Input Methods for Chinese/Japanese/Korean (CJK)

Dihong Jiang

d47jiang@uwaterloo.ca

July 22, 2021

Content

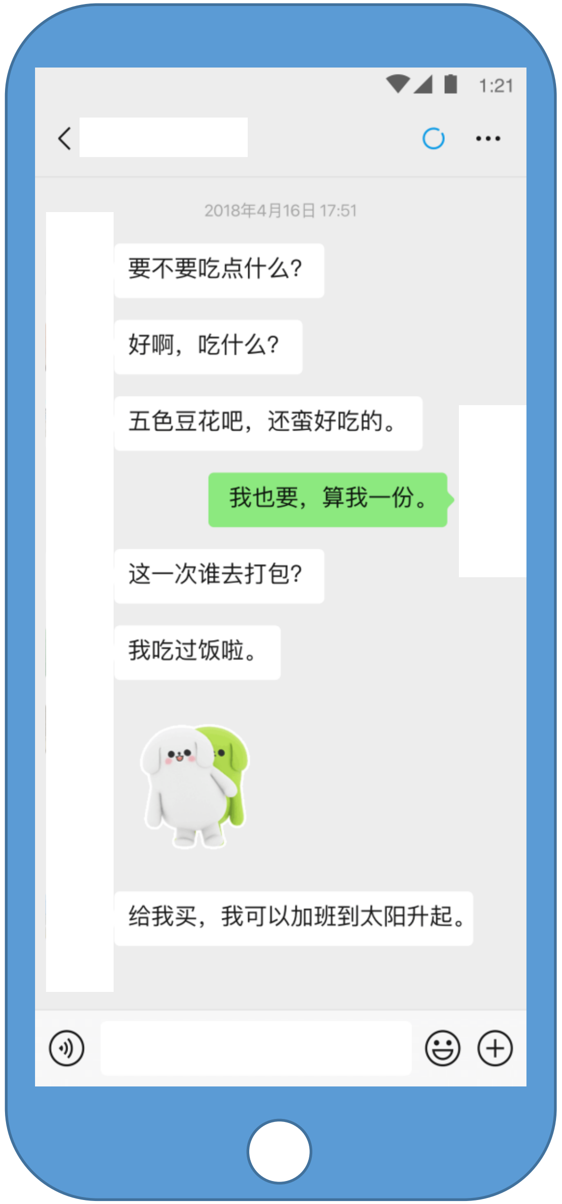
- Background
- Input Methods
  - Input hiragana / katakana
  - Input hangul
  - Input hanzi / kanji / hanja
  - How IME works
- Conclusions

Background

1

1 My motivation story for this topic

“How do you type in these characters?”



2018年4月16日 17:51

要不要吃点什么？

好啊，吃什么？

五色豆花吧，还蛮好吃的。

我也要，算我一份。

这一次谁去打包？

我吃过饭啦。

给我买，我可以加班到太阳升起。

1

Alphabetic vs. Ideographic

Human language (writing)

Alphabetic

- by the use of symbols expressing the sounds of speech
- E.g. English, French, German.

Ideographic

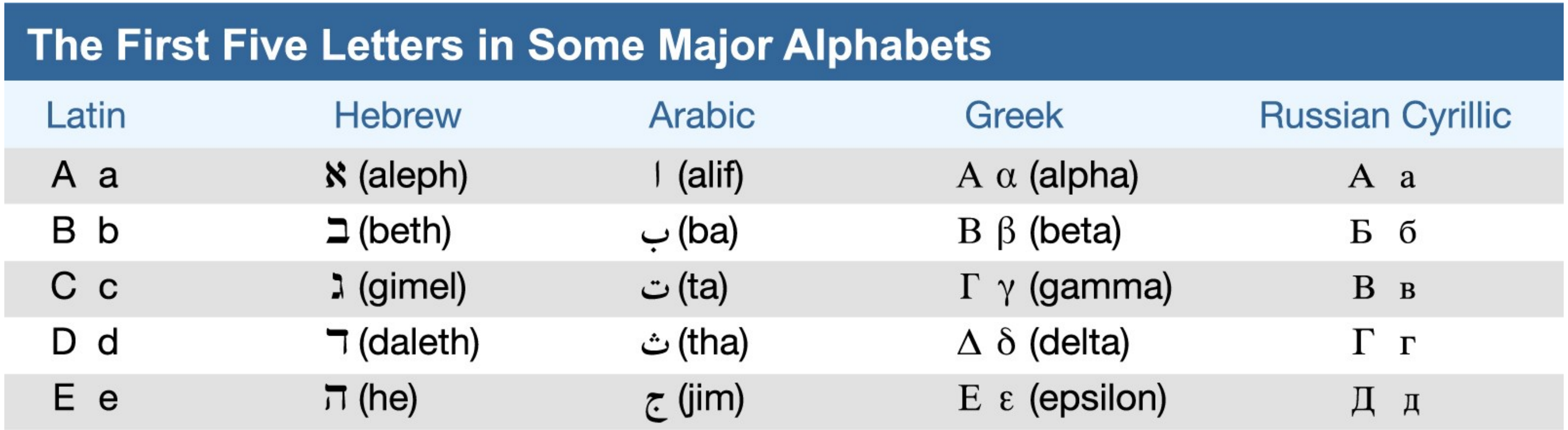
- by the use of symbols expressing the meanings
- E.g. Chinese, ancient Egyptian.

1

Alphabetic vs. Ideographic

Alphabetic:

- Alphabetic languages use a standardized set of letters – basic written symbols -- to compose words. Each of the letter roughly represents a phoneme, a spoken sound.
- Number of alphabets are limited, e.g. 26 in English.
- Whatever typed from the keyboard becomes part of a word.



The First Five Letters in Some Major Alphabets

Latin	Hebrew	Arabic	Greek	Russian Cyrillic
A a	א (aleph)	ا (alif)	Α α (alpha)	А а
B b	ב (beth)	ب (ba)	Β β (beta)	Б б
C c	ג (gimel)	ت (ta)	Γ γ (gamma)	В в
D d	ד (daleth)	ث (tha)	Δ δ (delta)	Г г
E e	ה (he)	ج (jim)	Ε ε (epsilon)	Д д

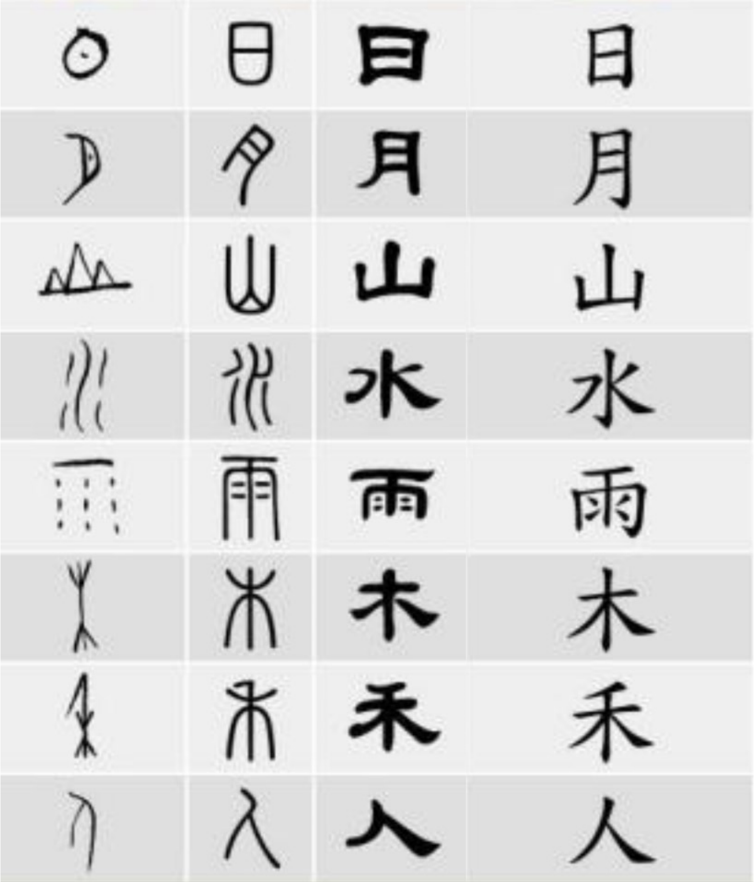
Image source: https://www.britannica.com/topic/alphabet-writing/Development-and-diffusion-of-alphabets

1

Alphabetic vs. Ideographic

Ideographic:

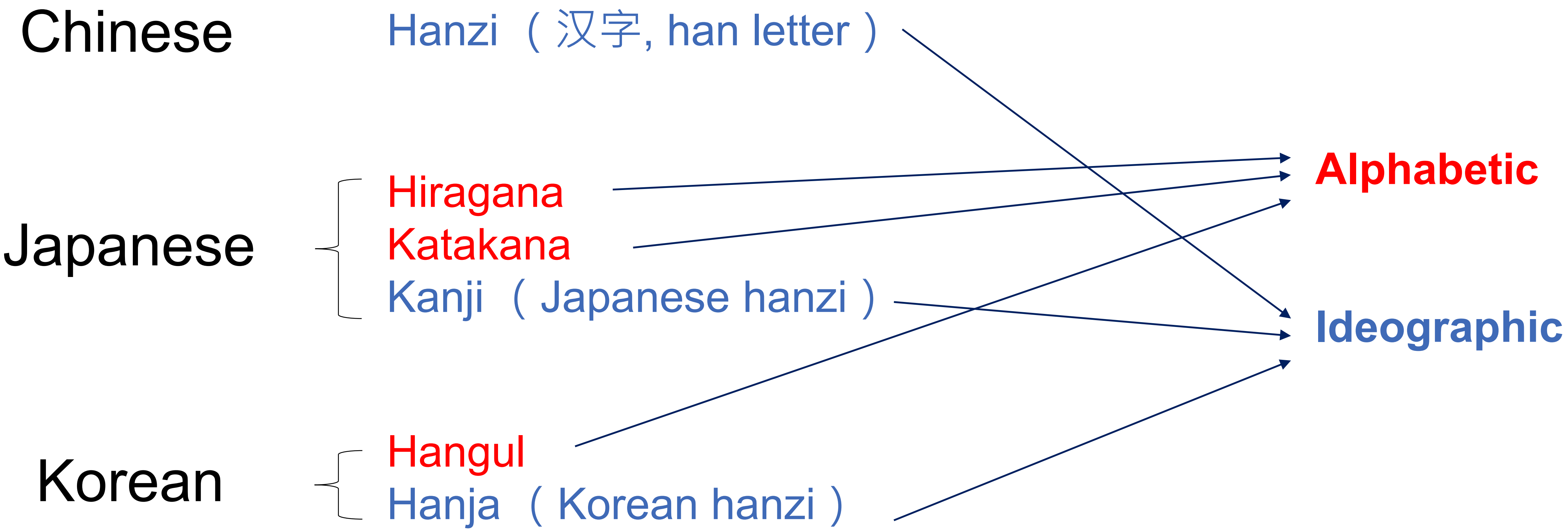
- Ideographic languages represent the language by means of ideograms, i.e. symbols representing ideas and concepts rather than sounds.
- Number of characters are usually unbounded, e.g. more than 50k Chinese characters => needs different input methods.



Meaning	Oracle bone script	Seal script	Clerical script	Regular script
Sun			日	日
Moon			月	月
Mountain			山	山
Water			水	水
Rain			雨	雨
Wood			木	木
Rice plant			禾	禾
Human			人	人

Image source: http://hanzibunan.weebly.com/36896233832786165306 3593724418.html

1 Connections among CJK



Chinese

Hanzi （汉字, han letter）

Japanese

Hiragana

Katakana

Kanji （Japanese hanzi）

Korean

Hangul

Hanja （Korean hanzi）

Alphabetic

Ideographic

1 Classification of input methods

- **Keyboard & mouse operations (mainly focus on this category)**
- Visual – Optical Character Recognition (OCR)
- Audio – Voice recognition
- Future – Brain wave?

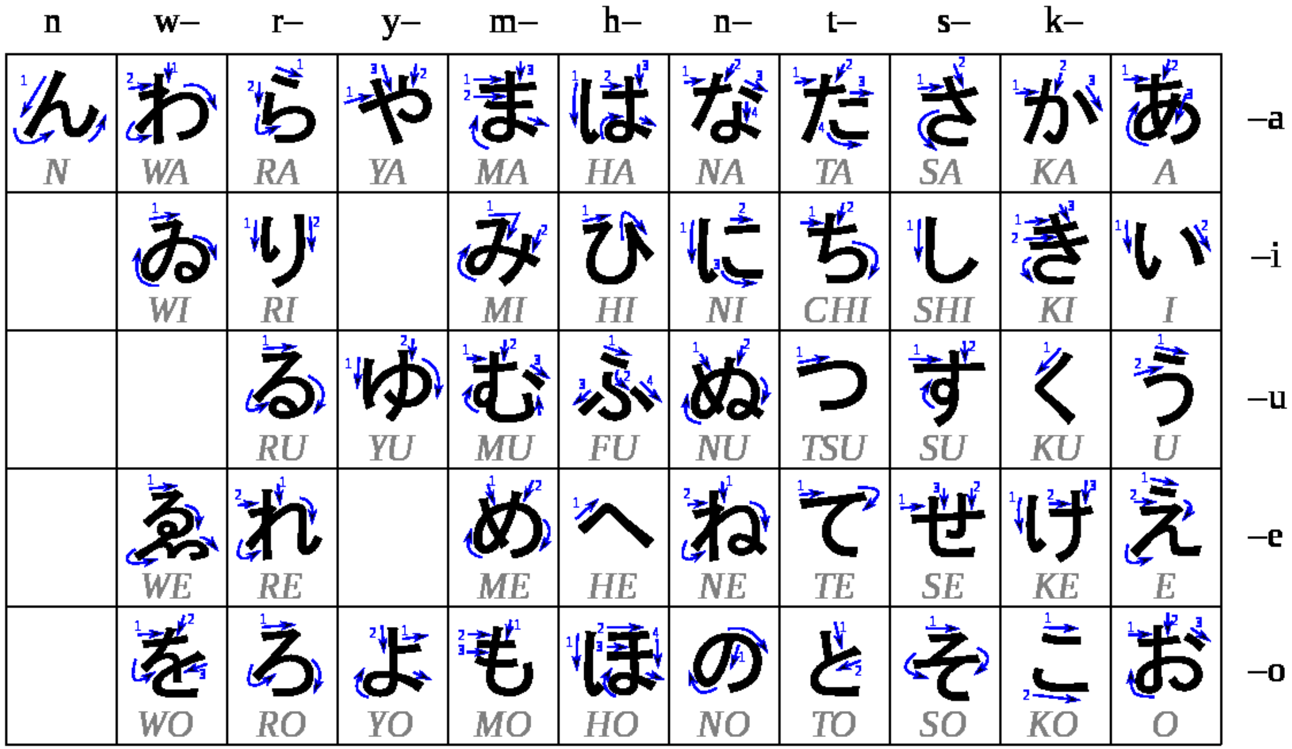
Input Methods

2

2 Input hiragana

Hiragana

- Mainly for spelling out Japanese originated words.
- More basic and common in Japanese writing.
- Input using romanization of Japanese (a.k.a. romaji, use of Latin alphabets to write Japanese language), or using Japanese keyboard to input directly.



n	w–	r–	y–	m–	h–	n–	t–	s–	k–		
ん N	わ WA	ら RA	や YA	ま MA	は HA	な NA	た TA	さ SA	か KA	あ A	–a
	ゐ WI	り RI		み MI	ひ HI	に NI	ち CHI	し SHI	き KI	い I	–i
		る RU	ゆ YU	む MU	ふ FU	ぬ NU	つ TSU	す SU	く KU	う U	–u
	ゑ WE	れ RE		め ME	へ HE	ね NE	て TE	せ SE	け KE	え E	–e
	を WO	ろ RO	よ YO	も MO	ほ HO	の NO	と TO	そ SO	こ KO	お O	–o

Image source: https://en.wikipedia.org/wiki/Hiragana

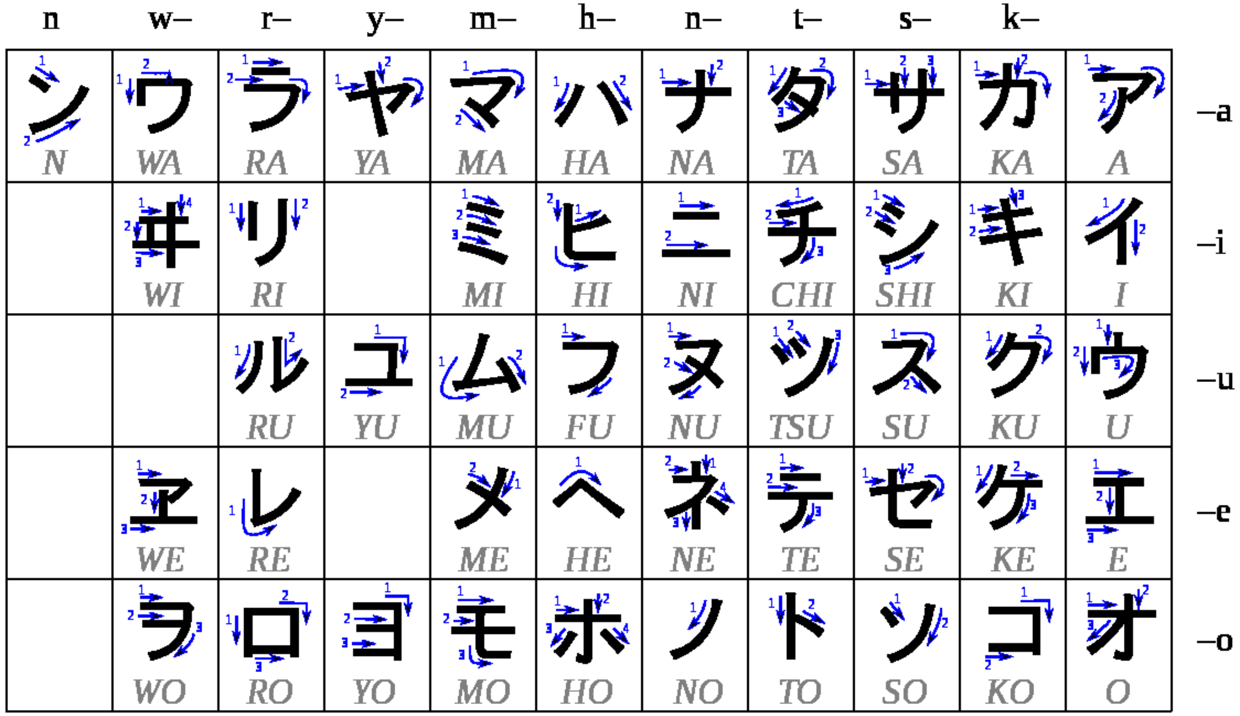
	romaji		hiragana
Example:	SA	->	さ
	KI	->	き

2 Input katakana

Katakana

- Mostly for spelling out foreign words.
- Sharing the same set of syllables as hiragana
- Can be transferred from hiragana, which alerts the reader to the fact that the word is an imported one
- Input by being transferred from hiragana, or using Japanese keyboard to input directly (switching modes between hiragana and katakana)

	romaji		hiragana		katakana
Example:	SA	->	さ	->	サ
	KI	->	き	->	キ



n	w–	r–	y–	m–	h–	n–	t–	s–	k–		
ン N	ワ WA	ラ RA	ヤ YA	マ MA	ハ HA	ナ NA	タ TA	サ SA	カ KA	ア A	–a
	ヰ WI	リ RI		ミ MI	ヒ HI	ニ NI	チ CHI	シ SHI	キ KI	イ I	–i
		ル RU	ユ YU	ム MU	フ FU	ヌ NU	ツ TSU	ス SU	ク KU	ウ U	–u
	ヱ WE	レ RE		メ ME	ヘ HE	ネ NE	テ TE	セ SE	ケ KE	エ E	–e
	ヲ WO	ロ RO	ヨ YO	モ MO	ホ HO	ノ NO	ト TO	ソ SO	コ KO	オ O	–o

Image source: https://en.wikipedia.org/wiki/Katakana

2 Input hiragana/katakana

- Using Japanese keyboard that can switch between hiragana and katakana

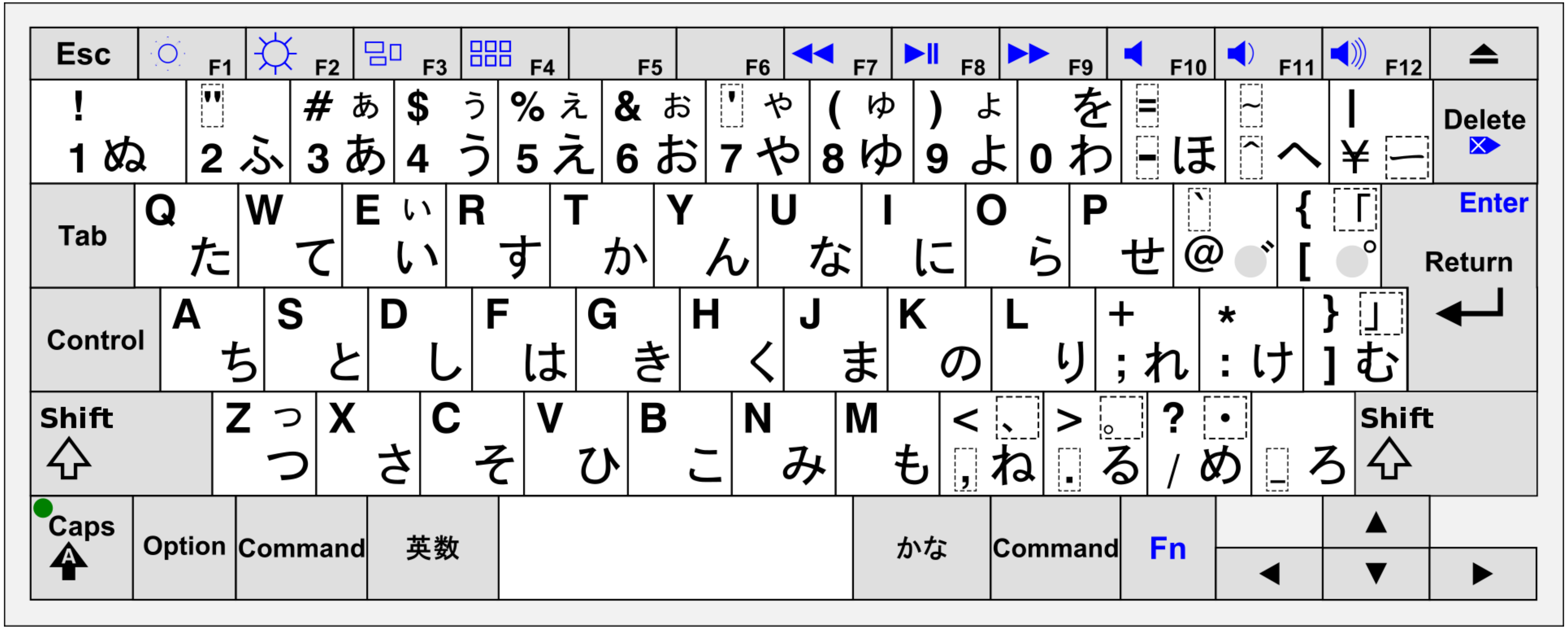


Image source: https://en.wikipedia.org/wiki/Language_input_keys#/media/File:KB_Japanese_Mac_-_Apple_Keyboard_(MB869JA).svg

2 Input hangul

Hangul

- 24 basic letters: 14 consonant letters (ㄱ ㄴ ㄷ ㄹ ㅁ ㅂ ㅅ ㅇ ㅈ ㅊ ㅋ ㅌ ㅍ ㅎ) and 10 vowel letters (ㅏ ㅑ ㅓ ㅕ ㅗ ㅛ ㅜ ㅠ ㅡ ㅣ). There are also 27 complex letters formed by combining the basic letters
- Hangul letters were designed to model the physical morphology of the tongue, palate and teeth, so they may look like ideographic but they are actually alphabetic
- South Korea developed a Romanization system, i.e. romaja, but it is rarely used by Koreans.

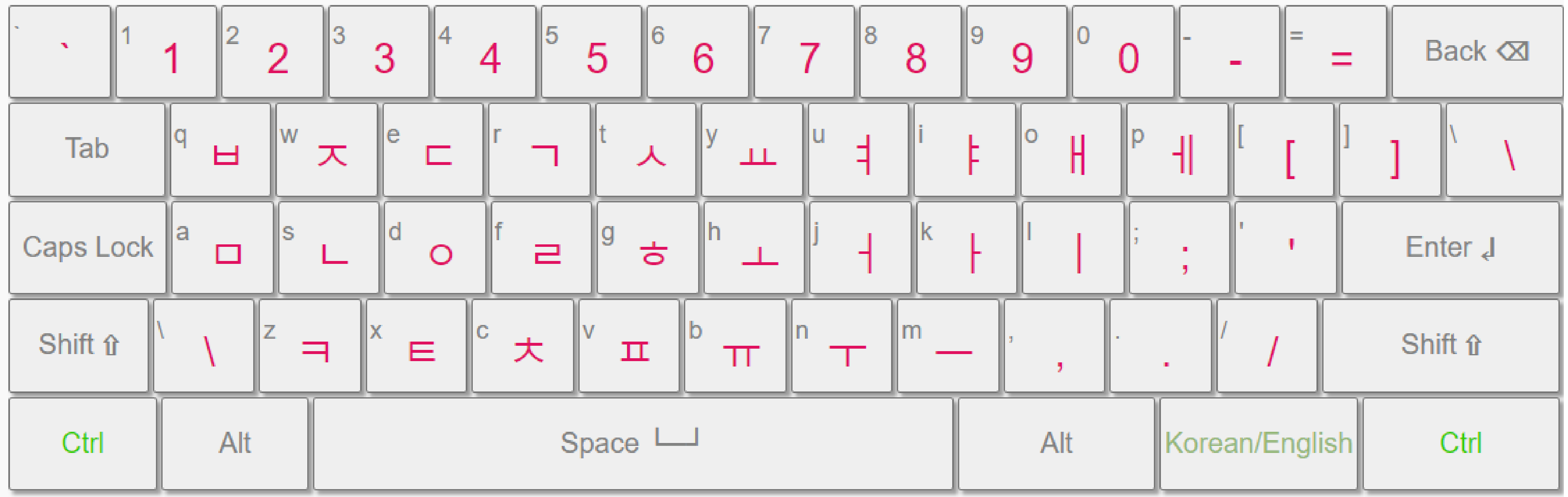
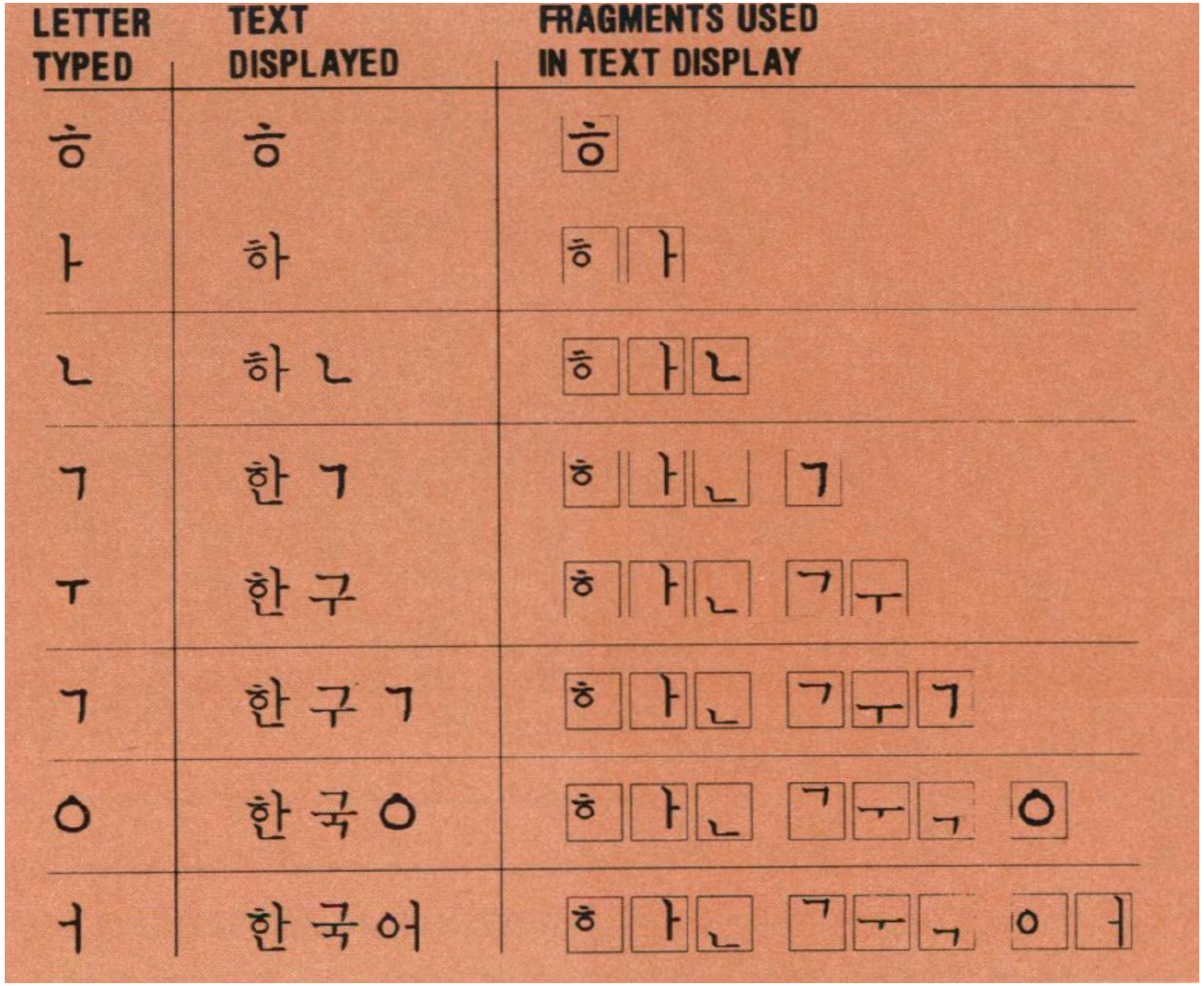


Image source: https://www.clavier-arabe.co/p/korean-keyboard.html

2 Input hangul

- A Hangul character is composed of a consonant followed by a vowel. A final consonant is optional.
- Syllables are organized into blocks according to some rules
- An example of typing in eight successive hangul letters, i.e. “han gug-eo”, which literally means “Korean”. [1]



	LETTER TYPED	TEXT DISPLAYED	FRAGMENTS USED IN TEXT DISPLAY
h	ㅎ	ㅎ	ㅎ
a	ㅏ	하	ㅎ ㅏ
n	ㄴ	하 ㄴ	ㅎ ㅏ ㄴ
g	ㄱ	한 ㄱ	ㅎ ㅏ ㄴ ㄱ
u	ㅜ	한 구	ㅎ ㅏ ㄴ ㄱ ㅜ
g	ㄱ	한 구 ㄱ	ㅎ ㅏ ㄴ ㄱ ㅜ ㄱ
-	ㅇ	한 국 ㅇ	ㅎ ㅏ ㄴ ㄱ ㅡ ㄱ ㅇ
eo	ㅓ	한 국 어	ㅎ ㅏ ㄴ ㄱ ㅡ ㄱ ㅇ ㅓ

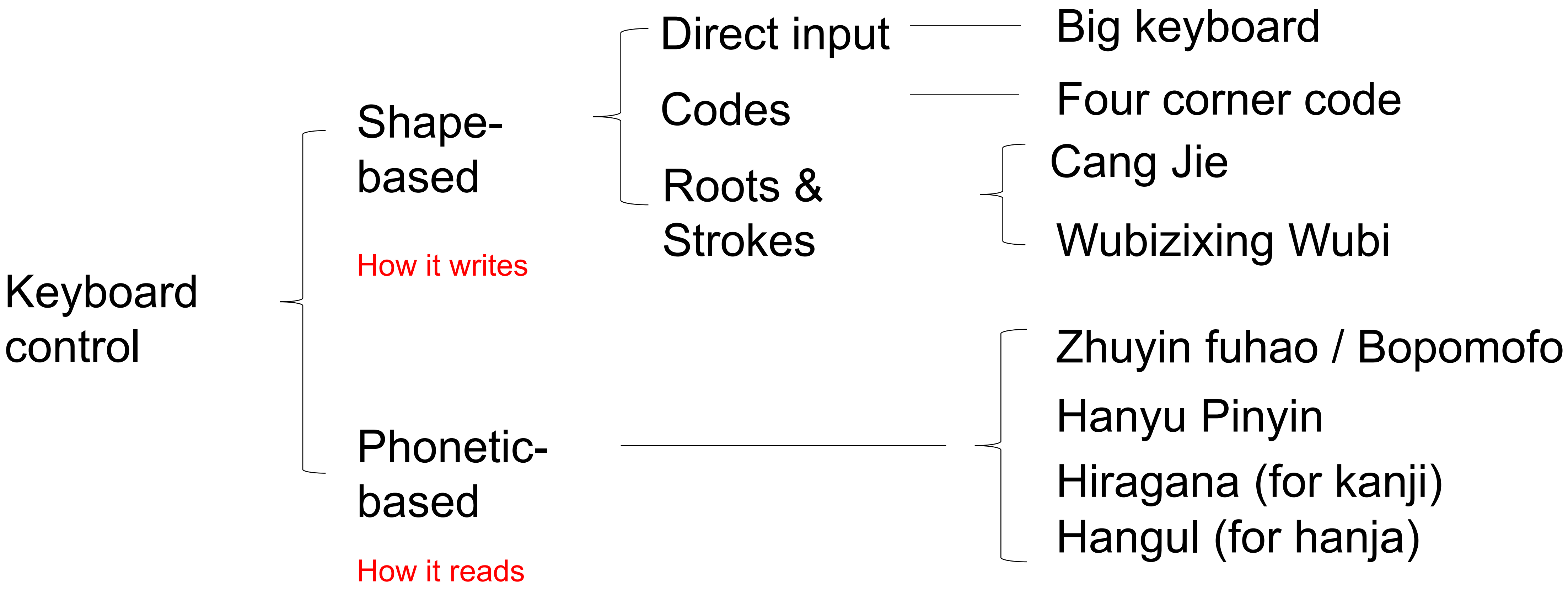
han gug –eo (Korean)

Image source: Ref. [1]

2

Input hanzi / kanji / hanja

- Hanzi: input methods can be categorized into two classes: (1) shape-based, and (2) phonetic-based.



Keyboard control

Shape-based

How it writes

Direct input

Big keyboard

Codes

Four corner code

Roots & Strokes

Cang Jie

Wubizixing Wubi

Phonetic-based

How it reads

Zhuyin fuhao / Bopomofo

Hanyu Pinyin

Hiragana (for kanji)

Hangul (for hanja)

2

Input hanzi / kanji / hanja – shape-based

Direct input: one character per key using a big keyboard

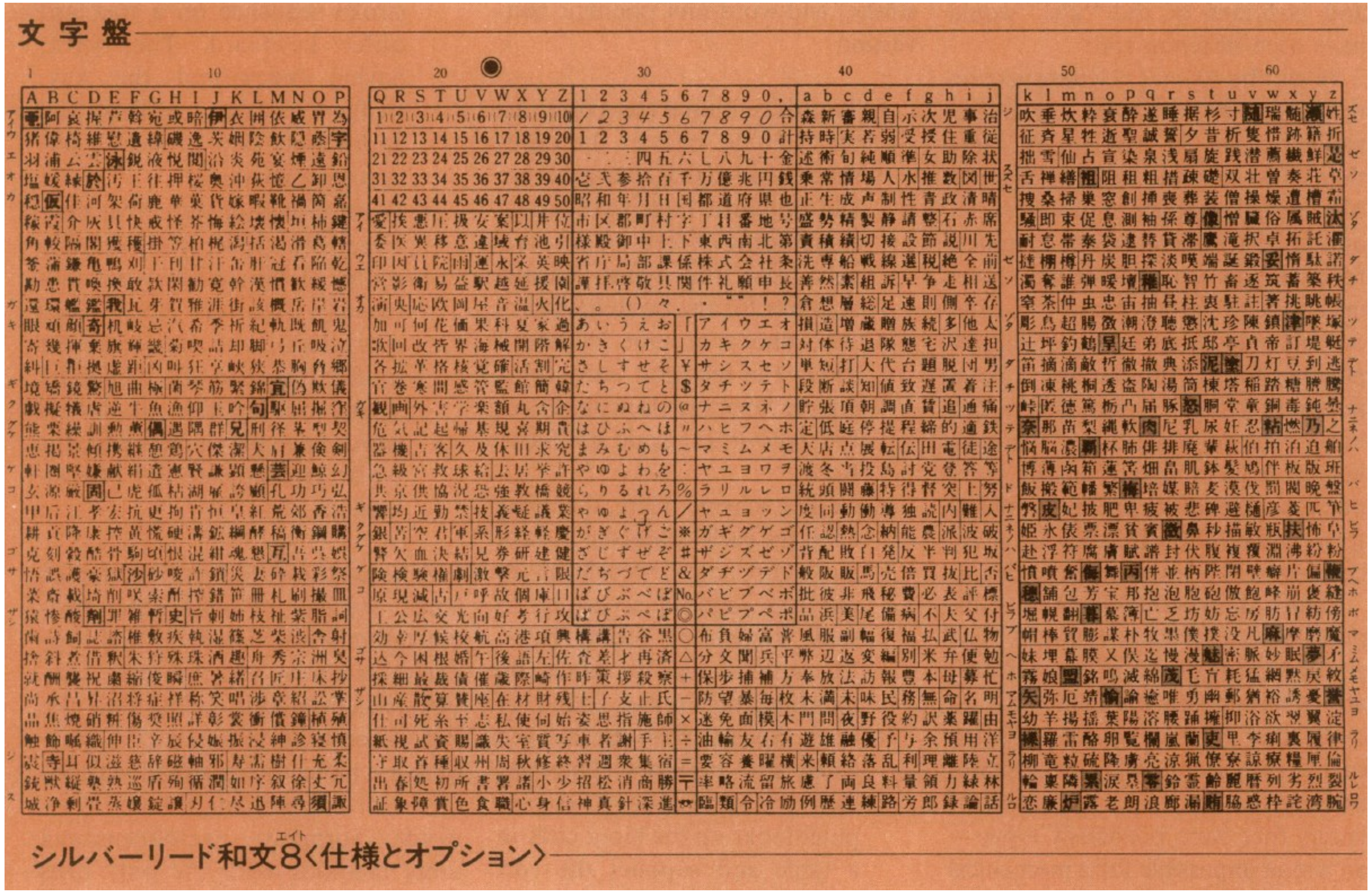
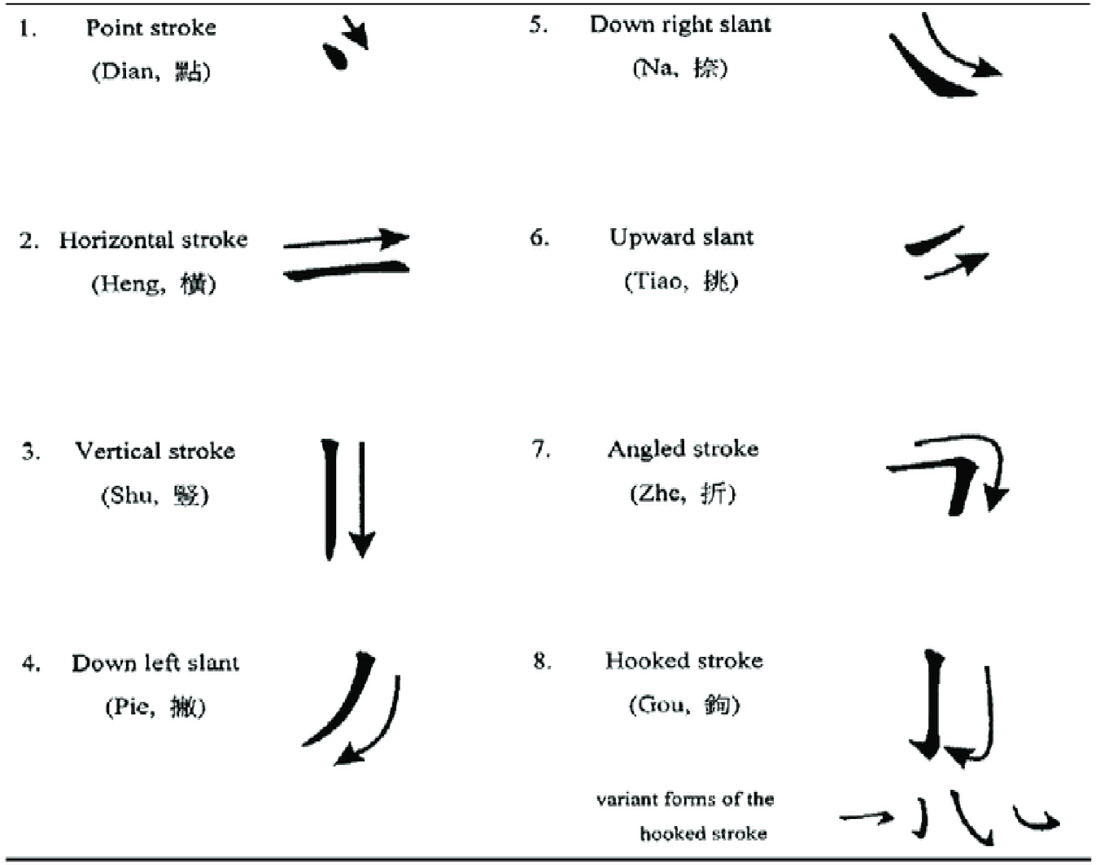


Image source: Ref. [1]

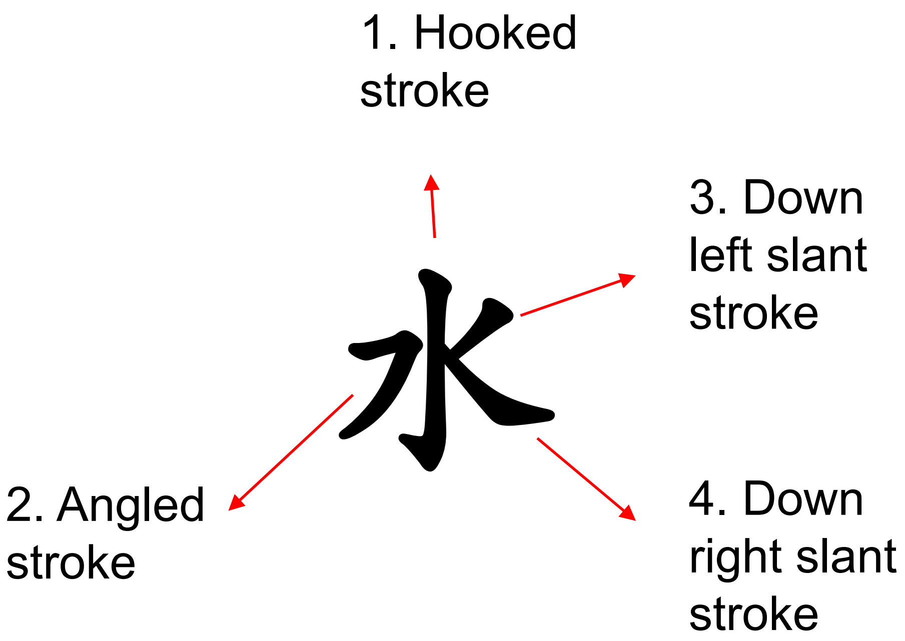
2 Chinese strokes

- Basic strokes of Chinese characters



1.	Point stroke (Dian, 點)	5.	Down right slant (Na, 捺)	
2.	Horizontal stroke (Heng, 横)	6.	Upward slant (Tiao, 挑)	
3.	Vertical stroke (Shu, 豎)	7.	Angled stroke (Zhe, 折)	
4.	Down left slant (Pie, 撇)	8.	Hooked stroke (Gou, 鉤)	
			variant forms of the hooked stroke	

Image source: Ref. [2]



1. Hooked stroke

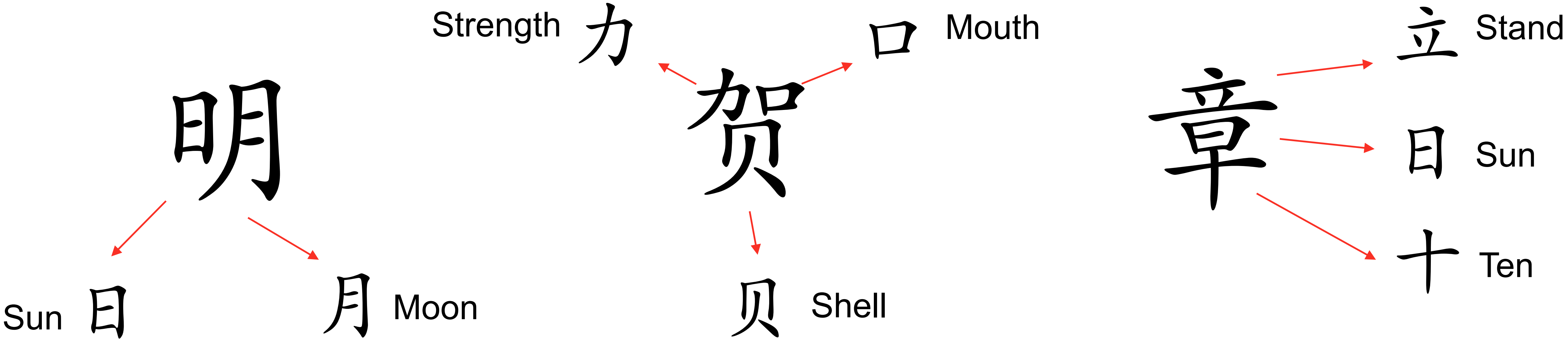
2. Angled stroke

3. Down left slant stroke

4. Down right slant stroke

2 Roots of Chinese characters

- Roots: the smallest component of Chinese characters



明
Sun 日
月 Moon

Bright

贺
Strength 力
口 Mouth
贝 Shell

Congratulate

章
立 Stand
日 Sun
十 Ten

Chapter

Note: Sometimes you can find relations between the character and its components, but this is not always true.

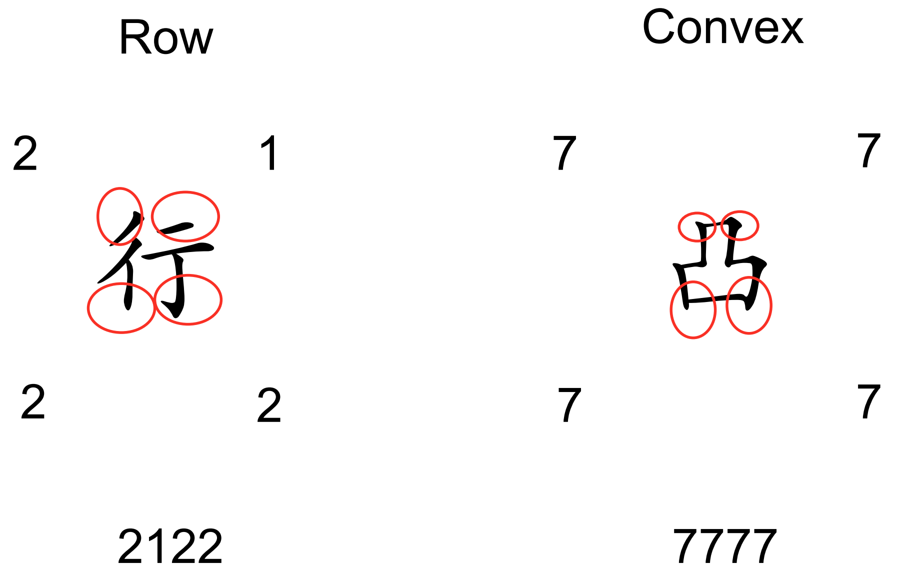
2

Input hanzi – shaped-based

- Shaped-based code: Four corner codes
  - Invented by Yunwu Wang in 1920s
  - Using 0-9 to denote 10 different kinds of strokes. The four digits encode the shapes found in the four corners of the symbol, top-left to bottom-right. Although this does not uniquely identify a Chinese character, it leaves only a very short list of possibilities. A fifth digit can be added to describe an extra part above the bottom-right if necessary.

Digit	Strokes
0	亠
1	一乚（horizontal, upward slant, right hook）
2	丨丿亅（vertical, down left slant, left hook）
3	丶（point, down right slant）
4	十乂（cross）
5	扌（multiple strokes inserted on one stroke）
6	口（surrounded four sides without extensions）
7	乛「」㇗（connection of horizontal and vertical）
8	八丷人（two strokes without cross）
9	小忄⺌（three strokes）

Examples:



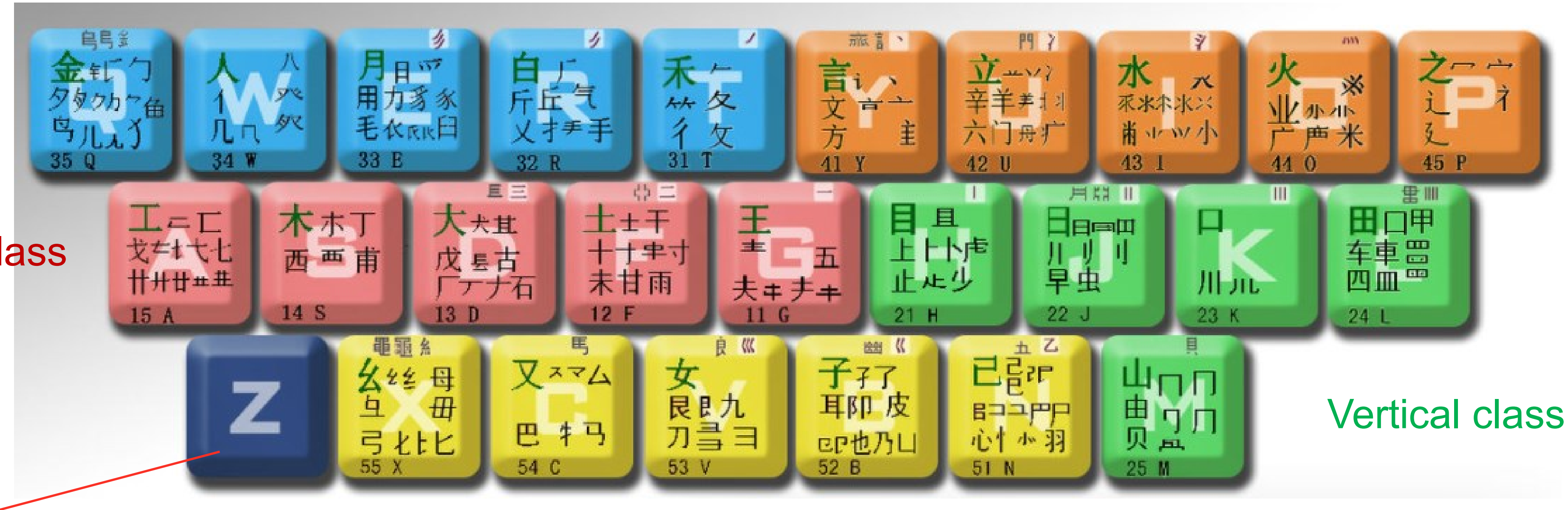
Input hanzi – shape-based

- Wubizixing / Wubi （五笔字型, five-stroke input）
  - Invented by Yongmin Wang in 1983
  - Group keyboard into five regions based on strokes and roots
  - More popular for simplified Chinese (also can be used for traditional Chinese)
  - Input as how you write the character. Need to know the stroke order.
  - Extremely popular in China in 1990s

Down left slant class

Point and down right slant class

Horizontal class



Vertical class

Hooked class

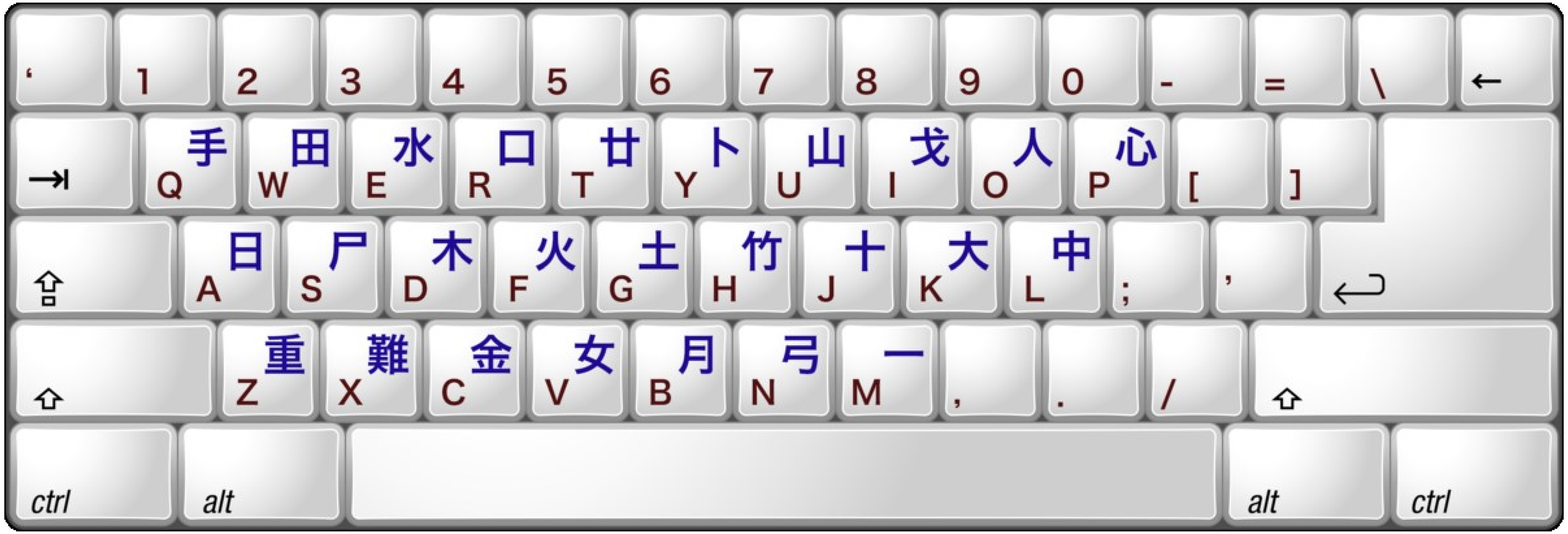
Wildcards

Image source: https://en.wikipedia.org/wiki/Wubi_method#/media/File:5strokes.jpg

2

Input hanzi – shape-based

- Cang Jie (仓颉输入法, Tsang Jie is a mythological god who was thought to be the inventor of hanzi.)
  - Invented by Bong-Foo Chu in 1976. He released the patent in 1982.
  - More efficient for traditional Chinese, thus still popular in Hong Kong, Taiwan, Macau
  - Every key represent a root and all its variants
  - You are playing jigsaw puzzles!



` 1 2 3 4 5 6 7 8 9 0 - = \ ←
Q 手 W 田 E 水 R 口 T 廿 Y 卜 U 山 I 戈 O 人 P 心 [ ]
A 日 S 尸 D 木 F 火 G 土 H 竹 J 十 K 大 L 中 ; '
Z 重 X 難 C 金 V 女 B 月 N 弓 M 一 , . /
ctrl alt alt ctrl

Image source: https://en.wikipedia.org/wiki/Cangjie_input_method#/media/File:Keyboard_layout_cangjie.png

2 Input hanzi – shape-based

- Example of Cang Jie Input: inputting the roots sequentially.

English	Bright	Chapter
Hanzi	明	章
CJ roots	日 月	卜 廿 日 十
Keystrokes	A B	Y T A J

Note: every root on the keyboard has many variants, so you may not find the exactly same root on the keyboard (need to learn Cang Jie code)

Input kanji / hanja – phonetic-based

- Input syllables using hiragana / hangul as phonetic alphabets
  - Attention: extra operation is needed when homophones occur

- Examples of inputting kanji:

romaji -> hiragana -> kanji

ko u shi yo u -> こうしよう

こうしよう	
交渉	negotiation
考証	research
公証	notarization
高尚	noble
校章	school badge
…	…

- Examples of inputting hanja:

hangul -> hanja

ㄴ ㅏ ㅁ -> 남

남	
南	south
男	male
藍	blue
襤	ragged
嵐	mist
…	…

Chinese phonetic systems

- Zhuyin fuhao / Bopomofo（注音符号, phonetic symbols）
  - Introduced by Chinese government in 1910s
  - The name Bopomofo comes from the first four letters of the system: ㄅ, ㄆ, ㄇ and ㄈ
  - Still popular in Taiwan

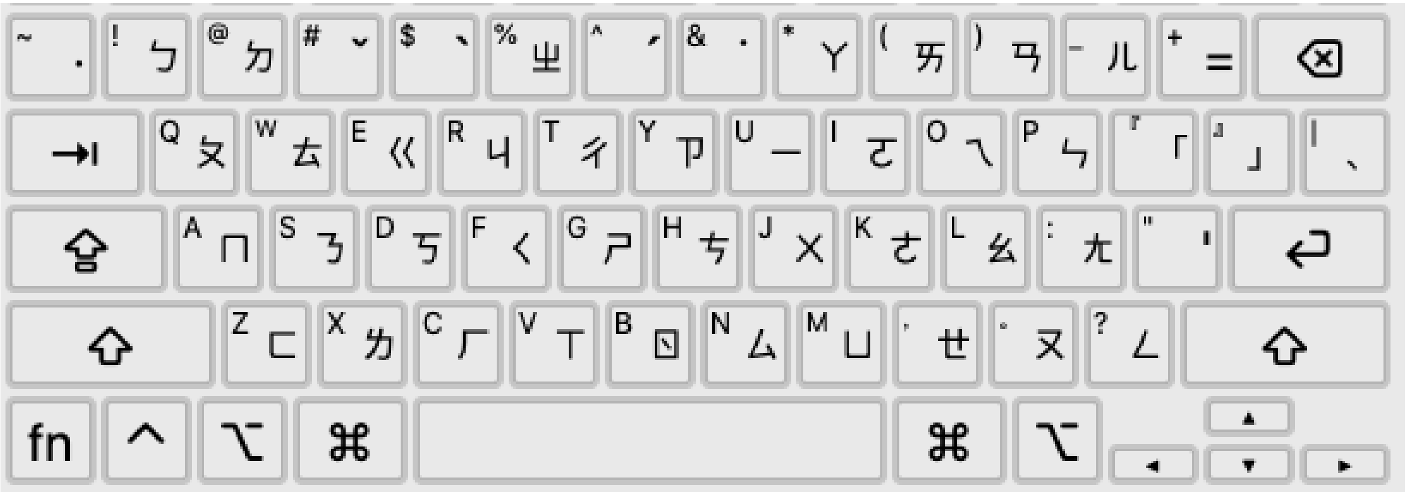
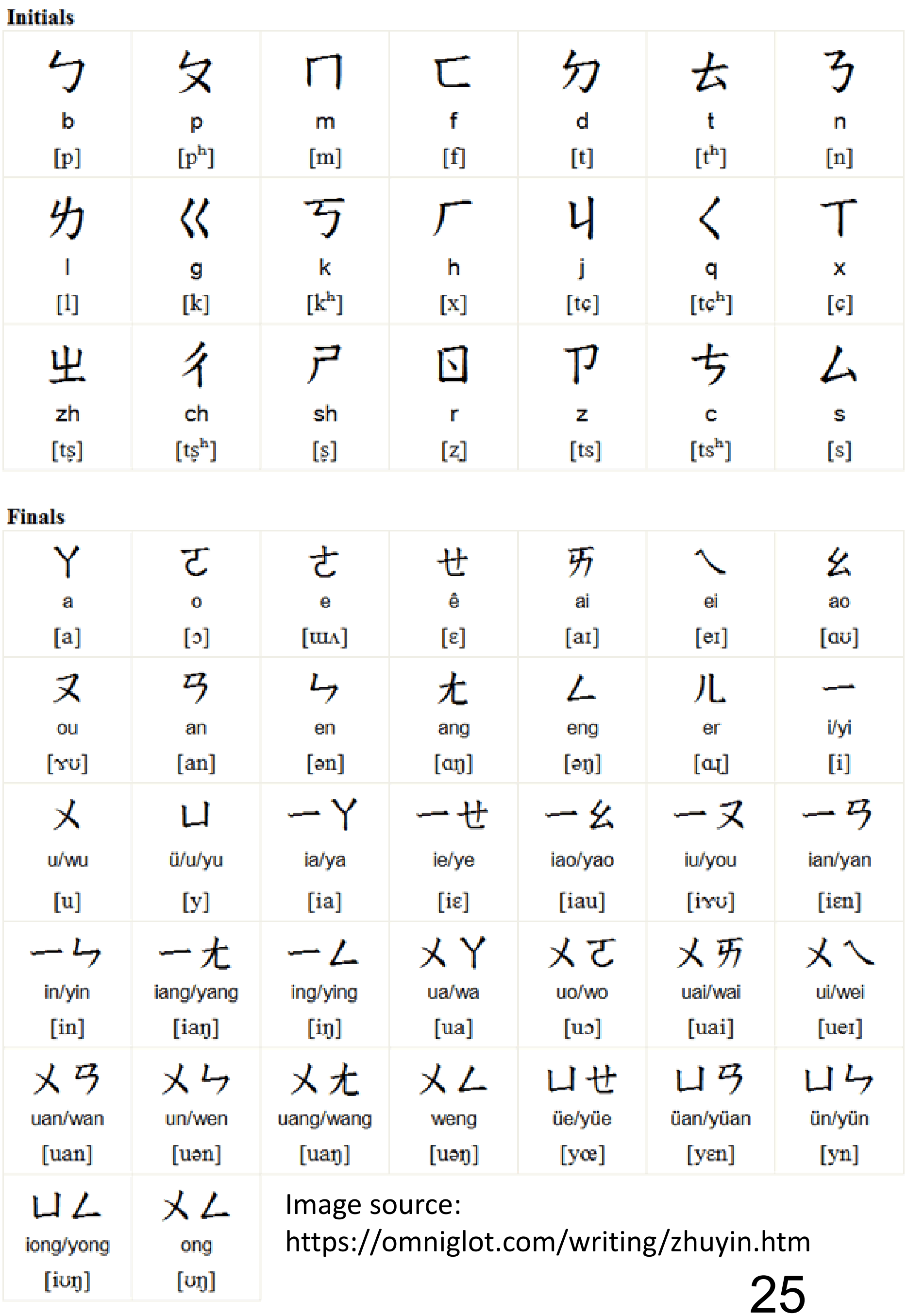

~ ˙ ! ㄅ @ ㄉ # ˇ $ ˋ % ㄓ ^ ˊ & ˙ * ㄚ (ㄞ) ㄢ - ㄦ + =
Q ㄆ W ㄊ E ㄍ R ㄐ T ㄔ Y ㄗ U ㄧ I ㄛ O ㄟ P ㄣ 「 」 、
A ㄇ S ㄋ D ㄎ F ㄑ G ㄕ H ㄘ J ㄨ K ㄜ L ㄠ : ㄤ " '
Z ㄈ X ㄌ C ㄏ V ㄒ B ㄖ N ㄙ M ㄩ , ㄝ . ㄡ ? ㄥ
fn ^ ⌥ ⌘ ⌘ ⌥

Image source: http://xahlee.info/kbd/chinese_input_methods.html



Initials

ㄅ b [p]	ㄆ p [pʰ]	ㄇ m [m]	ㄈ f [f]	ㄉ d [t]	ㄊ t [tʰ]	ㄋ n [n]
ㄌ l [l]	ㄍ g [k]	ㄎ k [kʰ]	ㄏ h [x]	ㄐ j [tɕ]	ㄑ q [tɕʰ]	ㄒ x [ɕ]
ㄓ zh [tʂ]	ㄔ ch [tʂʰ]	ㄕ sh [ʂ]	ㄖ r [ʐ]	ㄗ z [ts]	ㄘ c [tsʰ]	ㄙ s [s]

Finals

ㄚ a [a]	ㄛ o [ɔ]	ㄜ e [ɯʌ]	ㄝ ê [ɛ]	ㄞ ai [aɪ]	ㄟ ei [eɪ]	ㄠ ao [ɑʊ]
ㄡ ou [ɤʊ]	ㄢ an [an]	ㄣ en [ən]	ㄤ ang [ɑŋ]	ㄥ eng [əŋ]	ㄦ er [ɑɻ]	ㄧ i/yi [i]
ㄨ u/wu [u]	ㄩ ü/u/yu [y]	ㄧㄚ ia/ya [ia]	ㄧㄝ ie/ye [iɛ]	ㄧㄠ iao/yao [iau]	ㄧㄡ iu/you [iɤʊ]	ㄧㄢ ian/yan [iɛn]
ㄧㄣ in/yin [in]	ㄧㄤ iang/yang [iaŋ]	ㄧㄥ ing/ying [iŋ]	ㄨㄚ ua/wa [ua]	ㄨㄛ uo/wo [uɔ]	ㄨㄞ uai/wai [uai]	ㄨㄟ ui/wei [ueɪ]
ㄨㄢ uan/wan [uan]	ㄨㄣ un/wen [uən]	ㄨㄤ uang/wang [uaŋ]	ㄨㄥ weng [uəŋ]	ㄩㄝ üe/yüe [yœ]	ㄩㄢ üan/yüan [yɛn]	ㄩㄣ ün/yün [yn]
ㄩㄥ iong/yong [iʊŋ]	ㄨㄥ ong [ʊŋ]					

Image source: https://omniglot.com/writing/zhuyin.htm

2 Chinese phonetic systems

- Hanyu – pinyin （汉语拼音, Chinese language spelling）
  - Developed by Youguang Zhou in 1950s
  - A standard Romanization of Chinese
  - Four tones in Pinyin system **(but tone is not required for Pinyin input)**
  - Currently the official phonetic system in mainland China
  - Compulsory learning material for Chinese students
  - Prevailing input method in China

Initials	b, p, m, f, d, t, n, l, g, k, h, j, q, x, zh, ch, sh, r, z, c, s, y, w
Finals	a, o, e, i, u, ü (v), ai, ei, ui, ao, ou, iu, ie, üe (ve), er, an, en, in, un, ün, ang, eng, ing, ong

běi jīng

北京 Beijing

shàng hǎi

上海 Shanghai

2 Input hanzi – phonetic-based

- Do not need to worry about tones
  - Accents / Dialects
  - Extra operation is needed when homophones occur
  - Homophones decrease as word length increases, e.g. Beijing -> 北京， 背景
- Example of input Chinese characters using phonetic-based methods

Using Bopomofo:

ㄕˊ ㄐㄧㄢ ->

ㄕˊ ㄐㄧㄢ	
時間	time
事件	event
實踐	practice
世間	in the world
飾件	decoration
…	…

Using Pinyin:

shijian ->

shijian	
时间	time
事件	event
实践	practice
世间	in the world
饰件	decoration
…	…

2

Input hanzi – phonetic-based - summary

- Using romaji/hiragana in Japan

romaji -> hiragana -> kanji

ko u shi yo u -> こうしよう

こうしよう	
交渉	negotiation
考証	research
公証	notarization
…	…

- Using hangul in Korea

hangul -> hanja

ㄴ ㅏ ㅁ -> 남

남	
南	south
男	male
藍	blue
…	…

- Using Bopomofo in China:

ㄕˊ ㄐ ㄧ ㄢ ->

ㄕˊ ㄐㄧㄢ	
時間	time
事件	event
實踐	practice
…	…

- Using Pinyin in China:

shijian ->

shijian	
时间	time
事件	event
实践	practice
…	…

2 Shape-based vs. phonetic-based

In Japan and Korea, it is not hard to imagine that phonetic-based methods are dominant for typing in kanji and hanja, since their own language, i.e. hiragana and hangul provide natural tools as phonetic alphabets. How about China?
The table below shows the comparison.

	Shape-based (e.g. wubi)	Phonetic-based (e.g. pinyin)
Advantages	Fewer keystrokes (for single character)	Easy to learn
	Lower collision rate	Easy to use
	Faster	Modern pinyin is super intelligent!
Disadvantages	Need to know very well about hanzi	Higher collision rate because of homophones
	A lot of roots and keys to memorize	Higher possibility of typing error due to more keystrokes

Phonetic-based methods are more popular for inputting hanzi !

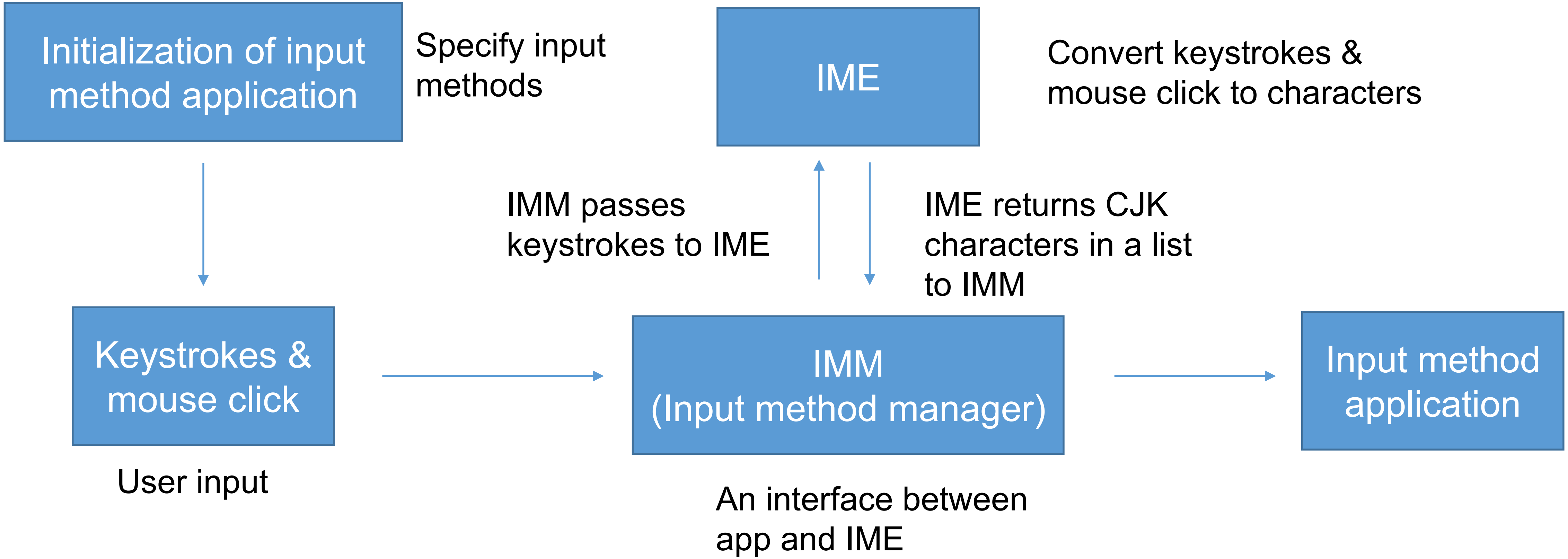
2 More about modern Pinyin input method

- Memorization
    - Remember your choice. Rank homophones based on your input frequency
- Automated input error correction
    - shenme -> 什么 (“what” in English). Type “shenem” will also succeed.
- Input a long sentence by only inputting first consonant letters (initials) of each character
    - Input “wsymyjs” -> 我是一名研究生 (“I am a graduate student” in English). Originally, I have to input complete pinyin of all characters, i.e. “woshiyimingyanjiusheng”.

According to [3], there are more than 97% people in China using Pinyin as the input method.

2 How input method editor (IME) works

- A closer look at the workflow of inputting CJK characters on computers:


Initialization of input method application
Specify input methods
IME
Convert keystrokes & mouse click to characters
IMM passes keystrokes to IME
IME returns CJK characters in a list to IMM
Keystrokes & mouse click
User input
IMM (Input method manager)
An interface between app and IME
Input method application

Note: The output CJK characters are encoded in Unicode on computers. Unicode assign certain ranges for CJK characters.

Conclusions

3

3

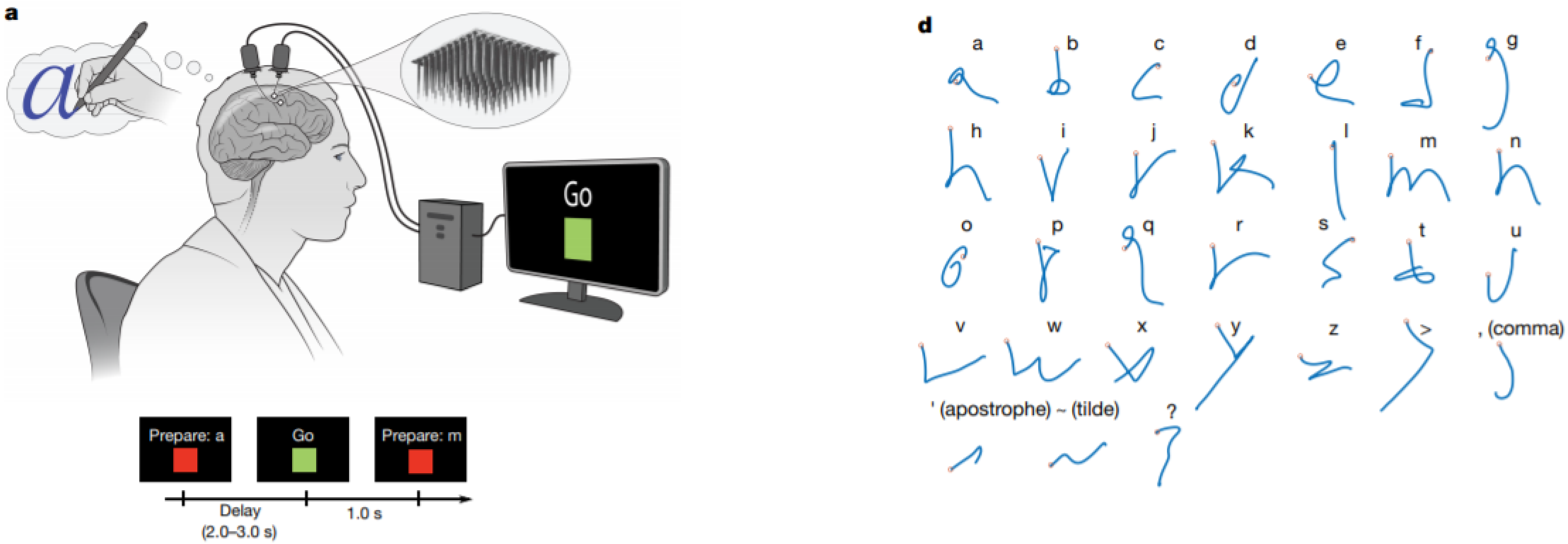
Conclusions

- For alphabetic characters, e.g. hiragana, katakana, hangul, they are efficient to input using current Latin keyboard.
- For ideographic characters, e.g. hanzi, kanji, hanja, phonetic-based input is more popular for its convenience of learning and usage.

3 Future trends

- Optical character recognition: convert images of printed / handwritten text into machine-encoded text
- Voice input: convert human voice into machine-encoded text
- Brain-computer interface (BCI): convert human minds into machine-encoded text?

3 Future trends



A BCI device decodes attempted handwriting movements from neural activity and translates it to text in real time [14]

3 References

1. Becker, Joseph D. "Typing Chinese, Japanese, and Korean." *Computer* 18.01 (1985): 27-34.
2. Guan, Connie Qun, Charles A. Perfetti, and Wanjin Meng. "Writing quality predicts Chinese learning." *Reading and Writing* 28.6 (2015): 763-795.
3. Chen, Zheng, and Kai-Fu Lee. "A new statistical approach to Chinese Pinyin input." *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*. 2000.
4. Huang, Timothy D., and Jack KT Huang. *An Introduction to Chinese, Japanese and Korean Computing*. Vol. 12. World scientific, 1991.
5. Wu, Zimin, and J. D. White. "Computer processing of Chinese characters: An overview of two decades' research and development." *Information Processing & Management* 26.5 (1990): 681-692.
6. Liu, Chao-Lin, and Jen-Hsiang Lin. "Using structural information for identifying similar Chinese characters." *Proceedings of ACL-08: HLT, Short Papers*. 2008.
7. Wang, Jingkang, et al. "Multiple character embeddings for chinese word segmentation." *arXiv preprint arXiv:1808.04963* (2018).

3 References

8. Hu, Chieh-Fang, and Hugh W. Catts. "The role of phonological processing in early reading ability: What we can learn from Chinese." *Scientific Studies of Reading* 2.1 (1998): 55-79.
9. Tham, Yiu-Man, and Tong Lee. "Four corner code based pre-classification scheme for Chinese character recognition." *Proceedings of ICSIPNN'94. International Conference on Speech, Image Processing and Neural Networks*. IEEE, 1994.
10. Yamada, Hisao, and Jiro Tanaka. "A human factors study of input keyboard for Japanese text." *Proceedings of the International Computer Symposium*. 1977.
11. DeFrancis, John. *The Chinese language: Fact and fantasy*. University of Hawaii Press, 1986.
12. Allen, Grant. "The Input Method Framework." *Beginning Android 4*. Apress, 2012. 103-112.
13. Liu, C-L., et al. "Visually and phonologically similar characters in incorrect Chinese words: Analyses, identification, and applications." *ACM Transactions on Asian Language Information Processing (TALIP)* 10.2 (2011): 1-39.
14. Willett, F. R., Avansino, D. T., Hochberg, L. R., Henderson, J. M., & Shenoy, K. V. (2021). High-performance brain-to-text communication via handwriting. *Nature*, *593*(7858), 249-254.

THANKS