DIGITALNI ZAPIS SLOVENSKIH ZNAKOV

Primož Peterlin,[1] Aleš Košir,[2] Tomaž Erjavec[3]

(1) Inštitut za biofiziko MF, Univerza v Ljubljani
Lipičeva 2, 1000 Ljubljana
primoz.peterlin@biofiz.mf.uni-lj.si

(2) Hermes SoftLab,
Litijska 51, 1000 Ljubljana,
ales.kosir@hermes.si

(3) Odsek za inteligentne sisteme, Institut Jožef Stefan
Jamova 39, 1000 Ljubljana
tomaz.erjavec@ijs.si

POVZETEK

Obdelana je tematika zapisa slovenskih znakov v računalništvu in informatiki. Predstavljen je pregled obstoječih zakonskih norm in uporabljenih praktičnih rešitev. Navedemo nekaj primerov slabih rešitev, jih komentiramo, in predlagamo boljše. Kot problem, ki še čaka na rešitev, pa so izpostavljeni problemi standardiziranega kodiranja pismenk, ki so jih uvedli slovenski slovničarji s prve polovice 19. stoletja.

ABSTRACT

The topic of digital coding of Slovene characters in computer is treated. An overview of the existing regulating norms is presented, along with solutions found in practice. Some examples of inferior solutions are quoted and commented, with suggestions for improvements. The problem of digital coding of glyphs introduced by the 19th century Slovene grammarians is presented as a case still waiting to be solved.

1 UVOD

Slovensko abecedo sestavlja petindvajset črk: A, B, C, Č, D, E, F, G, H, I, J, K, L, M, N, O, P, R, S, Š, T, U, V, Z in Ž. Treh črk med njimi, Č, Š in Ž, angleščina, *lingua franca* računalništva in informatike, ne pozna. Zakaj je s tem lahko težava? Črke in drugi znaki so v računalniku predstavljeni s številčno vrednostjo ali kodo. Običajna rešitev je vpeljava prireditvene tabele. Veliki črki A tako ustreza neka koda, veliki črki B neka druga koda, in tako dalje. Male črke imajo spet svoje kode, prav tako številke, ločila in drugi znaki. Takšni urejeni tabeli pravimo kodirani nabor znakov. Izmenjavo informacij olajša, če vsi uporabljamo enako prireditveno tabelo, oziroma isti kodirani nabor znakov. Kodirani nabori znakov so zato standardizirani, za usklajevanje standardov pa skrbijo standardizacijska telesa na državni in mednarodni ravni.

V nadaljevanju je predstavljenih nekaj kodiranih naborov znakov, ki se ali pa so se uporabljali na ozemlju Republike Slovenije. Zadržimo se ob trenutnem stanju in pokomentiramo nered, ki vlada na področju kodiranja naših znakov. Kodirani nabor znakov, ki se uveljavlja, je ISO 10646. Obeta rešitev cele vrste jezikovnih problemov, prinaša kopico novih tehničnih problemov, nazadnje pa ponuja tudi možnost kodiranja pismenk slovenskih slovničarjev 19. stoletja. Končamo s pregledom programov in metod prečrkovanja.

2 KODIRANI NABORI ZNAKOV

Prva zakonska rešitev kodiranja slovenskih znakov je bila sprejeta leta 1982 še v okviru tedanje Socialistične federativne republike Jugoslavije kot jugoslovanski standard JUS I.B1.001 (Ur. list SFRJ št. 72/82). Naslanjal se je na mednarodni standard ISO 646 iz leta 1973, ta pa sloni na še starejšem ameriškem standardu ANSI X3.4 iz leta 1968, bolj znanem pod kratico ASCII (American Standard Code for Information Interchange). Standardni nabor ASCII je sedembiten in kodira $2^7 = 128$ znakov; poleg šestindvajsetih velikih in malih črk angleške abecede ter desetih številk še 33 drugih izpisljivih znakov ter 33 krmilnih neizpisljivih znakov. Jugoslovanski standard je deset manj rabljenih znakov (@[\]^` {|}~) nadomestil z velikimi in malimi črkami Ć, Č. Đ, Š in Ž, s čimer je omogočil izražanje v slovenščini in latinični srbohrvaščini. Standard JUS je zrelost dosegel nekaj let kasneje v izdaji JUS I.B1.002, njegovo popularnost pa izpričujejo tudi »ljudska« imena zanj, kot npr. YUSCII, ter kasneje, odvisno od geografske širine in dolžine pišočega, SLOSCII oziroma CROSCII.

Podobno kot SFRJ so problem svojih znakov rešile tudi druge države, in v sedemdesetih letih je v Evropi in drugod po svetu na osnovi ISO 646 nastal dober ducat nacionalnih standardov. Že takrat so se kazale omejitve sedembitnega nabora znakov, saj v besedilu ni bilo možno kombinirati besedila v dveh različnih jezikih, npr. nemško in francosko. Na pobudo evropskega združenja proizvajalcev računalniške opreme ECMA je mednarodna organizacija ISO zato leta 1987 izdala prve štiri osembitne standarde iz družine ISO 8859 [1], ki so po vrsti urejali kodne razporede za jezike zahodne

Evrope, vzhodne oz. srednje Evrope, mediteranskega bazena z Južno Afriko, ter baltsko-skandinavskega območja. Prvi, »zahodni« ISO 8859-1, je hitro doživel uveljavitev v industriji, tudi »vzhodni« ISO 8859-2 je na sicer ekonomsko znatno šibkejšem območju srednje in vzhodne Evrope in delno navzlic prizadevanjem industrije dosegel relativno dober sprejem, medtem ko se je za »južni« ISO 8859-3 in »severni« ISO 8859-4 izkazalo, da nista dovolj domišljena. Družina osembitnih naborov ISO 8859 je medtem skupaj z novimi predlogi narasla že na štirinajst članov in se še vedno dopolnjuje.

Osembitni kodni nabor ISO 8859-2 je na mestih s kodami od 0 do 127 identičen standardu ISO 646 [2], na preostalih 128 mestih pa kodira vse potrebne znake za pisanje v albanščini, češčini, finščini, hrvaščini oz. srbohrvaščini, irščini, gornjelužiškosrbščini, madžarščini, nemščini, poljščini, romunščini, slovaščini in slovenščini. Od latiničnih jezikov srednje in vzhodne Evrope tako zaradi manjkajoče črke M z ostrivcem [3] manjka le spodnjelužiškosrbščina. SFRJ je standard nostrificirala 29. januarja 1988 kot jugoslovanski standard z oznako JUS I.B1.013 (Ur. list SFRJ št. 9/88). V veljavo je stopil dva meseca kasneje, ni pa imel statusa obveznega tehničnega predpisa.

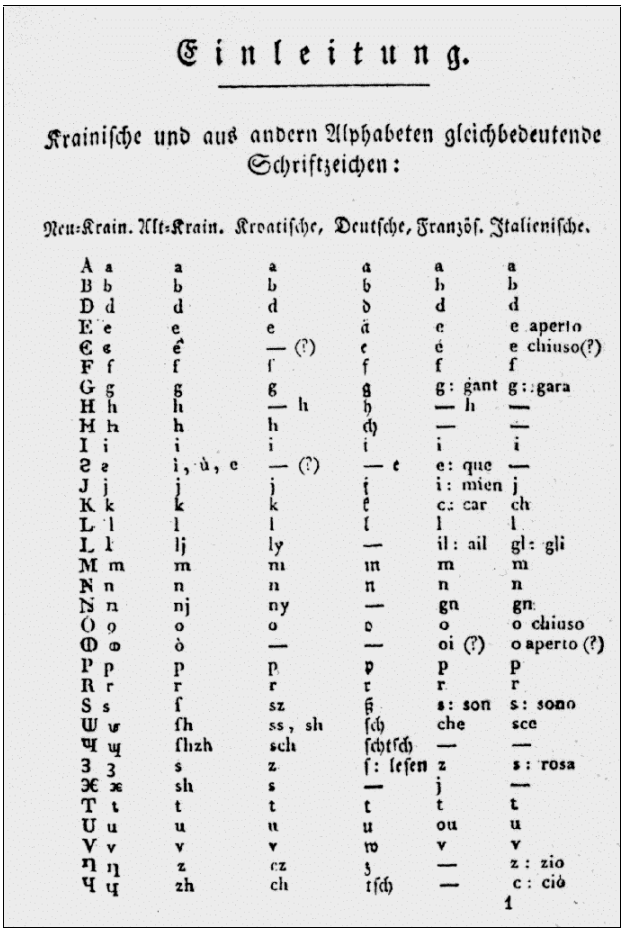
Devetdeseta so prinesla pravo poplavo kodnih razporedov. V vsega nekaj letih so multinacionalke IBM, Microsoft in Apple izdale vsaka svoj kodirani nabor znakov, vse z ambicijo, da pokrijejo potrebe pišočih v srednji in vzhodni Evropi. Vsi trije nabori so res vsebovali vse potrebne znake za pisanje v ciljnih jezikih, vendar na različnih mestih v tabeli. Ti kodni nabori so zato nezdružljivi tako med seboj, kot tudi s standardom ISO 8859-2 [4]. Glede na to, da so čisto vsi nastali po sprejemu mednarodnega standarda ISO 8859-2, za táko ravnanje težko najdemo razumljivo razlago.

3 SLOVENIJA IN ISO 8859-2

Značilno je, da je ISO 8859-2 začel pridobivati na veljavi s širitvijo Interneta, večanjem obsega komunikacije, in manjšo navezanost besedil na orodja, s katerimi so bila ustvarjena. Če je bilo v zaprtih okoljih še pogojno sprejemljivo, da uporabniki Microsoft Windows pišejo v kodnem naboru Codepage 1250, uporabniki IBM OS/2 v Codepage 852, tisti na Apple Macintosh v MacRoman CE, tisti na sistemih Unix večinoma v ISO 8859-2, vsi pa po malem še v JUS I.B1.002, je z odprtjem, ki jo je prinesel Internet, postalo to nevzdržno. Če pošljemo elektronski dopis ali objavimo stran na svetovnem spletu, je potrebno, da lahko naslovnik napisano tudi prebere, in to ne glede na to, kakšno strojno ali programsko opremo uporablja.

Problem se rešuje z različnimi tehničnimi rešitvami. V elektronski pošti in elektronskih novicah (USENET) prevladuje prečrkovanje z uporabo črk C, S in Z namesto Č, Š in Ž. Na spletnih straneh pa so razmere pestrejše. Precej strežnikov uporablja vmesnike CGI (Common Gateway Interface), ki na strani strežnika prečrkujejo iz enega kodnega nabora v drugega. Če se je to še pred nekaj leti zdela dobra rešitev, pa zdaj ni več tako. Oba najpogostejša spletna brskalnika, Netscape Navigator/Communicator in Microsoft Internet Explorer, podpirata kodni nabor ISO 8859-2 tudi v okoljih, ki naravno uporabljajo drug nabor znakov (npr. Microsoft Windows), če le strani v glavi MIME (Multipurpose Internet Mail Extensions) pravilno označujejo uporabljeni kodirani nabor znakov. Tudi sodobna orodja za pripravo spletnih strani, kot npr. Netscape Composer ali Microsoft FrontPage, podpirajo standard ISO 8859-2 in spletne strani korektno opremljata z glavo MIME. Način dela z vmesniki na strežniški strani je zato slabša rešitev, saj po nepotrebnem obremenjuje strežnik, takšnih navidez »dinamičnih« spletnih strani mrežni medpomnilniki (angl. cache) navadno ne shranjujejo, s čimer dodatno obremenjujemo tudi omrežje [5,6], in ne nazadnje so tudi oznake virov (URL, Universal Resource Locator) daljše in manj pregledne.

Čudi torej, da spletne strani pri nas še vedno tako radodarno ponujajo možnost izbire sicer statičnih dokumentov. Moteče je tudi, da večina njih kot privzeti kodni nabor postavlja Windows-1250 namesto ISO 8859-2. Obrazložitev, da se s tem prilagajajo uporabnikom, ki zvečine uporabljajo opremo Microsoft, ne zdrži več, vse odkar oba popularnejša brskalnika tudi v tem okolju brez težav bereta strani, zapisane v skladu s kodnim naborom ISO 8859-2. Celo če prepustimo tržnim ponudnikom storitev interneta, da se ravnajo po lastnem tržnem instinktu, pa bi želeli in pričakovali, da bodo vsedržavni projekti, kot je na primer slovenski knjižnični sistem COBISS, dajali prednost standardnim rešitvam. Ni izključeno, da je tovrstni nered na slovenskem delu interneta tudi posledica ignorantskega odnosa Urada za standardizacijo in meroslovje pri Ministrstvu za znanost in tehnologijo do tega medija. Naslovna stran, http://www.usm.mzt.si/, se v času pisanja tega prispevka kiti s cvetkami, kot je ponujen izbor kodnega nabora z oznako »437 [Slovenski jezik (VMS)]«, prek katere pridemo do strani, kodirane v skladu s kodnim naborom JUS I.B1.002, vendar pa brez značke MIME o uporabljenem naboru znakov (te pravzaprav nima nobena stran na omenjenem strežniku), kar implicira kodni nabor ISO 8859-1. Kodni razpored 437 se ni nikoli uporabljal na operacijskem sistemu VAX/VMS, ni primeren za prikaz slovenščine, saj ne vsebuje naših znakov, in seveda ni enak naboru JUS I.B1.002. Takšna zmeda ne bi bila v čast še tako zanikrnemu spletnemu strežniku, glavnemu standardizacijskemu telesu v državi pa je resnično v sramoto.



Einleitung.

Krainiſche und aus andern Alphabeten gleichbedeutende Schriftzeichen:

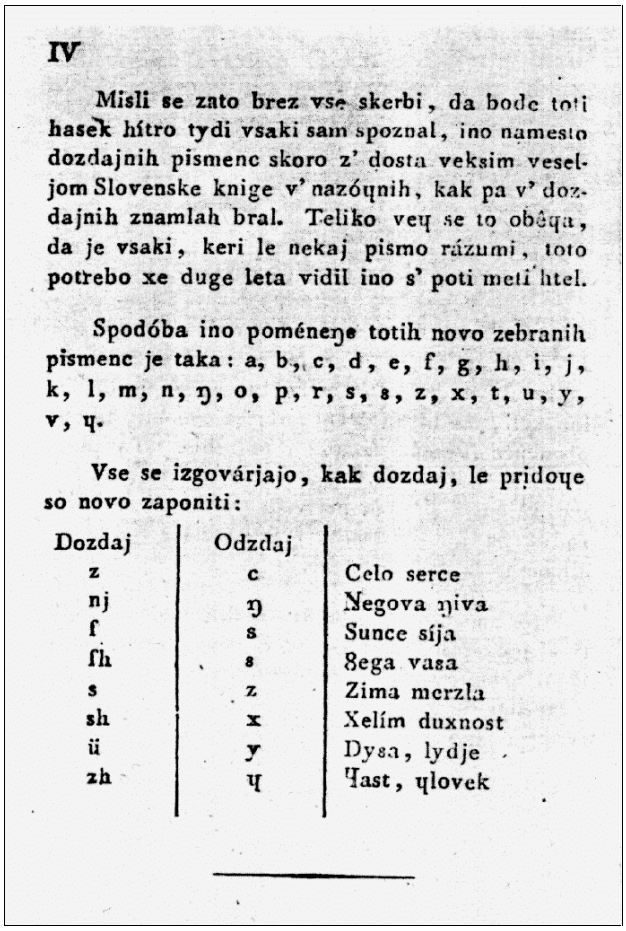
Neu-Krain.	Alt-Krain.	Kroatiſche,	Deutſche,	Franzöſ.	Italieniſche.
A a	a	a	a	a	a
B b	b	b	b	b	b
D d	d	d	d	d	d
E e	e	e	e	e	e aperto
Ê ê	ê	— (?)	ê	é	e chiuso(?)
F f	f	ſ	f	f	f
G g	g	g	g	g: gant	g: gara
H h	h	— h	h	— h	—
H h	h	h	ch	—	—
I i	i	i	i	i	i
Ə ə	i, ù, e	— (?)	— e	e: que	—
J j	j	j	j	i: mien	j
K k	k	k	k	c: car	ch
L l	l	l	l	l	l
L l	lj	ly	—	il: ail	gl: gli
M m	m	m	m	m	m
N n	n	n	n	n	n
Ń n	nj	ny	—	gn	gn
O o	o	o	o	o	o chiuso
Ò ò	ò	—	—	oi (?)	o aperto (?)
P p	p	p	p	p	p
R r	r	r	r	r	r
S s	ſ	sz	ß	s: son	s: sono
Š š	ſh	ss, sh	ſch	che	sce
Ꜩ ꜩ	ſhzh	sch	ſchtſch	—	—
Ž ž	s	z	ſ: leſen	z	s: rosa
Ꜫ ꜫ	sh	s	—	j	—
T t	t	t	t	t	t
U u	u	u	u	ou	u
V v	v	v	w	v	v
Ʒ ʒ	z	cz	ʒ	—	z: zio
Ч ч	zh	ch	tſch	—	c: ciò

1

Slika 1: Franc Serafin Metelko, Lehrgebäude der Slowenischen Sprache, Laibach 1825. Začetek uvoda s pregledom pismenk za foneme v slovenščini in bližnjih jezikih. Metelčica je v prvem stolpcu.

4 ISO 10646/UNICODE

V zadnjih letih se začenja uveljavljati kodirani nabor znakov ISO 10646 [7]. Ta 32-bitni kodni razpored je bil zasnovan z idejo, da bi normiral in kodiral pismenke čisto vseh pisav, ki se ali pa so se uporabljale na planetu. Naloga se je zaenkrat pokazala kot vendarle prezahtevna, in delovna skupina za njegovo pripravo se je osredotočila na prvo »stran«, to je prvih 65536 znakov, ter uskladila svoje delovanje s konzorcijem Unicode, ki se je istočasno loteval podobne naloge. Čeprav je standard še vedno v dodelavi in vseh 65536 mest še ni zasedenih, pa se obstoječi del ponekod že uporablja v praksi [8]. Res pa je izvedba 16-bitnega standarda tehnično zahtevna naloga, tako da bo verjetno poteklo še nekaj let, preden bo ta standard prevladal nad osembitnimi, ki danes predstavljajo večino. Več programskih hiš poskuša standardu olajšati pot v prakso tako, da daje prosto na razpolago pisave, kodirane v tem kodnem naboru, kot npr. Bitstream s svojo pisavo Cyberbit [9]. Družba Microsoft, ki je pred nekaj leti s svojim operacijskim sistemom Windows NT napravila pionirske korake pri uporabi tega nabora, je medtem sicer napravila korak nazaj, medtem ko so pisave iz družine Lucida, ki so jih prilagali k Windows NT, kodirale 1451 pismenk, uporabljajo njihove novejše pisave (sicer brezplačno dostopne prek interneta [10]) nekoliko okrnjen »hišni« nabor WGL4 (Windows Glyph List), latinično-ciriličnogrško podmnožico nabora ISO 10646 s 652 kodiranimi pismenkami. Za primerjavo navedimo, da kodira nabor MES (Minimum European Subset) [11] iz osnutka evropskega standarda ENV 1973:1995 926 znakov, nabor EES (Extended European Subset) [12] iz istega osnutka pa 3109 znakov.



IV

Misli se zato brez vse skerbi, da bode toti hasek hítro tydi vsaki sam spoznal, ino namesto dozdajnih pismenc skoro z' dosta veksim veseljom Slovenske knige v' nazóqnih, kak pa v' dozdajnih znamlah bral. Teliko veq se to obêqa, da je vsaki, keri le nekaj pismo rázumi, toto potrebo xe duge leta vidil ino s' poti meti htel.

Spodóba ino poménenje totih novo zebranih pismenc je taka: a, b, c, d, e, f, g, h, i, j, k, l, m, n, ŋ, o, p, r, s, ſ, z, x, t, u, y, v, q.

Vse se izgovárjajo, kak dozdaj, le pridoqe so novo zaponiti:

Dozdaj	Odzdaj	
z	c	Celo serce
nj	ŋ	Negova ŋiva
f	s	Sunce sija
ſh	ſ	Sega vasa
s	z	Zima merzla
sh	x	Xelím duxnost
ü	y	Dysa, lydje
zh	q	Qast, qlovek

Slika 2: Peter Dajnko, Kmet Izidor s svojimi otroki ino lydmi, Radgona 1824. Stran iz predgovora z razlago novih črk.

Kodirani nabor znakov ISO 10646 ima lepe izglede, da se v resnici uveljavi kot globalni nabor znakov. Svoj del pri krojenju svoje usode pa moramo pri tem prevzeti tudi Slovenci sami. Sodobno slovenščino ta kodni nabor povsem pokrije; tudi samoglasnike z razločevalnimi znamenji (ostrivec, krativec in cirkumfleks), s katerimi so v osembitnih naborih težave. Dodatna razločevalna znamenja, ki se uporabljajo v dialektologiji, lahko navadno sestavimo s kombinacijo kodiranih znakov. Zaenkrat pa v celoti manjkajo posebna znamenja, ki so jih

uvajali Franc Serafin Metelko (slika 1), Peter Dajnko (slika 2) in drugi slovenski slovničarji s prve polovice 19. stoletja. Kam pelje popolna odsotnost teh pismenk v mednarodnih standardih, si lahko bralka ali bralec, če ne drugje, ogleda v odgovarjajočem zvezku Enciklopedije Slovenije, kjer so stavci morali znake v stavljeno stran dorisati ročno, saj jih očitno tudi v ekspertnih tipografskih naborih ni najti. Te pismenke so za večino sveta neobstoječe, in bodo takšne tudi ostale, dokler se kdo ne spomni nanje. Glede na to, da je verjetno največja potreba po njih ravno pri nas, bi bilo na mestu, da se čimprej osnuje skupina strokovnjakov, ki naj pregleda, katere pismenke so samo tipografske izpeljanke že kodiranih pismenk, katere pa so v resnici izvirne, ter pripravi predlog za vključitev teh v standardni kodirani nabor znakov ISO 10646.

5 PREČRKOVANJE

V primeru, ko tehnične sposobnosti programske ali strojne opreme ne omogočajo vnosa, izpisa ali prenosa naših znakov, uporabimo prečrkovanje (angl. transliteration), navadno tako, da pri tem uporabimo le znake iz kodiranega nabora znakov ASCII, ki je nekakšen skupni imenovalec vseh sodobnih naborov znakov. Prečrkovanje gre lahko avtomatično: odposlano elektronsko pošto program avtomatično predela tako, da namesto črke Č v sporočilu stoji niz `=C8`, program na prejemnikovi strani pa izvede obrnjeno konverzijo. Različne sheme prečrkovanja uporabljajo tudi programi za obdelavo besedil. Sistem za elektronsko stavljenje LaTeX na primer opiše črko Č z nizom `\v{C}` (večina piscev sicer raje uporablja krajši makroukaz `"C` ali pa celo uporablja kar osembitne znake za vnos). Grafični jezik za opis strani PostScript za isto črko uporabi niz `/Ccaron`.

Posebej pazljivo so se lotili te teme snovalci standarda za označevanje besedil ISO 8879 [13], z imenom SGML (Standard Generalized Markup Language). Eden od osnovnih ciljev SGML je, da so v njem zapisani podatki prenosljivi z ene strojne in programske opreme na drugo brez izgube informacije. Označena besedila zato lahko vsebujejo le znake ASCII, zato pa SGML vsebuje splošen mehanizem za nadomeščanje nizov pri procesiranju dokumenta, ki je uporaben tudi za prečrkovanje. Nize, t.im. entitete SGML, ki naj se ob procesiranju nadomestijo, lahko definiramo sami, v standardu pa so tudi že definirana opisna imena za širok nabor znakov; za zapis črke Č je to npr. `&Ccaron;`, ki se nahaja v tabeli entitet z uradno oznako `ISO 8879:1986//ENTITIES Added Latin 2//EN`. V tem standardu je npr. tudi zapisan sedemjezični korpus projekta MULTEXT-East [14].

Obstaja tudi kar nekaj prosto dostopnih programov, ki znajo prečrkovati besedilo, bodisi med enakovrednimi kodnimi nabori (npr. iz Windows-1250 v ISO 8859-2), ali pa tudi prek ene od prečrkovalnih shem v revnejši nabor. Verjetno največ uporabnikov med njimi ima GNU recode [15]. Na njegovi osnovi je bil v okviru projekta MULTEXT razvit mt-recode [16]. V okviru delovne skupine za pretvorbo med kodiranimi nabori znakov pri združenju TERENA so razvili sistem C3 za pretvorbo med kodiranimi nabori znakov [17].

Povsem spontano pa so se razvile tudi »ročne« prečrkovalne sheme. Pisci, ki zaradi tehničnih nezmožnosti uporabljane opreme ne morejo vnašati naših znakov, si pač pomagajo, kakor vedo in znajo. Najpogosteje tako, da namesto Č, Š in Ž pišejo črke brez kljukic: C, S in Z. Nekateri imajo rajši »telegrafski slog«: CC, SS in ZZ. Spet drugi pišejo raje CX, SX, ZX ali C*, S*, Z*; ali pa uporabijo makrozapis iz LaTeXa: `"C`, `"S`, `"Z`. Bogato zbirko različnih načinov zapisa naših posebnosti najde bralec v [4]. Tovrstne sheme imajo povsem drugačne značilnosti od prej omenjenih strojnih. Medtem ko so pri prvih snovalci pazili, da je postopek prečrkovanja strojno obrnljiv, manj pomembna pa je podobnost nadomestnega niza z izvornim znakom (prim. Č in `=C8`), so druge izbrane tako, da inteligentnega odjemnika, bralca, čim močneje asociirajo na izvorne znake, pri čemer žrtvujejo obrnljivost zapisa (prim. »teza« in »teža« — bralec bo iz konteksta verjetno znal ugotoviti, katera beseda je prava, medtem ko avtomatično to ne gre, saj sta obe besedi slovnično pravilni. Pri posebej nesrečno izbranih primerih — npr. »problem je resen« — pa ima lahko tudi človek težave. S ali š?).

6 ZAKLJUČEK

Izpostavljene probleme lahko povzamemo v treh tematskih sklopih. Ugotovimo lahko, da obstoječi standardi za kodiranje znakov pokrivajo večino potreb pisane slovenščine. Kot problematično lahko omenimo le manjkajoča samoglasnika s krativcem (è, ò), ki ju ni ne v kodiranem naboru znakov ISO 8859-2, ne v nobenem drugem osembitnem naboru znakov z našimi znaki. Razmeroma malo razpoložljivih mest v kodni tabeli je pač inherentna slabost osembitnih naborov znakov. Zaradi te omejitve smo se pri aplikacijah, ki zahtevajo hkraten prikaz večjezičnega besedila (npr. slovenščina-francoščina, slovenščina-ruščina), prisiljeni zanašati na programsko izvedbo simultanega prikaza več naborov na zaslonu.

Internet in z njim povezana odprtost, večja mobilnost informacije ter manjša navezanost pisane beseda na orodje, s katerim je nastala, je še poudaril potrebo po enotnem kodnem naboru, ki olajša izmenjavo informacij. Stanje v Sloveniji zaenkrat še ni takšno, kot bi si ga želeli. Zanimiva je primerjava z drugo tranzicijsko državo, Poljsko, ki je, čeprav je standard ISO 8859-2 pet let za SFRJ nostrificirala kot državni standard PN-93 T-42118, precej doslednejša v njegovem izvajanju. Čeprav naša zakonodaja ne daje pravne podlage za reguliranje tovrstnega nereda, pa bi bilo morda za začetek spodbudno

že, če bi državni organi s svojo prisotnostjo na internetu lahko služili za zgled. Nevzdržno je na primer, da Ministrstvo za znanost in tehnologijo (http://www.mzt.si/) ponuja svoje spletne strani izključno v kodnem naboru Windows-1250.

V nekaj letih lahko pričakujemo uveljavitev novega, šestnajstbitnega standarda ISO 10646/Unicode. Ta standard nam ponuja možnost, da poskrbimo tudi za doslej zapostavljane pismenke slovenskih slovničarjev iz prve polovice 19. stoletja. Narobe bi bilo, če priložnosti ne bi izkoristili.

7 VIRI

[1] R. Czyborra, »The ISO 8859 alphabet soup«, spletna stran. http://czyborra.com/charsets/iso8859.html

[2] International Organization for Standardizaion, »ISO 8859-2:1987 — Information Processing — 8-bit Single Byte Coded Graphic Character Set — Part 2: Latin Alphabet No. 2«, 1987.

[3] J. Toporišič et al., ed., *Slovenski pravopis 1 — Pravila.* DZS, 4. izdaja, 1994.

[4] A. Košir, »Slovenščina in računalniki«, spletna stran. http://nl.ijs.si/gnusl/tex/tslovene/slolang/slolang.html

[5] M. Martinec, »Rešitev problema slovenskih šumnikov v HTML dokumentih«, spletna stran. http://www.ijs.si/doc/www-csz.html

[6] D. Vrtin, »Kako uporabljati slovenske črke v HTML dokumentih?«, spletna stran. http://www1.feri.uni-mb.si/~david/si/index.html-l2

[7] International Organization for Standardization, »ISO/IEC 10646-1:1993 — Information Technology — Universal Multiple-octet Coded Character Set (UCS) — Part 1: Architecture and Basic Multilangual Plane«, 1993.

[8] The Unicode Consortium, *The Unicode Standard: Version 2.0*. Addison-Wesley, 1996.

[9] Bitstream Cyberbit, spletna stran. http://www.bitstream.com/products/world/cyberbits/

[10] Microsoft Typography, spletna stran. http://www.microsoft.com/typography/

[11] Comité Européen de Normalisation, Technical Comittee 304 (CEN/TC304), »Minimum European Subset of ISO/IEC 10646-1, Technical contents of ENV 1973:1995«, spletna stran. http://www.indigo.ie/egt/standards/mes.html

[12] Comité Européen de Normalisation, Technical Comittee 304 (CEN/TC304), »Extended European Subset of ISO/IEC 10646-1, Technical contents of Annex E to ENV 1973:1995«, spletna stran. http://www.indigo.ie/egt/standards/ees.html

[13] International Organization for Standardization, »ISO 8879:1986 — Information Processing — Text and Office Systems — Standard Generalized Markup Language (SGML)«, 1986.

[14] T. Erjavec and C. de Loupy, »The MULTEXT-East Project: Practical experience in multilingual corpus coding and processing«, in *CEN/TC304 Workshop: Providing Multilingual Support for Middleware: Implementing the Universal Character Set ISO 10646 in the European Information Society*, str. 12. November 1996, Bled, Slovenia.

[15] F. Pinard, »GNU Recode«, FTP. ftp://ftp.iro.umontreal.ca/pub/recode/

[16] C. de Loupy, »MtRecode — The MULTEXT Character Translation Program«, spletna stran, 1996. http://www.lpl.univ-aix.fr/projects/multext/MtRecode

[17] Trans-European Research and Education Networking Association (TERENA), Coded Character Set Conversion Task-Force (C3-TF), »The C3 System for Confersion of Coded Character Sets«, spletna stran. http://www.nada.kth.se/i18n/c3/