

## 2. STANJE PRIMJENE KODNIH SUSTAVA U HRVATSKOJ

U zadnjih nekoliko godina već se više puta u našoj zemlji započinjala akcija koja je trebala pridonijeti normiranosti primjene kodnih sustava u informacijskoj tehnologiji. Niti jedna akcija nije do kraja provedena iako su pomogle da se o problemu počne govoriti na kompetentnim razinama.

Jedna od takvih akcija je i izrada dokumenta *Hrvatski grafemi u informatici* koju je na inicijativu Hrvatske asocijacije za softverske tehnologije - *HRAST*, izradila 1993. godine Sekcija za normizaciju.

Uspostavljanjem *Zavoda za normizaciju i mjeriteljstvo* i formalnim osnivanjem *Koordinacijskog odbora za informacijsku tehnologiju* 1994. godine nije se ništa promijenilo. Nedostatak inicijative i potrebnih formalizama u Zavodu za normizaciju i mjeriteljstvo pomogao je da se na području usvajanja hrvatskih normi za informacijsku tehnologiju ništa ne dogodi, odnosno da se zatečeni nered nastavi.

Analizom raspoložive dokumentacije ustanovljeno je da se u Hrvatskoj primjenjuje više različitih kodnih skupova za hrvatske grafeme.

Ustanovljeno je da ovisno o izvoru, postoje dvije osnovne grupe kodnih skupova. Prvu grupu čine standardi što su ih definirale internacionalne ili nacionalne organizacije (ISO, JUS), a drugu grupu čine standardi što su ih definirali proizvođači informatičke opreme (npr. IBM, Microsoft), a koji su zbog masovne upotrebe postali operativni *de facto* standardi.

Prema načinu primjene kodnih oblika postoje tri grupe.

Prvu grupu čine norme koje primjenjuju 7 bitova za kodiranje jednog grafema. Zasnivaju se na standardu **ISO 646:1973**, odnosno na US ASCII 7-bitnom kodnom sustavu i ISO preporuci za nacionalizaciju iz 1973. godine.

U tu grupu spada:

### **HRN I.B1.002:1982**

*Obrada podataka; Skup znakova za razmjenu podataka kodiranih sa 7 bitova....*

Usklađen sa standardom *ISO 646* iz 1973. godine.

Drugu grupu čine standardi koji primjenjuju 8 bitova za kodiranje jednog grafema a temelje se na standardu *ISO 8859-2* iz 1987. godine.

U tu grupu spadaju:

### **HRN I.B1.013:1988**

*Obrada podataka; Skup grafičkih znakova za razmjenu podataka kodiranih jednim bajtom za latinična pisma.*

Usklađen sa standardom *ISO 8859-2* iz 1987. godine.

### **ISO 8859-2:1987**

*Information processing; 8-bit single-byte coded graphic character sets; Part 2: Latin alphabet No. 2.*

Dio (1 do 9) serije *ISO 8859* standarda koja određuje kodne skupove za razne jezične grupe.

### **IBM CP-912**

*IBM-ova konsolidirana verzija kodnog skupa ("Code Page", kodna stranica) koja primjenjuje temeljne postavke o grafemima i kodovima prema normi ISO 8859-2.*

Usklađen sa standardom *ISO 8859-2* iz 1987. godine.

Treću grupu čine standardi koji primjenjuju 8 bitova za kodiranje jednog grafema a ne temelje se ni na jednom standardu kojeg su propisale međunarodne institucije za standardizaciju.

U tu grupu spadaju:

### **IBM CP-870:1989**

IBM-ova verzija kodnog skupa namijenjenog upotrebi u jezičnim skupinama srednje i istočne Europe. Na tržištu je poznatiji po komercijalnim nazivima *IBM Latin-2* ili *Latin-2 - Multilingual/ROECE*.

Ne odgovara standardu *ISO 8859-2*.

### **IBM / Microsoft CP-852:1990**

IBM-ova i Microsoft-ova verzija kodnog skupa koja podržava i hrvatski jezik, a namijenjena je prvenstveno osobnim računalima. Osim uobičajenih grafema sadržava i semigrafičke simbole. Na tržištu je prepoznatljiviji po komercijalnom nazivu *IBM PC Latin-2* i odgovara standardu *Microsoft C- 852*.

Ne odgovara standardima *ISO 8859-2*, *IBM C-870*.

**Microsoft CP-1250**

Microsoftova verzija kodnog skupa namijenjenog prvenstveno osobnim računalima koji rade pod radnim sustavom i skupine Microsoft Windows. (*MS-DOS/Windows 3.1 EE*, MS Windows 95, MS Windows NT) Ne odgovara standardima *ISO 8859-2*, *IBM C-870* niti *IBM/MS CP-852*.

Ovaj pregled i komentar standarda što se primjenjuju u hrvatskoj informatici, treba samo ilustrirati postojeću šarolikost. Poseban je problem što različiti standardi uključuju različite grafeme, pa nije samo pitanje kodnog oblika koji se primjenjuju za hrvatske grafeme nego i raspoloživost ukupnog grafemskog skupa i mogućnost konverzije među kodnim skupovima.

U postupku definiranja latiničnog grafemskog skupa za jezike koji se upotrebljavaju u zemljama srednje i istočne Europe, IBM i ISO nisu se služili istim pravilima, pa zbog toga nisu ni grafemski skupovi jednaki.

Pri usporedbi razmatranih standarda treba promatrati slijedeće karakteristike:

01. Kodni oblik koji je primijenjen za kodiranje hrvatskih palatala.
02. Ukupni skup grafema što ih obuhvaća kodni skup.
03. Grafemi raspoređeni na kodne pozicije 0 do 31 koje su izvorno bile namijenjene kontrolnim funkcijama.
04. Grafemi raspoređeni na kodne pozicije 128 do 159 koje su prema ISO 6429 preporuci namijenjene posebnim kontrolnim funkcijama.

Hrvatski palatali kodirani su ovisno o primijenjenom standardu. U 7-bitnom kodnom sustavu (*HRN I.B1.002:1982*) postavljeni su na zamjenske pozicije što ih je za nacionalizaciju preporučio standard ISO 646. To su kodne pozicije što ih u originalnom *ISO/ASCII* standardu zauzimaju grafemi navedeni u sljedećoj tablici.

HRN	Hex	Dec	ISO/ASCII
ć	7D	125	}
Č	5D	093	]
č	7E	126	~
Č	5E	094	^
đ	7C	124	
Đ	5C	092	\
š	7B	123	{
Š	5B	091	[
ž	60	096	`
Ž	40	064	@

Tablica 1: Kodovi za hrvatske palatale prema ISO 646

Ostali standardi rasporedili su hrvatske palatale na neke druge kodne pozicije. Poznatiji standardi navedeni su u sljedećoj tablici. U zaglavlju tablice namjerno su navedena imena pod kojima su komercijalno poznati pojedini kodni sustavi.

Korisno je uočiti da se često koristi naziv "Latin 2" za kodne tablice koje se međusobno razlikuju. Tako se kodna stranica CP-1250 koja se koristi u programskom proizvodu Microsoft Windows 3.1EE znatno razlikuje od kodne stranice CP-852, a za njih se često neispravno koristi naziv "Latin 2".

Grafemi obuhvaćeni kodnim skupom vrlo su važni iz aspekta primjene u informatici. Taj aspekt se često zanemaruje i razmatra se samo obuhvaćanje skupa latiničnih grafema hrvatske abecede.

Grafem	7-bit ASCII HRN LB1.002		8-bit Latin 2 IBM CP-852		8-bit ANSI ISO 8859-2		8-bit EBCDIC IBM CP-870		8-bit Latin 2 MS CP 1250	
	Hex	Dec	Hex	Dec	Hex	Dec	Hex	Dec	Hex	Dec
é	7D	125	86	134	E6	230	49	073	E6	230
Ć	5D	093	8F	143	C6	198	69	105	C6	198
č	7E	126	9F	159	E8	232	47	071	E8	232
Č	5E	094	AC	172	C8	200	67	103	C8	200
đ	7C	124	D0	208	F0	240	8C	140	F0	240
Đ	5C	092	D1	209	D0	208	AC	172	D0	208
š	7B	123	E7	231	B9	185	9C	156	9A	154
Š	5B	091	E6	230	A9	169	BC	188	8A	138
ž	60	096	A7	167	BE	190	B6	182	9E	158
Ž	40	064	A6	166	AE	174	B8	184	8E	142

Tablica 2: Usporedba kodova za hrvatske palatale

Pisanje teksta uz pomoć računala i pohrana tekstualnih informacija u baze podataka postali su dio poslovnih informacijskih sustava. Zbog toga ne smijemo unaprijed ograničiti tekstualnu informaciju samo na grafeme koji se nalaze u hrvatskoj abecedi, jer se često pojavljuju riječi iz ostalih jezika koje sadrže specifične grafeme npr. francuski grafemi s akcentima, njemački preglašeni samoglasnici ili posebni grafemi koji se upotrebljavaju u slavenskim jezicima.

Uobičajeno je da se skup grafema koji će se upotrebljavati za tekstualne informacije ograniči na latinični skup, pa se eventualna upotreba čiriličnih ili nekih drugih skupova (npr. grčki, hebrejski, japanski) transliterira u latinični skup prema nekom od usvojenih standarda.

Kodni sustavi koji koriste osam bitova ograničeni su na 256 kodnih pozicija, što nužno navodi na kompromis pri izboru ukupnog skupa obuhvaćenih grafema.

Temelj za izbor grafema čini standard ISO 6937-2:1983 *Information processing - Coded character sets for text communication*. U njemu su navedeni svi grafemi što se upotrebljavaju u jezicima koji se služe latiničnim pismom. Standardizirani su: naziv grafema, simbolički grafički oblik i skup grafema koji se upotrebljavaju u pojedinim jezicima. Na taj standard nastavile su se aktivnosti ISO i Unicode organizacija koje su u svojoj konačnoj i prihvaćenoj ediciji (ISO 10646-1, Unicode verzija 2) dokumentirale ukupni grafemski skup, pridruženi kodni sustav te međunarodno prihvaćene nazive grafema.

Naziv grafema posebno je važan jer se sam grafički oblik ne može standardizirati zbog velikog broja pisama upotrebljivanih pri oblikovanju tekstualnih informacija. Naziv koji je standardizirao ISO nije jedinstven, jer je IBM definirao vlastiti standard za nazive, a u upotrebi je standard što ga je uveo ADOBE preko POSTSCRIPT jezika, kao i neki drugi.

ISO i Unicode (konzorcij organizacija) započeli su akciju izrade standarda zasnovanog na kodnom obliku koji rabi više bajtova (okteta). Broj grafema što se može predstaviti 16-bitnim kodnim sustavom ( $2^{16} = 65536$ ) osigurat će istodobnu upotrebu svih grafema primijenjenih u jezicima što se danas upotrebljavaju za uobičajenu komunikaciju.

Rezultat tih napora je ujednačeni prijedlog i poznat je kao ISO 10646-1 odnosno Unicode verzija 2.

Pri izboru grafemskog skupa postignuti kompromis temelji se najčešće na podjeli jezika na grupe. Takav pristup imao je IBM pri izradi svojih kodnih skupova, a isto tako i ISO pri izradi standarda ISO 8859.

ISO standard je jezike grupirao prema vlastitom nahođenju, a sam popis jezika za koje se pojedini kodni skup može upotrebljavati nije dio standarda nego je tek samo dodatna neobavezna informacija.

Prema tako provedenom grupiranju hrvatski jezik se našao u skupini jezika s latiničnim pismom koji se upotrebljavaju u srednjoj i istočnoj Europi (kodni naziv: ISO 8859-2, puni naziv: ISO 8859 *Part 2: Latin alphabet No. 2*).

U standardu ISO 8859-2 su navedene slijedeće zemlje kao natuknica područja primjene (*Annex: Geographical areas of application of the coded character set of this part of ISO 8859 - This annex does not form part of the standard*): Albania, Austria, Czechoslovakia, Germany, Hungary, Poland, Romania, Switzerland, Yugoslavia).

Takva podjela primorala nas je da prihvatimo činjenicu da se uz primjenu kodnog skupa ISO 8859-2 ne mogu prikazati svi grafemi jezika što se danas često upotrebljavaju za poslovno komuniciranje sa zemljama Zapada - npr. francuski.

Time nije riješen problem svih grafema određenih hrvatskom abecedom. Glasovi dž, lj i nj prema našem pravopisu su jedinstveni grafem, dok su u informatici prešutno interpretirani kao dva grafema. Problem sortnog rasporeda za hrvatsku abecedu ostaje stalno otvoren.

Hrvatska abeceda odredila je raspored za slovni skup, ali informatička definicija sortnog rasporeda treba obuhvatiti sve grafeme koji se nalaze u nekom kodnom skupu. To je zadatak koji izlazi iz okvira predviđenog za rad na ovom dokumentu i potreban je zajednički rad stručnjaka za hrvatski jezik i stručnjaka za informatiku.

Konstrukcija standarda *ISO 8859* najbolje je opisana njegovom definicijom navedenom u samom standardu:

*ISO 8859 consists of several parts. Each part specifies a set of up to 191 graphic characters and the coded representation of each of these characters by means of a single 8-bit byte. The use of control functions for the coded representation of composite characters is prohibited by ISO 8859. Each set is intended for use for a group of languages.*

Iako je *ISO 8859* odnosno *ISO 6429* direktno zabranio upotrebu kodnih pozicija namijenjenih kontrolnim funkcijama za kodiranje grafema, ostali standardi to ne poštuju, odnosno odlučili su se na drugačiji kompromis.

Zbog skućenog broja grafema koji se mogu predstaviti u svakom kodnom skupu prema standardu *ISO 8859* (191), ostali standardi su dozvolili dvojaku upotrebu kodnih pozicija namijenjenih kontrolnim funkcijama. Time je ukupan broj raspoloživih kodnih pozicija porastao za 64, što je teoretski omogućilo predstavljanje ukupno 255 grafema.

IBM i Microsoft u svojim standardima kodnih skupova koriste taj kompromis.

Uključenjem Hrvatske u Internet i aktiviranjem komunikacijsko-računalnih mreža *CARNet* i *HPT Internet* problemi su postali još očitiji.

Programska podrška koja se primjenjuje nije ujednačena po pristupu i interpretaciji hrvatskih grafema i primijenjenih kodnih normi, a međunarodna zajednica nema interesa da rješava naše lokalne probleme.

Očiti je problem za upotrebu hrvatskog jezika i primjenu na Internetu postojanje dva direktno nekompatibilna kodna skupa: *ISO 8859-2* i *Microsoft CP-1250*. Uspostavljeni sustav Interneta u svojim dokumentima (npr. *RFC 1345 - Character Mnemonics & Character Sets*) definira upotrebu kodnih skupova. Posebno je važno zamijetiti da je komunikacijski sustav Interneta građen uz pretpostavku da se razmjenjuju podaci kodirani sa 7-bitima te da se ne rabe za kodiranje tekstova kodne pozicije namijenjene kontrolnim funkcijama.

Takav, danas već nesuvremen pristup, utječe na izbor kodnog skupa za primjenu u hrvatskoj informatici.

S druge strane Microsoft je primjenom svoje tehnologije i strategije za višejezičnu komunikaciju pomogao da se u Hrvatskoj može raditi s širokom paletom njegovih proizvoda ne narušavajući integritet i temeljni pravopis našeg jezika. Microsoft je u svojem pristupu kodnog sustava za Windows proizvoda slijedio Unicode preporuke, te je njegov Windows NT jedan od prvih proizvoda na svjetskom tržištu koji dosljedno rabi Unicode odnosno *ISO 10646* norme i omogućuje pisanje višejezičnih dokumenata (npr. *Microsoft Office 97* paket proizvoda). Ne može se i ne smije podcijeniti utjecaj Microsofta na uređivanje primjene kodnih sustava u Hrvatskoj jer su njegovi proizvodi prisutni praktički na svim osobnim računalima.

U svojoj strategiji razvoja i razvoja novih proizvoda Microsoft je u većem broju slučajeva zanemario norme *ISO 8859*. U zadnjim verzijama radnog sustava *MS-DOS* postojala je kodna stranica *CP-912* koja je po svojim temeljnim značajkama odgovarala preporukama *ISO 8859-2*, no ta se akcija nije nastavila i u seriji proizvoda *Microsoft Windows* nema podrške za kodnu normu *ISO 8859-2*.

Nedosljednost u primjeni kodne norme kod Microsofta vidi se po tome da neki komunikacijski proizvodi za npr. elektroničku poštu ili šetanje Internetom podržavaju kodni sustav *ISO 8859-2*.

Stoga je dio ove inicijative Hrvatske informatičke zajednice i prijedlog da se na hrvatskom tržištu izradi programski proizvod koji bi uklapanjem u postojeće *Microsoft Windows* proizvode – prvenstveno *Microsoft Windows 95*, omogućio primjenu kodne norme *ISO 8859-2*. Dio ovog dokumenta je i prijedlog minimalnog skupa funkcija i svojstava koje bi trebao imati takav proizvod.