Unicode に見る文字コード国際化の現状と課題

指導教員：　渡辺　恭人

0940125

氏名　李　蓮

提出日:2013 年 12 月 14 日

目次

1．背景・目的

1－1　背景

現在、国により承認している国の数は様々ですが，日本が承認している国は 194 か国で，それに日本を加えて 195 か国(2011 年)が、日本政府が言うところの国の数になります。これらに北朝鮮（朝鮮民主主義人民共和国）を加えた 196 か国（日本が承認している国＋日本＋北朝鮮）が独立国の数となります。その中、東アジアに属している国は日本、中国、韓国、北朝鮮の 4 か国であります。この 4 か国の間では、国の領域のことや、拉致問題や、歴史の捉え方の違いや、慰安婦のことなどで 20 世紀初めのような武器を持っての戦いではありませんが、政治や、歴史、技術の面で常に闘ってきました。

葛藤が続くなか、アジア固有の価値を高め、未来の世代に向けた交流の活性化を目的としている「日中韓 30 人会」があります。韓国メディアによると「日中韓 30 人会」は、今年 7 月に北海道で第 8 回会議を開き共通漢字 800 字を選定したそうです。

日本、中国、韓国、北朝鮮の 4 か国についての共通点の一つとしては、「漢字」があります。中国は５万字ほどの漢字が使われており、漢字の国とも言われています。中国には今から 1500 年くらい前、「千字文」という本がありました。この本は当時子供たちに字を教えるために作られた本だそうです。この本は後に朝鮮半島に伝わり、またそこから日本にも伝わって、東アジアの漢字文化に大きく貢献をしました。

日本語で使われる文字にはひらがなとカタカナと漢字があります。ひらがなは漢字をくずした文字、カタカナは漢字の一部を取り出した文字、漢字は日本で作られた文字と中国から入った文字の両方あります。韓国は 1948 年朝鮮半島から独立して一つの国家になりました。韓国で、漢字はニュースなど外国の名前を漢字で書いたり、必要に応じて自分の名前を漢字で書いたりしています。昔はもっと盛んに使われていましたが、グローバルの進出によって衰退しかけています。

私は、朝鮮半島の子孫である中国人で今は日本に住んでいます。あいうえおと言った日本語、ｂｐｍｆｄｔｎｌ（いわゆるピンイン）と言った中国語、ㄱㄴㄷㄹㅁㅂㅅㅇ（いわゆる字音）と言ったハングルもわかりますし、この国々に興味を持っています。それによって、この国々がますます発展することを心より願って

います。

1－2　目的

本研究では、まずこの国々の共通点の「漢字」から今現在インターネット時代の「文字コード」を見出し「文字コード」が持つ国際化について考えてみます。アメリカで誕生したコンピュータとインターネットの仕組みは英語圏の人々にとって必要な文字や記号をすべて表現することができ完璧な仕組みとも言われていますが、実は我々の漢字やいろいろな文字を使っている日本、中国、台湾、韓国、北朝鮮などの国々や地域にはいろいろ不便なところがあることに気付きその解決策について検討します。

次に「文字コード」から「インターネットによる情報化」に範囲を広げ、世界中国々や地域・隅々まで「インターネットによる情報化」に目を向いて頂けるよう声をかけます。

2．現状

2－1　文字コードとは

文字コードとは、コンピュータ上、あるいはネットワーク上で文章情報処理を行い、文字を処理するための文書の表現形式である。文字コードはファイル中やネットワーク上で使われ、プログラム間やホスト間で文書をやりとりするための処理コード、つまり内部コードに分類することもできます。

文字に限らずデジタルデータは、0 と 1 の 2 進数でやり取りされていますので、日本語やアルファベットなどの文字や記号も、数値と対応させたコードに変換してやり取りされています。つまり、「A」という文字を、例えば「1000001」という具合にコード化して処理するわけです。このように、文字とコードとの対応規則をまとめたものを文字コードと言います。全ての文字は、文字コードというコード体系に置き換えられて、さらに最終的には 2 進数に変換されてコンピュータで処理されます。

デジタルデータは 0 と 1 の 2 進数で成り立っています。したがって、データサイズの最小値は「0」もしくは「1」の数字 1 桁分のサイズが最小値になります。

このデジタルデータ 1 桁分の最小単位を、ビット（bit）といいます。ただし、bit（ビット）単位では、サイズが小さすぎて扱いやすいものではありません。1 ビットでは「0」か「1」の 2 通りのデータしか扱うことができないためです。そこで、通常はいくつかのビット単位をまとめて、ある程度の数値まで表現できるようにパック化したビットの集合体を最小単位として利用します。

このようなビットの集合体の単位規格はいくつか存在しますが、最も一般的な単位は、バイト（bite）と呼ばれる単位です。バイト（bite）のビット数は、8 ビット（1 バイト=8 ビット）になります。つまり、8 桁の 2 進数を 1 単位とするのが「バイト」です。この「バイト」はパソコンに関連する単位の中で、 私たち一般のユーザーが最も多く接する単位 となります。ビットは小文字の「b」や「bit」、バイトは大文字の「B」の表記がよく使われます。

では、なぜ 8 ビットなのかというと、半角英数字 1 文字分のデータサイズになるからです。コンピュータはアメリカ生まれなので、英語が基本となります。コンピュー

タに関することは英語の使用のみが前提になっていますので、半角英数字（アルファベットと数字）を表現することのできる最低限のビット数を最小単位として定めています。

8 桁の 2 進数では、0 から 255 までの、256 通りを表現することができます。256 通りを表現することができれば、半角英数字はすべて表現することができます。アルファベットは A から Z まで 26 文字で、小文字を加えても 52 文字しかありませんので、これに数字を 10 種類加えても 62 とおりです。さらに、&・<>などの記号を加えたとしても 256 文字なら十分に足りるので、1 バイトあれば、英語圏の人々にとって必要な文字や記号をすべて表現することができるというわけです。

もっとも、日本語や中国語などの漢字を扱う人々にとっては、256 種類ではすべての文字を表現することはできません。日本人が使う「常用漢字」だけでも軽く千を超えるからです。コンピュータをはじめとする IT 技術はアメリカ生まれアメリカ育ちなので、他の国の例えばアジア人のことなど想定していなかったと考えられます。

2－1－1　ASCII とは

ASCII（アスキー；American Standard Code for Information Interchange）は、アメリカの JIS に相当する団体である ANSI が 1963 年に制定した文字コードです。ASCII は世界中に広まり、ASCII をもとにして 1967 年に ISO R 646 という国際的な推奨規格が制定され、これをもとに各国が自国の標準文字コードをつくりました。この ISO R 646 には、文字コードを 6 ビットで表す案と 7 ビットで表す案が併記されていましたが、1973 年に 7 ビットに一本化されて ISO 646 という正式規格になりました。この ISO 646 には、BCT（Basic Code Table）と IRV（International Reference Version）の 2 種類のコード表が含まれています。BCT は全世界共通の文字からなり、12 文字からなる IRV にはナショナル・ユース・ポジションとして各国独自の文字を割り当てることができるようになっており、各国の通貨記号や、フランスではアクサンなどが、ドイツではウムラウトなどが割り当てられました。ISO 646 は 1991 年にさらに改正されています。世界中には様々な文字コードが存在しますが、英数字部分はほとんどがこの ASCII コードを基に作られており、ASCII はコンピュータ用の標準コードとして最も普及しています。

表 2－1：　ASCII コード

ASCII（アスキー）コード									
	0b	000	001	010	011	100	101	110	111
0b	0x	0	1	2	3	4	5	6	7
0000	0	制御文字群		SP	0	@	P	`	p
0001	1			!	1	A	Q	a	q
0010	2			″	2	B	R	b	r
0011	3			#	3	C	S	c	s
0100	4			$	4	D	T	d	t
0101	5			%	5	E	U	e	u
0110	6			&	6	F	V	f	v
0111	7			’	7	G	W	g	w
1000	8			(	8	H	X	h	x
1001	9			)	9	I	Y	i	y
1010	A			*	:	J	Z	j	z
1011	B			+	;	K	[	k	{
1100	C			,	<	L	＼	l	|
1101	D			−	=	M	]	m	}
1110	E			.	>	N	^	n	~
1111	F			/	?	O	_	o	DEL

(1)　ASCII コードの符号化

ASCII コードの符号化は、7 ビットの 2 進数で表現されています。たとえば、″J″という文字の符号化は 2 進数で 1001010 になりますが、それを 16 進数に変換して 4A と表記します。

表 2－2：「ASCII コードの富豪と文字の対応一覧」

	上位 3 ビット→								
下位4ビット↓		0	1	2	3	4	5	6	7
	0	NUL	DLE	SP	0	@	P	`	p
	1	SOH	DC1	!	1	A	Q	a	q
	2	STX	DC2	″	2	B	R	b	r
	3	ETX	DC3	#	3	C	S	c	s
	4	EOT	DC4	$	4	D	T	d	t
	5	ENQ	NAC	%	5	E	U	e	u
	6	ACK	SYN	&	6	F	V	f	v

	7	BEL	ETB	’	7	G	W	g	w
	8	BS	CAN	(	8	H	X	h	x
	9	HT	EM	)	9	I	Y	i	y
	A	LF/NL	SUB	*	:	J	Z	j	z
	B	VT	ESC	+	;	K	[	k	{
	C	FF	FS	,	<	L	＼	l	|
	D	CR	GS	−	=	M	]	m	}
	E	SO	RS	.	>	N	^	n	~
	F	SI	US	/	?	O	_	o	DEL

(2) ASCII コードの構成

印字可能な文字部分は、図形文字の 21～7E の範囲にある文字で、その他は「制御文字」です。制御文字は普通の文字ではなく、文字データの中に混じって特別な機能を表すものです。

表 2－3：制御文字の一覧

16 進	コード	解説		16 進	コード	解説
00	NUL	ヌル(空文字)		01	DC1	制御装置 1
01	SOH	ヘディング開始		02	DC2	制御装置 2
02	STX	テキスト開始		03	DC3	制御装置 3
03	ETX	テキスト終了		04	DC4	制御装置 4
04	EOT	伝送終了		05	NAC	否定応答
05	ENQ	問い合わせ		06	SYN	同期文字
06	ACK	肯定応答		07	ETB	伝送ブロック終了
07	BEL	ベル		08	CAN	取消
08	BS	バックスペース		09	EM	媒体終端
09	HT	水平タブ		0A	SUB	
0A	LF/NL	復帰/改行		0B	ESC	(制御コード)拡張
0B	VT	垂直タブ		0C	FS	ファイルセパレータ
0C	FF	改ページ		0D	GS	グループセパレータ
0D	CR	復帰		0E	RS	レコードセパレータ
0E	SO	シフトアウト		0F	US	ユニットセパレータ
0F	SI	シフトイン		20	SP	(半角)スペース
10	DLE	データリンクでの拡張		7F	DEL	削除

制御文字とはディスプレイやプリンタを制御する特別な文字で、画面に表示されることはありません。改行(CR)やエスケープ(ESC)、タブ(TAB)などが制御文字になります。ASCII コードでは 00～1F と 7F の範囲に制御文字が配置されています。

(3) 8 ビット拡張 ASCII コード

ISO646 で表現できるのはアルファベットを使う国々に限定され、文字体系が異なるロシアやギリシャ語などではさらに拡張が必要でした。ASCII コードは最初の桁を常にゼロとしているので、1 バイトで扱える文字は 128 種類、実質は、7 ビットしか使っていません。そこで、最初の桁も有効に使えるように ISO646 を 8 ビットに拡張し、 8/0～F/F の領域に文字を追加できるようにしました。この仕組みを「拡張 ASCII コード」 と呼びます。拡張 ASCII コードでは、拡張した領域を使って 96 種類の文字の追加削除が可能になっています。この 96 の領域を各国の必要に応じて追加文字のセットを順次制定しました。それが「ISO8859」です。

表 2－4：ISO の 8 ビットコードの構成

文字	1 バイト領域
制御文字	0x00 ～ 0x1F, 0x7F
空白	0x20
図形文字	0x21～0x7E
制御文字の領域	0x80～0x9F
追加の図形文字	0xA0～0xFF

2－1－2 マルチバイトコード

日本語や中国といった文字は数千字、数万字を上回り、最低でも 2 バイトを使って 1 文字を表現する必要があります。このような 1 バイトを超える文字コードを「マルチバイトコード」と呼びます。

日本語を表現できるマルチバイトコードの種類には、大きく分けて JIS、シフト JIS、 EUC、 UNICODE があります。

(1) JIS コード

表 2－5：JIS コード範囲

文字	1 バイト目の領域	2 バイト目の領域
制御コード	00～1F、7F	－
ASCII	20～7E	－
半角カタカナ	21～5F（7 ビット）/ A1～DF（8 ビット）	－
漢字	21～7E	21～7E
補助漢字	21～7E	21～7E

JIS コードは、1 バイト文字として ASCII と半角カタカナがあり、2 バイト文字として漢字が追加されています。JIS コード には、ASCII の倍の幅で表示するアルファベットや数字などの文字があります。このことから、JIS 漢字の半分の幅の ASCII 文字は「半角文字」、JIS 漢字は「全角文字」と呼ばれています。

ISO2022

マルチバイトコードによって、文字数の多い言語もカバーできるようにはなりましたが、複数の国の言語を混在させて使うということはできません。複数の国の言語を支障なく表現するには、一括してすべての言語を表現できる文字セットを制定するか、必要に応じて文字セットを切り替えて使えるようにするかの方法が考えられます。後者の文字セットを切り替える方式は、ISO によって符号拡張法「ISO2022」が標準化されています。JIS 規格では「JIS X 0202」 です。ISO2022 は文字セットを切り替えるための規格で、ASCII コードのような文字セットではありません。

ISO2022 は、「エスケープシーケンス」と呼ばれる特殊な符号を使って、文字セットを切り替える方法を定めた国際規格です。これによって、マルチバイト文字や ISO646 各国語版などの複数の文字セットを混在して使うことが可能になりました。先に解説した JIS コードも、ISO2022 準拠の符号拡張方式を適用することで、1 バイトの ASCII や JIS ローマ字と切り替えながら共に用いることができます。

ISO-2022-JP

多くのネットワークは 7 ビットでメッセージの交換を行うので、ネットワークでの JIS コードは 7 ビット版が標準です。7 ビット JIS コードは、ASCII と同じように下位 7 ビットだけを使い、7 ビットコード 2 文字を組み合わせて日本語 1 文字を表します。これが ISO によって国際化され、ISO-2022-JP となりました。

表 2－6：ISO-2022-JP のエスケープシーケンス

切り替えられる文字コード	コード	エスケープシーケンス
ASCII（1 バイト）の開始	1B 24 40	[ESC]（ B
JIS ローマ字（1 バイト）の開始	1B 24 42	[ESC]（ J
JIS-1978（漢字 2 バイト）の開始	1B 28 4A	[ESC] $ @
JIS-1983（漢字 2 バイト）の開始	1B 28 42	[ESC] $ B

(2) シフト JIS（MS 漢字コード）

MS-DOS で日本語文字を表すコードとして、マイクロソフト社とアスキー社が共同開発したのが「シフト JIS」です。MS 漢字コードとも呼ばれ、MS-DOS やウインドウズシリーズ、一部の UNIX 系 OS や、マッキントッシュなどでも使用されています。シフト JIS の名前の由来は、JIS コードを簡単な計算で変換していることにあるのですが、もともと JIS 規格ではなく、ISO2022 にも準拠していなかったです。

表 2－7：シフト JIS コード範囲

文字	1 バイト目の領域	2 バイト目の領域
制御コード	00～1F、7F	-
ASCII 文字	20～7E	-
半角カタカナ	A1～DF	-
漢字	81～9F、E0～FC	40～7E、80～FC

(3) EUC

「EUC」は AT&T 社によって定められ、UNIX 環境での事実上の標準日本語コードと

なっています。日本語 EUC も JIS コードと同じく JISX0208 の文字セット規格を ISO2022 に基づいて符号化しています。EUC 自身は、日本語だけでなく複数の文字セットを同じテキスト内で処理することが可能です。

CGI プログラムで発生する文字化けのほとんどは、文字コードを EUC にすることで解決できます。

表 2－8：EUC コード範囲

文字	1 バイト目の領域	2 バイト目の領域	3 バイト目目の領域
制御コード	00～1F、7F		
ASCII	21～7E	–	–
半角カタカナ	8E	A0～FF	–
漢字	A0～FF	A0～FF	–
補助漢字	8F	A0～FF	A0～FF

(4) Unicode

Unicode とは、一言でいうと全世界の文字を 16 ビットで表現することを目指す規格であります。

UCS-4

一般に UCS と呼ばれている符号化で、1 文字あたり 4 バイト使用するために UCS-4 とも呼ばれます。コード表は 0 ～ 7F FF FF FF の領域を使い、約 21 億文字を扱えます。UCS-4 では UCS-4 と UTF-8 の 2 種類のコーディングが可能です。

UCS-2

UCS-2 は ISO10646 を 2 バイトで符号化する方法です。単に Unicode といえばこの 16 ビットの符号化を指します。コード表は 0 ～ 10 FF FF の領域を使い、約 111 万文字を扱えます。USC-2 のエンコード方式として、UTF-16 と UTF-7 の 2 種類があります。

UTF

既存のネットワークなどは ISO2022 に準拠した文字データを処理することを前提に

していますので、UCS のデータをそのまま利用すると問題が生じています。そのため、UCS を問題なく利用できるようにいくつかのエンコード方式が考案されました。いくつかのエンコード方式を総称した呼び名が UTF です。

UTF のエンコーディングは、互換性を考慮して UCS の ASCII に当たる部分を通常の ASCII コードに変換し、それ以外のコードをそれぞれの方式で変換します。

UTF-7

UTF-7 は、インターネットで利用できるように ASCII に当たる部分以外の 2 バイト文字を、Base64 エンコードや uuencode と似た方式によって 7 ビットで符号化するエンコード方式です。

2－2　東アジアにおける主要言語と文字

2－2－1　日本

現在の日本で最も広く使用されている言語は日本語であり、使用人口は文部科学省によれば 1 億 2500 万人だそうです。

日本語で使われる文字にはひらがなとカタカナと漢字があります。漢字は中国から入った文字、ひらがなは漢字をくずした文字、カタカナは漢字の一部を取り出した文字です。

表 2－9：中国の漢字から今のひらがなが生まれる

発音	元の字とひらがな	発音	元の字とひらがな	発音	元の字とひらがな	発音	元の字とひらがな	発音	元の字とひらがな
″a″	安あ	″sa″	左さ	″na″	奈な	″ma″	末ま	″ra″	良ら
″i″	以い	″shi″	之し	″ni″	仁に	″mi″	美み	″ri″	利り
″u″	宇う	″su″	寸す	″nu″	奴ぬ	″mu″	武む	″ru″	留る
″e″	衣え	″se″	世せ	″ne″	禰ね	″me″	女め	″re″	礼れ
″o″	於お	″so″	曾そ	″no″	乃の	″mo″	毛も	″ro″	呂ろ
″ka″	加か	″ta″	太た	″ha″	波は	″ya″	也や	″wa″	和わ
″ki″	幾き	″chi″	知ち	″hi″	比ひ	–		″wi″	為ゐ
″ku″	久く	″tsu″	州つ	″fu″	不ふ	″yu″	由ゆ	″n″	尓ん
″ke″	計け	″te″	天て	″he″	部へ	–		″we″	恵ゑ
″ko″	己こ	″to″	止と	″ho″	保ほ	″yo″	与よ	″o″	遠を

表 2－10：中国の漢字から今のカタカナが生まれる

発音	元の字とカタカナ	発音	元の字とカタカナ	発音	元の字とカタカナ	発音	元の字とカタカナ	発音	元の字とカタカナ
″a″	阿ア	″sa″	散サ	″na″	奈ナ	″ma″	万マ	″ra″	良ラ
″i″	伊イ	″si″	之シ	″ni″	二ニ	″mi″	三ミ	″ri″	利リ
″u″	宇ウ	″su″	須ス	″nu″	奴ヌ	″mu″	牟ム	″ru″	流ル
″e″	江エ	″se″	世セ	″ne″	禰ネ	″me″	女メ	″re″	礼レ
″o″	於オ	″so″	曾ソ	″no″	乃ノ	″mo″	毛モ	″ro″	呂ロ
″ka″	加カ	″ta″	多タ	″ha″	八ハ	″ya″	也ヤ	″wa″	和ワ
″ki″	幾キ	″ti″	千チ	″hi″	比ヒ	–		″wi″	井ヰ
″ku″	久ク	″tu″	州ツ	″fu″	不フ	″yu″	由ユ	″n″	？ン
″ke″	介ケ	″te″	天テ	″he″	部ヘ	–		″we″	恵ヱ
″ko″	己コ	″to″	止ト	″ho″	保ホ	″yo″	与ヨ	″o″	乎ヲ

漢字には中国から伝来したものと、それを真似て日本で作ったものとがある。漢字は 270～310 年頃に「論語」、「千字文」が百済から到来したことが公式的に初めてとされています。

JIS：日本では、JIS（日本工業規格）が文字の規格を定めている。ISO 2022 の日本語の文字集合を ISO 2022-JP と呼んでいます。当初は、ASCII、JIS X 0201 のローマ字部分、JIS X 0208（JIS X 0208:1978 および JIS X 0208:1983）から構成されており、一般に JIS 漢字や JIS コードと呼ばれていました。後に JIS X 0212、JIS X 0213 が追加されました。

Shift JIS：Shift JIS は、MS 漢字コードとも呼ばれているもので、MS、つまりマイクロソフト社が作ったコード体系であります。パソコンで使われる漢字コードとして最も普及しています。

EUC ：UNIX では、EUC というコードが標準です。「拡張」というのは国際化に対応させるため、各国の言語を表示できるようにしたからです。但し、UNIX 機でも、IBM の AIX 機やヒューレッド・パッカード社の HP-UX などでは Shift-JIS が使われてい

ることがあります。EUC は日本語版、中国語版、韓国語版があり、日本語版の EUC を EUC-JP と呼んでいます。

JIS 拡張漢字（JIS X 0213:2004）には 11,233 字（漢字 10,050 字、非漢字 1,183 字）が収録されています。

２－２－２中国

現在、中国でよく使われている言語は中国語です。中国語を母語とする人は約 12 億人、第二言語としても約 2 億人が使用しているといわれており、世界最大の話者人口を有しています。ギネスブックによれば「現存する世界最古の言語」だそうです。

中国語の文字には漢字と拼音があります。漢字には、繁体字と、簡体字があります。繁体字は 1960 年代以前に中国で用いられていた伝統的な漢字で、伝統字や正体字とも呼ばれ、現在でも台湾や香港で用いられています。簡体字は、中国が繁体字を簡略化して 1964 年に「簡化字総表」としてまとめた漢字で、中国本土やシンガポールで用いられています。ちなみに、見て分かるように「漢字」と「汉字」だと「漢字」が繁体字で「汉字」が簡体字です。

拼音（ぴんいん）とは B p m f d t n l g k h j q x zh ch sh r z c sのことです。

1980 年に国家標準総局が簡体字の文字コード GB 2312（国家标准） 2312 = 信息交换用汉字编码字符集 基本集）が制定されました。GB コードや GB 基本漢字とも呼ばれています。その後もいくつかの国家規格が制定されている。GB 2312 を採用した EUC 文字コードを、EUC-CN（簡体字中国語 EUC）と呼んでいます。

GB18030-2005 には 70,244 字が収録されています。

GB 2312-80

GB 2312-80（信息交换用汉字编码字符集基本集）の表です。すべて Unicode に収録されており、数値文字参照で記述することができます。ただし、オペレーティングシステムやブラウザのバージョンなどの環境によっては、文字化けする可能性があります。また、Unicode のユニフィケーションにより日中で共通のコードが与えられて

いる文字は、日本語環境では日本の字形で表示されることがあります。

表 2－11：GB 2312-80

非漢字（01 区～09 区）

区位	GB	EUC	+0	+1	+2	+3	+4	+5	+6	+7	+8	+9	+A	+B	+C	+D	+E	+F
01 区	2120	A1A0			、	。	·	ˉ	ˇ	¨	〃	々	—	～	∥	…	‘	’
	2130	A1B0	“	”	〔	〕	〈	〉	《	》	「	」	『	』	〖	〗	【	】
	2140	A1C0	±	×	÷	∶	∧	∨	∑	∏	∪	∩	∈	∷	√	⊥	∥	∠
	2150	A1D0	⌒	⊙	∫	∮	≡	≌	≈	∽	∝	≠	≮	≯	≤	≥	∞	∵
	2160	A1E0	∴	♂	♀	°	′	″	℃	$	¤	¢	£	‰	§	№	☆	★
	2170	A1F0	○	●	◎	◇	◆	□	■	△	▲	※	→	←	↑	↓	=	

区位	GB	EUC	+0	+1	+2	+3	+4	+5	+6	+7	+8	+9	+A	+B	+C	+D	+E	+F
02 区	2220	A2A0																
	2230	A2B0		1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.	14.	15.
	2240	A2C0	16.	17.	18.	19.	20.	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
	2250	A2D0	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)	①	②	③	④	⑤	⑥	⑦
	2260	A2E0	⑧	⑨	⑩			㈠	㈡	㈢	㈣	㈤	㈥	㈦	㈧	㈨	㈩	
	2270	A2F0		Ⅰ	Ⅱ	Ⅲ	Ⅳ	Ⅴ	Ⅵ	Ⅶ	Ⅷ	Ⅸ	Ⅹ	Ⅺ	Ⅻ			

・

・（途中略）

・

区位	GB	EUC	+0	+1	+2	+3	+4	+5	+6	+7	+8	+9	+A	+B	+C	+D	+E	+F
08	2820	A8A0		ā	á	ǎ	à	ē	é	ě	è	ī	í	ǐ	ì	ō	ó	ǒ
	2830	A8B0	ò	ū	ú	ǔ	ù	ǖ	ǘ	ǚ	ǜ	ü	ê					
	2840	A8C0						ㄅ	ㄆ	ㄇ	ㄈ	ㄉ	ㄊ	ㄋ	ㄌ	ㄍ	ㄎ	ㄏ
	2850	A8D0	ㄐ	ㄑ	ㄒ	ㄓ	ㄔ	ㄕ	ㄖ	ㄗ	ㄘ	ㄙ	ㄚ	ㄛ	ㄜ	ㄝ	ㄞ	ㄟ
	2860	A8E0	ㄠ	ㄡ	ㄢ	ㄣ	ㄤ	ㄥ	ㄦ	ㄧ	ㄨ	ㄩ						
	2870	A8F0																

区位	GB	EUC	+0	+1	+2	+3	+4	+5	+6	+7	+8	+9	+A	+B	+C	+D	+E	+F
09	2920	A9A0					─	━	│	┃	┄	┅	┆	┇	┈	┉	┊	┋
	2930	A9B0	┌	┍	┎	┏	┐	┑	┒	┓	└	┕	┖	┗	┘	┙	┚	┛
	2940	A9C0	├	┝	┞	┟	┠	┡	┢	┣	┤	┥	┦	┧	┨	┩	┪	┫
	2950	A9D0	┬	┭	┮	┯	┰	┱	┲	┳	┴	┵	┶	┷	┸	┹	┺	┻
	2960	A9E0	┼	┽	┾	┿	╀	╁	╂	╃	╄	╅	╆	╇	╈	╉	╊	╋
	2970	A9F0																

第 1 級漢字（16 区～55 区）

区位	GB	EUC	+0	+1	+2	+3	+4	+5	+6	+7	+8	+9	+A	+B	+C	+D	+E	+F
16	3020	B0A0		啊	阿	埃	挨	哎	唉	哀	皑	癌	蔼	矮	艾	碍	爱	隘
	3030	B0B0	鞍	氨	安	俺	按	暗	岸	胺	案	肮	昂	盎	凹	敖	熬	翱
	3040	B0C0	袄	傲	奥	懊	澳	芭	捌	扒	叭	吧	笆	八	疤	巴	拔	跋
	3050	B0D0	靶	把	耙	坝	霸	罢	爸	白	柏	百	摆	佰	败	拜	稗	斑
	3060	B0E0	班	搬	扳	般	颁	板	版	扮	拌	伴	瓣	半	办	绊	邦	帮
	3070	B0F0	梆	榜	膀	绑	棒	磅	蚌	镑	傍	谤	苞	胞	包	褒	剥	

区位	GB	EUC	+0	+1	+2	+3	+4	+5	+6	+7	+8	+9	+A	+B	+C	+D	+E	+F
17	3120	B1A0		薄	雹	保	堡	饱	宝	抱	报	暴	豹	鲍	爆	杯	碑	悲
	3130	B1B0	卑	北	辈	背	贝	钡	倍	狈	备	惫	焙	被	奔	苯	本	笨
	3140	B1C0	崩	绷	甭	泵	蹦	迸	逼	鼻	比	鄙	笔	彼	碧	蓖	蔽	毕
	3150	B1D0	毙	毖	币	庇	痹	闭	敝	弊	必	辟	壁	臂	避	陛	鞭	边
	3160	B1E0	编	贬	扁	便	变	卞	辨	辩	辫	遍	标	彪	膘	表	鳖	憋
	3170	B1F0	别	瘪	彬	斌	濒	滨	宾	摈	兵	冰	柄	丙	秉	饼	炳	

·

·（途中略）

·

区位	GB	EUC	+0	+1	+2	+3	+4	+5	+6	+7	+8	+9	+A	+B	+C	+D	+E	+F
54	5620	D6A0		帧	症	郑	证	芝	枝	支	吱	蜘	知	肢	脂	汁	之	织
	5630	D6B0	职	直	植	殖	执	值	侄	址	指	止	趾	只	旨	纸	志	挚
	5640	D6C0	掷	至	致	置	帜	峙	制	智	秩	稚	质	炙	痔	滞	治	窒
	5650	D6D0	中	盅	忠	钟	衷	终	种	肿	重	仲	众	舟	周	州	洲	诌
	5660	D6E0	粥	轴	肘	帚	咒	皱	宙	昼	骤	珠	株	蛛	朱	猪	诸	诛
	5670	D6F0	逐	竹	烛	煮	拄	瞩	嘱	主	著	柱	助	蛀	贮	铸	筑	

区位	GB	EUC	+0	+1	+2	+3	+4	+5	+6	+7	+8	+9	+A	+B	+C	+D	+E	+F
55	5720	D7A0		住	注	祝	驻	抓	爪	拽	专	砖	转	撰	赚	篆	桩	庄
	5730	D7B0	装	妆	撞	壮	状	椎	锥	追	赘	坠	缀	谆	准	捉	拙	卓
	5740	D7C0	桌	琢	茁	酌	啄	着	灼	浊	兹	咨	资	姿	滋	淄	孜	紫
	5750	D7D0	仔	籽	滓	子	自	渍	字	鬃	棕	踪	宗	综	总	纵	邹	走
	5760	D7E0	奏	揍	租	足	卒	族	祖	诅	阻	组	钻	纂	嘴	醉	最	罪
	5770	D7F0	尊	遵	昨	左	佐	柞	做	作	坐	座						

第 2 級漢字（56 区～87 区）

区位	GB	EUC	+0	+1	+2	+3	+4	+5	+6	+7	+8	+9	+A	+B	+C	+D	+E	+F

	5820	D8A0		亍	丌	兀	丐	廿	卅	丕	亘	丞	鬲	孬	噩	丨	禺	丿
	5830	D8B0	匕	乇	夭	爻	卮	氐	囟	胤	馗	毓	睾	鼗	丶	亟	鼐	乜
56	5840	D8C0	乩	亓	芈	孛	嗇	嘏	仄	厙	厝	厣	厥	廝	靥	贗	匚	叵
	5850	D8D0	匭	匱	匾	賾	卦	卣	刂	刈	刎	剄	刳	劌	剴	剌	剞	剡
	5860	D8E0	剜	蒯	剽	劂	劁	劐	劓	冂	罔	亻	仃	仉	仂	仨	仡	仫
	5870	D8F0	仞	伛	仳	伢	佤	仵	伥	伧	伉	伫	佞	佧	攸	佚	佝	

区位	GB	EUC	+0	+1	+2	+3	+4	+5	+6	+7	+8	+9	+A	+B	+C	+D	+E	+F
	5920	D9A0		佟	佗	伲	伽	佶	佴	侑	侉	侃	侏	佾	佻	侪	佼	侬
	5930	D9B0	侔	俦	俨	俪	俅	俚	俣	俜	俑	俟	俸	倩	偌	俳	倬	倏
57	5940	D9C0	倮	倭	俾	倜	倌	倥	倨	偾	偃	偕	偈	偎	偬	偻	傥	傧
	5950	D9D0	傩	傺	僖	儆	僭	僬	僦	僮	儇	儋	仝	氽	佘	佥	俎	龠
	5960	D9E0	汆	籴	兮	巽	黉	馘	冁	夔	勹	匍	訇	匐	凫	夙	兕	亠
	5970	D9F0	兖	亳	衮	袤	亵	脔	裒	禀	嬴	蠃	羸	冫	冱	冽	冼	

・

・（途中略）

・

区位	GB	EUC	+0	+1	+2	+3	+4	+5	+6	+7	+8	+9	+A	+B	+C	+D	+E	+F
	7720	F7A0		鰲	鰭	鰨	鰥	鰩	鰳	鰾	鱈	鰻	鰵	鱅	鱖	鱔	鱒	鱧
	7730	F7B0	靼	鞅	鞑	鞒	鞔	鞯	鞫	鞣	鞲	鞴	骱	骰	骷	鶻	骶	骺
87	7740	F7C0	骼	髁	髀	髏	髂	髖	髕	髑	魅	魃	魘	魎	魈	魍	魑	飨
	7750	F7D0	餍	餮	饕	饔	髟	髡	髦	髯	髫	髻	髭	髹	鬈	鬏	鬢	鬟
	7760	F7E0	鬣	麽	麾	縻	麂	麇	麈	麋	麒	鏖	麝	麟	黛	黜	黝	黠
	7770	F7F0	黟	黢	黷	黧	黥	黲	黯	鼢	鼬	鼯	鼹	鼷	鼽	鼾	齄	

２－２－３台湾

現在、台湾でよく使われている言語は中国語です。文字は繁体字が用いられています。香港も同様です。

CCCII：

1980 年に CCCII（シーシーシーアイアイ；Chinese Character Code of Information Interchange = 中文資訊交換碼）が制定されました。当初は 4,808 文字を収録し、その後、改定を重ね、75,000 字以上を収録しています。日本・中国（簡体字)・韓国の文字も収録しており、これを基にアメリカでは東アジア地域の言語を扱う文字コードとして ANSI/NISO が 1989 年に EACC（イーエイシーシー：East Asia Coded Character、Z39.64-1989）を規格化されました。

CNS 11643：

1986 年には經濟部中央標準局が CNS 11643（シーエヌエスイチイチロクヨンサン；Chinese National Standards 11643 = 通用漢字標準交換碼）という国家規格を制定しました。CNS 11643 を採用した EUC 文字コードを、EUC-TW（イーユーシーティーダブリュー = 繁体字中国語 EUC）と呼でいます。CNS 11643-1992 には 48,711 字が収録されています。

２－２－４韓国

韓国における言語は韓国語です。文字はハングルです。

表２－12：ハングル子音字（ザウン、縦の軸）と、
母音字（モウン、横の軸）

	ㅏ	ㅑ	ㅓ	ㅕ	ㅗ	ㅛ	ㅜ	ㅠ	ㅡ	ㅣ
ㄱ	가 カ	갸 キャ	거 コ	겨 キョ	고 コ	교 キョ	구 ク	규 キュ	그 ク	기 キ
ㄴ	나 ナ	냐 ニャ	너 ノ	녀 ニョ	노 ノ	뇨 ニョ	누 ヌ	뉴 ニュ	느 ヌ	니 ニ
ㄷ	다 タ	댜 ティヤ	더 ト	뎌 ティヨ	도 ト	됴 ティヨ	두 トゥ	듀 ティユ	드 トゥ	디 ティ
ㄹ	라 ラ	랴 リャ	러 ロ	려 リョ	로 ロ	료 リョ	루 ル	류 リュ	르 ル	리 リ
ㅁ	마 マ	먀 ミャ	머 モ	며 ミョ	모 モ	묘 ミョ	무 ム	뮤 ミュ	므 ム	미 ミ
ㅂ	바 パ	뱌 ピャ	버 ポ	벼 ピョ	보 ポ	뵤 ピョ	부 プ	뷰 ピュ	브 プ	비 ピ
ㅅ	사 サ	샤 シャ	서 ソ	셔 ショ	소 ソ	쇼 ショ	수 ス	슈 シュ	스 ス	시 シ
ㅇ	아 ア	야 ヤ	어 オ	여 ヨ	오 オ	요 ヨ	우 ウ	유 ユ	으 ウ	이 イ
ㅈ	자 チャ	쟈 チャ	저 チョ	져 チョ	조 チョ	죠 チョ	주 チュ	쥬 チュ	즈 チュ	지 チ
ㅊ	차 チャ	챠 チャ	처 チョ	쳐 チョ	초 チョ	쵸 チョ	추 チュ	츄 チュ	츠 チュ	치 チ
ㅋ	카 カ	캬 キャ	커 コ	켜 キョ	코 コ	쿄 キョ	쿠 ク	큐 キュ	크 ク	키 キ
ㅌ	타 タ	탸 ティヤ	터 ト	텨 ティヨ	토 ト	툐 ティヨ	투 トゥ	튜 ティユ	트 トゥ	티 ティ
ㅍ	파 パ	퍄 ピャ	퍼 ポ	펴 ピョ	포 ポ	표 ピョ	푸 プ	퓨 ピュ	프 プ	피 ピ
ㅎ	하 ハ	햐 ヒャ	허 ホ	혀 ヒョ	호 ホ	효 ヒョ	후 フ	휴 ヒュ	흐 フ	히 ヒ

KS コード：

ハングルの文字コードには、個別に符号化した字母を組み合わせて 1 文字（音節）を表現する方法（例：ㄱ+ ㅏ=가）と、字母を組み合わせて構成した各音節文字を符号化する方法があります。 韓国で最も普及している文字コードは KS X 1001 という国家規格で、通称 KS コードと呼ばれています。韓国の漢字は原則的に 1 文字 1 音ですが、複数の発音を持つものもあり、それらを読みの数だけ重複して符号化しています。KS X 1001 を採用した EUC 文字コードを、EUC-KR（ 韓国語 EUC）と呼んでいます。

KS X 1001-2002 には 8,227 字が収録されています。

表 2－13：KS X 1001-2002

非ハングル・非漢字など（1 区～12 区）

区点	KS	EUC	+0	+1	+2	+3	+4	+5	+6	+7	+8	+9	+A	+B	+C	+D	+E	+F
01区	2120	A1A0			、	。	·	‥	…	¨	〃		—	∥	＼	~	‘	’
	2130	A1B0	“	”	〔	〕	〈	〉	《	》	「	」	『	』	【	】	±	×
	2140	A1C0	÷	≠	≤	≥	∞	∴	°	′	″	℃	Å	¢	£	¥	♂	♀
	2150	A1D0	∠	⊥	⌒	∂	∇	≡	≒	§	※	☆	★	○	●	◎	◇	◆
	2160	A1E0	□	■	△	▲	▽	▼	→	←	↑	↓	↔	=	≪	≫	√	∽
	2170	A1F0	∝	∵	∫	∬	∈	∋	⊆	⊇	⊂	⊃	∪	∩	∧	∨	￢	

・

・（途中略）

・

ハングル音節（16 区～40 区）

区点	KS	EUC	+0	+1	+2	+3	+4	+5	+6	+7	+8	+9	+A	+B	+C	+D	+E	+F
16区	3020	B0A0		가	각	간	갇	갈	갉	갊	감	갑	값	갓	갔	강	갖	갗
	3030	B0B0	같	갚	갛	개	객	갠	갤	갬	갭	갯	갰	갱	갸	갹	갼	걀
	3040	B0C0	걋	걍	걔	걘	걜	거	걱	건	걷	걸	걺	검	겁	것	겄	겅
	3050	B0D0	겆	겉	겊	겋	게	겐	겔	겜	겝	겟	겠	겡	겨	격	겪	견
	3060	B0E0	겯	결	겸	겹	겻	겼	경	곁	계	곈	곌	곕	곗	고	곡	곤
	3070	B0F0	곧	골	곪	곬	곯	곰	곱	곳	공	곶	과	곽	관	괄	괆	

区点	KS	EUC	+0	+1	+2	+3	+4	+5	+6	+7	+8	+9	+A	+B	+C	+D	+E	+F
	3120	B1A0		괌	괍	괏	광	괘	괜	괠	괩	괬	괭	괴	괵	괸	괼	굄
	3130	B1B0	굅	굇	굉	교	굔	굘	굡	굣	구	국	군	굳	굴	굵	굶	굻
17	3140	B1C0	굼	굽	굿	궁	궂	궈	궉	권	궐	궜	궝	궤	궷	귀	귁	귄
区	3150	B1D0	궐	궘	궙	궛	규	균	귤	그	극	근	귿	글	긁	금	급	긋
	3160	B1E0	긍	긔	기	긱	긴	긷	길	긺	김	깁	깃	깅	깆	깊	까	깍
	3170	B1F0	깎	깐	깔	깖	깜	깝	깟	깠	깡	깥	깨	깩	깬	깰	깸	

・

・（途中略）

・

漢字（42 区～93 区）

区点	KS	EUC	+0	+1	+2	+3	+4	+5	+6	+7	+8	+9	+A	+B	+C	+D	+E	+F
	4A20	CAA0		伽	佳	假	價	加	可	呵	哥	嘉	嫁	家	暇	架	枷	柯
	4A30	CAB0	歌	珂	痂	稼	苛	茄	街	袈	訶	賈	跏	軻	迦	駕	刻	却
42	4A40	CAC0	各	恪	慤	殼	珏	脚	覺	角	閣	侃	刊	墾	奸	姦	干	幹
区	4A50	CAD0	懇	揀	杆	柬	桿	澗	癎	看	磵	稈	竿	簡	肝	艮	艱	諫
	4A60	CAE0	間	乫	喝	曷	渴	碣	竭	葛	褐	蝎	鞨	勘	坎	堪	嵌	感
	4A70	CAF0	憾	戡	敢	柑	橄	減	甘	疳	監	瞰	紺	邯	鑑	鑒	龕	

・

・（途中略）

・

区点	KS	EUC	+0	+1	+2	+3	+4	+5	+6	+7	+8	+9	+A	+B	+C	+D	+E	+F
	7D20	FDA0		爻	肴	酵	驍	侯	候	厚	后	吼	喉	嗅	帿	後	朽	煦
	7D30	FDB0	珝	逅	勛	勳	塤	壎	焄	熏	燻	薰	訓	暈	薨	喧	暄	煊
93	7D40	FDC0	萱	卉	喙	毁	彙	徽	揮	暉	煇	諱	輝	麾	休	携	烋	畦
区	7D50	FDD0	虧	恤	譎	鷸	兇	凶	匈	洶	胸	黑	昕	欣	炘	痕	吃	屹
	7D60	FDE0	紇	訖	欠	欽	歆	吸	恰	洽	翕	興	僖	凞	喜	噫	囍	姬
	7D70	FDF0	嬉	希	憙	憘	戲	晞	曦	熙	熹	熺	犧	禧	稀	羲	詰	

２－３　問題点

Unicode とは、一言でいうと全世界の文字を 16 ビットで表現することを目指す規格であります。世界のあらゆる文字の表現を目指し、1993 年に ISO(国際標準化機構)と IEC(国際電気標準会議)との合同技術委員会で標準化されましたが、1 文字を 16 ビットで表現したところ最大 65536 文字しか収録できず、漢字の統合を試みる結果となりました。

日本・中国・韓国・台湾など、Unicode の 16 ビットでの制約では自国の言葉をきちんと表現できない国は多く存在するにもかかわらず ISO　10646 という国際標準化機構の規格となってしまいました。

CJK 統合漢字のベースとなった主な文字集合は、中国（C）の GB 2312、台湾（C）の CNS1,2,14 面、日本（J）の JIS X 0208 と JIS X 0212、韓国（K）の KS C 5601 になります。これらベースとなった文字集合の文字を足すと、20,902（Unicode1.0.1 時）にはとうてい収まりませんが、字形的に同一もしくは些細な字形差と認められる文字は、同じコードポイントに割り振ることで、20,902 という数にまとまりました。これを漢字統合（Han Unification）と呼んでいます。

Unicode は、漢字統合によって多くの文字を収録していますが、日本で一般的に使われている明朝体の字体と、中国大陸・台湾の字形が異なるにも関わらず、同じコードに「統合」されている場合もよくみられています。このことによって、本来区別したい字が区別できないという問題が発生しました。

私のお父さんは「李　春燮」という名前で「燮」の字は「変」の字と似ていることで「変」の字になりました。「燮」の字は「程よくする」の意味であります。それが「変わる、変化する」意味の「変」の字になって「春のような温かい性格を持った持ち主で苦労をしないで程よい人生を送る」の意味を持った「春燮」が「春が変わる」の意味を持った「春変」になりました。

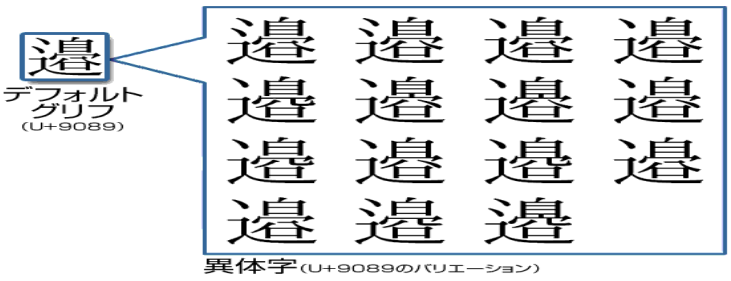
中国では名前が人の人生を左右するといわれており、わざわざお金をかけてまで名前を付けてくれる会社で名前を付けています。この頃は需要が高まったことでもあり、このような会社は色々なところでみられています。

私のお父さんは生まれたときに政府に紙ベースで登録した「李　春燮」からオンライン化したときに住民登録など正式な名前が「李　春変」になってしまいました。自

分の意志ではないことに最初お父さんも戸惑っていましたが、パスポートまでもこの名前になってしまった今はもうありのまま受け入れている。このように名前に因んだことはおそらく私のお父さんだけでないと考えられます。

名前のことは中国だけてなく、日本でも大きな波紋を広げていました。いわゆる「消えた５０００万件の年金記録」であります。給料から天引きされていて一定の年齢になれば受給されるはずの年金が受給できなくなってしまいました。その要因の一つに漢字カナ自動変換システムによる記録の誤りが挙げられています。渡辺の「辺」にしても以下のようにいろいろあります。

表２－14：「辺」の字色々



（出典：イースト株式会社）

このようにCJK統合漢字によって些細な字形差と認められる文字が、同じコードポイントに割り振られたことで区別できなくなり、またこの要因で名前が一致しないということで年金が受給できない方が多く確認できています。一所懸命働いて頂いたその一部を出したのに、老後に年金の受給ができないという方たちのショックは何にもたとえ難いです。その人たちの老後が懸念されます。

この様なことから、英語圏向きに作られたインターネット仕組みはどうせその地域に住んでいる人たちには不便なところがないので問題視されないのは当たり前ですが、問題が起こっている国々でもその問題に向けて力を合わせる必要があると考えられます。

3．解決への模索

3－1　CJK 統合漢字について

日本マイクロソフト（株）は去年 11 月 9 日、「Microsoft Office」で“Unicode IVD（UTS#37)”に対応した異体字を扱えるようにするアドイン「Unicode IVS Add-in for Microsoft Office」を公開しました。これによって、58000 の異体字、様々な異体字を含むデータの表示・印刷・編集などが可能となりました。

文字や言語はそれぞれの国が持っている特々のものです。この１０年くらいまでは文字の読み書きやパソコン上入力の練習をすることで言語を学ぶことができました。また、色々な言語を勉強することで益々世界中の多くの人とコミュニケーションをとることができ、いい商品を紹介しあったりしていたところ輸入、輸出が生まれたりして、またそれが世の中の経済の活性化にもつながりました。

今現在はインターネット技術が発展したことで、コンピュータ上利用しやすくなったし、スマートフォンも普及してきました。これらは最初文字コードを作った ISO のおかげであります。最初は多少厳しい面もありましたが、文字コードから国際化文字コードもできほとんどの国が自国の文字をパソコン上精確に入力することができるようになったことで、文字文化がいつまでも存続しうると考えられます。もちろん、コンピュータやスマートフォンの普及によって我々のコミュニケーションがとりやすくなったのも事実です。このごろはいつでもどこでも無料でビデオ通話ができるようになりましたし、わからない言葉をすぐ翻訳してくれるアプリケーションなどもでき、大変便利になりました。

3－2　世界の文字事情

中西印刷株式会社によれば、世界の現用文字は以下の 28 種類だそうです。

表３－１：世界現用文字

文字	使用地域	Unicode 場所	いつ編入したか
ラテン文字	ヨーロッパのほとんどの国をはじめアジアの一部、アフリカの大部分、南北アメリカとオセアニアの全部	0020-0024F	
ギリシャ文字	ギリシャ国内	0370-03FF	
ロシア文字	旧ソビエト連邦諸国　モンゴルなど	0400-04FF	
グルジア文字	グルジア共和国	10D0-10FF	
アルメニア文字	アルメニア	0530-058F	
ヘブライ文字	イスラエル	0590-05FF	
アラビア文字	中東　西アジア一帯　北部アフリカ	0600-06FF	
デーバナーガリー文字	北部インド	0970-097F	
グルムーキー文字	北部インドパンジャブ地方	0A00-0A7F	
グジャラート文字	西部インドグジャラート地方	0A80-0A8F	
オリヤー文字	東部インド　オリッサ地方	0B00-0B7F	
ベンガル文字	バングラデシュ　インドウエストベンガル地方	0980-09DF	
タミール文字	南インド　タミールナドゥ地方	0B80-0BFF	
テルグー文字	東インド　アンドラプラデシ地方	0C00-0C7F	
カンナダ文字	西インド　カルナタカ地方	0C80-0CFF	
マラヤラム文字	南インド　ケララ地方	0D00-0DFF	
シンハラ文字	スリランカ	0D80-0DFF	
ビルマ文字	ミャンマー		
クメール文字	カンボジア	1780-17FF	
タイ文字	タイ	0E00-0E7F	
ラオ文字	ラオス	0E80-0EFF	
漢字	中国、台湾、韓国、日本	4E00-9FFF	
チベット文字	中国西蔵自治区　青海省の一部	0F00-0FFF	
モンゴル文字	中国内蒙古自治区	1800-18AF	
朝鮮文字（ハングル）	朝鮮半島	3400-4DFF	
日本文字	日本	3040-30FF	

注音符号	現用文字　東アジア	3100-3120	
アムハラ文字（エチオピア文字）	エチオピア	1200-137F	

（出典：中西印刷株式会社）

中国には 55 の少数民族があり、その中には「満族」という民族があります。1000 万人を超える人口を持っていて、少数民族の中では 2 番目に多い民族です。17 世紀の半ばから 20 世紀の初期まで満族によって建てられた「清国」という国がありました。満族には自分たちの言語や文字を持っていて、「清国」が建てられたことによって、広く使われましたが、その後漢族によって「清国」が滅亡され、自分たちの文字や言語を隠さざるを得なかったです。このようなこともあり、20 世紀 80 年代ごろには少数の年配の者しか「満族」の言葉ができる人がいなくなりました。それから 30 年くらい経った今現在、「満族」という民族は書籍上残っていますが、その言語や文字は私たちの身の回りからすっかり消えていきました。

中国にインターネットが流行り始まったのは 2000 年ごろです。パソコンで自分たちの文字が見られる、若しくは文字が打てることによって、増々その文字に興味を持つようになります。その様になれば、いつまでも生き続けられます。

もし、中国でのインターネット時代が 20 年早くやってきていれば、「満族」の文字や言葉もその波に乗って今でも生き続けていたはずです。

パキスタンでも同じようなことが起こっています。

パキスタンの国語はウルドゥー語ですが、全人口の約 7%程度を占めるにすぎず、ウルドゥー語ができない国民が多くいるそうです。義務教育制度が確立しておらず、識字率は 55%程度です。2012 年のインターネットの普及率は 2001 年の 1.32%から 7.5 倍ほど上がって 9.96%になったものの、まだまだ国民の 100 人に 9 人しか利用していません。1 人当たり GNI は約 1,050 ドル（2010/11 年度）、総人口の約 4 人に 1 人が貧困と言われる開発途上国であり、外国援助・投資、国外からの郷里送金に大きく依存した経済構造となっているそうです。

わかりやすく言うと義務教育制度が確立していないことから識字率が低くなり、インターネットのことなど興味を持てなくなり、世界中がどう動いているかが把握でき

ず国際事情についていけなくなった事が経済など伸びない理由だと考えられます。

この事からは「インターネットによる情報化」と「経済」の関係性がよくわかります。

表 3－2：インターネット普及率が低い国々

189 位	シエラレオネ	1.30
190 位	ブルンジ	1.22
191 位	ミャンマー	1.07
192 位	東ティモール	0.91
193 位	エリトリア	0.80

（出典 ITU － ICT Statistics）

表 3－2 は 2012 年世界 193 ヶ国を対象とした国別インターネット普及率ランキングから 189 位～193 位を切り取ったものです。左側の列が順位、真ん中の列が国名、右側の列がインターネットの普及率です。この国々のその年の名目 GDP を調べると、シエラレオネ 155 位、ブルンジ 159 位、ミャンマー74 位、東ティモール 146 位、エリトリア 157 位とミャンマー以外は基本的に低いことがわかりました。

これに対して同じ年のアメリカのインターネットの普及率は 81.03%で 1 位で日本は 79.05%で 2 位です。

4．まとめと今後の検討内容

4－1　まとめ

本研究では、まず文字コードの今までの歴史を調べました。調べていくうちに文字コードは欧米人向きに作られていてその地域に住む人たちには問題はないが、日本、中国、台湾、韓国など漢字や数万字の文字を使う国や地域の人たちには問題が起こっていることがわかりました。そして、2012 年の 11 月に「Unicode IVS　Add-in for Microsoft Office」が開発され、漢字に関わるほとんどの問題が解決されたこともわかりました。今年で 8 回目を開いた「日中韓 30 人会」などもあり、文化の面ではお互いに力を合わせ、解決しようとする動きが見えました。

あとは中国の少数民族の中の一つ「満族」について調べました。今現在は 1000 万人の人口を持つ民族で、17 世紀の半ばから 20 世紀の初期まで満族によって建てられた「清国」という国が建てられたくらい大きく歴史に残る民族で自分たちの言葉や文字をもっていた民族ですが、今はほとんど失われています。「インターネット時代」に乗り遅れた事が原因だと考えられます。

その次はパキスタン国内の事情を調べました。パキスタンは義務教育制度が確立していないことから認知率が低くなり、インターネットのことなど興味を持てなくなり、世界中がどう動いているかが把握できず国際事情についていけなくなった事が経済など伸びない理由だと考えました。2012 年世界 193 ヶ国を対象とした「国別インターネット普及率ランキング」を調べたら、やはり普及率が低い国が GDP も低いということがわかりました。以上の様に「インターネット時代」に乗り遅れて痛い目に合っている民族や「インターネットの情報化」と経済の関係性がよくわかる国はまだまだ多くあると考えられます。

本研究はまだまだ未熟なものですが、国々が「インターネットの情報化」に目をむきそれに関心を持って頂く事も目的の一つとしています。世界中の国々の隅々まで「インターネットの情報化」が進み、同じ土俵で平等に使われるような日が一日でも早く来てくれればありがたいです。

4－2　中国の少数民族たちの現用文字とその現状

中国には 55 の少数民族が住んでおり、各民族が自分たちの民族語を所有するとは限らないし、民族語があるとしても文字を持っているとは限りません。調べによれば、55 の少数民族の中 22 種類の文字が使われているそうです。この２２種類の文字は全種 Unicode に載っているのか、もし載っていないのがあればなぜ載っていないのかを調べたいです。

4－3　文字コード研究者による最新の文献

2012 年の「Unicode IVS　Add-in for Microsoft Office」の公開以降の文字コード研究者による最新文献を調べ、その動きを把握することにより、ポスト Unicode ともいえる次世代国際化文字コードのあるべき姿を模索することに繋げたいです。

参考文献

[1] 太田昌孝 『いま日本語が危ない ： 文字コードの誤った国際化』 丸山学芸図書、1997 年、 15 ページ

[2] 小池和夫［ほか］『 漢字問題と文字コード 』 太田出版、1999 年、 22 ページ

[3] 安岡孝一，安岡素子 『文字コードの世界』 東京電機大学出版局、1999 年、36 ページ

[4] 小林龍生［ほか］ 『インターネット時代の文字コード 』共立出版、 2002 年、62 ページ

[5] 深沢千尋 『文字コード「超」研究』 ラトルズ、 2003 年、 54 ページ

[6] 萬宮健策 「パキスタンの諸言語資源をめぐる現状と課題」『アジア情報室通報』第 4 巻第 4 号 2006 年 12 月、142～156 ページ

[7] マイナビニュース
http://news.mynavi.jp/news/2012/11/12/004/index.html 2013 年 12 月 13 日

[8] Smart
http://rfs.jp/sb/perl/09/01-10.html 2013 年 12 月 13 日

[9] 韓国語あいうえお表
http://www.cinemart.co.jp/sejong/intro/win.html 2013 年 12 月 13 日

[10] 文部科学省 世界の母語人口
http://www.mext.go.jp/b_menu/shingi/chukyo/chukyo3/015/siryo/060

32708/003/001.htm 2013 年 12 月 13 日

[11] コンピュータ基礎講座
http://www.asahi-net.or.jp/~ax2s-kmtn/ref/ksx1001.html 2013 年 12 月 13 日

[12] 世界の文字
http://www.nacos.com/information/character/present.php 2012 年 12 月 13 日

[13] 日本年金機構
http://www.nenkin.go.jp/n/www/k-cam/index2.jsp 2013 年 12 月 13 日

[14] サーチナ
http://news.searchina.ne.jp/disp.cgi?y=2013&d=1121&f=national_1121_014.shtml 2013 年 12 月 13 日

[15] 世界経済のネタ帳
http://ecodb.net/ranking/icts_internet.html#PK 2013 年 12 月 13 日

謝辞

まず、本研究テーマの設定に当たり興味を持っている「東アジア事情」と今までのゼミで勉強したことをどう繋げていけばいいのかで悩んだ時に助けて頂いた朱　全安教授に大変感謝いたします。また、本研究に当たり最初から最後までご指導を頂いた渡辺　恭人准教授にも大変感謝いたします。感謝の気持ちでいっぱいです。本研究はまだまだ未熟なものですが、ここに至るまで研究のテーマや方向性がぶれることもありました。その都度、より良い方向へ導いて頂いたお陰で、ここまででも進めることができました。最後に私の卒業論文にかかわって頂いた全ての方にもう一度感謝を述べさせて頂きます。本当にありがとうございました。