

Character Sets and Characters: The Basis of Chinese Language Computing

Thomas A. Chan
Saint Paul, Minnesota

This article provides background information on the development of three lineages of character sets and encodings of relevance to Chinese language computing: the Guobiao family of character sets in China, the Big5 character set and its descendants in Taiwan and Hong Kong, and Unicode, the international character set. In addition, the state of support of the characters used in various orthographies and transcription systems as well as symbols of interest to Chinese language applications, including Han characters, Hanyu Pinyin, and the International Phonetic Alphabet, is assessed.

1. Character Sets and Encodings

The character, whether it be an element of a writing system or a symbol conveying non-linguistic information, resides in a character set, and as manifested in surface form through an encoding, forms the foundation of computer applications with text. As the technical use of the term “character” (*TUC* 2000: 985) conflicts with the occasional, everyday usage of the term to refer to those used in the Chinese writing system, the term “Han character,” as a translation of *hanzi* 漢字, is hereafter used for the latter.

1.1. ASCII and Single-byte Character Sets

At the basis of virtually every character set in use today is ASCII, a 7-bit character set encoding little more than the letters of the Latin alphabet necessary for writing English, hence its name, the American Standard Code for Information Interchange (Figure 1). ASCII is also known as the International Reference Version (IRV) of the ISO 646 standard from the International Organization for Standardization (ISO). In the past there had been other versions of ISO 646 for various countries, of which only a few remain today, such as Japan’s, where the most obvious difference is the substitution of the yen currency symbol ¥ (at code-point 5C in hexadecimal notation, or 92 in the more familiar, everyday decimal notation) in place of the backslash \. Other variations sometimes occurring in East Asia described by Lunde (1999: 76-78) are relatively minor, affecting at most

three symbols in each country's variant: the dollar currency symbol \$ (at code-point 24 hexadecimal, or 36 decimal), the backslash, and the tilde ~ (at code-point 7E hexadecimal, or 126 decimal).

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0	C0 Control Codes															
1																
2	[sp]	!	"	#	\$	%	&	'	()	*	+	,	-	.	/
3	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
4	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5	P	Q	R	S	T	U	V	W	X	Y	Z	[\]	^	_
6	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
7	p	q	r	s	t	u	v	w	x	y	z	{		}	~	[del]
8	Unused by ASCII															
9																
A																
B																
C																
D																
E																
F																

Figure 1. ASCII

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0	C0 Control Codes															
1																
2	[sp]	!	"	#	\$	%	&	'	()	*	+	,	-	.	/
3	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
4	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5	P	Q	R	S	T	U	V	W	X	Y	Z	[\]	^	_
6	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
7	p	q	r	s	t	u	v	w	x	y	z	{		}	~	[del]
8	C1 Control Codes															
9																
A	[del]	¡	¢	£	¤	¥	¦	§	¨	©	*	«	¬	-	®	¯
B	°	±	²	³	´	µ	¶	·	¸	¹	º	»	¼	½	¾	¿
C	À	Á	Â	Ã	Ä	Å	Æ	Ç	È	É	Ê	Ë	Ì	Í	Î	Ï
D	Ð	Ñ	Ò	Ó	Ô	Õ	Ö	×	Ø	Ù	Ú	Û	Ü	Ý	Þ	ß
E	à	á	â	ã	ä	å	æ	ç	è	é	ê	ë	ì	í	î	ï
F	ð	ñ	ò	ó	ô	õ	ö	÷	ø	ù	ú	û	ü	ý	þ	ÿ

Figure 2. ISO 8859-1 (West European)

ASCII and its variants use only 7 bits, which allows for 128 (2^7) codepoints, which only uses up half the space available in a byte, which has 8 bits, allowing for 256 (2^8) codepoints. The remaining space allowed the deficiencies in ASCII to be rectified by a proliferation of ASCII extensions created by standards organizations and computer vendors, such as the nearly ubiquitous ISO 8859-1 (Figure 2), for Western European languages such as French and German. Parallel to ASCII's avoidance of rows beginning with 00 and 01 hexadecimal, which are reserved for use by C0 control codes such as tabs, carriage returns, and line feeds, the rows beginning with 08 and 09 hexadecimal are avoided by ISO 8859-1 for C1 control

codes. Hence, there are really only ninety-six and 192 codepoints available for printable characters in 7-bit ASCII and a proper 8-bit extension of ASCII, respectively.

Just for Western European languages alone, other significant 8-bit extensions to ASCII in use today are both from computer vendors: Microsoft’s Windows CP1252 (CP = “codepage”) (Figure 3), a superset of ISO 8859-1, and Apple Computer’s MacRoman (Figure 4).

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
8	€		,	/	„	…	†	‡	^	%o	Š	<	Œ		Ž	
9		‘	’	“	”	•	—	~	™	š	>	œ		ž	ÿ	
A		ı	é	£	¤	¥	¦	§	¨	©	ª	«	¬	-	®	¯
B	°	±	²	³	´	µ	¶	·	¸	¹	º	»	¼	½	¾	¿
C	À	Á	Â	Ã	Ä	Å	Æ	Ç	È	É	Ê	Ë	Ì	Í	Î	Ï
D	Ð	Ñ	Ò	Ó	Ô	Õ	Ö	×	Ø	Ù	Ú	Û	Ü	Ý	Þ	ß
E	à	á	â	ã	ä	å	æ	ç	è	é	ê	ë	ì	í	î	ï
F	ð	ñ	ò	ó	ô	õ	ö	÷	ø	ù	ú	û	ü	ý	þ	ÿ

Figure 3. Windows CP1252 (West European) (excerpt)

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
8	À	Á	Ç	È	Ñ	Ó	Ù	á	à	â	ã	ä	å	ç	é	è
9	ê	ë	ı	ı	ı	ı	ñ	ó	ò	ô	õ	ù	ú	û	ü	
A	†	°	é	£	§	•	¶	ß	®	©	™	’	”	≠	Æ	Œ
B	∞	±	≤	≥	¥	µ	∂	Σ	Π	π	∫	ª	º	Ω	æ	ø
C	¿	ı	¬	√	/	≈	Δ	«	»	…	À	Á	Ï	Œ	œ	
D	—	—	“	”	‘	’	÷	∅	ÿ	ÿ	/	€	<	>	fi	fl
E	‡	·	,	„	%o	À	È	Á	É	Ê	Ì	Î	Ï	Ó	Ô	
F		Ò	Ù	Û	Ü	ı	^	~	—	—	·	°	”	”	”	”

Figure 4. MacRoman (excerpt)

While mutually incompatible with each other, Windows CP1252 and MacRoman both infringe upon the rows reserved for C1 control codes in order to encode thirty-two additional letters and symbols, such as Windows CP1252’s “curly quotes” (at codepoints 93 and 94 hexadecimal).

Eight-bit extensions to ASCII have also been created for other scripts and languages, including Arabic, Baltic, Central European, Cyrillic, Greek, Hebrew, Thai, Turkish, and Vietnamese. However, for some of these extensions, it has been difficult to fit all the necessary elements of the orthography, such as the VISCII (Vietnamese Standard Code for Information Interchange) character set devised by

Nguyen, et al. (1993), which infringes on rows reserved for both C0 and C1 control codes. The obvious solution to this kind of space problem is to introduce additional bytes, as has been done for East Asian character sets.

1.2. China's Guobiao Character Sets

The history of the *Guobiao* 國標 (< *Guojia biao zhun* 國家標準 'national standard') family of character sets in China can be traced through a chain of standards, beginning with the GB 2312 character set in 1980, to its expanded forms as the GBK (*Guobiao kuozhan* 國標擴展) specification in 1993, to the recent GB 18030 character set in 2000.

The GB 2312 character set consists of 7,445 characters, including 6,763 simplified-form Han characters, which are divided into two groups by frequency, a set of 3,755 and a set of 3,008. The former set is arranged by Pinyin reading,¹ while the latter is arranged by radical and number of strokes.

GB 2312 has since undergone a number of revisions, the first of which was GB 6345.1 in 1986, which adds 132 symbols, including additional Pinyin letters. According to Lunde (1999: 81-83), there was also GB 8565.2 in 1988, which provided a different set of revisions from GB 6345.1, and ISO-IR 165 in 1992, which in addition to its own revisions, also included those from GB 6345.1 and GB 8565.2. However, the GB 6345.1 revisions have been the most influential, and are incorporated into some "GB 2312" descriptions such as Lunde's "Appendix E GB 2312-80 Table" (1999: 638-652). Their acceptance has also been corroborated by their inclusion in the GB/T 12345 character set (T = *tuijian* 推薦 'recommended') released in 1990, which is essentially a traditional form Han character version of GB 2312, as well as GB 2312's successor, GBK, released in 1993.

Although GB 2312 may appear in surface forms such as the 7-bit safe HZ (*hanzi* 漢字) encoding devised by Lee (1995), it usually appears in EUC-CN (EUC = Extended Unix Code; CN = China) encoding, such that the latter is typically known as "GB 2312" encoding. While there is a technical distinction between a "character set" and the surface forms it appears in, called "encodings,"² it is not always observed, and

¹ Hence, a simple sort of GB 2312 data will have the effect of alphabetizing Han characters by their Pinyin readings, presuming they all belong to the first, more frequent set.

² For example, the Japanese character set JIS X 0208 can appear in EUC-JP (on Unix), ISO 2022-JP (for email), and Shift-JIS (on Windows) encodings, and it is common for software to convert between them.

unnecessary encumbrance when discussing a character set that only has one usual encoding; hence these terms will be used synonymously unless otherwise specified.

Unlike the 8-bit extensions to ASCII, GB 2312 employs the empty upper half of ASCII as the first of a two-byte sequence to encode Han characters and other symbols. As a result of the backwards compatibility with ASCII, the actual number of codepoints available is immediately reduced from 65,535 (256×256) to 32,768 (128×256). This figure is further reduced as follows: 1) by avoiding codepoints reserved for C0 and C1 control codes (00-1F and 80-9F hexadecimal); 2) by using only codepoints in the upper half of a byte (80-FF hexadecimal), complementing the use of codepoints in the lower half of a byte for ASCII characters; and 3) by restricting the space to a 94×94 square (in this case, A1-FE hexadecimal) for compatibility with some 7-bit encodings that cannot use the space of a 96×96 square. Since the maximum capacity is 8,836 (94×94), GB 2312's 7,445 characters and GB 6345.1's 132 additional characters fit comfortably within this space. A property of this design is that given any single byte, it is possible to determine if it is a single-byte ASCII character or part of a two-byte sequence (Figure 6).

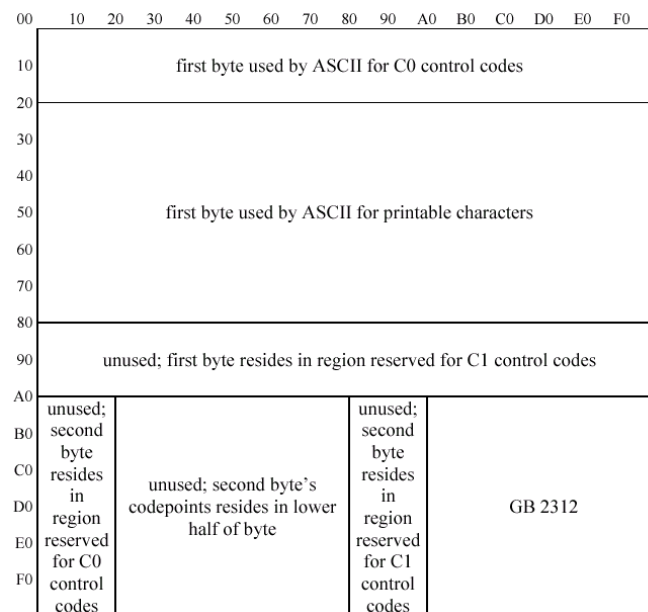


Figure 6. GB 2312

In 1993, the GB 13000.1 standard was published which, besides being essentially a translation of the Unicode 1.1 standard from that same year, also included an annex introducing GBK as the successor to GB 2312. By extending the possible ranges of two-byte sequences (from A1-FE hexadecimal in the first byte to 81-FE, and A1-FE in the second byte to 40-7E and 80-FE, skipping 7F), the capacity was increased to 23,940, with room to incorporate all 20,902 of Unicode 1.1's Han characters that were not already part of GB 2312, plus a small number of additional Han characters and other symbols (Figure 7). Essentially, one of features of the Unicode character set, a larger repertoire of Han characters, was gained, while retaining backwards compatibility with GB 2312.

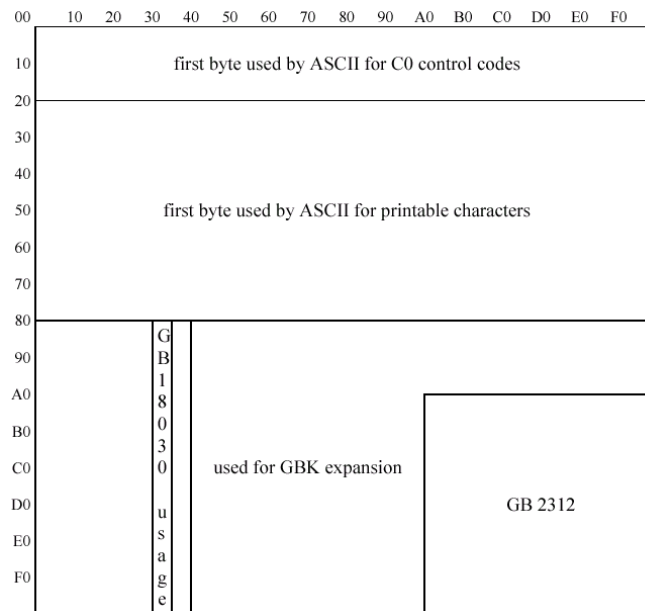


Figure 7. GBK and GB 18030

Just as GBK expanded GB 2312 to incorporate Han characters from Unicode 1.1, when the next version of the Unicode character set to include additional Han characters was published, the GB 18030 standard³ was published in 2000 to keep up. However, the 6,582 additional Han characters introduced as “CJK Extension A” in Unicode 3.0 in that same year could not fit into the remaining space

³ Further details of the standard may be found in Meyer (2001), who produced the first English translation of the standard, with commentary.

left in GBK, so four-byte sequences were introduced. These consisted of a pair of two-byte sequences (in the range of 81-FE hexadecimal in the first byte, and 30-39 in the second byte), creating over 1.5 million new codepoints. Not only did the 6,582 additional Han characters fit into this space, but provisions were made so that every Unicode character, including those not yet allocated, would have a mapping to GB 18030. This feature would soon see use, for in the following year, Unicode 3.1 was published, adding a staggering 42,711 Han characters. But more immediately, another consequence was that the writing systems of some ethnic minorities—Mongolian (classical script), Tibetan, Uighur (Arabic script), and Yi—became supported, at least in theory. Effectively, all the characters of Unicode were gained, while retaining backwards compatibility with both GBK and GB 2312.

The publication of GB 18030 also came with an important legal requirement; namely, products released in China after 1 September 2001 must support the standard, software released between 17 March 2000 and 31 August 2001 must be upgraded to support it, while software released prior to 17 March 2000 would be exempt. One consequence of this policy is that there is actually a lot of software that supports GB 18030, such as Microsoft's Windows 2000 and XP (Dr. Int'l 2001).

1.3. Taiwan's Big5 Character Set

The need for a character set in Taiwan was answered with the Big5 character set in 1984, which was not a national standard, but a *de facto* standard created by industry, hence one explanation (Lunde 1999: 89) for its name—referring to the five companies involved in its creation. Since then, its prolific development has diverged in different directions, including a separate lineage in Hong Kong.

Big5 consists of 13,494 characters, including 13,053 traditional form Han characters. Of that 13,053, 13,051 are unique, due to erroneous duplication of two Han characters, 兀 and 𪗇. The 13,053 Han characters are divided into two groups by frequency, a set of 5,401 and a set 7,652, each of which is arranged by total number of strokes and then radical.

Similar to the GB 2312 character set, Big5 employs the empty upper half of ASCII as the first of a two-byte sequence to encode Han characters and other symbols. However, its possible ranges of two-byte sequences had to be larger (A1-FE hexa-

decimal in the first byte, and 40-7E and A1-FE in the second byte, skipping 7F), for a capacity of 14,758, in order to accommodate the larger pool of characters (Figure 8).

The influence of Big5 was immediately felt in the design of the CNS 11643 character set⁴ (CNS = Chinese National Standard) in 1986, which included Big5's repertoire of Han characters, less the two duplicates, albeit in a different encoding, as its first two "planes." Although CNS 11643 had the blessing of being a national standard and would go on to develop in its own direction, growing to seven planes and 48,027 Han characters in its 1992 edition, Big5 remained a *de facto* standard.

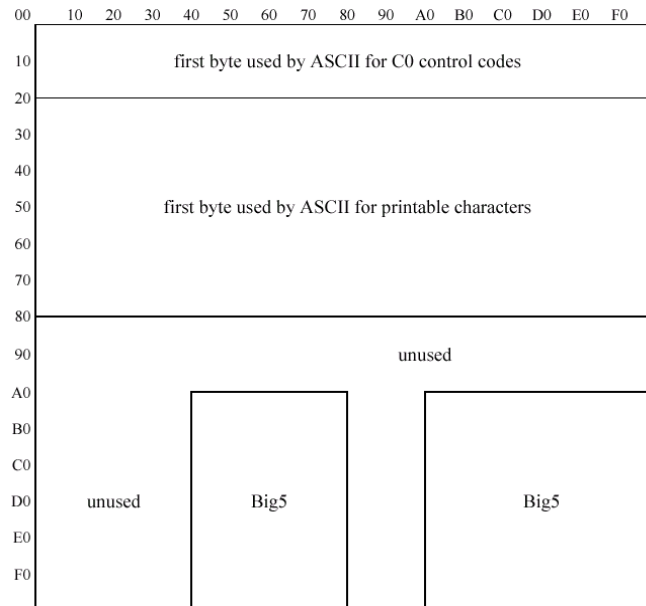


Figure 8. Big5

However, Big5 alone was not enough for some, and various extensions have been devised, such as ETen's 倚天, which provided Japanese kana and other symbols, plus seven additional Han characters (Lunde 1999: 559-561). The ETen extensions, in whole or in part, have been included in many implementations of Big5 such that it has become a *de facto* extension to a *de facto* character set, including Big5+ and Big5HKSCS, both prominent successors to Big5.

⁴ CNS 11643 website. (URL: <<http://www.cns11643.gov.tw/>>.)

At the end of the 1990s, CMEX⁵ made two attempts to expand Big5. The first was Big5+ in 1997, which arrived at similar results as GBK. By expanding the possible ranges of two-byte sequences (incidentally, to the same as GBK's), it could also include all the Han characters of Unicode 1.1. However, Big5+ has yet to meet the kind of success that GBK has; while GBK was defined in 1993 and implemented by the mid-1990s such as in the mainland Chinese edition of Windows 95; Lunde (1999: 91-92) observed in 1999 that Big5+ "... had yet to be fully implemented in any operating system," an observation which is still true.

The second attempt to expand Big5 was Big5E (E = "extended") in 1999, which expands Big5 in a different direction by including almost 4,000 Han characters from planes 3 and 4 of the 1992 edition of CNS 11643. The choice of a different set of additional Han characters was tailored for Big5E's use in government record-keeping. However, Big5E also has not emerged as a successful heir to Big5 in the same manner as the *Guobiao* line, which culminated with GB 18030.

1.4. Hong Kong's Big5 Extensions

The final years of the 1990s saw the rapid development of Big5 extensions to support characters used in Hong Kong, including those used in writing Cantonese. Using the same capability to define custom characters, *waizi* 外字, that allowed the creation and popularization of the now *de facto* ETen extensions, computer vendors and other organizations came out with a myriad of incompatible extensions.

Font vendors such as Taiwan-based DynaLab 華康 came out with two extensions, which Meyer (1998: 35-36) dubs "DynaLab HK A" and "DynaLab HK B," which encode 784 and 1,411 (including 664 Han characters) characters, respectively. Meanwhile, the Hong Kong branch of Monotype also came out with two character sets, dubbed by Meyer (1998: 36) as "Monotype-314" and "Monotype-471," where the latter is a superset of the former, and the number of characters (only Han characters in these cases) encoded is made explicit in Meyer's naming scheme. Although they are now obsolete and information about them is increasing harder to find, details of select extensions, "DynaLab HK A" and "Monotype-471," can be found in Meyer (1998: 35-36) and Lunde (1999: 564-569).

⁵ CMEX website. (URL: <<http://www.cmex.org.tw/>>.)

Organizations such as the Hong Kong University of Science and Technology came out with their own set in March 1996, consisting of 152 Han characters⁶, while around the same time period the Hong Kong-based Apple Daily 蘋果日報 newspaper used to provide its own downloadable font with additional characters for its online readership.

Even the Hong Kong government joined the bandwagon around 1994/1995 with the quiet release of the Government Chinese Character Set (GCCS) extension, featuring 3,049 characters (of which most, but not all, are Han characters, as commonly believed), for government-internal use. Meyer (1998: 36) was not able to discern "... when (or even whether) it has been officially published," but was able to determine that GCCS was an amalgamation of characters from various branches of the government.

The circumstances of its origins no doubt accounts for a number of its flaws, most of which have been inherited by its successor, HKSCS, including: 1) haphazard populating of the available space, where some areas are completely filled, while others are sparsely occupied with large gaps between characters; 2) no apparent order to the arrangement of Han characters, unlike the total strokes and radical scheme employed in Big5 proper; 3) duplication of Han characters that were already in Big5, detailed in Lunde (1991: 101); and 4) duplication of Han characters internally within GCCS due to lack of criteria for determining whether a character was significantly different enough to merit encoding it separately, and not within the boundaries of variation between font styles. There is also criticism that concerns the content of the standard, including: 1) miswritten Han characters; 2) characters of unverifiable origin or that cannot be found in dictionaries; and 3) "characters of the Cantonese dialect ... even foul language" (*HK Gov't* 2002). However, the explanation, as answered in various places in the HKSCS FAQ (*HK Gov't* 2002), is concerned solely about practical matters and not whether there are dialect characters or maledicta, as they do occur in publications and legal transcripts, nor the several hundred otherwise difficult-to-verify characters for names of people, companies, and places, which occur in government records and legal documents. Similar to other, earlier Big5 extensions, information about GCCS is also becoming harder to find; this is due to events such as the disappearance of the meager government site about GCCS, but a code chart may be examined in Lunde (1999: 804-813).

⁶ HKUST End-user Defined Characters. (URL: <<http://www.ust.hk/itsc/chinese/infra/eudc/>>.)

Due to the lack of definitive, official information about GCCS, a number of unusual implementations were engendered, including a GCCS-based add-on to GBK, the basis of the Pan-Chinese edition of Windows NT 4.0 (which crossed a traditional and simplified Chinese capable system with an English interface), which includes GCCS only in terms of repertoire and not encoding, providing a “faint image of the original GCCS” (Meyer 1998: 36-37). The exact contents of GCCS have also been in question, such as the “Monotype-3194” extension (Meyer 1998: 37), which includes an additional 145 Han characters, courtesy of the Hong Kong Department of the Judiciary (Lunde 1999: 37), but which in hindsight is understandable as post-1995 development of the standard.

Despite its defects, GCCS prevailed over the other extensions by virtue of its much larger repertoire and quasi-official status, as it was made mandatory in 1995 for computer products sold to the government to support it. As a consequence of the same GCCS-capable products being sold to the public, its hegemony was ensured.

In September 1999, the Information Technology Services Department (ITSD)—which had released GCCS—and the Chinese Language Interface Advisory Committee (CLIAC) came out with a revised version of GCCS called Hong Kong Supplementary Character Set (HKSCS), including a website⁷ to provide the specification, code charts, fonts, and input methods. This edition of HKSCS consisted of 4,702 characters (of which most are Han characters); however, this total was arrived at not only by adding 1,759 new characters, but also by correcting two of GCCS’s flaws by eliminating: 1) eighty-four Han characters that were duplicates, and 2) twenty-two Han characters that could not be verified. For backwards compatibility, however, the codepoints occupied by those 106 Han characters are reserved and were not reused (*HK Gov’t* 2001: Annex I: 1-5). Since 1999, an additional 116 Han characters have been added to HKSCS, for a total of 4,818 characters, which were published in the 2001 edition of HKSCS in December of that year.

Both the 1999 and 2001 editions of HKSCS, as well as its predecessor GCCS, are implemented completely by placing them within the user-defined regions of Big5, and hence can be conceptualized as a very large set of *waizi* 外字.

⁷ HKSCS website. (URL: <<http://www.info.gov.hk/digital21/eng/hkscs/>>.)

1.5. Unicode, the International Character Set

True to its claim of being a “universal character encoding scheme for written characters and text” (*TUC 2000*: 4), a common theme throughout the development of the Unicode standard has been the international nature of the collaboration and compromise. The earliest and most transforming instance was the 1991 merger between the Unicode 1.0 work done by the Unicode Consortium, an industry organization, and the ISO 10646.1 draft by ISO, a standards organization, both of whom recognized that they were working towards the same goal. Revisions were made over the following years to both standards to incorporate features from each other, which culminated in both the Unicode 1.1 standard and the ISO 10646-1 standard in 1993. Since then, efforts have been made to keep both standards synchronized with each other. As the Unicode standard has been historically more readily available than ISO’s, being distributed in book form, and since 2000, also online,⁸ a detailed history of their relationship as of 2000 may be found in “Appendix C Relationship to ISO/IEC 10646” (*TUC 2000*: 967-972).

The process by which the Han characters in Unicode were assembled, as described in “Appendix A Han Unification History” (*TUC 2000*: 961-962), is an archetypical example of the international contributions to Unicode. Beginning with the Chinese Character Code for Information Interchange (CCCII) character set developed in Taiwan, a progression is charted to its modification as the East Asian Character Code (EACC) character set from the United States for bibliographical use, to the Research Libraries Group’s (RLG) CJK Thesaurus used to maintain EACC, to work done at Apple’s based on RLG’s work, which led to a merger with work done at Xerox to form the first Unicode Han character draft in 1989. This was proposed for inclusion in ISO 10646, which created interest that ultimately led to the formation of the Chinese/Japanese/Korean Joint Research Group (CJK-JRG) in 1990, composed of East Asian countries and other organizations. A second Unicode Han character draft was merged with China’s GB 13000 character set, which was completed by the CJK-JRG in 1992, to form the initial set of 20,902 Han characters in Unicode. The following year, the CJK-JRG was renamed the Ideographic Rapporteur Group (IRG),⁹ reporting to a working group¹⁰ at ISO.

⁸ Online edition of Unicode standard book. (URL: <<http://www.unicode.org/book/u2.html>>.)

⁹ IRG website (URL: <<http://www.cse.cuhk.edu.hk/~irg/>>.)

¹⁰ ISO/IEC JTC1/SC2/WG2 website. (URL: <<http://std.dkuug.dk/JTC1/SC2/WG2/>>.)

The original set of 20,902 Han characters in Unicode 1.1 was a union of those from character sets in use in 1991, ensuring Unicode's viability. These were arranged by radical and stroke, which was culturally-neutral, according to the order given in four major dictionaries (in order of precedence): the *Kangxi Zidian* 康熙字典, the Japanese *Dai Kanwa Jiten* 大漢和辭典, the *Hanyu Da Zidian* 漢語大字典, and the Korean *Dae Jaweon* 大宇源. In 2000, Unicode 3.0 added 6,582 additional Han characters as "CJK Extension A," and in 2001, Unicode 3.1 added 42,711 additional Han characters as "CJK Extension B," including characters from non-character set sources such as the *Ci Hai* 辭海, *Ci Yuan* 辭源, *Hanyu Da Cidian* 漢語大詞典, *Hanyu Da Zidian* 漢語大字典, and *Kangxi Zidian* 康熙字典 dictionaries, as well as from works such as the *Siku Quanshu* 四庫全書 and the *Zhongguo Dai Baike Quanshu* 中國大百科全書. Currently, work is progressing at the IRG on a further "CJK Extension C1," which did not make it into Unicode 4.0 (April 2003).

It is clear that the Han characters in Unicode alone, which total over 70,000 as of Unicode 3.1, not to mention the characters of other scripts from around the world, cannot fit into a two-byte sequence the way that GB 2312 or Big5 do, even if backwards compatibility with ASCII were discarded and all 65,535 codepoints (256×256) were made available. The Unicode character set commonly appears in two encodings, UTF-16 (*TUC* 2000: 19-20) and UTF-8 (*TUC* 2000: 20, 47). UTF-16 has at its core an obsolete encoding once thought to be synonymous with Unicode, the ASCII-incompatible UCS-2, which uses only two bytes and thus is incapable of representing Unicode versions later than 3.0. However, UTF-16 extends UCS-2 with a mechanism similar to that used by GB 18030, in which a pair of two Unicode characters, called surrogates, is used to create room for approximately a million more characters. On the other hand, UTF-8 uses an ingenious mechanism that retains backwards compatibility with ASCII, while algorithmically encoding all of Unicode's characters as a variable-length sequence of one to four bytes.

2. Assessment of Support of Characters in Chinese Text

This section examines the state of support of select characters of interest to Chinese language applications.

2.1. Hanyu Pinyin

In addition to the letters of the Latin alphabet, the Hanyu Pinyin romanization system for Mandarin also uses two vowels with diacritics, *e* with a circumflex, *ê*, and *u* with an umlaut, *ü*. The letter transcribing each syllable's nucleus can be combined with one of four diacritics for tones, a macron, an acute accent, a caron, and a grave accent, respectively, e.g., *gā*, *gá*, *gǎ*, and *gà*. However, the existence of *ê* with or without any tone diacritic is not as widely recognized as that of *ü*, which is reflected in the lack of consensus of a common substitution as a typographic compromise or for typing purposes, like that of “uu,” “v,” and “yu” for *ü*, as well as its omission from some custom-encoded Pinyin fonts¹¹. This is despite its inclusion in dictionaries such as the *Xiandai Hanyu Cidian* 現代漢語詞典 (XHC 1995: 285), where it is used to transcribe interjections (*tanci* 嘆詞) in all four tones (usually written with 欸).

Even more often overlooked is the existence of *m* and *n* with diacritics, but they can be found in dictionaries such as the (PRC) *Ci Hai* 辭海 (1979: 733-734, 740, 746, 753, 755, 1101) and the *Xiandai Hanyu Cidian* 現代漢語詞典 (XHC 1995: 753, 811, 826), where they are used in syllables *m*, *n*, and *ng*. While some instances of unusual syllables may be dismissed as being beyond the scope of Pinyin or borderline cases in imitation of dialects, such as *m* in the first tone for 姆 of 媽媽, a word meaning ‘mother,’ and *m* in the second tone for 嘸, meaning ‘to not have,’ both Wu 吳 Chinese words; others, such as *m* in the second and fourth tones, as well as *n* in all but the first tone, are used in interjections (usually written with 欸).

The GB 2312 character set includes *a*, *e*, *i*, *o*, *u*, and *ü* in versions with all four tone diacritics, as well as an unadorned *ü*. However, only a plain *ê* with no tone diacritics is included. If the GB 6345.1 revisions are taken into account, as they were when they were incorporated into GB 2312's successor, GBK, then *n* with all tone diacritics except the first becomes available, as well as *m* in the second tone (but not the fourth tone).

In contrast, neither the Big5 character set nor its proposed successor, Big5+, include special letters used for Pinyin, but the HKSCS extension to Big5 is a different story. Similar to GB 2312, HKSCS includes *a*, *e*, *i*, *o*, *u*, and *ü* in versions

¹¹ Pinyin Fonts Online. (URL: <<http://www.foolshop.com/pfc/pinyinfonts.html>>.)

with all four tone diacritics, as well as an unadorned *ü*. But unlike GB 2312, *ê* with all four tone diacritics is included, as well as an unadorned *ê*. Uppercase versions of *a*, *e*, *o*, and *ê* in versions with all four tone diacritics, and an unadorned *Ê*, are also included, presumably for initially-capitalized (but not all-capitalized) text. However, these are concurrent with the lack of *m* and *n* with tone diacritics.

One character set which does not impose a decision over which interjections to not write is Unicode, which includes in precomposed form most of the Pinyin letters in both upper and lower case form. The few combinations that are not available in precomposed form can be represented by a base letter followed by combining diacritics (Figure 9), an option not available in other character sets. Alternatively, all Pinyin letters could be represented in decomposed form, e.g., *é* can be broken down into *ê* plus a combining acute accent (and *ê* can be broken down further into *e* plus a combining circumflex).

	a	e	ê	i	o	u	ü	m	n
	0061	0065	00EA	0069	006F	0075	00FC	006D	006E
̄	ā	ē		ī	ō	ū	ǔ		
0304	0101	0113		012B	014D	016B	01D6		
˘	á	é	ě	í	ó	ú	ǘ	ń	ň
0301	01E1	00E9	1EBF	00ED	00F3	00FA	01D8	1E3F	0144
ˇ	ǎ	ě		ǐ	ǒ	ǔ	ǚ		ň
030C	01CE	011B		01D0	01D2	01D4	01DA		0148
̀	à	è	è	ì	ò	ù	ù		ñ
0300	00E0	00E8	1EC1	00EC	00F2	00F9	01DC		01F9

	A	E	Ê	I	O	U	Ü	M	N
	0041	0045	00CA	0049	004F	0055	00DC	004D	004E
̄	Ā	Ē		Ī	Ō	Ū	Ǔ		
0304	0100	0112		012A	014C	016A	01D5		
˘	Á	É	Ě	Í	Ó	Ú	ǘ	Ŋ	Ň
0301	00C1	00C9	1EBE	00CD	00D3	00DA	01D7	1E3E	0143
ˇ	Ǻ	Ě		ǐ	ǒ	ǔ	ǚ		ň
030C	01CD	011A		01CD	01D1	01D3	01D9		0147
̀	À	È	È	Ì	Ò	Ù	Ù		Ñ
0300	00C0	00C8	1EC0	00CC	00D2	00D9	01DB		01F8

Figure 9. Pinyin letters in Unicode

2.2. Dialect Han Characters: A Case of Specialist Usage of Han Characters

Although specialists in history and literature certainly benefit from the inclusion of Han characters from sources such as the *Kangxi Zidian* 康熙字典 and the *Hanyu Da Zidian* 汉语大字典, another specialist group that gains from the

emergence of increasingly larger Han character repertoires in character sets are those who work with dialect characters. An upper bound for the degree of support of dialect characters may be determined by examining that of one of the more developed traditions, Yue 粵 Chinese, which is known from circumstantial evidence to have non-trivial support. Among the Han characters originally included in the Unicode character set is a small set of fifty-eight “Hong Kong” (along with ninety-two “Korean Idu” 吏讀) characters included as a “fictitious extension” (Lunde 1999: 131) of the GB/T 12345 character set when the latter was submitted for inclusion, while the HKSCS extension to the Big5 character set is known to include Cantonese dialect characters among its repertoire.

Since most sources do not adequately mark Cantonese dialect characters as such and for lack of a commonly-accepted definition and list, the list given in Chan (2001: 145-152) is adopted here. Chan (2001: 38-39) initially used the list given by Rao, et al. (1996: 377-380) as a data set of Cantonese dialect characters, but that list was soon found to include obscure words that were not familiar to contemporary Cantonese speakers and were not attested in other sources. In the interest of working with familiar words in contemporary use whose written form could be compared to those used in other sources, retained in that list were only those words that were also found in three prominent dictionary and dictionary-like sources: Lau (1977), Yue (1972), and Meyer (1947). That left words that were certainly in use in the past half century, representing one-third of the time period covered by the study (mid-eighteenth century to the present), and used commonly enough to be included in four out of the eight sources used. In addition, four words (and their characters) that did not meet that requirement had, nonetheless, been retained in Chan (2001) as exceptions, because they demonstrated important points.

There were 116 words total, of which 113 were monosyllabic, two disyllabic, and one trisyllabic. Between the aforementioned four sources, and four others—Williams (1856), Williams (1909 [1874]), Aubazac (1909), and O’Melia (1959)—there were 266 unique characters, of which seven were used to write two different words, while one character was used to write three different words.

The presence of the 266 Han characters in the Chan (2001) corpus were checked here against the repertoires of the GB 2312 (1980), Big5 (1984), CNS 11643 (1992 edition only), Big5HKSCS (2001 edition only), and the 1.1 (1993), 3.0 (2000), and 3.1 (2001) versions of the Unicode character sets. Since Unicode 3.2 subsumes all of the Han characters in the other character sets analyzed, the mappings between these and Unicode 3.2—as documented in the

Unihan 3.2 database file,¹² which contains a variety of information on the Han characters in Unicode—were used to determine their inclusion or omission from other character sets.

That the GB 2312 (1980) character set was found to be only capable of representing a dismal seventy-one out of 266, or 26.69%, of the Han characters in the corpus is not surprising, as it contains only simplified-form characters, and the corpus overwhelmingly consists of non-simplified forms, including characters drawn from historical sources that predate the mid-twentieth century simplification movement.

However, this does not account for the case of the Big5 (1984) character set, which, despite its traditional-form characters, only supported 137 out of 266, or 51.50%—little more than half. This is also in spite of the fact that the Big5 character set includes 13,053 (13,051 unique) Han characters in comparison to GB 2312's 6,763. As Big5 is the *de facto* character set used in Cantonese-speaking regions such as Hong Kong, it is no wonder that the HKSCS extension to Big5 remedies this situation by including Cantonese dialect characters among its over 4,000 Han characters, increasing the support of Han characters in the corpus drastically to 222 out of 266, or 83.46%.

In comparison, the Unicode 1.1 (1993) character set—whose 20,902 Han characters are a combination of those in commonly-used character sets as of 1991, including GB 2312 and Big5—only includes 194 out of 266, or 72.93%. The advent of Unicode 3.0 (2000), with the addition of “CJK Extension A” with 6,582 new Han characters, only brings the figure up to 207 out of 266, or 77.82%. It is not until the release of Unicode 3.1 (2001)—with the addition of “CJK Extension B” with 42,711 new Han characters, among them all of HKSCS’—that HKSCS’ 83.46% support is surpassed, with 246 out of 266, or 92.48%.

However, a larger repertoire of Han characters does not automatically imply greater support, as already shown by the cases of Unicode 1.1 (over 20,000 Han characters but only 72.93% support) versus Big5HKSCS (less than 18,000 Han characters but 83.46% support). Due to its obscurity, Taiwan’s actual national character set, CNS 11643, is not very relevant to the majority of people; nonetheless, it is enlightening to compare its support, as the 1992 edition of CNS 11643 contains 48,027 Han characters. Yet, it only supports 196 out of 266, or 73.68%, which is not very different from that of Unicode 1.1 (72.93%), despite being over

¹² Unihan 3.2 database (URL: <<http://www.unicode.org/Public/UNIDATA/Unihan.txt>>.)

twice its size. It should be noted, though, that the support that the 1992 edition of CNS 11643 does provide contributes to that of Unicode 3.1, which subsumes the 1992 edition of CNS 11643.

It may also be observed that fair to good (~75-90%) support of Cantonese dialect characters did not materialize until the end of the 1990s (Unicode 3.0 in 2000 with 77.82%, the 2001 edition of Big5HKSCS with 83.46%, and Unicode 3.1 in 2001 with 92.48%).

Since the corpus only includes Han characters used for common words, it is expected that these figures would be lower for rarer words. Additionally, although “unmarked phonetic loans” (so-called in Chan (2001: 58-62)) have been included in the corpus used here, e.g., usages such as 訓 for 訓, 左 for 左, and 野 for 嘢, etc., these are not the preferred forms seen in print for contemporary Cantonese dialect writing. They may, however, be used in the event of technical limitations, such as an online chatroom that uses Big5. Hence, the actual figures may be lower if non-preferred forms (which in many cases are also the historical forms) are excluded.

Since the inclusion of Cantonese dialect characters has been facilitated at least twice by systematic attempts to encode them, such as the “fictitious extension” to GB/T 12345 and HKSCS, as well as being one of the more well-known instances of dialect-writing and hence more likely to be documented in dictionaries that character sets draw upon, it is also expected that the support of the dialect characters used in other traditions will so far have been accidental. For example, 冇, a Min 閩 Chinese dialect character, meaning ‘hard’ (≠ soft), which is omitted in otherwise comprehensive dictionaries such as the *Hanyu Da Zidian* 漢語大字典, but included in Williams (1909 [1874]: 788) and in the *Hanyu Fangyan Cihui* 漢語方言詞匯 (1995: 499) as being another way to write 模, has been included in Unicode by virtue of its being in HKSCS (and often cited as an example of HKSCS’ contents), a character set extension which does not have obvious rationale to include such a character.

2.3. IPA and other Phonetic Transcription Systems

Among the character sets in common use, only Unicode offers the International Phonetic Alphabet (IPA). This presents an alternative to the commonly-used

but antiquated and custom-encoded IPA fonts¹³ (based on the 1990 and 1993 Kiel revisions) from the Summer Institute of Linguistics (SIL), who itself is moving away from proprietary encodings with the publication of a beta version of a Unicode-based IPA font¹⁴ in August 2002. Since IPA has been available in Unicode since the early 1990s, it was based on the 1989 revision, the latest available at the time, but additions have since been made, such as letters added in Unicode 3.0 for disordered speech from a 1997 revision of IPA. According to *The Unicode Standard Version 3.0* (TUC 2000: 164-165), obsolete and non-standard IPA letters are also included, as well as those used in Sinological, Americanist, and other traditions. Recent additions for Unicode 4.0 (April 2003)¹⁵ that will be published in hardcopy in September 2003 include the Uralic Phonetic Alphabet (UPA), proposed by Ruppel et al. (2002) and Everson et al. (2002); as well as four letters used by Sinologists and Sino-Tibetanists for palato-alveolars that are *t*, *d*, *l*, and *n* with the “curly-tail” diacritic seen in *ç* (Pinyin *x*) and two letters used for rounded apical vowels, the counterparts of unrounded [ɿ] and [ɿ̚] (used in some transcriptions for *i* as in Pinyin *shi* and *i* as in Pinyin *si*, respectively), proposed by Cook and Everson (2001)¹⁶.

Since the IPA has been considered part of the Latin alphabet (with the exception of a few letters unified with their counterparts in the Greek alphabet) and some of its letters have been adopted into the orthographies of African languages, it is not segregated into its own part of Unicode, but distributed among various blocks, including *C0 Controls and Basic Latin*, *C1 Controls and Latin-1 Supplement*, *Latin Extended-A*, *Latin Extended-B*, *IPA Extensions*, *Spacing Modifier Letters*, *Combining Diacritical Marks*, *Greek and Coptic*, and *Latin Extended Additional*. While they may be found by browsing the online Code Charts¹⁷ for those blocks and by recognizing them by appearance and/or description, many of them are deceptively similar to non-IPA letters and symbols. Although dated in relation to Unicode versions, an authoritative starting point for the letters and symbols to use is contained in “Appendix 2 Computer coding of IPA symbols” (161-185) of the International Phonetic Association’s *Handbook of the International Phonetic Association* (Cambridge: Cambridge University Press, 1999).

¹³ SIL Encore IPA Fonts. (URL: <<http://www.sil.org/computing/fonts/encore-ipa.html>>.)

¹⁴ SIL Unicode IPA Font beta. (URL: <http://www.sil.org/computing/fonts/ipa_unicode/>.)

¹⁵ Unicode 4.0.0 (April 2003). (URL: <http://www.unicode.org/standard/versions/enumeratedversions.html#Unicode_4_0_0>.)

¹⁶ Note that the codepoints suggested to be used in proposals are not necessarily the ones that are eventually adopted.

¹⁷ Unicode code charts (PDF version). (URL: <<http://www.unicode.org/charts/>>.)

2.4. Zhuyin Fuhao

Although *Zhuyin fuhao* 注音符號, originally known as *Zhuyin zimu* 注音字母 and alternatively as *Bopomofo*, is supported by both the GB 2312 and Big5 character sets (and their respective supersets), only Unicode includes the three letters that were excised from *Zhuyin fuhao* because they were later found to depict sounds that were not considered part of the emerging standard Chinese language (but still seen in sources as late as Mathews 1943): “v,” “ng,” and “gn,” which resemble the Han characters ㄩ, ㄥ, and ㄍ, respectively. A supplementary set of *Zhuyin fuhao* letters was later added in Unicode 3.0 (*TUC 2000: 278-279*) as the *Bopomofo Extended* block for the transcription of Southern Min and Hakka as spoken in Taiwan.

2.5. Yi Jing and Tai Xuan Jing Symbols

While the eight *Yi Jing* 易經 trigrams have been available in Unicode since the early 1990s, they were buried in the *Miscellaneous Symbols* block along with meteorological, zodiacal, chess, and other symbols. It was not until Unicode 4.0 that related complementary symbols were added, proposed by Cook et al. (2001). These consisted of two (2^0) monograms and four (2^2) digrams added to the *Miscellaneous Symbols* block, and sixty-four (2^6) hexagrams added to the new *Yi Jing* block, all of which are named for the number of whole or broken bars in each symbol.

A derivative set of symbols used in the more obscure *Tai Xuan Jing* 太玄經, consisting of bars that are whole, broken in two, or broken in three, were also added in Unicode 4.0, proposed by Cook et al. (2002). While all combinations would yield three (3^0) monograms, nine (3^2) digrams, and eighty-one (3^4) tetragrams, two monograms and four digrams are identical to those in the *Yi Jing* set. Due to their relative obscurity, the additional symbols used by the *Tai Xuan Jing* have instead been added to the new *Tai Xuan Jing* block in the Supplementary Multilingual Plane (SMP, also known as Plane 1), between the original Basic Multilingual Plane (BMP, also known as Plane 0) used by most scripts of the world, and the Supplementary Ideographic Plane (SIP, also known as Plane 2) used by the Han characters of CJK Extension B.

References

- Aubazac, Louis. 1909. *Liste des Caractères les Plus Usuels de la Langue Cantonnaise*. Hong Kong: Société des Missions-Étrangères.
- CH Ci Hai bianji weiyuanhui 辭海編輯委員會. 1979. *Ci Hai 辭海*. Small print edition. Shanghai: Shanghai cishu 上海辭書.
- Chan, Thomas. 2001. *Orthographic Change: Yue (Cantonese) Chinese Dialect Characters in the Nineteenth and Twentieth Centuries*. M.A. thesis. Columbus: The Ohio State University.
- Cook, Richard S., Jr., and Michael Everson. 2001. N2366R: Proposal to add six phonetic characters to the UCS. URL: <<http://linguistics.berkeley.edu/~rscook/pdf/UniProp-Final/n2366r-curly-tail.pdf>>.
- Cook, Richard S., Michael Everson, and John H. Jenkins. 2001. N2363: Proposal to add monogram, digram and hexagram characters to the UCS. URL: <<http://linguistics.berkeley.edu/~rscook/pdf/UniProp-Final/01283-n2363.pdf>>.
- Cook, Richard S., Michael Everson, and Michael Nylan. 2002. N2416: Proposal to add monogram, digram and tetragram Characters to the UCS. URL: <<http://linguistics.berkeley.edu/~rscook/pdf/UniProp-Final/02089-n2416.pdf>>.
- Dr. International. 2001. New Chinese encoding GB 18030. Ask Dr. International #15 (October). URL: <<http://www.microsoft.com/globaldev/DrIntl/columns/015/default.msp>>.
- Everson, Michael, Erkki Kolehmainen, Klaas Ruppel, and Trond Trosterud. 2002. N2442: Justification for placing the Uralic Phonetic Alphabet in the BMP. URL: <<http://www.dkuug.dk/jtc1/sc2/wg2/docs/n2442.pdf>>.
- HFC Beijing Daxue Zhongguo yuyan wenxue xi yuyanxue jiaoyanshi 北京大學中國語言文學系語言學教研室. 1995. *Hanyu Fangyan Cihui 漢語方言詞滙*. 2nd edition. Beijing: Yuwen 語文.
- HK Government. 2001. Hong Kong Supplementary Character Set—2001. URL: <<http://www.info.gov.hk/digital21/eng/hkscs/document.html>>.
- HK Government. 2002. Digital 21 HKSCS FAQ (March 26). URL: <http://www.info.gov.hk/digital21/eng/structure/cli_faq.html>.
- International Phonetic Association. 1999. *Handbook of the International Phonetic Association*. Cambridge: Cambridge University Press.
- Lau, Sidney 劉錫祥. 1977. *A Practical Cantonese-English Dictionary*. Hong Kong: The Government Printer.
- Lee, Fung Fung 李楓峰. 1995. RFC 1843: HZ—A data format for exchanging files of arbitrarily mixed Chinese and ASCII characters. URL: <<http://www.cis.ohio-state.edu/cgi-bin/rfc/rfc1843.html>>.

- Lunde, Ken. 1999. *CJKV Information Processing*. Sebastopol, CA: O'Reilly & Associates, Inc.
- Nguyen, Cuong T., Hoc D. Ngo, Cuong M. Bui, and Thanh van Nguyen. 1993. RFC 1456: Conventions for Encoding the Vietnamese Language Revision 1.1. URL: <<http://www.cis.ohio-state.edu/cgi-bin/rfc/rfc1456.html>>.
- Mathews, R.H., et al. 1943. *Mathews' Chinese-English Dictionary*. Revised American edition. Cambridge, MA: Harvard University Press.
- Meyer, Bernard F. and Theodore F. Wempe. 1947. *The Student's Cantonese-English Dictionary*. 3rd edition. New York: Field Afar Press. (First edition 1935.)
- Meyer, Dirk. 1998. Dealing with Hong Kong specific characters. *Multilingual Computing & Technology* 9.3 (April): 35-38.
- Meyer, Dirk. 2000. New Hong Kong character standard. *Multilingual Computing & Technology* 11.2 (March): 30-32.
- Meyer, Dirk. 2001. Summary, explanation, and remarks 1.4: GB 18030-2000. URL: <http://examples.oreilly.com/cjkvinfo/pdf/GB18030_Summary.pdf>.
- O'Melia, Thomas A. 1959. *First Year Cantonese*. 4th edition. Hong Kong: Catholic Truth Society. (First edition. 1938.)
- Rao Bingcai 饒秉才, Ouyang Jueya 歐陽覺亞, and Zhou Wuji 周無忌, eds. 1996. *Guangzhouhua Fangyan Cidian* 廣州話方言詞典. Hong Kong: Shangwu 商務. (Originally published in 1981.)
- Ruppel, Klass, Michael Everson, and Trond Trosterud. 2002. N2419: Uralic Phonetic Alphabet characters for the UCS. URL: <<http://www.dkuug.dk/jtc1/sc2/wg2/docs/n2419.pdf>>.
- TUC The Unicode Consortium. 2000. *The Unicode Standard Version 3.0*. Reading, MA: Addison-Wesley. Online edition at <<http://unicode.org/book/u2.html>>. See also Unicode Standard Annexes #27 (<<http://www.unicode.org/reports/tr27/>>) and #28 (<<http://www.unicode.org/reports/tr28/>>).
- Williams, Samuel Wells. 1856. *A Tonic Dictionary of the Chinese Language in the Canton Dialect*. Canton: Office of the Chinese Repository.
- Williams, Samuel Wells et al. 1909. *A Syllabic Dictionary of the Chinese Language Arranged According to the Wu-Fang Yüan Yin [五方元音] and Alphabetically Rearranged According to the Romanization of Sir Thomas F. Wade*. Tongzhou 通州: North China Union College. Originally published in 1874.
- XHC Zhongguo shehui kexueyuan yuyan yanjiusuo cidian bianjishi 中國科學院語言研究所詞典編輯室. 1983. *Xiandai Hanyu Cidian* 現代漢語詞典. 2nd edition. Beijing: Shangwu 商務. (First edition 1978.)
- Yue-Hashimoto, Oi-kan 余靄芹. 1972. *Studies in Yue Dialects 1: Phonology of Cantonese*. London: Cambridge University Press.