PROPOSAL FOR REGISTERING THE OLD SLAVIC CYRILLIC SCRIPT IN UNICODE

Adopted at the International Conference Held in the Serbian Academy of Arts and Sciences from 15-17 October 2007, Organized by the SASA Language and Literature Department and the SASA Institute for the Serbian Language, on the Subject:

STANDARDIZATION OF THE OLD SLAVIC CYRILLIC SCRIPT AND ITS REGISTRATION IN UNICODE

Participants in the Conference

Heinz Miklas, Johannes Reinhart (Austria); **Rumjan Lazov, Anisava Miltenova** (Bulgaria); **Stana Jankoska, Violeta Martinovska** (Macedonia); **Elena L. Aleksejeva, Viktor A. Baranov, Anatolij A. Turilov** (Russia); **Dušnica Grbić, Jasmina Grković-Major, Gordana Jovanović, Zoran Kostić, Katarina Mano-Zisi, Predrag Miodrag, Slobodan Pavlović, Viktor Savić, Gojko Subotić,Gordana Tomović, Đorđe Trifunović, Brankica Čigoja, Novka Šokica, Irena Špadijer** (Serbia); **Father Josif** (Hilandar Monastery); **Václav Čermák** (Czech Republic); **Jelica Stojanović** (Montenegro)

INTRODUCTION

The Unicode Consortium is a non-profit organization founded for the purpose of developing, disseminating, promoting and using UNICODE-STANDARD which defines text presentation in contemporary computer programs. Unicode-standard has been accepted by the leading world information technology companies such as: Apple, HP, IBM, Just System, Microsoft, Oracle, SAP, Sun, Sybase, Unisys and many others. Every language, i.e. its script, has its code page (code system), containing all the letters, numerals and punctuation marks characteristic of that script. Until Unicode appeared, code systems were in collision: different systems used either the same place (code) for different signs, or different places (codes) for the same signs. Conversion between different systems was not possible. Unicode-standard uses a single code for each character (letter, numeral, punctuation mark or other symbol) regardless of the platform (computer operative system), computer program and language. Unicode-standard has become universally accepted and almost all the languages of the world, i.e. their scripts, have been registered in it today.

When the Cyrillic script is in question, Unicode has registered the complete contemporary Cyrillic (covering all the scripts of all the nations using it), and only those Old Church Slavic letters which the contemporary Cyrillic does not contain [17 lowercase (ѡ, ѣ, ѥ, ѧ, ѩ, ѫ, ѭ, ѯ, ѱ, ѳ, ѵ, ѷ, ѹ, ѻ, ѽ, ѿ, ҁ), 17 uppercase (Ѡ, Ѣ, Ѥ, Ѧ, Ѩ, Ѫ, Ѭ, Ѯ, Ѱ, Ѳ, Ѵ, Ѷ, Ѹ, Ѻ, Ѽ, Ѿ, Ҁ), 4 diacritical marks, 3 symbols for numerals)– see at www.unicode.org – U0400 Cyrillic] They, in fact, belong to the civil (secular) script (the so-called: "grazdanka"), which was modified in the 18th century (Unicode – codes from 0460 to 0481). That is why the letter "a" from the modern Cyrillic and the letter az „а" from the Old Church Slavic have the same Unicode code (0430). That is also the case with the other coinciding letters. A consequence of this is that the modern and old Cyrillic cannot be employed in the same font at the same time. An **OLD CHURCH SLAVIC SCRIPT WHICH WOULD ADDITIONALLY COVER NATIONAL REDACTIONS AS WELL HAS ACTUALLY NOT BEEN REGISTERED AT ALL**. The use of the present solution for application to the Old Church Slavic script, even for the simplest of uses, is not possible due to different problems:

– Lacking are the standard letters: ђ, ꙋ, џ, ꙑ, є, ꙗ, ꙟ, as well as the less frequently used: α, ꙣ, ε, ι, ʌ, ꙧ, ꙥ, ρ, с, ꙡ, т, ꙋ, ꙑ, ꙑ, ѽ, ꙓ, undefined ones:ъ, є, ѫ, ѥ and letters necessary for transliteration from the Glagolitic script.

– Letter names (in the contemporary and old Cyrillic) and shapes are different.

– Some have the same form, the old djerv „ђ" and the modern „ћ" but the pronunciation is different.

– Some have a different form, the old „ꙗ" and the modern „я" but the same pronunciation. Additional confusion stems from the fact that the old „я" did not develop from „ꙗ" but from small jus „ѧ", so that its place in the old and new alphabets differs.

– Some of them have different positions in the azbuka /alphabet/, in the modern and old Cyrillic so that the sorting is wrong.

The full and user-friendly application of the Old Church Slavic script requires the registering of numerous letters, ligatures, superimposed letters with and without titlos, a large number of diacritical and punctuation marks and all the Old Slavic numerals. Since that has not been registered, i.e. since there is no standard, all the existing fonts have been made in "their own" code system, which differs from other ones. Therefore, the main problem is that of incompatibility in the use of Old Slavic fonts in different settings, this preventing the reading and use of old texts in electronic form outside the setting they were created in. The Internet, i.e. international communication, has practically been made impossible.

PRINCIPLES

1. The script of a language has been fully registered in Unicode if, and only if, all the attributes of that script have been registered. In the case of the Old Slavic script, as concluded at the Belgrade and according to the adopted **Standard Old Slavic Cyrillic script**, that means:

a) **All lowercase and uppercase letters** (not variants of writing the same letter, i.e. glyphs)

Although Unicode does not register glyphs but letters, Old Slavic is a different case. The writing of glyphs, even in the same line, was the standard way of writing (for example: „ѻ“, „о“, „ѡ“, or „ѧ“and „ѧ“, or „т“and „т“, or „ѹ“, „ѕ“ и „ѕ“ etc.). That was not an exception but a rule of writing (see an example of writing). The rest of the letters should be considered to be glyphs which belong to Unicode's Private Use Area. In that way even the most complex Old Slavic text would be a Unicode text, fully compatible with all fonts observing the standard (Unicode).

b) **All superimposed letters with or without titlos** (written above letters or between letters) for lowercase and uppercase letters in accordance with solution No. 2 (see Composite letters)..

c) **All diacritical marks** for lowercase and uppercase letters and punctuation marks in accordance with solution No. 2. (see Composite letters).

d) **All numerals** for lowercase and uppercase letters (see for example www.unicode.org – U10140-Ancient Greek Numbers)

2. The sequence of Unicode codes should be in accordance with the alphabet adopted at the Belgrade Conference, for sorting purposes.

Bearing in mind the principles, it is not possible to make additions to what has been registered so far, and that was also concluded by the Belgrade Conference. New registering is required for two reasons. The first is that in the present registration there is no space for inserting a large number of new codes. The other reason is that, even if there was enough space, the existing Cyrillic codes would have to be completely rearranged in order to enable sorting. It is a fact that it is now impossible to change the sequence because the Cyrillic code page is already being widely used.

Registering of the Old Slavic script separately and independently from the present situation, as the Cyrillic has been registered, enables the Old Slavic script to be registered fully, in accordance with both principles, and then we can have both the contemporary and Old Slavic script in the same font.

According to our proposal and bearing in mind the principles for registering scripts, examples of writing and the principles of making composite letters, it is necessary to register 1,303 characters (see attachment). Glyphs (in yellow fields) should be coded in Unicode's Private Use Area, for which an internal standard also needs to be made to enable Old Slavic fonts to be compatible.

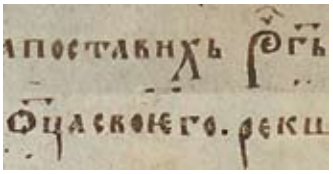
Summary of characters for registration

Lowercase	standard	glyphs
Letters	96	246
Superimposed letters without titlo (written over letters)	96	43
Superimposed letters without titlo (written between two letters)	96	43
Superimposed letters with titlo (written over letters)	96	43
Superimposed letters with titlo (written between two letters)	96	43
Diacritical marks and titlos (written over letter or between two letters)	49	37
Numbers	99	61
Punctuation marks and symbols	26	43
Uppercase		
Letters	96	124
Superimposed letters without titlo (written over letters)	96	43
Superimposed letters without titlo (written between two letters)	96	43
Superimposed letters with titlo (written over letters)	96	43
Superimposed letters with titlo (written between two letters)	96	43
Diacritical marks and titlos (written over letter or between two letters)	49	37
Numbers	99	61
Punctuation marks and symbols	21	60
Total number of unique characters	**1.303**	**1.013**

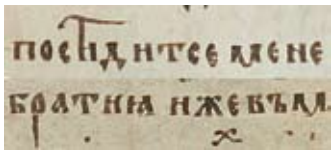
Ligatures for lowercase and uppercase letters (see for example www.unicode.org – UFB00 Ligatures for Latin, Armenian and Hebrew Ligatures), for now, remain in Unicode's Private Use Area

Lowercase ligatures	85	7
Uppercase ligatures	111	26

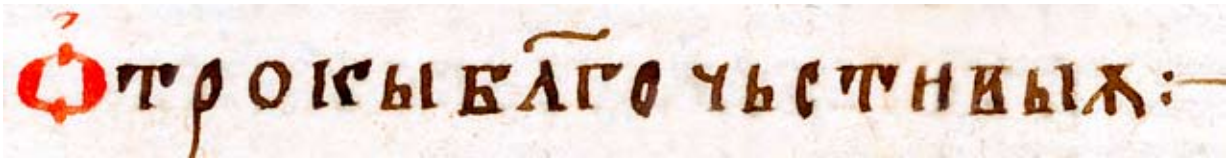
Examples:



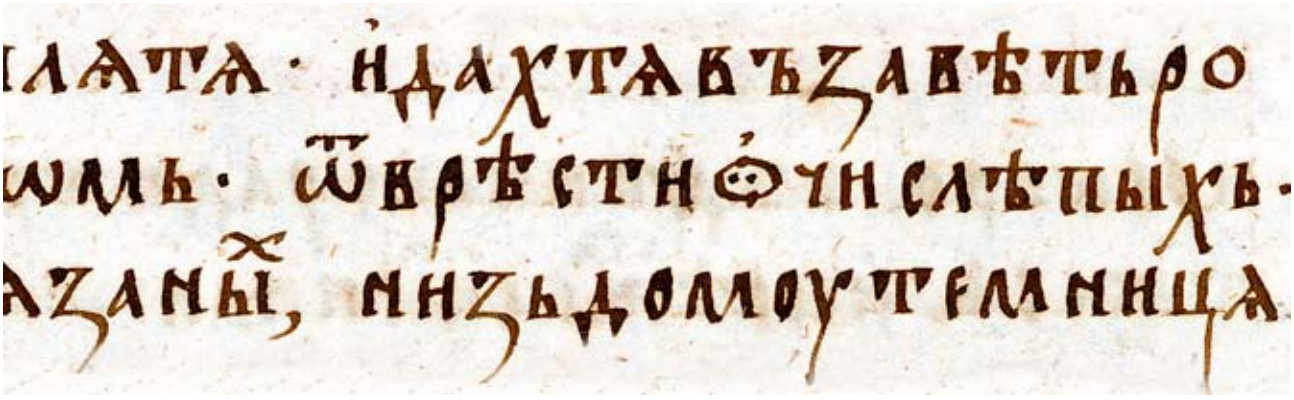
Ligature „pe“ ҏ, and letters „pe“ рє.



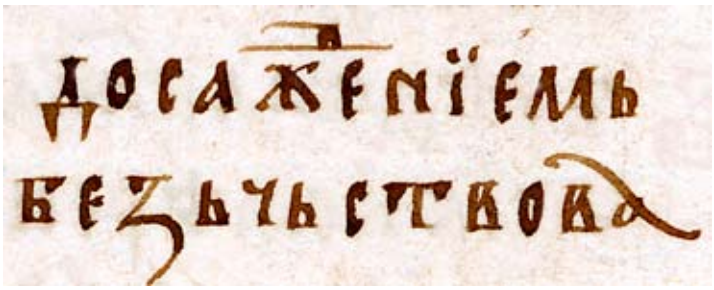
Ligature „ти“ ҥ, and letters „ти“ тн.



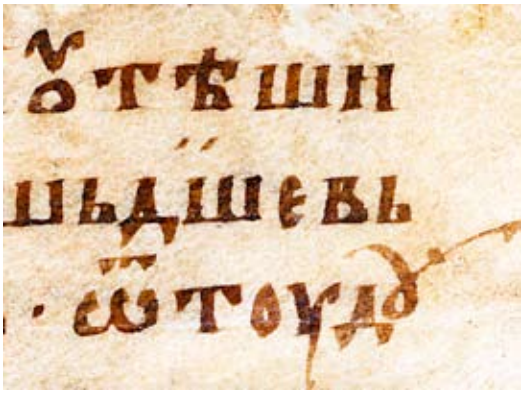
Letter ѻ "onSiroko2", о "onSiroko", ѳ "on".



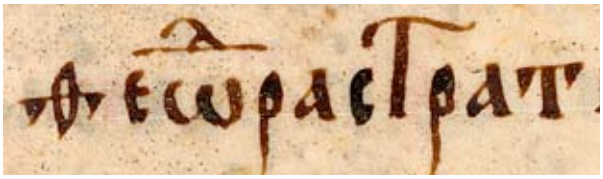
Letter о "onSiroko", ѽ "onDvooko2", ѳ "on".



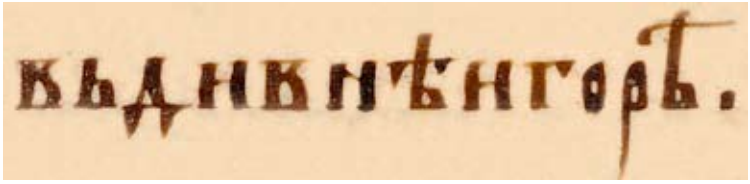
Letter а "az", ѧ "alfaCir5"..



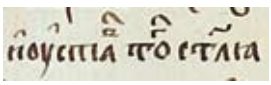
Letter ꙋ "uk", оу "onik", ѹ "ukDva2".



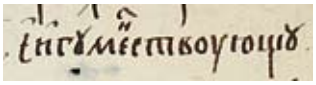
Letter т "tverdoVisoko", т "tverdo".



Letter ѣ "jat", ѣ "jatVisoki".



Letter „tverdo2“ т, and „tverdo3“



Letter „јота“ ї, and „ize“ н.

Composite characters

To construct a letter with diacritical mark over it, there are two possibilities. Let us show the principle on the example of lower case udieresis “ü” and upper case Udieresis “Ü”.

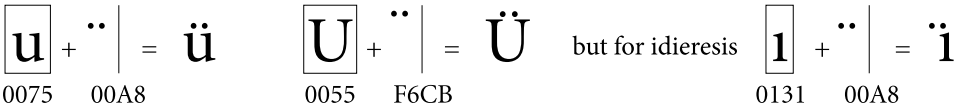
1. In Unicode we have letter and code for “u” (0075), diacritical mark and code for dieresis “¨” (00A8) and composite character with code for udieresis “ü” (00FC). In this composite character, typographer put dieresis exactly in right position over letter “u”.

For upper case “U” (0055) we use again the same dieresis as before. We shift dieresis and put over “U” in correct position in composite character Udieresis “Ü” (00DC). In both cases we do not alter position of two dots in the character dieresis itself, but we alter positions of dots in composite character.

All other letters with diacritical marks are constructed on the same way. When you type, you type directly “ü” with one keystroke. The consequence of this approach is that for all letters with dieresis you need two codes and two places (letter and composite) in font and only one code and place for dieresis. It means that for 12 letters (ä, ë, ï, ö, ü, ÿ, Ä, Ë, Ï, Ö, Ü, Ÿ) you need 25 codes and places in font.

2. The other way is simplified approach and for typographic point of view incorrect. We have code for “u” and code for dieresis “¨”, but we do not make composite character “ü”. Instead, when we make dieresis we put zero width for the character and position dots with offset on the left. When we type, we type first “u” and after that dieresis which come over the “u” because it has zero with and offset to the left. For upper case “Ü” we have code and place for “U” and now for upper case Dieresis (F6CB) “¨”. The dots are now in different position, higher and more to the left, with zero width of character. In both cases if we position dots exactly for the width of letter “u” and “U” we will have good result. So, for the same 12 letters (ä, ë, ï, ö, ü, ÿ, Ä, Ë, Ï, Ö, Ü, Ÿ) you need 14 codes and places in font (one for each letter and one for dieresis and one for Dieresis. This is considerable less codes and places than in first case.

What’s the trouble? As we make correct position of dieresis for the width of letter “u” and there is one dieresis for all letters only, the position of dieresis for letter “ı” (0131) will be incorrect.



To conclude:

- **Solution No1 is, from the typographic point of view, correct and all Roman scripts as well as the Cyrillic are constructed in this way. Where Old Slavic is concerned, several tens of thousands of code places would be required, because of the great number of vowels, diacritical marks and superimposed letters.**
- **Solution No2 is typographically incorrect but requires less than 10.000 codes.**
- **Solution No3 does not exist.**