

On a phonetic and structural encoding of Chinese characters in Chinese texts

Ch. Boitet & F. X. Tch  ou
GETA, IMAG-campus
(UJF & CNRS)
BP 53X, 38041 Grenoble Cedex, France

(submitted to *ROCLing-III*, Taipei, 19–23 September 1990)

Abstract

A phonetic and structural encoding of Chinese characters is proposed for representing Chinese texts on any computer, using universally available characters (all contained in the PL/I character set), in a readable, yet unambiguous way. More than 80% of the 14 872 characters of the *CIHAI* (      ) dictionary are unambiguously represented with "simple" PS encoding. "Iterated" PS encoding is designed to disambiguate the complete collection of Chinese characters since ancient times (  67 000). PS encoding is geared more toward exchange and processing than storage. Easy to construct with a standard dictionary and a table of keys, this scheme might also be used by non-experts to input Chinese texts while learning the Chinese characters. It may be adapted to the encoding of Chinese characters in Japanese and Korean texts.

Keywords

Text encoding, Chinese characters, phonetico-structural encoding, computer representation of Chinese texts.

Motivations

On computers, Chinese characters are coded internally in several ways. For the most part, 2 bytes are used to represent 1 character (GB2312-80 norm in the PRC, JIS code in Japan, etc.). Unfortunately, these representations are not portable, and utterly unreadable without special equipment, even when the codes are made of pairs of printable ASCII characters. The usual schemes (with the exception of that of Xerox [Becker 1984]) are also limited to a few other character sets. That poses a problem for multilingual texts.

At GETA, we initially hit on this problem while constructing a French-Chinese Machine Translation prototype under the Ariane MT generator. This system runs on IBM machines. At the time, under VMSP/CMS.3, we had access only to normal, 1-byte EBCDIC characters. Even now, although 2-byte characters have been introduced in VMSP/CMS.5, it is impossible to visualize them and to print them on conventional devices.

The question was how to write the dictionaries and generate the output texts? Following the previous work of [Feng Zhi Wei 1981] and [Yang Ping 1981], we used the phonetic pinyin transcription in the output texts (ex: YI), and concatenated to it a disambiguating number in the dictionaries in the (frequent) case of collisions (ex: YI8). However, this approach was not very satisfactory, because the pinyin transcription is so ambiguous that, even for a Chinese who knows it (apparently a rarity), it is very difficult to understand a text written in pinyin. Adding disambiguating numbers in the texts is no help, because a dictionary becomes necessary.

What we need, then, for Chinese as for all languages except a few ones like English, which use roman characters and no diacritics, is a "minimal" transcription, which is *portable* and *readable*. Portability may be achieved by using only universally available characters (roman uppercase characters without diacritics, digits, some special signs such as dollar, hyphen, usual punctuation marks, etc.), while readability requires using the pronunciation as part of the code. As a minimal base, we suggest the PL/I character set, which doesn't even contain lowercase characters. This is an advantage in many countries, where the equipment has been modified to associate signs of the local language to the ASCII or EBCDIC positions of the lowercase roman characters.

PL/I character set [Tucker 1986] :

space	.	(+	&	\$	*)	;	-	/	,	%	?	:	#	@	'	=	"						
A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
0	1	2	3	4	5	6	7	8	9																

It is well known that one Chinese character may have more than one pronunciation (a maximum of 5 in Mandarin Chinese and... about 20 in Japanese — 生 has 17 in [2]). Hence, we must allow a character to take on more than one code. The only property required of the encoding scheme is that a code uniquely determine a character.

I. Simple PS encoding

1. Principle

The longest simple PS code of a character is the string :

$\langle p0 \rangle \langle t0 \rangle - \langle n0 \rangle / \langle p1 \rangle \langle t1 \rangle - \langle n1 \rangle$

$\langle p0 \rangle$, $\langle t0 \rangle$, $\langle n0 \rangle$ are the pinyin, the tone, and the stroke count of the character, respectively, while $\langle p1 \rangle$, $\langle t1 \rangle$, $\langle n1 \rangle$ are those of the character's (semantic) key. A code is *valid* if it disambiguates the character with respect to a given collection of characters (for instance, that of the GB2312-80 norm, that of the CIHAI dictionary, or the complete collection since ancient times).

Examples :	HAO3-6/NU:1-3	好	longest simple PS code
	HAO3-6/N		shortest valid code (for CIHAI)

Pinyin is the official phonetic transcription of Mandarin Chinese. A pinyin is a string of 1 (M) to 6 (CHUANG, SHUANG, ZHUANG) roman letters representing a syllable. We make a small change to this transcription, by replacing Ü with U: . The column (:) is universally available, while the trema (¨) is not.

There are only 420 different pinyin in Mandarin Chinese. In the CIHAI set, we get an average of 35 characters per pinyin, a minimum of 1 (ENG, HM, HNG) and a maximum of 286 (YI).

Each pinyin may be pronounced with up to 5 different tones (high, rising, falling-rising, falling, light). Tones are essential for reading aloud and oral understanding. They are not marked in the official pinyin transcription. However, they are often added, especially for teaching purposes and in dictionaries. When they are indicated, the light tone is not marked, and the four others are marked by diacritics placed above the first vowel of the pinyin (ˊ, ˋ, ˊˊ, ˋˋ). HM, HNG, M, NG are the only pinyin without vowels. HM and HNG have no tone (rather, always the light tone), while M and NG bear the usual tones, marked on M and N.

To code a tone, we use an integer between 1 and 5 (t0) and concatenate it to the pinyin, rather than inserting it after the letter above which the diacritic is usually placed.

We arrive now at a total of roughly 1400 codes, because some pinyin don't take on all of the 5 tones. Some, like ENG, GEM, HM, HNG, NGU, take on only 1 (ENG1, GEM4, HM5, HNG5, NGU2). The majority takes 4. Now, the maximum number of characters per code is still high (154 for YI4). A good way to reduce the ambiguities is to use the number of strokes of the character (n0). With the same example, we get 1 character for YI4-2, YI4-3, YI4-4, YI4-23, an average of 7 for all YI4-n, and a maximum of 14 for YI4-13.

Hence, the elementary coding using p0, t0, n0 is by far not sufficient. We must refine it by decomposing the character : p1, t1, n1, will correspond to the key of the character.

All Chinese dictionaries classify the characters by key, by ascending order of their stroke count. The choice of the keys varies among different authors. Most Chinese dictionaries adopt the traditional set of 214 keys [CIHAI]. Some use a set of 250 keys. In his character dictionary, [Wieger 1972] has chosen a set of 224 keys. More variants may be found, especially in Japanese dictionaries [Nelson 1974]. Although it is sometimes difficult to determine the key of a character, we adopt here the classical set of 214 keys, because these keys are known by most Chinese, who use them to distinguish between homophones in oral conversation ("the HAO of NÜ" = 好 — GOOD).

With that choice, all keys but 10 are also full Chinese characters. All keys, even those 10, have a pronunciation (pinyin + tone). Hence, it is legitimate to extend our codes with p1, t1, n1. The complete list of keys, with pronunciations (178) and alternate graphic forms (246), is given in annex.

Take again the example of YI. In [CIHAI], YI1, YI2, YI3, YI4 correspond to 33, 68, 31, 154 characters, respectively. With simple PS encoding, YI1 and YI3 are fully disambiguated, while 3 ambiguous pairs remain in YI2, and 1 triple and 12 pairs in YI4. Out of 286 characters, 33, less than 12%, need further coding. We haven't yet carried out an exhaustive study on the entire CIHAI dictionary. By sampling, we estimated that 90% or more of these 14 872 characters, which represent more than twice the content of the GB2312-80 norm (6753), are unambiguously representable by simple PS coding.

2. Qualities of simple PS encoding

The simple PS encoding is *readable* by anybody knowing how to pronounce the pinyin. Pinyin is taught to all Chinese in the PRC and to all students of Chinese abroad. It is not necessary to know a single Chinese character to read a text coded in this way.

PS encoding is also *portable* to all computer systems because of the character set used.

Finally, PS encoding can be used to represent Chinese parts of multilingual texts, by simply using a tag such as :HAN, in the SGML fashion, to introduce the Chinese portions (we would have one such tag for each transcription : :RUSSIAN, :ARABIC, etc.).

3. Examples

Taking again YI, we now show the 65 simple PS codes of YI2 (68 characters). For each character, we show the character itself, the key, and the simple code. We also underline the necessary part of a code (smallest disambiguating prefix).

施 <u>ㄟ</u> YI2-5/FANG1-2	施 <u>方</u> YI2-9/FANG1-4	施 <u>酉</u> YI2-10/YOU3-7	羨 <u>羊</u> YI2-13/YANG2-6
台 <u>口</u> YI2-5/KOU2-3	嘆 <u>口</u> YI2-9/KOU3-3	蛇 <u>虫</u> YI2-11/CHONG2-6	蹠 <u>足</u> YI2-13/ZU2-7
色 <u>ㄣ</u> YI2-6/YI3-1	宦 <u>ㄣ</u> YI2-9/MIAN3-3	癩 <u>疒</u> YI2-11/CHUANG2-5	歛 <u>欠</u> YI2-14/QIAN4-4
夷 <u>大</u> YI2-6/DA4-3	拖 <u>木</u> YI2-9/MU4-4	移 <u>禾</u> YI2-11/HE2-5	疑 <u>疒</u> YI2-14/PI3-5
异 <u>艹</u> YI2-6/GONG3-3	姨 <u>女</u> YI2-9/NU:3-3	焉 <u>...</u> YI2-11/HUO3-4	飴 <u>食</u> YI2-14/SHI2-8
圪 <u>土</u> YI2-7/TU3-3	侈 <u>才</u> YI2-9/QIAN3-4	貽 <u>貝</u> YI2-12/BEI4-7	遺 <u>辶</u> YI2-15/CHUO4-3
拖 <u>木</u> YI2-7/MU4-4	侈 <u>才</u> YI2-9/SHOU3-4	滢 <u>氵</u> YI2-12/BING1-2	儀 <u>彳</u> YI2-15/REN2-2
佗 <u>亻</u> YI2-7/REN2-2	漢 <u>氵</u> YI2-9/SHUI3-4	蜨 <u>虫</u> YI2-12/CHONG2-6	蟻 <u>虫</u> YI2-16/CHONG2-6
沂 <u>氵</u> YI2-7/SHUI3-3	怠 <u>心</u> YI2-9/XIN1-4	詒 <u>言</u> YI2-12/YAN2-7	嶷 <u>山</u> YI2-17/SHAN1-3
迤 <u>辶</u> YI2-8/CHUO4-3	馳 <u>貝</u> YI2-10/BEI4-7	詭 <u>言</u> YI2-12/YAN2-7	鯪 <u>魚</u> YI2-17/YU2-11
宜 <u>ㄣ</u> YI2-8/MIAN3-3	羨 <u>艹</u> YI2-10/CAO3-6	羴 <u>羊</u> YI2-12/YANG1-6	鯪 <u>魚</u> YI2-17/YU2-11
標 <u>才</u> YI2-8/QIAN3-3	虜 <u>戶</u> YI2-10/HU4-4	褰 <u>衣</u> YI2-12/YI1-6	移 <u>禾</u> YI2-17/ZHU2-6
僕 <u>亻</u> YI2-8/REN2-2	撲 <u>木</u> YI2-10/MU4-4	拖 <u>木</u> YI2-13/MU4-4	彝 <u>彳</u> YI2-18/JI4-3
沂 <u>氵</u> YI2-8/SHUI3-3	侈 <u>木</u> YI2-10/MU4-4	嬰 <u>女</u> YI2-13/NU:3-3	謬 <u>言</u> YI2-18/YAN2-7
圪 <u>土</u> YI2-8/TU3-3	眈 <u>目</u> YI2-10/MU4-5	馳 <u>馬</u> YI2-13/RI4-4	髀 <u>角</u> YI2-21/JUE2-7
怡 <u>忄</u> YI2-8/XIN1-3	腴 <u>月</u> YI2-10/ROU4-6	滢 <u>氵</u> YI2-13/SHUI3-3	齧 <u>齒</u> YI2-23/CHI3-15
迤 <u>辶</u> YI2-9/CHUO4-3	訖 <u>言</u> YI2-10/YAN2-7	謬 <u>言</u> YI2-13/YAN2-7	議 <u>鳥</u> YI2-24/NIAO3-11

Simple PS codes of the 68 characters of CIHAI pronounced YI2

(The 3 ambiguous pairs have simple PS codes YI2-10/MU4-4, YI2-12/YAN2-7, YI2-17/YU2-11).

II. Iterated PS encoding

1. Incompleteness of the simple encoding

Our reference dictionary is quite large, containing about 3 times the number of characters mastered by University level students, and about 2 times the number of characters of the GB2312-80 norm. We have seen that about 10% of our 14 872 characters are not distinguished by simple PS coding, and that the ambiguous groups are very small (pairs, triples). At that point, we might resort to some arbitrary numbering to distinguish characters having the same simple PS code.

However, the results obtained on our reference dictionary may be somewhat misleading, because the total number of Chinese characters since ancient times is about 67 000 (incidentally, more than can be accommodated by a 2-byte coding scheme). Some authors mention even more. For literary or historical applications, we should be able to consider all of them. Our ambiguity rate using simple encoding will grow considerably. Hence, we should refine our coding in a principled way.

2. A solution



The idea of iterated PS encoding is very simple : extend the simple PS code with analogous strings of the form

/<p2> <t2> - <n2>.../<pk> <tk> - <nk>

where each triple corresponds







to a key appearing in the character. (p1, t1, n1) is associated with the primary (semantic) key, which is not necessary written first when drawing the character. (p2, t2, n2) is associated with the key appearing first in the writing and different from the primary key, (p3, t3, n3) with the second, etc.

Examples of complete iterated PS codes :


 YI2-18/JI4-3/MI3-6/MI4-6/GONG3-3
 YI2-8/REN2-2/DA4-3/GONG1-3

In the second example, we use the key DA4-3 first, because its first stroke is drawn first, although its last stroke is written after the key GONG1-3 is written.

Here are the 3 pairs of YI2, with their complete iterated codes :

 YI2-10/MU4-4/DA4-3/GONG1-3
 YI2-10/MU4-4/XI1-3/XI1-3
 YI2-12/YAN2-7/SI1-2/KOU3-3
 YI2-12/YAN2-7/4/YI3-1
 YI2-17/YU2-11/DA4-3/GONG1-3
 YI2-17/YU2-11/FANG1-2/4


In some rare cases, like the 4th and the 6th above, we have a subcharacter which is not completely decomposable into keys. The solution, then, is to indicate the number of strokes separating 2 keys. Remark that, in some cases, the stroke count may be less than the total of the stroke counts of the keys, because a stroke is common to two keys, as in the following example :

 YI2-15/REN2-2/YANG2-6/SHOU3-4/YI4-4

3. Estimates

Iterating once seems to solve easily all ambiguities remaining in our encoding of the 14 872 characters of the CIHAI dictionary. By "easily", we mean that <p2> is sufficient in the vast majority of cases, <p2> <t2> almost always, and <p2> <t2> - <n2> always.

It is extremely likely that 2 iterations will suffice for almost all known Chinese characters. Ultimately, iterating until the last key is described will almost surely identify uniquely a character, because of the historical genesis of the characters. Note that we may allow omission of the part <pi> <ti> - for any key in the iteration, if that is enough to disambiguate, and obtain shorter codes. With the previous example, we might accept

and even YI2-15/REN2-2/6/4/4  YI2-15/REN2-2/6/8

if we want to consider the complete subcharacter in the lower righthand portion of the character.

III. Other usages and perspectives

1. Text input

As constructed, our codes are names for sets of characters. A code is valid if the set is a singleton. It is quite easy to imagine a system of text input using PS encoding : routine use requires simply typing a prefix of a simple PS code. At each point, the system should show the number of characters selected, and offer them as a menu if this number falls below a specified limit.

Of course, this method will never beat any of the methods developed for fast typing by specialists or by educated Chinese people. However, non-experts might use it quite easily. Other methods may also be used by non-experts for text entry, like schemes relying on some numerical coding of the elementary strokes and encoding the very writing of the character. However, these methods lead to codes which are completely unreadable and which are often as long as far as the number of characters typed is concerned.

2. Character learning

PS encoding links pronunciation and structure. As such, it can be a useful aid in learning how to write Chinese, starting from some knowledge of spoken Chinese and pinyin.

On the screen or on paper, texts could be presented with their codes running parallel to the characters. Note that the computer representation used for storing the texts may be arbitrary, as long as the correspondence with the PS codes is stored somewhere.

Reading a Chinese text on the screen, the student could select portions to be shown with PS codes. Options could include :

- pronunciation only (p0, t0),
- simple PS code,
- shortest valid PS code,
- complete PS code.

For the few characters having more than one pronunciation, the set of codes of each kind should be accessible (pop-up window, on-place rotation,...).

Active learning could also be achieved by entering study texts into a micro, until the student is competent enough to switch to a faster input method.

3. Other languages using the Chinese characters

Japanese uses also Chinese characters, but has no tones. The <ti> - parts should hence be suppressed. The pronunciations should be transcribed in "romaji". For the keys, pronunciations of Chinese origin should be preferred, because they are shorter and less numerous than those of Japanese origin. Finally, the number of kanji to be considered seems not to exceed 8 000. For example, the dictionary by [Nelson 1974] contains only about 5 000 characters, which may occur both alone or in about 70 000 compounds.

The case of Korean is similar to that of Japanese, with the same remark on the set of characters actually used.

Conclusion

PS encoding of Chinese characters in Chinese texts is readable, portable, and fairly easy to learn and to put into practice. For usual characters, the shortest valid codes are shorter than complete PS codes, at least with respect to a collection of about 15 000 characters.

At GETA, we had previously developed similar schemes for encoding texts in other languages (written in roman characters or in other character sets, with or without diacritics and/or case distinctions), such as French, Russian, Greek, Arabic, Thai, etc., using almost the same reduced character set. We call this family of transcriptions *universal*. In the same vein, richer character sets, such as the various ISO national codes (e.g., ISO-025 for French), may be used to encode multilingual texts in *local* transcriptions, which are more natural, but less portable.

The problem of encoding multilingual texts for computer storage, exchange, and processing is crucial at a time where information becomes more international every day. This problem is more acute for the languages written with ideograms like Chinese, Japanese and Korean than for all others. We hope that our work may contribute to a general solution valid for all languages, past or present. As a matter of fact, we believe it is a responsibility of computer scientists not to treat languages in a purely utilitarian way, but to enable other people, possibly interested in rare or ancient languages, to work with them on the computer.

Acknowledgements

Many thanks should go to M. Embar and E. Blanc for helping us to improve earlier drafts of this paper, in form as well as in content. All remaining deficiencies are of course ours.

-0-0-0-0-0-0-0-0-0-

References

- BECKER J. D. (1984). *Le traitement de texte multilingue*. Pour la Science, sept. 1984, 66—76. Transl. from Science.
- CIHAI (1983) 辞海 *Large Dictionary of Chinese Characters and Words* (Title freely transl.). 4th edition, Ci Shu, Shanghai.
- CIYUAN (1984) 辞源 *Comprehensive Dictionary of Chinese Characters and Words* (Title freely transl.). 3rd edition, Shang Wu, Beijing.
- FENG Zhi Wei (1981) *Mémoire pour une tentative de traduction multilingue du chinois en français, anglais, japonais, russe et allemand*. Doc. GETA, Grenoble, 40 p. + annexes.
- HEISIG J. W. (1977) *Remembering the Kanji*. Japan Publications Trading, Tokyo.
- IEEE (1985) *Special Issue on Chinese/Kanji Text and Data Processing*. IEEE Trans. on Computers. Contributions by Chu Y., Huang J. K., Becker J. D., Matsuda R., Makino H., Cui W.
- KURATANI & al (1982) *A New Dictionary of Kanji Usage*. Gakken, Tokyo.
- NELSON A. N. (1974) *The Modern Reader's Japanese-English Character dictionary*. Tuttle, 2nd revised edition.
- ROSE-INNES A. & KOS W. (1980) *Beginners' dictionary of Chinese-Japanese Characters and Compounds*. Meiseisha, Tokyo.
- TUCKER A. (1986) *Programming Languages*. McGraw-Hill.
- WIEGER L. s.j. (1972) *Caractères chinois. Etymologie. Graphies. Lexique*. 8ème édition, Kuangchi Press, Taichung.
- YANG Ping (1981) *Un essai sur la génération du chinois*. Doc. GETA, Grenoble, 30 p. + annexes.

-0-0-0-0-0-0-0-0-0-

Annex : (p,t,n) codes of the 214 keys with their graphic and phonetic variants

八	BA1-2	寸	CUN4-3	井	GONG3-3	角	JUE2-7	正	PI3-5	豕	SHI3-7	夕	XI1-3	又	YIN3-3
白	BAI2-5	大	DA4-3	谷	GU3-7	凵	KAN3-2	片	PIAN4-4	士	SHI4-3	西	XI1-6	用	YONG4-5
勺	BAO1-2	歹	DAI3-4	骨	GU3-10	口	KOU3-3	丿	PIE3-1	氏	SHI4-4	乚	XI4-2	酉	YOU3-7
貝	BEI4-7	隶	DAI4-8	鼓	GU3-13	老	LAO3-6	攴	PU1-4	示	SHI4-5	香	XIANG1-9	又	YOU4-2
鼻	BI2-14	刀	DAO1-2	瓜	GUA1-5	来	LEI3-6	攴	PU1-4	才	SHOU3-3	小	XIAO3-3	魚	YU2-11
匕	BI3-2	丩	DAO1-2	龜	GUI1-16	里	LI3-7	齊	QI2-14	手	SHOU3-4	乚	XIN1-3	羽	YU3-6
比	BI3-4	鼎	DING3-13	龜	GUI3-10	力	LI4-2	气	QI4-4	首	SHOU3-9	心	XIN1-4	雨	YU3-8
采	BIAN4-7	斗	DOU3-4	丨	GUN3-1	立	LI4-5	了	QIAN3-3	攴	SHU1-4	小	XIN1-4	玉	YU4-5
彰	BIAO1-10	豆	DOU4-7	厂	HAN3-2	龍	LONG2-16	犬	QIAN3-4	泰	SHU3-12	辛	XIN1-7	聿	YU4-6
丿	BING1-2	門	DOU4-10	禾	HE2-5	角	LU3-11	欠	QIAN4-4	鼠	SHU3-13	行	XING2-6	曰	YUE1-4
水	BO1-5	而	ER2-6	黑	HEI1-12	鹿	LU4-11	井	QIANG2-4	彳	SHUI3-3	玄	XUAN2-5	月	YUE4-4
卜	BU3-2	耳	ER3-6	虎	HU1-6	馬	MA3-10	青	QING1-8	水	SHUI3-4	穴	XUE2-5	俞	YUE4-17
采	CAI3-8	二	ER4-2	户	HU4-4	麥	MAI4-11	人	REN2-2	冰	SHUI3-5	血	XUE4-6	爪	ZHAO3-4
𠂇	CAO3-4	匚	FANG1-2	黃	HUANG2-12	矛	MAO2-5	儿	REN2-2	人	SI1-2	牙	YA2-4	爪	ZHAO3-4
艸	CAO3-6	方	FANG1-4	火	HUO3-4	毛	MAO4-4	日	RI4-4	糸	SI1-6	西	YA4-6	支	ZHI1-4
長	CHANG2-8	非	FEI1-8	几	JI1-2	門	MEN2-8	肉	ROU2-5	女	SUI1-3	广	YAN2-3	久	ZHI3-3
電	CHANG4-10	飛	FEI1-9	己	JI3-3	龜	MENG3-13	月	ROU4-4	田	TIAN2-5	言	YAN2-7	止	ZHI3-4
車	CHE1-7	風	FENG1-9	丑	JIA4-3	米	MI3-6	肉	ROU4-6	入	TOU2-2	羊	YANG2-6	背	ZHI3-12
中	CHE4-3	岳	FOU3-6	子	JIA4-3	𠂇	MI4-2	入	RU4-2	土	TU3-3	羊	YANG2-6	至	ZHI4-6
臣	CHEN1-6	𠂇	FU4-3	女	JIA4-3	𠂇	MIAN3-3	色	SE4-6	瓦	WA3-5	么	YAO1-3	彳	ZHI4-7
長	CHEN2-7	父	FU4-4	見	JIAN4-7	面	MIANG4-9	山	SHAN1-3	尤	WANG1-3	文	YAO2-4	舟	ZHOUI-6
齒	CHI3-15	阜	FU4-8	𠂇	JIE2-2	𠂇	MIN3-5	彳	SHAN1-3	兀	WANG1-3	頁	YE4-9	竹	ZHU2-6
彳	CHI4-3	干	GAN1-3	己	JIE2-2	木	MU4-4	舌	SHE2-6	王	WANG2-4	一	YI1-1	𠂇	ZHU2-6
赤	CHI4-7	甘	GAN1-5	巾	JIN1-3	目	MU4-5	身	SHEN1-7	冫	WANG3-4	𠂇	YI1-5	、	ZHU3-1
虫	CHONG2-6	高	GAO1-10	斤	JIN1-4	鳥	NIAO3-11	生	SHENG1-5	𠂇	WANG3-5	衣	YI1-6	佳	ZHUI1-8
𠂇	CHUAN1-3	戈	GE1-4	金	JIN1-8	牛	NIU2-4	尸	SHI1-3	𠂇	WANG3-6	乙	YI3-1	子	ZI3-3
𠂇	CHUAN3-6	革	GE2-9	門	JIONG1-2	𠂇	NIU2-4	十	SHI2-2	𠂇	WEI2-3	𠂇	YI3-1	自	ZI4-6
𠂇	CHUANG2-5	𠂇	GE2-10	𠂇	JIU3-9	女	NU3-3	石	SHI2-5	韋	WEI2-9	乚	YI4-3	走	ZOU3-7
之	CHUO4-3	艮	GEN4-6	𠂇	JIU4-6	皮	PI2-5	食	SHI2-8	文	WEN1-4	𠂇	YI4-3	足	ZU2-7
之	CHUO4-4	工	GONG1-3	丩	JUE2-1	疋	PI3-5	食	SHI2-9	无	WU2-4	邑	YI4-7		
彳	CHUO4-7	𠂇	GONG1-3	𠂇	JUE2-1	疋	PI3-5	矢	SHI3-5	毋	WU2-4	音	YIN1-9		

There is a total of 246 graphic variants for the 214 keys, with 178 different pronunciations (there are 36 pairs, 13 triples, and 2 quadruples of keys ambiguous with respect to pronunciation), and 227 different codes (16 pairs and 2 triples are ambiguous). When several keys have the same code, they are almost always graphic variants of one of the canonical 214 keys. REN2-2 is an exception (man standing and man sitting). The canonical keys appear in bold face.