Mateja PETROVČIČ

Idejni razvoj obravnavanja pisave v petih vzhodnoazijskih regijah z vidika informacijskih tehnologij

Povzetek

Prispevek predstavi razvoj iskanja rešitev, kako definirati pisave azijskih jezikov, da jih bo mogoče računalniško obdelovati. Razširitev 7-bitnega ASCII-ja na 8-bitno različico je pokrila večino črkovnih pisav, vendar to še zdaleč ni zadostovalo za nečrkovne pisave. Enodimenzionalno razmišljanje je ustvarilo možnosti za zgolj nekaj manj kot 200 znakov, za razmah rešitev za azijske pisave pa je bil potreben preskok na dvodimenzionalno razmišljanje. Revolucionarno odkritje Japoncev so prevzeli tudi Kitajska, Tajvan, Hongkong in Koreja, pri čemer so svoje kodirane nabore znakov umestili v identično strukturo in s tem ustvarili lokalne različice istega načina kodiranja. Kmalu zatem je dvodimenzionalno razmišljanje rodilo nove zapuščinske nabore znakov in pripadajoče načine kodiranja. Na Tajvanu so dvema dimenzijama dodali še tretjo, s čimer so ustvarili zelo sistematičen, kompleksen in hkrati fleksibilen pristop h kitajski pisavi. Njegova veličina pa ni nikoli prišla v celoti do izraza, saj ga je zasenčila pojavitev Unicoda, ki je ustvaril nov koncept poenotenih pismenk ter s tem podrl meje med pisavami azijskih regij. Čeprav je Unicode že skoraj v celoti izpodrinil regionalne zapuščinske sisteme, so ti pomembna idejna dediščina, saj je večdimenzionalni način razmišljanja tudi predpogoj in osnova Unicoda.

Ključne besede: pisava, vzhodnoazijski jeziki, kodirani nabori znakov, zapuščinski kodirni sistemi, Unicode

Abstract – Conceptual Development of the Approaches to Writing Systems of the Five East Asian Regions from the Perspective of Information Technologies

This paper presents a search for solutions of how to define Asian writing systems in the scope of information technologies. While the extension of the 7-bit ASCII to the 8-bit ASCII covered most of the alphabetic writing systems, this was far from sufficient for non-alphabetic ones. A one-dimensional way of thinking led to solutions for less than 200 characters, and that was obviously not enough for writing systems of East Asian languages. A switch to the two-dimensional thinking was thus necessary. The first promising solutions were presented by Japanese scholars, and the other East Asian regions adopted their ideas with minor changes. China, Taiwan, Hong Kong and Korea became the new research centres for the next decades. This two-dimensional thinking gave birth to several new character sets and encoding methods. In Taiwan, even a third dimension was added to the previous two, creating a very systematic, complex and flexible approach to the Chinese script. The advantages of this system were never fully expressed, however because the newly emerged Unicode became the leading system. Unicode created a new concept of unifying characters, whereby the distinction of varieties between the scripts of the Asian regions became a secondary question. Although Unicode has almost completely replaced the regional legacy systems, they represent an important conceptual heritage. Their multidimensional way of thinking is nevertheless a prerequisite and the basis of Unicode.

Keywords: font, East Asian languages, encoded character sets, legacy encoding systems, Unicode

1 Pisava in sistem pisave

S konceptom pisave, razvojem različnih sistemov in njihovimi klasifikacijami se je ukvarjalo že veliko raziskav. Med prve obsežnejše študije sodijo Taylor (1883), Diringer (1948), Moorhouse (1953), Gelb (1969) in druge, ki so pomembno prispevale k razumevanju te tematike, vendar so njihovi pogledi do določene mere zastareli. Med sodobnejša dela uvrščamo raziskave, kot so Daniels in Bright (1996), Coulmas (2008), Gnanadesikan (2011), Borgwaldt in Joyce (2013), Daniels (2017) in druge, med deli v slovenskem jeziku pa se posebej na klasificiranje sistemov kitajske in japonske pisave osredotočata Bekeš (1999; 2019) in Hmeljak Sangawa (2019). Primerjava stališč posameznih študij bi bila na tem mestu preobširen zalogaj, zato se omejimo na definicije Konzorcija Unicode, saj je poudarek tega prispevka na obravnavanju pisave z vidika informacijskih tehnologij.

V okviru informacijskih tehnologij termin *sistem pisave* (ang. *writing system*) označuje dva različna koncepta. Po eni strani opisuje splošno načelo, kako posamezne skupine pisav grafično predstavijo izbrani jezik. S tega vidika Konzorcij Unicode sledi klasifikaciji, ki pisave deli na tri kategorije: abecede, zlogovnice in logozlogovnice.

Abecede ali črkovnice (ang. *alphabet*) so sistemi pisave, kjer so osnovni elementi črke, ki se uporabljajo za zapisovanje soglasnikov in samoglasnikov. Nam najbolj znana črkovnica je latinica, ki se z določenimi prilagoditvami uporablja za zapisovanje številnih jezikov. Stopnja ujemanja med glasovi in črkami je ločeno vprašanje, ki ga na tem mestu ne obravnavamo. Sistem pisave, ki zapisuje le soglasnike, se imenuje *abdžad* in ne črkovnica. Nam najbolj znan abdžad je sistem arabske pisave (The Unicode Consortium, The Unicode Standard, Version 11.0.0 2018, 256).[1]

Znaki zlogovnic zapisujejo zloge, kar največkrat pomeni kombinacijo soglasnika/-ov in samoglasnika/-ov. Sem sodita tudi japonski zlogovnici *hiragana* in *katakana*. Na Kitajskem v to skupino spada zlogovnica enega od jezikov Yi.[2] Korejski *hangul* ni niti črkovnica niti zlogovnica, saj so njegovi zlogi sestavljeni iz črk *jamo* (beri *džamo*), ki pa same po sebi niso samostojne enote korejske pisave. Zaradi teh značilnosti ga Unicode imenuje navidezna zlogovnica (The Unicode Consortium, The Unicode Standard, Version 11.0.0 2018, 257).

Sistem kitajske pisave je po naravi logografski, Unicode pa uporablja termin logozlogovnica. S tem izrazom označuje sisteme pisav, kjer najmanjše enote zapisujejo besede in/ali morfeme besed, te pa se lahko uporabljajo tudi kot grob zapis glasovne podobe. Osnovna enota logozlogovnic ima številna poimenovanja, na primer *ideograf, ideogram, logograf, logogram* in *sinogram*, laično pa tudi kar *črka* ali *znak*. Logografi/-mi so po definiciji enote, ki zapisujejo besedo ali morfem, ideografi/-mi pa enote, ki zapisujejo ideje ali koncepte. Meje med morfemi, besedami in koncepti so pogosto zabrisane (The Unicode Consortium, The Unicode Standard, Version 11.0.0 2018, 258).

Kitajske pismenke se v osnovi uporabljajo za zapisovanje kitajskega jezika, v svoje sisteme zapisovanja pa so jih integrirali tudi drugi vzhodnoazijski jeziki. Posamezne regije so po vzoru kitajskih pismenk ustvarile lastne pismenke, ki so v uporabi le na tistem območju. Krovni izraz za pismenke je v kitajščini *hanzi* (beri *handzi*), v japonščini *kanji* (beri *kandži*), v korejščini pa *hanja* (beri *handža*). Unicode, kot bomo videli v nadaljevanju, vse pismenke vrže v en koš

1 Poleg teh obstajajo še *abugide*, pri katerih se s primarnimi znaki zapisujejo soglasniki, ki implicirajo en samoglasnik, ostali samoglasniki pa se zapisujejo s sekundarnimi, dodanimi znaki, ki skupaj z znaki za soglasnike tvorijo skupek, ki zapisuje zlog (Bright 2000; Daniels 2017; Share 2016). Sem spadajo pisave indijske podceline, na primer *devanagari* (Pandey in Jha 2019).

2 Kitajska vlada priznava šest jezikov skupine Yi. Jeziki so med seboj nepovezani, a so v šamanske namene uporabljali skupno pisavo. Tradicionalni sistem pisave je bil logografski, sodobni pa je zlogovni. Zlogovnica Yi je od leta 1980 uradna pisava Severnega Yi.

in zabriše regionalne meje. To skupino znakov poimenuje *han* ali *poenoteni CJK ideografi*[3] (ang. *CJK Unified Ideographs*).

Po drugi strani se izraz *sistem pisave* nanaša na skupek ***pisav*** (ang. *script*), ki se uporabljajo za zapisovanje določenega jezika.[4] S tega vidika *sistem japonske pisave* uporablja štiri pisave, tj. pismenke *han*, *hiragano*, *katakano* in *latinico*[5] (The Unicode Consortium, The Unicode Standard, Version 11.0.0 2018, 256). Poleg teh so, tehnično gledano, dandanes v Vzhodni Aziji v rabi še pisave *bopomofo, hangul, yi, nüshu,*[6] *lisu,*[7] *miao*[8] in *tangut*[9] (The Unicode Consortium, The Unicode Standard, Version 11.0.0 2018, 691). Vizualno gledano je vsem tem pisavam skupno, da so videti kvadratne oblike. Vsaka grafična enota torej zavzema kvadratek prostora.[10]

2 Nekodirani in kodirani nabori znakov

O nekodiranih in kodiranih naborih znakov govorimo predvsem v povezavi s pisavami, ki so sestavljene iz velikega števila elementov. Sem sodijo v prvi

3 CJK je kratica za kitajsko (C), japonsko (J) in korejsko (K). Alternativni izraz je še CJKV, ki doda vietnamščino (V).

4 Izraza *script* ne smemo enačiti z izrazom *nabor znakov* (ang. *character set*), saj slednji označuje množico znakov, ki se načeloma razlikuje od znakov določenega sistema pisave. Če se omejimo zgolj na primer latinice, je nabor znakov ISO/IEC 8859-1 primeren za zapisovanje angleščine, nemščine, italijanščine in številnih drugih jezikov, ki uporabljajo latinico, vendar ne za slovenščino, ker ne vsebuje šumnikov. Po drugi strani je nabor znakov ISO/IEC 8859-2 ustrezen za zapisovanje slovenščine, slovaščine, češčine, madžarščine in še nekaterih drugih jezikov, vključno z angleščino.

5 O latinizaciji lastnih imen v Sloveniji gl. Hmeljak Sangawa 2000.

6 Pisavo *nüshu* so uporabljale ženske v provinci Hunan na jugovzhodu Kitajske. Znaki te pisave izhajajo iz pismenk, vendar pogosto zapisujejo samo glasovno podobo zlogov.

7 Pisava *lisu* je nastala v začetku 20. stoletja za zapis jezika *lisu* iz province Yunnan. Kitajska jo uradno priznava od leta 1992. Sestavljena je iz črk latinice, rotiranih črk latinice ter ločil, ki se uporabljajo za označevanje tonov (The Unicode Consortium, The Unicode Standard, Version 11.0.0 2018, 692).

8 Pisava *miao* je nastala leta 1904 in zapisuje istoimenski jezik iz severovzhodnega dela province Yunnan.

9 Pisava *tangut* zapisuje tangutski jezik, ki je bil od 11. do 16. stoletja v rabi na področju današnje severovzhodne Kitajske. Ponovno so ga odkrili konec 19. stoletja, dandanes pa je predvsem predmet akademskih raziskav (The Unicode Consortium, The Unicode Standard, Version 11.0.0 2018, 693).

10 Zaradi te značilnosti so nabori znakov za azijske pisave ustvarili tudi latinico cele širine, kjer vsaka črka tipografsko zaseda prostor enega kvadratka (jap. 全角ローマ字 *zenkaku roomaji*; kit. 全角字母 *quanjiao zimu*). Kot zgled vzemimo črko A v okviru Unicoda, ki uspešno združuje različne nabore znakov. Osnovna velika tiskana črka A je na kodni točki <U+0041>, njena ustreznica v *celi širini* (A) pa je na kodni točki <U+FF21>.

vrsti sistemi pisav azijskih jezikov z več tisoč pismenkami. Preden pa razjasnimo, na kaj se ta termina nanašata, moramo omeniti dva osnovna pristopa do take gmote pismenk.

Pregled zgodovinskih leksikografskih del pokaže, da želi del gradiva zajeti čim več pismenk, drugi del pa stremi k omejitvi na najbolj ključne pismenke. V prvo kategorijo sodijo normativni slovarji, ki poskušajo iz te množice zajeti čim več pismenk ter tako standardizirati pisavo. Odprto množico pismenk, ki se še vedno povečuje, poskušajo omejiti in izločiti alternativne, neuradne zapise iste pismenke. V drugo kategorijo sodijo različni krajši seznami pismenk, ki poskušajo iz obsežnega nabora znakov izluščiti najpomembnejše pismenke in ustvariti manjše, obvladljive podmnožice pismenk. Več deset tisoč pismenk je namreč za vsakdanjo rabo preveč, saj je med njimi precej zastarelih ali redkih (Zhao in Baldauf 2008, Yong and Peng 2008).

Reforme kitajske pisave segajo daleč v preteklost. V dinastiji Qin (221–207 pr. n. št.) je bila standardizacija pisave ena od ključnih nalog tedanjega časa. Izdelali so seznam 3.500 pismenk, ki naj bi se uporabljale kot uradni standard. V tem okviru je nastalo delo *Cangjiepian* (倉頡篇), ki pomeni poskus reforme kitajske pisave in vzpostavitev pravopisnih standardov za tedanjo malo pečatno pisavo (Zhao in Baldauf 2008, 25). Naslednje večje delo je slovar pismenk[11] *Shuowen jiezi* (说文解字) iz leta 100 n. št., v katerem je zbranih 9.353 pismenk. Tudi tu je določeno, katere pismenke naj bi se uporabljale za pravilen zapis pisave, vključene pa so še pogosto rabljene alternativne oblike pismenk. Slovarji so se skozi zgodovino ves čas dopolnjevali in zajemali vedno več pismenk, vendar so to v vseh primerih še vedno podmnožice obstoječih pismenk. Naslednji slovar, katerega vpliv je viden še dandanes, je *Kangxi zidian* iz leta 1716, ki je definiral 47.035 pismenk.[12] Kot pišeta Zhao in Baldauf (2008, 16), je najobsežnejši slovar do sedaj *Zhonghua zihai* (中华字海) iz leta 1994, ki obsega 85.000 pismenk. Vendar tudi zanj velja, da ni zaključena množica obstoječih pismenk, naj bo še tako obsežen.

Za izboljšanje pismenosti je kitajska vlada v 50. letih prejšnjega stoletja definirala zaključeno množico 7000 pismenk, kar poznamo pod imenom *splošno rabljene pismenke* (*Xiandai Hanyu tongyongzi biao* 现代汉语通用字表). Te se dalje delijo na 3500 *pogosto rabljenih pismenk* (*Xiandai Hanyu changyongzi biao* 现代汉语常用字表), kolikor jih predvidoma pozna oseba s srednješolsko izobrazbo. Seznam se še podrobneje deli na *2500 primarnih*

11 Današnji slovarji se delijo na slovarje pismenk in slovarje besed. V tem prispevku govorimo izključno o slovarjih pismenk, zato se izraz *slovar* v nadaljevanju nanaša nanje.

12 Nanj se opira tudi Unicode, vendar več o tem v nadaljevanju.

pismenk za osnovnošolsko raven in *1000 sekundarnih pismenk* za srednješolsko raven. Poleg tega je kitajska vlada objavila seznam poenostavljenih pismenk (*Jianhuazi zongbiao* 简化字总表), ki obsega 2200 pismenk. Sem sodijo pismenke, ki se grafično razlikujejo od tradicionalnih pismenk, cilj poenostavitve pismenk pa je bilo tudi zmanjšanje števila potez za več kot polovico, kar naj bi še dodatno pripomoglo k večji pismenosti (Zhao in Baldauf 2008, 48).

Tajvan je v letih 1982–1984 pismenke splošne rabe definiral v širšem obsegu. Pogosto rabljenih pismenk je 4.808 (*Changyong guozi biaozhun ziti biao* 常用國字標準字體表), seznam sekundarnih pismenk obsega 6.341 pismenk (*Ci changyong guozi biaozhun ziti biao* 次常用國字標準字體表), redkih pismenk pa je 18.480 (*Hanyong ziti biao* 罕用字體表). Poleg tega so definirali še seznam 18.609 različic pismenk (*Yiti guozi zibiao* 異體國字字表).

Na Japonskem nabor pogosto rabljenih pismenk od reforme leta 2010 obsega 2.136 pismenk (*Jōyō Kanji hyo* 常用漢字表[13]), od katerih jih 1.006 sodi v izobraževalni nabor (Kyōiku kanji 教育漢字), ta pa nadalje točno določa, v katerem razredu osnovne šole naj bi otroci usvojili določene pismenke.[14]

Preostalih 1.130 (od leta 2020 dalje 1.110) pismenk sodi med pogosto rabljene, ki presegajo osnovnošolsko raven. Poleg tega je določen še seznam 863 pismenk za osebna imena (*Jinmei-yō Kanji* 人名用漢字一覧表).[15]

Koreja in Vietnam sta skozi zgodovino prevzemala določene kitajske pismenke, vendar sta kasneje oblikovala svoji pisavi. Koreja je ustvarila pisavo *hangul* ter določila nabor pismenk, ki naj bi jih učenci usvojili v letih šolanja. Izobraževalni nabor tako obsega 1800 pismenk (*Hanmun Gyoyukyong Gicho Hanja* 한문교육용기초 한자/漢文教育用基礎漢字), od katerih naj bi se jih 900 naučili na srednješolski, 900 pa na visokošolski ravni. Korejsko vrhovno sodišče je določilo tudi seznam 2.964 pismenk, ki so sprejemljive za uporabo v osebnih imenih (*Inmyeong-yong Hanja* 인명용 한자/人名用漢字) (Lunde 2008, 84). Poznavanje pismenk pa za govorce korejskega jezika ni bistvenega pomena, saj dandanes večinoma vse zapisujejo s hangulom.

Vse zgoraj omenjene množice pismenk so nekodirani nabori znakov, torej podmnožice, ki so nastale izven okvira informacijskih tehnologij, neodvisno

13 Pridobljeno s strani Agencije za kulturo Ministrstva za izobrazbo, kulturo, šport, znanost in tehnologijo (Jōyōkanji-hyō 常用漢字表 [Seznam pogosto rabljenih pismenk] 2010).

14 S prenovo učnih načrtov, ki bo stopila v veljavo leta 2020, bodo v nabor za osnovno šolo vključili 20 pismenk, ki se uporabljajo za zapis imen prefektur, tako da bo v tem naboru 1026 pismenk. Tudi vrstni red usvajanja bo nekoliko spremenjen (Monbukagakushō 2017).

15 Pridobljeno s strani Ministrstva za pravosodje (Hōmushō 2017).

od premisleka, ali jih bomo lahko računalniško obdelovali. Poznavanje nekodiranih naborov znakov je pomembno za razumevanje kodiranih naborov znakov, saj so ti informatikom služili kot osnova za izdelavo kodiranih naborov znakov.

Izraz kodirani nabor znakov torej nakazuje, da gre za zbirko znakov, ki so predvideni za računalniško obdelavo. Vsak znak mora imeti svojo *kodno točko*, torej unikatno numerično vrednost. To je ključnega pomena za razumevanje problematike pisav azijskih jezikov. Zasnova računalnikov namreč ni jezikovno neodvisna, temveč se navezuje na angleščino in sistem angleške pisave, kar razberemo že iz zgradbe 8-bitnega bajta, ASCII-ja, zasnove tipkovnice in podobno. Kot bomo videli v nadaljevanju, se tu odraža tudi enodimenzionalni način razmišljanja, ki pa ni bil primeren za sisteme pisav azijskih jezikov.

3 Enodimenzionalni pristop

Amerika, zibelka računalniškega razvoja, je z ASCII-jem (ang. *American Standard Code for Information Interchange*) postavila standarde za kodiranje znakov. Prvotni, 7-bitni ASCII je s sedmimi biti definiral 2^7 oziroma 128 kodnih točk. To je zadostovalo za 33 kontrolnih znakov, presledek in 94 izpisljivih znakov. S to količino kodnih točk je bilo mogoče definirati 26 velikih tiskanih črk, 26 malih tiskanih črk angleške abecede, 10 števk ter 32 drugih pogostih znakov, kamor sodijo na primer ločila in matematični operatorji. To so obenem znaki tipične angleške tipkovnice.

Razmišljanje je povsem enodimenzionalno. Z enim bitom ustvarimo 2 kombinaciji (2^1), z dvema bitoma 4 kombinacije (2^2), s tremi biti 8 kombinacij (2^3) in tako naprej. Sedem bitov je zadostovalo za angleško pisavo, a ne za črkovne pisave drugih jezikov, ki uporabljajo znake, ki jih angleška abeceda ne pozna, na primer *č, š, ž, Č, Š, Ž* za slovenščino, *ä, ö ü, ß* za nemščino, *é, è, ê, ë, æ, œ, ç* itd. za francoščino in podobno.

Osmi bit je omogočil dodatnih 128 kodnih točk, kar je zadostovalo za večino črkovnih pisav. Črke angleške abecede so ostale na istih kodnih točkah, lokalno specifične črke pa so bile na vrednostih od 161 do 255 (decimalni zapis). Edina težava je bila, da je bilo tudi novih 94 mest premalo za vse posebne črke vseh pisav. V okviru standardov ISO/IEC 8859 je tako nastalo 15 delov oziroma različic, ki so se uporabljale v različnih regijah in so pokrivale pisave določene skupine jezikov.

Tabela 1: Primerjava kodnih točk 185, 232 in 248 v petih regionalnih različicah standarda ISO/IEC 8859.

ISO/IEC 8859	ime	primerno za pisave naslednjih jezikov	185	232	248
ISO/IEC 8859-1	Latin-1, zahodnoevropski	angleščina, nemščina, islandščina, italijanščina, portugalščina …	¹	è	ø
ISO/IEC 8859-2	Latin-2, srednjeevropski	slovenščina, slovaščina, madžarščina, poljščina …	š	č	ř
ISO/IEC 8859-3	Latin-3, južnoevropski	turščina, malteščina, esperanto	ı	è	ĝ
ISO/IEC 8859-5	Latin/cirilica	bolgarščina, makedonščina, ruščina, beloruščina …	Й	ш	j
ISO/IEC 8859-7	Latin/grščina	sodobna grščina	Ή	θ	ψ

V praksi je to pomenilo, da sta slovenski in slovaški uporabnik besedo *češnja* videla enako, nemškemu ali angleškemu uporabniku se je ta beseda prikazala kot *èe¹nja*, v Bolgariji so videli zapis *шеЙnja*, v Grčiji pa se je na tem mestu izpisala *ϑeΉnja*. V enostavnih urejevalnikih besedil (na primer Notepad++) lahko preklapljamo med različnimi kodiranji in opazujemo razlike med nabori znakov. Opazili bomo, da se bo angleški pangram *the quick brown fox jumps over the lazy dog* v vseh kodiranjih prikazoval pravilno. Slovenski pangram *v kožuščku hudobnega fanta stopiclja mizar in kliče* bo popačen samo pri šumnikih. Za nemščino lahko uporabimo *Victor jagt zwölf Boxkämpfer quer über den großen Sylter Deich.* Za prikaz estonske pisave je primeren *see väike mölder jõuab rongile hüpata* in tako naprej.[16]

16 Za pangrame drugih **črkovnih pisav** glej na primer http://clagnut.com/blog/2380/. Pangrami za **zlogovnice** so nekoliko daljši, a še vedno obvladljivi. Za hangul imamo primer 밤새컴퓨터로요약을해치우면좋겠다 (*BamSae KumPyooTuhRo YoYakEul HaeChiWooMyun JotGetDa*). Za hiragano lahko uporabimo とりなくこゑす ゆめさませ みよあけわたる ひんかしを そらいろはえて おきつへに ほふねむれゐぬ もやのうち (*torinakukowesu yumesamase miyoakewataru hinkashiwo sorairohaete okitsuheni hofunemurewinu moyanōchi*) (gl. https://camtsmith.com/articles/2016-11/pangrams) ali pangram, ki je predstavljen v Hmeljak Sangawa (2019).
Pangrami za kitajščino ne obstajajo, ideja pangrama pa je bila prisotna že v 6. stoletju, ko so v besedilu *Qianziwen* (千字文) uporabili 1000 tedaj pogosto rabljenih pismenk. To delo so se morali otroci naučiti na pamet. Podobno besedilo z naslovom *Sanzijing* (三字经) je nastalo v 13. stoletju, v dinastijo Song (960–1279) pa datira še besedilo *Baijiaxing* (百家姓), v katerem je v verze stkanih 472 priimkov. Vsa tri dela so znana pod skupnim imenom *San-bai-qian* (三百千) in so bila učno gradivo do približno leta 1930. Celotna besedila so na voljo na spletni strani https://baike.baidu.com/item/三百千.

Prvi poskus prilagoditve azijskim jezikom so bile azijske različice nabora ASCII. Kitajska različica se je imenovala GB-Roman, tajvanska CNS-Roman, japonska JIS-Roman in korejska KS-Roman. Prav tako kot ASCII tudi ti nabori obsegajo 94 natisljivih znakov. Edina razlika je bila v vrednosti znakov »$« in »\« (Lunde 2008, 91). Te rešitve skoraj niso omembe vredne, ker so spremenili le glif enega znaka.

Naslednji korak je bil 8-bitni standard JIS X 0201, ki je kodne točke osmega bita spremenil v katakano polovične širine. Zgoraj omenjena *češnja* bi se v okviru tega standarda prikazala kot ･ *eｶnja*. Za eno zlogovnico je bilo dovolj kodnih točk, za drugo pa je že zmanjkalo prostora, kaj šele, da bi sem vključili tudi pismenke.

Z linearnim pristopom torej z osmimi biti ustvarimo 256 kodnih točk, kar je občutno premalo za pisave azijskih jezikov. Tabela 2 prikazuje, koliko bi znašale maksimalne vrednosti različno dolgih nizov.

Tabela 2: Največja vrednost v binarnem in decimalnem zapisu glede na število bitov.

Število bitov	skupno število kodnih točk	binarni zapis	decimalni zapis
7	128	1111111	127
8	256	11111111	255
9	512	111111111	511
10	1024	1111111111	1023
11	2048	11111111111	2047
12	4096	111111111111	4095
13	8192	1111111111111	8191
14	16384	11111111111111	16383
15	32768	111111111111111	32767
16	65536	1111111111111111	65535
17	131072	11111111111111111	131071

Za postavitev 7000 znakov, kolikor je splošno rabljenih pismenk, bi potrebovali 13-bitne bajte. Ker je dolžina bajta, torej najmanjšega nosilca informacije, stvar dogovora, bi – teoretično gledano – azijski računalniki lahko delovali na osnovi daljših bajtov. Verjetno bi se kmalu pojavilo vprašanje, kako dolg naj bo bajt, da bo kodnih točk res zadosti. Če bi želeli računalniško prikazati vsebino slovarja *Zhonghua zihai* s 85.000 pismenkami, bi potrebovali 17-bitne bajte. To bi bilo sicer izvedljivo, vendar se je po drugi strani že leta 1964 uveljavila konvencija, da je en bajt skupek osmih bitov (Internet History 1962 to 1992).

Na tej točki je bilo videti, da pisav azijskih jezikov ne bo mogoče računalniško obdelovati. Do prve prave rešitve so prišli Japonci leta 1978 z razvojem standarda ISO-2022. Dandanes, ko Unicode 12.0[17] definira že več kot 100.000 pismenk, vprašanje azijskih pisav ni več tako problematično, vendar ni rečeno, da bi do današnjih rešitev sploh prišlo, če ne bi bilo vmes idejnega preklopa iz enodimenzionalnega na dvodimenzionalno razmišljanje.

4 Dvodimenzionalni pristop

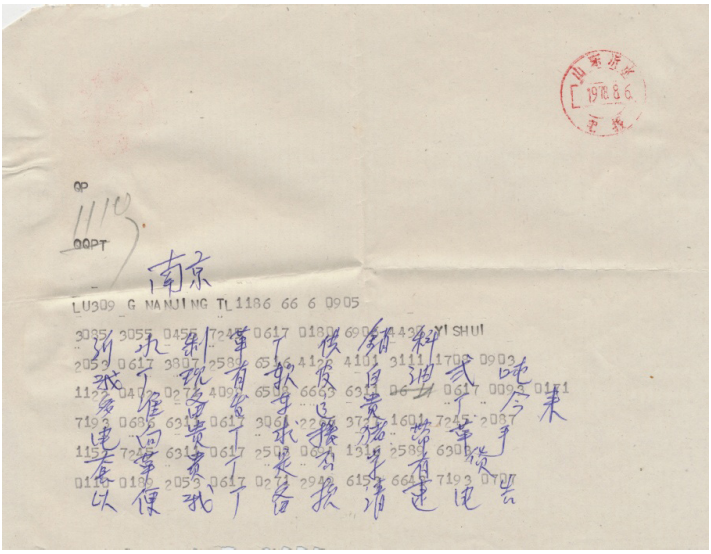
Osnove dvodimenzionalnega načina razmišljanja so bile nevede prisotne že v kitajskem poštnem sistemu. Skoraj neverjetno je, da je do prve računalniške rešitve preteklo toliko časa, saj so bili zametki rešitev že dani.

Podobno kot je ASCII zadoščal za izmenjavo angleškega besedila v okviru informacijskih tehnologij, je pred tem Morsejeva abeceda zadoščala za prenos angleškega besedila prek telegrafije. In podobno kot je bil ASCII premalo za potrebe azijskih pisav, tudi Morsejevi znaki niso zadoščali za azijske jezike. Sporočila bi načeloma lahko prenašali prek glasovne podobe, vendar bi tudi to zaradi enakozvočnic, črkovnega zapisa jezika, narečnih razlik in dolžine besedil prineslo veliko težav. Rešitev za nedvoumen prenos pismenk je bila objavljena leta 1881, ko je Zheng Guanying (鄭觀應) objavil priročnik *Dianbao xinbian* (电报新编) (Mair 2015).

Vsaka pismenka je imela 4-mestno kodo, od 0000 do 9999. Na vsaki strani priročnika je bila tabela velikosti 10 x 10, kar pomeni, da je bilo na eni strani 100 pismenk. Prvi dve cifri sta predstavljali stran v priročniku, tretja cifra je pomenila oznako vrstice, v kateri je pismenka bila, četrta cifra pa je pomenila oznako stolpca, v katerem je pismenka bila. Na primer, pismenki *zhongwen* 中文 imata kodo 0022 2429, kar pomeni, da je pismenka *zhong* 中 v drugem stolpcu druge vrstice na strani 00, pismenka *wen* 文 pa je na strani 24, v drugi vrstici, devetem stolpcu.

Delo telegrafistov je bilo večstopenjsko. Poštni uradnik je pismenke sporočila najprej s pomočjo telegrafskega priročnika pretvoril v štirimestne kode, te je nato pretvoril v Morsejevo abecedo in sporočilo telegrafiral poštnemu uradniku na drugi strani. Ta je sprejete Morsejeve znake zapisal kot neprekinjeno gmoto cifer, jih razdelil na štirimestne sklope ter nazadnje s pomočjo telegrafskega priročnika štirimestne kode ponovno pretvoril v pismenke. Slika 1 prikazuje primer telegrama:

17 Objavljen je bil 5. marca 2019. Verzija 13 je načrtovana za marec 2020.

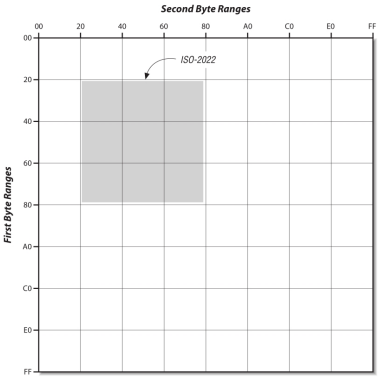


Slika 1: Primer kitajskega telegrama (Mair 2015).

Leta 1885 je telegrafski sistem od Kitajske prevzela tudi Koreja, vključno s pismenkami. Korejski telegrami so bili pisani bodisi v pismenkah bodisi v črkah angleške abecede, ne pa v hangulu (Tomokiyo 2014).

10.000 različnih telegrafskih kod torej ne razumemo linearno oziroma enodimenzionalno, kot bi šteli od 0 do 9999, temveč ploščinsko oziroma dvodimenzionalno. Pismenke so v mreže 10 x 10 razvrstili po slovarskih načelih. Za osnovo je služilo 214 radikalov iz slovarja *Kangxi zidian*. Radikali z manj potezami so bili uvrščeni pred radikali z več potezami, znotraj posameznega radikala pa so bile pismenke zopet razvrščene po številu in obliki potez. Priročniki s kitajskimi telegrafskimi kodami so še dandanes v uporabi, na voljo so tudi na spletu, pri čemer med kitajsko, tajvansko in hongkonško verzijo obstajajo določene razlike (Biaozhun dianmaben (Zhongwen shangyong dianma) [標準電碼本(中文商用電碼)] 2004-2018).

Kot je bilo omenjeno na začetku tega poglavja, so rešitve glede kodiranja prvi našli Japonci leta 1978 z razvojem standarda ISO-2022. Različne nabore znakov so razvrstili v mrežo dimenzije 94 x 94. To je namreč število izpisljivih znakov v okviru ASCII-ja. Na ta način so ustvarili 8.836 kodnih točk. Slika 2 prikazuje območje mreže, kamor so umestili kodirane nabore znakov. Oznake osi uporabljajo šestnajstiški zapis, kar pomeni, da je desetiška vrednost 128 tu prikazana kot 80, vrednost 255 pa kot FF.



Slika 2: Območje kodnih točk v kodiranju ISO-2022 (Lunde 2008, 231).

Iz Slike 2 je razvidno, da so vsi znaki v rangu prvih sedmih bitov in da je osmi bit neizkoriščen. To pomeni, da je bil ta sistem kodiranja zelo priročen za izmenjavo informacij med računalniki. Ker pa se je območje mreže prekrivalo s črkami angleške abecede, je moral obstajati sistem preklapljanja med enobajtnim in dvobajtnim procesiranjem podatkov. Modalna kodiranja, kamor sodi tudi ISO-2022, to rešujejo z ubežnimi sekvencami ali drugimi posebnimi znaki, ki nakazujejo preklapljanje med nabori znakov ali različnimi verzijami istega nabora znakov (Lunde 2008, 195). Ker bi bilo spuščanje v podrobnosti za namen tega prispevka preveč kompleksno in dolgovezno, naj uporabim enostavnejšo primerjavo. Predstavljajmo si, da so nizi enic in ničel tirnice, procesor pa je vlak, ki potuje po njih. Položaj kretnic (izbrana ubežna sekvenca) ga usmeri na prvi tir (enobajtno branje) ali drugi tir (dvobajtno branje). Na koncu odseka so znova kretnice (ubežne sekvence), ki vlak preusmerijo v novo smer. Na ta način je bilo možno s 7-bitnim bajtom ustvariti 128 + 8.836 kodnih točk.[18]

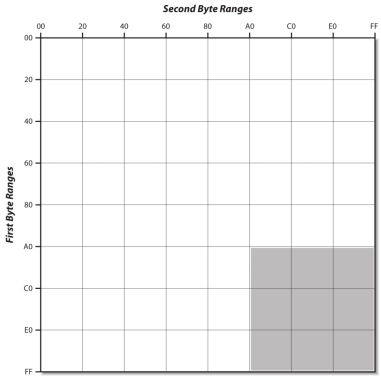
Japonska izvedba ISO-2022-JP je lahko v toliko kodnih točk zajela več naborov znakov: ASCII, JIS-Roman (oziroma japonsko različico ASCII), JIS X 0208 (tu so bili posebni znaki, cifre, latinica, hiragana, katakana, grška abeceda, cirilica, oznake tabel ipd.) in JIS X 0208-1983 (razširitve iz leta 1983).

Te rešitve so nato prevzeli še v drugih regijah Vzhodne Azije. Kitajska je v svoji različici ISO-2022-CN seveda ohranila ASCII, na mrežo 94 x 94 pa umestila nabore znakov GB 2312-80 (GB-Roman oziroma kitajsko različico ASCII, hiragano, katakano, grško abecedo, cirilico, pinyin, bopomofo, 6.763 pismenk, posebne znake, oznake tabel ipd.) in prvi dve ravni tajvanskega standarda CNS 11643-1992.

18 7-bitnega bajta se je kasneje posluževal še UTF-7, ki pa danes ni več v rabi.

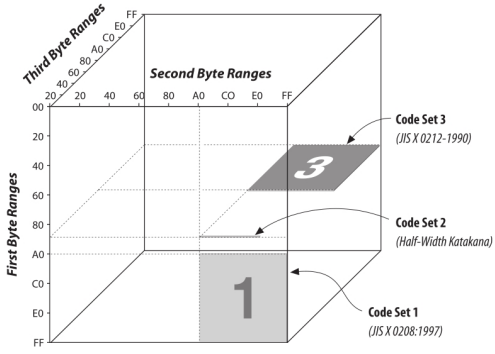
V razširjeno izvedbo ISO-2022-CN-EXT so dodali še ravni 3 do 7 tega tajvanskega standarda (Lunde 2008, 229-230). Tudi Korejci so izdelali svojo različico ISO-2022-KR.

Ko je bil prvi korak v dvodimenzionalno razmišljanje storjen, ni bilo treba več veliko do prehoda na nemodalno kodiranje, ki za preklop med eno-, dvo- ali večbajtnim procesiranjem uporablja numerične vrednosti kodnih točk. Na osnovi kodiranja ISO-2022 je nastalo kodiranje EUC (*Extended Unix Code*). Slika 3 prikazuje položaj dvobajtne mreže v kodiranjih EUC. Vsaka regija je v to območje tudi tokrat postavila svoje nabore znakov.



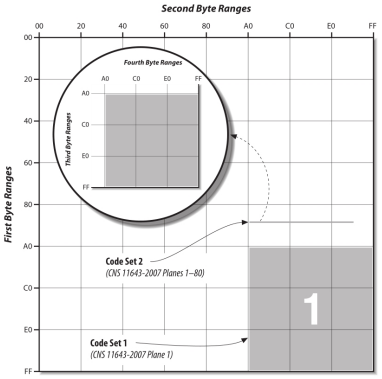
Slika 3: Dvobajtno območje v okviru kodiranja EUC (prirejeno po Lunde 2008, 246).

Japonci so dodatne nabore znakov podaljšali na tri bajte, kar je bil že zametek tridimenzionalnega pristopa h kodiranju. V tretjo dimenzijo oziroma 3-bajtno branje je sicer vodila ena sama točka prvega bajta, tj. 0x8F. Slika 4 prikazuje japonsko različico EUC-JP.



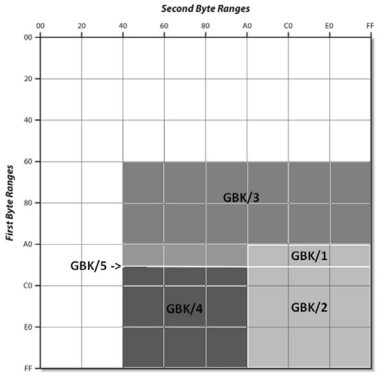
Slika 4: Dvo- in tribajtno območje v okviru kodiranja EUC-JP (Lunde 2008, 250).

Tajvanci so večbajtne znake umestili na podobno območje kot Japonci, le da so bili znaki 4-bajtni. Ustvarili so 80 ravni, kot je okvirno razvidno iz Slike 5. Območje 4-bajtnih znakov je tam, kjer je označena tanka siva črta. Razširitev tega sistema se je uporabljala tudi v Hongkongu, ki je praktično prevzel tajvanske rešitve. Obe regiji namreč uporabljata tradicionalne pismenke, pri čemer je Hongkong moral dodati nekatere lokalno specifične pismenke. V nasprotju s Tajvanom, ki je pismenke uredil po skrbno premišljenih načelih, so pismenke v razširjenem naboru znakov HKSCS (Hong Kong Supplementary Character Set) dodane brez posebnega sistema.



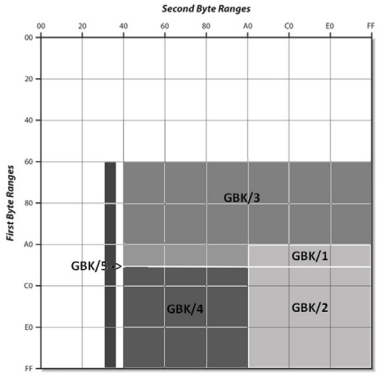
Slika 5: Dvo- in štiribajtno območje v okviru kodiranja EUC-TW (Lunde 2008, 248).

Poleg različic sistemov ISO-2022 in EUC, ki so se uporabljale v več azijskih regijah, so posamezne regije ustvarile lastne, regionalno pogojene sisteme kodiranja, ki pa niso prinesli korenito drugačnih pristopov. Za primer vzemimo Kitajsko. Sistem kodiranja GBK je izhajal iz ISO-2022-CN. Slika 6 prikazuje, v katere smeri so se širili za pridobitev novih kodnih točk.



Slika 6: Območje kodnih točk v sistemu kodiranja GBK.

Tudi standard GB18030, ki ga morajo od leta 2006 dalje podpirati vse naprave, namenjene kitajskemu trgu, je sistemu kodiranja GBK le še dodal nov pas kodnih točk, kot prikazuje Slika 7.

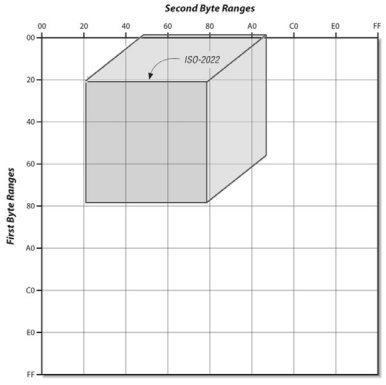


Slika 7: Območje kodnih točk v sistemu kodiranja GB18030.

5 Tridimenzionalni pristop

Najkompleksnejši pristop k strukturiranju informacij in razporeditvi pismenk se kaže v naboru znakov CCCII (Chinese Character Code for Information Interchange 中文資訊交換碼). Prva verzija datira v leto 1980, popravki pa v leti 1982 in 1987 (Lunde 2008, 122).

CCCII temelji na kodiranju ISO 2022 in je razdeljen na 16 slojev (ang. *layer*). Vsak sloj razen zadnjega se dalje deli na 6 ravni (ang. *plane*), kar je skupno 94 ploskev. Ko to združimo s konceptom mreže 94 x 94, nastane kocka dimenzij 94 x 94 x 94, kot prikazuje Slika 8.



Slika 8: Struktura kodiranja CCCII.

Vsak sloj je namenjen določenemu tipu znakov, kot prikazuje Tabela 3. Tabela 4 nato še natančneje prikaže strukturo prvega sloja.

Tabela 3: Kategorije znakov po slojih v kodiranju CCCII (Lunde 2008, 122).

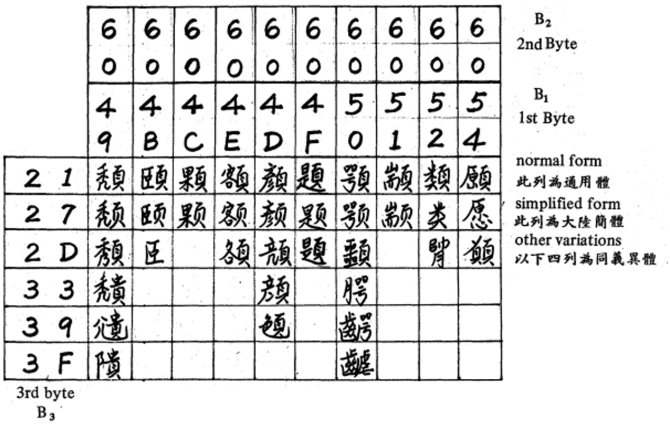
sloj	raven	tip znakov
1	**1–6**	**znaki, ki niso pismenke, in tradicionalne pismenke**
2	7–12	poenostavljene pismenke
3–12	13–72	različice pismenk s sloja 1
13	73–78	hiragana, katakana in japonske pismenke *kanji*
14	79–84	jamo, hangul in korejske pismenke
15	85–90	rezervirano območje
16	91–94	ostali znaki

Tabela 4: Struktura prvega sloja po ravneh v kodiranju CCCII (prirejeno po Lunde 2008, 122).

raven	vrstica (dec)	število znakov	tip znakov
1	1		rezervirano za kontrolne znake
1	2–3		
1	4–10	0	nedodeljeno
1	11	35	kitajska ločila
1	12–14	214	214 Kangxijevih radikalov
1	15	78	številke in zhuyin (bopomofo)
1	16–67	4.808	pogosto rabljene pismenke[19]
1–3	68_1–64_3	17.032	sekundarne pismenke
3–6	65_3–5_6	20.583	ostale pismenke
6	6–94	0	nedodeljeno

Kocka dimenzij 94 x 94 x 94 ponuja 830.584 kodnih točk, kar je skoraj desetkrat toliko, kolikor pismenk obsega najobširnejši kitajski slovar. Tako lahko ugotovimo, da je veliko kodnih točk praznih. Izjemnost te rešitve se kaže predvsem v medsebojnem odnosu slojev in posledično povezanosti pismenk. Kot prikazuje Slika 9, imajo različice iste pismenke enako vrednost prvih dveh bajtov, razlikujejo pa se v tretjem bajtu.

19 Glej poglavje 0 *2 Nekodirani in kodirani nabori znakov*, odstavek o Tajvanu.



Slika 9: Kodne točke sorodnih pismenk v kodiranju CCCII (Hsieh in drugi 1981, 135).

Ker CCCII deluje v okviru 7-bitnih bajtov, je primeren za bibliotekarske sisteme. Kongresna knjižnica Amerike ga je priredila v EACC (East Asia Coded Character). Na spletnih straneh Kongresne knjižnice je na voljo seznam kodnih točk za 13.478 pismenk (Code Table East Asian Ideographs ('Han') 2007). Iz vrednosti kodne točke lahko razberemo, ali je določena pismenka primarna tajvanska pismenka (x1xxxx), poenostavljena pismenka (x7xxxx) ali ena od različic te pismenke (NNxxxx). Oglejmo si primer v Tabeli 5.

Tabela 5: Primer umestitve uradne pismenke in njenih različic v sistemu EACC.

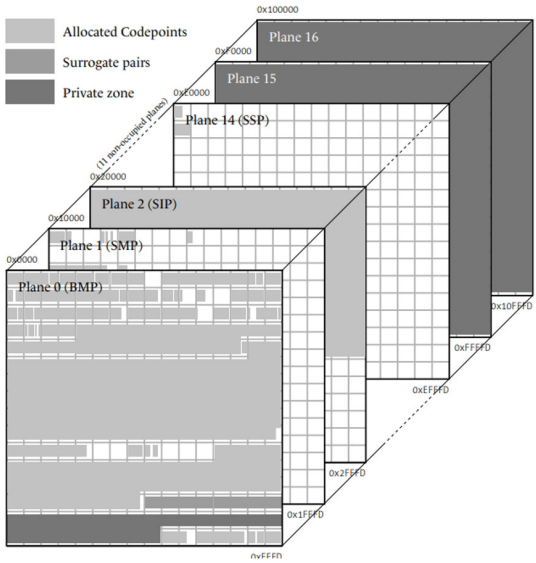
kodna točka	pismenka	raba
213421	劍	uradna tradicionalna pismenka na Tajvanu
273421	剑	poenostavljena pismenka
2D3421	劒	neuradna različica
333421	釼	neuradna različica
453421	劒	neuradna različica
4B3421	劍	uradna japonska pismenka
513421	釰	neuradna različica

CCCII je bil zamišljen zelo sistematično in celovito, vendar se ni povsem prijel. Eden od razlogov je morda v tem, da je po sistemu CCCII vsak znak dolg tri bajte, kar je bilo bolj potratno od tajvanskega zapuščinskega sistema Big5. Poleg tega je tajvanska vlada kot uradni standard določila CNS 11643.

6 Unicode

Unicode je ubral novo pot razvrščanja pisav, pri čemer je vse pismenke vrgel v isti koš. Ne glede na to, ali se določena pismenka uporablja v vseh sistemih pisave, ki danes še uporabljajo pismenke (Japonska, Kitajska, Tajvan, Hongkong), ali pa je omejena zgolj na določeno regijo, spadajo pismenke v enotno kategorijo *han* oziroma *poenoteni CJK ideografi*, kot je bilo omenjeno že v prvem delu tega prispevka. Regionalne razlike med pismenkami so se tako zabrisale.

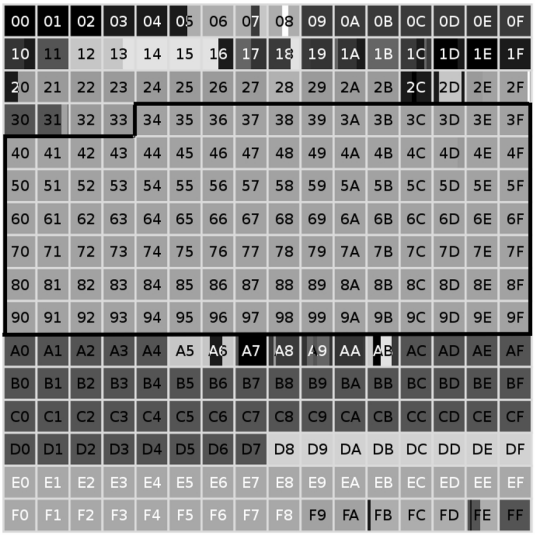
Tudi Unicode je strukturiran tridimenzionalno, pri čemer uporabi skoraj vse točke mreže 256 x 256. Spomnimo se, da so osnovni zapuščinski sistemi definirali le del mreže, običajno v območju izpisljivih znakov.



Slika 10: Osnovna struktura Unicoda (Haralambous in Horne 2007, 69).

Primarna mreža se imenuje *osnovna večjezikovna raven* (ang. Basic Multilingual Plane, BMP, Plane 0), ki s 65.536 kodnimi točkami zadostuje za večino sistemov pisav. Znaki za zapisovanje kitajščine, japonščine in korejščine so v obsegu od 2E80 (dodatki k radikalom) do 9FFF (pismenke), pri čemer so pismenke definirane v črno obrobljenem polju Slike 11.[20]

20 Določeni sklopi znakov, na primer zlogi hangula kot celota, so definirani na drugih odsekih (AC00-D7AF).



Slika 11: Položaj pismenk na osnovni večjezikovni ravni.

Poleg osnovne ravni je na razpolago še 16 dodatnih ravni, ki služijo kot prostor za bodoče širitve. Med novodefinirane znake ne sodijo le elementi pisav v klasičnem smislu, temveč tudi čustvenčki.

Če se omejimo zgolj na pisave japonščine, kitajščine in korejščine, vidimo, da je predzadnja različica (Unicode 11.0.0) dodala tri pismenke za kemijske elemente in dve pismenki za japonska lastna imena.[21]

Tabela 6: Nove pismenke v verziji Unicode 11.0.0.

kodna točka	pismenka	opis
U+9FEB	鿫	ào (ang. *oganesson*) ununoktij, element v periodnem sistemu z atomskim številom 118
U+9FEC	鿬	tián (ang. *tennessine*) tenesin, element v periodnem sistemu z atomskim številom 117
U+9FED	鿭	nǐ (ang. *nihonium*) ununtrij, element v periodnem sistemu z atomskim številom 113. To je prvi element periodnega sistema, ki so ga odkrili v Aziji, natančneje, na Japonskem. Od tod tudi ime *nihonium*. Opomba: pismenka 鉨 s tradicionalnim radikalom za kovino obstaja že od verzije 1.1 (junij 1993), vendar je na kodni točki U+9268.[22]

21 https://www.unicode.org/versions/Unicode11.0.0/ in https://www.unicode.org/charts/PDF/U4E00.pdf.

22 Primerjaj https://www.compart.com/en/unicode/U+9268.

kodna točka	pismenka	opis
U+9FEE		japonsko lastno ime
U+9FEF		japonsko lastno ime[23]

V verziji Unicode 12.0.0, ki je bila izdana 5. marca 2019, se na azijske pisave nanašajo le obrobne spremembe, na primer manjši znaki za hiragano in katakano za zapis stare japonščine. Manjša posodobitev – in s tem tudi verzija 12.1.0 – je sledila maja 2019, ko je bil kodni točki U+32FF pripisan znak ㋿, ki z eno pismenko označuje novo obdobje *Reiwa* 令和 v japonskem koledarju. Obdobje *Heisei* (8. 1. 1989–30. 4. 2019) namreč pišemo z dvema pismenkama 平成 ali z eno samo ㍻ (U+337B) (The Unicode Consortium, Japanese Era 2018).

7 Kratke 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 enovite številčne vrednosti, s čimer so omogočili na primer vnašanje kitajskih pismenk in slovenskih šumnikov v isto besedilno datoteko. Za to je bilo treba izstopiti iz enodimenzionalnega načina dojemanja kodnih točk in začeti razmišljati na drugačen način. Čeprav je Unicode dandanes že skoraj v celoti izpodrinil regionalne zapuščinske sisteme, so ti pomembna idejna dediščina. Tudi konceptualne podobnosti med Unicodom in predstavljenimi zapuščinskimi sistemi kodiranj verjetno niso zgolj naključne.

Izumiti sistem, ki dopušča ustvarjanje novih kodnih točk, je bil predpogoj in prvi korak proti uspešnemu razmahu informacijskih tehnologij v vzhodnoazijskih jezikih. Sledila so vprašanja, kako te pisave čim bolj uspešno in ekonomično vnašati, prikazovati, tiskati ali prenašati med različnimi platformami. Značilnosti posameznih jezikov in njihovih pisav namreč prinašajo težave, ki jih v jezikih s črkovnimi pisavami ne poznamo. To se je na primer kmalu pokazalo v okviru knjižničnih baz podatkov. Še pred popularizacijo Unicoda je prav za potrebe knjižnic nastal Code for Chinese Character Sets for Information Interchange (CCCII), ki ga je Kongresna knjižnica leta 1989 sprejela kot ameriški standard za jezike CJK (okrajšava za Chinese-Japanese-Korean) (Kent 1993). Tudi brez podrobnega poznavanja problematike na številne zagate naletimo že, če se poslužimo slovenskega knjižničnega informacijskega

23 Primerjaj https://www.unicode.org/L2/L2017/17396-n4831-japan-unc.pdf.

sistema COBISS, ki ga uporabljajo knjižnični sistemi Slovenije in nekaterih sosednjih držav. Težave nastanejo ne le pri popisu in vnosu del, temveč tudi pri iskanju ustreznih zadetkov. Če ne drugje, ostaja na tem področju prostor za izboljšave.

Zahvala

Pričujoči članek je rezultat raziskovalnega dela v okviru raziskovalnega programa Azijski jeziki in kulture (P6-0243), ki ga financira Javna agencija za raziskovalno dejavnost Republike Slovenije iz državnega proračuna.

Bibliografija

Bekeš, Andrej. 2019. »Kam so šle kitajske pismenke: transformacije pisav v državah kitajskega kulturnega kroga.« V *Procesi in odnosi v Vzhodni Aziji*, uredili Andrej Bekeš, Jana Rošker in Zlatko Šabič, 217–233 Ljubljana: Znanstvena založba Filozofske fakultete UL.

———. 1999. »Pojmovni okvir za klasificiranje sistemov kitajske in japonske pisave.« *Azijske in afriške študije* 3 (1999): 218–238.

Biaozhun dianmaben (Zhongwen shangyong dianma) 標準電碼本(中文商用電碼). 2004–2018. Dostop 20. 12. 2018. http://code.web.idv.hk/cccode/cccode.php?i=1.

Borgwaldt, Susanne R. in Terry Joyce. 2013. *Typology of writing systems.* Amsterdam: John Benjamins.

Bright, William. 2000. »A Matter of Typology: Alphasyllabaries and Abugidas.« *Studies in the Linguistic Sciences* 30, št. 1 (2000): 63–71.

Bunkachô 文化庁 [Agencija za kulturo]. *Jōyōkanji-hyō* 常用漢字表 [Seznam pogosto rabljenih pismenk]. 30. november 2010.

»Code Table East Asian Ideographs ('Han').« 2007. The Library of Congress. http://memory.loc.gov/diglib/codetables/9.1.html.

Coulmas, Florian. 2008. *Writing systems: an introduction to their linguistic analysis.* Cambridge: Cambridge University Press.

Daniels, Peter T. 2017. »Writing Systems.« V *The Handbook of Linguistics*, urednika Mark Aronoff in Janie Rees-Miller. New York: Oxford University Press.

Daniels, Peter T. in William Bright. 1996. *The world's writing systems.* New York: Oxford University Press.

Diringer, D. 1948. *The alphabet: A key to the history of mankind.* New York: Philosophical Library.

Gelb, Ignace J. 1969. *A study of writing.* Chicago: University of Chicago Press.
Gnanadesikan, Amalia E. 2011. *The Writing Revolution Cuneiform to the Internet.* New York: John Wiley & Sons.
Haralambous, Yannis in P. Scott Horne. 2007. *Fonts & encodings. From Unicode to advanced typography and everything in between.* Sebastopol, CA: O'Reilly Media.
Hmeljak Sangawa, Kristina. 2000. »Al' prav se piše kaisha ali kajša: o latinizaciji lastnih imen v Sloveniji.« *Azijske in afriške študije* 4, št. 1 (2000): 75–89.
———. 2019. »Makrostruktura predmodernih japonskih slovarjev: kitajski vzori in japonske inovacije.« V *Procesi in odnosi v Vzhodni Aziji*, uredili Andrej Bekeš, Jana Rošker in Zlatko Šabič, 191–215 Ljubljana: Znanstvena založba Filozofske fakultete UL.
Hōmushō 法務省 [Ministrstvo za pravosodje]. 2017. *Jinmeiyō kanji (Kosekihō shikō kisoku dai 60 jō beppyō dai 2 »Kanji no hyō«)* 人名用漢字（戸籍法施行規則第６０条別表第２「漢字の表」[Pismenke za osebna imena (Zakon o popisu prebivalstva, člen 60, 2. tabela. Tabela pismenk)].
Hsieh, Ching-chun, Jack Kai-tung Huang, Chung-tao Chang in Chen-chau Yang. 1981. »The Design and Application of the Chinese Character Code for Information Interchange (CCCII).« *Journal of Library and Information Science* 1981: 129–143.
Internet History 1962 to 1992. Dostop 10. 1. 2019. https://www.computerhistory.org/internethistory/1960s/.
Kent, Allen. 1993. *Encyclopedia of library and information science.* Vol. 51, suppl. 14: 51. New York: Dekker.
Lunde, Ken. 2008. *CJKV Information Processing (2nd edition).* ZDA: O'Reilly Media.
Mair, Victor. *Chinese Telegraph Code (CTC).* 24. april 2015. Dostop 12. 2. 2019. http://languagelog.ldc.upenn.edu/nll/?p=19175.
Monbukagakushō 文部科学省 [Ministrstvo za izobraževanje, kulturo, šport, znanost in tehnologijo]. 2017. *Shōgakkō Gakushū shidōyōryō (Heisei 29 nen 3 gatsu kokuji*）小学校 学習指導要領（平成 29 年 3 月告示）[Osnovna šola. Učni načrt (objavljen marca 2017)].
Moorhouse, A. 1953. *The triumph of the alphabet: A history of writing.* New York: H. Schuman.
Pandey, Krishna Kumar in Smita Jha. 2019. »Tracing the Identity and Ascertaining the Nature of Brahmi-Derived Devanagari Script.« *Acta Linguistica Asiatica* 9, št. 1 (2019): 59–73.
Share, D. L. in Daniels, P. T. 2016. »Aksharas, alphasyllabaries, abugidas, alphabets and orthographic depth: Reflections on Rimzhim, Katz and Fowler (2014).« *Writing Systems Research* 8, št. 1: 17–31.

Taylor, I. 1883. *The alphabet: An account of the origin and development of letters.* London: Kegan Paul, Trench.

The Unicode Consortium. 2018. *Japanese Era.* Dostop 12. 3. 2019. http://blog.unicode.org/2018/09/new-japanese-era.html.

———. 2018. *The Unicode Standard, Version 11.0.0.* ISBN 978-1-936213-19-1. Mountain View, CA: The Unicode Consortium.

———. 2019. *The Unicode Standard, Version 12.0.0.* ISBN 978-1-936213-22-1. Mountain View, CA: The Unicode Consortium.

Tomokiyo, S. 2014. *Chinese Telegraph Code (CTC), or A Brief History of Chinese Character Code (CCC).* 10. januar 2014. Dostop 14. 1. 2019. http://cryptiana.web.fc2.com/code/chinese_e.htm.

Yong, Heming in Jing Peng. 2008. *Chinese Lexicography. A History from 1046 BC to AD 1911.* Oxford: Oxford University Press.

Zhao, Shouhui in Richard B. Baldauf. 2008. *Planning Chinese characters reaction, evolution or revolution?* Dordrecht: Springer.