Mateja PETROVČIČ

Conceptual Development of the Approaches to Writing Systems of the Five East Asian Regions from the Perspective of Information Technology

Abstract

This paper presents a search for solutions for how to define Asian writing systems in the context of information technology. 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 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 then became the new research centres for this issue during the following decades. This two-dimensional thinking gave birth to several new character sets and encoding methods. In Taiwan, a third dimension was even 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 different varieties between the scripts of the five Asian regions became a secondary question. Although Unicode has almost completely replaced the regional legacy systems, they represent an important conceptual heritage. Moreover, their multi-dimensional way of thinking is a prerequisite for and the basis of Unicode.

Keywords: font, East Asian languages, Coded Character Sets, legacy encoding systems, Unicode

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

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

1 Scripts and writing systems

Much research has already been done on the concept of scripts, the development of different systems and their classifications. The first extensive studies include Taylor (1883), Diringer (1948), Moorhouse (1953), and Gelb (1969), among others, who contributed significantly to the understanding of this topic, but their views are somewhat outdated. More modern works include research such as Daniels and Bright (1996), Coulmas (2008), Gnanadesikan (2011), Borgwaldt and Joyce (2013), and Daniels (2017). In the Slovenian language, Bekeš (1999; 2019) and Hmeljak Sangawa (2019) focus on the classification of the Chinese and Japanese scripts. A comparison of the positions of individual studies would be too extensive at this

point, so let us limit ourselves to the definitions of the Unicode Consortium, since the focus of this paper is on the treatment of scripts from the point of view of information technology.

In the context of information technology, the term writing system denotes two different concepts. On the one hand, it describes the general principle of how individual groups of scripts graphically represent the selected language. From this point of view, the Unicode Consortium follows a classification that divides scripts into three categories: alphabets, syllabaries and logosyllabaries.

Alphabets are writing systems where the basic elements are letters, which are used to write consonants and vowels. The alphabet we are most familiar with is the Latin alphabet, which with certain adaptations is used to write many languages. The degree of correspondence between sounds and letters is a separate issue that we do not consider here. A writing system that records only consonants is called an *abjad*, not an alphabet. To us the most familiar *abjad* is the Arabic writing system (The Unicode Consortium, The Unicode Standard, Version 11.0.0 2018, 256).[1]

Syllabary signs record syllables, which most often means a combination of consonant(s) and vowel(s). This also includes the Japanese *hiragana* and *katakana* syllabaries. In China, the syllabary of one of the Yi languages belongs to this group.[2] The Korean *Hangul* is neither an alphabet nor a syllabary, as its syllables are composed of letters called *jamo*, which are not independent units of the Korean script in themselves. Because of these characteristics, Unicode calls it a featural syllabary (The Unicode Consortium, The Unicode Standard, Version 11.0.0 2018, 257).

The Chinese writing system is logographic in nature, and Unicode uses the term logosyllabary. This term refers to writing systems where the smallest units represent words and/or word morphemes, which can also be used as rough recordings of the acoustic image. The basic unit of logosyllabaries has many names, for example, *ideograph*, *ideogram*, *logograph*, *logogram*

1 In addition to these, there are also abugids, in which consonants that imply one vowel are written with primary signs, while the rest of the vowels are written with secondary, added signs, which, together with signs for consonants, form groups that record syllables (Bright 2000; Daniels 2017; Share 2016). These include the scripts from the Indian subcontinent, such as Devanagari (Pandey and Jha 2019).

2 The Chinese government recognizes six languages of the Yi group. The languages are unrelated, but they used a common script for shamanic purposes. The traditional writing system was logographic, but the modern one is syllabic. The Yi syllabary has been the official script of Northern Yi since 1980.

and *sinogram*, and in layman's terms, even a *letter* or *sign*. By definition, logographs are units that represent a word or morpheme, while ideographs are units that represent ideas or concepts. The boundaries between morphemes, words and concepts are often blurred (The Unicode Consortium, The Unicode Standard, Version 11.0.0 2018, 258).

Chinese characters are primarily used to write the Chinese language, but other East Asian languages have also integrated them into their writing systems. Moreover some individual regions have created their own characters based on the Chinese ones, which are only used in that specific area. The umbrella term for the characters is *hanzi* in Chinese, *kanji* in Japanese, and *hanja* in Korean. Unicode, as we will see below, throws all the characters into one set and blurs regional boundaries. It calls this group of characters *Han* or CJK Unified Ideographs.[3]

On the other hand, the term *writing system* refers to a set of **scripts** that are used to write down a certain language.[4] From this point of view, the Japanese writing system uses four scripts, i.e. *Han*, *hiragana*, *katakana* and Latin[5] (The Unicode Consortium, The Unicode Standard, Version 11.0.0 2018, 256). In addition to these, technically speaking, the scripts *Bopomofo, Hangul, Yi, Nüshu*,[6] *Lisu*,[7] *Miao*[8] and *Tangut*[9] are in use today in East Asia (The Uni-

3 CJK stands for Chinese (C), Japanese (J) and Korean (K). An alternative term is CJKV, which adds Vietnamese (V).

4 The term *script* should not be equated with the term *character set*, as the latter denotes a set of characters that, in principle, differs from the characters of a specific writing system. If we limit ourselves to the example of the Latin alphabet, the ISO/IEC 8859-1 character set is suitable for writing English, German, Italian and many other languages that use the Latin alphabet, but not for Slovenian, because it does not contain sibilants. On the other hand, the ISO/IEC 8859-2 character set is suitable for writing Slovenian, Slovak, Czech, Hungarian and some other languages, including English.

5 On the Latinization of proper names in Slovenia, see Hmeljak Sangawa 2000.

6 The *Nüshu* script was used by women in the Hunan province of southeastern China. The characters of this script come from Chinese characters, but they often only represent the acoustic image of the syllables.

7 The *Lisu* script was created at the beginning of the 20th century for recording the *Lisu* language of the Yunnan province. It has been officially recognized by China since 1992. It consists of Latin letters, rotated Latin letters, and punctuation used to denote tones (The Unicode Consortium, The Unicode Standard, Version 11.0.0 2018, 692).

8 The *Miao* script was created in 1904 and is used for recording the language of the same name from the northeastern part of the Yunnan province.

9 The *Tangut* script records the Tangut language, which was in use from the 11th to the 16th century in what is now northeastern China. It was rediscovered at the end of the 19th century, and nowadays it is mainly the subject of academic research (The Unicode Consortium, The Unicode Standard, Version 11.0.0 2018, 693).

code Consortium, The Unicode Standard, Version 11.0.0 2018, 691). Visually speaking, what all these scripts have in common is that they look square. Each graphic unit therefore occupies a square of space.[10]

2 Uncoded and Coded Character Sets

We talk about Uncoded and Coded Character Sets mainly in connection with scripts, which consist of a large number of elements. These primarily include the writing systems of Asian languages with thousands of characters. But before we clarify what these terms refer to, we must mention two basic approaches to this mass of characters.

A review of historical lexicographic works shows that part of the material aims to cover as many characters as possible, while the other part strives to limit itself to the most important ones. Normative dictionaries belong to the first category, as they try to capture as many characters as possible from this mass and thus standardize the writing. However, they also try to limit and exclude alternative, unofficial records of the same character from the open set of characters, which is still increasing. In the second category are the various shorter lists of characters that attempt to extract the most important ones from the vast character set and create smaller, manageable subsets. Tens of thousands of characters are too many for everyday use, as some of them are quite outdated or rare (Zhao and Baldauf 2008, Yong and Peng 2008).

Chinese script reforms go back a long way, and the standardization of writing was one of the key tasks during the Qin dynasty (221–207 BCE), when a list of 3,500 characters was created to be used as an official standard. In this context the work *Cangjiepian* (倉頡篇) was also made, which represents an attempt to reform the Chinese script and establish orthographic standards for the then small seal script (Zhao and Baldauf 2008, 25). The next major work is the character dictionary[11] *Shuowen jiezi* (说文解字) from 100 CE, in which 9,353 characters are listed. It also defines the characters to be used for correct writing, while including their frequently used alternative forms.

10 Because of this characteristic, the character sets for Asian scripts also created a full-width Latin alphabet, where each letter typographically occupies the space of one square (Japanese: 全角ローム字 *zenkaku roomaji*; Chinese: 全角安全全像 *quanjiao zimu*). As an example, let's take the letter A in the Unicode framework, which successfully combines different character sets. The base capital letter A is located at code point <U+0041> and its full width equivalent (Ａ) is at code point <U+FF21>.

11 Present-day dictionaries are divided into character dictionaries and word dictionaries. In this paper we only discuss character dictionaries, so the term dictionary below refers to the latter.

Dictionaries have been added to throughout history and have thus included more and more characters, but in all cases these are still only subsets of all existing characters. The next dictionary whose influence is still visible today is the *Kangxi zidian* from 1716, which defined 47,035 characters.[12] As Zhao and Baldauf (2008, 16) write, the most comprehensive dictionary to date is the *Zhonghua zihai* (中华字海) from 1994, which includes 85,000 characters. However, the truth is that it too is not a complete set of all existing characters, no matter how extensive it is.

In order to improve literacy, in the 1950s the Chinese government defined a closed set of 7,000 characters, which we know as the *commonly used characters* (*Xiandai Hanyu tongyongzi biao* 现代汉语通用字表). Of these, 3,500 are further listed as *frequently used characters* (*Xiandai Hanyu changyongzi biao* 现交汉语常用字表), which is the number that a person with a high school education is expected to know. The list is further divided into 2,500 primary characters for the elementary school level and 1,000 secondary characters for the high school level. In addition, the Chinese government has published a list of simplified characters (*Jianhuazi zongbiao* 简化字总表), which includes 2,200 characters. These include characters that are graphically different from traditional characters, with the aim of reducing the number of strokes by more than half, with the idea that this should further contribute to greater literacy (Zhao and Baldauf 2008, 48).

In 1982–1984, Taiwan defined general-purpose characters in a broader scope. There are 4,808 frequently used characters (*Changyong guozi biaozhun ziti biao* 常用國字標準字體表), the secondary characters list includes 6,341 characters (*Ci changyong guozi biaozhun ziti biao* 次常用國字標準字體表), and the rare characters list includes 18,480 (*Hanyong ziti biao* 罕用字體表). In addition, a list of 18,609 versions of characters (*Yiti guozi zibiao* 異體國字字表) was also defined.

In Japan, the list of frequently used characters since the 2010 reform includes 2,136 characters (*Jōyō Kanji hyō* 常用漢字表[13]), of which 1,006 belong to the educational set (Kyōiku kanji 教育漢字), which further defines exactly in which primary school grade children should learn certain characters.[14]

12 Unicode is also based on it, but more on this below.

13 Retrieved from the Agency for Cultural Affairs of the Ministry of Education, Culture, Sports, Science and Technology (Jōyōkanji-hyō 常用漢字表 [List of Frequently Used Kanji Characters] 2010).

14 With the revision of the curricula, which will enter into force in 2020, 20 characters used to write the names of prefectures will be included in the set for primary school, so that it will then include 1,026 characters. The order of adoption will also be slightly changed (Monbukagakushō 2017).

The remaining 1,130 (1,110 from 2020 onwards) characters are among the frequently used ones, which exceed the primary school level. In addition, there is a list of 863 characters for personal names (*Jinmei-yō Kanji* 人名用漢字覧表).[15]

Both Korea and Vietnam adopted certain Chinese characters throughout history, but later developed their own scripts. Korea created the *Hangul* script and established a set of characters that students should learn during their school years. The educational set thus comprises 1,800 characters (*Hanmun Gyoyukyong Gicho Hanja* 한문교이용기초 한자/漢文教育用基礎漢字), of which 900 are to be learned at the secondary school level and 900 at the university level. The Supreme Court of Korea also established a list of 2,964 characters acceptable for use in personal names (*Inmyeong-yong Hanja* 인명용 한자/人名用漢字) (Lunde 2008, 84). However, knowledge of *hanja* characters is not essential for Korean speakers, as nowadays almost everything is written in *Hangul*.

All of the above-mentioned sets of characters are uncoded, i.e. subsets that were created outside the framework of information technology, regardless of whether they can be computerized. Knowledge of Uncoded Character Sets is important for understanding Coded Character Sets, as the former served as the basis for computer scientists to construct the latter.

The term Coded Character Set therefore indicates that it is a collection of characters intended for computer processing. Each character must have its own code point, i.e. a unique numerical value. This is crucial for understanding the issue of Asian language scripts. Namely, the design of computers is not language-independent, but is linked to English and the English writing system, which can be seen from the structure of the 8-bit byte, ASCII, keyboard design, and the like. As we will see below, this also reflects a one-dimensional way of thinking, which was not suitable for the writing systems of Asian languages.

3 One-dimensional approach

America, the cradle of computer development, set the standards for character encoding with ASCII (*American Standard Code for Information Interchange*). The original 7-bit ASCII with seven bits defined 2^7 or 128 code points. This was enough for 33 control characters, a space and 94 printable characters. With this amount of code points, it was possible to define the

15 Retrieved from the Ministry of Justice (Hōmushō 2017).

26 uppercase and 26 lowercase letters of the English alphabet, 10 digits and 32 other common characters, such as punctuation marks and mathematical operators. These are also the characters of a typical English keyboard.

This thinking is completely one-dimensional. With one bit we create two combinations (2^1), with two bits four combinations (2^2), with three bits eight combinations (2^3), and so on. Seven bits were sufficient for the English script, but not for the alphabets of other languages that use characters unknown to the English alphabet, for example *č*, *š, ž*, *Č*, *Š*, *Ž* for Slovenian, *ä*, *ö*, *ü*, *ß* for German, *é*, *è*, *ê*, *ë*, *æ*, *œ*, *ç*, etc. for French, etc.

The addition of an eighth bit allowed for an additional 128 code points, which was sufficient for most alphabet scripts. The letters of the English alphabet remained at the same code points, and the locally specific letters were at values from 161 to 255 (decimal notation). The only problem was that even the 94 new places were not enough for all the special letters of all the scripts. As part of the ISO/IEC 8859 standards, 15 parts or versions were thus created, which were used in different regions and covered the scripts of a certain group of languages.

Table 1: Comparison of code points 185, 232 and 248 in the five regional versions of the ISO/IEC 8859 standard.

ISO/IEC 8859	Name	suitable for the scripts of the following languages	185	232	248
ISO/IEC	Latin-1, Western European Latin-2, Central European Latin-3, South European Latin/Cyrillic Latin/Greek	English, German, Icelandic, Italian, Portuguese... Slovenian, Slovak, Hungarian, Polish... Turkish, Maltese, Esperanto Bulgarian, Macedonian, Russian, Belarusian... Modern Greek	¹ š ı Й Ή	è č è ш θ	øřĝj ψ

In practice, this meant that Slovenian and Slovak users saw the word *češnja* (cherry) the same way, German or English users saw it as *èe¹nja*, in Bulgaria they saw the word *шеЙnja*, and in Greece it was printed out as *ϑeΉnja*. In simple text editors (for example Notepad++) we can switch between different encodings and observe the differences between character sets. We will notice that the English pangram *the quick brown fox jumps over the lazy dog* will be displayed correctly in all encodings. The Slovenian pangram *v kožuščku hudobnega fanta stopiclja mizar in kliče* will be distorted only where there

are sibilants. For German we can use *Victor jagt zwölf Boxkämpfer quer über den großen Sylter Deich*. To display the Estonian script, *see väike mölder jõuab rongile hüpata* is suitable, and so on.[16]

The first attempt to adapt to Asian languages was the Asian versions of the ASCII set. The Chinese version was called GB-Roman, Taiwan's CNS-Roman, Japan's JIS-Roman, and Korea's KS-Roman. Like ASCII, these sets contain 94 printable characters. The only difference was in the value of the "$" and "\" signs (Lunde 2008, 91). These fixes are hardly worth mentioning because they only changed the glyph of one sign.

The next step was the 8-bit JIS X 0201 standard, which converted eighth-bit code points to half-width *katakana*. Within the scope of this standard the cherry (*češnja*) mentioned above would appear as ・ｅﾅnja. There were enough code points for one syllabary, but there was already not enough space for the other, let alone to include kanji characters.

Using the linear approach, we therefore create 256 code points with eight bits, which is significantly too few for the scripts of Asian languages. Table 2 shows what the maximum values of strings of different lengths would be.

Thirteen-bit bytes would be needed to display 7,000 characters, the number of commonly used characters. Since the length of a byte, i.e. the smallest carrier of information, is a matter of agreement, Asian computers could – in theory – operate on the basis of longer bytes. The question would probably soon arise as to how long a byte should be in order to really have enough code points. If you wanted to display the contents of the 85,000-character *Zhonghua zihai* dictionary on a computer, you would need 17-bit bytes. This would be feasible, but on the other hand, the convention that one

16 For pangrams of other **alphabets** see, for example, http://clagnut.com/blog/2380/. Pangrams for **syllabaries** are a bit longer, but still manageable. For *Hangul*, we have an example of 밤새컴퓨터로요약을해치우면좋겠다 (*BamSae KumPyooTuhRo YoYakEul HaeChiWooMyun JotGetDa*). For *hiragana*, we can use 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 moyanōch*i) (see https://camtsmith.com/articles/2016-11/pangrams) or the pangram presented in Hmeljak Sangawa (2019).
There are no pangrams for Chinese, but the idea of a pangram was already present in the 6th century, when the text *Qianziwen* (千字文) used 1,000 characters that were commonly used at the time, and children had to learn this work by heart. A similar text entitled *Sanzijing* (三字经) was created in the 13th century, and the text *Baijiaxing* (百家姓), in which 472 surnames are woven into verses, also dates to the Song dynasty (960-1279). All three works are collectively known as *San-bai-qian* (三百千) and were included in teaching materials until about 1930. The full texts are available at https://baike.baidu.com/item/三百千.

byte is a set of eight bits had already been established in 1964 (Internet History 1962 to 1992).

Table 2: Maximum value in binary and decimal notation according to the number of bits.

Number of bits	Total number of code points	Binary notation	Decimal notation
7	128	1111111	127
8	256	11111111	255
9	512	111111111	511
10	1,024	1111111111	1023
11	2,048	11111111111	2047
12	4,096	111111111111	4095
13	8,192	1111111111111	8191
14	16,384	11111111111111	16383
15	32,768	111111111111111	32767
16	65,536	1111111111111111	65535
17	131,072	11111111111111111	131071

At this point, it appeared that the scripts of Asian languages would not be computerizable. The Japanese came up with the first real solution in 1978 with the development of the ISO-2022 standard. Today, when Unicode 12.0[17] already defines more than 100,000 Chinese characters, the issue of Asian scripts is no longer as problematic, but who is to say that today's solutions would have come at all if there had not been for the switch from one-dimensional to two-dimensional thinking.

4 Two-dimensional approach

The basics of a two-dimensional way of thinking were unknowingly already present in the Chinese postal system. It is thus almost unbelievable that so much time passed before the first computer solution was found, since the seeds of that solution had already been sown.

Just as ASCII was sufficient for the exchange of English text in the context of information technology, Morse code was previously sufficient for the trans-

17 Translator's Note: When the original article was written, Unicode 12.0 had already been released on March 5, 2019. Unicode 13.0 was scheduled for release in 2020 and was officially launched on March 10, 2020.

mission of English text via telegraphy. And just as ASCII was inadequate for the needs of Asian scripts, Morse was inadequate for Asian languages. In principle, messages could be transmitted via acoustic image, but even this would bring many problems due to homophones, the lettering of the language, dialectal differences and the length of the texts. The solution to the unambiguous transmission of Chinese characters was published in 1881 when Zheng Guanying (鄭觀應) published the manual *Dianbao xinbian* (电报新编) (Mair 2015).

Each character had a four-digit code, from 0000 to 9999. On each page of the manual was a 10 x 10 table, meaning there were 100 characters on one page. The first two digits represented the page in the handbook, the third digit referred to the row mark in which the letter was, and the fourth digit referred to the column mark in which the letter was. For example, the characters *zhongwen* 中文 have the code 0022 2429, which means that the character *zhong* 中 is in the second column of the second row on page 00, and the character *wen* 文 is on page 24, in the second row, the ninth column.

The work of the telegraph operators was multi-staged. The postal clerk first converted the characters of the message into four-digit codes with the help of a telegraph manual, then converted them into Morse code and telegraphed the message to the postal clerk on the other side. The clerk then wrote down the accepted Morse codes as a continuous sequence of digits, divided them into four-digit sets, and finally, with the help of a telegraph manual, converted the four-digit codes back into characters. Figure 1 shows an example of a telegram:

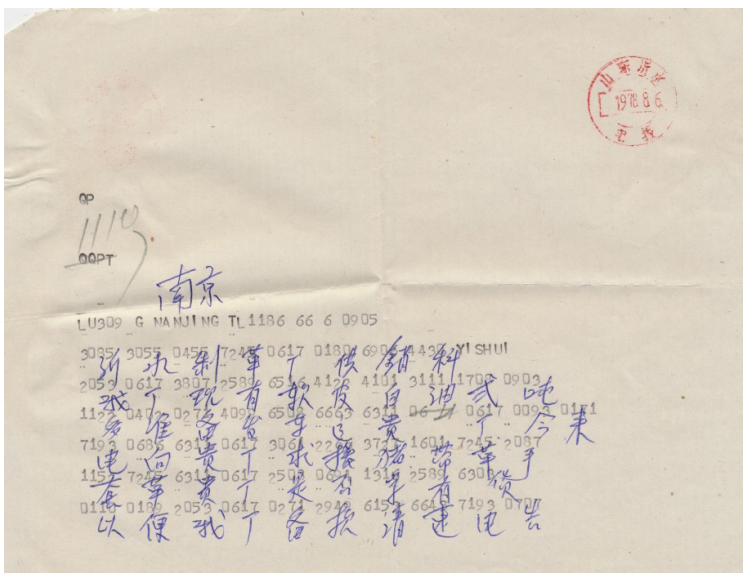

Figure 1: An example of a Chinese telegram (Mair 2015).

In 1885, Korea also adopted the telegraph system from China, including characters. Korean telegrams were written either in Chinese characters or in letters of the English alphabet, but not in *Hangul* (Tomokiyo 2014).

The 10,000 different telegraph codes are therefore not understood in a linear or one-dimensional way, as we would count from 0 to 9999, but in a flat or two-dimensional way. The characters were classified into 10 x 10 grids according to dictionary principles. Two hundred and fourteen radicals from the *Kangxi zidian* dictionary served as the basis for this. Radicals with fewer strokes were ranked before radicals with more strokes, and within each radical the characters were again classified according to the number and shape of strokes. Chinese telegraph code manuals are still in use today, and are available online, with some differences between the Chinese, Taiwanese and Hong Kong versions (*Biaozhun dianmaben* (*Zhongwen shangyong dianma*) [標準電碼本(中文商用電碼)] 2004- 2018).

As mentioned at the beginning of this chapter, coding solutions were first found by the Japanese in 1978 with the development of the ISO-2022 standard. Different sets of characters were classified into a grid of 94 x 94 points, as this is the number of printable characters within ASCII. In this way, they created 8,836 code points. Figure 2 shows the area of the grid where the Coded Character Sets were placed. The axis labels use hexadecimal notation, which means that a decimal value of 128 is displayed here as 80 and a value of 255 as FF.

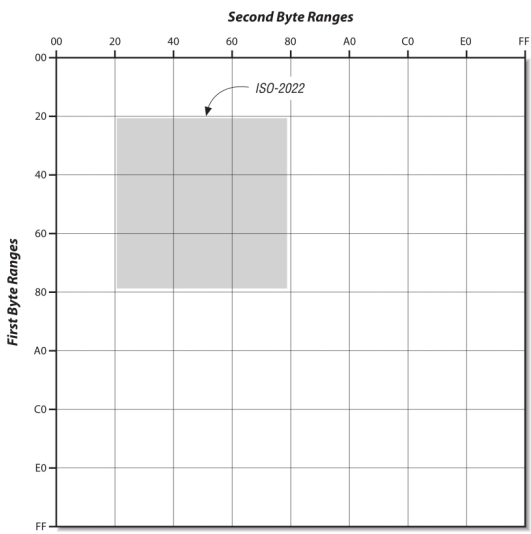

Figure 2: Area of code points in the ISO-2022 encoding (Lunde 2008, 231).

We can see from Figure 2 that all characters are in the rank of the first seven bits and that the eighth bit is unused. This means that this coding system was very convenient for exchanging information between computers. However, since the grid area overlapped with the letters of the English alphabet, there had to be a system for switching between single-byte and double-byte data processing. Modal encodings, including ISO-2022, solve this with escape sequences or other special characters that indicate switching between character sets or different versions of the same character set (Lunde 2008, 195). Since going into detail about this would be too complex and long-winded for the purpose of this paper, let me use a simpler comparison. Imagine that the strings of ones and zeros are the tracks, and the processor is the train that travels along them. The position of the switches (the selected escape sequence) directs it to the first track (single-byte read) or the second track (double-byte read). At the end of the section, there are switches (escape sequences) that redirect the train to a new direction. In this way, it was possible to create 128 + 8,836 code points with a 7-bit byte.[18]

In that many code points the Japanese version of ISO-2022-JP could accommodate several character sets: ASCII, JIS-Roman (or the Japanese version of ASCII), JIS X 0208 (with special characters, digits, Latin, *hiragana*, *katakana*, Greek alphabet, Cyrillic, table markings, etc.) and JIS X 0208-1983 (1983 expansion).

These solutions were then adopted in other regions of East Asia. In its ISO-2022-CN version, China of course kept ASCII, and placed GB 2312-80 character sets (GB-Roman or the Chinese version of ASCII, *hiragana*, *katakana*, Greek alphabet, Cyrillic, *pinyin*, *Bopomofo*, 6,763 Chinese characters, special characters, table markings, etc.) and the first two levels of the Taiwanese standard CNS 11643-1992 on the 94 x 94 grid. Levels 3 to 7 of this Taiwanese standard were added to the expanded version ISO-2022-CN-EXT (Lunde 2008, 229-230). The Koreans also produced their own version, ISO-2022-KR.

Once the first step into two-dimensional thinking was taken, it did not take long to switch to non-modal encoding, which uses numeric code point values to switch between single-, double-, or multi-byte processing. The EUC (*Extended Unix Code*) coding was created based on the ISO-2022 coding. Figure 3 shows the position of the double-byte grid in EUC encodings. This time each region also placed their own character sets in this area.

18 7-bit bytes were later used by UTF-7, which is no longer in use today.

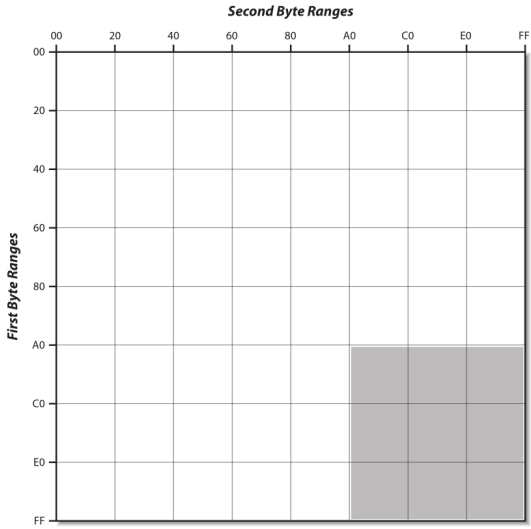

Figure 3: Two-byte area in EUC encoding (adapted from Lunde 2008, 246).

The Japanese extended the additional character sets to three bytes, which was already the beginning of a three-dimensional approach to coding. A single point of the first byte led to the third dimension or 3-byte reading, i.e. 0x8F. Figure 4 shows the Japanese version of EUC-JP.

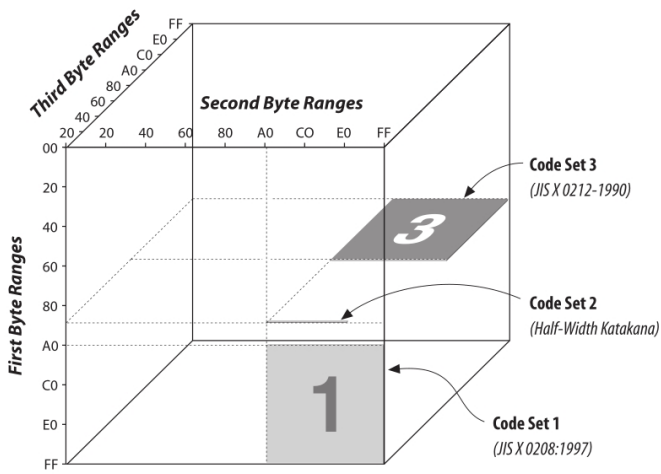

Figure 4: Two- and 3-byte area within the EUC-JP encoding (Lunde 2008, 250).

The Taiwanese placed multibyte characters in a similar area to the Japanese, except that the characters were four bytes. They created 80 levels, as roughly seen in Figure 5. The 4-byte character area is where the thin grey line is marked. An extension of this system was also used in Hong Kong, which more or less directly adopted the Taiwanese solutions. Indeed, both regions use traditional Chinese characters, with Hong Kong only having to add some locally specific ones. In contrast to Taiwan, which arranged the characters according to carefully considered principles, the letters in the expanded HKSCS (Hong Kong Supplementary Character Set) were added without a specific system.

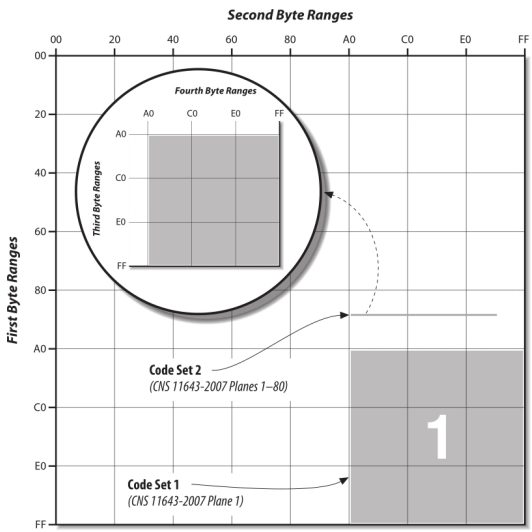

Figure 5: Two- and 4-byte area in EUC-TW encoding (Lunde 2008, 248).

In addition to the versions of the ISO-2022 and EUC systems that were used in several Asian regions, individual regions created their own, regionally conditioned coding systems, which did not result in radically different approaches. Take China for example. The GBK coding system was derived from the ISO-2022-CN. Figure 6 shows in which directions they expanded to obtain new code points.

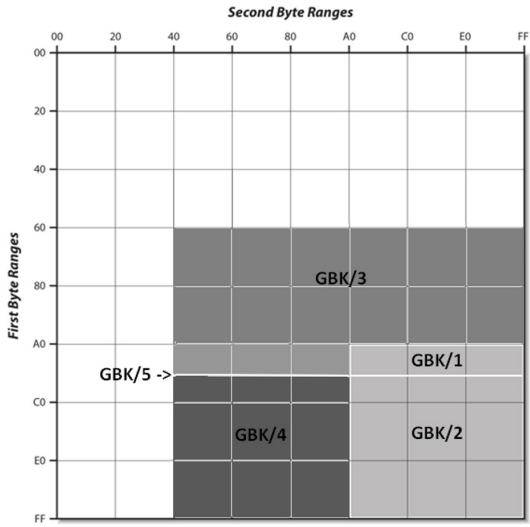

Figure 6: Area of code points in the GBK coding system.

Even the GB18030 standard, which all devices intended for the Chinese market must support since 2006, only added a new band of code points to the GBK coding system, as shown in Figure 7.

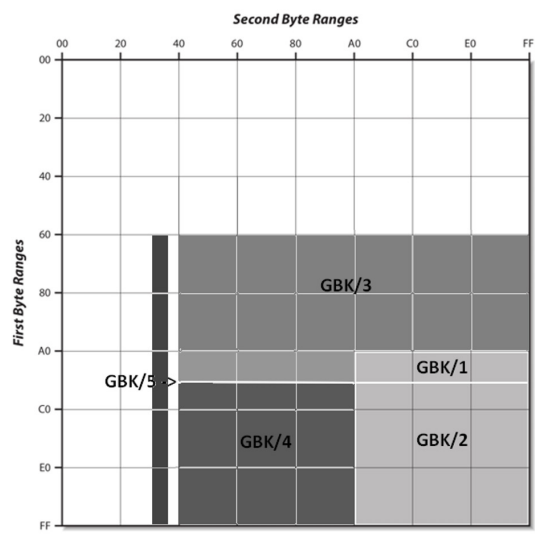

Figure 7: Area of code points in the GB18030 coding system.

5 Three-dimensional approach

The most complex approach to the structuring of information and the arrangement of characters is manifested in the CCCII character set (Chinese Character Code for Information Interchange 中文資訊交換碼). The first version dates back to 1980, with revisions in 1982 and 1987 (Lunde 2008, 122).

CCCII is based on the ISO 2022 coding and is divided into 16 layers. Each layer except the last is further divided into six planes, which makes a total of 94 planes. Combining this with the concept of a 94 x 94 grid results in a 94 x 94 x 94 cube as shown in Figure 8.

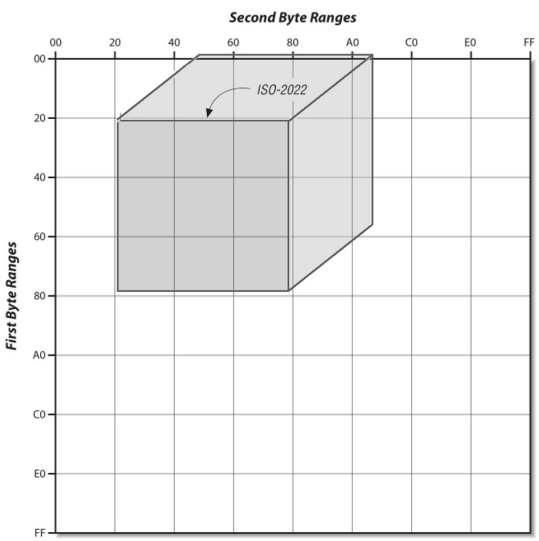

Figure 8: CCCII encoding structure.

Each layer is dedicated to a certain character type, as shown in Table 3. Table 4 then shows the structure of the first layer in more detail.

Table 3: Character categories by layers in CCCII coding (Lunde 2008, 122).

Layer	Plane	Character type
1	**1–6**	**Non-Chinese characters and traditional Chinese characters**
2	7–12	Simplified Chinese characters
3–12	13–72	Character versions from layer 1
13	73–78	*Hiragana*, *katakana* and Japanese *kanji* characters
14	79–84	*Jamo*, *Hangul* and Korean *hanja* characters
15	85–90	Reserved area
16	91–94	Other characters

Table 4: Structure of the first layer by levels in CCCII coding (adapted from Lunde 2008, 122).

Plane	Line (dec)	Number of characters	Character type
1	1		Reserved for control characters
1	2–3		
1	4–10	0	Unassigned
1	11	35	Chinese punctuation
1	12–14	214	214 *Kangxi* radicals
1	15	78	Numbers and *zhuyin* (*Bopomofo*)
1	16–67	4,808	Frequently used characters[19]
1–3	68_1–64_3	17,032	Secondary characters
3–6	65_3–5_6	20,583	Other characters
6	6–94	0	Unassigned

The 94 x 94 x 94 cube offers 830,584 code points, which is almost ten times as many as the most comprehensive Chinese dictionary. For this reason we can find that many code points are empty. The exceptional nature of this solution can be seen above all in the interrelationship of the layers and the resulting connection of the characters. As Figure 9 shows, versions of the same character have the same value of the first two bytes, but differ in the third byte.

19 See Chapter 2 *Uncoded and Coded Character Sets*, the paragraph on Taiwan.

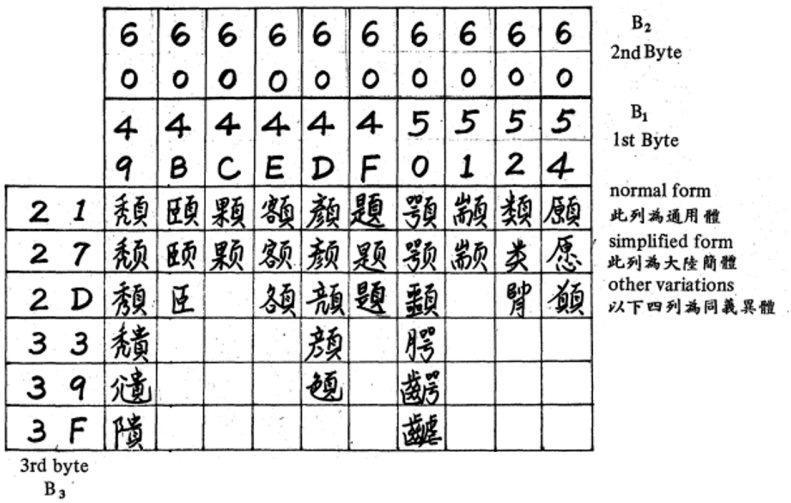

Figure 9: Code points of characters in the CCCII encoding (Hsieh et al. 1981, 135).

Because CCCII works within 7-bit bytes, it is suitable for library systems, and the Library of Congress of America has adapted it to the EACC (East Asia Coded Character) system. A list of code points for 13,478 characters is available on the Library of Congress website (Code Table East Asian Ideographs (*Han*) 2007). From the value of the code point, we can immediately tell whether a particular character is a primary Taiwanese character (x1xxxx), a simplified character (x7xxxx), or one of the variants of the character (NNxxxx). Let's look at the example in Table 5.

Table 5: Example of placement of official character and its variants in the EACC system.

Code point	Character	Use
213421	劍	The official traditional character in Taiwan
273421	剑	Simplified character
2D3421	劒	Unofficial version
333421	釼	Unofficial version
453421	劍	Unofficial version
4B3421	劍	The official Japanese character
513421	釰	Unofficial version

The CCCII was conceived very systematically and comprehensively, but it did not quite catch on. One reason may be that under the CCCII system each character is three bytes long, which is more wasteful than Taiwan's legacy Big5 system. In addition, the Taiwan government has set CNS 11643 as the official standard.

6 Unicode

Unicode took a new path in classifying scripts, throwing all characters into the same group. Regardless of whether a specific character is used in all writing systems that still use characters today (Japan, China, Taiwan, Hong Kong), or whether it is limited only to a certain region, characters belong to the single category of *Han*, or the *Unified CJK Ideographs*, as was already mentioned in the first part of this paper. Regional differences between characters were thus blurred.

Unicode is also structured three-dimensionally, using almost all points of a 256 x 256 grid. Recall that the underlying legacy systems defined only part of the grid, usually in the area of printable characters.

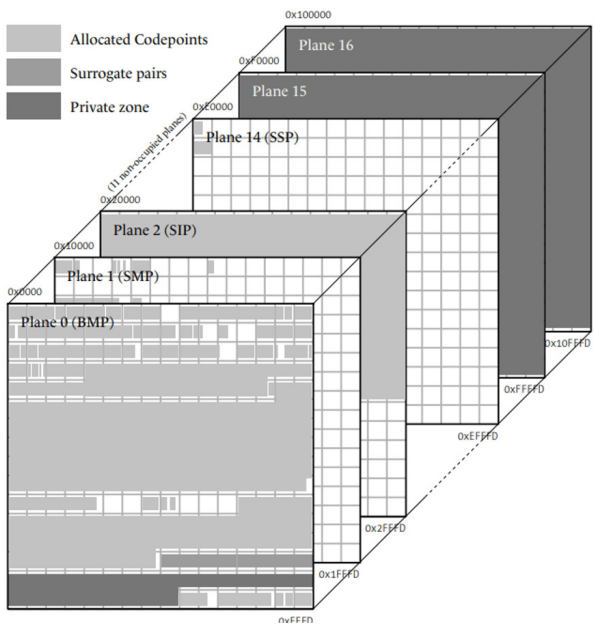

Figure 10: Basic structure of Unicode (Haralambous and Horne 2007, 69).

The primary grid is called the *Basic Multilingual Plane* (BMP, Plane 0), which with 65,536 code points is sufficient for most writing systems. The characters for writing Chinese, Japanese, and Korean range from 2E80 (additions to radicals) to 9FFF (characters), with the characters defined in the black-lined box in Figure 11.[20]

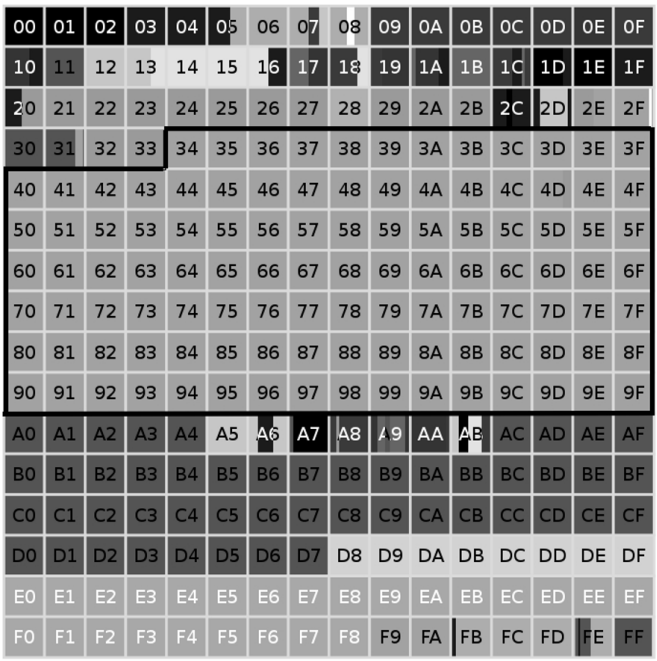

Figure 11: Position of Chinese characters on the basic multilingual plane.

In addition to the basic plane, there are 16 additional planes available, which serve as space for future expansions. The newly defined characters include not only script elements in the classical sense, but also emoticons.

If we limit ourselves to Japanese, Chinese and Korean scripts, we see that the penultimate version (Unicode 11.0.0) added three characters for chemical elements and two characters for Japanese proper names.[21]

20 Certain sets of characters, such as *Hangul* syllables as a whole, are defined in other sections (AC00-D7AF).

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

Table 6: New characters in Unicode version 11.0.0.

Code point	Character	Description
U+9FEB	龒	ào (eng. *oganesson*) the element in the periodic table with atomic number 118
U+9FEC	鿬	tián (eng. *tennessine*) the element in the periodic table with atomic number 117
U+9FED	钅尔	nǐ (eng. *nihonium*) the element in the periodic table with atomic number 113. It is the first element of the periodic table that was discovered in Asia, more precisely, in Japan. Hence the name *nihonium*. Note: The character 鈰 with the traditional radical for metal has existed since version 1.1 (June 1993), but at code point U+9268.[22]
U+9FEE	龐	Japanese proper name
U+9FEF	鿯	Japanese proper name[23]

In Unicode version 12.0.0, released on March 5, 2019, only marginal changes apply to Asian scripts, such as smaller *hiragana* and *katakana* characters for Old Japanese. A minor update – and with it version 12.1.0 – followed in May 2019, when the code point U+32FF was assigned the character ㋿, which with one letter indicates the new *Reiwa* 令和 period in the Japanese calendar. The *Heisei* period (January 8, 1989–April 30, 2019) is written with two letters 平成 or with a single ㍻ (U+337B) (The Unicode Consortium, Japanese Era 2018).

7 Brief concluding thoughts

The simultaneous use of different scripts is not as self-evident as we think today. It was only at the end of the 1980s that experts found a solution to assign uniform numerical values to thousands of characters, which made it possible, for example, to enter Chinese characters and Slovenian sibilants into the same text file. For this, it was necessary to get out of the one-dimen-

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

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

sional way of perceiving code points and start thinking in a different way. Although Unicode has now almost entirely supplanted regional legacy systems, the latter remain an important conceptual legacy. Even the conceptual similarities between Unicode and the presented legacy encoding systems are probably not purely coincidental.

The invention of a system that allows the creation of new code points was a prerequisite and the first step towards the successful expansion of information technology in East Asian languages. What followed were questions about how to enter, display, print or transfer these scripts between different platforms as efficiently and economically as possible. The characteristics of individual languages and their scripts bring problems that are unknown in languages with alphabetic scripts. This soon became apparent, for example, in the context of library databases. Even before the popularization of Unicode, the Code for Chinese Character Sets for Information Interchange (CCCII) was created precisely for the needs of libraries, which was adopted by the Library of Congress in 1989 as the American standard for CJK languages (abbreviation for Chinese-Japanese-Korean) (Kent 1993). Moreover, even without detailed knowledge of the issue, we encounter many problems if we use the Slovenian library information system COBISS, which is used by the library systems of Slovenia and some neighbouring countries. Problems arise not only in listing and entering works, but also in finding relevant hits. If nothing else, there remains room for improvement in this area.

Sources

Bekeš, Andrej. 2019. “Kam so šle kitajske pismenke: transformacije pisav v državah kitajskega kulturnega kroga.” [Where Have the Chinese Characters Gone? Modernization of Writing Systems in the Periphery of the Sinographic Cosmopolis.] In *Procesi in odnosi v Vzhodni Aziji*, edited by Andrej Bekeš, Jana Rošker and Zlatko Šabič, 217–233. Ljubljana: Znanstvena založba Filozofske fakultete UL.

———. 1999. “Pojmovni okvir za klasificiranje sistemov kitajske in japonske pisave.” [A conceptual framework for classifying Chinese and Japanese writing systems] *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. accessed 20 Dec. 2018. http://code.web.idv.hk/cccode/cccode.php?i=1.

Borgwaldt, Susanne R. and 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, no. 1 (2000): 63–71.

Bunkachô 文化庁 [Agency for Cultural Affairs]. *Jōyōkanji-hyō* 常用漢字表 [List of Frequently Used Kanji Characters]. 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.” In *The Handbook of Linguistics*, edited by Mark Aronoff and Janie Rees-Miller. New York: Oxford University Press.

Daniels, Peter T. and 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 and 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.” [Is it correct to spell kaisha or kajša: about the Latinization of proper names in Slovenia.] *Azijske in afriške študije* 4, št. 1 (2000): 75–89.

———. 2019. “Makrostruktura predmodernih japonskih slovarjev: kitajski vzori in japonske inovacije.” [Macrostructure of Pre-modern Japanese Dictionaries: Chinese Models and Japanese Innovations.] In *Procesi in odnosi v Vzhodni Aziji*, edited by Andrej Bekeš, Jana Rošker and Zlatko Šabič, 191–215 Ljubljana: Znanstvena založba Filozofske fakultete UL.

Hōmushō 法務省 [The Ministry of Justice]. 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」[Kanji Characters for Personal Names (Ordinance for Enforcement of the Family Register Act, Article 60, Appended Table 2 “Tables of Kanji Characters”)]

Hsieh, Ching-chun, Jack Kai-tung Huang, Chung-tao Chang and 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. Accessed 10 Jan. 2019. https://www.computer- history.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. Accessed 12 Feb. 2019. http://languagelog.ldc.upenn.edu/nll/?p=19175.

Monbukagakushō 文部科学省 [Ministry of Education, Culture, Sports, Science and Technology]. 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） [Elementary School Curriculum Guidelines (published in March 2017)].

Moorhouse, A. 1953. *The triumph of the alphabet: A history of writing.* New York: H. Schuman.

Pandey, Krishna Kumar and Smita Jha. 2019. "Tracing the Identity and Ascertaining the Nature of Brahmi-Derived Devanagari Script." *Acta Linguistica Asiatica* 9, no. 1 (2019): 59–73.

Share, D. L. and Daniels, P. T. 2016. "Aksharas, alphasyllabaries, abugidas, alphabets and orthographic depth: Reflections on Rimzhim, Katz and Fowler (2014)." *Writing Systems Research* 8, no. 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.* Accessed 12 Mar. 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 January 2014. Accessed 14 Jan. 2019. http://cryptiana.web.fc2.com/code/chinese_e.htm.

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

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