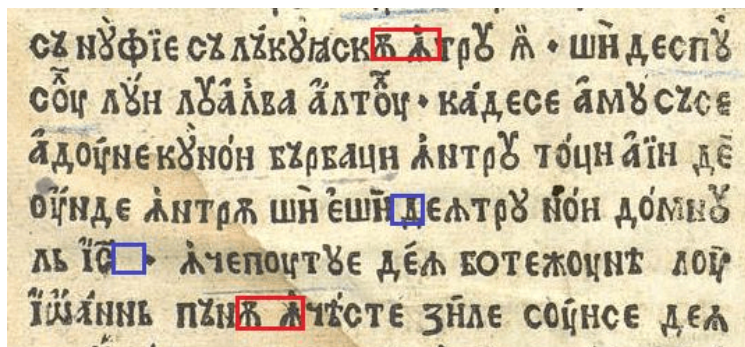

Figure 4: Normal and narrow spaces in Coresi's *Apostle*, around 1566

Narrow spacing can also be used in transcribing other texts belonging to Coresi. One can argue that in the Psalters from 1570 and 1577 we can identify traces of the grammaticalization of the Romanian definite article. I. Coteanu (*apud* Nicolae, 2013, p. 16) had signalled the existence of a free proclitic definite article in phrases such as "ispitire a ei voastre credinţă" [temptation of your faith]. The dative clitic is generally already fused with the noun in the stage of the first Romanian texts, yet some graphic particularities suggest the existence of a previous stage, between the one described by Coteanu and the one belonging to the current norm, where the demonstrative pronoun was not yet completely grammaticalized and had not lost its accent. We can exemplify with graphic forms such as *мънїе ей* in PsVor (Fig. 1), *пелúце êu / поýстиïu êu* in Ps1577 (Fig. 5 top) or *пелúцеéu / поýстïuéu* in Ps1570 (Fig. 5 bottom). All these three examples contain graphic conventions used for rendering the initial vowel: *e lat* (PsVor), *psili* (Ps1577) and *iso/psili+oxia* (Ps1570). In addition, in two of these cases—PsVor and Ps1577—there seems to be a space between the lexeme and the article.

The last type of space that can be used in transcription is U+200B, Zero Width Space (break or non-break). We can use it to separate words in texts with *scriptio continua*. We can thus preserve the initial text while allowing the computer to analyze the lexical units separately.
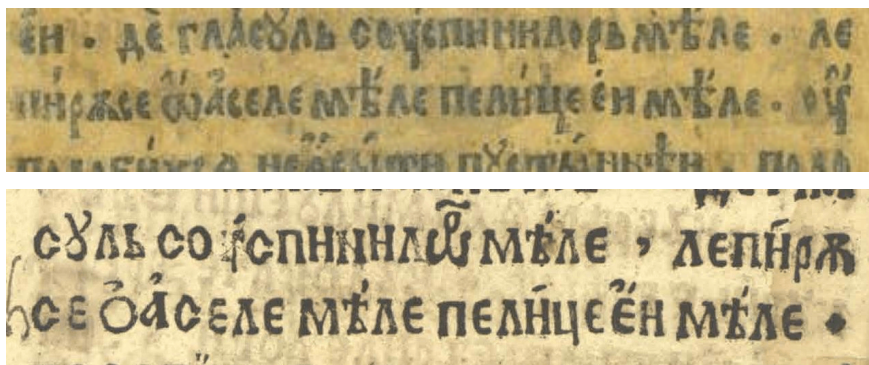
Figure 5: Ps1577 (top) and Ps1570 (bottom)

This type of space can be used for automatic analyses of early Slavonic corpora transcribed accurately.

Non-break spaces were used in the eRomLex project to transcribe the cases in which in a manuscript several superscript letters occur on the same basic letter and the font does not offer pre-set solutions (as it does, however, for the superscript combination slovo-tverdi). The first superscript letter was noted on the basic letters, while the remaining letters were placed after the basic letter, each on a non-break space.

5.2. Musical manuscripts

It is worth mentioning briefly that the website of the Ponomar project provides a series of solutions and explanations for the transcription of musical manuscripts. Their processing is a difficult task, as they contain both text and musical notes and in order to transcribe such material a dedicated soft or a more refined solution are definitely needed. Moreover, most of these manuscripts are not Slavonic–Romanian, but Greek–Romanian, and consequently out of our area of interest. Useful resources that can facilitate the work process are available to researchers who are interested in church music in the above-mentioned source.

6. Conclusions

A couple of decades ago, a perfectly diplomatic Slavonic electronic edition was difficult to achieve. Things are totally different today. We are pleased to see that we are currently well equipped in terms of the set of tools needed for editing Slavonic and Slavonic–Romanian texts. Keyboard layouts are available for all operating systems, along with fonts with almost exhaustive character inventories, with instructions for use that no longer scare uninitiated users. All these tools allow us to produce text editions with a high degree of fidelity to the originals. The editing of Slavonic text and studies based on Slavonic text grow in diversity and have a whole range of objectives. Those interested in editing Slavonic texts, researchers or enthusiasts, can decide which of the existing tools best suits the proposed goals and the means at their disposal. A better understanding of the technical data can shorten the working time, improve the transcription, accelerate learning steps and, and generally make an editor's work much easier. We are hopeful that our study will help our readers in this respect.

Bibliography

A. Referințe

Giuglea, G. (1910). *Psaltirea Voronețeană: (mss. 693 dela Bibl. Acad. Rom.)* in "Revista pentru Istorie, Arheologie și Filologie", 7, p. 444–467, [online].

Macarie (1512). *Tetraevanghel slavonesc*, Târgoviște. Incunable nr. CRV3 of the Romanian Academy Library, available in the online catalogue of the institution.

Nicolae, A. (2013). *Gramaticalizarea articolului hotărît românesc. Noi rezultate*, in "Limbă și literatură", **LIV** (1–2), p. 7–19.

Paliga, S. (2009). *Culegerea textelor vechi slave în format electronic*, in "Romanoslavica", **XLV**, p. 179–188.

Ps1570 = Coresi (1570). *Psaltire românească*, incunable nr. BRV 16/1 of the Central University Library Cluj [online].

Ps1577 = Coresi (1577). *Psaltire slavo-românească*, two copies kept at the Romanian Academy Library.

PsVor = *Psaltirea Voroneţeană*, ms. rom. 693 of the Romanian Academy Library.

Rosetti, A. (1956). *Limba română în secolele al XIII-lea – al XVI-lea*, Editura Academiei, Bucureşti.

Toma, S. (1976). *Psaltirea slavo-română (1577) în comparaţie cu psaltirile coresiene din 1570 şi din 1589*, Editura Academiei, Bucureşti.

B. Resurse web

obshtezhitie.net
ponomar.net
slavistik.phil-fak.uni-koeln.de
unicode.org