Chinese information processing

Yucheng Liu

University of Nevada, Las Vegas

Follow this and additional works at: https://digitalscholarship.unlv.edu/rtds

Repository Citation
https://digitalscholarship.unlv.edu/rtds/544

This Thesis is brought to you for free and open access by Digital Scholarship@UNLV. It has been accepted for inclusion in UNLV Retrospective Theses & Dissertations by an authorized administrator of Digital Scholarship@UNLV. For more information, please contact digitalscholarship@unlv.edu.
INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI
A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
313/761-4700  800/521-0500
Chinese Information Processing

by

Yucheng Liu

A thesis submitted in partial fulfillment
of the requirements for the degree of

Master of Science

in

Computer Science

Department of Computer Science
University of Nevada, Las Vegas
December 1995
The thesis of Yucheng Liu for the degree of Master of Science in Computer Science is approved.

Chairperson, Junichi Kanai, Ph.D.

Examine Committee Member, Thomas A. Nartker, Ph.D.

Examine Committee Member, Kia Makki, Ph.D.

Graduate Faculty Representative, Xin Li, Ph.D.

Interim Dean of the Graduate College, Cheryl L. Bowles, Ed.D.

University of Nevada, Las Vegas
December 1995
ABSTRACT

A survey of the field of Chinese information processing is provided. It covers the following areas: the Chinese writing system, several popular Chinese encoding schemes and code conversions, Chinese keyboard entry methods, Chinese fonts, Chinese operating systems, basic Chinese computing techniques and applications.
# Contents

1 Overview of Chinese Information Processing .................................................. 1
   1.1 Introduction .......................................................................................... 1
   1.2 Background of Chinese Language ......................................................... 3
   1.3 Basic Concepts and Terminology ............................................................ 3
       1.3.1 Software Related Concepts ......................................................... 4
       1.3.2 Chinese Computing Based Concepts ............................................. 4
       1.3.3 Encoding Standards ................................................................. 5

2 The Chinese Writing System .............................................................................. 7
   2.1 Roman Characters ................................................................................ 7
   2.2 Symbols and punctuation ......................................................................... 7
   2.3 Hanzi Character .................................................................................... 8
       2.3.1 The Structure of Hanzi .............................................................. 8
       2.3.2 Pronunciation ........................................................................... 12
   2.4 Chinese Typefaces, Type Styles and Type Sizes ...................................... 17
       2.4.1 Typefaces ............................................................................... 17
       2.4.2 Type Styles ............................................................................ 19
       2.4.3 Type Sizes ............................................................................ 19
   2.5 Chinese Text Style ................................................................................. 20

3 Chinese Character Set Standards and Encoding Methods ............................... 21
   3.1 Chinese Character Set Standards ............................................................ 21
   3.2 Chinese Encoding Methods ...................................................................... 23
       3.2.1 GB 2312-80 Encoding .............................................................. 23
       3.2.2 HZ Encoding ........................................................................ 26
       3.2.3 Big-5 Encoding ...................................................................... 27
       3.2.4 CNS 11643 Encoding ............................................................... 28
       3.2.5 CCCII .................................................................................... 30
       3.2.6 International Encoding Methods ............................................... 32

4 Chinese Input .................................................................................................. 36
   4.1 Big Keyboard input Method .................................................................... 36
   4.2 Small Keyboard Input Method ................................................................ 38
       4.2.1 Input by Pronunciation .............................................................. 39
       4.2.2 Input by Structure ................................................................... 40
       4.2.3 Input by Encoding Value .......................................................... 43
       4.2.4 Input by Other Criteria ............................................................. 44
   4.3 Optical Character Recognition ............................................................... 44
       4.3.1 Online/Offline Handwritten and Handprinted CCR ................... 45
       4.3.2 Printed Chinese Character Recognition .................................... 46
### Table of Contents

4.4 Audio Input Method ........................................ 47
4.5 Chinese Character Dictionaries ...................... 47

5 Chinese Output .............................................. 49
5.1 Chinese Fonts .............................................. 49
5.1.1 Bitmap Font .............................................. 50
5.1.2 Vector Fonts .............................................. 51
5.1.3 Outline Fonts .............................................. 51
5.1.4 Evaluation .................................................. 52
5.2 Font Generation Methods ................................. 52
5.2.1 Vector Pattern .............................................. 52
5.2.2 Dot Matrix Pattern ....................................... 54
5.2.3 Evaluation .................................................. 55
5.3 Font Storage .................................................. 55
5.4 Printer Output ................................................ 56
5.5 Screen Output ................................................ 57
5.6 Chinese Terminal ............................................ 57

6 Chinese Information Processing Techniques ............ 58
6.1 Chinese Operating Systems ............................... 58
6.1.1 Chinese Operating Systems on MS-DOS ............ 59
6.1.2 Chinese Operating Systems on MS-Windows ....... 60
6.1.3 Chinese Operating Systems on Unix ................. 60
6.1.4 Chinese Operating Systems on Macintosh ......... 62
6.2 Code Conversion ............................................. 62
6.2.1 GB ↔ HZ Conversion ..................................... 62
6.2.2 GB ↔ Big-5 Conversion .................................. 63
6.2.3 GB/Big-5 ↔ Unicode ...................................... 63
6.2.4 Input code → Internal code ............................. 63
6.3 Chinese Code Detection ..................................... 64
6.4 Character Search ............................................ 65
6.4.1 Search in font .............................................. 65
6.4.2 Search in a file ............................................. 65
6.5 Character Sorting ............................................ 66
6.6 Memory Management ....................................... 66
6.6.1 Memory Management for Dictionary ................. 66
6.6.2 Font Buffer ................................................ 67
6.7 Compression Techniques ................................. 68

7 Chinese Applications ......................................... 69
7.1 Chinese text processing systems ....................... 69
7.1.1 General Chinese text processing system .......... 69
7.1.2 Chinese Typesetting/Editing system ............... 70
7.2 Chinese programming languages ....................... 70
7.3 Chinese on Internet ......................................... 71
7.3.1 Using Chinese E-mail and News ..................... 71
7.4 Sending Chinese FAX ........................................ 74
7.5 Machine Translation ......................................... 74
7.6 Processing Chinese History and Classical literature 74
7.7 Chinese Library Information Network .................. 75

8 Summary ...................................................... 76
List of Tables

1.1 A Simple Chinese Sentence ......................................................... 3
2.1 Roman Characters in Chinese Text ............................................. 7
2.2 Special Symbols in Chinese Text ............................................... 8
2.3 Statistics of Chinese Character Structure .................................. 11
2.4 Chinese Type Styles ................................................................. 19
2.5 Chinese Character Sizes .......................................................... 20
3.1 Content of GB 2312-80 ............................................................. 24
3.2 Content of Big-5 ...................................................................... 27
3.3 Content of CNS 11643 .............................................................. 29
3.4 Content of CCCII's 16 Layers .................................................. 31
3.5 Content of CCCII's First Layer ............................................... 32
3.6 Basic Multilingual Plane .......................................................... 33
3.7 Content of ISO 10646/Unicode ............................................... 34
3.8 ISO 2022 Standard for GB Set ................................................ 35
4.1 Shuang Pin Keyboard Mapping ................................................ 39
4.2 Zhu Yin Keyboard Mapping ..................................................... 40
6.1 Multi-mapping from GB to Big-5 .............................................. 63
6.2 Incorrect Match Made by ‘grep’ ............................................... 66
List of Figures

2.1 The Stroke Structure of Hanzi character ‘sky’ ........................................ 8
2.2 Five Basic Strokes ............................................................................... 9
2.3 34 Complex Strokes ............................................................................... 9
2.4 The Tree Structure of the Hanzi Character “bay” ................................ 10
2.5 Traditional, Simplified and Over-simplified Characters ...................... 12
2.6 Orthographic and Variant Characters ................................................ 12
2.7 Zhu Yin Symbol Table ......................................................................... 13
2.8 Homonymic Characters ......................................................................... 16
2.9 Polyphonic Characters ......................................................................... 17
2.10 Four Common Type Faces ................................................................... 18
2.11 Five Other Type Faces ......................................................................... 18

3.1 Three 7-bit Byte Extension ................................................................... 31

4.1 Classification of Chinese Input Methods ............................................. 37
4.2 Wu Bi Keyboard Layout ........................................................................ 41
4.3 Cang Jie Keyboard Layout .................................................................... 42
4.4 Fundamental Symbols of 3-Corner Coding Method ............................ 43
4.5 Basic Steps in On-line Handwritten CCR ........................................... 45

5.1 Three Font Types .................................................................................... 50
5.2 Irregular-looking Character .................................................................. 51
5.3 An Example of Chiao-Tung Radical System ......................................... 53
5.4 Dot Matrix pattern of Hanzi character ‘behind’ .................................... 54

6.1 Layer Structure of CCDOS .................................................................... 60
6.2 Chinese Shell on UNIX ......................................................................... 61
Acknowledgments

This work would not have been possible without the help of my advisor, Dr. Junichi Kanai. He had contributed tremendous time and efforts in this project, and had provided many valuable ideas.

This research was supported by the United States Department of Energy and the United States Department of Defense. I wish to thank Dr. Thomas Nartker of Information Science Research Institute (ISRI) who provided much encouragement and a favorable work environment for my whole project. I am very grateful to Robert Cray, Andy Bagdanov for their valuable help in setting up systems. I thank Professor George Nagy of Rensselaer Polytechnic Institute whose careful reviews of the manuscript were greatly appreciated. Many faculty members in the University of Nevada, Las Vegas (UNLV) computer science department have willingly provided their time and expertise. Lively and informative discussions carried out by members of the Chinese Computing Network (CCNET-L@uga.cc.uga.edu) have also been very helpful. Last but not least, I would like to thank my friend Robert Brown, who proofread this paper.
Chapter 1
Overview of Chinese Information Processing

1.1 Introduction

With rapid economic expansion and strong government support for high-tech ventures, China is rushing into the Information Age, largely bypassing many intermediate infrastructural stages such as communications network development and experience with large central computing centers. Since an ideographic writing system is used in Chinese daily life, special handling on computer systems is required for Chinese information processing. A mystique surrounds this field due to a lack of information, especially outside the Chinese language. But the Chinese writing system has deeply influenced many Asian countries. It is clear that Chinese information processing theory and practice are pertinent to both the internationalization as well as localization of character processing in the Information Age.

Reform movements have been historically an integral part of the evolution of ideographic writing systems, just as with orthographic simplification in English speaking countries. However, Chinese reform movements are quite different in Mainland China and Taiwan for historic reasons. This paper will completely review various Chinese ideographic systems and provide important information for internationalization and localization efforts and will cover other related Chinese applications. Since there are
very few books covering Chinese information processing in English, this survey should also provide extremely useful information and insights for English language audiences. The introduction of early computer systems made English information processing a reality. Adapting computer software to handle more complex writing systems, such as Chinese, requires significant research and development efforts. Chinese information processing is more difficult than English processing for the following reasons:

- The Chinese writing system is much more complicated than English.

- Unlike English, which is always read from left to right horizontally, Chinese text can be type-set horizontally or vertically and can be read from left to right as well as from right to left.

- Even the smallest Chinese character set contains over 7,000 commonly used characters, many more than used in English.

- There is no universally recognized Chinese character set standard.

- There is no universally recognized encoding method for Chinese. Conversion is needed when using different encoding schemes.

- The Chinese writing system uses ideograms rather than phonograms. Keyboard entry is an important nontrivial issue in this regard.

- Many Chinese characters consist of complex strokes. Large disk spaces are required for storing fonts.

Chinese information processing is based on and significantly restricted by English language computer system conventions. Understanding how English to Chinese adaption are made is perhaps the key issue in Chinese information processing field.
1.2 Background of Chinese Language

The Chinese language is one of the oldest human languages in the world. The written form used today is developed from the ancient pictographic form. The earliest remaining records known today are engraved on ox bone or tortoise shell dated back to 1,400 – 1,200 BC. Unlike most Western languages, which started as pictographic and then turned to alphabetical, Chinese maintained its pictographic nature.

Table 1.1 shows a sentence in English and Chinese:

<table>
<thead>
<tr>
<th>English</th>
<th>The price of this book is 36 dollars.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chinese</td>
<td>这本书的价格是36美元。</td>
</tr>
</tbody>
</table>

Table 1.1: A Simple Chinese Sentence

Each entity, which has its own unique shape, is called a character. Each character, somewhat equivalent to an English syllable, has its own meaning, shape and pronunciation. There are no phonetic characters in the Chinese language.

Chinese characters are called Hanzi in Chinese. Some of these characters are also used in Japanese and Korean but usually have quite different meanings from their Chinese counterparts. They are known as Kanji in Japanese and Hanja in Korean. There are over 6,000 Kanji and over 1,800 Hanja.

More information about Chinese characters will be introduced in chapter 2.

1.3 Basic Concepts and Terminology

This section will define basic concepts which will apply throughout.
1.3.1 Software Related Concepts

- **Internationalization**: A blanket term referring to the process of preparing software so that it can be used by more than one culture or region. It is often abbreviated as *I18N*.

- **Localization**: A process of adapting software to one specific culture or region. It is often abbreviated as *L10N*.

- **Chinazation**: A specific instance of localization. It requires special character set handling because Chinese character sets require more than one byte to represent all characters.

1.3.2 Chinese Computing Based Concepts

Following are some basic concepts used in Chinese computing:

- **Code Point (or Code Position)**: The numeric code within an encoding method that is used to refer to a specific character. For two-byte characters, this refers to the row and the cell.

- **Character Set**: A mapping between code points and characters. It is implementation independent.

- **Internal Code (or Encoding Scheme)**: The presentation of a character set in a certain implementation. It is implementation dependent and should be used only within a given environment.

- **Exchange Code (or Information Interchange Code)**: Designed for codes transmitted between computer systems with different implementations.

- **Glyph**: A specific instance of a character. A classic example is that *f* and *i* are two separate glyphs, but they can be fused into a single glyph *fi* (called a
ligature).

- **Typeface**: The printed style of a character set.

- **Type Style**: The printed style of a glyph.

### 1.3.3 Encoding Standards

Some popular encoding standards are:

- **ASCII** (*American Standard Code for Information Interchange*): 7-bit code, 34 control characters and 94 graphic characters. It was developed to simplify and standardize machine-to-machine and system-to-system communication in English.

- **EBCDIC** (*Extended Binary-Coded-Decimal Interchange Code*): Internal code for IBM computers. 8-bit code, 66 control characters and 190 graphic characters.

- **EUC** (*Extended UNIX Code*): 8-bit code, compatible with EBCDIC. Implemented as the internal code for most UNIX workstations configured to support multiple-byte encoding, such as Chinese, Japanese, etc.

- **ISO 646** (*the International Standard for 7-bit character sets*): Defines ASCII as an International Reference Version (IRV) with twelve positions for placing alternate characters to create National Replacement Character Sets (NRCS).

- **ISO 8859** (*the International Standard for 8-bit character sets*): A series of several standard character sets that handle all the European Latin-script languages. It currently has nine parts. All of them contain ASCII in their first 128 positions. (e.g. 8859-1 is the well-known *ISO Latin 1*).
• **GB**: Abbreviation for *GuoBiao*, which means *National Standard* in English. It was established in Mainland China and is commonly used for simplified Chinese characters in China and Singapore.

• **Big-5**: A very popular standard used for traditional Chinese characters in Taiwan and Hong Kong.

• **HZ**: Abbreviation for *Hanzi*, which means *Chinese Character* in English. It is a variation of GB and is used on InterNet for simplified characters.
Chapter 2

The Chinese Writing System

Chinese text is typically composed of two different writing systems: Roman characters and Hanzi characters. Hanzi can be in simplified or traditional forms.

2.1 Roman Characters

Roman characters used in Chinese text are the same as those used in Western texts.

<table>
<thead>
<tr>
<th>Name</th>
<th>Character</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lowercase</td>
<td>abcdefghijklmnopqrstuvwxyz</td>
</tr>
<tr>
<td>Uppercase</td>
<td>ABCDEFGHIJKLMNOPQRSTUVWXYZ</td>
</tr>
<tr>
<td>Numerals</td>
<td>0123456789</td>
</tr>
</tbody>
</table>

Table 2.1: Roman Characters in Chinese Text

English words may appear in technical papers. Letters are sometimes used as anonymous names in newspaper or magazine articles.

2.2 Symbols and punctuation

In addition to the symbols and punctuation marks used in Western texts, some other symbols and punctuation marks are also used. The most commonly used symbols are listed in table 2.2.
Chinese Symbol | Meaning
--- | ---
. | Sentence
' | Break
— | —
..... | ...
《 》 | Book Title (Used in Horizontal Text)
『 』 | Book Title (Used in Vertical Text)
々 | Repeat Previous Character
(Not Used in Formal Text)

Table 2.2: Special Symbols in Chinese Text

2.3 Hanzi Character

A Hanzi character is usually referred to as a Chinese character, although strictly speaking, Chinese characters include both Hanzi and Roman characters.

2.3.1 The Structure of Hanzi

Chinese characters look very similar to, and are confusing for, Western people, because they are not familiar with the Chinese character structure — stroke structure. Every Hanzi character is composed of line segments called strokes. Figure 2.1 shows the stroke decomposition of the character 天 (sky).

$$
\text{天} = - + --- + \smallfrown + \backslash
$$

Figure 2.1: The Stroke Structure of Hanzi character ‘sky’

2.3.1.1 Stroke Structure

Strokes can be classified into five basic strokes: dot, horizontal, vertical, horizontal upward and downward drag.
However, knowing only basic strokes is not enough to distinguish similar looking characters. For example, \( \pm \) (soldier) is different from \( \pm \) (earth) and \( \mp \) (dry) is different from \( \mp \) (Yu: last name). Thus stroke types, position, and length are three important features for recognizing and entering Chinese characters. Even though these are not the only methods for character construction, they are fundamental keys to Chinese character recognition (see chapter 4).

There is no absolute standard for classifying strokes. Strokes can be divided into more complex strokes. Scholars have tried to find a complete stroke sub set that generates all Chinese characters, but no consensus has been reached yet. Figure 2.3 shows 34 different complex stroke forms given by the Chinese Character Analysis Group of Taiwan [30] as one suggested complex stroke subset.

The writing order of strokes within a character is generally from left to right and
from up to down. But people are more likely to follow their own older cultural habits, which do not always follow such rules. The stroke decomposition sequence of Hanzi character 天 (sky) showed in Figure 2.1 is also its normal writing order.

2.3.1.2 Graphic Structure

Chinese characters can also be treated as a group of graphic units, which are called root, radical or subradical. Some characters are roots themselves while others are composed of radicals and subradicals.

Figure 2.4 shows the decomposition of the character 灎 (bay):

![Figure 2.4: The Tree Structure of the Hanzi Character “bay”](image)

The structure of a character falls into one of the following four types according to the relationship among the component radicals:

1. Horizontal: 例子 are: 虾, 机, 懒 (shrimp, machine, lazy).

2. Vertical: 例子 are: 男, 草, 染 (man, grass, dye).
3. Surrounded: □ examples are: 回, 国, 周 (come back, country, hear).

4. Inseparable: □ examples are: 月, 母, 天 (moon, mother, field).

Table 2.3 shows a statistics of character structure:

<table>
<thead>
<tr>
<th></th>
<th>Ci Hai Dictionary</th>
<th>GB2312-80</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total char.</td>
<td>16,339</td>
<td>6,763</td>
</tr>
<tr>
<td>Horizontal</td>
<td>68.35%</td>
<td>62.59%</td>
</tr>
<tr>
<td>Vertical</td>
<td>28.77%</td>
<td>22.98%</td>
</tr>
<tr>
<td>Surrounded</td>
<td>8.08%</td>
<td>9.56%</td>
</tr>
<tr>
<td>Inseparable</td>
<td>2.60%</td>
<td>4.87%</td>
</tr>
<tr>
<td>Ave. # of Strokes</td>
<td>12.71</td>
<td>10.64</td>
</tr>
<tr>
<td>Stroke Peak</td>
<td>8-15</td>
<td>7-13</td>
</tr>
</tbody>
</table>

Table 2.3: Statistics of Chinese Character Structure

2.3.1.3 Simplified and Traditional Hanzi

The simplification of Chinese characters started in 1922. In 1956, the Chinese government in Mainland China approved 515 simplified characters and 54 simplified radicals, subradicals and roots. Since then, these simplified characters have been used routinely in Mainland China. But traditional characters are still used in Taiwan and Hong Kong. In 1977, another Hanzi simplification proposal (*Scheme for the Second-Time Simplification of the Chinese Character*) was published in Mainland China in order to reduce stroke number of very complex characters and many frequently used characters. However the 1977 simplification effort was dropped after a short time. One reason was that many characters were over-simplified (i.e. character lost elegance from the calligraph view). Newspaper, magazines, official/government documents and books published during that time in Mainland China were printed using over-simplified characters.
2.3.1.4 Orthographic and Variant Characters

Many Chinese characters have more than one written form for various reasons, such as geographic, ethnic, time, etc. These characters have the same meaning, same pronunciation, but different forms. The one having the most commonly used form is called the Orthographic Character, and the rest are called Variant Characters.

<table>
<thead>
<tr>
<th>Orthographic</th>
<th>Simplified</th>
<th>Variant1</th>
<th>Variant2</th>
</tr>
</thead>
<tbody>
<tr>
<td>倉</td>
<td>仓</td>
<td>全</td>
<td>全</td>
</tr>
</tbody>
</table>

Figure 2.6: Orthographic and Variant Characters

2.3.2 Pronunciation

There are many major dialects (i.e. Mandarin, Shanghai, Cantonese, Sichuanese, Amoyese, Harka, etc) with more than one hundred minor ones all over China. Verbal communication is very difficult for people who speak different dialects. The spoken form was not standardized until Mandarin was selected and published as the official standard by the Ministry of Education of China in 1918. Today, Mandarin is the official language in Mainland China, Taiwan and a few other Asian countries. For political and historical reasons associated with British colonization, people in Hong
Kong speak Cantonese as their official language.

2.3.2.1 Zhu Yin and Pin Yin Symbols

Like English, Chinese also has pronunciation symbols. Although Mandarin is the official language in Mainland China and Taiwan, each uses different set of diacritical symbols. The set used in Taiwan is called Zhu Yin (also spelled Ju Yin) symbols. It contains 37 phonetic symbols which has 21 consonants and 16 vowels with 5 tones (but only 4 tone symbols).

Consonants:

Vowels:

Combination Vowels:

Tones:
1. Even: High and unwavering tone.

2. Rising: Questioning tone.

3. Dipping: Falling and rising tone.


5. Light: Untoned and unaccented tone.

The Zhu Yin symbols are very difficult for Western people to learn. In order to Romanize Chinese spelling, the Chinese Romatzy Spelling Research Committee was established in 1923. The Romanization was adopted by the Mainland Chinese government in February 1958. This set of pronunciation is called *Pin Yin* symbols which has the same 37 phonetic symbols and 4 tones as Zhu Yin symbols.

Consonants:

<table>
<thead>
<tr>
<th>b</th>
<th>p</th>
<th>m</th>
<th>f</th>
<th>d</th>
<th>t</th>
<th>n</th>
<th>l</th>
</tr>
</thead>
<tbody>
<tr>
<td>ㄅ</td>
<td>ㄆ</td>
<td>ㄇ</td>
<td>ㄈ</td>
<td>ㄉ</td>
<td>ㄊ</td>
<td>ㄋ</td>
<td>ㄌ</td>
</tr>
<tr>
<td>g</td>
<td>k</td>
<td>h</td>
<td>j</td>
<td>ㄎ</td>
<td>ㄉ</td>
<td>ㄕ</td>
<td>ㄖ</td>
</tr>
<tr>
<td>zh</td>
<td>ch</td>
<td>sh</td>
<td>r</td>
<td>ㄓ</td>
<td>ㄔ</td>
<td>ㄕ</td>
<td>ㄖ</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ㄆ</th>
<th>ㄆ</th>
<th>ㄇ</th>
<th>ㄈ</th>
<th>ㄉ</th>
<th>ㄊ</th>
<th>ㄋ</th>
<th>ㄌ</th>
</tr>
</thead>
<tbody>
<tr>
<td>ㄅ</td>
<td>ㄆ</td>
<td>ㄇ</td>
<td>ㄈ</td>
<td>ㄉ</td>
<td>ㄊ</td>
<td>ㄋ</td>
<td>ㄌ</td>
</tr>
<tr>
<td>g</td>
<td>k</td>
<td>h</td>
<td>j</td>
<td>ㄎ</td>
<td>ㄉ</td>
<td>ㄕ</td>
<td>ㄖ</td>
</tr>
<tr>
<td>zh</td>
<td>ch</td>
<td>sh</td>
<td>r</td>
<td>ㄓ</td>
<td>ㄔ</td>
<td>ㄕ</td>
<td>ㄖ</td>
</tr>
</tbody>
</table>
Vowels:

<table>
<thead>
<tr>
<th></th>
<th>i</th>
<th>u</th>
<th>ü</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>ia</td>
<td>ua</td>
<td></td>
</tr>
<tr>
<td>Y</td>
<td>iY</td>
<td>xY</td>
<td></td>
</tr>
<tr>
<td>o</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>e</td>
<td>i</td>
<td></td>
<td></td>
</tr>
<tr>
<td>e</td>
<td>i</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ai</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ei</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ao</td>
<td>iao</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ou</td>
<td>iou</td>
<td></td>
<td></td>
</tr>
<tr>
<td>an</td>
<td>ian</td>
<td>uan</td>
<td></td>
</tr>
<tr>
<td>en</td>
<td>in</td>
<td>uen</td>
<td></td>
</tr>
<tr>
<td>a</td>
<td>i</td>
<td>u</td>
<td></td>
</tr>
<tr>
<td>e</td>
<td>ing</td>
<td>ueng</td>
<td></td>
</tr>
<tr>
<td>o</td>
<td>iong</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Tones:

1. Even (unwavering): ã
2. Rising (questioning): ˊ
3. Dipping (falling & rising): ˇ
4. Falling (emphasis): ̀
2.3.2.2 Homonymic and Polyphonetic Characters

Among all possible combination of 37 phonetic symbols and 4 tones, many have no corresponding characters (just as the combination of letters e, r, p in English does not correspond to any actual English word or syllable). According to Revised Chinese Dictionary of Words and Phrases published by Shang Wu in 1981, only 1,335 pronunciations contain at least one character. This means that there exist many Homonymic (with same pronunciation) characters. If the 4 tones are disregarded, there are only 410 different pronunciation combinations. However the distribution of character/pronunciation is very uneven since some pronunciations (i.e. nenɡ) have only one character (能), but others (i.e. yi) may have over one hundred (see Figure 2.8). This should be taken into consideration while implementing input methods based on phonetics.

<table>
<thead>
<tr>
<th>Pin Yin</th>
<th>Tone</th>
<th>Character1</th>
<th>Character2</th>
</tr>
</thead>
<tbody>
<tr>
<td>yi</td>
<td>even</td>
<td>衣 (clothes)</td>
<td>— (one)</td>
</tr>
</tbody>
</table>

Figure 2.8: Homonymic Characters

On the other hand, some characters may have more than one pronunciation. These characters are called Polyphonetic characters. The actual pronunciation of a given Polyphonetic character is determined by the context of its usage. According to the statistics by the Chinese Character Analysis Group of Taiwan, there exists 1,816 Polyphonetic characters.
2.4 Chinese Typefaces, Type Styles and Type Sizes

2.4.1 Typefaces

There are over nine different typefaces used in Chinese writing. Currently, the following four typefaces are most commonly used in Chinese printing and documents (see Figure 2.10).

1. **Song Font**: It is originally developed during the Song Dynasty. The horizontal strokes are about half thinner than the vertical strokes. This font is the most widely used font. It is called the Ming Font by Japanese since it was introduced into Japan during the Ming Dynasty. It is used as a major font in books, newspaper and magazines.

2. **Fang Song (Imitated Song) Font**: It is a derivative of the Song Font with: differences in width between the vertical and horizontal strokes being very small, characters are elongated along vertical side, and there is a slight upper-right tendency on the horizontal strokes. Compared with Song Font, this font is more attractive. It is used in the printing of short articles, some literature books and documents. In Mainland China, it is also the required font for government documents.

3. **Hei (Black) Font**: In this font, the strokes show a great deal of strength because they are blackened. Generally it is used in the title, to catch the attention of readers.
4. **Kai (Writing) Font**: This font was derived during Han Dynasty. Each stroke has its own unique starting and ending features. It shows the strength of the Chinese painting brush. It is usually used as an associate font in newspapers and magazines.

![Figure 2.10: Four Common Type Faces](image)

Other typefaces include: Li Shu, Xing Shu, Wei Bei, Round, Xiao Yao, etc (see Figure 2.11). These fonts are usually used in titles for calligraph art.

![Figure 2.11: Five Other Type Faces](image)
2.4.2 Type Styles

Many type styles are used in newspaper and magazines, usually in titles. Commonly used type styles are: compressed, outline, tilted, underline, shadow, etc. The bold style is usually done by using Hei Font.

<table>
<thead>
<tr>
<th>Font Style</th>
<th>Chinese</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>正常字</td>
</tr>
<tr>
<td>Boldface (Hei Font)</td>
<td>北大</td>
</tr>
<tr>
<td>Vertically-compressed</td>
<td>扁字</td>
</tr>
<tr>
<td>Horizontally-compressed</td>
<td>窄字</td>
</tr>
<tr>
<td>Tilted</td>
<td>斜体</td>
</tr>
<tr>
<td>Contract</td>
<td>反白字体</td>
</tr>
<tr>
<td>Shadow</td>
<td>网格效果</td>
</tr>
<tr>
<td>Outline</td>
<td>网络</td>
</tr>
<tr>
<td>3-D</td>
<td>电子</td>
</tr>
</tbody>
</table>

Table 2.4: Chinese Type Styles

2.4.3 Type Sizes

In English, the character size is measured by point. One point is 1/72 inch. In Chinese, there are 18 commonly used sizes for printing with the following mapping to point (see Table 2.5).

Some other slightly different sizes may still be used. There is no standard on font sizes for political and historical reasons. For phototypesetter, there is a national standard called GB3937-83, where grade (1 grade = 0.25mm) is used as basic unit. There are 30 grades in GB 3937-83, various from grade 1 to 100. Computer programs, i.e. Desk Top Publishing System (DTP), usually define character sizes in points. For
<table>
<thead>
<tr>
<th>Chinese character size</th>
<th>Points (1 point = 1/72inch)</th>
<th>mm</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. 0</td>
<td>42</td>
<td>14.70</td>
</tr>
<tr>
<td>small 0</td>
<td>36</td>
<td>12.60</td>
</tr>
<tr>
<td>No. 1</td>
<td>27.5</td>
<td>9.625</td>
</tr>
<tr>
<td>No. 2</td>
<td>21</td>
<td>7.35</td>
</tr>
<tr>
<td>small 2</td>
<td>18</td>
<td>6.30</td>
</tr>
<tr>
<td>No. 3</td>
<td>16</td>
<td>5.60</td>
</tr>
<tr>
<td>No. 4</td>
<td>13.75</td>
<td>4.8125</td>
</tr>
<tr>
<td>small 4</td>
<td>12</td>
<td>4.2</td>
</tr>
<tr>
<td>No. 5</td>
<td>10.5</td>
<td>3.675</td>
</tr>
<tr>
<td>small 5</td>
<td>9</td>
<td>3.15</td>
</tr>
<tr>
<td>No. 6</td>
<td>8</td>
<td>2.80</td>
</tr>
<tr>
<td>small 6</td>
<td>6.875</td>
<td>2.416</td>
</tr>
<tr>
<td>No. 7</td>
<td>6</td>
<td>2.10</td>
</tr>
</tbody>
</table>

Table 2.5: Chinese Character Sizes

More details, see [31] and [39].

2.5 Chinese Text Style

Many Western people believe that Chinese read from right to left. This is only partially true. In ancient times, books were printed from top to bottom, right to left. However, after the revolutions of the 20th century in Mainland China, books are printed horizontally, from left to right, just like in English. Vertically printed articles are used in newspapers and magazines only for the purpose of making the whole paper look more interesting. In Hong Kong, Taiwan and some northern American countries, some literature books are still printed vertically. But science and technology books are printed horizontally because they usually contain lots of formulas and English terminologies, which will not look good if printed vertically.
Chapter 3

Chinese Character Set Standards and Encoding Methods

3.1 Chinese Character Set Standards

The alphabetical languages typically contain less than 256 characters. One 7-bit or 8-bit byte is enough to represent the character set in these languages. But for Chinese there are more than 256 characters, and one byte is not enough for complete coding. If two bytes are used, \(2^{16} = 65,536\) characters can be coded and is enough for practical Chinese character sets. However, for most information interchange environments, such as e-mail and newsgroups, only 7-bit per byte can be used. In this case, only \(2^{14} = 16,384\) potential characters are available. Since some Chinese character sets approach or surpass \(2^{14}\), either practical character sets have to be reduced below this number or the coding space has to be expanded to provide for the larger sets and future characters.

To establish a character set standard, the following issues must be considered:

- The maximum size of the set
- The length of the character code in the set (byte/char)
- The specific character content of the set
• The order of the characters in the set

The major differences between the Chinese character set and the ASCII set are:

1. The number of Chinese character is huge; there are over 74,000 characters according to the latest statistics by the Chinese Character Analysis Group in Taiwan [20].

2. The frequency of each character is quite different: the most frequently used characters are fewer than 5,000, covering 99% of usage [32]. In addition, the most commonly used 2,000 characters cover 97% of usage [25]; and the 100 most frequently used characters cover 45% of usage. Compared with the total of 74,000 characters, the usage distribution is very uneven. However, we can not ignore the rest of over 60,000 characters just because of their low usage frequency, as one can not discard the least used English letter z without causing chaos.

3. The frequency of each character can change according to the content of the file.

4. There are many variates of common character and many idle (rarely used) characters in Chinese.

5. The total number of Chinese characters is not fixed, and will likely significantly grow in the future.

6. No character set can be big enough to cover every Chinese characters. Set designer must leave some code spaces for the inevitable adding of new characters for special purposes and as a result of general cultural evolution.

For general purpose Chinese computers, the minimum number of characters required are still in dispute. The Chinese Character Analysis Group in Taiwan recommends a total of 33,357 characters, including 4,808 most frequently used and 17,032
less frequently used orthographic and 11,517 variants which includes 3,625 simplified characters used in mainland China. Several popular character sets all have different sizes for various reasons.

3.2 Chinese Encoding Methods

Encoding is the mapping of a character to a numeric value. The most commonly used encoding methods for Chinese are: GB and Big-5. GB code is used in Mainland China; Big-5 is used in Taiwan and Hong Kong. Since these codes are originally designed as internal codes, they are not suitable for information interchange due to the difficulties caused by the different protocols of control characters. Thus some other encoding methods such as ISO2022, HZ, CCCII have been developed.

There are three types of encoding methods:

1. Monolingual character sets: ASCII, GB, Big-5, CNS

2. Multilingual character sets: ISO 10646/Unicode, CCCII

3. Methods of combining monolingual character sets: ISO 2022, EUC, HZ

They can also be classified into two types: 7-bit and 8-bit. All 7-bit encodings are based on ISO 646. Two 7-bit ASCII compatible encoding gives $94 \times 94^1$ coding space. All 8-bit encoding are based on EBCDIC. Two 8-bit gives $190 \times 190 = 36,100$ coding space.

3.2.1 GB 2312-80 Encoding

GB (GuoBiao) is a national character set standard established by the government of Mainland China in 1981. The official name for it is *Code of Chinese Graphic Character Set for Information Interchange — Primary Set* [3]. It is the only standard used in

---

1To avoid 34 ASCII control characters, only 94 graphic codes are usable.
Mainland China. The current version is called GB 2312-80, derived from older version GB2311-80. It is a $94 \times 94$ matrix, in which contains 6,763 simplified Chinese Hanzi characters and 682 letters, numbers and symbols. 1,291 blanks are reserved or user defined. Table 3.1 shows the content of GB.

<table>
<thead>
<tr>
<th>Row</th>
<th># of chars</th>
<th>Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-2</td>
<td>166</td>
<td>Punctuation and Misc Symbols</td>
</tr>
<tr>
<td>3</td>
<td>94</td>
<td>ASCII Set</td>
</tr>
<tr>
<td>4</td>
<td>83</td>
<td>Japanese Hiragana</td>
</tr>
<tr>
<td>5</td>
<td>86</td>
<td>Japanese Katakana</td>
</tr>
<tr>
<td>6</td>
<td>48</td>
<td>Greek Letters</td>
</tr>
<tr>
<td>7</td>
<td>66</td>
<td>Russian Letters</td>
</tr>
<tr>
<td>8</td>
<td>63</td>
<td>26 Pin Yin Accented Vowels &amp; 37 Zhu Yin symbols</td>
</tr>
<tr>
<td>9</td>
<td>76</td>
<td>Table-drawing Elements</td>
</tr>
<tr>
<td>16-55</td>
<td>3755</td>
<td>Level 1 Hanzi</td>
</tr>
<tr>
<td>56-87</td>
<td>3008</td>
<td>Level 2 Hanzi</td>
</tr>
<tr>
<td>10-15,88-94</td>
<td></td>
<td>User Defined</td>
</tr>
</tbody>
</table>

Table 3.1: Content of GB 2312-80

Each row of the GB table is called a “Qu”, and each column is called a “Wei”. The coordinates of a character in the table is called its “Qu-Wei” matrix code. There is a simple relation between GB code and “Qu-Wei” matrix code:

$$7\text{-bit GB First byte} = Qu + 32$$

$$7\text{-bit GB Second byte} = Wei + 32$$

$$8\text{-bit GB} = 7\text{-bit GB} + 128 \text{ (for both first and second bytes)}$$

Since Qu and Wei range from 01 to 94, while ASCII ranges from 33 to 126, ‘32’ is added in order to match the GB byte into the ASCII range. In most application systems, for example, under CCDOS (*Chinese Character Disk Operating System*), ‘128’ is added to distinguish GB and ASCII by marking the highest bit of each GB
byte with '1'. This variation is called 8-bit GB code. The original GB 2312-80 is called 7-bit GB or pure GB code.

7-bit GB is designed to be used as exchange code while 8-bit GB is usually used as internal code. 8-bit GB code can not be used as exchange code on non-8-bit clean systems.

```
Oxxxxxxx Oxxxxxxx 7-bit GB code (two bytes)
1xxxxxxx 1xxxxxxx 8-bit GB code (two bytes)
0xxxxxxx ASCII (one byte)
```

Not every cell in the 94 × 94 GB table is used. The range for the first GB byte is from 0x21 to 0x77, the range for the second byte is from 0x21 to 0x7E. All GB codes avoid the range of ASCII control characters — this is important for the compatibility with ASCII.

The Chinese characters in GB table are divided into two levels: Level One from B0A1 to D7F9 contains 3,755 most commonly used characters. Level Two from D8A1 to F7FE contains 3,008 less commonly used characters (including radicals).

All these 6,763 Chinese characters are selected based on Table of Printed Commonly Used Chinese Character, jointly published by Committee of Reforming the Chinese Written Language and the Ministry of Culture of the Mainland China. The characters in Level One are arranged by pronunciation (Pin Yin). The characters in level Two are arranged by radical then by increasing stroke count.

The latest GB standard is called GB-13000. It was announced by Chinese National Standard Organization in December 1993. GB-13000 is actually equivalent to the CJK portion of ISO 10646/Unicode (see Section 3.2.6.1).
3.2.2 HZ Encoding

In order to exchange Chinese text file through Electronic Mail (E-mail) on ASCII-based computer systems, HZ (Hanzi) is introduced as an encoding method for encoding arbitrarily mixed Chinese and ASCII characters in a 7-bit format. HZ is also intended for the design of terminal emulators that display and edit mixed GB and ASCII text files in real time. Strictly saying, HZ is not another encoding method — it is just a method of combining 7-bit GB (pure GB) set and ASCII set. The format of HZ is described in the following:

1. The default mode is ASCII mode. In ASCII mode, a byte is interpreted as an ASCII character, unless a ‘~’ is encountered.

2. The character ‘~’ is an escape character. By convention, it must be immediately followed only by ‘~’, ‘{’, ‘}’ or ‘\n’, with the following special meaning:
   - The sequence ‘~’ is interpreted as a character ‘~’.
   - The escape-to-GB sequence ‘~{’ switches the mode from ASCII to GB.
   - The escape-from-GB sequence ‘~}’ switches the mode from GB to ASCII.
   - The sequence ‘~\n’ is a line continuation marker to be consumed with no output produced.

In GB mode, characters are interpreted two bytes at a time as 7-bit GB code until the escape-from-GB ‘~\}’ is encountered.

Three things must be noticed: First, the escape-from-GB code ‘~\}’ (7E7D) is outside the defined 7-bit GB range. Second, an ASCII ‘~’ is always encoded as ‘~‘. Third, a sequence of GB code should be enclosed in ‘~{’ and ‘~}’.

HZ is now very popular on Internet for GB users. The encoding process is very straightforward. Lots of software have been developed for reading E-mail, news,
electronic magazines in HZ on ASCII computer systems. One disadvantage of HZ is that it does not support Big-5 code (a popular code for traditional characters used in Taiwan and Hong Kong, see Section 3.2.3). Several proposals for adapting HZ code to support Big-5 have been discussed on the Chinese Computing Network. However, there is no agreement on this subject yet.

3.2.3 Big-5 Encoding

Big-5 is the most popular encoding standard used in Taiwan and Hong Kong, where traditional characters, rather than simplified characters, are used daily. It was established in May 1984 by the Institute for Information Industry of Taiwan. Although Big-5 is not a national standard, it represents more than 90% of the Chinese computer system market in Taiwan.

The Big-5 character set standard is set in a 94 × 157 matrix, which contains 13,053 Hanzi characters and 470 non-Hanzi characters. The Hanzi characters are divided into two levels: Level One contains 5,401 commonly used characters and level Two contains 7,652 less commonly used characters. The Hanzi character in both levels are arranged by increasing stroke count, then by radical. Table 3.2 shows the content of Big-5. The first byte of Big-5 is in the upper ASCII range: 0xA1–0xFE. The second byte (0x40–0x7E) is not restricted to the same range.

<table>
<thead>
<tr>
<th>Row</th>
<th>Code</th>
<th># of chars</th>
<th>Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-3</td>
<td>470</td>
<td>5,401</td>
<td>Level 1 Hanzi</td>
</tr>
<tr>
<td>4-38</td>
<td>A140-C87E</td>
<td>7,652</td>
<td>Level 2 Hanzi</td>
</tr>
<tr>
<td>41-89</td>
<td>C940-FEFE</td>
<td></td>
<td>User Defined</td>
</tr>
</tbody>
</table>

Table 3.2: Content of Big-5

Big-5 also has two forms: 7-bit form and 8-bit form. The 8-bit form is very popular and is widely used in many Chinese Operation Systems such as ETen, Cxterm,

One big problem of Big-5 is that it has no officially approved table standard. It is difficult to find two identical tables across different systems. For example, the ranges from C6A1 to C8FE and F9DD to F9FE are used by ETen, but not by Chinese Windows. Characters encoded in this range by ETen can not be correctly shown in Chinese Windows. Also, the glyph shapes are very poorly defined: different characters may have the same code point.

3.2.4 CNS 11643 Encoding

CNS was established as the national standard by Taiwan government in 1986. It is officially registered as an ISO 2022 character set and follows the ISO 646 and ISO 2022 communication protocols, which means all control codes in ASCII set are avoided. The range from 00 to 20 and 7F of the 7-bit code are not used. The latest CNS version is called CNS 11643. Two earlier versions are called CNS 5205 and CNS 7654. CNS 11643-1986 contains 13,051 Hanzi while CNS 11643-1992 was expanded to more than 60,000 Hanzi.

The CNS 11643-1986 character set has only two planes with only 13,051 Chinese characters. The CNS 11643-1992 character set is expanded into 16 "planes". There are $94 \times 94 = 8836$ characters on each plane. The first seven planes are defined in Tabel 3.3.

In each plane, like Big-5, the characters are arranged by the number of strokes and then by radical. This scheme is less desired because of the following reasons:

1. The distribution of character strokes is heavily concentrated in the range between 10 to 17 strokes (over 3,000 characters in each category), while others are very low. For example, 1 stroke category only contains 16 characters.
Table 3.3: Content of CNS 11643

<table>
<thead>
<tr>
<th>Plane</th>
<th>Range</th>
<th># of chars</th>
<th>Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2121-427E</td>
<td>684</td>
<td>Roman Chars, Radicals, Misc Symbols</td>
</tr>
<tr>
<td></td>
<td>4421-7D4B</td>
<td>5401</td>
<td>Level 1 Hanzi</td>
</tr>
<tr>
<td>2</td>
<td>2121-7244</td>
<td>7650</td>
<td>Level 2 Hanzi</td>
</tr>
<tr>
<td>3</td>
<td>2121-6246</td>
<td>6148</td>
<td>Frequently Used Hanzi</td>
</tr>
<tr>
<td>4</td>
<td>2121-6E5C</td>
<td>7298</td>
<td>Rarely Used Hanzi (part 1)</td>
</tr>
<tr>
<td>5</td>
<td>2121-7C51</td>
<td>8603</td>
<td>Rarely Used Hanzi (part 2)</td>
</tr>
<tr>
<td>6</td>
<td>2121-647A</td>
<td>6388</td>
<td>Variant Character (part 1)</td>
</tr>
<tr>
<td>7</td>
<td>2121-6655</td>
<td>6539</td>
<td>Variant Character (part 2)</td>
</tr>
<tr>
<td>8-16</td>
<td></td>
<td>8290</td>
<td>Undefined</td>
</tr>
</tbody>
</table>

2. Not everyone knows the correct number of strokes for each character. Some complex strokes may be counted differently by different people.

There are two forms of CNS11643-1986: 7-bit and 8-bit forms. The 8-bit CNS can be found in Eten’s CNS mode and DEC stations in Taiwan. The 7-bit form uses certain escape sequences to distinguish plane 1 and 2. It was used in the mail and news systems in the SeedNet of Taiwan. However, SeedNet switched to Big-5 in May 1994 because of the popularity of Big-5.

During the transmission of information, if a character belongs to different planes, a plane shift control character must be sent first to signal the change of plane. In order to increase the efficiency, the most frequently used characters are placed in the first plane.

CNS is quite different from Big-5 in the following aspects:

- The order of some of the alphabets and symbols is different

- The order of the Hanzi is different because of corrected stroke counts. 14 Hanzi (Five in plane 1 and nine in plane 2) are relocated.

- There is no Non-Hanzi character in CNS
• Two duplicated characters in Big-5 set were deleted

• CNS uses two 7-bit bytes plus a plane number to define a character while Big-5 uses two 8-bit bytes.

• CNS is an official standard while Big-5 is a non-official but popular standard.

3.2.5 CCCII

CCCII (Chinese Character Code for Information Interchange) was developed by the Chinese Character Analysis Group in Taiwan. It is also known as REACC/EACC (RLIN Eastern Asian Character Code).

The first CCCII version was published in 1980. It contains 4,807 most frequently used Chinese characters, 214 radicals, 76 symbols and punctuation marks, and 41 phonetic symbols [2]. The second edition was published in 1982. It contains 33,544 characters (21,885 orthographic characters and 11,660 variant characters including simplified characters). The latest version was released in 1987, contains 53,940 characters (4,808 most frequently used characters, 17,032 next frequently used characters and 20,583 rarely used characters).

CCCII is based on the ISO 646 communication 7-bit coding standard. Its technique of code extension and the identifying location of the escape sequence are based on the ISO 2022 standard. It uses three bytes per character, which gives a $94 \times 94 \times 94$ coding space (see Figure 3.1).

The coding space can also be treated as three dimensional vectors: \((plane, section, position)\). It is illustrated in Table 3.4. A layer is defined as a group of six consecutive planes. There are total of 16 layers \(^2\). Table 3.5 shows the first layer of CCCII.

The Chinese characters in CCCII are ordered by radical (same sequence as used in Kang Xi Dictionary), first by ascending stroke count, second by the precedence

\(^2\)The last layer contains only four planes (plane 91–94).
A variant character is placed in the same section and position as its corresponding orthographic character, but in a different plane (layer). By such an arrangement, variant forms of a given character can be identified with the orthographic forms of a given character by simply adding a plane (layer) offset number.

Since CCCII was designed to be a universal tool for all Chinese, Japanese and Korean information interchange, all symbols used in these languages are collected and coded. However, not every user needs such a big set, and CCCII also allows users to select a proper subset of the code.
<table>
<thead>
<tr>
<th>Plane:Section</th>
<th>Content</th>
<th># of Characters</th>
</tr>
</thead>
<tbody>
<tr>
<td>1:1</td>
<td>Reserved for Control Codes</td>
<td></td>
</tr>
<tr>
<td>1:2-3</td>
<td>Arithmetic and ASCII Symbols</td>
<td></td>
</tr>
<tr>
<td>1:4-10</td>
<td>User’s Space</td>
<td></td>
</tr>
<tr>
<td>1:11</td>
<td>Punctuation Marks</td>
<td>35</td>
</tr>
<tr>
<td>1:12-14</td>
<td>Radicals</td>
<td>214</td>
</tr>
<tr>
<td>1:15</td>
<td>Numerals and Phonetic Symbols</td>
<td>78</td>
</tr>
<tr>
<td>1:16-67</td>
<td>Most Frequently Used Hanzi</td>
<td>4808</td>
</tr>
<tr>
<td>1:68-3:64</td>
<td>Next Most Frequently Used Hanzi</td>
<td>17,032</td>
</tr>
<tr>
<td>3:65-6:5</td>
<td>Other Hanzi</td>
<td>20,583</td>
</tr>
<tr>
<td>6:6-94</td>
<td>Undefined</td>
<td></td>
</tr>
</tbody>
</table>

Table 3.5: Content of CCCII’s First Layer

There are some methods for compressing 3-byte CCCII code into 2-byte code. Also, a Chinese Character Data Base was developed in order to make the usage of CCCII easy and efficient [20].

3.2.6 International Encoding Methods

The International Standard Organization (ISO) has established several standards related to the international information interchange. Two important ones, ISO 10646/Unicode and ISO 2022, are briefly described here.

3.2.6.1 ISO 10646/Unicode Encoding

The ISO and the Unicode Consortium have both developed a multilingual character set designed to combine the majority of the world’s character set standards into one larger set. These two standards are named ISO 10646 and Unicode.

ISO 10646 is a 4-byte/32-bit code. It has 128 groups of 256 planes. Each plane contains $256 \times 256 = 65,536$ coding space. However, 4-byte is hard for most people to accept.

Unicode is a 2-byte/16-bit fixed-width representation. It encodes the characters from the world’s principal scripts and languages, allowing the conversion to and from
all national, international and vendor code set standards.

In 1993, ISO 10646:1990 and Unicode Version 1.0 were merged into ISO 10646-1:1993 with Unicode Version 1.1 as a compliant subset. The first plane (plane 00 of group 00) of ISO 10646 is changed to a 2-byte form. It is called the Basic Multilingual Plane (BMP) (see Table 3.6).

<table>
<thead>
<tr>
<th>Zone</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>A-zone</td>
<td>Alphabets, Symbols, CJK Auxiliary, Hangul,...</td>
</tr>
<tr>
<td>I-zone</td>
<td>Unified CJK Ideographs (20992)</td>
</tr>
<tr>
<td>O-zone</td>
<td>Reserved for Future Use (16384)</td>
</tr>
<tr>
<td>R-zone</td>
<td>Private Use(6K), Compatibility Area, Arabic Presentation Forms,... (8192)</td>
</tr>
</tbody>
</table>

Table 3.6: Basic Multilingual Plane

Unicode (version 1.1) is a subset of ISO 10646. It fits into the BMP of ISO 10646. It tries to unify approximately 21,000 Han characters (Hanzi, Kanji and Hanja) from various national standards into a single set of approximately 20,902 characters. This effort is known as “Han Unification”. The unified set is known as CJK (Chinese, Japanese and Korean). There are many Chinese characters in the different standards which share both a common form and meaning. Those that have slight form differences are unified into a single code point under the process of Han Unification. The Unicode CJK set is a 256 x 256 matrix. It is defined in Table 3.7:
Table 3.7: Content of ISO 10646/Unicode

<table>
<thead>
<tr>
<th>Row</th>
<th>Codes</th>
<th># of Chars</th>
<th>Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-17</td>
<td>0000-1FFF</td>
<td>2,300</td>
<td>Alphabets</td>
</tr>
<tr>
<td>18</td>
<td>240</td>
<td></td>
<td>Hangul Alphabet</td>
</tr>
<tr>
<td>31-32</td>
<td>478</td>
<td></td>
<td>Latin &amp; Greek Precombined Forms</td>
</tr>
<tr>
<td>33-48</td>
<td>2000-2FFF</td>
<td>1376</td>
<td>Symbols</td>
</tr>
<tr>
<td>49-52</td>
<td>3000-3FFF</td>
<td>842</td>
<td>CJK Auxiliaries</td>
</tr>
<tr>
<td>53-78</td>
<td>6,656</td>
<td></td>
<td>Hangul</td>
</tr>
<tr>
<td>79-160</td>
<td>4000-E7FF</td>
<td>20,902</td>
<td>CJK Unified Ideographs</td>
</tr>
<tr>
<td>250-251</td>
<td>302</td>
<td></td>
<td>CJK Unified Compatibility</td>
</tr>
<tr>
<td>252-254</td>
<td>650</td>
<td></td>
<td>Presentation Forms</td>
</tr>
<tr>
<td>255-256</td>
<td>422</td>
<td></td>
<td>Compatibility &amp; Specials</td>
</tr>
</tbody>
</table>

3.2.6.2 ISO 2022

ISO 2022 is called the *Code Extension Techniques for Use with the ISO 7-bit Coded Character Set*. It is based on the ISO 646 standard, but was extended to a multiple byte code. All coding schemes must avoid the 34 control characters, which means one 7-bit byte can have only 94 coding spaces for graphic characters.

Strictly speaking, ISO 2022 is not a character set. It is just a collection of escape sequences and conventions for switching between character sets that are already defined. It is an open standard and only defines global meanings for the escape sequences. It can only be used with compatible character sets. The ISO compatibility is defined as following conditions:

- A fixed number of bytes in the range 0x21-0x7E or 0x20-0x7F.
- The high bits of both bytes must be identical.
- Special escape sequence are used to identify appropriate character set.

It is very important that an acceptable coding scheme follow the ISO coding rules. Character sets to be used with ISO 2022 should be registered and thus assigned escape
sequences. GB 2312-80 and CNS level 1 and 2 are registered. The GB set also conforms to the byte range of ISO 2022.

<table>
<thead>
<tr>
<th>Size</th>
<th>Character Set</th>
<th>Official (final char)</th>
<th>Escape Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>94</td>
<td>ASCII</td>
<td>'B'</td>
<td>ESC (B)</td>
</tr>
<tr>
<td>94 × 94</td>
<td>GB2312</td>
<td>'A'</td>
<td>ESC $ A</td>
</tr>
<tr>
<td>94 × 94</td>
<td>CNS 11643 plane 1</td>
<td>'G'</td>
<td>ESC $( G</td>
</tr>
<tr>
<td>94 × 94</td>
<td>CNS 11643 plane 2</td>
<td>'H'</td>
<td>ESC $( H</td>
</tr>
</tbody>
</table>

Table 3.8: ISO 2022 Standard for GB Set
Chapter 4

Chinese Input

Unlike English, in which words can be composed by typing in letters, Chinese has no alphabet to be entered to compose a given character. Thus, special input methods for addressing a given character must be derived.

Chinese input method can be classified in many different ways. The classification diagram showed in Figure 4.1 is based on the human/machine interface. All these methods, except for the big keyboard input method, are very active research areas. They are going to be discussed in this chapter.

4.1 Big Keyboard input Method

The big keyboard input method originated from the Chinese typewriter. The operation of a Chinese typewriter proceeds in a hunt-and-peck fashion. The printing head holds the most frequently used characters which are stored in a grid ordered by radical. Less frequently used characters will be inserted into the grid when needed. One example of this direct input method is the input device made by General Computer and Data System Co. It is similar to a Chinese typewriter. An optical sensor is used to read the internal code on the printing head into memory.

In the 1970s, IBM developed a touch key input device called IBM 5255 system. The keyboard consists of 256 character keys and 12 control keys. There are total of
256 \times 12 = 3,172 \text{ Chinese characters available on the keyboard to be entered directly.}

In the late 1970s, based on the high resolution of 3277 display terminal, IBM gave an implementation of Chinese input. Hundreds of Chinese characters were displayed on the screen. The desired character is chosen by touching the position on the screen using a special optical pen. This was the beginning of the pen touch input technique. Later, another pen touch big keyboard input approach was developed by the Japanese. Instead of using the screen, a special panel was designed. Over 3,000 characters were fixed on the panel. For a professional typist, the average speed was 40–70 characters per minute.

Some of the big keyboards are very useful in certain areas. For example, pen touch
keyboards are used in the computer inquiry system at a Japanese airport. However, they are rarely used in most places due to the following disadvantages:

- The device is heavy and bulky.
- There is a limit of the number of characters a big keyboard device can represent.
  It is incapable of adding new characters.
- It is difficult to memorize the keyboard layout, hence can not be commonly used in daily life.

4.2 Small Keyboard Input Method

Since the appearance of microcomputer, hundreds of input methods on small keyboard (i.e. the English ASCII keyboard) have been developed. These methods can be divided into three types:

- Input by pronunciation
- Input by structure
- Input by encoding value

All these input methods allow the user to input text phonemically with the use of Roman characters. One big problem with these methods is the high collision rate, i.e. different characters result in the same index. To solve this, the number of keyboard strokes has to be increased, which means more key entry time is needed. All input methods are designed in an effort to reduce the number of key strokes. Generally, the number of key strokes is between one and six per character.

More than 500 input methods have been invented. An evaluation standard has been set up to compare different input methods. For details about the rules and implementation, refer to [18].
4.2.1 Input by Pronunciation

Every Chinese character’s phonetic code (which consists of a consonant and a consonant-vowel) can be represented by English letters. However, in Chinese, different characters may have the same pronunciation. Therefore, a phonetic code is mapped to a group of Chinese characters that share the same pronunciation. The wanted character can be selected from the list of the group.

Two well known pronunciation-based methods are Pin Yin and Zhu Yin. Pin Yin is used in Mainland China while Zhu Yin in Taiwan. The keyboard layout for Pin Yin method is the same as for the English keyboard (also called QWERTY keyboard) since Pin Yin symbols are already Romanized. There are usually one to six keyboard strokes for each character. An improved Pin Yin method is called Shuang Pin method. It reduced the number of keyboard strokes by mapping complex vowels or constants into a single letter. Table 4.1 shows a carefully designed mapping.

![Table 4.1: Shuang Pin Keyboard Mapping](image)

For example, the keystroke for Pin Yin ‘zhang’ using Pin Yin method is ‘z-h-a-n-g’, a total of five keystrokes. However, if Shuang Pin method is used, the keystroke sequence will be reduced to ‘a-h’, only two strokes.

The Zhu Yin input method works in a very similar way as the Pin Yin input method. There are several proposals for the Zhu Yin keyboard layout. Table 4.2 shows one of the layouts used by TwinBridge Chinese System:

Pronunciation-based methods are provided by every Chinese system since they are very easy to use and need the minimum training. However, they are also relatively
slow because of the large number of homonym characters.

One way to solve this problem is the use of a phrase input method, which is based on the fact that the homonym phrases (of more than one character) are very rare. A successful implementation of this phrase input is the real-time, on-line information storage and retrieval system for telephone information services by the Telecommunication Laboratories in Taiwan. A phonetic encoding system was implemented at Academia Sinica of Taiwan in 1973. The system resolves the collision problem by typing in the pronunciation of a two-character-phrase that contains the desired character. Based on the same principle, a Multilingual Word Processing System was implemented by Becker [14]. This system can use both Pin Yin and Zhu Yin.

Another way to solve homonym collision is to focus on the structure of Chinese characters, rather than on their pronunciation. It will be discussed in the following section.

4.2.2 Input by Structure

This group of methods is based on the form of the Chinese character. It can be further divided into three types:

1. Input by radical

2. Input by stroke
3. Input by corner.

4.2.2.1 The Radical Type

All Chinese characters are made up of a major and some minor radicals. These major or minor components can be represented by English letters. The mapping from Chinese radicals to English letters is an n-to-1 mapping and is carefully chosen so that the number of characters with the same stroke combination is optimal.

One popular method is called "Wu Bi" (means five basic strokes) Input Method. It is used in Mainland China for all characters in GB2312-80 and is regarded as the most efficient input method. A professional Wu Bi typist can type at approximately 120 characters per minute. However, one has to memorize the special Wu Bi keyboard layout (see Figure 4.2) and practice often. It is not very hard to memorize the keyboard layout, but is easy to forget after a short period of inactivity.

The 26 keys are divided into five areas corresponding to five basic strokes. 130 basic radicals are chosen and assigned to a key according to their starting basic stroke. The letter 'z' is used as wildcard when the user is uncertain about which key should be chosen. The system will give a list of all possible characters. The maximum number of keyboard strokes is four. Also 2,300 most frequently used phrases (composed of two or more characters) can be entered with at most four strokes.
Another popular method is called “Cang Jie” which is widely used in Taiwan and Hong Kong. It uses 24 radicals/roots. Like Wu Bi, this method also has its special keyboard layout (see Figure 4.3).

4.2.2.2 The Stroke Type

The stroke-based methods describe a character by its stroke form and number of total strokes. Similar to an English word being spelled out by a string of letters, a Chinese character is addressed by a stroke sequence code according to their writing order. One example is called 'Bi Hua’ method. These methods are not so efficient because the average number of basic strokes or complex strokes per character is large (11 basic strokes for traditional characters); and because the writing order of strokes is difficult to decide under some situations.

4.2.2.3 The Corner Type

This type of methods works by assigning a certain number to each of the two (diagonal), three or four corners of a given character according to the stroke form in that corner. A set of conversion rules is needed in order to match a number to a stroke form.
One typical example is the 3-corner Coding Method by Jack K. T. Huang et al [30]. Figure 4.4 shows its symbol table.

![Figure 4.4: Fundamental Symbols of 3-Corner Coding Method](image)

Ninety-nine major corner codes and 201 subradicals are used in this method. Each one is represented by two digits from 00 to 99. The collision rate for this method is about 3% among all commonly used characters. It requires one to remember the code for radicals.

Another type is called 4-corner method invented in the 1930s as a dictionary lookup index. The roots (major radicals) are numbered from 0 to 9. This method is relatively easy to learn, but the collision rate is high.

The 2-corner input method is also called the ‘Diagno’ or ‘Shou Wei’ method.

### 4.2.3 Input by Encoding Value

As introduced in the previous chapter, Chinese characters are encoded by GB code, or Big-5 code. Some input methods are implemented by using these internal codes.
These methods are similar to entering extended ASCII on US keyboard. They are usually used for the characters that fail to be entered by other commonly used methods.

Besides internal codes, some systems also provide Telex code as an input method. Telex code was implemented for the Chinese Telephone and Telegraph Company in 1911. It used to have four digits which encompassed at most 10,000 Chinese characters. Now it has been enlarged to five digits.

### 4.2.4 Input by Other Criteria

In some systems, special input methods such as *English-Chinese dictionary* (enter the English word and choose one from the corresponding Chinese words), *phrase association* (given a single character, list all possible words started with that character), are provided. These association methods are usually used to improve input speed.

Multiple input criteria can also be combined to supplement one another. For example, *Dai-E* [4] method works by typing in the phonetic symbols together with diagonal digit numbers, which is a combination of pronunciation type and corner type. Many systems provide Pin Yin Method combined with phrase association to improve key enter speed for Pin Yin method users, it is called *Lian Xiang* method.

### 4.3 Optical Character Recognition

An efficient way to enter printed or written Chinese text is Optical Character Recognition (OCR). The research on Chinese Character Recognition (CCR) started in the 1960s. It has been a very difficult task due to the large number of characters and the complexity of character structures. CCR can be divided into two subfields:

1. Handwritten (connected) and handprinted (isolated) CCR

2. Machine printed CCR
4.3.1 Online/Offline Handwritten and Handprinted CCR

Online handwritten and handprinted CCR systems recognize input characters according to their stroke features. Figure 4.5 shows the basic principle:

![Diagram of basic steps in online handwritten CCR](image)

Figure 4.5: Basic Steps in On-line Handwritten CCR

Commonly used tablets are: Pressure-sensitive tablets, Electrostatic tablets, and Resistance tablets.

Practical systems have been developed in Mainland China, Japan and USA. They require strict restrictions on handwriting of each stroke if they are to work efficiently. Therefore, speed and accuracy are relatively low and recognition can be tedious when reading large amounts of handwriting.

One on-line handwritten CCR system was developed in 1988. In this system, a fuzzy-attributes automaton for stroke recognition is applied and a structure line segment ordering method is also adopted to tolerate irregular writing patterns.

Offline handwritten recognition is more difficult since information on stroke sequence is lost. It is still in the research stage. Proposed research approaches include the following:

1. **2-D extended attribute grammar method** (2-D EAGs). It uses key structures (which distinguish a character from others) for recognition and disregards redundant structure. To distinguish similar characters, the concepts of polysemous grammars and parallel grammars have been introduced [45].
2. **Dynamic decision tree method**: Each character can be parsed into distinct paths. An attention area is selected from this pattern and a key feature is extracted from this area. All characters that have the same key feature are put into the candidate set of the root node of the decision tree. The recognition result is obtained in an object-directed manner [33].

3. **Sub-structure detector method**: A Chinese character is divided into sub-structures. A neural network is adopted to recognize Chinese characters [24].

A large sample base for handwritten CCR was built by the Institute of Automation of Academia Sinica in 1988, consisting of 4,000,000 sample characters written by 1,000 people.

### 4.3.2 Printed Chinese Character Recognition

There are three types of input devices for printed CCR: Camcorders, FAX facsimiles, and Scanners. Scanners are the most popular type. The earliest research on printed CCR was by Casey and Nagy in 1966 [15]. Much basic work was done in late 70s. Methods included: large tree classifier with heuristic search and global training [42]; feature extraction of printed CCR [9]; constrained connection graph [23]; attributed relational graph [11]; and linear regression [29]. Most of the feature extraction research is based on the concept of stroke information since Chinese characters can be represented by set of strokes. The matching techniques depend highly on the type of feature extraction method applied.

The first printed Chinese OCR system was developed by the Shen Yang Institute of Automation in Mainland China in 1984. It was designed for a single font, and only 2,500 characters could be recognized. In 1989, the first multifont printed OCR system was developed by Qin Hua University in Beijing. It can recognize the 3,755 most frequently used characters defined in GB Level One plus some other symbols.
Current commercial printed Chinese OCR systems utilize TIFF images in DOS, UNIX and MS-Windows.

4.4 Audio Input Method

Voice input in Chinese started in the 1980s. The earliest paper was by Yu T C in 1983 [43]. It described an experimental work with only thirty-one words. Different Mandarin Speech Recognition Systems are described by Shi [6], Chang and Yang [5], Chen and Pao [17], and Yu [8].

Some innovative algorithms for analyzing Mandarin homonyms was developed in 1986 by Pao and Tan [7].

4.5 Chinese Character Dictionaries

Dictionaries are very useful resource for entering Chinese characters, especially old-fashioned books. There are two types of dictionaries:

- Conventional dictionary
- Specialized dictionary

Dictionaries typically use character structure or pronunciation as an index. Commonly used indices are: radical, stroke number, 4-corner code, Pin Yin, and Zhu Yin.

The Earliest dictionary is Shuo Wen Jie Zhi (Read and Explain Chinese) by Xu Shen in 121 AD. It contains 9,353 character. In 1716, the Kang Xi Dictionary was commissioned by the Emperor Kang Xi. Over 47,000 characters were collected. Nobody knows the exact number of Chinese characters today. The Chinese Character Analysis Group in Taiwan is trying to collect all Chinese characters. Currently over 74,000 Chinese characters have been collected, identified, sorted and coded.
Several ongoing projects are building computerized Chinese–English dictionaries. University of Massachusetts and University of Foreign Studies in Beijing are compiling a contemporary computerized Chinese-English dictionary. Their ultimate goal is to collect more than 85,000 entries. Dr. Cheng Chungying, (University of Hawaii at Manoa) is making a computerized English dictionary of Chinese philosophy which will cover all information on the history of Chinese philosophy from 2,000 BC to the present.
Chapter 5

Chinese Output

There are two major output devices: monitors and printers. In the case of computer monitors, the most basic unit (picture element) of output is called a *pixel*, while in the case of printers, it is called a *dot*. No matter which output system is used, fonts form the basis of document writing.

5.1 Chinese Fonts

One of the most important aspects of Chinese output is the availability of fonts. Fonts can be stored in a variety of formats. A typical Chinese font consists of all characters from one of the two most popular encoding methods: GB2312-80 and Big-5. The size of the font depends on the size of the character set. The speed of output depends on the size of the font bitmap. Usually for alphabet letters, $7 \times 9$ bitmaps are sufficient. But for complex Chinese characters, at least a $15 \times 16$ bitmap is required. Many tools have been developed for designing fonts. These tools can be grouped into general types, corresponding to three forms in which typefaces are manipulated:

1. **Bitmap type**: Fonts are described and manipulated as explicit bitmaps.

2. **Vector type**: Typefaces are represented by vectors.
3. **Outline type:** Typefaces are described and manipulated as outlines composed of curves, splines, etc.

Figure 5-1 shows three forms of character "k" using different methods. Although the internal representation of characters in these formats differs considerably, the final result, whether printed or displayed on a computer monitor, is simply a bitmapped font.

![Figure 5.1: Three Font Types](image)

### 5.1.1 Bitmap Font

The first Chinese fonts were bitmapped. Each character is constructed from a matrix. Each cell of the matrix could be turned on or off (1 or 0).

There are many different bitmap font formats. One of them is called HBF, short for "Hanzi Bitmap Font". The format of a HBF file is based on that described in X11R5 X Logical Font Description (XLFD) Conventions version 1.4 and X11R5 Bitmap Distribution Format (BDF) version 2.1. A separate HBF file is used to specify the HBF properties and to record the index (file offset) information to access the separately existing bitmap file(s) of the HBF. The bitmap file exists independently
of the HBF file. HBF is the suggested standard format on most UNIX based systems. The adoption of a standard such as HBF alleviates the problem of different software being distributed with their own Hanzi bitmap files. It also reduces the need to keep different formats of the same Hanzi bitmap fonts, thus reducing disk space requirements.

The limitation of Bitmap fonts is that they are usually restricted to a certain size. Any enlarging may produce irregular-looking (jaggy) results (see Figure 5.2).

![Figure 5.2: Irregular-looking Character](image)

5.1.2 Vector Fonts

Vector fonts are also known as *parametric fonts*. They improve the internal representation of bitmap fonts by using vectors to define their shapes mathematically. Each character is broken down to primitive elements, and these elements form a library of instructions for describing the shape of a character. Vector fonts are scalable. However, they do not provide very pleasing typographic results, especially when printed at large point sizes.

5.1.3 Outline Fonts

More recent Chinese fonts are constructed from outlines. In outline fonts, each character is described mathematically as a set of line segments, arcs and curves. A single outline font can be used at any conceivable size and resolution. The designer need
not worry about the point size of the font. Currently, the two most commonly used outline font technologies are *Postscript* and *TrueType*. Their major difference lies in that Postscript uses Bezier curves, while TrueType uses quadratic splines.

5.1.4 Evaluation

Bitmap fonts need more storage space than vector and outline fonts. Vector fonts need the least storage space. Usually 7,000 – 20,000 Chinese characters can be stored in 40 – 60 KB EPROM.

To output a character, bitmap fonts need no additional processing, while vector and outline fonts need some special technique to compose a character. So bitmap fonts are the fastest.

Unlike vector and outline fonts, bitmap fonts are not dependent on the structure of Chinese character.

5.2 Font Generation Methods

There are two ways to generate good looking Chinese fonts for displaying on raster or vector display devices and for printing on dot matrix or plotter printers: vector pattern and dot matrix pattern.

5.2.1 Vector Pattern

Vector generation is based on the fundamental strokes of Chinese character. According to the study by the Chinese Character Analysis Group of Taiwan, at least 34 different fundamental strokes are needed to generate adequate-looking Chinese characters in computer (See Figure 2.3).

Each of these stroke shapes can be simply described using a set of vector endpoints, such as $(x_0, y_0)$ to $(x_1, y_1)$. Else, some mathematics scheme, such as cubic spline curve or square spine curve, can be used to produce good looking strokes.
Theoretically, every Chinese character can be decomposed and described by a set of these fundamental strokes. A font generation system called ACCFONT (Automatic Chinese Character Font) was developed to describe the definition of every stroke for complete character generation by utilizing B-Splines curve [16]. However, this is seldom used in practical implementations because B-splines are computation intensive and require large amounts of extra programming to accelerate display speeds to acceptable levels. B-splines also require huge memory spaces to handle characters with many strokes.

The most common way to implement vector generation method is by defining a set of roots/radicals rather than strokes. One important consideration is how to choose each root/radical. To compose a character, the bigger the root set used, the less the number of roots needed to compose a character and vise versa. One implementation is the Chiao-Tung Radical System, in which a set of 496 roots/radicals were chosen after the analysis of over 49,000 characters from the Kang Xi dictionary. Every Chinese character is defined by a radical equation, which contains their radicals/roots and inter-relational operators. For example, the character ‘後’ (behind) can be described by the equation in Figure 5.3.

\[
\text{後} = \begin{array}{c}
\text{part1} \\
\text{part2} \\
\text{part3}
\end{array}
\begin{array}{c}
\text{Three radicals that composed the character} \\
\text{Horizontal and vertical radical operators} \\
\text{The relative size ratio of each root/radical}
\end{array}
\]

Figure 5.3: An Example of Chiao-Tung Radical System
5.2.2 Dot Matrix Pattern

This method uses a matrix of binary values to describe a character. It is usually done by CAD (Computer Aided Design) system. When a certain character is selected, a single routine will map the indexed dot matrix patterns to the output device, i.e. a displaying Cathode Ray Tube (CRT) or a dot matrix printer. Figure 5.4 shows the dot matrix of character ‘後’ (behind).

![Dot Matrix pattern of Hanzi character ‘behind’](image)

The major sizes of dot matrix are listed below:

- 15 × 16 (or 16×16): the smallest bitmap. The display for characters with more than 8 horizontal or vertical strokes is very poor. It can only be used for low resolution display.

- 24 × 24: The most commonly used bitmap size.

- 32 × 32: better quality than 24×24. for higher printer quality.

- 48 × 48: good quality font. But occupies too much space. It has to be stored in disk. So speed is low.

- 64 × 64 and higher: compression and decompression is needed. It is used in typesetting systems.
When choosing a bitmap size for a printing system, font quality, machine speed and space must be considered.

5.2.3 Evaluation

Both methods discussed above are generally used in computers and have their own advantages as well as disadvantages. Following is a brief evaluation:

1. Memory requirement: The vector generation method requires less memory than the dot matrix method.

2. Flexibility: For vector generated fonts, it is very easy to change the character size by simply applying some simple mathematical operations to the x, y coordinates. However, for dot matrix fonts, character size is relatively inflexible. Changes to the character tend to require huge amounts of extra processing. Also, enlargements of a character from $24 \times 24$ to $96 \times 96$, for example, produce stepwise zigzag results, as described earlier, that are very noticeable.

3. Execution Speed: Vector generation method requires more execution time than dot matrix methods. Using vector patterns, each stroke of the character will require many steps of calculation prior to actual output. On the other hand, using dot matrix pattern could be as simple as moving a stream of data from memory to output devices, hence it is intrinsically faster.

5.3 Font Storage

There are three methods for font storage:

1. Disk Storage: A font can be set as an independent device shared with several devices. This is a cheap but less efficient way.
2. ROM (EPROM) and RAM and Disk: ROM (EPROM) is favored for most frequently used characters, RAM for frequently used characters when reading from disk by DMA upon system initialization. Infrequently used characters can be read from disk as needed.

3. Chinese Card: The principle is the same as for ASCII character generation. It can be used for printer and monitor, but the order for reading fonts is different for printer and screen. For most row major scan systems and laser/ink-jet printers, the bitmap should be row major. However, for column major printers, such as dot matrix printers, it should be column major. Therefore the monitor and laser printer can not share the same font with the dot matrix printer unless an algorithm for converting between row-major and column-major is used.

5.4 Printer Output

No matter what kind of font is used, ultimately every font is resolved into a bitmap. Printers can range from low-resolution, dot-matrix printers to high quality laser printers that contain a Postscript interpreter. Physical and storage limitations tend to restrict laser printers to 300–600 dot per inch. Usually different programs are needed for printing on different printers.

The speed for printing Chinese differs from different devices. For example, the speed on 24 × 24 dot matrix printer is 50/100 characters per second without/with a special font card installed on the printer.

The output printer could be: a dot matrix printer, plotter, hardcopier, electrostatic writing device, digital CRT typesetter, optical fibet tube, or laser writing device (laser comp and laser printer).
5.5 Screen Output

High resolution monitors (CRTs) can display 100 pixels per inch. Usually one screen can display 11–26 lines with 40–80 characters per line depending on the resolution of the screen. The more characters displayed on one screen, the slower the screen switches from page to page. Suppose the speed rate for display is 50 HZ, for displaying a full screen of 25 × 40 characters, $25 \times 40 \times 50 = 50,000$ characters/second are needed.

5.6 Chinese Terminal

A Chinese terminal has communication functions, a RS232 serial port and can be controlled remotely through networks. There are two kinds of Chinese terminals:

1. Ordinary Chinese terminal: Functions include Chinese input/output, Chinese code conversion and Chinese editor.

2. Intelligent Chinese terminal: Functions include Chinese I/O, Code conversion, editor, file management, table processing, and graphic processing. It can also be used as an independent workstation.

The highest resolution Chinese terminal is $1024 \times 1024$. The 53241 Chinese terminal by Wang An Company can display 24 lines with forty $24 \times 24$ bitmap characters per line. The Chinese terminal is still undergoing intensive research and development, especially in the following areas: multi-windows, color displays, multistyles, and miniaturization.
Chapter 6

Chinese Information Processing Techniques

Many people think that input-output methods or coding schemes form the whole picture of Chinese computing. However, without the support of good Chinese language processing techniques, other efforts in Chinese computing research and development will be in vain. In this chapter, information processing techniques are organized into three levels:

1. System level: Chinese operating systems

2. Code level: Code conversion, detection, searching, and sorting

3. Memory level: Memory management

6.1 Chinese Operating Systems

To design a Chinese operating system, the following functions must be guaranteed:

- Reserve original operating system functions for standard ASCII

- Chinese I/O management

- Chinese device drive modules
Chinese operating systems are usually composed of an ASCII operating system, Chinese fonts, and system loaded fonts and management modules. Chinese operating systems are designed for either Big-5 (in Taiwan and Hong Kong) or GB (in Mainland China) code. The only difference is that they use different encoding methods (Big-5 and GB) and fonts (traditional and simplified). Basic operating principles are the same for the two systems.

6.1.1 Chinese Operating Systems on MS-DOS

Japanese researchers have modified Microsoft’s MS-DOS to be able to process Japanese directly. This is more difficult for Chinese systems since Chinese culture and language requires multiple input methods, (for example, Wu Bi method should be provided for professional Hanzi typist and Pin Yin method is required for general user), while the Japanese language requires just one input method. Currently the Chinese Operating Systems are running on top of MS-DOS, other than with modified versions of MS-DOS. Examples include: CCDOS, developed by the Electronic Engineer Industry Department No. 6 of Mainland China; UCDOS, developed by Beijing Hope company; Lian Xiang, developed by Lian Xiang Company; Golden Mountain System, developed by Golden Mountain Company.

CCDOS (Chinese Character Disk Operating System) runs on MS-DOS. It adds Chinese processing function to MS-DOS’s file management system (IBMDOS.COM) and Basic I/O system (BIOS).

‘CCBIOS’ is the kernel of CCDOS. It has modules for Chinese input, Chinese display, Chinese printing, font management, etc. The internal code used in CCDOS is 8-bit GB code. It set the highest bit of each byte of GB2312-80 to ‘1’ in order to distinguish GB code from ASCII code (refer to Section 3.2.1). The following two
parts are loaded into memory while running the following:

1. The Program Area: Mainly CCBIOS, used for replacing the device driver.

2. The Data Area: Fonts and dictionaries

6.1.2 Chinese Operating Systems on MS-Windows

Microsoft Company has announced the Asian versions of Win95 (Chinese, Japanese and Korean) will be out in January 1996. Traditional Chinese Win95 will be available in Nov 1995 according to Microsoft's schedule. Since the core code of the Asian version will be identical to English Win95, all Win95 applications will run without problems on CJK Win95. Besides, Microsoft also have a Chinese Win95 with English interface designed for U.S. developer and users. So it will be a standard Win95 with Chinese capabilities.

6.1.3 Chinese Operating Systems on Unix

To build a Chinese operating system on Unix is more difficult. The Chinese operating system on UNIX usually runs as a Chinese Shell.
Chinese shell 'csh' produces input process 'in' and output process 'out', input pipe 'pipein' and output pipe 'pipeout'. The function of 'in' is reading from keyboard, converting input code to internal code. The result is sent to 'pipein'. 'csh' produce new subprocess 'cld' to process data from 'pipein'. The output of 'csh' and 'cld' is sent to 'pipeout'. 'out' is in charge of processing the output data from 'pipeout'. Since 'csh' is a shell outside UNIX kernel, it reserved the compatibility to UNIX. For more details, refer to [22].

One Chinese Unix/Xenix operating system was developed by Taiwan. It amounted to a single patch up job to process Chinese characters. Currently 'cxtterm' together with 'hztty' is very popular on UNIX for Chinese processing. It can process GB, HZ and Big-5 code. Also, DEC\textsuperscript{1} have localized their DECwindows (Motif based) to provide Chinese environment (both simplified and traditional Chinese).

\textsuperscript{1}The name was changed to Digital.
6.1.4 Chinese Operating Systems on Macintosh

The Macintosh operating environment is the first one that put international language compatibility into the design stage. The operating system for the recently announced Macintosh II font not only handles text from left to right and from right to left, but can also handle multiple bytes per character. In 1987, Apple released two versions of a Chinese operating system, one for simplified characters and one with orthographic for traditional characters.

6.2 Code Conversion

For historical reasons, several different Chinese character set standards were developed (see Section 3.2). Some efforts have been made to make the encoding conventions for these standards more compatible. Conversion between internal code and external code is necessary because Chinese character conversion is not built into the operating systems of many Internet mail or news servers. Conversion has to take place at the user application level. Also, conversion between input code and the internal code is necessary.

6.2.1 GB ↔ HZ Conversion

As mentioned in Chapter 3, HZ is just a variation of GB code. The conversion between them is very simple:

Let: $g_1 =$ 1st byte of GB code, $g_2 =$ 2nd byte of GB code;

$h_1 =$ 1st byte of HZ code, $h_2 =$ 2nd byte of HZ code.

To convert between GB file and HZ file,

Let: $h_1 = g_1 - 0x80$

$h_2 = g_2 - 0x80$

and $\{h_1h_2\} = g_1g_2$
6.2.2 GB ↔ Big-5 Conversion

Since there are more characters in Big-5 than in GB, the conversion is not one-to-one. It is possible that some Big-5 characters can not be matched to GB and a single GB character may be matched to more than one Big-5 character (see Table 6.1) and vice versa. Several n-to-one mappings exist in both directions, but happen more in matching Big-5 to GB. Conversions between GB and Big-5 require a lookup table.

<table>
<thead>
<tr>
<th>GB Form</th>
<th>Big-5 Form1</th>
<th>Big-5 Form2</th>
</tr>
</thead>
<tbody>
<tr>
<td>面</td>
<td>麥</td>
<td>面</td>
</tr>
<tr>
<td>flour, face</td>
<td>flour</td>
<td>face</td>
</tr>
</tbody>
</table>

Table 6.1: Multi-mapping from GB to Big-5

6.2.3 GB/Big-5 ↔ Unicode

Hanzi characters in Level One and Two are designed as 21xxxx and 22xxxx, respectively, in the Unicode mapping tables, where xxxx is a four-hexadecimal-digit in the range of 2121–7E7E.

6.2.4 Input code → Internal code

A Chinese system should support various input methods. It is important to design a general purpose 'input code' to 'internal code' conversion module. There are two major methods:

1. Static method: For each input method, save its mapping table of input codes and internal codes on disk or ROM. When a certain input method is chosen, the corresponding table is loaded into memory.

2. Active type: A general purpose index program and dictionary program are provided by the conversion model. A system sets up the object dictionary
automatically and saves it on disk according to the user defined code table. When a certain input method is chosen, the system loads the corresponding object dictionary into memory. These dictionaries are huge. In order to save space, compression is used. Usually a 5-byte input code (1-5 ASCII characters) can be compressed to three byte.

6.3 Chinese Code Detection

Following is an example of how to detect GB code in a file.

/**** Function: Detect if a 2-byte character is in the range of GB code
**** c_hi : high byte of the character
**** c_low: low byte of the character
**** return: 0 not in the GB range
**** 1 in the GB range
****/*

int IsGB(c_hi, c_low)

int c_hi, c_low;
{
    if (((c_hi>=0xAl)&&(c_hi<=0xFA))
        /* check high byte */
        (c_low>=0xAl)&&(c_low<=0xFE))
        /* check low byte */
    return 1; /* GB code */
    else
    return 0; /* non GB code */
}
6.4 Character Search

6.4.1 Search in font

How do you search for a single character in a font? Direct search is inefficient for large font sizes. Usually not every character of the font is loaded into ROM or RAM due to limited memory space, although this is becoming more practical as memory sizes increase. Therefore the memory resident font is typically modified to handle such conditions more efficiently. Following are three algorithms:

1. First in First out (FIFO): When replacement from slow to fast memory happens, the first loaded page of the font set is taken out of fast memory. This is a simple algorithm, but not very efficient. Sometimes even frequently used pages are taken out of fast memory if they were loaded first from slow memory.

2. Least Recently Used (LRU): This is an improvement of FIFO algorithm. When the replacement happens, the least recently used page is taken out instead of the first loaded page.

3. Least Frequently Used (LFU): This is another way of improving FIFO algorithm. The least frequently used page is taken out when the replacement happens.

6.4.2 Search in a file

The ‘grep’ (Global Regular Expression Printer) utility is one of the most commonly used utilities on UNIX. It performs searches based on pattern matching. Some versions of ‘grep’ handle multiple-byte characters. However, there is a problem in this: the current index into a file must either advance one or two bytes depending on whether the current position is for a one-byte or two-byte character. This means that
mismatches can be made on the second byte of a two-byte character plus, extending to the first byte of the next character. An example is given in Table 6.2.

| Search String | D6D0CEC4D0CECABD (中文形式) |
| Search Character | D0CE (形) |
| Correct Match | D6D0CEC4D0CECABD |
| Incorrect Match | D6D0CEC4D0CECABD |

Table 6.2: Incorrect Match Made by 'grep'

### 6.5 Character Sorting

For English, sorting is simple. The alphabetical sequence is the only indexing method. For Chinese, there are many ways to sort Chinese data: by pronunciation, by internal code, by radical, by the total number of strokes, etc. The implementation of these various sorting methods is limited to the database of available information. For example, name lists are usually sorted by total number of strokes, while characters in dictionaries are sorted by pronunciation. In computers, the easiest sorting approach is by internal code.

### 6.6 Memory Management

Memory is always a big concern in writing Chinese programs. Two items consume a lot of memory: the dictionary and the font.

#### 6.6.1 Memory Management for Dictionary

The minimum memory required by a decent input method is approximately 70K. A typical input sequence of one to five key length is normally compressed to a 4-byte long integer (3-byte is seen in some systems to save memory space).
Following is a simple calculation for the memory required by the GB code system. As we learned in Chapter 3, the GB internal code of a dictionary will be no longer than \(94 \times 94 = 8,836\) individual codes. If only Hanzi characters are considered, then 6768 codes (or total of \(2 \times 6768\) bytes) are used by the dictionary. Suppose the input code is compressed to three bytes, then the total size is \((3 \times 6768) + (2 \times 6768) = 33,840\) bytes.

To search the dictionary efficiently, a binary tree can be implemented. Suppose each node of the tree contains three 16-bit pointers (i.e. left, right, itself), then the size of the binary tree will be \(6768 \times 3 \times 2 = 40K\) bytes. So the minimum size for the dictionary is \(40K + 30K = 70K\).

Of course, if other 'intelligent' input methods (e.g., 'phrase association') are used, the memory required will be much more than 70K. Memory required will also be greater where code space is greater than 6768, e.g., with the larger fonts discussed earlier. Extra space is needed for the 'phrases'. The memory space can also be decreased through use of a better dictionary searching algorithm.

6.6.2 Font Buffer

A Font buffer is a memory block in which the font data resides. How much memory should be allocated for font buffer? With the Dynamic Font Buffer (DFB), the size of the memory required depends on the number of character codes used. Statistics shows that a typical Chinese article uses only about 500–1,000 character codes. Assuming use of a \(16 \times 16\) bitmap font, \(\text{DFB} = 1000 \times 16 \times 2 = 32K\) bytes are required.

A table of pointers is needed to match the character code to the corresponding font address. Using the simple memory model, where each character code is used as an index to a font address, the table size = \(94 \times 94 \times 4 = 35K\) bytes (for 32-bit pointers in the GB case). This can be improved by using a segmented memory approach, where two tables are maintained. The first table contains ninety-four 32-bit pointers. The
second has $94 \times 94$ 8-bit offsets. The font address is calculated as follow:

\[
\text{Font address} = 32\text{-bit-pointer } [\text{HI byte of GB code}] \\
+ (8\text{-bit-offset } [\text{GB code}]) \times \text{size of bitmap}
\]

Therefore, the size of the tables $= 94\times4 + 94 \times 94 = 9K$ bytes.

### 6.7 Compression Techniques

For ordinary bitmap fonts (i.e. 24x24), each character can be represented by 72 bytes. The memory needed is less than 720KB for 10,000 Chinese characters. Although tolerable, this is still very large for present devices. On the other hand, for high quality fonts, such as 64x64, nearly 4MB memory is needed for 10,000 characters. In such cases obviously, compression techniques are needed.

Using outline or vector approaches to describing character fonts can save ten times space of bitmap fonts. The most commonly used compress technique is called black and white method. It records the number of black or white points line by line:

rowmark of repeated line][white][black]...[white][black]  
* \(N\ W1\ B1\ ...\ Wn\ Bn\)

The average compression rate by this method is about 24 times.

Due to the fact that there are many slanted lines in Chinese characters, another method called ‘ZengLiang’ compression method was invented based on black and white method. The average compression rate for ZengLiang can be increased to 200-fold. For details about this method, refer to [46].
Chapter 7

Chinese Applications

The development of Chinese tools makes computerized Chinese Information Processing possible.

7.1 Chinese text processing systems

7.1.1 General Chinese text processing system

This is commonly used in daily life. The output can be a $24 \times 24$ dot matrix printer or laser printer. More complex $24 \times 24$, $32 \times 32$ or $48 \times 48$ bitmap fonts, or TrueType fonts are also typically provided.

Following is a list of several popular general Chinese text processing systems:

- **WS (WordStar):** This is a modified version of WordStar. This needs to be run on CCDOS. It has three components: WS.COM, WSMGS.OVR and WSOVLY1.OVR. The instructions provided by WSMGS.OVR are translated into Chinese. Since Chinese characters are treated as two 1-byte characters, errors such as one Chinese character being cut by EOL can happen.

- **WPS (Word Processing System):** Needs CCDOS. This is a strong and popular edit/print system in Mainland China. The potential errors mentioned for WS are avoided. If a Chinese card is used, 256KB memory can be saved.
- **Eten**: Developed by Eten Company in Taiwan. It supports Big-5 code.

- **Poorman's latex**: A share ware that runs as a patch of Latex compiler. Chinese GB code or Big-5 code can be used directly in latex files.

### 7.1.2 Chinese Typesetting/Editing system

The first Desk Top Publishing (DTP) system was that used for publishing the United Daily News in Taiwan in 1982. It established a standardized Chinese character set and a Chinese text editing module. This module provides various editing, proofreading, and re-editing capabilities.

The strength of these systems is the storage and output of high quality fonts. Since many typefaces and type styles are used, the bitmap for each character can be very large. The output device is usually laser printer. Examples include:

- Hua Guang Publishing System (Model IV).
- Fang Zheng System (by Beijing University).
- 4-S Super Science Setting System (by Si Tong Company).
- Large screen publishing system for newspapers and magazines.
- M-6403 Desktop publish system: This needs the M-6403 Chinese card and runs on CCDOS. Functions are: typesetting, edit, graphic and table make, character make (128 x 128), and virus detect.

### 7.2 Chinese programming languages

Many applications are patch jobs from existing English microcomputing programs. Examples include:

- Chinese DBMS
• Dai-E Chinese FORTH
• Chinese COBOL
• Chinese FORTRAN
• Chinese BASIC
• Chinese LOGO

7.3 Chinese on Internet

7.3.1 Using Chinese E-mail and News

The improvement of Chinese encoding, (especially the implementation of HZ encoding method) and the development of Chinese tools, makes it possible to use E-mail and News in Chinese as easily as in English.

7.3.1.1 Chinese on Internet/Bitnet

All Sun Systems (SunOS, SunView, OpenWindows, Solaries...) have had CJK (Chinese, Japanese and Korean) support for years. The first version (SunOS 4.0.3) was released in early 1990. On all SunOS 4.1.1 and above systems, nearly all programming utilities (vi, ed, sed, grep, awk, etc) and communication software (tip, ftp, telnet, rlogin, etc) are made 8-bit clean. The same situation is also true for current workstations based on UNIX SysV Release 4 (IBM, DEC, HP, etc).

However, there are still many older UNIX machines that are not 8-bit clean, e.g., machines where the kernel TTY driver does not provide for an 8-bit mode and where UUCP news feed is used. E-mail has more problems. Many sites do not give users any way to mail 8-bit text. The solution is using uuencode or MIME.

For more information, call 1-800-USA-4SUN.
Because of the 7-bit limitation of Usenet, files in 8-bit Chinese code must be converted to the 7-bit representation. There are two frequently used approaches:

1. uuencode/uudecode
2. conversion to HZ code

'uuencode/uudecode' has no official standard and varies from vendor to vendor. It can get corrupted in certain mail gateways. To solve this, BASE 64 encoding was invented, standardized and implemented in MIME.

7.3.1.2 Sending/Receiving Chinese E-mail

If a file is in 8-bit GB or Big-5 code, then it has to be treated as a binary file. It has to be "uuencode"d before sending and "uudecode"d after receiving. If a file is in HZ code or 7-bit pure GB code, then it is compatible with ASCII file. Thus it can be sent directly.

To view/edit/print a file, try different tools according to the computer systems and the file code used. Various tools for DOS, UNIX (X-windows), Macintosh, and MS-Windows can be obtained by anonymous FTP. The addresses on Internet are ftp.ifcss.org, cnd.org[132.249.229.100] and nctuccca.edu.tw[140.111.3.21].

7.3.1.3 Posting/Reading Chinese News

The Chinese newsgroup on Usenet is alt.chinese.text. This group uses the HZ encoding method. You can post HZ files and read it using ZWDOS, CXTERM/Hzty, etc.

For Big-5 user, there is a mirror group called: alt.chinese.text.big-5. You can post Big-5 files to this group, but must make sure that the system you are using is 8-bit clean.

The imperfect mirroring between alt.chinese.text and alt.chinese.text.big5 is unavoidable due to the inaccuracy of conversions between HZ and Big-5. Some other
Chinese posting newsgroups are:
  chinese.comp.software
chinese.flame
chinese.newsgroups.announce
chinese.newsgroups.answers
chinese.newsgroups.discussion
chinese.newsgroups.newusers
chinese.rec.misc
chinese.rec.sports
chinese.talk.misc
chinese.talk.politics
chinese.test
chinese.text.Unicode

7.3.1.4 Others

A few precautions should be taken before transmitting a Chinese file through an ASCII computer system:

- One Chinese character is two bytes, and occupies two columns.

- Break long lines to fewer than 80 columns. If you are using HZ, please notice that each escape-in and escape-out sequence occupies two columns.

- Do not break one character (2-byte) into two 1-byte parts (by <SPACE>, <RETURN>, etc).

There are many Chinese articles, books and software available on Internet through anonymous FTP. Also, there are more than twenty electronic Chinese magazines that
can be subscribed to through Internet. Chinese is also on World-Wide-Web now. In order to read Chinese on WWW, Chinese environment has to be installed first. Browsers, such as 'lynx', 'Netscape' or 'Mosaic', are executed under the Chinese environment. If the WWW is running under X-windows, special settings are needed for browsers.

7.4 Sending Chinese FAX

Chinese text files cannot be sent directly by FAX. They have to be converted to an image file using special software. A program called gb2pic.exe can convert GB files to three popular image formats: TIFF, PCX and BMP. It can be obtained by anonymous ftp from:

ftp.ifcss.org: //pub/software/dos/print/qcsprint.txt qcsprint.zip

7.5 Machine Translation

The earliest machine translation was the translating of Russian to Chinese. The Chinese code used is a 4-digit telegraph code.

In 1992, an Intelligent English-Chinese Translation System, IMT/EC863, was developed by Zhaoxiong Chen. It is based on the theory he presented at the 12th International Conference on Linguistics in 1988.

7.6 Processing Chinese History and Classical literature

One of the most difficult challenges for sinologists is how to locate a piece of important information from the countless volumes of Chinese historical and classical literature. The earliest record of the collecting of classical books was from the Qing Dynasty. A comprehensive collection was published. It is called 'Si Ku Quan Shu'. There are
about one billion characters in those 3,460 volumes of 79,339 books. Although only half of them remain today, they still contain invaluable information for sinologists. However, searching for some types information in this data base by hand would be very difficult. To solve this problem, the answer is to computerize the data base. Kyoto University of Japan, Research Institute for Humanistic Studies is undertaking a computerization of Chinese classics [21]. They have finished inputting Kang Xi dictionary (49,188 characters), Fan Nan Wen Ji (500,000 characters) into a computerized format. At the Academia Sinica of Taiwan, a full text processing system is under development. The object document in the full-text database consists of the Shi Huo Zhi and twenty-four Dynasties (a total of 559,820 characters). This system provides both a controlled vocabulary search and free-text search for that text database.

7.7 Chinese Library Information Network

HyperCard is a microcomputer usage environment introduced by Apple Computer. Users can easily navigate through the intertwined HyperCard stacks which contain the desired information in a multimedia form. Some HyperCard stacks, such as learning, writing, speaking Chinese and Chinese dictionaries are under development [27].
Chapter 8

Summary

There are so many interesting topics covered in this paper. However, due to the time limitation and the lack of reference information, some of the areas, such as Chinese on World-Wide-Web, new Chinese output methods, Chinese on MS-Windows, and new Chinese applications are still not fully addressed.

There are also lots of challenges in the Chinese Information Processing area, listed as following:

- The open Chinese character set makes Chinese encoding difficult. No character set existing now can satisfy all kinds of implementation.

- Input speed is still very slow compared to English one.

- There are lots of culture differences exist between Mainland China and Taiwan, not only in the character forms (simplified vs traditional), but also in the phrase habits.

- Unlike Western languages, where alphabetic order is the only sorting order, Chinese characters have more than one sorting orders, usually depends on user implementation.

- Communication on Networks requires 8-bit clean, otherwise, special tools must be used.
Bibliography


