30

A)

CED -- A MACHINE READABLE CHINESE DICTIONARY

Shih-Shyeng Tseng*, Meng-Yuan Chang*, Ching-Chun Hsieh**, Keh-Jiann Chen**
*Computing Center, Academia Sinica
**Institute of Information Science, Academia Sinica

Keywords: Textual database, Chinese electronic dictionary, Chinese information processing.

Abstract: A Chinese dictionary stored in computer memory and retrieved by a set of software is called a Chinese electronic dictionary. The CED is an experimental Chinese electronic dictionary developed by the authors at the Computing Center of Academia Sinica, Taipei, ROC.

The development of CED is intended for building a set of machine readable tools for various Chinese information processing applications. Additionally, CED also provides users with available information of Chinese vocabularies, so that the users will get an easy learn, easy use, and friendly retrieval system.

The CED currently contains about 40,000 Chinese vocabularies. It provides two kinds of information of a vocabulary. The first kindof information consists of Chinese character strings including the vocabulary, the pronounciations and the meaning descriptions of the vocabulary. They are obtained from the Gwoyeu Ryhbaw Tsyrdean. The second kind of information consists of syntactic categories of the vocabulary and annexed attributes such as preceded measures for nouns, case frames for verbs, etc. The latter is obtained through a current research work accomplished by our research team.

The software tools for the CED includes a parsing program, a user interface, and a number of management programs. The parsing program scans the source text files to recognize all text elements, and translates those source text files into a set of formatted machine readable data files. The management programs assist researchers to insert, delete or modify the information to/from the data files within the CED. In additional, they also provide a data retrieving mechanism to the user interface. The user interface consists of a set of programs which can help the users retrieve any information of the CED.

In this paper, the design concept and the organization of CED will be presented. The applications of CED will be given by some examples. The furture development of CED will be also discussed.

1. INTRODUCTION

In Taiwan, the researchs on Chinese information processing, which included the theoretical studies and the applicational developments, had been proceeded more than 15 years [1]. During the past years, most of the theoretical studies on Chinese information processing concerned Chinese characters, such as the studies and implementations of Chinese character sets[2,3], the coding techniques of Chinese characters [4,5], the input methods and output mechanisms of Chinese characters[6],etc. Since the Chinese words (vocabularies) are the smallest meaningful units of the Chinese language, all the Chinese informaiton processing systems based on the Chinese characters can only process the Chinese information as a number of character strings. We need more theoretical studies on Chinese vocabularies to develop the more intelligent Chinese information processing systems. For that reason, the researchers of both Computing Center of Academia Sinica and ERSO (the Electronic Research and Service Organization) have started a cooperative project to study the syntax and semantics of the most-frequently-used Chinese vocabularies. The development of CED is basically intended for buliding a computerized tool to assist our research work.

Because the Gwoyeu Ryhbaw Tsyrdean is one of the most popular Chinese dictionaries in Taiwan, and it collected a large valume of most-frequently-used Chinese vocabularies used in Taiwan. We chose it as the source of the Chinese vocabularies to our study. The CED currently contains about 40,000 Chinese vocabularies. It provides two kinds of information of one word. One kind of information consists of Chinese character strings including the vocabulary, the pronounciations and meaning descriptions of the vocabulary. They are obtained from the Gwoyeu Ryhbaw Tsyrdean. The other kind of information consists of syntactic categories of the vocabulary and annexed attributes such as preceded measures for nouns, case frames for verbs, etc. The latter is obtained through a current research work accomplished by our research team.

In addition to play the role of the computerized tool to our researchers, the CED has potentially various applications. The first, CED provides users with available information of Chinese vocabularies, so that the users will get an easy learn, easy use, and friendly retrieval system. That is, the CED can be regarded as a Chinese dictionary provided by the computer system. The second, CED can support a useful basis for the researchs on Chinese computational linguistics, Chinese artificial intelligence, and Chinese office automation.

2. THE ORGANIZATION OF CED

The CED is composed of a set of MD files and index files, a number of management programs, and a user interface, as shown in fig. 1. In CED, there are a set of machine readable data files called MD files which store up about 40,000 Chinene vocabularies. Each MD file consisted of a variable number of formatted records. The information of one Chinese word, which had been introduced in section 1, was stored up in a record. Those records, which were stored up in MD files, corresponded with the orderly vocabularies of Gwoyeu Ryhbow Tsyrdean. The present storage space of MD files are more than 3.1M bytes, and that storage space are increasing with our research.

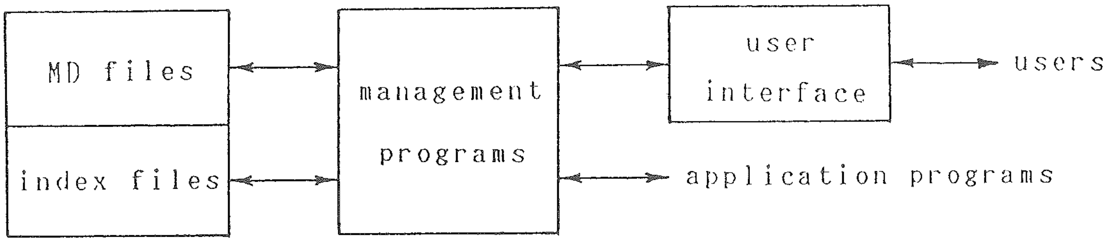


Fig. 1 The Organization of CED

The CED also contained a set of index files. Those index files included various indices which provided a group of accessing links from various items, such as pronunciations, radix and stroke counts, and syntactic categories, etc., to all of the corresponding records. A program can retrieve any necessary information fast and efficiently through those indices.

The CED provided a number of management programs as the medium between application programs and MD files (the user interface can be regarded as one of application programs). The managemet programs consisted of a set of maintenance programs and data-retrieving programs. They are actually a group of subroutines written by C programming language, in which each subroutine performed a specified operation. The maintenance subroutines can be called by any application program to perform the operations of insertion, deletion and modification in MD files. The data-retrieving subroutines can retrieve a group of records corresponding to a given item. Any additional subroutines can be added to CED if necessary.

The last part of present CED is a user interface. It provided a user-friendly menu-driven mechanism to help the users retrieve any information of the CED. The user interface also provided the operations of insertion, deletion and modification, but those operations just allowed for our researchers. This constraint is necessary to protect the CED from some misuses made by the end users.

3. THE CREATION OF MD FILES

The source text of Gwoyeu Ryhbaw Tsyrdean contains about 40,000 paragraphs in which each includes the information of one Chinese vocabalary. The MD files might be built on the source text by means of a traditional procedure. That is, a group of researchers laboriously analyzed each paragraph to obtain a set of elements of the vocabulary, such as the radix and stroke count of the first character, pronounciations, and meaning descriptions. The researchers also had to write those elements into a pre-printed form. Then the data in the filled forms should be entened into a computer system by a group of typists to create MD files. Since the Gwoyeu Ryhbaw Tsyrdean contains more than 1.6 million Chinese characters, the traditional method will lead a time-consuming and laborious work.

Instead of the traditional procedure, an ingenious method had been applied to the creation of MD files. Principally, the logical organization of the source text will be explictly represented by its typesetting format. As an example, fig. 2 shows the logical organization of Gwoyeu Ryhbaw Tsyrdean, and fig. 3 shows a piece of the dictionary. When a Chinese looks for a word in a Chinese dictionary, the logical organization as fig. 2 will occur to his mind. The fact leads the idea that the text elements (e.g. elements described in previous paragraph) will be recognized by a program if the source texts with associated information of typesetting formats had been stored in computer files.

The new method which used to build the MD files on the source text is composed of two steps:

(1) The source text and associated information of typesetting format were entered into computer files by a group of typists. As an example, there are several different fonts of Chinese characters appeared in fig. 3. In order to represent those different fonts, some markup symbols had been used to enclose the characters whose appeared in each of text element of radices, stroke counts and vocabularies.

(2) A program called CED parser scaned those source text files to recognize every text element and then to create the MD files. In order to recognize the text elements and then to construct each formatted data record as result, a pattern-matching and solt-filling algorithm had been applied to the CED parser.

Since the operations of CED parser replaced the work of researchers who analyzed the element and filled them into pre-printed form, the new method was faster and more effictent than the traditional method. It took about three man-month to design and implement the CED parser.

4. THE USER INTERFACE AND THE MANAGEMENT PROGRAMS

As that had been described in section 1, the CED essentially played the role of a computerized tool to our researchers. So we need an interactive user interface to assist us to retrieve and sometimes to list out any information from MD files and moreover to maintain the contents of MD files if we want to. The user interface served a menu-driven operation for the users. This menu-driven operation were looked like it appeared in some packages of databases. The menu should be displayed in a terminal whenever the user interface activated. The user can choose one of the functions such as retrieval, listing, insertion, deletion and modification from the menu. Then the user interface responds and the user has to enter some data with the key board to accomplish this function step by step. For example, if we need to add a new syntactic category to a vocabulary. We first choose the function of modification from the menu, and then the user interface will ask which vocabulary we want. We have to locate the vocabulary we need by specifying some retrieving key(s). When the object appeared in the screen, we then identify the item of category. Following that, we select some descriptions of meanings which correspond to the new category and enter the new category and annexed attributes. After those steps, the informaton of the new syntactic category should be added into the vocabulary.

All of the elementary operations involved in the user interface are performed by a set of management programs. They are actually a set of

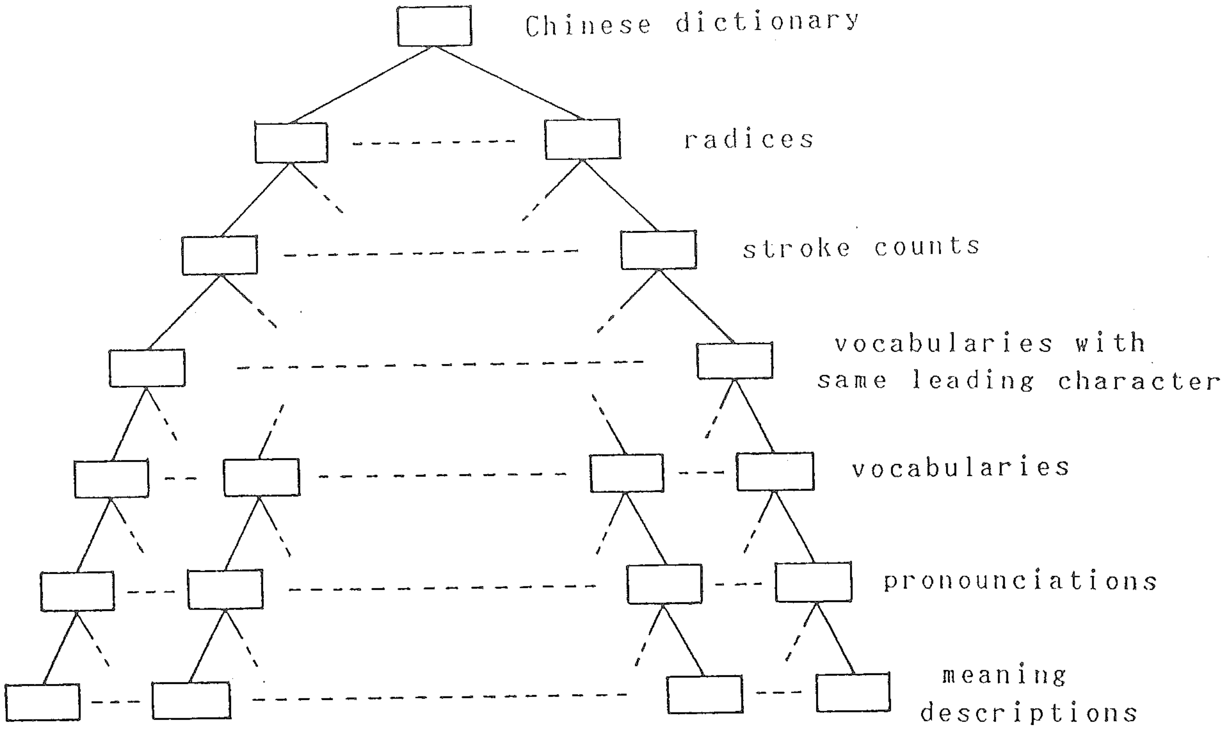
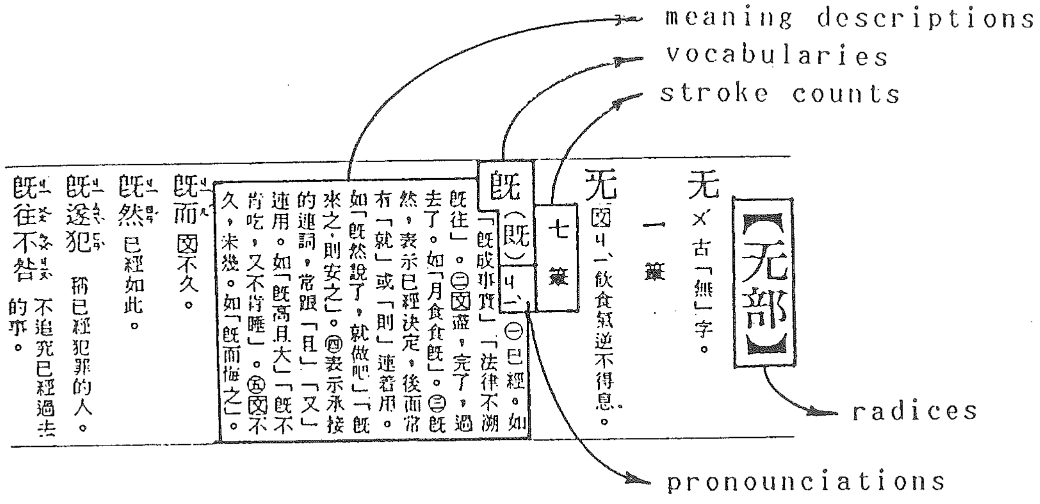


Fig. 2 The Logical Organization of Traditional Chinese Dictionary



【无部】

无 ㄨˊ 古「無」字。

一畫

无 因ㄐㄧˋ 飲食氣逆不得息。

七畫

既（既）ㄐㄧˋ ㊀已經。如「既成事實」「法律不溯既往」。㊁因盡，完了，過去了。如「月食食既」。㊂既然，表示已經決定，後面常有「就」或「則」連着用。如「既然說了，就做吧」「既來之，則安之」。㊃表示承接的連詞，常跟「且」「又」連用。如「既高且大」「既不肯吃，又不肯睡」。㊄因不久，未幾。如「既而悔之」。

既而ㄐㄧˋ ㄦˊ 因不久。

既然ㄐㄧˋ ㄖㄢˊ 已經如此。

既遂犯ㄐㄧˋ ㄙㄨㄟˋ ㄈㄢˋ 稱已經犯罪的人。

既往不咎ㄐㄧˋ ㄨㄤˇ ㄅㄨˋ ㄐㄧㄡˋ 不追究已經過去的事。

Fig. 3 A Sample Piece of Gwoyeu Ryhbaw Tsyrdean

subroutines, and the user interface can invoke them by means of a series of subroutine calls. This manner provided a wide feasibility to build up another application program in future.

5. CONCLUSION

The CED has some potential applications, as dicussed in section 1. In this section, those applications should be explained more detail.

(1) A computerized dictionary which developed for the users is principally a read-only database of vocabularies. Since the total storage space of CED can be limited to smaller than 5 million bytes and the user interface of CED includs a easy learn, easy use, and friendly retrieval mechanism, so the CED will be easily developed into a commercial product on some personal computers.

(2) Each of the information processing systems which are concerned with natural languages, such as machine translators, natural language query processors, expert systems, etc., has to include a dictionary of vocabularies. Since the CED contains about 40,000 Chinese vocabularies and associated information and a set of management programs, so it has ability to provide the necessary basis for various Chinese information processing systems which are concerned with Chinese language.

An important feature of CED is that it is an open system. That is, any new information for some vocabularies and moreover any new vocabularies with associated information can be easily added into the CED. Reversely, a subset of vocabularies with optional a subset of information of those vocabularies can be easily abstracted from the CED to serve various applications.

In additional, the experience with the design and development of the CED is very useful to the development of information processing systems. For example, how to build up the data files of the textual database from a large volume of source texts is an important but difficult problem. The experience in the project on the automation of Chinese History Literatures which is another research project currently worked by researchers at Academia Sinica, as well as the experience in the development of CED had lead us to develop a new technique to solve this problem.

※ REFERENCES ※

1. Tseng, shih-shyeng, << The Design of Chiese Character Characteristic Data Base (CCCDB) >>, ch.1, Master Thesis, 1982, NTIT, ROC.
2. Lin, shuh, << A Statistical Study on Chinese Character Set for Computer Uses >>, NCTU Technical Report CC-601, Mar. 29, 1972, NCTU, ROC.
3. The Chinese Character Analysis Group, << Symbol and Character Tables of CCCII >>, May 1983, Taipei, ROC.
4. Hsieh, ching-chun, et.al., "The Design and Application of the Chinese Character Code for Information Interchange (CCCII)", International Workshop on Chinese Library Automation, Feb. 14-19, 1981, Taipei, ROC.
5. Tseng, shih-shyeng, et.al., "An Universal Coding System for Multi-lingual Environment", 52nd IFLA General Conference, Aug. 24-29, Tokyo, Japan.
6. 資訊工業策進會，<<中文電腦發展調查分析報告(二)>>，技術通報C18號，Jul. 10, 1982, Taipei, ROC.