11 Računalniška obdelava azijskih pisav

Mateja Petrovčič

11.1 Uvod

Danes s sodobnimi operacijskimi sistemi in urejevalniki besedil v nekaj potezah namestimo osnovne načine za vnos azijskih pisav in z nekoliko prakse uspešno tipkamo besedila. A računalniška obdelava besedil v azijskih pisavah ni tako samoumevna, kot se zdi. V 70. letih prejšnjega stoletja je kljub desetletju raziskav veljalo prepričanje, da pisav azijskih jezikov ne bo mogoče vključiti v računalniške sisteme. Glavna težava ni bila v vprašanju, kako naj z nekaj deset tipkami na tipkovnici vnesemo nekaj tisoč pismenk, kot bi s stališča uporabnika verjetno pričakovali, temveč vprašanje, kako naj vsem znakom določimo številčne vrednosti oziroma kodne točke.

Ameriški standardni nabor znakov (*American Standard Code for Information Interchange*, ASCII), ki je bil izhodišče za druge nabore znakov, je s sedmimi biti definiral 2^7 oziroma 128 kodnih točk, od katerih je bilo samo 94 znakov izpisljivih. S to količino znakov je bilo mogoče izpisati le 32 simbolov, 10 števk, 26 velikih in 26 malih tiskanih črk angleške abecede. Razširjeni ASCII je z osmim bitom določil skupno 2^8 oziroma 256 kodnih točk, kar je zadostovalo za različne črkovne pisave, a še to ob uporabi kodnih tabel. Kodna tabela je napreč predpis, ki določa preslikavo vrednosti od 0 do 255 v znake. Vrednosti od 0 do 128 so vedno enake in ustrezajo standardu ASCII, preslikave kodnih točk od 128 do 255 pa se regionalno razlikujejo, kar uporabnik vidi kot različne znake, kot prikazuje Tabela 1.

Tabela 1: *Kodne točke 185, 190 in 232 v različnih kodnih tabelah ISO/IEC 8859.*

Kodna tabela	Pisave za naslednje jezike	Kodna točka		
		185	190	232
ISO/IEC 8859-1 (Latin-1)	zahodnoevropski jeziki	¹	¾	è
ISO/IEC 8859-2 (Latin-2)	srednje- in vzhodnoevropski jeziki	š	ž	č
ISO/IEC 8859-3 (Latin-3)	južnoevropski jeziki	ı		è
ISO/IEC 8859-5 (Latin/ Cyrillic)	cirilica	Й	О	ш
ISO/IEC 8859-6 (Latin/ Arabic)	arabska pisava			و
ISO/IEC 8859-7 (Latin/ Greek)	grška pisava	Ή	΄Υ	θ

Če vsak znak predstavimo z enim bajtom, lahko hkrati uporabimo največ 256 znakov, od katerih je izpisljivih manj kot 190 znakov, ker so določene kodne točke rezervirane za kontrolne znake. To pa še zdaleč ni dovolj za več tisoč znakov, ki se uporabljajo v azijskih pisavah.

Za rešitev tega problema je bilo potrebno izstopiti iz prvotnega okvira razmišljanja, da je vsak znak predstavljen z enim bajtom. Novi pogled na kodiranje znakov je omogočil hiter razvoj informacijske tehnologije v Vzhodni Aziji, pri čemer je nastalo nekaj močnih razvojnih centrov. Kitajska, Tajvan, Hongkong in Koreja so po vzoru Japonske razvili svoje regionalne nabore znakov in načine kodiranja. Čeprav jih Unicode postopoma izpodriva, so izredno pomembna zapuščina.

Preden si podrobneje ogledamo nabore znakov in različne sisteme kodiranja azijskih pisav, si moramo razjasniti, kaj izraz *pisava* pravzaprav pomeni. Zavedati se moramo, da slovenski izraz pisava označuje različne pojme.

V najširšem pomenu besede je pisava (ang. *writing system*) sistem grafične predstavitve elementov določenega jezika. V tej rabi govorimo o *kitajski pisavi*, *japonski pisavi* ali *korejski pisavi*. Vsak od teh sistemov uporablja različne podmnožice pisav (ang. *script*), kar najbolje vidimo na spodnjem primeru iz japonščine.

Legenda:
P pismenke
H hiragana
K katakana
L latinica

K	K	L	H	P	H
マドンナ	の	SP	が	転	んだ。
Madonna	no	esupii	ga	koro	-nda
Madonna	čl.	varnostnik	čl.	pasti	GLAG. OBR.

Madonnin varnostnik je padel.[1]

Pri glagolu »pasti« je uporabljena pismenka, slovnična obrazila so zapisana v hiragani, lastno ime »Madonna« je zapisana v katakani, tujka »SP« (ang. *Security Police*) pa se poslužuje latinice. V tem primeru so torej uporabljene štiri pisave: pismenke *kanji* (漢字), *hiragana* (平仮名), *katakana* (片仮名) in latinica. Kombinacija pisav je značilna tudi za korejščino in kitajščino, vendar v manjši meri.

Nenazadnje je v okviru urejanja besedil pisava, ki ji poljudno pravimo *font*, definirana na naslednji način:

1 Vir primera: Hmeljak Sangawa in drugi (2003, 5).

> **Pisava** je zbirka znakov, ki ima enako črkovno družino in isto velikost. Pisava je na primer helvetica velikosti 12 pik. Helvetica je črkovna družina, velikost pa je 12 pik. […] Nekateri programi imenujejo pisavo zbirko znakov, ki imajo isto črkovno družino, isto velikost in isti slog. Tako sta v tem primeru kurzivna helvetica pri 12 pikah in polkrepka helvetica pri 12 pikah različni pisavi. Drugi programi gledajo na slog pisave kot na dodatno lastnost (Kraynak, 1994, 134).

Za azijske pisave se pogosto uporabljajo črkovne družine *SimSun* (Kitajska), *PMingLiU* (Tajvan), *MS Mincho* (Japonska) in *Batang* (Koreja).

11.2 Oris pisav jezikov Vzhodne Azije

Azijski jeziki uporabljajo mešanico pisav, od katerih si bomo najprej podrobneje ogledali pismenke, nato zlogovni pisavi hiragana in katakana, korejski *hangul* (한글) in njegove gradnike *jamo* (자모; 字母), tajvanski *zhuyin* (注音) ter orisali sisteme transkripcije azijskih pisav. Tabela 2 poda pregled pisav v kitajščini, japonščini in korejščini.

Tabela 2: *Pregled pisav v kitajščini, japonščini in korejščini. Povzeto po Lunde (2008, 5).*

Regija	Kombinacija pisav
Kitajska	poenostavljene pismenke, latinica
Tajvan	tradicionalne pismenke, zhuyin, latinica
Japonska	japonske pismenke (kanji), hiragana, katakana, latinica
Koreja	korejske pismenke (hanja), hangul, jamo, latinica

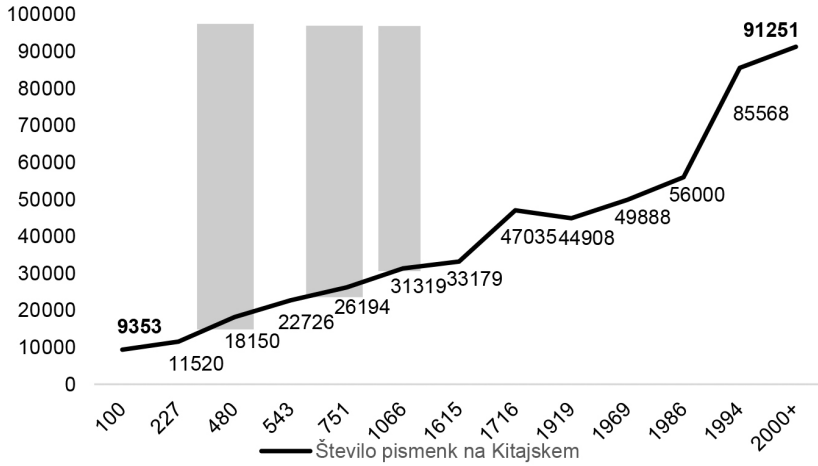
11.2.1 Pismenke

Pismenke so zaradi svoje dolge zgodovine in regionalne razpršenosti zelo heterogena skupina znakov. Ker so nastale na Kitajskem, jih pogosto poimenujemo *kitajske pismenke*. Kot bomo videli v nadaljevanju, se pismenke regionalno razlikujejo, a so zaradi svoje številnosti obravnavane enotno. Unicode jih poimenuje *CJK Unified Ideographs*, pri čemer je CJK kratica za Chinese-Japanese-Korean. V kitajščini jim pravimo *hanzi* (汉字/漢字), v japonščini *kanji* (漢字) in v korejščini *hanja* (한자/漢字) (Lunde, 2008, 60).

V literaturi zasledimo različna mnenja o domnevni prvi uporabi pismenk na Japonskem. Avtorji se načeloma strinjajo, da je kitajska pisava prvič prišla na Japonsko preko Korejskega polotoka v 4. ali 5. stoletju, ko so se tja razširili budistični in filozofski spisi (Halpern, 2001). Drugi val prevzemanja kitajskih pismenk se je

zgodil v letih 618–907, v času razcveta kitajske dinastije Tang. Prevzete pismenke so segale na področje vladajočih krogov in konfucijanske misli. Tretji val je sledil med leti 960 in 1279, v času dinastije Song. Prevzete pismenke tega obdobja so segale na področje zen budizma (Lunde, 2008, 71).

Graf 1 prikazuje število pismenk na Kitajskem (črtni prikaz) in obdobja, ko je Japonska prevzemala kitajske pismenke (barvni pasovi v ozadju). Kot vidimo, se je število pismenk na Kitajskem v dva tisoč letih povečalo za približno desetkrat.[2]



Graf 1: *Okvirno število vseh pismenk na Kitajskem v različnih časovnih obdobjih.*

Japonci so kitajsko pisavo sprva uporabljali za pisanje v kitajščini ali pa so si pismenke sposodili zgolj fonetično za zapis japonskih besed. Zaradi različnega ustroja japonskega in kitajskega jezika zgolj pismenke niso bile ustrezen način zapisovanja japonske pisave. Postopoma se je del pismenk uporabljal pomensko, torej za poimenovanje enakih ali podobnih stvari kot na Kitajskem, del pismenk pa za slovnična obrazila. Slednje so se poenostavljale in od tod sta se razvili zlogovni pisavi hiragana in katakana (Halpern, 2001).

Pomensko prevzete pismenke so imele bodisi prevzeto kitajsko branje (*onyomi* 音読み) bodisi domače japonsko branje (*kunyomi* 訓読み). Poleg tega so določene pismenke prevzeli večkrat, v različnih obdobjih, kar je privedlo še do večjega števila branj za vsako pismenko (Halpern, 2001). Pismenke so načeloma prevzemali v sklopu večzložnih besed, zato je še danes branje pogosto odvisno od tega, v kateri besedi se pismenka nahaja (Lunde, 2008, 72).

2 Graf temelji na podatkih v Lunde (2008, 71), ti pa se sklicujejo na glavna referenčna dela posameznih obdobij.

Koreja je pismenke prevzela še pred Japonsko, a se tam dandanes skorajda ne uporabljajo več. Nadomestila jih je pisava hangul.

Skozi zgodovino se je število kitajskih pismenk večalo, tudi Koreja in Japonska sta na osnovi kitajskih pismenk ustvarjali lastne pismenke (jap. *kokuji* 国字, dob. »nacionalne pismenke«). Na drugi strani je prišlo do poenostavljanja pismenk, modifikacij, številne pismenke so utonile v pozabo ipd. V današnjem času se posledično japonske pismenke v določeni meri razlikujejo od kitajskih, in tudi med pismenkami celinske Kitajske, Tajvana in Hongkonga obstajajo regionalne razlike.

Če štejemo samo pismenke, je za pismenost na Japonskem potrebno poznati okrog 1000 pismenk, na Kitajskem okrog 2500–3500 pismenk, v Koreji pa znanje pismenk ni več predpogoj za pismenost.

11.2.2 Hiragana in katakana

Predpogoj za pismenost na Japonskem sta zlogovni pisavi hiragana in katakana, z eno besedo *kana* (仮名). Hiragana se uporablja za slovnične besede in obrazila, ali pa kot mašilo za zapis japonskih polnopomenk, če ne poznamo pismenk zanje. Katakana je vzporedna zlogovna pisava, ki se uporablja za zapisovanje besed tujega izvora (*gairaigo* 外来語) in onomatopoetskih besed (Lunde, 2008, 53–54). Hiragana in katakana obsegata skupaj 92 znakov.

11.2.3 Hangul

Za zapisovanje korejskega jezika se uporablja hangul. Leta 1443 ga je ustvaril kralj Sejong (世宗), v veljavo pa je stopil leta 1446. Z razliko od kane to ni zlogovna pisava, temveč je vsak znak (kar glasovno soupada z zlogom) sestavljen iz črkovnih gradnikov. Vsak zlog hangula se lahko razstavi na posamezne gradnike jamo, ki predstavljajo samoglasnike in soglasnike. Hangul obsega 24 črk in 27 digrafov.

11.2.4 Zhuyin

Fonetični sistem zhuyin (*zhuyin fuhao* 注音符号) je nastal na začektu 20. stoletja in je metoda označevanja izgovarjave pismenk. Je torej neke vrste polčrkovna transkripcija kitajske pisave, ki se uporablja predvsem na Tajvanu. Prav tako kot hiragana ali katakana se je tudi zhuyin razvil iz poenostavitve pismenk. Obsega 37 znakov, od tega 21 soglasnikov in 16 samoglasnikov ali samoglasniških sklopov.

11.2.5 Sistemi transkripcij azijskih pisav

Pod izrazom transkripcija največkrat razumemo črkovni zapis, čeprav sodi sem tudi zhuyin, ki ni izključno črkovna pisava. V tiskanih publikacijah poznamo več sistemov transkripcij azijskih pisav. Na tem mestu omenimo le najbolj znane, saj je osnovno poznavanje pomembno tudi za razumevanje vnosov azijskih pisav.

Mednarodno standardiziran črkovni zapis kitajske pisave je od leta 1958 *Hanyu pinyin* (汉语拼音), na Tajvanu se uporablja *Tongyong pinyin* (通用拼音), večina pomembnih publikacij s kitajsko tematiko, ki so nastale v letih med 1912 in 1979, se je posluževala sistema *Wade-Giles*, znan je tudi sistem *Yale*, ki obsega transkripcije kitajščine, kantonščine, korejščine in japonščine. Danes se uporablja le za kantonščino (Petrovčič, 2015, 13–15).

Med najbolj znanimi črkovnimi zapisi japonske pisave so: Hepburnov sistem (*hebon shiki* ヘボン式), ki ga je leta 1885 zasnoval ameriški misionar James Curtis Hepburn; sistem Kunrei (*kunrei shiki* 訓令式), ki ga je leta 1937 japonska vlada razglasila za uradni sistem transkripcije; japonski sistem (*nippon shiki* 日本式), ki ga je leta 1881 razvil Tanakadate Aikitsu (田中館愛橘), eden prvih zagovornikov latiničnega zapisa japonskega jezika (predhodnik sistema Kunrei, skoraj identičen, a najmanj rabljen); standardni sistem (*hyōjun shiki* 標準式), popravljena verzija Hepburnove transkripcije iz leta 1908 in nato prenovljen leta 1946; in sistem za urejevalnike besedil (*wāpuro shiki* ワープロ式), ki so ga zasnovali razvijalci japonskih urejevalnikov besedil in sistemov vnosov japonske pisave (Lunde, 2008, 37).

Za črkovni zapis korejske pisave so v veljavi naslednji sistemi: prenovljena transkripcija korejske pisave (*gugeoui romaja pyogibeop* 국어의로마자표기법/國語의로마字表記法) iz leta 2000 v Republiki Koreji; sistem Ministrstva za izobraževanje McCune-Reischauer (*mungyobu* 문교부/文敎部) iz leta 1984 v Demokratični ljudski republiki Koreji; sistem Združenja za korejski jezik (*hangeul hakhoe* 한글학회/한글學會) iz leta 1984; in ISO/TR 11941:1996 iz leta 1996. Obstajajo še drugi sistemi, npr. Yale, Lukoff ali Horne, a niso več zelo razširjeni (Lunde, 2008, 43).

11.3 Nabori znakov

Kot smo videli, je pismenk več deset tisoč. Preden se posvetimo vprašanju, kako je mogoče takemu številu znakov pripisati številčne vrednosti oziroma kodne točke, moramo poznati koncepta *nekodirani* in *kodirani nabor znakov*.

Izraz nekodirani se nanaša na nabor znakov, ki ne odgovori na vprašanje, kako bodo znaki obdelani v računalniških sistemih. Izraz kodirani nabor znakov po drugi strani nakazuje, da gre za zbirko znakov, ki so predvideni za računalniško obdelavo. Odkar so računalničarji ugotovili, kako je to tehnično izvedljivo, so nekodirani nabori znakov podmnožica kodiranih naborov znakov (Lunde, 2008, 79). Zgodovinsko gledano pa so prav nekateri nekodirani nabori znakov služili kot osnova za izdelavo kodiranih naborov.

11.3.1 Nekodirani nabori znakov

Nekodirani nabori znakov se pogosto uporabljajo v izobraževalne namene. Nanje lahko gledamo kot na poskus, kako iz množice pismenk izluščiti tiste, ki so pogosto v rabi. Iz Grafa 1 zgoraj je razvidno, da je samo kitajskih pismenk kumulativno blizu sto tisoč, za vsakdanjo rabo pa jih je dovolj bistveno manj.

Kitajska je leta 1988 objavila seznam pismenk sodobne kitajščine, ki se deli na podkategorije, prikazane v Tabeli 3.

Tabela 3: *Kitajski nekodirani nabori znakov.*

<table>
<tr><td colspan="3">Seznam splošno rabljenih pismenk (7000 pismenk)
Xiandai Hanyu tongyongzi biao (现代汉语通用字表)</td></tr>
<tr><td colspan="2">Seznam pogosto rabljenih pismenk (3500 pismenk)
Xiandai Hanyu changyongzi biao
(现代汉语常用字表)</td><td rowspan="2">preostalih 3500 pismenk sodi med splošno rabljene, a ne pogosto rabljene pismenke</td></tr>
<tr><td>primarne pismenke
(2500 pismenk)</td><td>sekundarne pismenke
(1500 pismenk)</td></tr>
<tr><td>raven osnovne šole</td><td>raven srednje šole</td><td>zunaj šolskega sistema, glede na usmeritev posameznika</td></tr>
</table>

Poleg Seznama splošno rabljenih pismenk in Seznama pogosto rabljenih pismenk je kitajska vlada objavila tudi Seznam poenostavljenih pismenk (*Jianhuazi zongbiao* 简化字总表), ki obsega 2200 pismenk. Če to število primerjamo s številom splošno ali pogosto rabljenih pismenk, vidimo, da je število poenostavljenih pismenk manjše od števila pismenk v omenjenih seznamih. To pomeni, da določen del pismenk ni bil poenostavljen. S poenostavitvami se je število potez v povprečju zmanjšalo za polovico, kar naj bi pripomoglo k boljši pismenosti (Petrovčič, 2012, 172).

Tajvan je v letih 1982–1984 določil svoj obseg pismenk splošne rabe, kot je razvidno iz Tabele 4.

Tabela 4: *Tajvanski nekodirani nabori znakov.*

48.238 pismenk	Seznam pogosto rabljenih pismenk (4808 pismenk)	*Changyong guozi biaozhun ziti biao* (常用國字標準字體表)
	Seznam sekundarnih pismenk (6341 pismenk)	*Ci changyong guozi biaozhun ziti biao* (次常用國字標準字體表)
	Seznam redkih pismenk (18.480 pismenk)	*Hanyong ziti biao* (罕用字體表)
	Seznam različic pismenk (18.609 pismenk)	*Yiti guozi zibiao* (異體國字字表)

Japonska je pismenke razdelila v več naborov, jih spreminjala in jih tudi danes še vedno dopolnjuje. Najnovejši dokument o prenovljeni verziji pogosto rabljenih pismenk je bil objavljen 26. maja 2015.[3] Najožji, izobraževalni nabor natančno določa pismenke za posamezne razrede osnovne šole. Ta nabor pismenk je podmnožica nabora pogosto rabljenih pismenk, ki se pojavljajo v splošni rabi. Poleg tega obstaja še dopolnilni nabor pismenk, ki se jih lahko uporablja v osebnih imenih (Lunde, 2008, 82). Določitev pogosto rabljenih pismenk je v domeni Ministrstva za izobraževanje, kulturo, šport, znanost in tehnologijo, seznam dovoljenih pismenk za osebna imena pa v pristojnosti Ministrstva za pravosodje.

Tabela 5: *Japonski nekodirani nabori znakov.*

<table>
<tr><td colspan="6">Pogosto rabljene pismenke (1945 pismenk)
Jōyō Kanji (常用漢字)</td></tr>
<tr><td colspan="6">Izobraževalni nabor (1006 pismenk)
Kyōiku kanji (教育漢字)</td><td rowspan="3">preostalih 939 pismenk sodi med pogosto rabljene, ki presegajo osnovnošolsko raven</td></tr>
<tr><td colspan="6">število pismenk po posameznih razredih OŠ</td></tr>
<tr><td>1.r
80</td><td>2.r
160</td><td>3.r
200</td><td>4.r
200</td><td>5.r
185</td><td>6.r
181</td></tr>
<tr><td colspan="7">Seznam pismenk za osebna imena (983 pismenk)
Jinmei-yō Kanji (人名用漢字一覧表)</td></tr>
</table>

Koreja je določila nabor pismenk, ki naj bi jih učenci usvojili v letih šolanja. Prvo polovico pismenk naj bi se naučili v srednji šoli, drugo polovico pa na visokošolski ravni. Korejsko vrhovno sodišče je določilo tudi seznam pismenk, ki so sprejemljive za uporabo v osebnih imenih (Lunde, 2008, 84). Poznavanje pismenk pa za govorce korejskega jezika ni bistvenega pomena, saj je vse mogoče zapisati s hangulom.

3 Vir: Kaitei jōyōkanji-hyō' ni kansuru shian (「改定常用漢字表」に関する試案).

Tabela 6: *Korejski nekodirani nabori znakov.*

Izobraževalni nabor (1800 pismenk) *Hanmun Gyoyukyong Gicho Hanja* (한문교육용기초 한자/漢文敎育用基礎漢字)	
900 pismenk na srednješolski ravni	900 pismenk na visokošolski ravni
Seznam pismenk za osebna imena (2964 pismenk) *Inmyeong-yong Hanja* (인명용 한자/人名用漢字)	

11.3.2 Kodirani nabori znakov

Predpogoj za računalniško obdelavo azijskih pisav je bil ustvariti kodirane nabore znakov, kar pomeni, da je bila vsakemu znaku pripisana številčna vrednost. Sprva je vsak proizvajalec računalniške opreme ustvaril svoje standarde, kar pa je posledično pomenilo, da datoteke niso bile berljive na računalnikih drugih proizvajalcev. Za izmenjavo datotek je bilo potrebno ustvariti skupen standard. Prvi državni standard za CJK pisave je nastal na Japonskem leta 1978. Imenoval se je JIS C 6226-1978. Kitajska, Tajvan in Koreja so sledili temu vzoru in z manjšimi ali večjimi odstopanji izdelali podobne kodirane nabore znakov (Lunde, 2008, 85).

Vsi **kitajski** standardi imajo ozako GB, kar je kratica za »državni standard«. Dva najpogosteje rabljena kitajska standarda sta GB 2312-80 in GB 18030-2005. Na prvem je temeljila večina kitajskih internetnih strani, drugega pa je kitajska vlada razglasila za obveznega leta 2005. Vsi izdelki, ki so namenjeni kitajskemu trgu, morajo temeljiti na tem standardu.

Tajvan uporablja večje število znakov kot druge regije, kar je razvidno že iz nekodiranih naborov, predstavljenih v Tabelah 2 in 3. Oblikovanje standardov je vzelo v svoje roke pet velikih podjetij, od koder tudi ime znanega standarda Big Five. Uradni standard na Tajvanu je sicer CNS 11643-2007, ki definira kar 69.134 pismenk oziroma 70.939 znakov, vendar je bil Big5 bistveno bolj razširjen.

Hongkong je z vidika pisave samostojna regija. Za osnovo so vzeli standard Big Five in izdelali regionalne razširitve za pismenke, ki jih Tajvan ne pozna. Takih pismenk je čez 3000. Za razliko od drugih naborov znakov tu pismenke niso razvrščene po nobeni logiki. Hongkonški standardi imajo oznako HKSCS. Z vidika Unicoda se številne pismenke iz tega nabora nahajajo šele v razširitvi B, kar običajno pomeni, da si mora bralec za uspešno prikazovanje znakov dodatno namestiti podporo za to razširitev.

Japonska je uporabljala šest naborov znakov, med katerimi je najbolj poznan JIS X 0208:1997. Obsega 6879 znakov, od tega 6379 pismenk. Je naslednik standarda JIS C 6226-1978, ki je bil osnova za druge standarde.

Najbolj pogost **korejski** standard je KS X 1001:2004 in obsega 8227 znakov, od tega 4888 pismenk in 2350 zlogov hangula. Ker se določene pismenke podvajajo, je dejansko različnih pismenk mnogo manj, in sicer 4620.

Tabela 7: *Najpogostejši azijski regionalni standardi.*

Standard	Število znakov	Zastopane pisave in nabori znakov
GB 2312-80	7445	številke, GB-Roman (kitajska verzija ASCII), hiragana, katakana, grška abeceda, cirilica, pinyin, zhuyin, oznake tabel, pismenke prvega nivoja, pismenke drugega nivoja
Big Five	13.053	okrajšave, različni simboli, merske enote, oznake tabel, števila, latinica, grška abeceda, zhuyin, diakritični znaki, pismenke prvega nivoja, pismenke drugega nivoja
JIS X 0208:1997	6879	ločila, matematični simboli, razni simboli, številke, latinica, hiragana, katakana, grška abeceda, cirilica, oznake tabel, JIS kanji raven 1, JIS kanji raven 2, dodatni kanji
KS X 1001:2004	8227	razni simboli, KS-Roman (korejska verzija ASCII), jamo, številke, grška abeceda, elementi za črte in tabele, okrajšave, posebne oblike črk in jamo, hiragana, katakana, cirilica, hangul, hanja

Vsem standardom je skupno, da obsegajo več pisav. Do neke mere je to tudi razumljivo, saj so se regionalni standardi zgledovali po japonskem sistemu. Podrobnejši pogled razkrije predvsem razlike v razporeditvi pismenk. Spomnimo se, da so nekodirani nabori znakov posredno vplivali na kodirane nabore.

Kitajska je na prvi nivo postavila pogosto rabljene pismenke, na drugi nivo pa druge splošno rabljene pismenke. Pismenke prvega nivoja so razporejene fonetično, pri čemer je osnova za zapis izgovarjave pinyin. Iz tega jasno vidimo, kakšno težo ima pinyin v kitajskem prostoru. Pismenke drugega nivoja so primarno razporejene po radikalih in sekundarno po številu potez.

Tajvanski nabori so vse pismenke razporedili primarno po številu potez in sekundarno po radikalih, kar se v dveh pogledih razlikuje od kitajskih standardov. Po eni strani pinyin ni igral nobene vloge, po drugi strani pa so tudi radikali izgubili na svoji razvrščevalni vrednosti. Poleg tega Big Five ne podpira cirilice. Izdelovalci standarda Big Five so vključili številne različice japonskih, kitajskih in korejskih pismenk, ki so sicer na prvi pogled podobne, a se v podrobnostih razlikujejo (Lunde, 2008, 113).

V **japonskih** standardih so pismenke prve stopnje razvrščene glede na prevzeto kitajsko branje. Če pismenka nima prevzetega branja, je razvrščena glede na domače japonsko branje. Druga raven pismenk je razvrščena po radikalih in nato po številu

potez (Lunde, 2008, 133). Na ta način je razvrščenih 6355 pismenk, kar močno presega število pogosto rabljenih pismenk.

Korejski standardi so posebni v dveh pogledih. Na eni strani obsegajo na tisoče zlogov hangula, na drugi strani pa ima ista pismenka z različnimi branji pripisane različne kode. Razlog za tako število zlogov je v tem, da so celotni zlogi obravnavani kot samostojni znaki, podobno kot pismenke. Glede na to, da vsak zlog zavzema kvadratek prostora, je taka umestitev po svoje razumljiva. Korejski standardi so edini, ki pismenko z različnimi branji definirajo tolikokrat, kolikor je izgovarjav.

11.4 Načini kodiranja

Standardni nabori znakov so tesno povezani z načini kodiranja. Šele ko vemo, koliko znakom lahko pripišemo številčne vrednosti, lahko izdelamo kodirne nabore znakov. Obenem pa se moramo zavedati, da sta nabor znakov in način kodiranja različna pojma. Isti nabor znakov je lahko kodiran na različne načine, kar jasno vidimo na primeru Unicoda. Na primer, UTF-32, UTF-16 in UTF-8 kodirajo isti nabor znakov na tri različne načine. Po drugi strani lahko z istim načinom kodiramo različne nabore znakov. Nam najbližji primer so regionalne različice razširjenega ASCII, česar se starejše generacije verjetno tudi spominjajo. ISO 8859 je razdeljen na 15 delov, pri čemer je prvih 128 vrednosti vedno enakih (ASCII), dodatnih 128 vrednosti, ki jih je omogočil osmi bit, pa je prirejenih za potrebe regionalnih pisav. Slovenski uporabnik je moral izbrati ISO 8859-2, ki ga poznamo tudi kot Latin-2 Central European. Znak z isto kodo v drugi regionalni različici predstavlja drugo črko (Tabela 1).

128 numeričnih vrednosti 7-bitnega bajta zadostuje za zapis angleščine, vendar je to premalo za druge latinične pisave. Ne smemo pozabiti, da je od 128 znakov samo 94 natisljivih. Preostalih 34 znakov je rezerviranih za kontrolne znake. 8-bitni bajt je omogočil skupno 256 znakov, kar je dovolj za druge črkovne pisave, a še zdaleč ne za več tisoč znakov, ki se uporabljajo v azijskih pisavah. Prav zaradi tega dejstva je do 80. let prejšnjega stoletja veljalo, da azijskih pisav ni mogoče kodirati.

Prvi poskus prilagoditve azijskim jezikom so bile azijske različice nabora ASCII. Kitajska različica je bila GB-Roman, tajvanska CNS-Roman, japonska JIS-Roman in korejska KS-Roman. Skupni izraz za ta paket je CJK-Roman. Prav tako kot ASCII tudi ti nabori obsegajo 94 natisljivih znakov (Lunde, 2008, 91). Edina razlika z ASCII je bila v vrednosti znakov »$« in »\«.

Naslednji korak je bil 8-bitni standard JIS X 0201, ki je podpiral le ASCII in s pomočjo osmega bita še polovično katakano oz. katakano polovične širine. Zaradi majhnega števila glifov je bil to še obvladljiv nabor znakov.

Tabela 8: *JIS X 0201, katakana polovične širine (Elias, 2001).*

	8-bitni bajt															
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**A**	**B**	**C**	**D**	**E**	**F**
0	NUL	SOH	STX	ETX	EOT	ENQ	ACK	BEL	BS	HT	LF	VT	FF	CR	SO	SI
1	DLE	DC1	DC2	DC3	DC4	NAK	SYN	ETB	CAN	EM	SUB	ESC	FS	GS	RS	US
2	SP	!	»	#	$	%	&	»	(	)	*	+	,	-	.	/
3	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
4	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5	P	Q	R	S	T	U	V	W	X	Y	Z	[	¥	]	^	_
6	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
7	p	q	r	s	t	u	v	w	x	y	z	{	|	}	¯	DEL
8																
9																
A		。	「	」	、	・	ヲ	ァ	ィ	ゥ	ェ	ォ	ャ	ュ	ョ	ッ
B	ー	ア	イ	ウ	エ	オ	カ	キ	ク	ケ	コ	サ	シ	ス	セ	ソ
C	タ	チ	ツ	テ	ト	ナ	ニ	ヌ	ネ	ノ	ハ	ヒ	フ	ヘ	ホ	マ
D	ミ	ム	メ	モ	ヤ	ユ	ヨ	ラ	リ	ル	レ	ロ	ワ	ン	゙	゚
E																
F																

Pravi pomembnejši preboj se je zgodil leta 1978 z razvojem standarda ISO-2022. To je bilo modalno kodiranje, pri čemer so se določene ubežne sekvence preklapljale med nabori znakov in enobajtnim ali dvobajtnim zapisom. ISO-2022 je na območju natisljivih znakov postavil mrežo velikosti 94x94, pri čemer je nastalo 8.836 celic oziroma kodnih točk, kar je v določeni meri zadostovalo za obdelavo azijskih pisav. Ta način kodiranja je bil uporaben predvsem za izmenjavo informacij, saj je za celotno mrežo zadostovalo že 7 bitov informacij.

ISO-2022 sodi med regionalno neodvisna kodiranja, kar je v praksi pomenilo, da je bila ista struktura uporabljena za različne regionalne različice, kamor sodijo ISO-2022-JP, ISO-2022-CN, ISO-2022-CN-EXT, ISO2022KR ipd.

Drugi pomembni dosežek je bilo kodiranje EUC, saj je s svojo nemodalno strukturo služilo kot osnova za vsa poglavitna nacionalna kodiranja, npr. GBK, GB

18030, Big Five, Big Five Plus, Shift-JIS ali Johab. Razvoj informacijske tehnologije je doživel velik razmah, nastala so številna kodiranja, kar se je upočasnilo šele s pojavom Unicoda. Že prvotna osnovna večjezikovna raven je rešila vprašanje sočasne rabe različnih azijskih pisav, poleg tega pa je Unicode preko pretvorbenih tabel kompatibilen z več lokalnimi standardi. Podatki W3Techs v Tabeli 9 nazorno kažejo prevlado Unicoda nad drugimi kodiranji (Web Technology Surveys, 2016), vendar to ne more izničiti pomena in vrednosti regionalnih kodiranj.

Tabela 9: *Razširjenost načinov kodiranja skozi obdobja (Elias, 2001).*

	1. jan. 2010	1. jan. 2011	1. jan. 2012	1. jan. 2013	1. jan. 2014	1. jan. 2015	1. jan. 2016
UTF-8	50.6 %	59.8 %	68.0 %	74.7 %	78.7 %	82.3 %	85.7 %
ISO-8859-1	28.6 %	22.0 %	17.2 %	13.5 %	10.8 %	9.3 %	7.1 %
Windows-1251	4.3 %	3.7 %	3.3 %	2.8 %	2.7 %	2.2 %	1.9 %
Shift JIS	3.1 %	2.2 %	1.7 %	1.4 %	1.4 %	1.3 %	1.1 %
Windows-1252	3.2 %	2.3 %	1.7 %	1.3 %	1.3 %	1.1 %	1.0 %
GB2312	3.5 %	4.4 %	3.6 %	2.5 %	2.0 %	1.4 %	1.0 %
EUC-KR	0.3 %	0.2 %	0.2 %	0.2 %	0.2 %	0.4 %	0.4 %
EUC-JP	0.7 %	0.5 %	0.4 %	0.4 %	0.4 %	0.3 %	0.3 %
GBK	0.7 %	0.9 %	0.9 %	0.8 %	0.6 %	0.4 %	0.3 %
ISO-8859-2	0.9 %	0.7 %	0.6 %	0.5 %	0.4 %	0.3 %	0.3 %
ISO-8859-15	0.5 %	0.4 %	0.4 %	0.4 %	0.3 %	0.2 %	0.2 %
Windows-1256	1.2 %	1.2 %	0.8 %	0.5 %	0.3 %	0.2 %	0.2 %
Windows-1250	0.4 %	0.3 %	0.3 %	0.2 %	0.2 %	0.2 %	0.2 %
ISO-8859-9	0.7 %	0.5 %	0.3 %	0.3 %	0.2 %	0.2 %	0.1 %
Big5	0.4 %	0.3 %	0.2 %	0.2 %	0.1 %	0.1 %	0.1 %
Windows-1254	0.4 %	0.3 %	0.2 %	0.2 %	0.2 %	0.1 %	0.1 %
Windows-874	0.2 %	0.2 %	0.1 %	0.1 %	0.1 %	0.1 %	0.1 %
US-ASCII	0.2 %	0.1 %	0.1 %	0.1 %	0.1 %	0.1 %	<0.1 %
TIS-620	0.1 %	0.1 %	0.1 %	<0.1 %	<0.1 %	<0.1 %	<0.1 %
Windows-1255	0.1 %	0.1 %	0.1 %	<0.1 %	<0.1 %	<0.1 %	<0.1 %
ISO-8859-7	0.1 %	0.1 %	0.1 %	0.1 %	<0.1 %	<0.1 %	<0.1 %
Windows-1253	0.1 %	0.1 %	<0.1 %	<0.1 %	<0.1 %	<0.1 %	<0.1 %

11.5 Sklepne misli

Sočasna raba različnih pisav ni tako samoumevna, kot se nam danes zdi. Šele konec 80. let prejšnjega stoletja so strokovnjaki našli rešitev, kako več tisoč znakom pripisati

različne številčne vrednosti. Za to je bilo potrebno izstopiti iz prvotnega pojmovanja »en bajt informacij za en znak«. V Aziji je nekaj regionalnih centrov razvijalo svoje rešitve. Tehnične izvedbe posameznih proizvajalcev računalniške opreme so zamenjali nacionalni standardi. Ti so bili znotraj meja posamezne države ali regije kompatibilni, a ne tudi v okviru širšega azijskega prostora. Unicode je rešil številne zagate v zvezi s sočasno rabo različnih pisav, zato uspešno izpodriva preostale regionalne načine kodiranja. Kljub temu pa je poznavanje nacionalnih standardov pomembno, saj odraža neposredno povezavo s tradicionalnimi pogledi na pisavo.

Literatura in viri

Elias, Alexandre, 2001: *Encodings of Japanese.* Dostopno na naslovu: http://www.sljfaq.org/afaq/encodings.html (citirano 16. julij 2015).

Halpern, Jack, 2001: *Outline of Japanese Writing System*. Dostopno na naslovu: http://www.kanji.org/kanji/japanese/writing/outline.htm (citirano 8. avgust 2015).

Historical yearly trends in the usage of character encodings for websites. *Web Technology Surveys.* Dostopno na naslovu: http://w3techs.com/technologies/history_overview/character_encoding/ms/y (citirano 20. avgust 2015).

Jinmei-yō Kanji (人名用漢字一覧表). Dostopno na naslovu: http://www.moj.go.jp/content/001131003.pdf (citirano 20. avgust 2015) ali http://www.sljfaq.org/afaq/jinmeiyou-list.html (628 pismenk) (citirano 20. avgust 2015).

Jōyō Kanji (常用漢字). Dostopno na naslovu: http://www.sljfaq.org/afaq/jouyou-list.html (2134 pismenk) ali http://kokugo.bunka.go.jp/kokugo_nihongo/joho/kijun/naikaku/pdf/joyokanjihyo_20101130.pdf (citirano 20. avgust 2015).

Kaitei jōyōkanji-hyō' ni kansuru shian (「改定常用漢字表」に関する試案). Dostopno na naslovu: http://www.bunka.go.jp/seisaku/bunkashingikai/sokai/sokai_9/49/pdf/kaitei_kanjihyoshian.pdf (citirano 20. avgust 2015).

Kraynak, Joe, 1994: *Računalniški slovar.* Ljubljana: Mladinska knjiga.

Lunde, Ken, 2008: *CJKV Information Processing (2nd edition)*. O'reilly Media.

Petrovčič, Mateja, 2012: Spremembe kitajskega jezika v zadnjih sto letih. V: *Tradicija v objemu modernosti: Stoletje kitajskega preporoda* (ur. Rošker, Jana in Vampelj Suhadolnik, Nataša). Ljubljana: Znanstvena založba Filozofske fakultete. 165–180.

Petrovčič, Mateja, 2015: *Sodobna kitajščina 1*. Ljubljana: Znanstvena založba Filozofske fakultete.

Hmeljak Sangawa, Kristina in drugi, 2003: *Uvod v japonsko pisavo. Hiragana, katakana in prvih 854 pismenk.* Ljubljana: Znanstvena založba Filozofske fakultete.