TR86-005

Computational Approaches in the Research of Chinese Computer-Topological and Geometric Descriptions for Chinese Characters

By Keh-Jiann Chen
Associate Research Fellow
Institute of Information Science
Academia Sinica, Taipei,
Taiwan, Rep. of China
Tel:(02)782-2002

ABSTRACT

This paper suggests computational approaches in the research of Chinese computer via utilization of the topological and geometric structure information of Chinese characters. A description language is designed for constructing the characters from radicals and the radicals from strokes. The parsing of the descriptions of a character provides the structure of this character which represented as the control points of composed strokes. Many statistical results about Chinese characters can be easily derived and hence support the research of Chinese computer. Furthermore, it is also capable of generating various fonts stroke by stroke in the writing sequence. The supported research areas include 1. character font generation 2. Chinese input and coding 3. character recognition 4. CAI on character spelling, and 5. data compression.

Key words: Chinese computer, font generation, Chinese input, CAI, computational approach.

1. Introduction

The study of computational approaches in the research of Chinese computer was motivated by the vast amount of effort had been paid in the design of Chinese keyboard and the dot matrices for character fonts manually. With the large capacity of character set, many simple routine jobs become tedious and time consuming. For instances, (1) the dot matrices design, for each Chinese character, it needs a lots of work to layout the dots on a n x m rectangular grids by peoples to mimic the image of this character. (2) Researchers pay much of their time to design the Chinese keyboard in order to optimize the encoding. (3) On the character recognition, people tend to fail in the selection of better features due to lack of statistical data. The above tasks have common characteristics, i.e. they need to know the geometric and topological characteristic of all Chinese characters and then from which some common properties of Chinese characters can be found. The difficulties of finding the important properties of the characters is not caused by the complicate searching procedure but by the large set of search domain. Such kind of task is typically suit for the computational approach. Therefore we designed a description language which is used for describing the geometric and topological structure of Chinese characters in terms of radicals and strokes composition. Although, there were many description languages had been designed for composing radicals into character [1,2]. However, our language is the only one to describe radical structures in terms of strokes.

2. The Description Language

Basically, the Chinese characters are composed from radicals and the radicals are composed from strokes. First, we describe the way of constructing characters from radicals. For instance, the character 枷 is constructed from two radicals 木 and 加 from left to right. If we device an operator ⿰ which means compose two operands from left to right, then we are able to express 枷 as ⿰ (木) (加). However radical 加 , itself is also an composed character as 加 = ⿰ (力) (口). So, 枷 = ⿰ (木) (⿰ (力) (口)). The syntax structure of this description language is as simple as follows

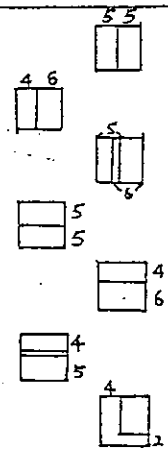
Tree ⟶ operator (Tree) (Tree).
Tree ⟶ radical

There are 3 basic types of operators

1. ⿰ : compose two radicals from left to right.
2. ⿱ : compose two radicals from top to bottom.
3. ⿻ : compose two radicals in overlapping.

Every Chinese character can be described from above three operators and its radicals.

However, for the convenience and the precision, the actual operators, we adopt, are in Figure 1.

Operators with parameters	Pictorial representation	Examples
Ø1		杜 =Ø1(木)(土)
Ø1[4]		唱 =Ø1[4](口)(昌)
Ø1[5,6] *		林 =Ø1[5,5.5](木)(木)
Ø2		早 =Ø2(日)(十)
Ø2[4]		号 =Ø2[4](口)(亏)
Ø2[4,5]		昌 =Ø2[4,5](日)(日)
Ø3[0,2,4,0] **		起 =Ø3[0,2,4,0](走)(巳)

Note :* Every character are supposed to fit into an unit square. The parameters of each operator denote the potion of the space occupied by the operand. The value of parameters is ranged between 0 and 1.

** The parameters of Ø3 denote the boundary of second operand from top down, bottom up, left to right and right to left respectively.

Figure 1. The operators for composing radicals into characters

For instance, the character 榴 has the following tree structure.

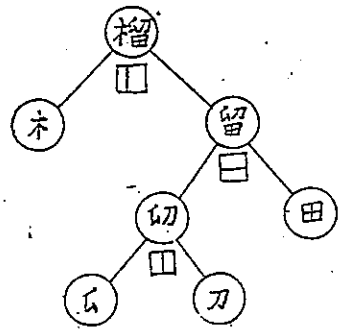


Figure 2. The tree structure of character 榴

The linear description of 榴 is Ø1[4](木) (Ø3 (Ø1 (疒)(刀)) (田)). Actually, each radical is coded as its radical code. Therefore, 榴 is Ø1[4](0423) (Ø3 (Ø1 (0359) (0217)) (0514)). If the area is empty, we have ( ) denotes null radical.

The description of each character will be parsed to find the windows for the radicals in this character. If the window for a radical is smaller than the size of this radical, then this radical will be scaled to the proper size. Otherwise this radical will be shift to the center of the window. Usually the standard size of a radical is a unit square. For the purpose of character fonts generation, the representations of radicals are the geometric location of their respective containing strokes.

Each stroke in a radical is governed by three control points, i.e. head middle and tail point of the stroke. The geometric value of three control points for the strokes in a radical were derived from parsing of radical descriptions.

Compose strokes into radicals

No matters how the description language is designed, it should have the following functions (1) describing the length (x-direction), the height (y-direction), and the rotation angle of strokes, (2) describing the connection relations between strokes.

There are 19 different basic strokes as shown in Fig.3.

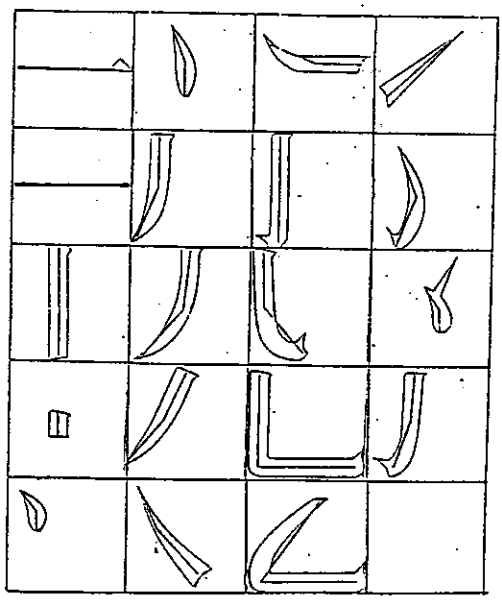
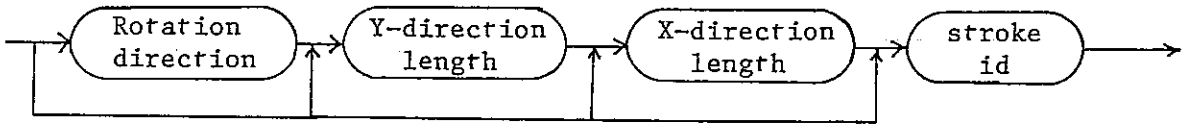


Figure 3. The 19 basic strokes of Shong style characters

However, the variations of each stroke are numerous. Therefore, the variations of a stroke should be able to be described from a standard one by appling length, height, and rotation operators. Every basic stroke has a standard size. For the convenience of character generation, we define the largest normal shape as the standard. The other variations are derived from the scaling and rotating operations.

The syntax diagram of the stroke representation is:



The detail notations for rotation and length as well as stroke identifiers see [3].

The relations between strokes may be connected or nonconnected. For the nonconnected relations, such as '⿰', we use the operators as in the composition of radicals into characters. Therefore, "⿰" is ⿲ (⿲ (丿) (丶)) (⿲ (丶) (丶)). For the connection relations, we devise an binary operator to determine the connected point. The reference points for a stroke are its control points namely H,M,T. For example, H denotes head point. MH denotes the center point between middle and head. We use the bisection operator to determine the exact connected location. Partial examples are shown in Figure 4.

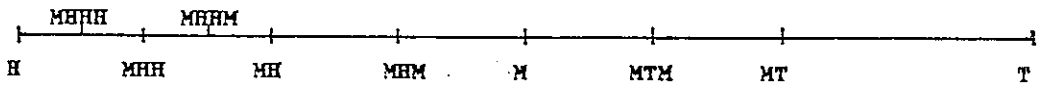


Figure 4. Bisection operators and their relative locations

Now, we are able to describe a complicate radical, for instance '鳥'.

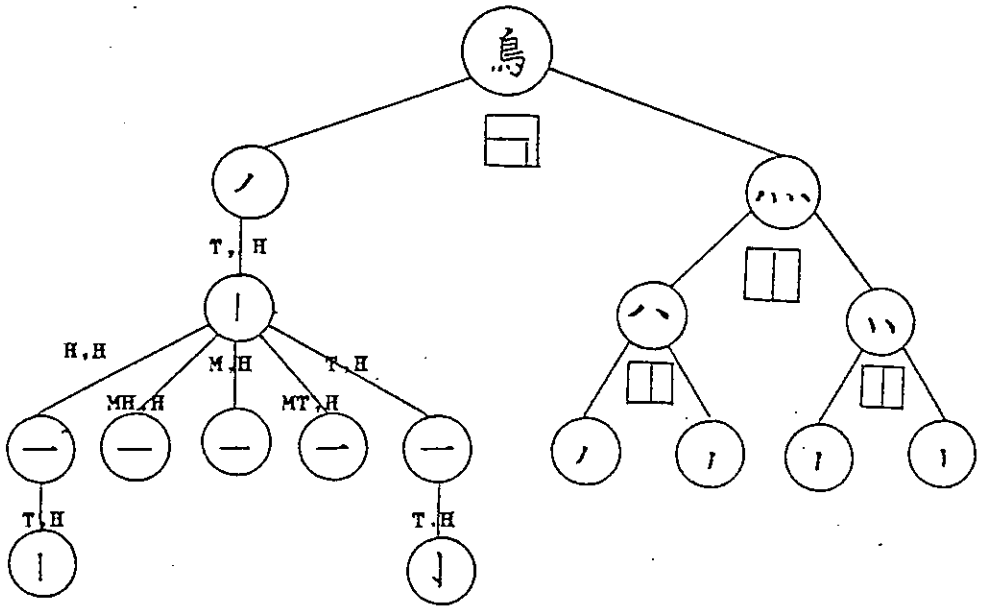


Figure 5. The tree representation of the radical '鳥'

Parsing of the description language

As we state before, the description of a character after parsed will derive the geometric information of each strokes in this character. The geometric information of the strokes is represented by the x,y coordinates of control points. For example, the control points of the character "鳥" is as shown in Figure 6.

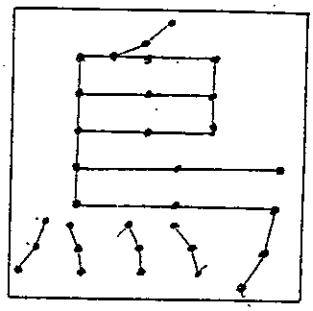


Figure 6. The control points of the character '鳥'

Since, characters are described as a composition of radicals by appling operators in the Figure 1 to form a tree structure. The parser looks the operators from top level down. Recurvely divides the windows into subwindows according to the operator. The top level window will be assigned by the user, usually a unit square. Then, the control points of the radicals are scaled and translated into the appropriated windows.

The parsing of radical description is a bit more complicate. The basic steps are as follows.

1. For individual stroke, get the control points of the stroke by scaling, or rotating the control points of standard strokes according to the operators.

2. If the strokes are connected by the operators, then find the connected points.

3. Each stroke is translated to the connection point according to the coordinates of the first stroke.

4. Find the minimax window of the radical, and shift the radical to the lower left corner of the unit square. If the radical size is greater than the unit; then scale to the unit size. Otherwise, keep it as it is.

After the completion of the parsing steps, each radical is represented by the coordinates of the control points of the contained strokes and the size of this radical. Most of radicals have size equal to unit square; some others may have size less than unit square. For example, the radical "一" has the size nearly zero at Y-direction and about 0.9 at X-direction. The size of radical is for the case when fitting the radicals into windows of a character. If the size of the window is less than the size of the corresponding radical than this radical will be scaled down to the appropriate size; otherwise, the radical will only be shifted to the center of the window. Here, 'the size' we mean the size in both X-direction and Y-direction.

3. Information provided by character descriptions

The descriptions of the Chinese characters provide the topological and geometric structures of each character from strokes to radicals then radicals to characters. Furthermore, we arrange the sequence of radicals and strokes in the description according to the writing sequence. Therefore descriptions of characters contain the following information.

1. The skeleton structure of each character serves the purpose for the generation of character fonts and partially for the character recognition.

2. The radical components of the characters serves the purpose for the studying of Chinese input or coding methods. Also, the representation of characters by radical compositions saves the storage in fonts generation[4] and character generation[5]. Figure 7 shows some sample results of Shong style font.

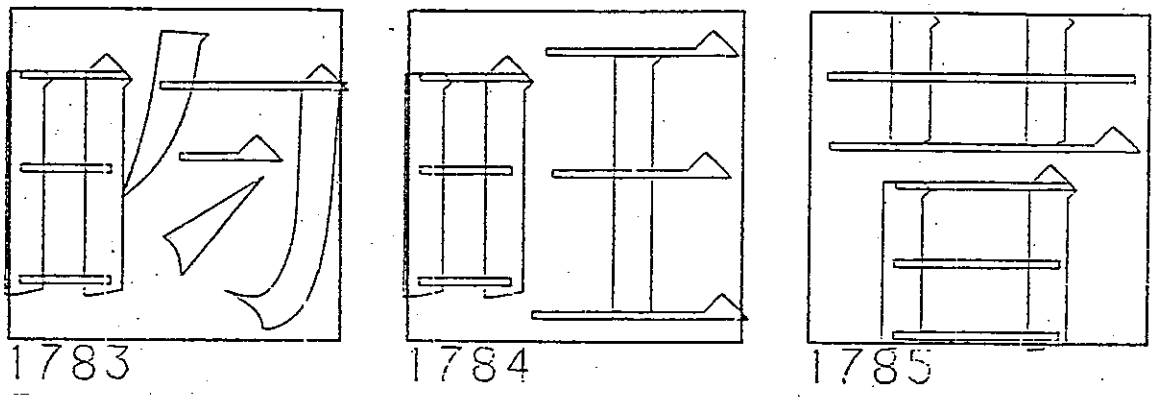


Figure 7. Some sample results of Shong style font.

3. Stroke sequence of each character provides a means of implementing low cost computer-aided instruction for Chinese character spelling [6]. Also, it provides the important sources of stroke coding method for Chinese which is useful in Chinese input and on-line character recognition [7].

The descriptions of character set can also provide many important statistical results about Chinese character such as, radical occurrence frequency, stroke distributions, stroke connection relations, corner shapes. In fact, any statistical result concerning about the structure of Chinese characters may be derived from these descriptions. Thus, they become very useful tool in the computational approaches for the research in Chinese computer.

4. Conclusion

We had completed around 12,000 descriptions for Chinese characters. Such descriptions had been very useful in many researches conduct in our institute. Besides the fonts and dot matrices generation, many statistical results about the structure of Chinese characters had been completed from them. Here, we name a few.

1. The frequency distribution of radicals from 8000 characters, Table 1 shows the first set of 20 radicals which is quite agree with the result of Suen and Huang [8].

2. The character sets of the same stroke sequences, Table 2 shows the result. Here, we group the basic strokes into 7 different stroke types as shown in Table 3.

3. The character sets of the same strokes without sequence restriction, we found 363 groups. Table 4 shows the partial result.

5. References

1. Shi-Kuo Chang, "an interactive system for Chinese character generation and retrieval", IEEE Trans. on Systems, Man, and Cybernetics, Vol.3, 257-265, May 1973.

2. 謝清俊等"中文字根的貯存和中文字的合成"，交大學刊，第六卷第一期，122-131, 1973.

3. 陳克健，鄭國揚，"ACCFONT - 中文字自動產生系統使用者手册"，中央研究院資訊所技術報告 TR-83-005,1983.

4. K.J. Chen, K.Y. Cheng & C.N. Chen, "On Construction and Generation of Chinese Characters", Intelligent System Imaging Technology and Software Engineering, Edit by S.P. Wang, P.99-114, 1984.

5. K.J. Chen & K.Y. Cheng, "A Model for Low Cost Chinese Character Generator", Proceedings of the International Conference of chinese Computing'85, San Francisco, U.S.A., 1985, P.F-1.1-P.F-1.10.

6. K.J. Chen & S.Y. Lin, "Computer Aided Instructions for Chinese Character Spelling with Compressed Coding", Proceedings of ICS'84, Taipei, Taiwan, R.O.C., P.306-310.

7. T. T. Hsieh, On-Line Recognition of Hand-written Chinese Characters, Master Thesis, Department of Electrical Engineering and Technology, National Taiwan Institute of Technology, Taipei, Taiwan, R.O.C.

8. C. Y. Suen and E.-M. Huang, "Computational Analysis of the Structural Compositions of Frequently Used Chinese Characters". Computer Processing of Chinese & Oriental Languages, Vol.1 Number 3 May, 1984, P.163-176.

ROOT NUMBER	ROOT1	ROOT2	ROOT3	ROOT4	ROOT5	ROOT6	ROOT7	ROOT8	ROOT9	TOTAL NUMBER
口	301	314	303	574	336	86	8	2	1	0	1624
一	101	37	192	365	172	103	34	6	2	0	911
氵	328	557	47	0	0	0	0	0	0	0	604
日	419	120	191	166	75	19	5	0	0	0	576
木	423	395	36	8	0	0	0	0	0	0	439
艹	410	295	107	10	0	1	0	0	0	0	413
月	421	153	54	147	43	10	3	1	0	0	411
亻	204	324	72	12	0	0	0	0	0	0	408
土	303	31	106	126	65	19	2	0	0	0	349
扌	361	330	12	0	0	0	0	0	0	0	342
九	205	1	27	127	114	14	26	13	1	0	323
灬	605	223	19	50	22	7	0	1	0	0	322
亠	202	82	193	24	10	3	0	0	0	0	312
田	514	44	92	82	63	22	1	0	0	0	304
口	302	36	86	91	55	8	19	1	0	0	296
讠	702	210	35	15	6	15	1	0	0	0	282
金	801	273	5	1	2	0	0	0	0	0	281
又	229	7	61	102	68	36	3	2	0	0	279
艹	343	185	58	23	7	1	1	0	0	0	275
宀	323	97	167	9	2	0	0	0	0	0	275

Table 1. The first 20 radicals with the highest frequency distribution computed from 8000 characters

1	几 九	19	仞 仂	37	午 牛
2	刀 力	20	加 召	38	作 件
3	土 士 工	21	诏 迦	39	提 捏
4	千 广	22	笆 笳	40	景 暝
5	口 口	23	弔 引	41	茹 茹
6	己 巳 已	24	余 余	42	晏 晚
7	大 犬	25	旰 旱	43	峴 嶌
8	反 木	26	杜 杠	44	細 紬
9	六 亢	27	味 杲	45	叵 言
10	日 曰	28	音 巨	46	丸 弋
11	且 旦	29	沫 沫		
12	北 匕	30	肚 肛		
13	右 石	31	姐 坦		
14	叶 甲 申	32	胄 冒 昀		
15	失 矢	33	唱 唱		
16	田 由	34	毗 昆		
17	由 已	35	棕 柬		
18	另 叨 叻	36	疱 疱		

Table 2. There are 46 groups of Chinese characters (found from 8000 characters) with same spelling sequences

	Strokes
1	一, ㇀
2	㇏, ㇄
3	丶, ㇀, 丶
4	丨, 亅, ㇁
5	丿, ㇓
6	㇇, ㇆, ㇖, ㇆
7	㇄, ㇙, ㇄

Table 3. 7 different stroke types for on-line character recognition

1.	丈又太犬	26	仁午牛	51	倚勑架
2	上土士工	27	仇仉尼	52	倻褫
3	不六宀	28	令夫本	53	颐說
4	丑五正	29	介爪	54	兄四
5	丕半平立	30	仍仿	55	元走
6	世占叶	31	仔竹行	56	充祝
7	丙内	32	伎床	57	入八
8	仞伪	33	仵件	58	冂卩
9	佮囷	34	伊后	59	冒喧
10	丸凡尤	35	佯伻位	60	凡尤
11	主玉	36	住社社	61	囟古
12	乃方	37	何向	62	刀力
13	久夂夭	38	余佘永	63	分夯
14	丑宋	39	佟来米	64	切厄
15	九兀几允	40	佯柈全	65	刋币
16	了勹才	41	併祈	66	刑布
17	互巨	42	例初	67	列兵
18	亠十	43	侮帝	68	到柯固
19	亡口口	44	侶徊	69	加另叨召叻
20	交仗伏	45	促保味	70	助肯
21	亦仞	46	侠狱	71	劼周
22	享伺	47	倌宫	72	匈材
23	京咐	48	倬信	73	匚于
24	亭帝	49	俯础	74	匡羌
25	亮栖	50	倖社	75	匣阜卣吉豆

Table 4. The first 75 groups of characters with same stroke combinations.

Appendix: The sets of Chinese characters with the same stroke combinations

1	丈 叉 太 犬	26	仁 午 牛	51	倚 勅 架
2	上 土 士 工	27	仇 仉 庇	52	僻 褲
3	不 六 亠	28	今 未 本	53	歐 訊
4	丑 五 正	29	介 爪	54	兄 四
5	丕 半 平 立	30	仍 仿	55	先 老
6	世 占 叶	31	仔 竹 行	56	兌 祝
7	丙 內	32	伎 床	57	入 八
8	仞 仂	33	作 件	58	冂 尸
9	佮 囷	34	伊 后	59	冒 喧
10	丸 凡 尤	35	伴 佯 位	60	凡 尤
11	主 玉	36	住 杜 杠	61	凼 古
12	乃 方	37	何 向	62	刀 力
13	久 夕 夭	38	余 佘 永	63	夯 夯
14	乏 米	39	佟 來 采	64	切 厄
15	九 兀 几 尢	40	佯 杵 卑	65	刊 币
16	了 勹 才	41	併 析	66	刑 布
17	互 巨	42	例 初	67	列 舟
18	亠 十	43	侮 帝	68	剄 柯 囿
19	亡 口 口	44	侶 徊	69	加 另 叻 召 叻
20	交 仗 伏	45	促 保 咻	70	助 肯
21	亦 仞	46	俠 秋	71	劼 周
22	享 伺	47	倌 宮	72	匈 材
23	京 咐	48	倬 借	73	匸 干
24	亭 帚	49	俯 姊	74	匠 芫
25	亮 栖	50	倖 衽	75	匣 卓 卣 昔 茸

76	卞斗	101	唾豇	126	圩玎
77	印芍	102	唰枷	127	圬拑
78	厝廉	103	咽莫	128	坐巫
79	厠哨	104	哀衷	129	坭屁
80	叉大	105	哈桌	130	坷拈
81	叩引西	106	員唄	131	垣珥
82	叭八合	107	唰倀	132	埃致
83	叮可弓	108	哺圃	133	培蚱
84	右石	109	哼師	134	堞塔
85	司吁	110	哿弨	135	錯鋙
86	吉吐	111	唐砧匿	136	天夫
87	吟合味東	112	售唯	137	失矢
88	吪囫	113	唾甜	138	夯奇
89	吭祀	114	啐桮粘	139	奚杕
90	呈告佢	115	啻啼	140	奶妨
91	吳茵	116	喃椁	141	如妄呿
92	吸度	117	善辜	142	妊妣
93	吻囫	118	喆罟	143	妓茇
94	呆困杏束亲足	119	喑楫軟	144	妗妹枝
95	咒哂	120	嗃鞘	145	妾姅
96	咂呫草革	121	嗇棺營	146	媧弩
97	呶姻	122	喤睡	147	姪痒
98	呷呻	123	嗡翰	148	娶娞
99	命囿	124	圉害	149	婦婷
100	咎屎	125	在年	150	婪案

151	嫗嬋	176	局昉昜	201	惶惺
152	妞妊	177	居屈	202	愿窩
153	孑手	178	屑朔	203	抹抖
154	孳孳	179	屏矽	204	抨拉拌
155	孌孿	180	峨峩	205	拐招
156	究杌	181	崚崎	206	捏捏
157	守村	182	崐崑	207	捽竚
158	安妤	183	己已已	208	援舲
159	宋林	184	帕帛	209	掉措
160	完罕	185	希府	210	掌第
161	宕柘	186	帙芬	211	支攴
162	宙柚	187	幀景	212	敉笶
163	客格	188	平立半	213	區單
164	宣桓軍	189	幸枉	214	廾广
165	室桱釭	190	店若	215	升斤
166	宰梓	191	庚奔	216	卑昕昇
167	宴晏焯焟	192	彄彈	217	卒杯
168	宵梢	193	得昌	218	整槩
169	寂椒	194	從綀	219	斮蓳
170	寄椅	195	志忐	220	日曰
171	富楀	196	忸忯	221	旦耳
172	寐楩	197	怙忸	222	旭酉
173	寠楼	198	悃棟	223	旰旱車
174	寵櫳	199	悼惜	224	昌盲
175	尖氷	200	惲愾	225	昔茸

226	昉昜	251	桀粧	276	泙泣洋
227	昕昇	252	桎釭	277	泳涂
228	星皇重	253	桓軍	278	涑涎
229	昧是查	254	桄秘	279	渣湜
230	昱音	255	桿章	280	湩湦
231	晏焯焟	256	梏桯結	281	煇煊
232	晰晢	257	梧粘	282	焯焟
233	暄暈暉	258	棄竦蛛	283	父爻
234	暝賠	259	棋粗	284	玖玫辰
235	朝腊閞	260	棗棘	285	異豈
236	未本	261	棠棟	286	玾釨釣
237	朱耒	262	棹唔	287	百臼
238	杆辛	263	棔筥	288	皇重
239	杈杖杕	264	楫軟	289	晢稗
240	杜杠	265	楨睟	290	相貞
241	杲杳	266	楸袂	291	省眇
242	杤杓	267	樟莫	292	眷睽
243	枘柄	268	櫝糟	293	眸項
244	枴桐	269	櫝鞏	294	蓄唔
245	枸的	270	毗毘	295	睞貶
246	某芩	271	汲汰	296	祁祊
247	柑苣苹	272	沫沫	297	祏祐
248	柯圉	273	沾泄	298	棻棨
249	校茭	274	沿洒	299	稈耜
250	株秣	275	況洒	300	醫醉

301	站 蚌	326	苙 苹	351	鄖 隕
302	竝 釺	327	莩 荷	352	鄰 隣
303	竦 蛛	328	荽 茯	353	酲 酷
304	竹 行	329	荇 荐	354	鈴 鉢
305	笞 笳	330	草 革	355	鈕 鉦
306	鈷 辝	331	萸 蒿	356	鍾 鍠
306	糊 稹	332	菜 萊	357	隍 隉
308	糦 豁	333	萱 葷	358	階 鄁
309	紊 紋	334	蕃 祼	359	馭 馱
310	紃 絅	335	薛 藉	360	駕 駋
311	紝 素	336	衿 袜	361	鳩 鳧
312	索 絆	337	評 証	362	麦 俊
313	紬 累 細	338	誄 誅	363	甲 由 申 田
314	綢 鋼	339	誦 飼	364	
315	缶 舌	340	貴 貼 貰	365	
316	羜 舴	341	踔 踣	366	
317	肚 肛	342	軌 軏	367	
318	胃 胄	343	迢 迦	368	
319	胖 胚	344	逞 造	369	
320	脅 脇	345	邢 阱	370	
321	脤 脣	346	邡 防	371	
322	腊 閒	347	部 陪	372	
323	腥 腫	348	郵 陲	373	
324	舒 鈞	349	都 陼	374	
325	芃 芁	350	鄔 隖	375	