

# 동아시아 문자 처리

2012-10-20

토요일(土曜日,星期六,Saturday)

성대현(成大鉉,成大铉,DaeHyun Sung)

# 문자란 무엇인가?

## What Is a Character?

“The smallest component of written language that has semantic value; refers to the abstract meaning and/or shape, rather than a specific shape.”

—The Unicode Consortium

# CJKV

- CJKV – Chinese-Japanese-Korean-Vietnamese
  - 중국어(chinese)
    - 번체자[繁體中文, Traditional Chinese], 간체자[简体中文, Simplified Chinese]
    - 한어병음(汉语拼音, Hanyu Pinyin), 주음부호(注音符號, ㄅ ㄆ ㄇ ㄉ, bofomopo)
  - 일본어(Japanese)
    - 히라가나(ひらがな, Hiragana) 카타카나(カタカナ, Katakana)
  - 한국어(Korean)
    - 한글(Hangul)
  - 베트남어(Vietnamese)
    - 쯔놈, Chữ Nôm (字喃/𐄓喃/𐄓喃)

대만(臺灣, 中華民國/中华民国,  
Taiwan, Republic of China)

중화인민공화국  
中华人民共和国/中華人民共和國  
People's Republic of China)

대한민국(大韓民國/大韩民国,  
Republic of Korea)  
조선민주주의인민공화국  
(朝鮮民主主義人民共和國  
朝鮮民主主义人民共和国,  
Democracy People's Republic of Korea)

베트남(越南, Socialist Republic of Vietnam  
Cộng hòa xã hội chủ nghĩa Việt Nam,  
共和社會主義越南)

일본(日本国/日本國, Japan)

漢字文化圈

汉字文化圈

한자문화권  
漢字文化圈

漢字文化圈

漢字文化圈

# 한자

- Chinese Character
- 漢字/汉字
- ㄏㄢˋ ㄓˋ /hànzì
- 한자
- かんじ
- hán tự

漢字  
汉字  
漢字

# 중국어(中國語,中国語,漢語,汉语) Chinese

- **한어병음[汉语拼音] 표기법**

- 1955년~1957년 중국 문자 개혁 위원회(中国文字改革委员会)는 한어병음방안(汉语拼音方案)을 채택
- 1958년 2월 전국인민대표대회(全国人民代表大会)에서 이 안을 비준, 시행.
- 1982년 국제표준ISO 7098이 되었다.

- **웨이드-자일스식 표기법**

- 19세기 웨이드와 자일스가 중국어의 라틴어표기를 고안함.

- **주음부호[注音符號] 표기법**

- 1913년, 중국독음통일회에 의해 제정
- 1918년, 베이징(北京) 북양정부가 공표. 난징(南京) 국민정부에서도 공인되었지만, 1930년에 '주음부호(注音符號)'로 개칭되고 한자의 발음기호로 축소. 현재 대만에서 사용

# 중국어(中國語,中国語,漢語,汉语) Chinese

Table 2-2. Chinese transliteration—consonants

Zhuyin/bopomofo	Pinyin	Wade-Giles
ㄅ	B	P
ㄆ	P	P'
ㄇ	M	M
ㄈ	F	F
ㄉ	D	T
ㄊ	T	T'
ㄋ	N	N
ㄌ	L	L
ㄍ	G	K
ㄎ	K	K'
ㄏ	H	H
ㄐ	J	CH <sup>a</sup>
ㄑ	Q	CH <sup>a</sup>
ㄒ	X	HS <sup>a</sup>
ㄗ	ZH	CH
ㄘ	CH	CH'
ㄙ	SH	SH
ㄖ	R	J
ㄗ	Z	TS
ㄘ	C	TS'
ㄙ	S	S

a. Only before *i* or *ü*

# 중국어(中國語,中国語,漢語,汉语) Chinese

Table 2-3. Chinese transliteration—vowels

	ㅣ ㅣ	ㄨ u	ㄩ ü
ㅑ A	ㅣ ㅑ IA	ㅓ ㅑ UA	
ㅓ O		ㅓ ㅓ UO	
ㅕ E	ㅣ ㅕ IE		ㅕ ㅕ UE
ㅗ AI		ㅓ ㅗ UAI	
ㅜ EI		ㅓ ㅜ UEI	
ㅛ AO	ㅣ ㅛ IAO		
ㅟ OU	ㅣ ㅟ IOU		
ㅓ AN	ㅣ ㅓ IAN	ㅓ ㅓ UAN	ㅕ ㅓ ÜAN
ㅜ EN	ㅣ ㅜ IN	ㅓ ㅜ UEN	ㅕ ㅜ ÜN
ㅕ ANG	ㅣ ㅕ IANG	ㅓ ㅕ UANG	
ㅛ ENG	ㅣ ㅛ ING	ㅓ ㅛ UENG or ONG	ㅕ ㅛ IONG



# 중국어(中國語,中国語,漢語,汉语) Chinese

- Chinese Tone(성조,聲調)

Tone	Number	Example	Meaning
경성(輕聲 / 轻声)	None	嗎 / 吗 ma	"question particle"—neutral
1성(陰平 / 阴平)	1	媽 / 妈 mā	"mother"—high level
2성(陽平 / 阳平)	2	麻 má	"linen" or "numb"—high rising
3성(上聲 / 上声)	3	馬 / 马 mǎ	"horse"—low falling-rising
4성(去聲 / 去声)	4	罵 / 骂 mà	"scold"—high falling

# 중국어(中國語,中国語,漢語,汉语) Chinese

- 번체자(Traditional Chinese, 繁體中文)
- 간체자(Simplified Chinese, 简体中文)

# 일본어(日本語, 日本語) Japanese

- **The Hepburn System(ヘボン式)**
  - 미국 선교사 James Curtis Hepburn이 1887년도에 고안한 로마자 표기 방법
- **The kunrei System(訓令式)**
  - 1937년 일본정부가 발표한 로마자 표기 방법
- **The Nippon System(日本式)**
  - 田中館愛橘(tanakadate aikitsu)가 1881년도에 고안한 로마자 표기 방법. 훈령식과 비슷함.
- **The Word Processor System(ワープロ式)**
  - 몇 십년전 일본의 워드프로세서 전용 기계에서 사용했던 일본어 입력 방법

# 일본어(日本語, 日本語)

# Japanese

- 히라가나(hiragana, ひらがな)

*Table 2-18. The hiragana syllabary*

[illegible]

# 일본어(日本語, 日本語)

# Japanese

- 카타카나(カタカナ, Katakana)

*Table 2-20. The katakana syllabary*

[illegible]

# 한국어(韓國語, 韓国語, 韩国語)

## Korean

- 초성(初聲, Initials)+중성(中聲, Vowels)+종성(終聲, Finals)
- South Korea(남한)
  - Consonant letters:
    - ㄱ (ㄲ) ㄴ ㄷ (ㄸ) ㄹ ㅁ ㅂ (ㅃ) ㅅ (ㅆ) ㅇ ㅈ (ㅉ) ㅊ ㅋ ㅌ ㅍ ㅎ
  - Vowel letters:
    - ㅏ ㅑ ㅓ ㅕ ㅗ ㅛ ㅜ ㅠ ㅡ ㅟ ㅢ ㅛ ㅝ ㅞ ㅟ ㅠ ㅡ ㅣ ㅤ
- North Korea(북한)
  - Consonant letters:
    - ㄱ ㄴ ㄷ ㄹ ㅁ ㅂ ㅅ -ㅇ (= final) ㅈ ㅊ ㅋ ㅌ ㅍ ㅎ ㄲ ㄸ ㅃ ㅆ ㅉ ㅇ- (= initial)
  - Vowel letters:
    - ㅏ ㅑ ㅓ ㅕ ㅗ ㅛ ㅜ ㅠ ㅡ ㅣ ㅞ ㅟ ㅠ ㅡ ㅣ ㅤ ㅛ ㅝ ㅞ ㅟ ㅠ ㅡ ㅣ ㅤ
- [http://en.wikipedia.org/wiki/Hangul\\_consonant\\_and\\_vowel\\_tables](http://en.wikipedia.org/wiki/Hangul_consonant_and_vowel_tables)

# 한국어(韓國語, 韓国語, 韩国語)

## Korean

- 국어의 로마자 표기법(The Re-vised Romanization of Korean, abbreviation RRK)
  - 2000년 7월 7일 공표된 로마자 표기법
- 매쿤-라이샤워 표기법(Ministry of Education (문교부) derived from and sometimes referred to as McCune-Reischauer)
  - 1984년 1월 13일 공표된 로마자 표기법
- 한글학회(Korean Language Society)
  - 1996년도에 발표된 로마자 표기법

# 한글 소리마디의 자모 결합하는 6가지 방법

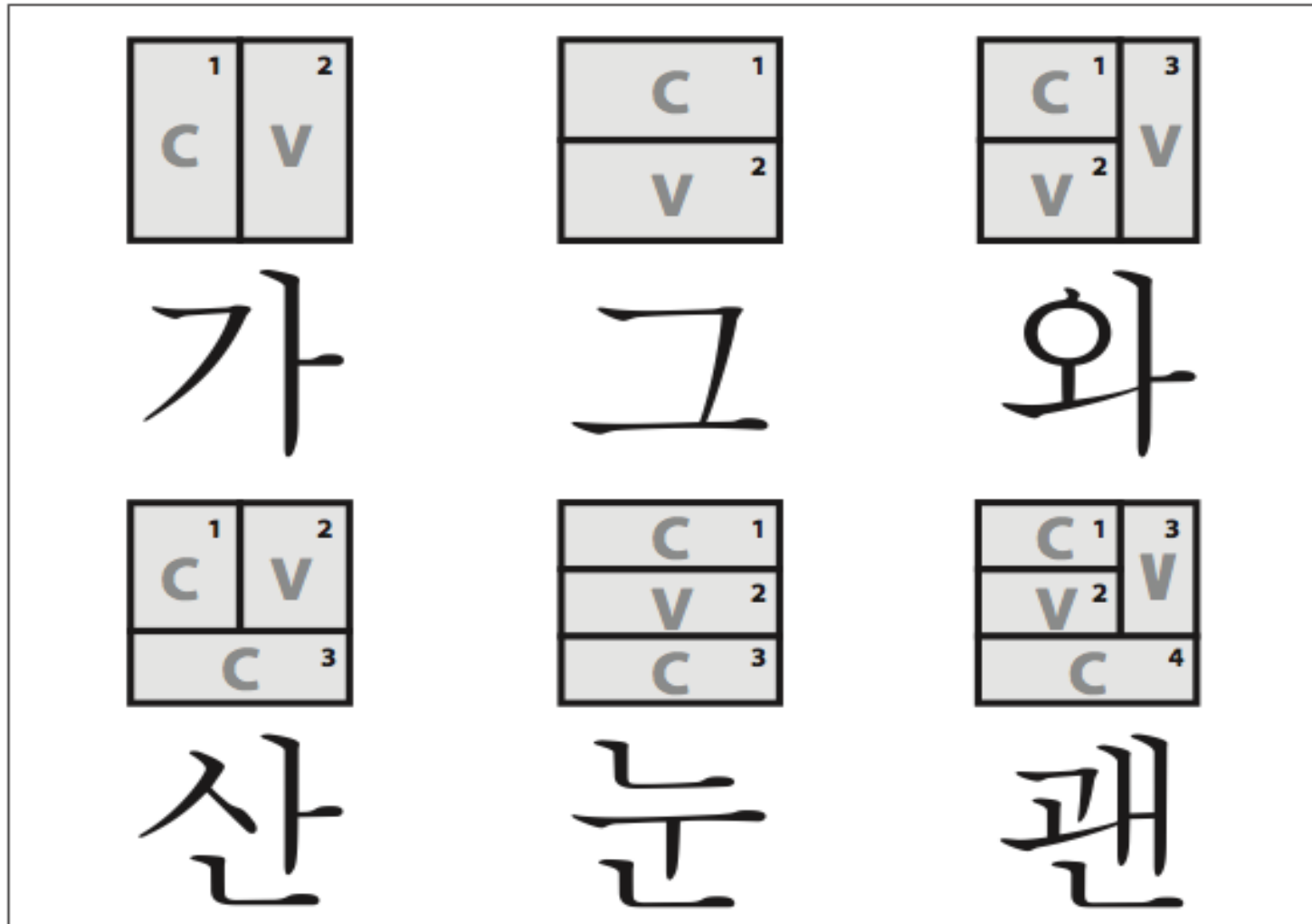


Figure 2-1. Six ways to compose jamo into modern hangul syllables



Table 2-9. Korean transliteration—consonants

Jamo	RRK <sup>a</sup>	MOE	KLS	ISO (DPRK)	Final	ISO (ROK)	Final
ㄱ	G/K <sup>b</sup> —G	K/G	G	K	K	G	G
ㄴ	N	N	N	N	N	N	N
ㄷ	D/T <sup>c</sup> —D	T/D	D	T	T	D	D
ㄹ	R/L <sup>d</sup> —L	R/L	L	R	L	R	L
ㅁ	M	M	M	M	M	M	M
ㅂ	B/P <sup>e</sup> —B	P/B	B	P	P	B	B
ㅅ	S	S/SH	S	S	S	S	S
ㅇ	None/NG	None/NG	None/NG	None	NG	None	NG
ㅈ	J	CH/J	J	C	C	J	J
ㅊ	CH	CH'	CH	CH	CH	C	C
ㅋ	K	K'	K	KH	KH	K	K
ㅌ	T	T'	T	TH	TH	T	T
ㅍ	P	P'	P	PH	PH	P	P
ㅎ	H	H	H	H	H	H	H
ㄲ	KK	KK	GG	KK	KK	GG	GG
ㄸ	TT	TT	DD	TT	n/a	DD	n/a
ㅃ	PP	PP	BB	PP	n/a	BB	n/a
ㅆ	SS	SS	SS	SS	SS	SS	SS
ㅈㅈ	JJ	TCH	JJ	CC	n/a	JJ	n/a

# 한국어(韓國語,韓国語,韩国语)

## Korean

*Table 2-10. ISO/TR 11941:1996 compound jamo transliteration*

Jamo	DPRK	ROK
ㄴ	KS	GS
ㄴㅈ	NJ	NJ
ㄴㅎ	NH	NH
ㄹㄱ	LK	LG
ㄹㅁ	LM	LM
ㄹㅂ	LP	LB
ㄹㅅ	LS	LS
ㄹㅈ	LTH	LT
ㄹㅎ	LPH	LP
ㄹㅌ	LH	LH
ㅂㅅ	PS	BS

Table 2-11. Korean transliteration—vowels

Jamo	RRK	MOE	KLS	ISO (DPRK and ROK)
ㅏ	A	A	A	A
ㅑ	YA	YA	YA	YA
ㅓ	EO	Ŏ	EO	EO
ㅕ	YEO	YŎ	YEO	YEO
ㅗ	O	O	O	O
ㅛ	YO	YO	YO	YO
ㅜ	U	U	U	U
ㅠ	YU	YU	YU	YU
ㅡ	EU	Ŭ	EU	EU
ㅣ	I	I	I	I
ㅞ	AE	AE	AE	AE
ㅟ	YAE	YAE	YAE	YAE
ㅠ	E	E	E	E
ㅡ	YE	YE	YE	YE
ㅢ	WA	WA	WA	WA
ㅣ	WAE	WAE	WAE	WAE
ㅤ	OE	OE	OE	OE
ㅥ	WO	WŎ	WEO	WEO
ㅦ	WE	WE	WE	WE
ㅧ	WI	WI	WI	WI
ㅨ	UI	ŬI	EUI	YI

# 베트남어(越南語, 越南语) Vietnamese

- 라틴문자(*Quốc ngữ*,꾸옥응우)
- 한자(*chữ Hán*: 𡵓漢, *Hán tự* 漢字 <汉字>)
- 베트남제(製) 한자(*chữ Nôm*)

mẹ tôi thường ăn chay ở chùa mọi chủ nhật  
媿 僻 常 餛 齋 於 𡵓 每 主 日

媿 (mẹ)	媽媽、母親。어머니, 엄마. 母、年配の女性、お母さん。
僻 (tôi)	我。나. わたし。
餛 (ăn)	吃。먹다. 食べる、飲食。
𡵓 (chùa)	寺、廟。불사, 사원. 仏寺、寺院。

# 입력방법

Locale(로케일)	Writing System(입력 시스템)
중국(China, People's Republic of China, 中华人民共和国)	Latin, hanzi(simplified chinese)[简体中文]
대만(Taiwan, 臺灣, Republic of China, 中華民國)	Latin, zhuyin[注音符號, ㄅ ㄆ ㄇ ㄉ] and hanzi(traditional chinese)[繁體中文]
일본(Japan, 日本国)	Latin, hiragana[ひらがな], Katakana[カタカナ], and kanji[漢字, かんじ]
한국(Korea, Republic of Korea, 대한민국, 大韓民國)	Latin, jamo[자모] hangul[한글] and hanja[漢字, 한자]
베트남(Vietnam, Socialist Republic of Vietnam)	Latin(Quốc Ngữ), chữ Nôm, chữ Hán

# CJKV Characters

- Latin characters
- Zhuyin 注音符號 ㄅ ㄆ ㄇ ㄉ, 주음부호, bopomofo
- Kana(hiragana ひらがな, Katakana カタカナ)
- Hangul 한글
- Chinese characters 漢字 汉字 한자
- Non-Chinese Chinese characters
  - Japanese kokuji[国字]
  - Korean 한국식한자, 국자
  - Vietnamese chữ Nôm

# CJK Variant glyphs

- [http://en.wikipedia.org/wiki/Han\\_Unification](http://en.wikipedia.org/wiki/Han_Unification)

*Table 3-99. CJKV character form differences*

Unicode code point	China	Taiwan	Japan	Korea
U+4E00	一	一	一	一
U+4E0E	与	与	与	与
U+5224	判	判	判	判
U+5668	器	器	器	器
U+5B57	字	字	字	字
U+6D77	海	海	海	海
U+9038	逸	逸	逸	逸
U+9AA8	骨	骨	骨	骨

# Variations

龜 龜 龜 龜 龜 龜 龜 龜 龜 龜 龜  
龜 龜 龜 龜 龜 龜 龜 龜 龜 龜 龜

*Table 1-10. Standard versus variant forms*

Standard Form	Variant forms	Additional variant forms
辺	邊 邊	邊邊邊邊邊邊邊邊邊邊邊邊邊 邊邊邊邊邊邊邊

*Table 1-11. Prototypical glyph changes over time—Japan*

Code point	1978	1983	1990	1997	2000	2004
36-52	辻	辻	辻	辻	辻	辻



# 각 나라마다 다른 한자

- 한국어 亮 U+F977

<http://www.unicode.org/cgi-bin/GetUnihanData.pl?codepoint=f977&useutf8=false>

- CJKV 통합문자 亮 U+4EAE

<http://www.unicode.org/cgi-bin/GetUnihanData.pl?codepoint=4EAE&useutf8=false>

# Japanese Kokuji(国字)

Table 2-40. Kokuji examples

Kokuji	Readings	Meanings
鰯 16-83 U+9C2F	<i>iwashi</i>	sardine
桑 23-09 U+7C82	<i>kume</i>	Used in personal names
込 25-94 U+8FBC	<i>komu</i>	(to) move inward
榊 26-71 U+698A	<i>sakaki</i>	A species of tree called <i>sakaki</i>
働 38-15 U+50CD	<i>hataraku, dō<sup>a</sup></i>	(to) work
峠 38-29 U+5CE0	<i>tōge</i>	mountain pass
畑 40-10 U+7551	<i>hata, hatake</i>	dry field
枠 47-40 U+67A0	<i>waku</i>	frame
凧 49-62 U+51E9	<i>kogarashi</i>	cold, wintry wind

a. Considered an On reading.

# 한국식 한자(Hanguksik hanja, 韓國式漢字)

*Table 2-41. Hanguksik hanja reading elements*

Hanguksik hanja element	Reading
乙	L
ㄱ	G
叱	D
○	NG

# 한국식 한자(Hanguksik hanja, 韓國式漢字)

Table 2-42. Hanguksik hanja examples

Hanguksik hanja	Reading	Meaning
𪛗 42-65 U+4E6B	갈 <i>gal</i>	Used in personal names
畚 51-44 U+7553	답 <i>dap</i>	paddy, wet field
𪛘 52-44 U+4E6D	돌 <i>dol</i>	Used in personal and place names
畹 56-37 U+551C	말 <i>mal</i>	Used in place names
鐡 64-54 U+9425	선 <i>seon</i>	Used in place names
箕 72-04 U+7B7D	오 <i>o</i>	Used in place names
岾 79-32 U+5CBE	점 <i>jeom</i>	mountain pass <sup>a</sup>

a. Compare this hanguksik hanja with the (Japanese) kokuji 峠 (*tōge*) in Table 2-40. I find it fascinating that Japan and Korea independently coined their own ideograph meaning “mountain pass.”

# 베트남 추놈(Vietnamese chữ Nôm)

Table 2-43. Chữ Nôm and chữ Hán examples

Chữ Nôm	Reading	Chữ Hán	Reading	Meaning
𠬞 21-47 U+20027	<i>ba</i>	三 42-06 U+4E09	<i>tam</i>	three
𠬞 29-55 U+219F2	<i>giữa</i>	中 42-21 U+4E2D	<i>trung</i>	center, middle
𠬞 34-02 U+21A38	<i>chữ</i>	字 50-30 U+5B57	<i>tự</i>	character
𠬞 35-77 U+24F93	<i>trăm</i>	百 64-02 U+767E	<i>bá</i>	hundred

# 문자집합(Character Set)

- Character Set(문자집합)
  - Character Code(문자코드)
  - Character Encoding(문자 인코딩)

# Character Code(문자코드)

- 문자를 숫자로 매핑(Mapping, 사상, 寫像)을 정의
- 문자를 표현하는 데이터값

예) A => 65(10진수)

B => 66(10진수)

C => 67(10진수)

- 문자 코드의 각각의 값을 코드 포인트(code point)

# Character Encoding(문자인코딩)

- 문자나 기호들의 집합을 컴퓨터에서 저장하거나 통신에 사용할 목적으로 부호화하는 방법을 가리킨다.
- 예) 아스키(ASCII, 1963년)
  - 65(10진수) => A
  - 66(10진수) => B
  - 67(10진수) => C



# ASCII

- American Standard Code for Information Interchange
  - Character code: 128( $2^7$ ) code points
  - Character encoding: 7bits each

ASCII Code Chart

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0	NUL	SOH	STX	ETX	EOT	ENQ	ACK	BEL	BS	HT	LF	VT	FF	CR	SO	SI
1	DLE	DC1	DC2	DC3	DC4	NAK	SYN	ETB	CAN	EM	SUB	ESC	FS	GS	RS	US
2		!	"	#	\$	%	&	'	(	)	*	+	,	-	.	/
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

# Extended ASCII

- 유럽에서 사용하는 문자들 입력은?
  - Ä Ç Ø Ö Ñ 등을 입력하기 위해 1bit 추가
  - Character code: 256( $2^8$ ) code points
  - Character encoding: 8bits each
  - ISO와 IEC의 공동 표준이다.
    - ISO/IEC 8859-1 (서유럽 언어)
    - ISO/IEC 8859-2 (중부유럽 언어)
    - (...)
    - ISO/IEC 8859-15 (ISO/IEC 8859-1 의 개정판)
    - ISO/IEC 8859-16 (남동유럽언어)
    - [http://en.wikipedia.org/wiki/ISO/IEC\\_8859](http://en.wikipedia.org/wiki/ISO/IEC_8859)

# 일본의 경우?

- 동아시아 국가(한국, 일본, 중국, 대만, 홍콩, 베트남)중에서 제일 경제가 발달되며 정보처리(Information Processing)를 먼저함
- JIS X0201
  - 7ビット及び8ビットの情報交換用符号化文字集合
  - 1969년도에 제정
  - 8bit 1byte 문자집합
  - 1byte에서 일본어를 넣자
  - 일본어 중에서 횡수가 많은 히라가나(hiragana, ひらがな)구현제외. 횡수가 간단한 카타카나(Katakana, カタカナ)만 입력가능
  - 글자들을 "반각(半角)문자 , Half-Width Katakana"로 부름
- 이후, 일본에서는 2byte로 정보처리를 구현. 한국, 대만, 중국, 홍콩에서 모두 일본을 따라 문자집합을 구성함.

# JIS X0201

第1面(JIS ラテン文字)

	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	NAK	%	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	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

第2面(半角カナ文字)

	0	1	2	3	4	5	6	7
0				ー	タ	ミ		
1			。	ア	チ	ム		
2			「	イ	ツ	メ		
3			」	ウ	テ	モ		
4			、	エ	ト	ヤ		
5			・	オ	ナ	ユ		
6			ヲ	カ	ニ	ヨ		
7			ア	キ	ヌ	ラ		
8			ィ	ク	ネ	リ		
9			ゥ	ケ	ノ	ル		
A			エ	コ	ハ	レ		
B			オ	サ	ヒ	ロ		
C			ャ	シ	フ	ワ		
D			ュ	ス	ヘ	ン		
E			ョ	セ	ホ	。		
F			ッ	ソ	マ	、		

# MBCS

- Multi Byte Character Set

Encoding (인코딩)	Encoding length(인코딩길이)	Locale(지역)
ISO-2022	1-, 2-byte	CJKV
EUC	1~4byte, depending on locale(지역에 의존적)	CJKV
GBK	1-, 2-byte	China(중국)
GB18030	1-, 2-, 4-byte	China(중국)
BigFive	1-, 2-byte	Taiwan(대만) Hong Kong(홍콩)
Shift-JIS	1-, 2-byte	Japan(일본)
Johab	1-, 2-byte	Korea(한국)
UHC	1-, 2-byte	Korea(한국)

# Unicode?

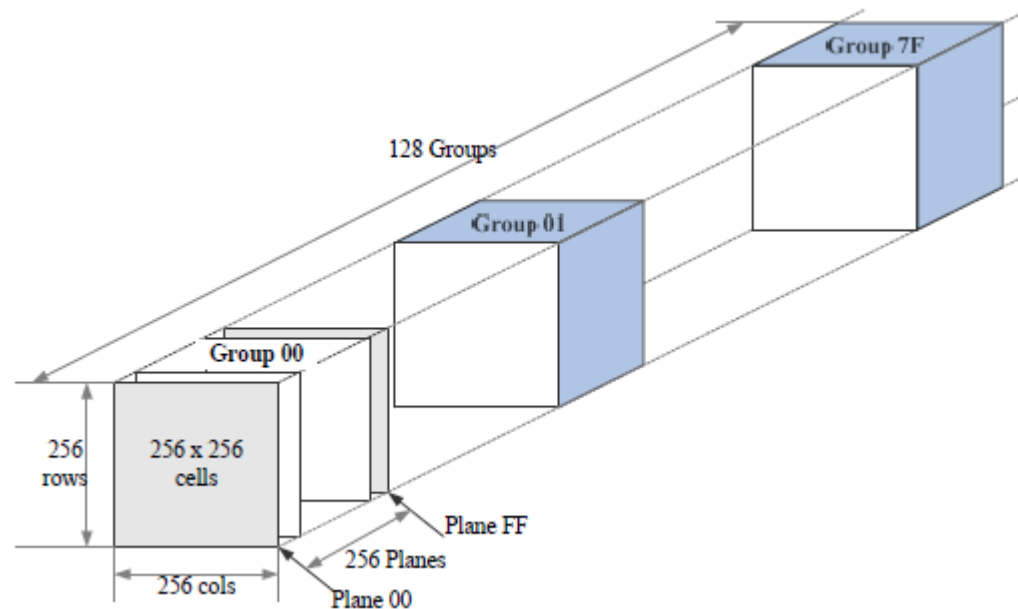
- 한국어(Korean) - 대한민국(KSX1001, EUC-KR), 조선민주주의인민공화국(KPS-9566)
- 일본어(Japanese) - 일본(Shift JIS,EUC-JP)
- 중국어(번체)[Traditional Chinese] - 대만(Big5, EUC-TW)
- 중국어(간체)[Simplified Chinese] - 중국(GB, EUC-CN)
- 동시에 한글, 일본어, 중국어 간체 번체 표현할수 없을까?
- 유로화(€)를 어떻게 표현하나?
- 이럴바에 동시에 표현해버릴까?
- => 문자를 통일하자! => Unicode

# Unicode

- Unicode 1.0.0 - 1991년 8월 제정
- (생략)
- Unicode 6.0 - 2010년 10월 11일 제정
- Unicode 6.1 - 2012년 2월 1일 제정

<http://ko.wikipedia.org/wiki/%EC%9C%A0%EB%8B%88%EC%BD%94%EB%93%9C>

# Unicode



- Cell: 한개의 문자가 할당되는 공간
- Plane:  $256 \times 256 (=65536)$ 개 Cell 묶음
- Group: 256개의 Plane 묶음(00~7F)



# Unicode

- 유니코드 값을 표현 - 코드포인트(code point)사용
- U+ 붙여 사용함.
  - ex) 'A' - U+0041 (또는 \u0041)로 표기함.
- 31비트 문자 집합
- 논리적으로 평면(plane)이라는 개념을 이용, 구획을 나누고, 평면 개수는 총 17개
- Wikipedia - [유니코드평면](#)

# Unicode

- 0번 평면 - 기본 다국어 평면(BMP; Basic Multilingual Plane) ~ 16번 평면
- 대부분의 문자는 U+0000~U+FFFF 범위에 있는 기본 다국어 평면에 속함.
- 일부 한자는 보조 다국어 평면(SMP, Supplementary Multilingual Plane)인 U+10000~U+1FFFF 범위에 속함
- 한글
  - 한글 자모 영역 - U+1100~U+11FF
  - 한글 소리 마디 영역 - U+AC00~U+D7AF

# Unicode

- Unicode Encoding 방식
  - 코드 포인트를 코드화
    - UCS-2, UCS-4
  - 변환 인코딩 형식(UTC, UCS Transformation Format)
    - UTF-7, UTF-8, UTF-16, UTF-32

주로 UTF-8 인코딩을 가장 많이 사용함.

# UTF-8 인코딩 방식

코드 포인트 범위	비트 수	인코딩
U+0000~U+007F	7	그대로 인코딩
U+0080~U+07FF	11	110XXXXX 10XXXXXX
U+0800~U+FFFF	16	1110XXXX 10XXXXXX 10XXXXXX
U+10000~U+1FFFFF	21	11110XXX 10XXXXXX 10XXXXXX 10XXXXXX

- "한글"
  - 코드포인트 - U+D55C U+AE00
  - UTF-8 인코딩 - 0xED 0x95 0x9C 0xEA 0xB8 0x80
  - UTF-8 인코딩에서 한글은 3바이트 인코딩.

# 유니코드에서 한글을 표현하는 방법

## 유니코드 범위 목록에서의 한글 관련 범위

이름	처음	끝	개수
한글 자모(Hangul Jamo)	1100	11FF	256
호환용 한글 자모(Hangul Compatibility Jamo)	3130	318F	96
한글 자모 확장 A(Hangul Jamo Extended A)	A960	A97F	32
한글 소리 마디(Hangul Syllables)	AC00	D7AF	11184
한글 자모 확장 B(Hangul Jamo Extended B)	D7B0	D7FF	80

# Unicode equivalence

## 정규화 방법

## 예

NFD (정준분해)  
Normalization Form Canonical Decomposition

À (U+00C0) → A (U+0041) + ` (U+0300)  
위 (U+C704) → ㅇ (U+110B) + ㅍ (U+1171)

NFC (정준 분해한 뒤 다시 정준 결합)  
Normalization Form Canonical Composition

A (U+0041) + ` (U+0300) → À (U+00C0)  
ㅇ (U+110B) + ㅍ (U+1171) → 위 (U+C704)

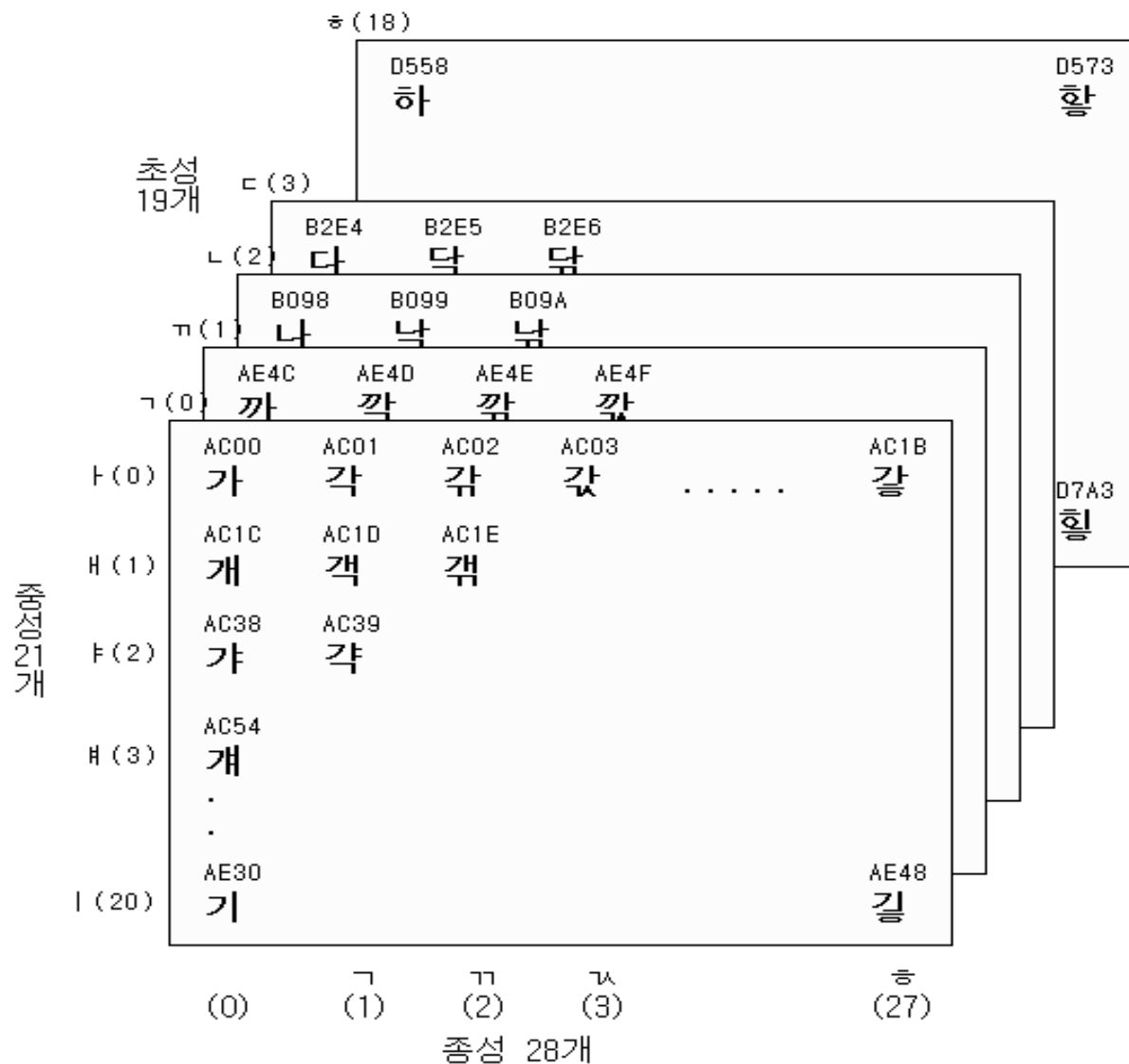
NFKD (호환 분해)  
Normalization Form Compatibility Decomposition

fi (U+FB01) → f (U+0066) + i (U+0069)

NFKC (호환 분해한 뒤 다시 정준 결합)  
Normalization Form Compatibility Composition

樂 (U+F914), 樂 (U+F95C), 樂 (U+F9BF) → 樂 (U+6A02)

# 유니코드에서의 한글 소리 마디



	110	111	112	113	114	115	116	117	118	119	11A	11B	11C	11D	11E	11F
0	ㄱ 1100	ㄴ 1110	ㄷ 1120	ㄹ 1130	ㅁ 1140	ㅂ 1150	<div>HJ F</div> 1160	ㅅ 1170	ㅇ 1180	ㅈ 1190	ㅊ 11A0	ㅋ 11B0	ㅌ 11C0	ㅍ 11D0	ㅎ 11E0	ㅇ 11F0
1	ㄱ 1101	ㄴ 1111	ㄷ 1121	ㄹ 1131	ㅁ 1141	ㅂ 1151	ㅅ 1161	ㅇ 1171	ㅈ 1181	ㅊ 1191	ㅊ 11A1	ㅋ 11B1	ㅌ 11C1	ㅍ 11D1	ㅎ 11E1	ㅇ 11F1
2	ㄱ 1102	ㅎ 1112	ㅅ 1122	ㅈ 1132	ㅊ 1142	ㅋ 1152	ㅌ 1162	ㅍ 1172	ㅈ 1182	ㅊ 1192	ㅊ 11A2	ㅋ 11B2	ㅌ 11C2	ㅍ 11D2	ㅎ 11E2	ㅇ 11F2
3	ㄱ 1103	ㄴ 1113	ㄷ 1123	ㄹ 1133	ㅁ 1143	ㅂ 1153	ㅅ 1163	ㅇ 1173	ㅈ 1183	ㅊ 1193	ㅊ 11A3	ㅋ 11B3	ㅌ 11C3	ㅍ 11D3	ㅎ 11E3	ㅇ 11F3
4	ㄱ 1104	ㄴ 1114	ㄷ 1124	ㄹ 1134	ㅁ 1144	ㅂ 1154	ㅅ 1164	ㅇ 1174	ㅈ 1184	ㅊ 1194	ㅊ 11A4	ㅋ 11B4	ㅌ 11C4	ㅍ 11D4	ㅎ 11E4	ㅇ 11F4
5	ㄱ 1105	ㄴ 1115	ㄷ 1125	ㄹ 1135	ㅁ 1145	ㅂ 1155	ㅅ 1165	ㅇ 1175	ㅈ 1185	ㅊ 1195	ㅊ 11A5	ㅋ 11B5	ㅌ 11C5	ㅍ 11D5	ㅎ 11E5	ㅇ 11F5
6	ㄱ 1106	ㄴ 1116	ㄷ 1126	ㄹ 1136	ㅁ 1146	ㅂ 1156	ㅅ 1166	ㅇ 1176	ㅈ 1186	ㅊ 1196	ㅊ 11A6	ㅋ 11B6	ㅌ 11C6	ㅍ 11D6	ㅎ 11E6	ㅇ 11F6
7	ㄱ 1107	ㄴ 1117	ㄷ 1127	ㄹ 1137	ㅁ 1147	ㅂ 1157	ㅅ 1167	ㅇ 1177	ㅈ 1187	ㅊ 1197	ㅊ 11A7	ㅋ 11B7	ㅌ 11C7	ㅍ 11D7	ㅎ 11E7	ㅇ 11F7
8	ㅁ 1108	ㄴ 1118	ㅅ 1128	ㅈ 1138	ㅊ 1148	ㅋ 1158	ㅌ 1168	ㅍ 1178	ㅈ 1188	ㅊ 1198	ㅊ 11A8	ㅋ 11B8	ㅌ 11C8	ㅍ 11D8	ㅎ 11E8	ㅇ 11F8
9	ㅅ 1109	ㅈ 1119	ㅊ 1129	ㅊ 1139	ㅋ 1149	ㅌ 1159	ㅍ 1169	ㅈ 1179	ㅊ 1189	ㅊ 1199	ㅊ 11A9	ㅋ 11B9	ㅌ 11C9	ㅍ 11D9	ㅎ 11E9	ㅇ 11F9
A	ㅅ 110A	ㅈ 111A	ㅊ 112A	ㅊ 113A	ㅋ 114A	ㅌ 115A	ㅍ 116A	ㅈ 117A	ㅊ 118A	ㅊ 119A	ㅊ 11AA	ㅋ 11BA	ㅌ 11CA	ㅍ 11DA	ㅎ 11EA	ㅇ 11FA
B	ㅇ 110B	ㅎ 111B	ㅅ 112B	ㅈ 113B	ㅊ 114B	ㅋ 115B	ㅌ 116B	ㅍ 117B	ㅈ 118B	ㅊ 119B	ㅊ 11AB	ㅋ 11BB	ㅌ 11CB	ㅍ 11DB	ㅎ 11EB	ㅇ 11FB
C	ㅅ 110C	ㅈ 111C	ㅊ 112C	ㅊ 113C	ㅋ 114C	ㅌ 115C	ㅍ 116C	ㅈ 117C	ㅊ 118C	ㅊ 119C	ㅊ 11AC	ㅋ 11BC	ㅌ 11CC	ㅍ 11DC	ㅎ 11EC	ㅇ 11FC
D	ㅅ 110D	ㅈ 111D	ㅊ 112D	ㅊ 113D	ㅋ 114D	ㅌ 115D	ㅍ 116D	ㅈ 117D	ㅊ 118D	ㅊ 119D	ㅊ 11AD	ㅋ 11BD	ㅌ 11CD	ㅍ 11DD	ㅎ 11ED	ㅇ 11FD
E	ㅅ 110E	ㅈ 111E	ㅊ 112E	ㅊ 113E	ㅋ 114E	ㅌ 115E	ㅍ 116E	ㅈ 117E	ㅊ 118E	ㅊ 119E	ㅊ 11AE	ㅋ 11BE	ㅌ 11CE	ㅍ 11DE	ㅎ 11EE	ㅇ 11FE
F	ㅅ 110F	ㅈ 111F	ㅊ 112F	ㅊ 113F	ㅋ 114F	<div>HC F</div> 115F	ㅅ 116F	ㅇ 117F	ㅈ 118F	ㅊ 119F	ㅊ 11AF	ㅋ 11BF	ㅌ 11CF	ㅍ 11DF	ㅎ 11EF	ㅇ 11FF



# Windows – Character Map



# Ubuntu - Character Map


문자 표

Ubuntu 굵게(B) 기울이기(I) 22

문자 범위

- 크메르 문자
- 키릴 문자
- 키프로스어
- 타갈로그 문자
- 타그반와 문자
- 타너 문자
- 타밀 문자
- 타이 문자
- 타이레 문자
- 타이비엣 문자
- 타이샴 문자
- 타크리 문자
- 텔루구 문자
- 티베트 문자
- 티피나그 문자
- 파르티아어 새김 ...
- 파스파
- 팔레비어 새김 문...
- 페니키아 문자
- 하누누 문자
- 한글
- 한자**
- 황실 아람어 알파...
- 히라가나
- 히브리 문자

문자 표(R) 문자 자세히(D)



**U+4EBA CJK UNIFIED IDEOGRAPH-4EBA**

**일반 문자 속성**

유니코드에 포함된 시점: 1.1  
유니코드 범위: 문자, 기타

**다른 표기법**

UTF-8: 0xE4 0xBA 0xBA  
UTF-16: 0x4EBA

C언어 8진수로 표기한 UTF-8: \344\272\272  
XML 10진수 엔티티: &#20154;

**한중일 한자 정보**

뜻을 영어로 풀어놓기: man; people; mankind; someone else  
북경어 표기: rén  
광둥어 표기: jan4  
일본어 음독: JIN NIN  
일본어 훈독: HITO  
중국 고어 표기: \*njin njin  
한글 표기: IN

목사할 텍스트(T): 목사(C)

U+4EBA CJK UNIFIED IDEOGRAPH-4EBA man; people; mankind; someone else

# Perl

- CPAN modules
  - **Unicode::UniHan**
  - `Lingua::JA::Romanize::Japanese`
  - `Lingua::ZH::Romanize::Pinyin`
  - `Lingua::ZH::Romanize::Cantonese`
  - `Lingua::KO::Romanize::Hangul`

# Perl – Unicode::Normalization

- UTF-8 encoded input
- Decode
- NFD(Normalization Form Canonical Decomposition)
- Perl Unicode string
- NFC(Normalization Form Canonical Composition)
- Encode
- UTF-8 encoded output

# Perl – Unicode::Normalization

```
#!/usr/bin/perl
use utf8;
use Unicode::Normalize;
#NFD can be helpful on input
my $str = NFD("위");
#NFC is recommended on output
my $output = NFC($str);
Print "=>";
print $output;
print "\n";
```

=> 위

# Java - Normalizer

```
import java.text.Normalizer;
public class NormalizerTest {
    public static void sprintIt(String test) {
        System.out.println(test);
        for (int i = 0; i < test.length(); i++) {
            System.out.println(String.format("U+%4X ", test.codePointAt(i)));
        }
        System.out.println();
    }
    public static void main(String args[]) {
        String ui = "위"; sprintIt(ui);
        String nfd = Normalizer.normalize(ui, Normalizer.Form.NFD); sprintIt(nfd);
        String nfc = Normalizer.normalize(nfd, Normalizer.Form.NFC); sprintIt(nfc);
    }
}
```

- 위
- U+C704 ,
- 위
- U+110B ,
- U+1171 ,
- 위
- U+C704 ,

# Perl – Unicode::Unihan

```
#!/usr/bin/perl
use utf8;
use v5.12;
use strict;
use Unicode::Unihan;
my $str    = "韓國";
my $unhan = Unicode::Unihan->new;
for my $lang (qw(Mandarin Cantonese Korean JapaneseOn JapaneseKun
)) {
    printf "CJK $str in %-12s is ", $lang;
    say $unhan->$lang($str);
}
```

- CJK 韓國 in Mandarin is HAN2GUO2
- CJK 韓國 in Cantonese is hon4gwok3
- CJK 韓國 in Korean is HANKWUK
- CJK 韓國 in JapaneseOn is KANKOKU
- CJK 韓國 in JapaneseKun is IGETAKUNI

# Perl – Unicode::UniHan

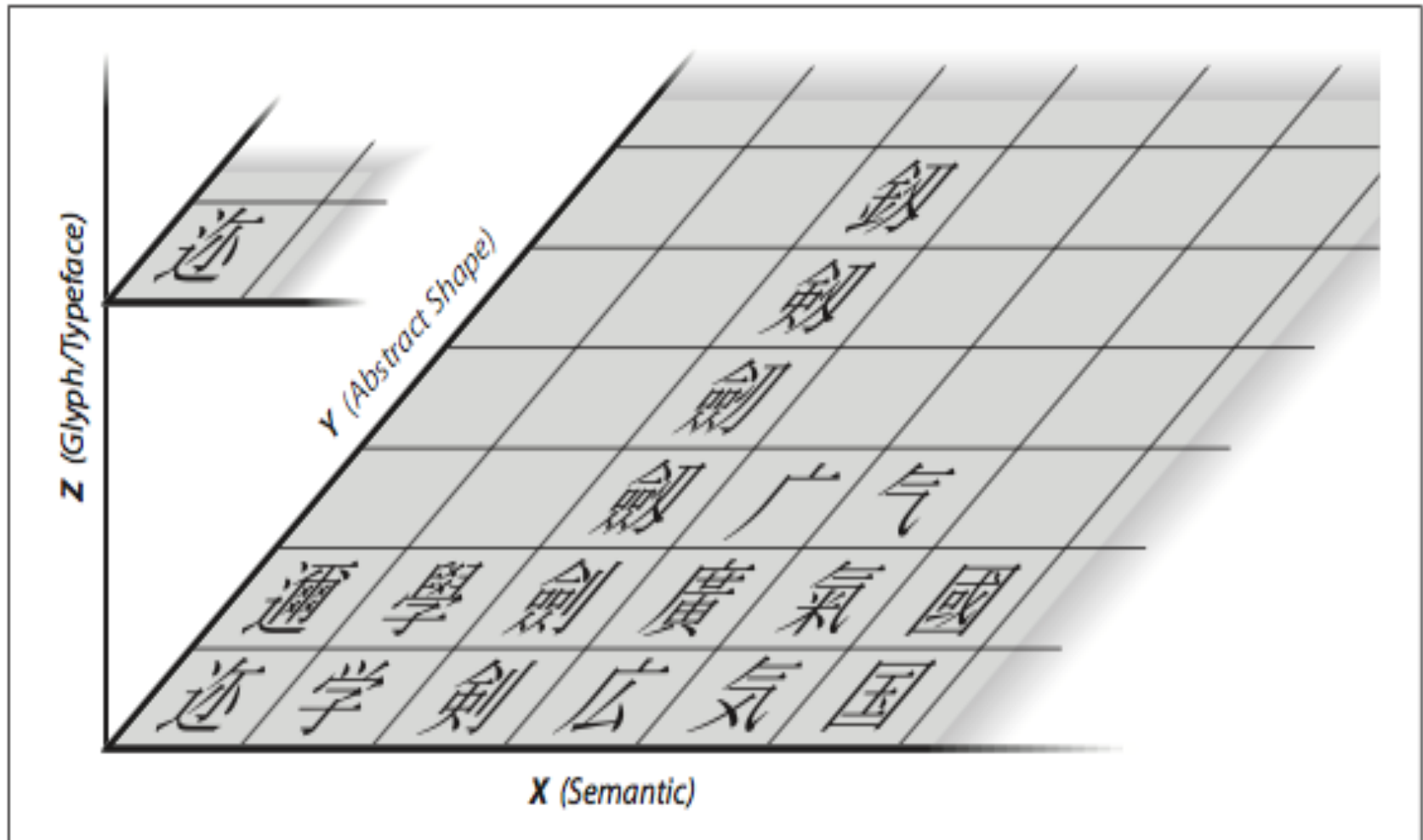


Figure 3-1. Three-axis model for comparing glyphs for ideographs



# Perl – Unicode::UniHan

```
#!/usr/bin/perl
use Unicode::UniHan;
use Encode;
$uh = Unicode::UniHan->new();
print $uh->ZVariant(decode(q(utf8), q(亮)));
print "\n";
## U+4EAE
print $uh->ZVariant(decode(q(utf8), q(亮)));
print "\n";
## U+F977
```

# Python - Unicodedata

- 한자 알아내기

- 수: 망치 마

- Python – Unicodedata

<http://docs.python.org/library/unicodedata.html>

```
>>> u"수"
```

```
u'\u3403'
```

```
>>> import unicodedata;
```

```
>>> unicodedata.name(u"수");
```

```
'CJK UNIFIED IDEOGRAPH-3403'
```

- <http://www.unicode.org/cgi-bin/GetUnihanData.pl?codepoint=3403>

# 참고(Reference)

- CJKV Information Processing, 1<sup>st</sup> Edition(Ken Lunde)
- CJKV Information Processing, 2<sup>nd</sup> Edition(Ken Lunde)
- CJK Compatibility Ideographs  
<http://www.isthisthingon.org/unicode/index.phtml?page=0F&subpage=9>
- Unicode
  - [http://www.novonetworks.com/jamestic/Unicode\\_1.0.pdf](http://www.novonetworks.com/jamestic/Unicode_1.0.pdf)
  - Unicode 이해하기 <http://www.slideshare.net/parkpd/unicode-4796992>
- 한글 인코딩의 이해 1편: 한글 인코딩의 역사와 유니코드  
<http://helloworld.naver.com/helloworld/19187>
- 한글 인코딩의 이해 2편: 유니코드와 Java를 이용한 한글 처리  
<http://helloworld.naver.com/helloworld/76650>
- JIS X 0201  
[http://www.cqpub.co.jp/interface/toku/2002/200212/toku1\\_3.htm](http://www.cqpub.co.jp/interface/toku/2002/200212/toku1_3.htm)



# 참고(Reference)

- Chinese Japan Korean Unicode
  - 전각, 반각 <http://nlp.solutions.asia/?p=39>
  - Unicode Table 유니코드 테이블 모음 <http://www.unicode.org>
  - CJK Unified Ideographs Extension A  
<http://www.unicode.org/charts/PDF/U3400.pdf>
  - CJK Unified Ideographs Extension B  
<http://www.unicode.org/charts/PDF/U20000.pdf>
  - CJK Unified Ideographs Extension C  
<http://www.unicode.org/charts/PDF/U2A700.pdf>
  - CJK Unified Ideographs Extension D  
<http://www.unicode.org/charts/PDF/U2B740.pdf>
  - CJK Compatibility Ideographs Supplement  
<http://www.unicode.org/charts/PDF/U2F800.pdf>
  - Hiragana <http://www.unicode.org/charts/PDF/U3040.pdf>

# 참고(Reference)

- Perl
  - [Perl] Lingua::\*::Romanize::\* 日中韓／ローマ字変換モジュール  
<http://www.kawa.net/works/perl/romanize/romanize.html>
  - perlunicook – cookbookish examples of handling Unicode in Perl  
<http://98.245.80.27/tcp/cripts/perlunicook.html>
  - Unicode CJK Compatible Variations  
<http://www.slideshare.net/lamp.purl/unicode-cjk-compatible-variations>

# Q&A

- 질문과 답변
- 質問と回答
- 問答/问答
- Thanks [@ganadist](#) (He lend CJKV 1<sup>st</sup> edition to me.) and [@JEEN LEE](#)
- [@studioego](#)