Jim Mathias

COMPUTER-AIDED PROCESSING OF CHINESE LEXICOGRAPHIC MATERIALS

Introduction

The CETA (Chinese-English Translation Assistance) Group, of which I am Executive Secretary, was organized in 1971 by a group of China specialists that perceived the need for cooperation in development of Chinese-English translation aids. CETA was created as an independent organization that would draw upon services of Chinese language and computer specialists from government institutions and the academic and private communities.

At the outset, it was determined that the primary effort would be the production of a computer-generated, loose-leaf 'living dictionary' of general terms which would be capable of being updated as current usage changed and new terms were added to the language. In addition, the CETA Group has been working on technical glossaries of which there are more than twenty in process and a bibliography, all in computer (machine-readable) form as well as printed form.

The CETA Group dictionary files are developed, sorted, and printed by computer. The principal advantages of using the computer are the ability to print Chinese characters without typesetting and the speed and economy with which dictionary data can be manipulated to make corrections or additions, to sort in different sequences (e.g. radical or phonetic), to extract particular subsets of information (e.g. idioms or place names), or to issue substitute pages to users of the printed version of the computer file.

Special conditions

In order to understand the design of CETA's methods and files it is necessary to remember that it is a voluntary participatory group in which most of the services, at least in the initial years, were contributed by academic Chinese-language specialists and government Chinese-language and computer programming specialists. The Secretariat staff initially operated exclusively as a switching mechanism to obtain information, organize it, put it in machine-readable form, and return it to editors to review. It was only in more recent years that the Secretariat staff has undertaken tasks of review, consolidating language information, editing and proofing.

Another important concept to keep in mind is that the general dictionary file was created not in the usual method in which each entry is exhaustively researched before going on to the next but rather one in which all entries are revised to an intermediate level and then the review cycles are repeated. Three or four complete cycles of review were planned. The first cycle weeded out transparent errors in Chinese or English based on top-of-the-head knowledge of the volunteer linguists. The second

cycle was intended to document the existence of the entries in one or more printed sources and further expand the English as much as possible (requiring some research). The third cycle was then initiated whose purpose is to:

(1) validate every entry in a Chinese source;

(2) research and include all available meanings;

(3) include grammatical information, some personal and place names and usage labels.

This third cycle of the general dictionary file will create what is more familiarly thought of as a dictionary. The reasons these tasks were approached in three cycles rather than one is that it produced usable information early, allowed greater flexibility using volunteer manpower, a meagre budget, and non-priority processing; and allowed us to formulate better the Chinese content of the file quickly.

Language problems

Major problems in computer processing arise from the fact that the Chinese writing system differs greatly from the writing system of Western languages. Western writing is alphabetic; Chinese by contrast has an inventory of many thousands of characters known as ideograms or logograms, each of which represents a unit of meaning. It also contains a so-called phonetic element which offers some clue to pronunciation, but with limited consistency. The actual sound of a given character may differ from dialect to dialect, and from language to language, but the meaning remains more or less the same. It can thus be said that China has several spoken languages, but one written language.

The compilation of dictionaries of Chinese requires a set of over 10,000 unique Chinese characters. This is not a measure of all Chinese characters since over 50,000 such characters exist. Also, Chinese characters are written equally spaced in a string without indication of word boundaries, i.e. there are no spaces between character groups to indicate words. Finally, if in our dictionary we wish to reflect the so-called standard spoken form of the character, i.e. Mandarin or putonghua, we must include a romanized spelling plus a symbol for one of the five major tones which is imposed on the basic sound of each syllable.

The computer files designed by the CETA Group represent the Chinese characters internally in an alpha-numeric code. The characters are displayed or printed by computer using vector-drawing methods. The characters are coded using a matrix and an encoding system. The Chinese characters in machine-readable form are identified by a four-digit code (or telegraphic code). This is the code used to send telegraphs in China. The code list has been expanded by us where necessary.

A page format was designed for the dictionary files which would allow ease of maintenance and organization of information. A page format shown in Fig. 1 is designed to achieve maximum legibility, and at the same time make efficient use of space

01 天作孽，猶可爲，自作孽，不可活
TIAN-ZUO\NIE\, YOU/KE~WEI/, ZI\ZUO\NIE\, BU\KE HUO/ 1131 0155 5642 9976 3730 0668 3634 9976 5261 0155 5642 9976 0008 0668 3172 IVFI IF DISASTERS COME FROM NATURE, SOMETHING CAN BE DONE TO COUNTER THEM; BUT IF THEY ARE OF ONE'S OWN MAKING, ONE IS DONE FOR. IVFI

02 從羣衆中來，到羣衆中去，集中起來，堅持下去
CONG/QUN/ZHONG\ZHONG-LAI/, DAO\QUN/ZHONG\ ZHONG-QU\, JI/ZHONG-QI~LAI/, JIAN~CHI/ XIA\QU\ 1783 5028 5883 0022 0171 9976 0451 5028 5883 0022 0637 9976 7162 0022 6386 0171 9976 1017 2170 0007 0637 INSI FROM THE MASSES, TO THE MASSES. CONCENTRATE THEM, AND HOLD FAST TO THEM IWWI

03 紙老虎 ZHI~LAO~HU~ 4786 5071 5706 IGEI PAPER TIGER. FROM MAO'S "TALK WITH THE AMERICAN CORRESPONDENT ANNA LOUISE STRONG", AUGUST 1946, WHICH SAYS

一切反動派都是紙老虎．看起來，反動派的樣子是可怕的，但是實際上並沒有甚麼了不起的力量．從長遠的觀點看問題，眞正強大的力量不是屬於反動派，而是屬於人民．

ALL REACTIONARIES ARE PAPER TIGERS. IN APPEARANCE, THE REACTIONARIES ARE TERRIFYING, BUT IN REALITY THEY ARE NOT SO POWERFUL, FROM A LONG-TERM POINT OF VIEW, IT IS NOT THE REACTIONARIES BUT THE PEOPLE WHO ARE REALLY POWERFUL IKSI

04 經濟核算制度 JING-JI\HE/SUAN\ZHI\DU\ 4842 3444 2702 4615 0455 1653 ICWI BUSNESS ACCOUNTING SYSTEM (A CONCEPT IN MARXIST ECONOMICS, KNOWN IN THE U.S.S.R. AS KHOZRASCHETNOST. ХОЗ

РАСЧЕТ
НОСТЬ) ICJ,CWI

05 負芻 FU\CHU/ 6298 5368 IGMI CARRYING GRASS FOR FIRING -- FOMENTORS OF DISORDER, FROM MENCIUS: 昔沈猶有負芻之禍 FORMERLY WHEN SHEN YU WAS EXPOSED TO THE OUTBREAK OF THE GRASS-CUTTERS. IGMI

06 贋鼎 YAN\DING~ 6372 7844 IGMI A SPURIOUS TRIPOD (THE STATE OF TS'I, 齊, DEFEATED THAT OF LU, 魯, AND DEMANDED FROM THEM THE 讒鼎, A FAMOUS HISTORICAL RELIC; THE STATE OF LU SENT THEM A SPURIOUS IMITATION); USED IN THE SENSE OF TO FORGE, TO COUNTERFEIT, ETC. IGMI

760906 PAGE 2 190 髟 + 尽

07 車書 CHE-SHU~ 6508 2579 IGMI UNIFORMITY, FROM THE LINES:

車同軌，
書同文，
行同倫． ALL CARRIAGE WHEELS ARE OF ONE SIZE, ALL WRITING IS WITH THE SAME CHARACTERS, AND IN CONDUCT THERE ARE THE SAME RULES.) IGMI

08 軋姘頭 GA/PIN-TOU/ 6509A 8280 7333 IOYI TO COMMIT ADULTERY (WU DIALECT) IWBI

09 風木 FENG-MU\ 7354 2606 IGMI NOT LONG ABLE TO CARE FOR PARENTS; (FROM THE LINES WRITTEN BY 皐魚 AT THE DEATH OF HIS PARENTS

木欲靜而風不止，
子欲養而親不待

THE TREES WOULD BE STILL BUT THE WIND CEASES NOT, THE SON WOULD CARE FOR HIS PARENTS BUT THEY TARRY NOT) IGMI

10 雄樣子 XIONG/YANG\ZI~ 7546 2876 1311 OF OLD FASHIONED STYLE, WEAK SPIRITED

11 鸊鷉 PI~QI/ 7754 7744 IVWI CHINESE LITTLE GREBE IVWI

12 髮妻 FA\QI~ 7569 1189 IGMI A MAN'S FIRST WIFE, TO WHOM HE WAS BETHPOTHED AS A CHILD. (FROM THE LINE: 結髮爲夫妻

恩愛兩不疑 MADE MAN AND WIFE IN THEIR PRIME (I.E. WHEN THE HAIR IS FIRST PUT UP), THEIR LOVE AND AFFECTION CANNOT BE DOUBTED) IGMI

Fig. 1

Note: These are sample entries showing special features. It is not an actual dictionary page.

to reduce the bulk of the dictionary to manageable size. The dictionary can accommodate terms of almost any length. Each entry is identified by page and line number. The sources of an entry are indicated by one or more source digraphs, i.e. two letters between vertical lines.

Other information incorporated to provide aid to the user is the telecode number, the pinyin romanization of the sound with tone, the English meaning, and, within the English meaning, subject labels and usage labels. These labels are invaluable for computer extraction of subsets of entries.

Batch-processing method

The CETA file creation and maintenance system was established in the early '70s as a batch-processing operation. The batch-processing method was chosen over an on-line approach because it would function with a very limited budget while the incorporation of an on-line method would require a heavy investment in equipment for the CETA Secretariat office.

Under the batch-processing method, the CETA Group was able to draw on contributed computer processing time and on contributed programming services. The processing time and program design service was offered only on a non-priority, time-available basis so that we had to design the handling of the lexicographic information to fit vagaries of such a non-priority, off-site, batch-processing operation. Nevertheless, it was the only way to initiate the system with the funds available. At the beginning in 1971 and 1972, the Secretariat staff was composed of a part-time Executive Secretary and a part-time office typist. Gradually, the budget has expanded to include keyboard operator and Chinese-language specialists so that the Secretariat staff now represents approximately three man-years per year.

The batch-processing method of preparing glossaries is a relatively complex series of steps - more complex because of the need to display graphic output whenever editing or proofing of Chinese characters is required. The system was designed on the assumption that keypunch cards would be the basic machine-readable media to initiate or change information. As many of the readers are aware, there are many problems that can arise from use of keypunch cards, such as cards warped, cards out of order, duplicate cards, cards missing, and so on. We experienced all the problems and would not like to return to them.

Another major problem over the years was the sudden cessation of computer processing by contributors for any one of several reasons. Occasionally, we would find that after a Systems Generation (change in operating system) was performed our programs would no longer run. At one time, it took three months to resolve the problem during which time no processing was done.

At other times, as personnel were shifted or transferred, errors would arise because the new person had not yet learned the idiosyncracies of selecting parameter cards intended to control elements of the programs. Frequent reruns were required. Also, the unexpected absence of processing personnel due to vacations,

illness, etc. produced unpredicted delays. These, along with the fact that every transmission of tape, cards, microfiche or paper, required days in transit, and you have a description of a system that is truly batched. When we look back, we wonder that it worked at all. Nevertheless, as difficult as it was, it was all there was and a great deal was accomplished. The alternative was to do nothing.

In spite of these difficulties, we managed over a ten-year period to create and review a general dictionary file through two complete cycles; the first cycle, approximately 90,000 entries, and the second, approximately 106,000. We are now in the third cycle and expect over 130,000 entries. We have also initiated a considerable number of technical glossaries amounting to over 100,000 entries to add to the file of 500,000 entries in the S&T (Scientific and Technical) collection.

On-line methods

By the early 1980s the cost of computer systems had dropped sufficiently that we were able to consider installing a small system in the Secretariat office. However, running the files to produce the COM (computer output microform) tapes which in turn produced microfiche and paper output, is the most expensive of the processing steps. This element could not be undertaken within the Secretariat office; therefore, whatever system we selected had to produce a file in mag-tape media compatible with standard IBM mag-tape format. This was a serious hindrance for some time.

Accepting the current dependence on outside COM generation of paper and microfiche copy, we ultimately selected the Mohawk Data Systems data entry equipment. It allows us to prepare files, printout alpha-numeric versions for proofing, correct them with a word-processing type of function, sort them as needed, dedupe for combination of multiple entries, and eventually produce an IBM formatted mag tape compatible with the programs and the equipment that produces the COM output for paper and microfiche versions of the files. The on-line system requires far fewer steps. The turn-around is much shorter; it also provides a far greater control of the files in several ways. For instance, the lexicographer sees his changes as he makes them. We do not risk loss of cards in transporting or running them. We also avoid introducing new errors in rewriting entries to replace ones with error since we now correct only the error rather than rekey the entire entry. We can work on files according to our own priorities rather than the priorities of a contributed batch service at another location. It requires a larger CETA Group budget since the budget is now paying for some services that were formerly contributed; however, the on-line system is quite cost-effective when one takes into account the cost of all services. It is also more efficient.

The primary deficiency of the MDS on-line system is the lack of graphic output to display or print the Chinese characters for proofing on the screen, or on paper. Another deficiency is that the system is designed primarily for data entry rather than word processing. For these reasons, and the reasons of lowering computer costs in general, we hope within a few years to replace this system with another that will allow display and output of Chinese

characters as well as sophisticated word-processing functions for even easier maintenance of the files.

One of the most likely candidates for this type of operation is the Xerox 8010 Star processing system. The Star is currently available for processing the Roman alphabet with symbols required for most of the West European languages, and Russian, Japanese and Chinese. It is a very sophisticated, efficient document-storage retrieval and preparation system intended for a paperless office. The J Star (Japanese) system was marketed in Japan last autumn. I have seen demonstrations of the coming C Star (Chinese) system which is quite impressive but not yet released. It is my understanding that it may be as much as another year before the C Star is available. From what I have seen, the coming Star equipment is very fast and very sophisticated and is designed for omni-language processing.

Conclusion

To recapitulate, the CETA Group has very successfully introduced a cooperative voluntary system of building lexicographic files, Chinese to English. This has been accomplished over a decade using largely batch-processing methods. The cost of on-line systems has fallen sufficiently that the CETA Group currently uses an on-line alpha-numeric system for initial file preparation and maintenance. However, batch processing on large computers is still required for creation of formatted tapes for use on COM (computer output microform) equipment to print graphic images on paper or fiche. We look forward to sophisticated multi-language systems such as the Xerox Star System, available with graphic display and laser printer output.