28

A)

# On the Automation of Chinese History Literatures

Ching-Chun Hsieh [*]
Zy-Kaan Ding
Yuan-Hua Wang
Chuai-Jung Yeh
Chi-Chou Shu
Shu-Fen Tung
Shi Lin
Fung-Hua Wang
Research Fellow, Inst. of Information Science,
Academia Sinica [*]
Computing Center, Academia Sinica

## Abstract

This paper contains two parts. In the first part, a brief introduction to the full-text processing system of Chinese history literatures will be presented. In the second part, we like to share some experience of processing history literatures with readers. Comments and discussions on the character set, retrieval mechanisms, mark-up language and further applications will be presented.

# Introduction

In order to study the feasibility of using computer as a tool for humanities study, a long-term project namely "Automation of History Literatures" (史籍自動化) has been launched since July 1st, 1984. Under this project, a software package called "Chinese Text Processor" (abbr. CTP) has been developed. It is a full-text data-base based system with controlled vocaburary, free-term, and text content retrieval mechanisms. CTP has three versions. Some highlights and characteristics of these versions are listed in Table 1.

CTP has three subsystems, they are：⑴Creation Module, ⑵Retrieval Module, and ⑶Application Module. The creation module has two major functions. The first one is a data-entry functional module which turns the original text into machine readable form called source text files. Then, the source text files will be processed by a text structurization module, the second functional module, and will produce structurized text files. A functional block diagram of the creation module is shown in Figure 1. In Figure 1, readers can easily find out that the original text after been processed by the structurization module is disassembled into four related files. The text string file keeps the original text string. The illustration file keeps the figures, graphes, equations and notes of the original text. The text structure file is a tree structured logical representation of the text content, ie. chapters, sections, paragraphs etc. And, lastly, the format file stores all type-setting information of the original text, such as paging, character font/size etc.

For mark-up rules used by CTP, please refer to Appendix A. In Appendix B, a BNF representation of the structurization parser is presented. For more information about CTP, please refer to reference [1], [2], [3], and [4].

| items / version | CTP 1.0 | CTP 2.0 | CTP 3.0 |
|---|---|---|---|
| 1. Hardware | Dragon 570(A 16-bit, 8086 based micro) | micro-VAX II with Dragon 570 as terminal | 3B2/3B5/3B15 of AT&T with Dragon 570 as terminal |
| 2. Operating System | CP/M | VMS 4.1 | UNIX 5.0/BINIX |
| 3. language used for development | BASIC (Compiler) | C | C |
| 4. Query form | Manual Driven, Chinese | Manual Driven, English | Manual Driven, English |
| 5. retrieval machemism | controlled vocaburary search through hashing or indexing files | free-term search by pattern matching through text string | ①content search<br>②free-term search by pattern maching through text-structure trees |
| 6. Basic text element handled | page | page | paragraph |
| 7. text structure representation | inverted files | inverted files | trees |
| 8. Proximity function | by volume, by page | by volume, by page | by volume, by page, by sub-tree |
| 9. text tested (No. of characters) | 食貨志(史記至隋)(200K) | 食貨志(史記至隋)(200K) | 食貨志(全)(600K)<br>三民主義講稿(200K)<br>先總統蔣公言論集--文化與教育(150K) |
| 10. No. of users | Single user | from 2 to 4 | from 8 to 48 |
| 11. Typical Response time in Seconds | 120-200 | 40 | 60-90 |
| 12. Date of Issue | March,1985 | March,1986 | Sept. 1986 |

Table 1. The characteristics of different versions of CTP

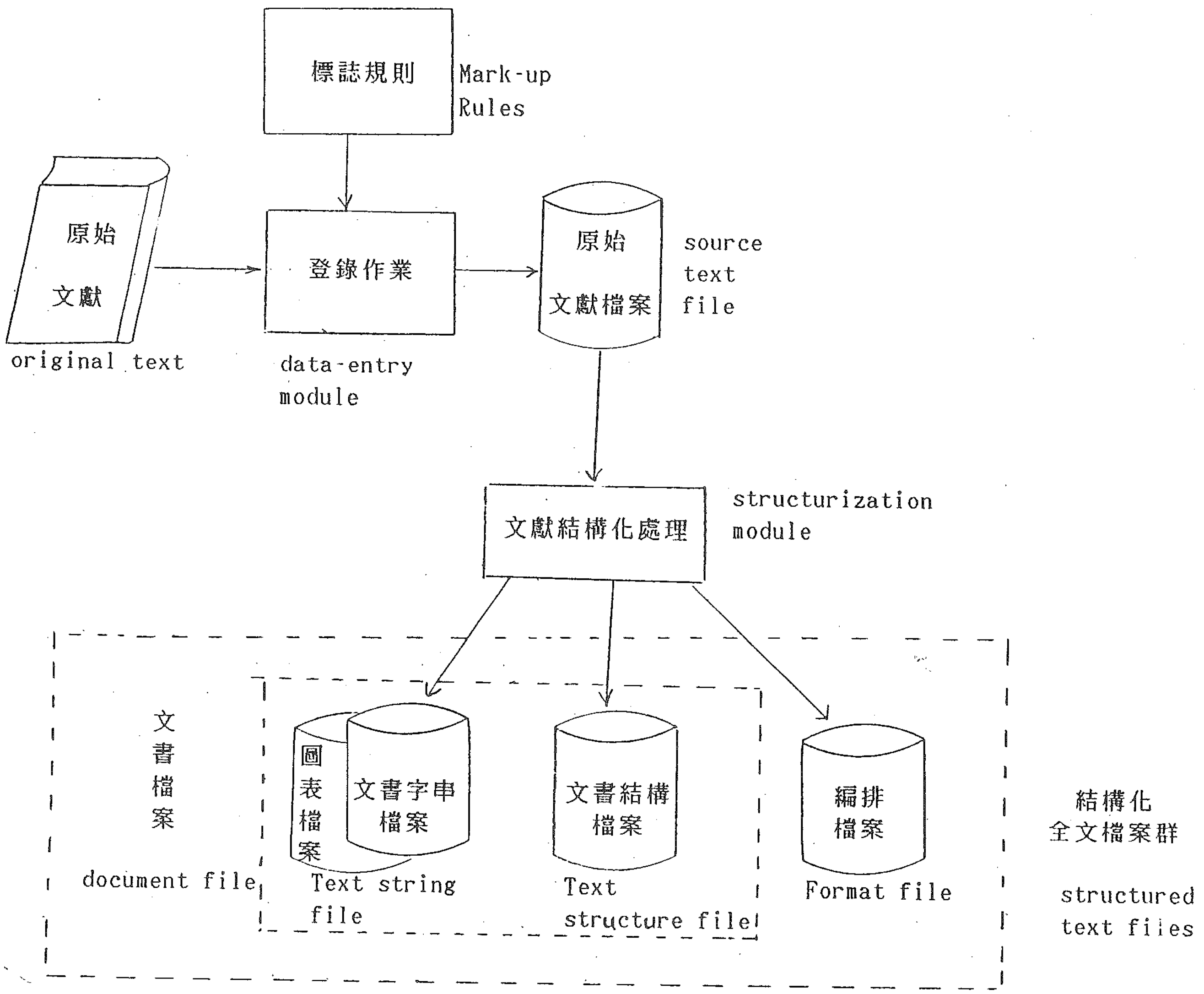


圖 1：全文處理系統之建檔作業部份方塊圖

Figure 1. A block diagram of the creation module

A functional block diagram of the retrieval module and the application module in shown in Figure 2. This part of the CTP can operate independently if structurized text files have already been produced by the creation module. In order to give you some ideas of the operations of the CTP, let me spend a few minutes in showing some slides of CTP for you.

The CTP version 3.0 is now ready to go to the public. In this paper, we like to present some worth-while considerations which we learned during the development of CTP. Your comments and discussions are hearty welcome.

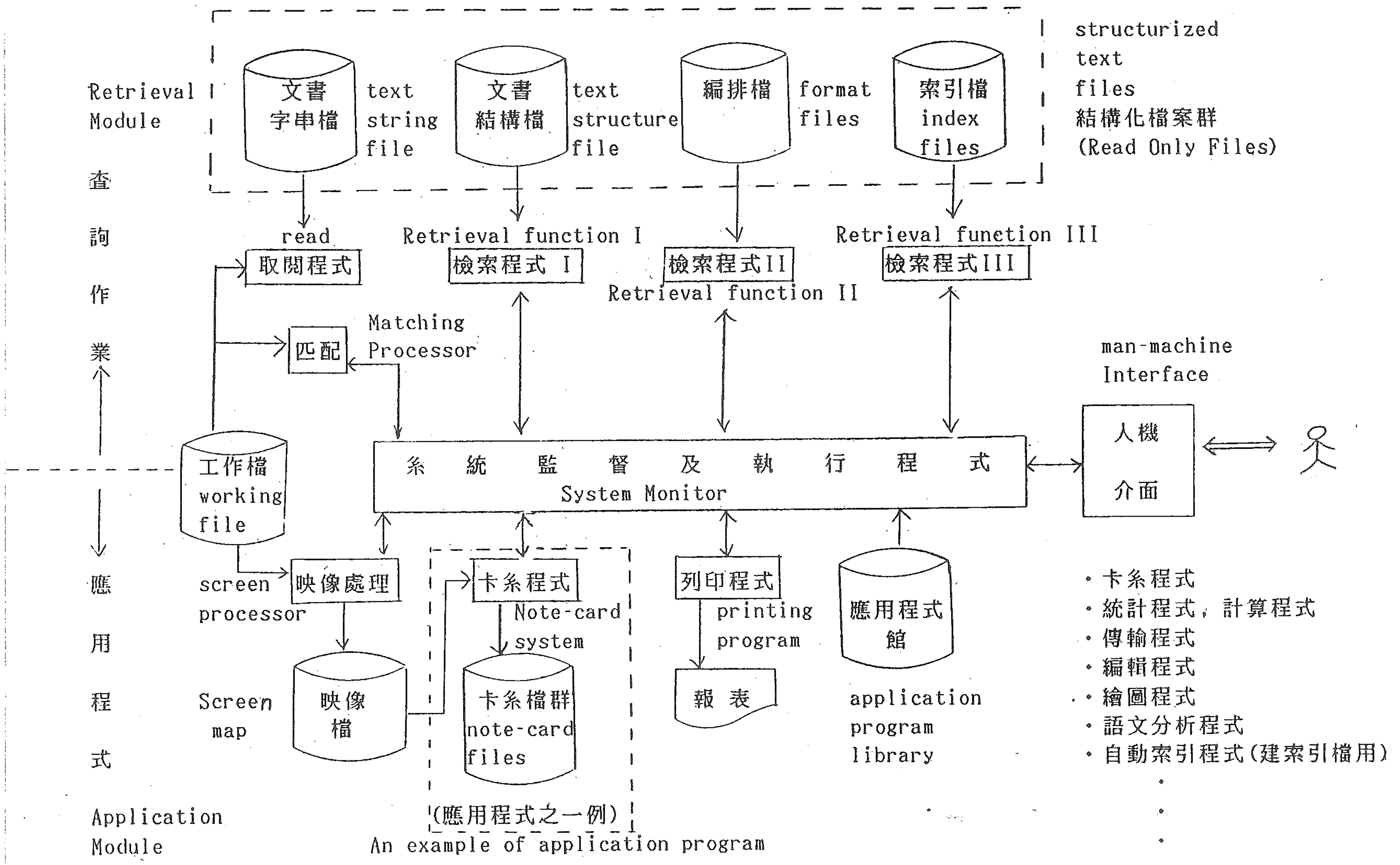
## Comments and Discussions on Processing History Literatures

While presenting history literatures in machine readable form, the information contained in the source text should not be distorted, added, and deleted by any means. But in real life, the above requirement can not be fulfilled with the present technology. Therefore, some efforts must be made to provide somewhat compromised solutions to this problem.

In this part of the paper, we will discuss the problems associated with character set, retrieval mechanisms, mark-up language, and furth applications in the subsequent sections.

### Character set

There are two character sets provided by vendor. One set contains about 13,000 characters which is identical to the set used by 通用漢字交換碼, a proposed standard set for commercial applications. The other set has more than 17,000 characters and it is a super set of the set mentioned above. In the first phase of our project while CTP 1.0 was developed,



圖二　全文處理系統之查詢及應用部份結構方塊圖

Figure 2. A block diagram of the retrieval module and the application module

the set of 13,000 characters was used. During the second phase of CTP 2.0, we shifted to the set of 17,000 characters. We have painful experience with these two character sets, because the coverage and completeness of these sets are obvious not capable enough for processing history literatures.

From Table 2 to Table 6, the reader can easily find that hundreds of characters are missing in these sets for processing history literatures. Therefore, we have to update the system whenever a character is to be added to the existing system. This is surely a time consuming and bothersome operation.

Another serious problem is the presence of variant forms of characters. The character patterns in ancient text can not always accommodate the present-day standard. So, in order not to distort the original text, we are asked to creat new characters whenever a mismatch of character pattern has been found. Even worse is in retrieval operation where both normal form and variant forms have to be used interchangeably. In other words, the capability of processing variant forms of characters is fatally needed for processing history literatures.

According to the facts we collected in Table 2 to Table 6, which is not limited to history literaturos, we conclude that:

(1) The 通用漢字交換碼 is not capable enough for processing literatures.

(2) For processing literatures, CCCII [5] is highly recommended.

Controlled vocaburary search verses free-term search

It is a long argument between controlled vocaburary search and free term search since the full-text data-base became available to the public [6]. To our experience, we like to make the following comments:

| No. | ch. | A | B | No. | ch. | A | B | No. | ch. | A | B | No. | ch. | A | B | No. | ch. | A | B | No. | ch. | A | B | No. | ch. | A | B | No. | ch. | A | B |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 001 | 閷 |  |  | 023 | 嵎 |  |  | 045 | 麪 |  | ✓ | 067 | 厨 |  |  | 089 | 括 |  |  | 111 | 庄 |  | ✓ | 133 | 嶺 |  |  | 155 | 礴 |  |  |
| 002 | 譶 |  |  | 024 | 开 |  |  | 046 | 竈 |  | ✓ | 068 | 亂 |  |  | 090 | 殷 |  |  | 112 | 褱 |  | ✓ | 134 | 狀 |  | ✓ | 156 | 硫 |  | ✓ |
| 003 | 鏁 |  |  | 025 | 謟 |  |  | 047 | 叶 |  | ✓ | 069 | 呪 |  | ✓ | 091 | 冲 |  | ✓ | 113 | 醋 |  | ✓ | 135 | 璕 |  |  | 157 | 磚 |  | ✓ |
| 004 | [illegible] |  |  | 026 | 撨 |  |  | 048 | 擡 |  | ✓ | 070 | 誦 |  | ✓ | 092 | 鏟 |  |  | 114 | 皐 |  | ✓ | 136 | 曼 |  | ✓ | 158 | 醻 |  | ✓ |
| 005 | 碙 |  |  | 027 | 垾 |  |  | 049 | 决 |  | ✓ | 071 | 牀 |  | ✓ | 093 | 敔 |  | ✓ | 115 | 瘕 |  |  | 137 | 玜 |  | ✓ | 159 | 酵 |  |  |
| 006 | 鬻 |  |  | 028 | 情 |  |  | 050 | 勅 |  | ✓ | 072 | 扱 |  | ✓ | 094 | 吳 |  | ✓ | 116 | 譌 |  | ✓ | 138 | 琺 |  | ✓ | 160 | 寓 |  |  |
| 007 | 蕟 |  |  | 029 | 峕 |  |  | 051 | 鹻 |  |  | 073 | 邅 |  |  | 095 | 恭 |  | ✓ | 117 | 鴈 |  | ✓ | 139 | 璒 |  | ✓ | 161 | 恡 |  |  |
| 008 | 扰 |  |  | 030 | 记 |  |  | 052 | 裒 |  | ✓ | 074 | 綺 |  | ✓ | 096 | 搭 |  |  | 118 | 稿 |  |  | 140 | 珐 |  | ✓ | 162 | 蕉 |  |  |
| 009 | 禀 |  | ✓ | 031 | 陀 |  |  | 053 | 𧺆 |  | ✓ | 075 | 儃 |  | ✓ | 097 | 牂 |  |  | 119 | 寳 |  |  | 141 | 璹 |  | ✓ | 163 | 廚 |  |  |
| 010 | 糺 |  |  | 032 | 崇 |  |  | 054 | 搣 |  | ✓ | 076 | 敍 |  |  | 098 | 呣 |  | ✓ | 120 | 瓮 |  |  | 142 | 珮 |  | ✓ | 164 | 朶 |  |  |
| 011 | 責 |  |  | 033 | 离 |  |  | 055 | 榘 |  | ✓ | 077 | 鑛 |  | ✓ | 099 | 壄 |  | ✓ | 121 | 鍮 |  | ✓ | 143 | 球 |  | ✓ | 165 | 杞 |  |  |
| 012 | 専 |  |  | 034 | 吵 |  |  | 056 | 檃 |  | ✓ | 078 | 亘 |  | ✓ | 100 | 币 |  |  | 122 | 澮 |  |  | 144 | 珏 |  | ✓ | 166 | 臬 |  |  |
| 013 | 迵 |  |  | 035 | 輾 |  |  | 057 | 槩 |  |  | 079 | 抵 |  |  | 101 | 堉 |  | ✓ | 123 | 綫 |  | ✓ | 145 | 琇 |  | ✓ | 167 | 詪 |  |  |
| 014 | 隸 |  |  | 036 | 廄 |  |  | 058 | 迴 |  | ✓ | 080 | 穽 |  | ✓ | 102 | 厩 |  | ✓ | 124 | 淮 |  | ✓ | 146 | 琉 |  | ✓ | 168 | 雨 |  |  |
| 015 | 髙 |  |  | 037 | 飧 |  |  | 059 | 熇 |  | ✓ | 081 | 賫 |  | ✓ | 103 | 悳 |  | ✓ | 125 | 瑇 |  | ✓ | 147 | 垩 |  | ✓ | 169 | 堢 |  |  |
| 016 | [illegible] |  |  | 038 | 悮 |  |  | 060 | 腃 |  |  | 082 | 恒 |  |  | 104 | 鼓 |  | ✓ | 126 | 无 |  | ✓ | 148 | 碍 |  | ✓ | 170 | 芝 |  |  |
| 017 | 敓 |  |  | 039 | 臬 |  |  | 061 | 煑 |  |  | 083 | 皓 |  | ✓ | 105 | 鼉 |  | ✓ | 127 | 瓯 |  |  | 149 | 府 |  | ✓ |  |  |  |  |
| 018 | 肥 |  |  | 040 | 污 |  |  | 062 | 艄 |  |  | 084 | 麯 |  |  | 106 | 飤 |  |  | 128 | 瓌 |  | ✓ | 150 | 矶 |  | ✓ |  |  |  |  |
| 019 | 稄 |  |  | 041 | 瓣 |  | ✓ | 063 | 藁 |  |  | 085 | 狷 |  | ✓ | 107 | 阮 |  |  | 129 | 溜 |  |  | 151 | 戬 |  | ✓ |  |  |  |  |
| 020 | 毹 |  |  | 042 | 异 |  |  | 064 | 臯 |  | ✓ | 086 | 怜 |  |  | 108 | 鉸 |  |  | 130 | 廬 |  |  | 152 | 礦 |  | ✓ |  |  |  |  |
| 021 | 觪 |  |  | 043 | 鉄 |  | ✓ | 065 | 餘 |  | ✓ | 087 | 糇 |  | ✓ | 109 | 櫂 |  |  | 131 | 絝 |  | ✓ | 153 | 矾 |  | ✓ |  |  |  |  |
| 022 | [illegible] |  |  | 044 | 粃 |  | ✓ | 066 | 鹽 |  | ✓ | 088 | 葸 |  | ✓ | 110 | 势 |  |  | 132 | 垍 |  |  | 154 | 砝 |  | ✓ |  |  |  |  |

A : The character set of 17,000 characters

B : The character set of 13,000 characters

Table 2. Missing Characters. The listed characters are used in 食貨志, but can not be found in the set of 13,000/17,000 characters

Missing Characters

Table A : 170

B : 89

Table 3. Variant forms.

The characters listed are used in 食貨志 and 史記. Their corresponding variant forms are also shown. But, those variant forms can not be found in the 17,000 set, and tense, they are also missing characters.

Variant forms found from 史記

| No. | Normal | Variant |
|---|---|---|
| 001 | 硇 | 碙 |
| 002 | 隸 | 隷 |
| 003 | 開 | 开 |
| 004 | 模 | 橅 |
| 005 | 起 | 迉 |
| 006 | 弛 | 㢮 |
| 007 | 崇 | 崈 |
| 008 | 廄 | 廏 |
| 009 | 污 | 汚 |
| 010 | 异 | 异 |
| 011 | 煮 | 煑 |
| 012 | 廚 | 厨 |
| 013 | 亂 | 亂 |
| 014 | 氈 | 氊 |
| 015 | 敍 | 敘 |
| 016 | 騑 | 騑 |
| 017 | 璿 | [illegible] |
| 018 | 栝 | 栝 |
| 019 | 廏 | 廐 |
| 020 | 宇 | 寓 |
| 021 | 朵 | 朶 |
| 022 | 杞 | 杞 |

Variant forms found from 食貨志

| No. | Normal | Variant | | | |
|---|---|---|---|---|---|
| 001 | 鰥 | 鱞 | 023 | 莧 | 莧 |
| 002 | 棗 | 棘 | 024 | 遷 | 䙴 |
| 003 | 渚 | 渚 | 025 | 衷 | 衷 |
| 004 | 箸 | 箸 | 026 | 侍 | 侍 |
| 005 | 笄 | 筓 | 027 | 毒 | 毒 |
| 006 | 盡 | 尽 | 028 | 持 | 持 |
| 607 | 咢 | 咢 | | | |
| 008 | 嗢 | 嗢 | | | |
| 009 | 悖 | 恃 | | | |
| 010 | 微 | 微 | | | |
| 011 | 都 | 都 | | | |
| 012 | 斥 | 斥 | | | |
| 013 | 羨 | 羨 | | | |
| 014 | 即 | 卽 | | | |
| 015 | 時 | 時 | | | |
| 016 | 篡 | 篡 | | | |
| 017 | 者 | 者 | | | |
| 018 | 諸 | 諸 | | | |
| 019 | 赧 | 赧 | | | |
| 020 | 屖 | 屖 | | | |
| 021 | 赭 | 赭 | | | |
| 022 | 著 | 著 | | | |

Table 4。Missing Characters

The following characters are used in 漢書/史記, but are not found in the set of 17,000 characters。(This table is not yet completed)

| No. | Ch. | No. | Ch. | No. | Ch. | No. | Ch. | No. | Ch. | No. | Ch. | No. | Ch. | No. | Ch. | No. | Ch. | No. | Ch. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 001 | 冄 | 023 | 槔 | 045 | 皃 | 067 | 紂 | 089 | 懽 | 111 | 僊 | 133 | 灊 | 155 | 陻 | 177 | 擥 | 199 | 丱 |
| 002 | 庑 | 024 | 咼 | 046 | 虵 | 068 | 笲 | 090 | 蚴 | 112 | 逌 | 134 | 覂 | 156 | 帟 | 178 | 𢜔 | 200 | 壥 |
| 003 | 鸖 | 025 | 佸 | 047 | 𣅜 | 069 | 毨 | 091 | 臭 | 113 | 豉 | 135 | 襢 | 157 | 关 | 179 | 獂 | 201 | 洼 |
| 004 | 鄐 | 026 | 叜 | 048 | 調 | 070 | 畼 | 092 | 陷 | 114 | 蘔 | 136 | 敦 | 158 | 屮 | 180 | 慽 | 202 | 皷 |
| 005 | 阠 | 027 | 畾 | 049 | 槿 | 071 | 鶮 | 093 | 莿 | 115 | 斥 | 137 | 遫 | 159 | 掔 | 181 | 鱣 | 203 | 騀 |
| 006 | 詧 | 028 | 隞 | 050 | 叟 | 072 | 傁 | 094 | 巂 | 116 | 羨 | 138 | 蒷 | 160 | 捥 | 182 | 擣 | 204 | 迼 |
| 007 | 鰥 | 029 | 阢 | 051 | 浸 | 073 | 鯢 | 095 | 郊 | 117 | 邞 | 139 | 鍚 | 161 | 截 | 183 | 搜 | 205 | 瑀 |
| 008 | 闚 | 030 | 隠 | 052 | 絰 | 074 | 蘄 | 096 | 骯 | 118 | 异 | 140 | 穲 | 162 | 聃 | 184 | 邃 | 206 | 振 |
| 009 | 芉 | 031 | 郱 | 053 | 濳 | 075 | 額 | 097 | 荓 | 119 | 即 | 141 | 鄹 | 163 | 春 | 185 | 敍 | 207 | 岼 |
| 010 | 禞 | 032 | 欶 | 054 | 抖 | 076 | 摢 | 098 | 鼛 | 120 | 時 | 142 | 煥 | 164 | 劓 | 186 | 籞 | 208 | 崲 |
| 011 | 漬 | 033 | 䍐 | 055 | 鯌 | 077 | 尽 | 099 | 婎 | 121 | 賫 | 143 | 乂 | 165 | 頭 | 187 | 慮 | 209 | 呹 |
| 012 | 陁 | 034 | 霤 | 056 | 鰣 | 078 | 錗 | 100 | 姘 | 122 | 𠮷 | 144 | 翽 | 166 | 醑 | 188 | 絭 | 210 | 髎 |
| 013 | 偵 | 035 | 廇 | 057 | 擐 | 079 | 陖 | 101 | 袣 | 123 | 者 | 145 | 蔽 | 167 | 蚨 | 189 | 黿 | 211 | 嵾 |
| 014 | 痠 | 036 | 繠 | 058 | 淯 | 080 | 僢 | 102 | 鸃 | 124 | 諸 | 146 | 罔 | 168 | 廩 | 190 | 褱 | 212 | 侍 |
| 015 | 駢 | 037 | 恭 | 059 | 箸 | 081 | 揁 | 103 | 踰 | 125 | 赧 | 147 | 爕 | 169 | 傛 | 191 | 訇 | 213 | 歫 |
| 016 | 騚 | 038 | 頷 | 060 | 徐 | 082 | 舥 | 104 | 懋 | 126 | 屖 | 148 | 适 | 170 | 衰 | 192 | 罕 | 214 | 紘 |
| 017 | 騟 | 039 | 鴶 | 061 | 郐 | 083 | 粤 | 105 | 谺 | 127 | 芈 | 149 | 趩 | 171 | 廊 | 193 | 夌 | 215 | 獖 |
| 018 | 偓 | 040 | 雉 | 062 | 鉤 | 084 | 壹 | 106 | 觛 | 128 | 赭 | 150 | 湆 | 172 | 鵬 | 194 | 戴 |  |  |
| 019 | 擇 | 041 | 曎 | 063 | 嚊 | 085 | 峭 | 107 | 陓 | 129 | 看 | 151 | 厥 | 173 | 罦 | 195 | 胻 |  |  |
| 020 | 壍 | 042 | 擲 | 064 | 衝 | 086 | 嵴 | 108 | 恃 | 130 | 奎 | 152 | 髙 | 174 | 褻 | 196 | 鉇 |  |  |
| 021 | 鬴 | 043 | 僸 | 065 | 邢 | 087 | 㬐 | 109 | 㣲 | 131 | 罽 | 153 | 毫 | 175 | 轜 | 197 | 䶂 |  |  |
| 022 | 敲 | 044 | 嶂 | 066 | 僺 | 088 | 俞 | 110 | 都 | 132 | 赦 | 154 | 盈 | 176 | 躝 | 198 | 嵊 |  |  |

Table 5. Missing character(normal form)
The following characters are used in Mandarin Daily News Dictionary but are not found in the set of 17,000 characters

| No. | ch. | No. | ch. | No. | ch. | No. | ch. | No. | ch. | No. | ch. | No. | ch. | No. | ch. | No. | ch. | No. | ch. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 001 | 丄 | 023 | 箎 | 045 | 銥 | 067 | [illegible] | 089 | 揞 | 111 | [illegible] | 133 | 摩 | 155 | [illegible] | 177 | [illegible] | 198 | [illegible] |
| 002 | 𠄞 | 024 | [illegible] | 046 | 鑢 | 068 | 駯 | 090 | [illegible] | 112 | [illegible] | 134 | [illegible] | 156 | [illegible] | 178 | 乚 | 199 | 丨 |
| 003 | [illegible] | 025 | [illegible] | 047 | [illegible] | 069 | [illegible] | 091 | [illegible] | 113 | 姘 | 135 | 乍 | 157 | 錂 | 179 | [illegible] | 200 | 刂 |
| 004 | 攴 | 026 | [illegible] | 048 | 飫 | 070 | 攉 | 092 | 痃 | 114 | 冣 | 136 | 三 | 158 | 鑥 | 180 | [illegible] | 201 | 川 |
| 005 | C | 027 | [illegible] | 049 | [illegible] | 071 | 凂 | 093 | 瘌 | 115 | 唅 | 137 | 嘞 | 159 | 鍁 | 181 | [illegible] | 202 | X |
| 006 | 乛 | 028 | 鮗 | 050 | 骶 | 072 | 蓁 | 094 | 眹 | 116 | 收 | 138 | 亠 | 160 | 鍉 | 182 | 麂 | 203 | 8 |
| 007 | V | 029 | [illegible] | 051 | 骻 | 073 | 鰿 | 095 | 碇 | 117 | 氪 | 139 | 闅 | 161 | [illegible] | 183 | [illegible] | 204 | 扌 |
| 008 | ㄠ | 030 | 蒄 | 052 | [illegible] | 074 | [illegible] | 096 | 駘 | 118 | 咩 | 140 | [illegible] | 162 | [illegible] | 184 | 否 | 205 | [illegible] |
| 009 | 巜 | 031 | 蝲 | 053 | [illegible] | 075 | [illegible] | 097 | 氵 | 119 | 啾 | 141 | 鞋 | 163 | 圖 | 185 | 蒯 | 206 | 覷 |
| 010 | 卩 | 032 | 蜆 | 054 | 譋 | 076 | [illegible] | 098 | 達 | 120 | 爪 | 142 | 鞴 | 164 | [illegible] | 186 | 夊 | 207 | 言 |
| 011 | 巿 | 033 | 蠔 | 055 | [illegible] | 077 | 鼪 | 099 | 三 | 121 | 唣 | 143 | 鬍 | 165 | 卌 | 187 | [illegible] |  |  |
| 012 | 丿 | 034 | [illegible] | 056 | 榝 | 078 | 齌 | 100 | 羓 | 122 | 抝 | 144 | 鰔 | 166 | 謔 | 188 | 孑 |  |  |
| 013 | 亚 | 035 | 鉡 | 057 | 憹 | 079 | [illegible] | 101 | 箄 | 123 | 麼 | 145 | [illegible] | 167 | 蛇 | 189 | 屐 |  |  |
| 014 | ㄒ | 036 | [illegible] | 058 | 臁 | 080 | [illegible] | 102 | 栓 | 124 | [illegible] | 146 | 夂 | 168 | 蛼 | 190 | 惧 |  |  |
| 015 | 瘂 | 037 | [illegible] | 059 | 桧 | 081 | 鐦 | 103 | 棒 | 125 | [illegible] | 147 | [illegible] | 169 | 裛 | 191 | 瑨 |  |  |
| 016 | 癵 | 038 | 鎂 | 060 | 焔 | 082 | 鏂 | 104 | [illegible] | 126 | 碐 | 148 | [illegible] | 170 | [illegible] | 191 | 廿 |  |  |
| 017 | 噔 | 039 | 鋝 | 061 | 嗞 | 083 | 鎮 | 105 | 逾 | 127 | 笵 | 149 | 鮌 | 171 | 鉅 | 192 | 三 |  |  |
| 018 | 睺 | 040 | 鋁 | 062 | 壅 | 084 | 薛 | 106 | 毵 | 128 | 腚 | 150 | [illegible] | 172 | 鉼 | 193 | 刂 |  |  |
| 019 | 瞘 | 041 | 蜳 | 063 | 鉣 | 085 | 辶 | 107 | [illegible] | 129 | 徃 | 151 | [illegible] | 173 | 鉡 | 194 | 扌 |  |  |
| 020 | 䛌 | 042 | 軞 | 064 | 亼 | 086 | 踤 | 108 | 滈 | 130 | 捏 | 152 | 竝 | 174 | 鈧 | 195 | 口三 |  |  |
| 021 | 砇 | 043 | 酶 | 065 | 陟 | 087 | [illegible] | 109 | 皂 | 131 | 栫 | 153 | 漫 | 175 | 酺 | 196 | [illegible] |  |  |
| 022 | 碙 | 044 | 醅 | 066 | 陗 | 088 | 擀 | 110 | 簋 | 132 | 揢 | 154 | [illegible] | 176 | [illegible] | 197 | [illegible] |  |  |

| No. | Normal | Variant form | No. | N | V | No. | N | V | No | N | V | No. | N | V | No. | Normal | Variant form | No. | N | V | No. | N | V | No | N | V | No. | N | V |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 編號 | 正體 | 異體 | 編號 | 正體 | 異體 | 編號 | 正體 | 異體 | 編號 | 正體 | 異體 | 編號 | 正體 | 異體 | 編號 | 正體 | 異體 | 編號 | 正體 | 異體 | 編號 | 正體 | 異體 | 編號 | 正體 | 異體 | 編號 | 正體 | 異體 |
| 001 | 當 | 当 | 022 | 華 | 華 | 043 | 爲 | 為 | 064 | 舉 | 挙 | 084 | 灰 | 灰 | 105 | 駢 | 駢 | 126 | 養 | 养 | 147 | 譽 | 誉 | 169 | 甚 | 甚 | 190 | 喬 | 乔 |
| 002 | 齊 | 斉 | 023 | 藝 | 蓺 | 044 | 為 | 为 | 065 | 擧 | 挙 | 085 | 燴 | 烩 | 106 | 鳳 | 凤 | 127 | 蘿 | 萝 | 148 | 變 | 変 | 170 | 戀 | 恋 | 191 | 研 | 研 |
| 003 | 癱 | 瘫 | 024 | 蟢 | 蟢 | 045 | 殺 | 杀 | 066 | 瓦 | 瓦 | 086 | 燭 | 烛 | 107 | 點 | 点 | 128 | 餜 | 餜 | 149 | 變 | 变 | 171 | 攜 | 携 | 192 | 賭 | 賭 |
| 004 | 競 | 竸 | 025 | 蟄 | 蟄 | 046 | [illegible] | [illegible] | 067 | 俠 | 侠 | 087 | 營 | 営 | 108 | 點 | 奌 | 129 | 曆 | 厤 | 150 | 象 | 象 | 172 | 攜 | 携 | 193 | 辦 | 办 |
| 005 | 樸 | 撲 | 026 | 兀 | 兀 | 047 | 片 | 片 | 068 | 儘 | 侭 | 088 | 燼 | 烬 | 109 | 斷 | 断 | 130 | 廣 | 広 | 151 | 覺 | 覚 | 173 | 攜 | 携 | 194 | 辭 | 辞 |
| 006 | 篩 | 篩 | 027 | 郎 | 郎 | 048 | 狀 | 状 | 069 | 儕 | 儕 | 089 | 直 | 直 | 110 | 於 | 扵 | 131 | 度 | 度 | 152 | 覺 | 覔 | 174 | 攤 | 摊 | 195 | 辭 | 辞 |
| 007 | 箠 | 箠 | 028 | 蠅 | 蝿 | 049 | 獅 | 獅 | 070 | 儔 | 儔 | 090 | 禮 | 礼 | 111 | [illegible] | [illegible] | 132 | 處 | 䖏 | 153 | 覽 | 览 | 175 | 攣 | 挛 | 196 | 巡 | 巡 |
| 008 | 緊 | 緊 | 029 | 頂 | 頂 | 050 | 獨 | 独 | 071 | 劃 | 刬 | 091 | 禱 | 祷 | 112 | 解 | 觧 | 133 | 處 | 處 | 154 | 觀 | 覌 | 176 | 攪 | 搅 | 197 | 過 | 过 |
| 009 | 緊 | 緊 | 030 | 賬 | 賬 | 051 | 獪 | 狯 | 072 | 劍 | 剣 | 092 | 簫 | 箫 | 113 | 簡 | 简 | 134 | 處 | 処 | 155 | 劉 | 刘 | 177 | 攪 | 搅 | 198 | 過 | 過 |
| 010 | 縱 | 縦 | 031 | [illegible] | [illegible] | 052 | 留 | 畱 | 073 | 壯 | 壮 | 093 | 算 | 祘 | 114 | 軀 | 躯 | 135 | 處 | 处 | 156 | 參 | 叅 | 178 | 會 | 会 | 199 | 適 | 适 |
| 011 | 網 | 網 | 032 | 竈 | 竃 | 053 | 留 | 畄 | 074 | 獵 | 獵 | 094 | 籌 | 筹 | 115 | 還 | 還 | 136 | 虛 | 虚 | 157 | 劫 | 刼 | 179 | 略 | 畧 | 200 | 邊 | 边 |
| 012 | [illegible] | [illegible] | 033 | 許 | 許 | 054 | 畝 | 畆 | 075 | 獸 | 獣 | 095 | 籤 | 籖 | 116 | 還 | 还 | 137 | 間 | 间 | 158 | 怨 | 怨 | 180 | 畫 | 画 | 201 | 邊 | 邉 |
| 013 | 緒 | 緒 | 034 | 歐 | 欧 | 055 | 瘵 | 瘵 | 076 | 穩 | 穏 | 096 | 負 | 負 | 117 | 閉 | 闭 | 138 | 陳 | 陈 | 159 | 贖 | 贖 | 181 | 們 | 们 | 202 | 邏 | 逻 |
| 014 | 縣 | 県 | 035 | 歟 | 歟 | 056 | [illegible] | [illegible] | 076 | 撲 | 樸 | 097 | 黽 | 黾 | 118 | 開 | 閞 | 139 | 保 | 保 | 160 | 嘔 | 呕 | 182 | 嬌 | 娇 | 203 | 醫 | 医 |
| 015 | 縕 | 緼 | 036 | 歆 | [illegible] | 057 | 衆 | 眾 | 077 | 擋 | 挡 | 098 | 提 | 提 | 119 | 類 | 类 | 140 | 驕 | 骄 | 161 | 回 | 回 | 183 | 對 | 对 | 204 | 釀 | 醸 |
| 016 | 縰 | 縰 | 037 | 歲 | 歳 | 058 | 衆 | [illegible] | 078 | 曠 | 旷 | 099 | 摳 | 抠 | 120 | 鰛 | 鰛 | 141 | 選 | 选 | 162 | 回 | 囬 | 184 | 彌 | 弥 | 205 | 釋 | 释 |
| 017 | 總 | 総 | 038 | 毆 | 殴 | 059 | 等 | 䒭 | 079 | 樹 | 樹 | 100 | [illegible] | [illegible] | 121 | 齊 | 斉 | 142 | 高 | 髙 | 163 | 夠 | 够 | 185 | 彎 | 弯 | 206 | 醫 | 毉 |
| 018 | 繩 | 縄 | 039 | 歷 | 歴 | 060 | 聃 | 聃 | 080 | 羌 | 羗 | 101 | 喪 | 丧 | 122 | 荊 | 荊 | 143 | 鹽 | 塩 | 164 | 弁 | 弁 | 186 | 惡 | 悪 | 207 | 鑄 | 鋳 |
| 019 | 繼 | 継 | 040 | 歸 | 归 | 061 | 烏 | 烏 | 081 | 攬 | 攬 | 102 | 壓 | 压 | 123 | 荔 | 茘 | 144 | 麗 | 丽 | 165 | 師 | 师 | 187 | 俞 | 兪 | 208 | 匯 | 滙 |
| 020 | 續 | 続 | 041 | [illegible] | [illegible] | 062 | 與 | 与 | 082 | 澁 | 渋 | 103 | 學 | 学 | 124 | 鑑 | 鉴 | 145 | 輕 | 軽 | 167 | 尷 | 尴 | 188 | 萃 | 萃 | 209 | 壽 | 寿 |
| 021 | 氣 | 気 | 042 | 歸 | 帰 | 063 | 舉 | 挙 | 083 | 游 | 游 | 104 | [illegible] | [illegible] | 125 | 餅 | 餠 | 146 | 輕 | 軽 | 168 | 壹 | 弌 | 189 | 蕭 | 萧 | 210 | 臺 | 臺 |

Table 6. Variant forms of characters found in the Mandarin daily News Dictionary and not found in the set of 17,000 characters. (to be continued)

| No.<br>編號 | N<br>正體 | V<br>異體 | No.<br>編號 | N<br>正體 | V<br>異體 | No.<br>編號 | N<br>正體 | V<br>異體 | No.<br>編號 | N<br>正體 | V<br>異體 | No.<br>編號 | N<br>正體 | V<br>異體 | No.<br>編號 | N<br>正體 | V<br>異體 | No.<br>編號 | N<br>正體 | V<br>異體 | No.<br>編號 | N<br>正體 | V<br>異體 | No.<br>編號 | N<br>正體 | V<br>異體 | No.<br>編號 | N<br>正體 | V<br>異體 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 211 | 僆 | 𫢸 | 232 | 稱 | 称 | 253 | 罷 | 罢 | 274 | 春 | 旾 | 295 | 策 | 筞 | 316 | 斆 | 斅 | 337 | 肉 | 月 | 358 | 鄉 | 郷 | 379 | 逕 | 迳 | 400 | 藝 | 蓺 |
| 212 | 鏈 | 链 | 233 | 關 | 关 | 254 | 羅 | 罗 | 275 | 梁 | 梁 | 296 | 管 | 管 | 317 | 盡 | 尽 | 338 | 懸 | 悬 | 359 | 鄙 | 鄙 | 380 | 斷 | 断 | 401 | 蟄 | 蛰 |
| 213 | 喬 | 乔 | 234 | 笑 | 关 | 255 | 羽 | 羽 | 276 | 榮 | 荣 | 297 | 節 | 莭 | 318 | 亂 | 乱 | 339 | 顯 | 㬎 | 360 | 遷 | 迁 | 381 | 操 | 撡 | 402 | 蟅 | 䗪 |
| 214 | 壜 | 壜 | 235 | 竹 | KK | 256 | 翍 | 翄 | 277 | 橋 | 桥 | 298 | 聯 | 联 | 319 | 享 | 亯 | 340 | 喪 | 丧 | 361 | 遷 | 迀 | 382 | 鳧 | 凫 | 403 | 郵 | 邮 |
| 215 | 強 | 强 | 236 | 縷 | 缕 | 257 | 看 | 看 | 278 | 權 | 权 | 299 | 聯 | 聨 | 320 | 以 | 叺 | 341 | 熱 | 热 | 362 | 趙 | 赵 | 383 | 麤 | 麁 | 404 | 鏊 | 鏊 |
| 216 | 噹 | 当 | 237 | 總 | 総 | 258 | 壽 | 夀 | 279 | 涅 | 湼 | 300 | 聯 | 聮 | 321 | 備 | 俻 | 342 | 職 | 职 | 363 | 趨 | 趋 | 384 | 鐵 | 鐡 | 405 | 裉 | 褃 |
| 217 | 甯 | 𡨴 | 238 | 垛 | 垛 | 259 | 耋 | 耋 | 280 | 濤 | 涛 | 301 | 聲 | 声 | 322 | 鑼 | 锣 | 343 | 棗 | 枣 | 364 | 劫 | 刼 | 385 | 餐 | 湌 | 406 | 蟹 | 蠏 |
| 218 | 亭 | 亭 | 239 | 蟻 | 蚁 | 260 | 肯 | 肎 | 281 | 濟 | 済 | 302 | 齶 | 腭 | 323 | 鑾 | 銮 | 344 | 應 | 譍 | 365 | 協 | 协 | 386 | 蘿 | 蘺 | 407 | 蠟 | 蜡 |
| 219 | 亭 | 亭 | 240 | 醉 | 酔 | 261 | 肸 | 肹 | 282 | 柳 | 桺 | 303 | 皻 | 皶 | 324 | 關 | 関 | 345 | 時 | 时 | 366 | 隨 | 随 | 387 | 虎 | 虎 | 408 | 雞 | 鳮 |
| 220 | 實 | 实 | 241 | 閏 | 闰 | 262 | 胼 | 腁 | 283 | 淵 | 渊 | 304 | 蓋 | 盍 | 325 | 關 | 関 | 346 | 時 | 旹 | 367 | 環 | 环 | 388 | 躅 | 躅 | 409 | 飯 | 飰 |
| 221 | 康 | 康 | 242 | 髓 | 髄 | 263 | 胸 | 胷 | 284 | 滾 | 滚 | 305 | 硜 | 硻 | 326 | 隱 | 隠 | 347 | 臨 | 临 | 368 | 傷 | 伤 | 389 | 贓 | 賍 | 410 | 飪 | 餁 |
| 222 | 廉 | 廡 | 243 | 雖 | 虽 | 264 | 問 | 问 | 285 | 囪 | 囱 | 306 | 囟 | 顖 | 327 | 騙 | 騗 | 348 | 還 | 还 | 369 | 墾 | 垦 | 390 | 稹 | 稹 | 411 | 歸 | 歸 |
| 223 | 膾 | 脍 | 244 | 體 | 躰 | 265 | 啟 | 啓 | 286 | 甌 | 瓯 | 307 | 堪 | 堪 | 328 | 騮 | 驑 | 349 | 囟 | 顖 | 370 | 懇 | 恳 | 391 | 矚 | 瞩 | 412 | 毆 | 殴 |
| 224 | 臍 | 脐 | 245 | 體 | 骵 | 266 | 圖 | 啚 | 287 | 發 | 發 | 308 | 場 | 场 | 329 | 騮 | 駵 | 350 | 類 | 類 | 371 | 懵 | 懞 | 392 | 蹩 | 蟞 | 413 | 犛 | 氂 |
| 225 | 蛾 | 䖸 | 246 | 鵝 | 䳘 | 267 | 溫 | 温 | 288 | 發 | 发 | 309 | 塊 | 块 | 330 | 驅 | 駆 | 351 | 顧 | 頋 | 372 | 晝 | 昼 | 393 | 蹴 | 蹵 | 414 | 猶 | 犹 |
| 226 | 謠 | 謡 | 247 | 鶇 | 鶫 | 268 | 啓 | 啔 | 289 | 皞 | 暤 | 310 | 器 | 噐 | 331 | 驕 | 骄 | 352 | 麥 | 麦 | 373 | 暖 | 㬉 | 394 | 瞅 | 矁 | 415 | 疣 | 肬 |
| 227 | 譣 | 譣 | 248 | 殼 | 殻 | 269 | 噁 | 唖 | 290 | 皺 | 皱 | 311 | 擬 | 拟 | 332 | 鬢 | 鬂 | 353 | 賓 | 賔 | 374 | 沿 | 㳂 | 395 | 種 | 秐 | 416 | 儒 | 仔 |
| 228 | 謳 | 讴 | 249 | 答 | 荅 | 270 | 牾 | 啎 | 291 | 矯 | 矫 | 312 | 撟 | 挢 | 333 | 辣 | 辢 | 354 | 講 | 講 | 375 | 垂 | 埀 | 396 | 簡 | 𥳑 | 417 | 稨 | 藊 |
| 229 | 識 | 訳 | 250 | 紝 | 絍 | 271 | 旜 | 旃 | 292 | 硎 | 硎 | 313 | 擺 | 摆 | 334 | 并 | 幷 | 355 | 蘇 | 苏 | 376 | 坎 | 壏 | 397 | 哈 | 般 | 418 | 曄 | 曅 |
| 230 | 覆 | 覄 | 251 | 四 | 四 | 272 | 既 | 旣 | 293 | 窮 | 穷 | 314 | 擴 | 扩 | 335 | 這 | 这 | 356 | 迎 | 迊 | 377 | 嶄 | 嶃 | 398 | 網 | 䋄 | 419 | 蘗 | 櫱 |
| 231 | 譜 | 譜 | 252 | 网 | 冈 | 273 | 明 | 眀 | 294 | 夾 | 夹 | 315 | 數 | 数 | 336 | 應 | 应 | 357 | 陸 | 陆 | 378 | 花 | 苍 | 399 | 縣 | 县 | 420 | 欂 | 榑 |

Table 6. (to be continued)

| No.<br>編號 | N.<br>正體 | V.<br>異體 | No.<br>編號 | N<br>正體 | V<br>異體 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 421 | 氂 | 犛 | 442 | 參 | 叅 | | | | | | | | | | |
| 422 | 蜡 | 䄍 | 443 | 慎 | 昚 | | | | | | | | | | |
| 423 | 捦 | 擒 | 444 | 懼 | 愳 | | | | | | | | | | |
| 424 | 靦 | 腼 | 445 | 曹 | 曺 | | | | | | | | | | |
| 425 | 以 | 㠯 | 446 | 花 | [illegible] | | | | | | | | | | |
| 426 | 竊 | 窃 | 447 | 慎 | [illegible] | | | | | | | | | | |
| 427 | 虧 | 亏 | 448 | 僛 | 欹 | | | | | | | | | | |
| 428 | 蚊 | 蟁 | 449 | 摠 | 捻 | | | | | | | | | | |
| 429 | [illegible] | 鎄 | 450 | 嬛 | [illegible] | | | | | | | | | | |
| 430 | 鏞 | [illegible] | | | | | | | | | | | | | |
| 431 | 廋 | 庋 | | | | | | | | | | | | | |
| 432 | 廟 | 庿 | | | | | | | | | | | | | |
| 433 | 臠 | 脔 | | | | | | | | | | | | | |
| 434 | 蛇 | 虵 | | | | | | | | | | | | | |
| 435 | 蛔 | 蛕 | | | | | | | | | | | | | |
| 436 | 遞 | 遆 | | | | | | | | | | | | | |
| 437 | 遞 | 递 | | | | | | | | | | | | | |
| 438 | 陰 | 阴 | | | | | | | | | | | | | |
| 439 | 豔 | 艳 | | | | | | | | | | | | | |
| 440 | 親 | 亲 | | | | | | | | | | | | | |
| 441 | 刃 | 刄 | | | | | | | | | | | | | |

Table 6.

(1)free-term search is more suitable for history literature processing than the controlled vocaburary search.

(2)In order to improve the system performance, we suggest to establish a controlled vocaburary index by the following:

- Set a counter for each free-term whenever being used for search.
- For a definite period of time, the above counts will be examined. If a term is found to be frequently used, then, it will be considered as a key word for search, and hence, its index shall be established.

In other word, we think a combination of both retrieval mechanisms will optimize the performance of the system. Besides, we suggest to establish the controlled vocaburary search on the foundation of free-term search and the frequency of usage of each terms.

### The needs of standard mark-up language

A very good acticle that explains the concept and needs of a standard mark-up language can be found in the Appendix A of ISO standard DIS 8879 [7] . The mark-up language serves as a bridge that allows data sharing among many application areas such as full-text database, printing industry, word processing, etc. In our application, mark-up language is a necessity for text element recognition.

During the development of CTP, it is sorry to find out that no such a standard mark-up language for Chinese text is available. Therefore, we have to work out a replacement, because we do not have enough resource to develop a standard generalized mark-up language parallel to that of DIS 8879. This replacement, aiming for text element recognition, is the BNF representation of 食貨志 text shown in Appendix B. So far, this abstract BNF model has been successfully used to represent 食貨志, 孫中山先生三民主義講稿, and 蔣中正先生言論集. It seems powerful for certain "regular" structured literatures. However, to design a generalized mark-up language for Chinese text is surely a necessity for furthes development in the Chinese text processing field.

## Further Applications

In a traditional information retrieval environment, where information is organized as formatted record type, the most important function of such a system is retrieval. This situation is no longer true when full-text can be retrieved from a database. Enumerous applications can be developed upon retrieved text, to name a few, such as statistical analysis, computing, graphics, word processing/text generation, etc. Therefore, the development of full-text database and full-text processing technology provide an excellent opportunity to integrate information retrieval systems, text/word processing systems, and computation/analytical systems all together. Besides, many new applications may be created, such as a work-station for humanities study. Therefore, we conclude that full-text processing is a very promising field for us to looking for further development.

# References

1. 毛漢光：<史籍自動化第一年總報告>，中央研究院 歷史語文研究所。 Institute of History & Language, Academia Sinica, Taipei, July 1985

2. 謝清俊等：<中文全文處理系統的設計與製作>，中央研究院 計算中心 Computing Center, Academia Sinica, Taipei, Sept. 1986

3. 謝娟娟：<中文全文處理系統使用手冊>，中央研究院 計算中心 Computing Center, Academia Sinica, Taipei, Sept. 1986

4. C. C. HSIEH, "Full-Text Processing of Chinese Language— —an experimental system for studying Chinese History Literatures" the SEAL sub-committee on Library Technology, AAS. 38th Aunual Conference, Chicago, U.S.A. March, 1986

5. <Chinese Character Code for Information Interchange> vol.1, vol.2 Chinese Character Analysis Group, 行政院文化建設委員會, Taipei, 1980, 1983.

6. Carol Tenopir, "Full-Text databases", <Annual Review of Information Science and Technology>, vol.19, pp215-246, ASIS, Nov. 1984

7. ISO Draft International Standard DIS 8879, <Information Processing—Text and Office Systems—Standard Generalized Markup Language(SGML)>, ISO/TC 97, International Organization for Standardization, Geneva, Switzerland, Oct. 1985.

# Appendix A

# Mark-up rules used by CTP

# 全文標誌規則

## 壹．通則

一、凡需利用本中心發展的「中文全文處理糸統」(CTPS)來處理的文獻，都必須嚴格依照本規則的規定做資料登錄的工作。若否，CTPS將產生錯誤而無法正常工作。

二、依照本規則登錄的文獻中不能有外國文字、字母、及阿拉伯數字。本規則亦不能用之於圖形、表格、公式等。故凡有上列資料之文獻，不宜直接以本規則處理。

三、文獻登錄時以頁為單位，亦即每頁建一檔案，此檔案之名稱為：「xy.src」之格式，其中x為原始文獻之代碼，y為以數字表示之頁碼，而src表示原始檔案之類別，文獻之代碼由糸統工程師設定，不可任意取用。

四、資料登錄時，應盡量依照原稿之版面排列形式。凡有空頁、空行、空白之處均須比照登錄。除在細則中有規定者外，不得任意改變版面之形式。

五、所有文字及符號之鍵入，除細則中另有規定者外，均取全形。

六、關於此標註規則之實務詳如細則。凡對此規則之使用有任何疑問時，請立刻與糸統工程師連絡，不可以隨自己意思選擇處理之方式。

## 貳．細則

一、檔案結構

檔案之第一行應先以阿拉伯數字打入頁次(即通則三中之y)，然後按換行鍵，自第二行輸入正文。

二、正文

⑴若為空白頁，則其正文以"a"表示之。

⑵正文前至少有二個全形空白(bb)。

三、書、卷

⑴書卷起處加 "d"，迄處加 "e"。

⑵書卷之標題起處加 "f"，迄處加 "g"。

⑶標題群之間以一個或一個以上之全形空白隔開。

⑷每個標題之起處加 "l"，迄處不加特別標誌。

四、字形大小

小字起處加 "★"，迄處加 "■"。

五、註釋及校勘

⑴註釋起處加 "h"，迄處加 "i"。

⑵校勘起處加 "j"，迄處加 "k"。

⑶註釋和校勘同時出現處，校勘出處加 "*"。

# Appendix B
# BNF representation of the structurization module

## 食貨志的制式ＢＮＦ表示法

(1)meta symbols of the BNF notation:

```
   < >  ::=  |   { }*   { }n   [ ]   " "
```

(2)constructs of the BNF notation:

```
   <…><…>        means sequence expression concatentation
   ::=           means "is defined as"
   |             means "or"
   {…}*          means "repeatition option times"
   {…}n          means "repeatition n times" n great than 0.
   […]           means "optional elements"
   "…"           means "terminal symbols"
```

(3)邏輯結構(Logical Structure):

```
    <卷      > ::= <標頭> { [<標題>] <大段> }* [<校勘記>]

    <標    頭> ::= "d" <卷頭> { [<附卷頭>] }* "e" "¥n"

    <卷    頭> ::= <卷別> "¥n"

    <卷    別> ::= {<詞>}*

    <附 卷 頭> ::= {<詞>}* "¥n"

    <標    題> ::= "f" "¥n" { {<全形空白>}* <標題群> "¥n" }* "g" "¥n"

    <標 題 群> ::= <標題詞> [ { {<全行空白>}* <標題詞> }* ]

    <標 題 詞> ::= <詞> ["★" { <子標題> }* "■"]

    <子 標 題> ::= {<詞>}*

    <大    段> ::= <段> [<註釋群>] | <前敘正文> {<引言內容>}* <後敘正文>
                  " l " {<大段>}*  | <統計條列句子>

    <段      > ::= {<全形空白>}2 <正文直到換行後接兩個全形空白以前為止>

    <正文直到換行後接兩個全形空白以前為止> ::= { <句子> }*

    <前敘正文> ::= [ {<句子>}* ] { <中文字> }* "：" [ "〔" <勘碼> "〕" ] "¥n"

    <引言內容> ::= [ {<全形空白>}0 { <中文字> }* "¥n" ]
                  {<全形空白>}4 { <中文字> }* "¥n"
                  { {<全形空白>}2 { <中文字> }* "¥n" }*

    <後敘正文> ::= {<全形空白>}0 {<句子>}*
```

```
<統計條列句子> ::= {"¥n"}* { {<中文字>}* "¥n" }*
<註釋群    > ::= "h" "¥n" {<註釋>}* "i" "¥n"
<註      釋> ::= {<全形空白>}1 "〔" <註碼> "〕" <註釋正文>
<註釋正文> ::= {<中文字>}* "¥n" [ { {<全形空白>}5 {<中文字>}* "¥n" }* ]
<註        碼> ::= "０" | "::" | "二" | "三" | "::" | "五" | "六" |
                 "七" | "八" | "九"
<校勘記    > ::= "j" "¥n" "校勘記" "¥n" {<校勘>}* "k" "¥n"
<校      勘> ::= "〔" <勘碼> "〕" <校勘正文>
<校勘正文> ::= {<中文字>}* "¥n" [ { {<全形空白>}4 {<中文字>}* "¥n" }* ]
<勘        碼> ::= "０" | "::" | "二" | "三" | "::" | "五" | "六" |
                 "七" | "八" | "九"
<詞          > ::= {<中文字>}*
<句      子> ::= {<中文字>}* <::句的標號>
<全形空白> ::= "　"
< 中文字 > ::= <宏碁公司出::之字形產生器DPC-24/60 之字集> |
               <新造字> | <教::部發佈的標點符號>
```