An automatic indexing experiment in Chinese is described. The first very large volume of modern Chinese concordances (two sets of one million-line KWIC index) has been compiled and materialized automatically with a modified kanji printer for Japanese.

INTRODUCTION

This paper describes an experiment to compile Chinese concordances automatically. A very large volume of KWIC indexes for modern Chinese (one million lines per set) has been compiled successfully with a kanji printer for Japanese. This paper discusses the purposes of the experiment, selection and input of the Chinese data, some statistics on Chinese characters (vs. kanji) and the concordance compilation process. Finally, examples from the computer-generated concordances are shown.

THE PURPOSES

The idea of machine-processing modern Chinese data originally came from Professor Yuen Ren Chao, Agassiz Professor Emeritus of Oriental Languages at the University of California at Berkeley, before one of the authors (Hashimoto) took over the directorship of the Princeton Chinese linguistics project. Chao served as the chief of the advisory committee to the project since its foundation. The idea, in short, was: so much has been said about the Chinese pai-hua-wen -- a written language of modern China -- yet nobody has ever clarified what it really was, i.e.; what the basic vocabulary was, what the major syntactic structure was, etc.: in other words the every detail of the reality of pai-hua-wen. Certain quantitative surveys were done before us, but even the most extensive one in those days was based on data consisting of no more than 100,000 characters. In addition, the selection was very poorly done -- most of the materials were primary school textbooks. We did not believe that school textbooks reflected the reality of the language, even in its written form. We chose one digit more than the previous one, namely 1,000,000 characters, though for various reasons, the actual data contained in our tape include several thousands more than one million [1, 2].

After completion of the computer input and editing of the million-character file at Princeton, researches towards statistical aspects of the data have been conducted [4]. As stated in [4], tables of character frequency can tell us various aspects of the Chinese, such as the basic character set, transient states of character strings and so on. This can be summarized as the first step of computer-processing modern Chinese data. However, in order to understand the reality of a language, besides statistics, concordances are the necessities which illustrate the contexts where and how those characters are used.

On the other hand, computer applications to Chinese have very limited background so far. No computer-generated concordances on Chinese have been reported yet. Thus the concordance generation project would not only be valuable to the understanding of Chinese pai-hua-wen, but also contribute to the development of the methodology to manipulate Chinese automatically. Consequently, a project to compile concordances of the Princeton million-character file was conducted at the Electrotechnical Laboratory during 1977-1979. This constitutes the second important stage of computer-processing modern Chinese.
disregarding the position of a radical within a character, so that the results were often a hopeless mess.

The Selection of the Data

It was tried, at the selection of the data, to cover every conceivable category and style of writings in China since her modernization, the so-called May 5 Movement period, from ordinary novels to philosophical writings, from political speeches to newspaper articles, etc. etc. These categories and styles were classified and were assigned appropriate marks to show the genre. The partial list of these writings follow:

For a complete list of all these writings and of the genre marks, see [3]. All the proper nouns were also marked, as they may not correctly contribute to any statistical measurement of the written language except for these proper nouns themselves. These nouns were marked in the original texts by research assistants with enough command of the language to make correct judgment. Anything else, including punctuation marks of all sorts, in the texts were properly processed. Every sentence, including some vocative phrases, was numbered within the writing piece quite mechanically, though occasionally it was necessary for specialists to make certain judgment for segmenting sentences.

The Code System

The Chinese standard telegraphic code system includes some 9500 codes for Chinese characters. A code consists of a set of 4 digits, which represents one Chinese character. Among those 9500, 5231 have been used.

Statistics

Statistical analysis of this million-character file can be found in [4]. Some additional statistics are provided here. Fig. 1 shows the 10 most frequently used characters with their frequencies. These 10 characters occupy 17.1% of the total amount. Fig. 2 is a table of character frequencies vs. the number of character types. Fig. 3 shows the cumulative percentage of character occurrences as a function of the number of character types (in descending order of frequency). It indicates, for example, only 92 characters represent 47% of the entire data. There are 1170 characters each of which are used more than 100 times and they occupy 92.8% of the whole data.

| Character | Frequency |
|-----------|-----------|
| 的         | 46531     |
| 一         | 18077     |
| 是         | 17874     |
| 了         | 16390     |
| 我         | 16138     |
| 在         | 12827     |
| 人         | 11096     |
| 有         | 11057     |
| 他         | 10727     |
| 们         | 10332     |

Fig. 1. List of High Frequency Characters

| Frequency | No. of Character Types |
|-----------|-----------------------|
| - 10001   | 10                    |
| 10000 - 5001 | 13                  |
| 5000 - 3001 | 32                   |
| 3000 - 2001 | 37                   |
| 2000 - 1001 | 106                  |
| 1000 - 501  | 176                  |
| 500 - 301   | 208                  |
| 300 - 201   | 191                  |
| 200 - 101   | 397                  |
| 100 - 81    | 150                  |
| 80 - 61     | 230                  |
| 60 - 41     | 294                  |
| 40 - 21     | 574                  |
| 20 - 11     | 563                  |
| 10 - 1      | 2250                 |

Fig. 2. Frequency Distribution of Chinese Character Types
CHINESE CHARACTERS VS. KANJI

Chinese characters were imported into Japan sometime in the 5th century. Since then, they have been extensively used with a few additional characters created in Japan (this modified set of Chinese characters is called "kanji"), although hiragana and katakana (two sets of pure Japanese characters with their origin also in the forms of Chinese characters) were invented early in the 9th century.

"Chinese characters for daily use" established by the Ministry of Education for modern Japanese includes a 1850 kanji set, however several thousand more are still in use especially for proper nouns. The Japanese Industrial Standard (JIS) "Code of the Japanese Graphic Character Set for Information Exchange (C6226)" established in 1978 includes a 6349 kanji set, hiragana, katakana, Roman alphabet, Greek letters, Russian letters and other symbols. The kanji set is grouped into 2 levels, the first level a 2965 kanji set and the second level a 3384 kanji set. This means some 3000 kanji are considered to be enough for basic information exchange in Japanese. In this experiment, the kanji printer system T4100 (Syowa Zyoho, Co., Ltd.) was used. A total of 8182 characters was available for this printer including 7560 kanji, hiragana, katakana, Roman alphabet, and other miscellaneous symbols. The system was developed 5 years before the establishment of JIS C6226.

As mentioned before, the million-character file included 5231 different Chinese characters. Among them, 295 were found to be unprintable (because they were not found in the T4100 system). The fonts of those 295 characters were designed and incorporated into the T4100 system. Later, when JIS C6226 was established, some of those 295 characters were found in the second level of the kanji set, namely 1 (frequency 773), 2 (581), 3 (563), 4 (345), 5 (343), 6 (189), 7 (178), and 8 (158). Fig. 4 shows the frequency of the remaining 287 characters. Their total frequency numbers 1100, which is 0.1% of the million-character file. This fact indicates that Chinese characters and kanji still overlap closely in modern Chinese and Japanese. (It should be noticed that the simplified Chinese characters are out of this scope since they did not exist at the so-called May 5 Movement period.)

THE CONCORDANCES

Besides the text itself, the Princeton million-character file contained information on the title, the author, the sentence numbers, and other miscellaneous editorial symbols (such as

| Frequency | No. of Character Types |
|-----------|------------------------|
| 554       | 1 ( )                  |
| 228       | 1 ( )                  |
| 134       | 1 ( )                  |
| 128       | 1 ( )                  |
| 100 - 51  | 7                      |
| 50 - 31   | 7                      |
| 30 - 21   | 8                      |
| 20 - 11   | 21                     |
| 10 - 5    | 37                     |
| 4         | 11                     |
| 3         | 34                     |
| 2         | 41                     |
| 1         | 117                    |

Fig. 4. Frequency Distribution of Chinese Characters which are not Found in the Kanji Set
marks to indicate proper nouns). Extensive work had to be done to interpret and reform editorial symbols. Fig. 5 shows the edited text sentences from the million-character file. After this editorial step and incorporation of Chinese character fonts to the T4100 kanji printing system, the concordance compilation process was started. Since we have had experience with the automatic compilation of one-million line concordances in Japanese [5], not many technical difficulties were encountered, except some malfunctions of our old kanji printer. Discussions on the salient features of those Chinese concordances follow.

Key Words

KWIC index style has been adopted as the form of Chinese concordances, since it is one of the most fundamental styles for computer-generated concordances. Because there is no clear segmentation of words in Chinese, and because one character represents a fairly sizable amount of information, each character was chosen as a "key word". Furthermore, no elimination of "non-key words" were made. Every character (including punctuation) was chosen as a key character. In this sense, the concordance may be named as "All characters in context" index. Consequently, one million character data required one million lines of index.

Contexts

One of the deficiencies of the KWIC index style is that the context each line can show is limited to its line length. We could afford 55 characters for the context. Since one or two Chinese characters represent a word, this length can accommodate more than 30 words of information in English.

Reverse Sorted Index

Two types of KWIC index have been produced. One is for the normal type, in which all lines are sorted in the ascending order of the Chinese standard telegraphic code of key characters (plus 7 succeeding characters). Fig. 6 shows an example page from this type of index. The other is the so called "reverse sorted" index. The major key for this type is the same as that of the normal type. The minor sort keys are, the characters immediately preceding the major key. Thus all lines for one key character are listed in the ascending order of the code for the character immediately preceding the key character and so on. Fig. 7 shows an example page from the reverse sorted concordance.

CONCLUDING REMARKS

The two sets of modern Chinese concordances can be reached at the National Inter-university Research Institute of Asian and African Languages and Cultures, Tokyo University of Foreign Studies. It should be noted that a concordance of one million lines amounts to over 25,000 pages (actually it counts for 27,341) or 50 volumes of a 5cm-width paper file. Before printing the whole index, engineers recommended linguists to use COM technique, but in vain. A microfiche version should have been produced for portability. Analysis of the concordances have just got off the ground. The resulting papers are expected to follow.

Fig. 5. An Example from the Edited Text
Fig. 6. A Page Example from the Chinese Concordance (Normal Style)
Fig. 7. A Page Example from the Chinese Concordance (Reverse Sorted Style)
REFERENCES

1. Kierman, F.A. and Barber, E.: "Computers and Chinese linguistics", Unicorn, No. 3 (1968)
2. Boltz, W.G., Barber, E. and Kierman, F.A: "Progress report on Pai-hua-wen computer count and analysis", Unicorn, No. 7, pp. 94-138 (1971)
3. Hashimoto, M.J., et al.: A grammatical analysis of the Princeton million-character computer file", Bulletin of the Chinese Language Society of Japan, No.222, pp. 1-16,36 (1975)
4. Hashimoto, M.J.,: "Computer count of modern Chinese morphemes", Computational Analysis of Asian and African Languages, No. 7, pp. 29-41 (1977)
5. Uemura, S.: "Automatic Compilation and Retrieval of Modern Japanese Concordances", Journal of Information Processing, Vol. 1, No. 4, pp. 172-179 (1979)