Romanoslavica XLV

CULEGEREA TEXTELOR VECHI SLAVE ÎN FORMAT ELECTRONIC

Sorin PALIGA

Introducere

La începutul anilor ’90 ai secolului trecut am asistat la abandonarea treptată a tipografiei „clasice” Gutenberg, bazată pe culegerea textelor cu plumb şi adoptarea formatelor electronice ale textelor. Avantajele sunt evidente: textul este „cules” (respectiv „introdus”) în calculator, i se fac toate modificările necesare, apoi este „adus în pagină” şi pregătit pentru tipar. În mod uzual, acum tipografiile acceptă doar formate „gata de tipar” sau, folosind un termen englez, *camera-ready copy*.

Consecinţele sunt evidente: răspunderea privind acurateaţea şi corectitudinea textului revine integral autorului/autorilor, respectiv celor care au în grijă textul, cu tot ceea ce implică acest lucru, de la banala eroare de scriere (litere „sărite”, litere lipsă, neglijenţă în redactare) până la erori de paginare, în limba română sau în limba străină, dacă e cazul.

În afara acestor chestiuni generale, o problemă încă şi mai mare o pun – între altele – textele slave vechi[1]. De ce? Pentru a răspunde aceste întrebări, trebuie – mai întâi – să clarificăm unele chestiuni generale ori de detaliu.

Ce este *Unicode*? Unicode şi limba română

„Democratizarea” de după anul 1989 a însemnat, între altele, şi „dreptul de cetate”, respectiv „dreptul” de a fi folosite pe calculator, al unor „limbi noi”, cum ar fi limbile vorbite în ţările foste comuniste. Dar nu numai: limbile clasice (greaca veche, ebraica), limbile moderne (care folosesc alfabetul chirilic, dar şi alte limbi ale globului, cum ar fi limbile orientale, chineza, japoneza, coreeana, dar şi limbile care se scriu de la dreapta la stânga, aşa numitele limbi RTL – *right-to-left* – cum ar fi araba şi ebraica) etc. etc. Această avalanşă de limbi, fiecare cu specificul său, a făcut presiuni asupra

[1]. Nu este în orizontul acestui succint articol să ia în discuţie problemele ridicate de textele scrise în alte limbi, precum limbile orientale ori limbile antice „moarte” (akkadiana, sumeriana etc.) Ne vom opri aici numai acelor chestiuni legate de culegerea textelor vechi slave şi nu de toate problemele care se pun sau se pot pune în cazul altor limbi.

producătorilor de software. Dacă, inițial, se lua un font anume, i se eliminau unele caractere și i se adăugau altele, de exemplu, caracterele folosite în chirilic, pentru a se obține un font rezonabil pentru nevoi curente, curând această metodă și-a dovedit limitele: s-a ajuns la sute și la mii de variante posibile, când fiecare creator de font punea aleator caracterele necesare unde i se părea mai potrivit. Consecința era inevitabilă: au apărut sute și mii de variante posibile, astfel că, dacă un text chirilic era scris folosind un anume font, să-i spunem fontul A, el trebuia citit tot folosind acel font A. Curând, autorii înșiși au rătăcit fontul folosit inițial, astfel că, la ora actuală, sunt mii și mii de pagini indescifrabile, autorii uitând să-și noteze ce font/fonturi au folosit în urmă cu 15-20 de ani. În fapt, aceste texte au devenit inutilizabile!

Era și normal astfel ca Unicode, un consorțiu chemat să pună ordine în acest haos lingvistic, să-și impună treptat regulile:

– fiecărui caracter i se alocă un cod numeric sau un cod numeric și alfabetic;

– acest cod trebuie supus discuției și apoi aprobării unui grup de specialiști;

– odată aprobat, acea codificare trebuie urmată de toți creatorii de fonturi.

– nu se admit variante grafice ale aceluiași grafem istoric, acestea fiind permise doar ca variante ale fonturilor, nu ca variante ale unor caractere în cadrul aceluiași font; cu alte cuvinte, un font va conține – în limitele astfel definite – un singur caracter astfel definit.

La ora actuală, majoritatea fonturilor care vin instalate odată cu sistemele de operare, indiferent care sunt acestea (MAC OS, Linux ori Windows), respectă aceste norme. Este însă posibil ca un astfel de font să nu cuprindă, spre exemplu, caracterele necesare scrierii chirilice. Fonturile care cuprind însă aceste caractere trebuie să respecte normele Unicode și, în unele cazuri, le respectă. La ora la care scriem aceste rânduri, există relativ puține fonturi care au implementat deja noile norme, dar – deși puține – sunt suficiente nevoilor curente ale slaviștilor. Este evident că, pe măsura trecerii timpului, noi și noi fonturi vor veni să completeze lista actuală. Sunt de menționat aici fonturile create de doi dintre participanții la discuții, respectiv de Ralph Cleminson (Marea Britanie), creatorul fontului Dilyana (cu varianta mai nouă Neon); și Aleksei Kriukov, autorul seriei reprezentate de fonturile Old Standard, Tempora și Theano[1].

Nu intrăm în detalii „istorice”, totuși dorim a nota câteva situații nefirești, pentru a înțelege dificultatea și complexitatea problemei. Limba română a cunoscut și ea anomalii de codificare. Cauza principală a fost că, *la începutul anilor ’90, România nu avea un standard național privind folosirea caracterelor specifice limbii române*, respectiv *ă â î ș ț*, cărora li se adaugă semnele de punctuație specifice. Deși acestea sunt folosite și în alte limbi, notăm – fie și în treacăt – că, în conformitate cu normele academice, semnele citării sunt „citat” (nu „citat“), deși încă și azi, mai ales în subtitrarea filmelor, se întâlnesc forme specifice limbii engleze (de exemplu ‘citat’ ori 'citat'). De asemenea, pe atunci, Academia – prin lucrările de referință specifice,

[1] Acest fonturi sunt gratuite și disponibile spre descărcare de pe paginile autorilor.

Îndreptarul ortografic, DEX etc. – nu definea clar ce semne diacritice TREBUIE folosite pentru a culege corect un text românesc. Precizarea a venit foarte târziu, abia odată cu DOOM 2, la pagina XXVI: „[...] căciula ˇ desupra lui *a: ă*[1]; circumflexul ˆ deasupra lui *a* și *i: â* și *î*; virgulița sub *s* și *t: ș* și *ț*, cu precizarea din nota 9 la aceeași pagină: !Și nu sedila, care se folosește sub *c* în alte limbi: *ç*. *În programele de calculator, ș, spre deosebire de ț, apare în mod greșit cu sedilă*".

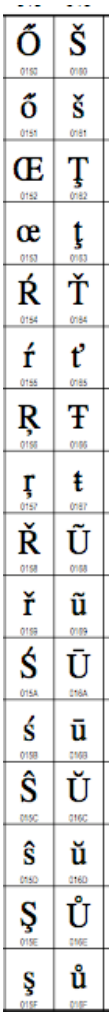
Ar fi de precizat că *s* cu sedilă (*s* cedilla) apare alături de *t* cu sedilă (*t* cedilla) în unele fonturi, în alte fonturi lipsind complet și *s* cu virgulă dedesubt (*s* comma below sau, cum scriu autorii DOOM 2, cu virgulița) și *t* cu virgulă dedesubt. Acest lucru s-a datorat, cum spuneam, în primul rând faptului că România nu avea, la începutul anilor '90, un standard, iar ulterior definirea diacriticelor s-a făcut confuz, neclar, iar principala companie creatoare de software, Microsoft, a introdus eronat s/t cu sedilă pentru limba română. În fapt, s cu sedilă este norma pentru limba turcă, iar t cu sedilă nu există ca atare în niciun standard[2].

În lipsa unor definiții clare, marile companii producătoare de software, preponderent Microsoft și Apple, ulterior și tot mai numerosele distribuții Linux, au adoptat (a se citi „improvizat") norme deduse din documentele accesibile la ora aceea. De exemplu, înaintea clarificărilor aduse de ASRO (Asociația română de Standardizare) în anul 2004, nu era clar dacă semnul diacritic de sub literele *s* și *t* era virgula (,) ori sedila (¸), era așadar *ș* și *ț* ori *ş* și *ţ*? Deși detaliul poate părea irelevant, cele două caractere fiind asemănătoare, în realitate, în sistemul Unicode, au codificări diferite:

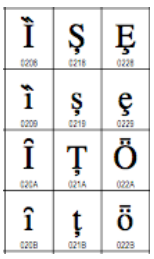
s cu virgulă dedesubt (s comma below)	ș, Ș	U+0219	U+0218
s cu sedilă (s cedilla)	ş, Ş	U+015F	U+015E
t cu virgulă		ț, Ț	U+021B U+021A
t cu sedilă		ţ, Ţ	U+0163 U+0162

[1] De fapt, în text apare semnul caron, ceh *háček* (ˇ), nu semnul scurtimii, *breve* (˘), cum definește textul, dovadă clară a faptului că autorii s-au lovit de aceeași problemă a redării corecte a caracterelor în format electronic, chiar în cazul unei limbi care folosește alfabetul latin și chiar dacă, față de alte limbi (ceha, polona, lituaniana etc.), româna are relativ puține semne diacritice.

[2] Nu intrăm aici în detalii privind discuțiile interminabile privitoare la definirea lui t cu sedilă ca folosit în transcrierea unor semne ebraice. Textul Unicode era ambiguu, se referea probabil la grafemul TZAV, deși, după știința noastră, nicăieri nu este definit ca fiind transcris prin t cu sedilă.



Locul lui s/S cu sedilă şi respecttiv locul lui t/T cu sedilă în cadrul aşa-numitei liste Latin Extended A. Codurile alocate secvenţiei ŞşŢţ sunt respecttiv U+015E U+015F U+0162 U+0163



Locul lui s/S cu virgulă şi respectiv locul lui t/T cu virgulă în cadrul listei Latin Extended B. Codurile alocate secvenţei ŞşŢţ sunt respectiv U+0218 U+0219 U+021A U+021B.

Din dorinţa de a simplifica, probabil, inventarul de caractere, la începutul anilor ’90, Microsoft a ales *s* şi *t* cu virgulă pentru limba română (incorect), în timp ce Apple,

corect, a ales *s* şi *t* cu virgulă. Această anomalie a persistat până recent, fiind corectată de compania Microsoft abia odată cu lansarea sistemului de operare Vista, deocamdată puțin răspândit. Ca atare, folosirea limbii române încă pune probleme utilizatorilor de Windows, care văd „pătrățele" în loc de s/t cu virgulă dedesubt. Deşi există şi un aşa-zis update care rezolvă chestiunea în sistemul Windows XP (European Union Bulgarian and Romanian Update, identificabil drept EUUpdate.exe pe site-ul Microsoft), acesta este, se pare, puțin cunoscut, astfel că majoritatea utilizatorilor continuă să aibă probleme. Nici nu există o precupare „națională" în acest sens, majoritatea utilizatorilor preferând fie să nu folosească diacritice (mai ales în poşta electronică), fie să le folosească eronat, lăsând-se pe seama companiilor pruducătoare de software.

Cum problema folosirii corecte a semnelor diacritice în limba română, conform normelor academice, este teoretic rezolvată, nu insistăm asupra altor probleme conexe. Întârzirea în luarea unor decizii şi-a făcut simțit efectul şi aici.

Unicode şi alfabetul chirilic

Deşi alfabetul chirilic a fost între primele codificate de Unicode, lista caracterelor era inițial incompletă, abia foarte recent, în cursul anului 2008, fiind corectate unele erori mai vechi şi fiind completată lista caracterelor lipsă. La începutul anilor '90, alfabetul chirilic era între primele alfabete nelatine adoptate şi codificate de consorțiul Unicode. S-a pornit, cum era firesc, de la caracterele folosite în limbile slave moderne care folosesc chirilicul, apoi s-a completat lista cu acele caractere chirilice folosite în limbile neslave vorbite şi folosite în diverse republici foste sovietice. Ulterior, li s-a adăugat o listă minimală de „semne chirilice arhaice" (Archaic Cyrillic), adică semnele slave vechi.

Lista nu era totuşi completă. Lipseau, pe de o parte, unele caractere folosite în unele limbi moderne neslave din spațiul fost sovietice, dar lipseau şi unele caractere vechi slave, iar câteva erau definite incorect. Ca atare, deşi cu mare întârziere, în cursul anilor 2005-2006 s-a inițiat, sub coordonarea lui Michael Everson[1], un proces de revizuire şi de completare a listei de caractere definibile ca „vechi slave" ori „chirilice", incluzând aici atât caractere specifice textelor medievale, cât şi textelor moderne. Țelul era ca, în decurs de circa un an, să se ajungă la consens în ceea ce priveşte lista de caractere lipsă ce trebuie definite drept „chirilice", vechi ori moderne precum şi lista de caractere inițial eronat definite (şi, implicit, eronat desenate în diverse fonturi). Autorul acestor rânduri a făcut de asemenea parte din acest colectiv.

[1] American de origine, Michael Everson s-a stabilit de mulți ani în Irlanda, de unde coordonează ample lucrări de implementare a unor standarde IT ce au ca scop principal folosirea unor caractere specifice limbilor rare.

~~ISO/IEC JTC1/SC2/WG2~~ ~~N3194~~
~~L2/06-xxx~~
~~2006-12-26~~

Universal Multiple-Octet Coded Character Set
International Organization for Standardization
Organisation internationale de normalisation
Международная организация по стандартизации

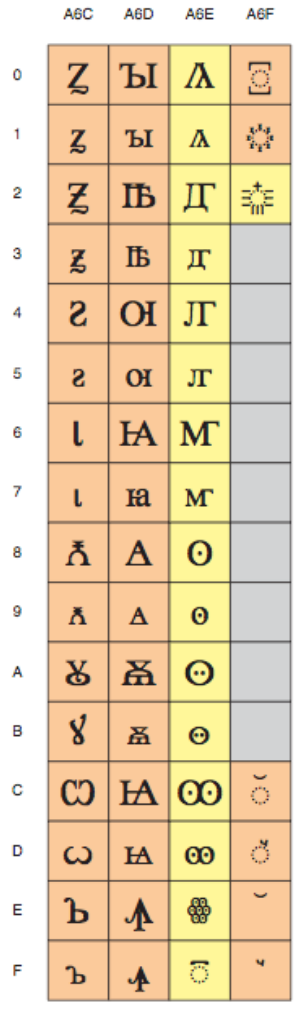
Doc Type: **Working Group Document**
Title: **Proposal to encode additional Cyrillic characters in the BMP of the UCS**
~~**Source:**~~ ~~**UC Berkeley Script Encoding Initiative (Universal Scripts Project)**~~
Authors: **Michael Everson, David Birnbaum (University of Pittsburgh), Ralph Cleminson (University of Portsmouth), Ivan Derzhanski (Bulgarian Academy of Sciences), Vladislav Dorosh (irmologion.ru), Alexej Kryukov (Moscow State University), Sorin Paliga (University of Bucharest), Klaas Ruppel (Research Institute for the Languages of Finland)**
Status: **Individual Contribution**
Action: **For consideration by JTC1/SC2/WG2 and UTC**
Replaces: **N3184, L2/06-359, N1744**
Date: **2006-12-26**

1. Introduction. This document requests the addition of a number of Cyrillic characters to be added to the UCS. It also requests clarification in the Unicode Standard of four existing characters. This is a large proposal. While all of the characters are either Cyrillic characters (plus a couple which are used with the Cyrillic script), they are used by different communities. Some are used for non-Slavic minority languages and others are used for early Slavic philology and linguistics, while others are used in more recent ecclesiastical contexts. We considered the possibility of dividing the proposal into several proposals, but since this proposal involves changes to glyphs in the main Cyrillic block, adds a character to the main Cyrillic block, adds 16 characters to the Cyrillic Supplement block, adds 10 characters to the new Cyrillic Extended-A block currently under ballot, creates two entirely new Cyrillic blocks with 55 and 26 characters respectively, as well as adding two characters to the Supplementary Punctuation block, it seemed best for reviewers to keep everything together in one document.

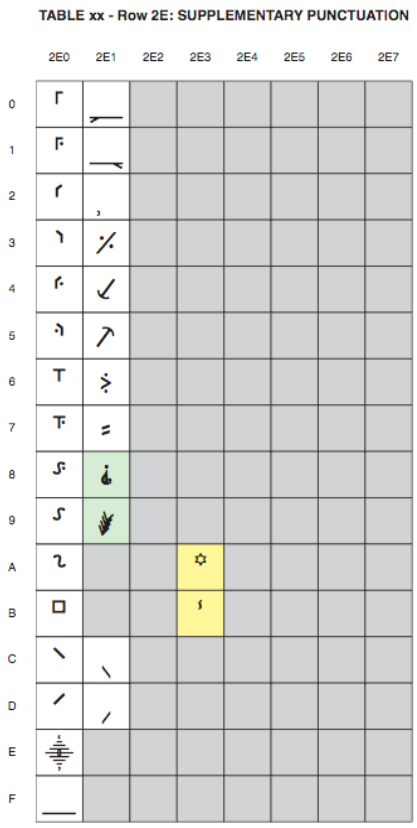
Prima schiță, datată 26 decembrie 2006, întocmită de Michael Everson în urma discuțiilor purtate privind corectarea unor erori și privind îmbogățirea caracterelor definite ori definibile drept chirilice (Cyrillic). Se poate vedea lista participanților precum și rezumatul lucrării.

Rezultatele finale au fost puse pe hârtie și apoi înaintate consorțiului Unicode de Michael Everson. Spre bucuria colectivului de lucru, aproape toate propunerile au fost aprobate, inclusiv mult așteptatele caractere specifice chirilicului românesc, nazala în (grafiată yn în documentele Unicode): ꙋ Ꙋ (așa apare în fonturile Dilyana și Neon, create de Ralph Cleminson), cu unele variante grafice ce încep să apară deja în unele fonturi, cum ar fi cele create de Aleksei Kriukov ꙋ Ꙋ sau în Everson Mono, creat de Michael Everson: ꙋ Ꙋ.

Rezolvarea acestor probleme legate de culegerea textelor slave vechi este benefică și lumii științifice de la noi, deschizând astfel drumul publicării – în condiții superioare – textelor slave vechi, inclusiv prin reeditarea unor lucrări azi epuizate.



Pagina referitoare la unele caractere inițial lipsă și care au fost introduse și definite în noua listă

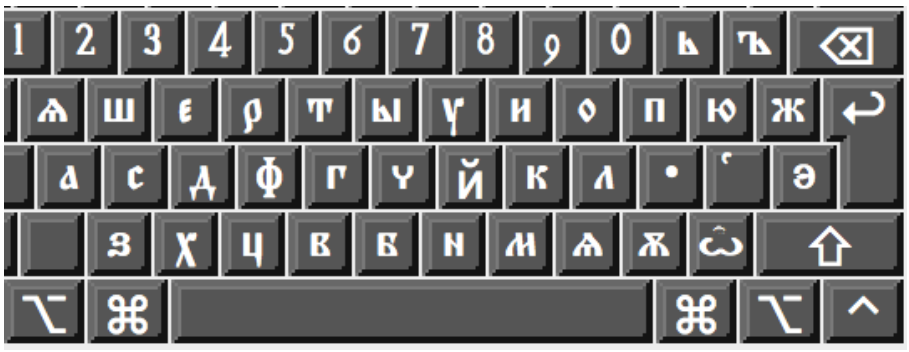


Pagina referitoare la semnele de punctuație

O pagină din forma finală adoptată de Consorțiul Unicode în versiunea 5.

Trebuie precizat că, în esență, mai trebuie rezolvată o singură problemă pentru a se putea redacta relativ ușor și rapid un text slav vechi: elaborarea unor tastaturi

(*keyboard layouts*) care să permită accesarea caracterelor slave vechi. Deocamdată, nicio companie producătoare de software, preponderent Apple şi Microsoft, nu oferă tastaturi care să acopere aceste nevoi. Pentru glagolitic, autorul acestor rânduri a publicat, pe paginile sale (vezi referinţele), o tastatură perfect funcţională pentru MAC OS X. Cum timpul nu este prea generos, nu am putut crea şi o variantă pentru Windows, dar suntem bucuroşi să împărtăşim din experienţa noastră tuturor celor dornici să facă acest lucru. Operaţiunea, precizăm, nu este atât dificilă, din punct de vedere tehnic, ci migăloasă. Iată aici o propunere de tastatură pentru slava veche şi propunerea de tastatură pentru glagolitic. Ultima dintre acestea este, sperăm, forma finală şi perfect utilizabilă pentru editarea oricărui text glagolitic.



O propunere a autorului de tastatură pentru slava veche, încă nefinalizată

Tastatura autorului pentru glagolitic, forma finală

Concluzii

La ora actuală, procesul de definire şi de implementare a tuturor caracterelor folosite în textele slave vechi s-a încheiat. Rămâne însă relativ limitată lista fonturilor disponibile şi, mai ales, lipsa unor tastaturi (*keyboard layouts*), absolut necesare pentru a lucra rapid şi comod cu asemenea texte. Pentru glagolitic, autorul acestor rânduri a finalizat tastatura glagolitică, fiind în fază relativ avansată tastatura pentru ceea ce se numeşte chirilic. Date fiind problemele ridicate de un repertoriu relativ mare de caractere, rezolvarea nu poate fi rapidă.

Tabula Gratulatoria

Aducem aici calde mulţumiri tuturor celor care, de-a lungul anilor, ne-au sprijinit în elaborarea unor lucrări şi în clarificarea unor chestiuni de detaliu. Lista ar fi prea lungă, citez aici doar câteva nume: Ralph Cleminson, Alex Eulenberg, Michael Everson, membri ai echipei Apple din Cupertino, California.

Referinţe

Cleminson, Ralph, http://web.ceu.hu/medstud/ralph.htm; Autor a numeroase studii privind culegerea textelor slave vechi. Autorul fonturilor Dilyana şi Neon

Everson, Michael, http://www.evertype.com/; http://evertype.com/mailman/ listinfo/ cyrillic_evertype.com. Probleme generale legate de fonturi, unicode, fonturi cu repertoriu bogat pentru culegerea textelor în limbi arhaice şi moderne

Linguistic List, http://linguistlist.org/

Medieval Unicode Font Initiative, http://helmer.aksis.uib.no/mufi/

Paliga, Sorin, *Despre Macintosh şi sistemul de operare MAC OS X*, Bucureşti, Meteor Press, 2008

Paliga, Sorin, http://www.unibuc.ro/ro/cd_sorpaliga_ro. Descărcare gratuită de tastaturi pentru glagolitic şi limbi vechi italice, discuţii privind standardele în era IT

SIL International, http://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&id=input-resources

Unicode Consortium, http://www.unicode.org/. Pagini de referinţă pentru tot ce înseamnă codificarea Unicode şi soluţiile adoptate

Vintilă-Rădulescu, Ioana (coordonator), *DOOM. Dicţionarul ortografic, ortoepic şi morfologic al limbii române*, ed. a II-a revăzută şi adăugită, Bucureşti, Univers Enciclopedic, 2005

Old Church Slavonic Texts in Electronic Format

The paper briefly presents the current situation with Unicode Consortium, resumes the situation specific to Romanian in the wake of the recent changes brought by the Romanian Academy and the implementation of the Academic norms in the operating systems (Linux, MAC OS, Windows).
The paper focuses on the specific problems raised by Old Church Slavonic documents. The author presents his experience within a quite large group who, in 2006–2007 put together all the facts and arguments in order to enhance the Unicode Cyrillic characters, and also correct former errors.
Finally, the author suggests practic ways to set Mediaeval Cyrillic and Glagolitic documents in electronic form, and exemplifies with some specific data.