JIM MATHIAS

COOPERATIVE FILE IMPROVEMENT AND USE OF A COMPUTER-BASED CHINESE/ENGLISH DICTIONARY

The CETA (Chinese-English Translation Assistance) Group is an independent organization formed to coordinate development of Chinese to English translation aids and data analysis techniques. It began as an ad hoc body of individuals from State, Commerce, Labor, Office of Education, Defense, Intelligence, Voice of America, Foreign Service Institute, Defense Language Institute, National Science Foundation and Library of Congress. Extension of interest into the scholarly community has broadened academic dimensions to include 43 US and international universities. CETA is developing a computer-based Chinese-English dictionary of current standard terms. It is also exploring tangential topics such as computer processing of Chinese research data, machine translation, and use of the CETA Dictionary file in an on-line computer aid system.

Academic research and development of computer operations in United States' universities has led to capability of computer generation of Chinese characters. Using this capability, CETA printed a 90,000 term dictionary file of Chinese-English entries and has developed a cooperative international process for refining and enriching the file. This process is called the CETA File Improvement System. It is founded on government/academic/private cooperation, designed to edit existing material and add new material. The improvement system is based on collective improvement of the file through a wide sharing of linguistic tasks and the use of computers to store the data and process changes. Thus far, thirty-seven government and forty-three academic linguists and language specialists have committed themselves to review an improvement of the file in return for which they receive the printed copy of the dictionary plus change pages as they are generated. Over 51,000 suggested improvements have been submitted and evaluated and are awaiting update. The File Improvement System proceeds by cycles in which progressively more rigid standards of review are applied.

*

The file will be reprinted in three to five year cycles with change pages issued during interim periods so that participants can share maximum benefits at all times.

When CETA examined the problem of producing a dictionary, it was concluded that significant results could be achieved only by sharing the many tasks involved. It was a forbidding problem, however, the potential for improving dictionaries without waiting 20 years for new editions was a meaningful incentive. The CETA Group issued a hard copy of the 90,000 term Chinese-English listing called The CETA Computer-Based Chinese-English Dictionary. It was produced as a "living" file that could be changed constantly. It was printed by computer – the principal advantages of which were ability to print Chinese characters without typesetting and economy of effort in manipulating the data. The computer could sort in different sequences, make corrections or additions at will, extract particular subsets, and produce a hard copy image of file materials. In a word, it was possible to take the present computer-produced manuscript and give parts of it to volunteers to review and correct or add information. Also it was possible to develop methods for the reviewer to easily prepare changes and for CETA to evaluate and then update the manuscript file.

The first cycle of file review for gross error and duplication has been completed. The reviewers were given a set of instructions to guide them in review of the dictionary material and the preparation of changes or additions. The steps required to process improvements to the CETA Dictionary are, briefly stated, receipt of suggestions for change or addition, preparation for keypunch, computer generation of a prooflist showing original as well as changed entries, manual review of the prooflist, computer selection of approved changes, and update of the computer dictionary file. The application of these steps assures that all changes to the master file will be examined at least once and questionable changes can be held for later review to avoid delaying update actions. As mechanism, the improvement system is quite smooth and under ideal conditions it is possible to change the computer file in a matter of minutes. Under the less than ideal conditions that usually prevail, it is still possible to update and provide current information within a few months rather than the usual 10 year dictionary building and 20 year reissue cycles.

Currently CETA has received and prepared for update a total of 51,000 changes to the 90,000 term file. Since there are more additions than deletions, the new file will be larger by a few percent. More im-

1	CP	沒有充分利用的	-NOT PUT TO FULL USE	22	CP	沒有階級的社會	-CLASSLESS SOCIETY
2	CP	沒有出路	-WITHOUT A WAY OUT	23	CP	沒有預料的	-UNEXPECTED
3	CP	沒有別的辦法	-TO HAVE NO CHOICE BUT	24	GT	沒材料儿	-GOOD FOR NOTHING
4	CP	沒有前例的	-UNPRECEDENTED	25	GT	沒根基	-NO FOUNDATION IN
5	CP	沒有原則的	-UNPRINCIPLED	26	GT	沒武藝	-CANNOT BE HELPED
6	CP	沒有把握	-NOT SURE OF, NOT CONFIDENT	27	GT	沒死活的	-DILIGENTLY
7	CP	沒有擊中目標	-TO MISS ONES TARGET	28	GT	沒气病	-EPILEPSY
8	CP	沒有改變	-UNCHANGED	29	GT	沒水衣服	-A DIVING SUIT
9	CP	沒有根据	-WITHOUT ANY BASIS	30	GT	沒治儿	-NO ALTERNATIVE, HOPELESS
10	CP	沒有止境	-THERE IS NO LIMIT TO	31	GT	沒法儿辦	-CANNOT BE HELPED, UNAVOIDABLE
11	CP	沒有沒能	-FAIL TO, DEVOID OF	32	GT	沒法子	-CANNOT BE HELPED, UNAVOIDABLE
12	CP	沒有理由	-UNREASONABLE, WITHOUT MERIT	33	GT	沒深沒浅	-CARELESS SPEECH AND BEHAVIOR
13	CP	沒有用的	-OF NO USE	34	GT	沒理人家	-AN UNREASONABLE PERSON
14	CP	沒有系統的	-UNSYSTEMATIC	35	GT	沒理矯理	-TO DRAW A FORCED INFERENCE
15	CP	沒有結果	-WITH NO RESULT	36	GT	沒甚么	-NONE, NOTHING, DONT MENTION IT
16	CP	沒有緊張局面的	-TENSION-FREE	37	GT	沒用	-USELESS, OF NO USED, FUTILE
17	CP	沒有被否認	-UNCHALLENGED	38	GT	沒的	-WITHOUT ANY REASON, USELESSLY
18	CP	沒有誠意	-INSINCERE	39	GT	沒皮沒臉的	-SHAMELESS
19	CP	沒有課決的問題	-UNSETTLED QUESTION	40	GT	沒皮賴臉	-SHAMELESS
20	CP	沒有資格	-NOT QUALIFIED TO	41	GT	沒眼色儿	-STUPIDLY HONEST
21	CP	沒有達到	-UNABLE TO REACH	42	GT	沒短的	-CEASELESSLY

Fig. 1. *Computer Printed Chinese Characters.*

portant, the greatest error will have been removed and the file will be prepared for the next cycle which will emphasize the further enrichment of the lexical content, addition of grammatic information, incorporation of restrictive and stylistic labels, and identification of agglutinated phrases. The second printing of the dictionary manuscript will include Pin Yin romanization with tone and telecode numbers as well as the customary English gloss and source information. The character vector file has been significantly updated so that it now contains capabilities of drawing approximately 10,500 characters. It will be continually updated through the dictionary review cycles. See Figure 1 Computer Printed Chinese Characters.

The file will also be available as the core of an on-line computer aid. Prototype computer aid functions have been developed which illustrate the ways in which a computer file can be used in an interactive mode to help a translator. They use input by telecode and romanization and graphic input is simulated. A cathode ray tube is used to display Chinese characters, romanizations (Pin Yin, Wade-Giles, Yale), the radical number plus residual stroke count, English meaning for the string and meaning for segments of the strings. Also developed is an automatic segmenting function which is the operation of breaking a string of characters into single characters and into segments of continuous characters (for synthesis of meaning form component parts). See Figure 2 Graphic Display.

	CHARACTER SEQUENCE				SEGMENT	ENGLISH
	(1)	(2)	(3)	(4)	IDENTIFIER	MEANING
STC	2693	3111	1714	2348	1-4	Diesel Engine
					2-4	-
	柴	油	引	擎	3-4	Engine
					4	To Raise (W) Yin...Engine
					1-3	-
P	1CHAI	2YOU	3YIN	2QING	2-3	-
W	CH'AI	YU	YIN	CH'ING	3	Lead, Draw, Attract
Y	CHAI	YOU	YIN	CHING	1-2	Diesel Oil, Fuel in
R	75.5	85.5	57.1	64.13		General
T	9	8	4	17	2	Oil, Grease
					1	Fire Wood

Fig. 2. *Graphic display.*

CETA hopes to test this system further using a refined data base for evaluation of its potential for shared access by a wide government and academic community.

CETA started with a poor dictionary but it was machineable. There are a lot of good dictionaries that are not machine readable and, therefore, difficult to change or consolidate. CETA is putting these things together by use of a wholly unique method; the voluntary cooperation of interested government and academic scholars and language specialists. The reward to participants is: 1) awareness of contribution to a worthwhile effort, 2) an up-to-date hard copy of the CETA computer file containing all the latest contributions by all participants and 3) use of the CETA Secretariat Office to search out and exchange information of common concern. The only cost is willingness to share in the work of CETA.