Automatic Compilation and Retrieval of Modern Japanese Concordance

SYUNSUKE UEMURA*

This paper describes an automatic indexing project in Japanese. Very large volume of modern Japanese concordance has been compiled automatically and its retrieval systems were provided. Japanese texts were manipulated directly by computers without using Romanization. The data of up to 750,000 words contains sampled sentences from 3 daily newspapers in Japan. Since the whole concordance reaches more than 750,000 lines in Japanese KWIC index form, traditional book style printing could not be adopted. The concept of "concordance system," which is not a concordance generating system but the concordance itself as a system is introduced. The concordance is recorded on both magnetic tapes and a set of microfiche. Mini-computer based concordance retrieval systems are developed for the users to retrieve and read any necessary part of the concordance either through a kanji (Chinese character) display terminal or through a microfiche viewer. The whole concordance was materialized successfully as the "concordance system".

1. Introduction

This paper describes an automatic indexing project in Japanese. Very large volume of modern Japanese concordance (more than 750,000 lines of KWIC index) has been compiled automatically and its retrieval systems were developed. The goals of the project were; (1) to develop technology for processing this amount of Japanese text, and (2) to demonstrate the feasibility of the concept of "concordance system".

The National Language Research Institute (NLRI) of Japan contributed the data of up to 750,000 words (chootan'i; 長単位), which contained sampled sentences from 3 daily newspapers in Japan. An automatic indexing project was carried out by the author at Electrotechnical Laboratory, to compile a Japanese concordance of the whole data, an inverted word list and others. The project included experiments on specially tailored information retrieval systems for the Japanese concordance, a COM experiment in Japanese, a code conversion system between kanji (Chinese character) code sets, and so on.

Though concordances of those modern newspaper vocabularies would be highly useful in understanding Japanese, no such concordances had been compiled before, partly because of the difficulties in kanji input/output processing, and partly because of the lack of understanding in automatic indexing in Japan. Various technical aspects of Japanese (as described later) have also prevented us from the idea. However recent improvements in kanji output devices and the fundamental studies on automatic indexing by the author made it practical to realize the automatic indexing of this amount of data. Not only the concordance generated would be useful, but also it is expected that such experiments will be beneficial to the progress of the researches on natural language (especially Japanese) processing by computers.

Since the concordance of the whole data reaches more than 750,000 lines it is inconvenient to print them out in traditional book style. The reader may not be able to carry this huge volume of books. For such huge volume of concordance, the author proposes the concept of "concordance system", which is not a concordance generating system, but the concordance itself as a system. The concordance system consists of magnetic tape files that contain the whole concordance, a set of microfiche that contains the whole data reduced in size, and at least one information retrieval system to browse or retrieve the necessary part of the concordance. It is expected that this type of system shall construct a new book concept in future libraries.

The project attained its goals by materializing the Japanese "concordance system" of such huge volume as 750,000 lines. Among the technical contributions to natural language processing by the project are:

(1) The first, very large volume of modern Japanese concordance was compiled and materialized.

(2) Mini-computer based concordance retrieval systems were developed with special design considerations to end user interfaces (menu and other interactions all in Japanese, and so on).

(3) The feasibility of microfiche concordance through Japanese COM technology was demonstrated.

(4) Experiments on automatic ordering of Japanese concordance entries, general Japanese code conversion system and some others have been carried out. An algorithm to order Japanese entries was proven to produce sufficiently natural ordering.

*Computer Science Division, Electrotechnical Laboratory, 2-6-1, Nagata-cho, Chiyoda-ku, Tokyo, Japan.

2. Historical Background

When Dr. H. P. Luhn coined a new term "automatic indexing" in late 1950's, it's meaning was twofold, namely, index compilation by computers and automatic assignment indexing [1]. Since then, much effort has been devoted to the latter by various research groups [2], but not so much to the former target. Especially in Japan, little work has been done so far on automatic compilation of concordances. This was partly because of misunderstanding on the term (only the assignment indexing aspects were highlightened), and partly because of unique features of Japanese, to be emphasized among them are a large character set including Chinese characters, no segmentation, and special ordering of dictionary entries.

Electrotechnical Laboratory is one of the largest national research institutes specializing in electricity and electronics in Japan. Since 1960's it has been active in the field of computer application to natural languages, especially in mechanical translation. As a fundamental research to support it, the current author has been working for the successful compilation of Japanese concordances by computers. He proposed various new techniques for segmentation, entry ordering and so on, and demonstrated their feasibility in [3, 4]. In additon, recent rapid progress in kanji (Chinese character) input/output devices made it possible to compile Japanese texts directly. Here "directly" means not by using Roman alphabets, but by using Japanese character set.

On the other hand, the National Language Research Institute has been active on the researches on vocabularies and kanji in Japanese. Three big research projects have been conducted manually, namely vocabularies and Chinese characters in women's magazines, in cultural reviews, and in 90 magazines today. Then a study of vocabularies in modern Japanese newspapers began in 1965 utilizing a computer. Their research was oriented towards statistical aspects on modern Japanese vocabularies. A series of reports has been published so far [5]. After extensive work by the NLRI, one third of the data (750,000 words) were open to the interested groups in Japan in 1972.

The current author immediatly initiated a project to compile a Japanese concordance of the whole data with the concept of "concordance system". The project began in April 1972 and completed in August 1975. The first, very large volume of Japanese concordance has been automatically compiled successfully and its retrieval system with kanji keyboard-display was provided. Later in 1976, another information retrieval system with microfiche retrieval facility was developed and integrated into the concordance system.

3. Automatic Indexing Project Based upon the NLRI Data

3.1 Motivation to the Japanese Concordance

The needs to "concordance" are quite clear. In the area of automatic processing of natural languages, the first and the most primitive stage is computer application to statistics. A table of word frequency can tell us much about the target language, such as basic vocabularies or transient aspect of the vocabulary. However, in order to understand a live language as it is, we need information not only on the frequency of words but also on "how" they are used. Thus the next step would be concordance compilation to capture word usages. There can be seen various types of concordances, e.g. a concordance of the Bible, of a specific novel, of an author's entire works, and of a language in general. Here we focus our attention to the last type. If we want to understand what a word means or how it is used, a concordance is a must. A wide variety of concordances compiled either manually or automatically has been published so far.

Since the appearance of electronic computers, researches on automatic compilation of concordances have been performed in Eulopean countries and in the U.S.A. However those research activities were somewhat limited towards the usage of Roman alphabets. No computer generated concordance had been available that used ideographs such as Japanese or Chinese, before our project.

Also stressed is the huge volume of the raw data, 750,000 words. Even if they had been Romanized, the volume would have caused much trouble. Since $750{,}000 = 50 \times 500 \times 30$, it would require 30 volumes of 500 page books (50 lines per a page), if we print it out in conventional book style. This problem of voluminous output would be essential to concordances themselves, especially when computers are utilized. Computers can easily produce fairly large volume of output, but how should it be materialized? A new concept or switch-over is required.

Therefore, the project on automatic compilation of Japanese concordance had the following motivation and significance.

(a) It will widen computer application to natural language processing (formally it was restricted to Roman alphabets, now we need new techniques to handle ideographs effectively).

(b) The volume of 750,000 lines of concordance itself requires new techniques or even new concepts.

(c) The concordance, final production of the project, would be of significance to the understanding of natural languages, especially that of Japanese.

3.2 Technical Difficulties and Their Solutions to Manipulate Japanese

Discussion on some characteristics unique to Japanese

and their solutions follow:

(a) Large character set and its code set—Japanese character set includes some 3,000 kanji (most of which are from Chinese characters but with different pronunciation), syllabic characters in two styles, katakana and hiragana. One of the most urgent needs to be resolved in Japanese computer society has been and is even now to develop an easy and efficient way to input or output this character set. Extensive investigation from various approaches is being carried out on this problem and is yielding fairly good results in Japan. In this project, we utilized a kanji keyboard-display terminal (by OKI Electric, Inc.) and a kanji COM system (by JEM Computer Systems, Inc.). It is expected to apply these techniques to the computer processing of Asian-African languages, such as Chinese, Korean and many others. Another related difficulty in this area comes from the fact that there is no standard computer code set for Japanese. A code for Japanese usually consists of 12 to 16 bits, but each device in each manufacturer uses its own. In this project, a general Japanese code conversion system was developed. It can convert Japanese data of a code set into another. At most 18 code sets can be handled. The NLRI data was converted into kanji keyboard-display code set and Japanese COM code set. Since more than 3000 different characters are used, code conversion is laborious and tedious. By the establishment of the Japanese Industrial Standard "Code of the Japanese Graphic Character Set for Information Interchange" early in 1978, the environment is expected to change in near future.

(b) No segmentation—Japanese sentences are continuous sequence of one or more words without any spaces between them. Words are simply concatenated to form sentences. This causes much trouble in automatic processing of Japanese since if one wants to manipulate a word, one will have to determine it before processing. This is one of the reasons why automatic indexing could not survive in Japan. Situations are the same in Chinese, Korean, and in some other languages.

The author proposed a method for automatic segmentation based upon the analyses of character classes and transition between these classes [3]. The rules are roughly described as follows.

Suppose we have a Japanese character string.

国産の COBOL コンパイラ研究

We can easily obtain a string of character classes from it, namely, kanji, kanji, hiragana, alphabet, . . . , and kanji.

(i) The leading position of kanji string has high possibility to be the beginning of a word, which indicates the segmentation information.

国産の COBOL コンパイラ研究
↑ ↑

A long kanji string usually consists of the series of 2 kanji phrases.

情報|処理|研究

There are also some groups of kanji that are used as prefixes or suffixes.

超|高速，計算|機

Any long kanji strings can be divided into meaningful units by the rules above, and they show how we should segment kanji strings.

超|高速|電子|計算|機

The rules for kanji strings can be formalized as BNF rules in Fig. 1.

```
<K-string> ::= <K-string> <K-string>
             | <K-prefix> <K-unit>
             | <K-unit> <K-suffix>
             | <K-unit>
<K-unit>   ::= 容量 | 導体 | 計算 | ···
<K-prefix> ::= 大 | 半 | ···
<K-suffix> ::= 機 | 器 | ···
```

Fig. 1 Construction rule for kanji strings.

(ii) Also the leading position of Roman alphabet string has high possibility to be the beginning of a word, especially when they consist of more than one character.

国産の COBOL コンパイラ研究
↑

(iii) The leading position of katakana string is sure to be the beginning of a word in usual Japanese sentences.

国産の COBOL コンパイラ研究
↑

(iv) The transition point to hiragana from others (kanji, katakana, . . .) has to be carefully examined, because it has high possibility to be a zyosi or zyodosi. Japanese titles of 148 scientific reports were investigated and the number of combination of actually used zyosi and zyodosi was proved to be less than 30. Some of them are;

に，による，と，としての，の，のための，が，を，および

The rest has some possibility to be keywords.

These rules require only small tables for segmentation. Fig. 2 shows an example of automatic segmentation

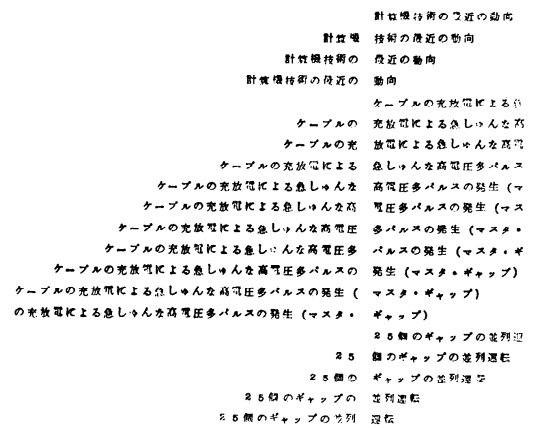

Fig. 2 An example of automatic segmentation of Japanese.

by this method. This is one of the most practical methods in producing Japanese concordance, because the processing time is much shorter than the traditional longest-match method.

On the other hand, staffs of the NLRI defined their own "word (chootan'i)" and segmented the raw data manually when preparing computer input. Consequently in our project, this manually segmented data was used.

(c) Special type of entry ordering—Entries are ordered by their pronunciation in Japanese dictionaries or concordances. However a word can be represented by different ways. If they are pronounced the same, they will appear next-door on a page. Syllabic types or character types slightly affect the ordering within same pronunciation word group. Both "あおい", "青い" and "アオイ" appear consecutively because they are pronounced the same. Another problem is that voiced sound (sonant) and semi-voiced sound (p-sound) also affect the entry ordering. Solution is to sort the entries by a series of key strings that represent pronunciation, sound information, and syllabic styles. A format and construction method of sorting key proposed by the author is roughly described as follows [4].

A sorting key for both hiragana and katakana string consists of a major key which represents basic pronunciation (voiceless sound only), and two minor keys, one for sound information and the other for the styles of character sets.

The proposed construction of sorting keys for kana strings is;

Major key: hiragana representation, all converted into voiceless sound.

Minor key 1:
- 0 for space
- 1 for special character
- 2 for contracted or assimilated sound (yoo-on, sokuon)
- 4 for voiceless sound (seion)
- 6 for voiced sound (dakuon)
- 8 for semi-voiced sound (handakuon)

Minor key 2:
- 0 for space
- 2 for hiragana
- 4 for katakana

A sorting key for Roman alphabet string consists of a major key which represent basic alphabet information, and two minor keys one for special characters, the other for the capitalization.

The proposed construction of sorting keys for Roman alphabet strings is;

Major key: All capitalized string, special characters such as–or. being eliminated.

Minor key 1:
- 0 for space
- 2 for alphabet
- 4 for special character

Minor key 2:
- 0 for space
- 2 for upper case character
- 4 for lower case character

Examples are shown in Fig. 3. For Chinese character strings, if we can get readings by any dictionary, then the strings can be sorted naturally. Since the NLRI data included a dictionary for all vocabularies in the raw data, it was possible to apply the method successfully in this project. Figs. 6 and 8 in the following sections show the outcome ordering obtained by this method.

Original String	Major Key	Minor Key1	Minor Key2
かなづかい	かなつかい	446440	222220
かんれいしゃ	かんれいしや	444442	222222
サーチング	さちんく	414460	444400
ちらし	ちらし	444000	222000
デリミッタ	てりみつた	644240	444440
とじしろ	とししろ	464400	222200
⋮	⋮	⋮	⋮
はさみ	はさみ	444000	222000
はす	はす	440000	220000
はず	はす	460000	220000
ハズ	はす	460000	440000
パス	はす	640000	440000
ばず	はす	660000	220000
パス	はす	840000	440000
はな	はな	440000	220000
⋮	⋮	⋮	⋮

(a) Sorting Key for Kana Strings

Original String	Major Key	Minor Key1	Minor Key2
Am	AM	220000000000	240000000000
am	AM	220000000000	440000000000
Am.	AM	224000000000	240000000000
am.	AM	224000000000	440000000000
A. M.	AM	242400000000	220000000000
Amabel	AMABEL	222200000000	244444000000
⋮	⋮	⋮	⋮
atomical	ATOMICAL	222222220000	444444440000
atomic clock	ATOMICCLOCK	222222022222	444444444440
atomic proof	ATOMICPROOF	222222022222	444444444440
atomic-proof	ATOMICPROOF	222222422222	444444444440
atomics	ATOMICS	222222200000	444444400000
⋮	⋮	⋮	⋮

(b) Sorting Key for Roman Alphabet Strings

Fig. 3 Proposed sorting key construction for Kana and Roman Alphabet strings.

3.3 Towards a Concordance System

The raw data published by the National Language Research Institute consists of 4748 article units sampled from 3 daily newspapers in Japan. A sampled article unit consists of an average 500 Japanese characters. The total amount of the data is 2,360,000 characters or 754,000 words. A dictionary for the whole data was also published, which included reading information, word classes, and so on.

As the format of concordance, we decided to choose KWIC index style since it is one of the most successful format in automatic indexing. Variations such as KWOC or so would fundamentally be the same as KWIC. New experimental formats were also investigated, but the detail will be found elsewhere [4]. KWIC index

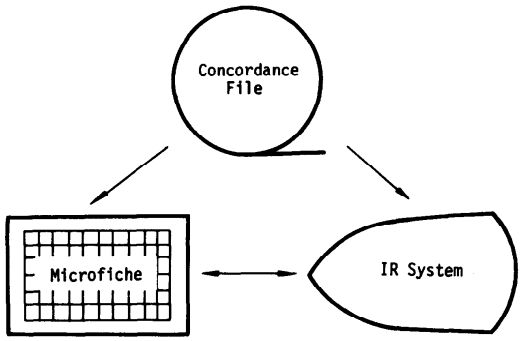

Fig. 4 The concept of "Concordance System".

of 750,000 words constructs 750,000 lines of concordance. Since this is a concordance for the whole language (Japanese), all words that appear in the data should be the subject. Usages should be found as many as possible. Elimination of specific words or usages would reduce the practical value of the concordance. Then, can the concordance of 750,000 lines be realized? Some 30 volumes of 500 page books would be required if we print it out. This is possible but certainly inconvenient to it's readers. For such huge volume of concordance, the concept of "concordance system" is proposed. The concordance system is not a concordance generating system, but the concordance itself as a system. The concordance system consists of the followings (see Fig. 4):

(a) Magnetic tape file (or any other auxiliary storage deveice) that contain the whole contents of the concordance. This component is for computer processing. It is also the base for both (b) and (c).

(b) A set of microfiche that contains the whole concordance reduced in size. This component is for portability and easiness of reproduction. It should be produced directly from (a) to avoid voluminous printing.

(c) An information retrieval system to browse or retrieve the necessary part of the concordance. This component is for those who wants to browse the concordance as if it were a book. The information retrieval system can either be based on (a) with display screen terminal, or on (b) with microfiche retrieval facility.

Possibly there can be many variations. This automatic indexing project realized a type of the concordance system as described below, and demonstrated the feasibility of the concept. The author believes that future libraries would adopt at least some of those concepts in near future.

3.4 Concordance Compilation

General flow of concordance compilation process is shown in Fig. 5. The computer system this project used to compile the concordances is as follows:

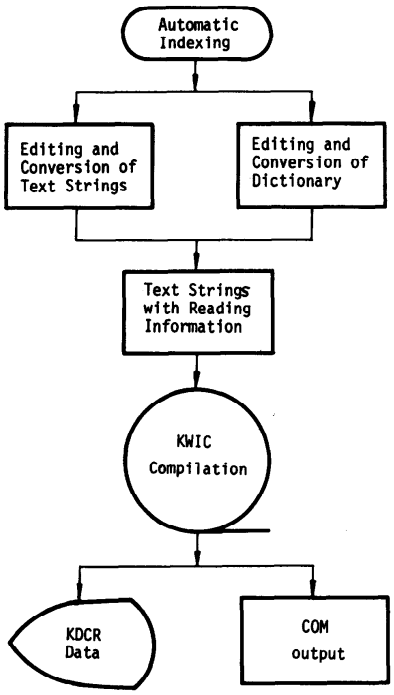

Fig. 5 General flow of concordance compilation process.

HITAC 8410 (compatible to IBM/360 model 40)
Main memory 262,144 bytes
1.44 micro second/2 bytes
Auxiliary Storage
6 magnetic tape units (32 r/mm)
8 disk drivers (each 7.25 MB/pack)

Some 30 programs (including a kanji code conversion system) were developed. All programs were written in COBOL. Working data sometimes exceeded 10 magnetic tapes (each with 732 m, 32 r/mm). Concordance of the whole data, more than 750,000 lines of KWIC index has been compiled successfully. It was stored on6 x 732 m magnetic tapes, and was then diverted to COM experiment and to the retrieval systems. As mentioned before, byproducts of the project, an inverted word list, experimental type of indexes have also been produced but the detail will not be discussed here (See [4] for reference).

3.5 Japanese COM Experiment

Instead of producing very large volume of books by printing out the whole concordance, they were photographed on 275 sheets of microfiche. The microfiche was produced by a kanji COM technique developed at JEM Computer Systems, Inc. JEM system 3100 PT-150F consists of a kanji display screen and a set of camera equipments. Contents of magnetic tapes are shown on the kanji display screen and then it is photo-

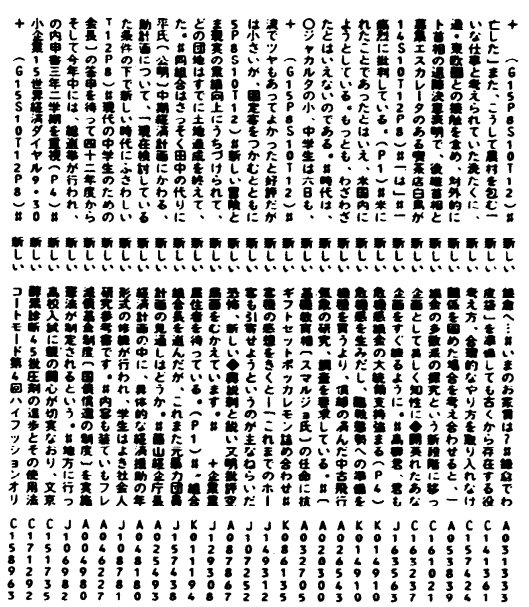

Fig. 6 A sample page from the microfiche concordance (Vertical printing style).

graphed. This means that the contents are microphotographed directly from magnetic tapes to sheets of microfiche. No printed matter are used in between.

Two technical points should be emphasized here. First, various standard microfiche formats has been examined, and Type Al format was accepted. A sheet of type Al microfiche contains 98(=7×14) pages of concordance, each page containing 1344 characters (=48×28) respectively. Reduction rate is 1/24 for traditional A4 form sheet. The reason for Al format is as follows:

(a) Most kanji printers in Japan use A4 or B4 type sheet for printing instead of 11 inches × 14 inches line printer sheets. Thus Type A3 format (for line printer sheets) were rejected.

(b) Type A1 format fits the vertical printing style in Japanese.

(c) Type A3 is becoming popular in the U.S.A., overwhelming old Type A2.

Another point to be emhasized is that the concordance is vertically printed out on the microfiche. Vertical printing (top-to-bottom, right-to-left) is traditional in Japan. Although horizontal style is becoming popular in various fields, novels and most newspapers are usually printed out vertically. So why not concordances! Since the microfiche production machine can print data only horizontally, the vertical printing was done solely by software. Fig. 6 shows an example page from the microfiche concordance.

The user who wants to read the concordance can either read the microfiche through casual viewers, or use the microfiche retrieval system described in the next section.

3.6 Concordance Retrieval Systems

To browse and read necessary parts of the concordance recorded on magnetic tapes or sheets of microfiche, the development of information retrieval systems was included in this project. Two types of retrieval systems have been developed. Though the two systems are on the same mini-computer OKITAC 4300C, they are independent from each others and the user may choose whichever he wants (see Fig. 7).

(a) Concordance retrieval system using a kanji keyboard-display terminal (KDCR)—This system has been constructed around a kanji keyboard-display device developed at OKI Electric, Inc. The basic features of the KDCR system are;

(i) It allows the user to handle the contents of the magnetic tapes (concordances) as if they were printed materials.

(ii) The interface between the system and the user fully utilizes Japanese so that end users, including linguists, can use it easily.

(iii) The interface includes special convention for end users, namely they can input keywords by hiragana representation.

The user specifies a keyword (a word he wants to

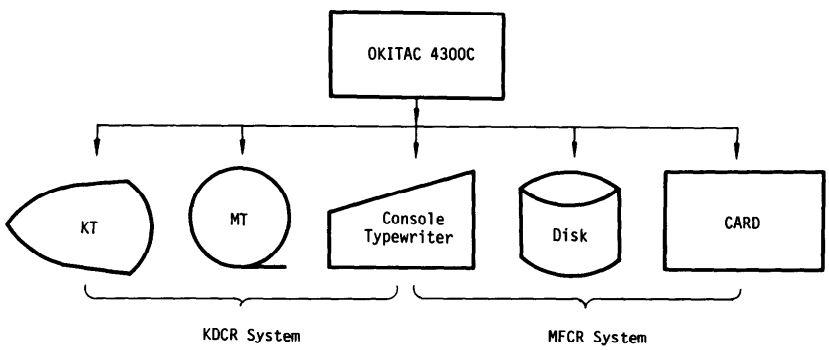

Fig. 7 Concordance retrieval systems environment.

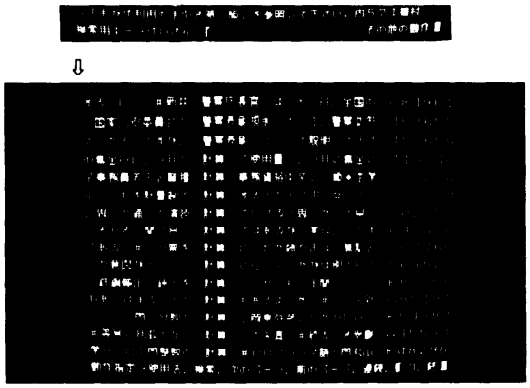

Fig. 8 A part of the concordance on the Kanji display screen (Horizontal Style).

examine it's usage) through the kanji keyboard-display. He may use syllabic representation to specify it. A keyword with syllabic representation (hiragana) is used as a matching key with the major sorting key of the concordance tape file. Therefore, in this case, sorting key should be kept remaining in the file. The system on the mini-computer then searches magnetic tapes sequentially and throw the retrieved part of the concordance onto the kanji display screen. Fig. 8 shows a part of the concordance on this stage. The part is retrieved by providing the keyword "けいさん". The last (16th) line of the screen is always so-called a "menu", by which the user and the system interact. Instead of typing in commands, the user can give it to the system by pointing proper position on the menu with the light pen. He may ask the system to retrieve the first page that contains the usages of the specified keyword, to show the succeeding or preceeding pages page by page, to produce a hard copy of the display screen, and to indicate him how to use the system. Here "page" refers to a screen image of the display terminal. A page consists of 32 characters × 15 lines (+1 line for the menu).

Though the "menu" approach is not novel for interactive systems, the one used here includes some new features.

(a) Not only the commands in menu but also all the interface with the user fully utilize Japanese (kana-kanji strings).

(b) The user can input keywords by hiragana representation.

(c) The menu is specially tailored to the "book browsing" feelings.

A trouble with this system is that it sometimes takes too long to retrieve a part of the concordance, since the data is on magnetic tapes.

The KDCR system environment is as follows:

Kanji Keyboard-display	OKI KT system
Mini-computer	OKITAC 4300C (24 K words × 16 bits, cycle time 0.6 microsecond)
Auxiliary Storage	Magnetic tape devices (TEAC MT-8A, 32 r/mm)
Software Size	14,800 words including buffer areas. Programming is with OKITAC 4300C assembly language. The number of statements is about 4800.

The KT system is controlled by OKITAC 4300C. The KT system is a kanji I/O terminal with a refresh type cathode ray tube and a teletypewriter type keyboard. The display screen can throw 32 × 16(=512) Japanese (or Chinese) characters, each character in turn being represented by 18 × 18(=324) dots. The dot images of kanji are on specially designed UI core translater memory, which is also contained in the KT system itself.

(b) Microfiche Concordance Retrieval System (MFCR)—The user can look for any part of the microfiche concordance by usual microfiche handling methods. Moreover an on-line interactive microfiche concordance retrieval system was designed and developed on the same mini-computer with commercial microfiche retrieval device CARD by the Image Systems, Inc.

CARD can handle 750 sheets of microfiche. The Japanese concordance reaches a little more than one third of the capacity. CARD itself has simple retrieval capability through the microfiche serial number and positional information within it. It was connected to the mini-computer, and an inverted file type concordance retrieval system was developed.

Inverted files to retrieve a microfiche page through user provided keywords are on the mini-disk. The user types a keyword into the console typewriter of the mini-computer. MFCR system searches the pertinent inverted file, determines the microfiche serial number and horizontal and vertical position within it, and send signals to CARD system to make it throw the corresponding page on the viewer screen. Less than 3 seconds are required to retrieve a page. The user can also specify to show him the succeeding or the preceeding pages, to produce a hard copy of the page, and to indicate him how to use the system. Here "page" refers to a viewer screen image of the CARD system. A page consists of 48 characters × 28 lines (see **3.5**). Though the page size of the MFCR system differs from that of the KDCR, both are generated from the same magnetic type file (component (1) in **3.3**). The basic commands are analogous to the "menu" of the KDCR system. Moreover he can specify the microfiche serial number and positional information within it directly if he knows.

The MFCR system environment is as follows:

Microfiche Retriever	CARD Model 201
Minicomputer	OKITAC 4300C
Auxiliary Storage	Mini-disk device (Mitsubishi M801F, 6 MB)

Software Size	194,000 words including buffer areas. Programming is with OKITAC 4300C assembly language. The number of statements is about 9400.

4. Summary

An automatic indexing project with Japanese data was described. The project had two levels of contributions. First, technical problems it faced and solved contributed to the progress of natural language processing by computers, especially automatic processing of those languages that use ideographs. The automatic compilation of Japanese index with this volume was shown to be feasible with the concept of "concordance system". Secondly, the concordance—the final product of the project—would be beneficial to researches on Japanese. Some research groups in Japan have already begun utilizing it.

Computers in numeric computation have turned the computation of pai (π) from mathematician's life work into "less-than-a-second" short job, thus made mathematicians free from this kind of fundamental but primitive computations. Computers in natural language processing is now changing the compilation of "concordances" from linguist's life work into "less-than-a-month" or "less-than-a-year" scale work. In addition, the volume of concordance that can be compiled at one time increased drastically. Linguists or those who are interested in language processing can now concentrate their efforts to more "intelligent" or "humanistic" aspects of their researches.

The author coins here a new term "CAL (Computer Assisted Linguistics)" to emblematize the concept of researches on natural language processing assisted by this kind of huge computer generated concordance system.

Acknowledgment

The author is greatly indebted to the staffs of the National Language Research Institute for providing him the raw data. He gratefully acknowledges the supports and suggestions given by the staffs of the Electrotechnical Laboratory, especially, Drs. O. Ishii, H. Nishino, K. Torii and H. Nishimura. He wishes to extend his sincere appreciation to Professors T. Sakai and M. Nagao of Kyoto University for their guidance and encouragement.

References

1. Luhn H. P. Keyword-in-Context Index for Technical Literature (KWIC INDEX), *Am. Doc.*, **11**, 4 (1960), 288–295, also in H. P. "Luhn Pioneers of Information Science Selected Works", Schults, C. K. ed., Spartan Books (1968).
2. Stevens, M. E. Automatic Indexing: A State-of-the-art Report, *NBS Monograph*, **91**, U.S. Department of Commerce, 1965, (Revised, 1970).
3. Uemura, S. A Japanese KWIC Indexing System, *Information Processing in Japan*, **10** (1970), 1–4.
4. Uemura, S. Studies on Automatic Indexing by Computers (Volume one: Fundamental Studies on Automatic Indexing, Volume two: Automatic Indexing Experiments with Large Quantities of Data in Japanese), *Researches of the Electrotechnical Laboratory* 743, 747 (1974, in Japanese).
5. Studies on the Vocabulary of Modern Newspapers, **1–4**, *National Language Research Institute Reports*, 37, 38, 42, 46, Syuei Shuppan Co., (1970–1973, in Japanese).

(Received December 21, 1977; revised November 1, 1978)