[From: *Overcoming the language barrier*, 3-6 May 1977, vol.1 (München: Verlag Dokumentation, 1977)]

COMPUTER TRANSLATION OF CHINESE SCIENTIFIC JOURNALS

S.C. Loh & L. Kong

Machine Translation Project

The Chinese University of Hong Kong

Abstract

A natural language processor, called CULT, (Chinese University Language Translator), capable of translating Chinese scientific texts into readable English has been developed during the past six years at the Chinese University of Hong Kong. The system has been modified, improved and rigorously tested, and its potential and capabilities are amply demonstrated and realised.

Since January 1975, the CULT system has been used, on a regular basis, to translate the Chinese mathematical journal, namely, ACTA MATHEMATICA SINICA, which is published by the Chinese Academy of Sciences at Peking.

The design of CULT is briefly discussed in the paper and the system is written in USA Standard FORTRAN. This program may be run on any computer possessing a core memory of about 32K words and a backing storage of about 1 megabyte.

1. INTRODUCTION

Machine translation (MT) research at the Chinese University was initiated by the authors in late 1969. The aim was to study the possibility of automatic translation from Chinese into English by computer techniques. The preliminary survey, after studied the sentence structures and linguistic complexity of the Chinese scientific texts, particularly in the field of mathematics, indicates that mathematics being an exact science, is relatively easy to translate and the Fully Automatic High-Quality Translation (FAHQT), as suggested by Bar-Hillel, is extremely difficult, if not impossible, to achieve, but some practical machine translation system involving the collaboration of man and machine is feasible.

The first natural language processor, CULT, was designed and tested in October 1972. In subsequent years, many improvements and modifications have been made to the system. Difficulties encountered in linguistics and programming have been identified and rectified. As a result, CULT has become more refined in structure and efficient in operation.

Since January 1975, CULT has been used on a regular basis, in translating the Chinese mathematical journal, namely, ACTA MATHEMATICA SINICA, published quarterly by the Academy of Sciences, Peking. Also ACTA PHYSICA SINICA has been translated.

2. TRANSLATION PROCEDURES

The natural language processor, CULT, consists of a main program and subroutines.

The Chinese sentences are first converted into telegraphic codes and these are input into the machine together with any quotations or equations. A subroutine substitutes the quotations with 4-digit dummy codes. The numeric codes are stored in a disc file and the quotations in another file. One record in the file contains one sentence (Each sentence is separated by a 5-digit code ranging from 10100 to 10800).

The sentences are analysed one at a time. A subroutine is used to make comparisons between the telegraphic codes and the entries in the machine translation dictionary. Words are found from individual characters of the input by the 'Largest Match' principle. Another subroutine is called to extract the function codes of each word from the dictionary. At this stage a preliminary analysis is performed on the words, e.g. words that can act as a noun or a verb are analysed and one unique function is assigned to it in the particular sentence under consideration.

A second pass is now made to check for punctuation marks. An interrogative sentence is translated differently from one which is affirmative in nature. This information is stored in the program for later use. A third pass is made to search for pre-positional phrases. Each word in the phrase is translated by an appropriate subroutine (according to the parts of speech) and stored. A fourth pass is then made to search for the verb. At the moment only sentences with at least one verb can be translated.

The MT program makes a fifth pass and checks the first word of the sentence. If it is a conjunction it is translated and stored. If it is *a* verb, then the predicate is analysed. If a subject exists, the phrase before the verb is taken out and translated. Each word is analysed by the appropriate subprogram depending on the function codes. The most suitable form for the target word is chosen, and order, articles and auxiliary words are also determined. Any adverbs immediately before the verb, however, are left alone. These adverbs are passed to that part of the system which analyses the predicate before they are translated.

The analysis of the predicate determines first the adverbs. The adverbs are translated and their positions in the sentence determined, i.e. whether it should be placed before or after the verb. Next the program searches for auxiliary verbs. The auxiliary verbs are translated and their order rearranged. Then the MT program makes another pass to determine the main verb. At this point the voice and tense of the sentence is determined. The appropriate verb form is selected from the dictionary.

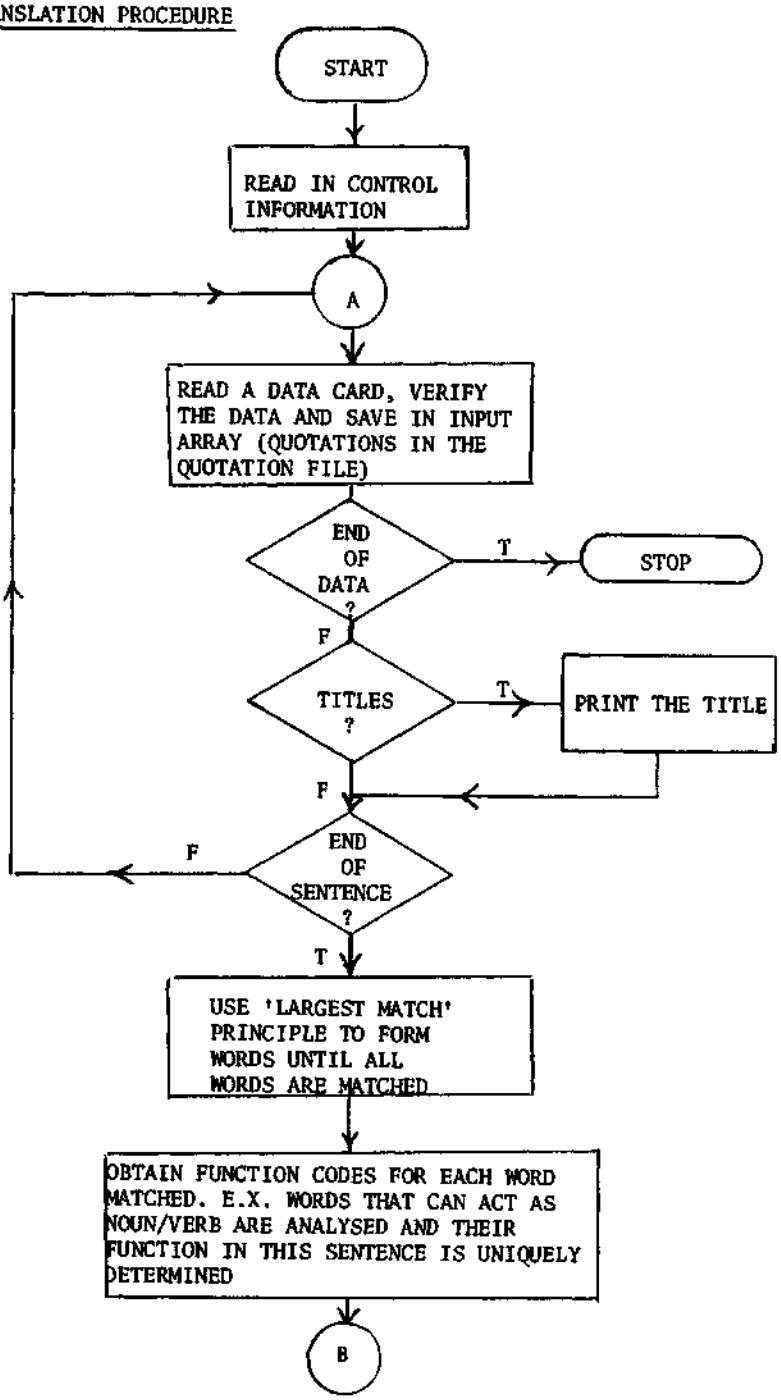
Now the program makes a second check on the punctuation. Auxiliary words are added or the verb position is changed if an interrogative sentence is being translated. Then the verb is checked to determine if it is transitive or intransitive. A transitive verb directs the program to analyse the objects. The number of objects is checked and each is translated separately. An intransitive verb is checked to see whether it possesses any complements. A complement is translated if it exists; otherwise the punctuation is checked again.

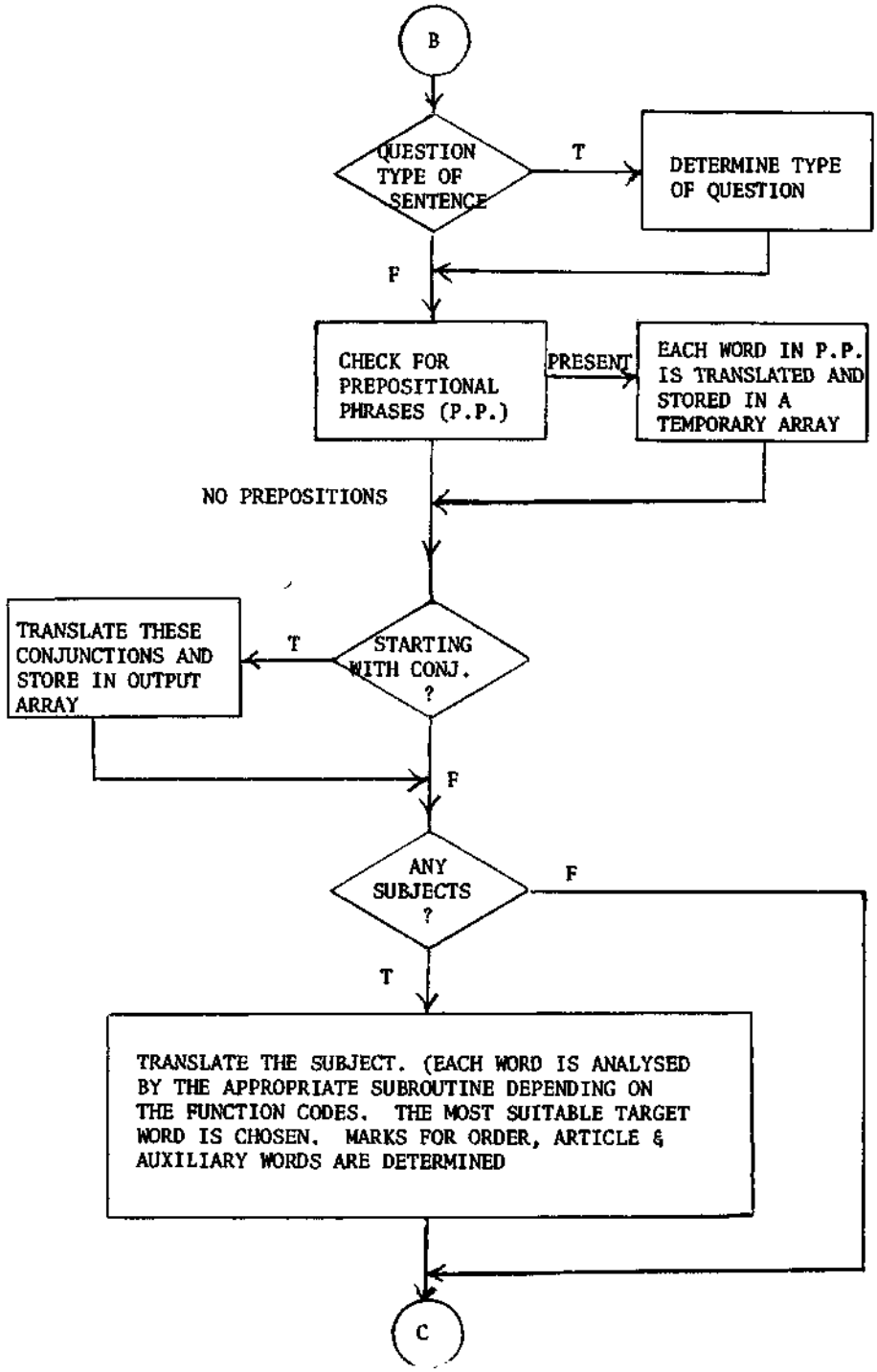
This checking of the punctuation determines the mark for the end of the sentence or just a phrase separator. Three cases are considered at this point. If it is a phrase separator, the MT program goes back to the point of translating objects. If it is an end of sentence marker it checks to see whether a question is being translated again and word positions are changed if necessary. If it is neither a phrase separator nor a marker, the program checks whether the word is another verb or not. If it is a verb, it is translated (in the infinitive form). If it is not a verb, then somewhere an error has occurred and the program prints out an error message and goes on to the next sentence. The absence of errors sends the program back again to the point of the analysis of objects.

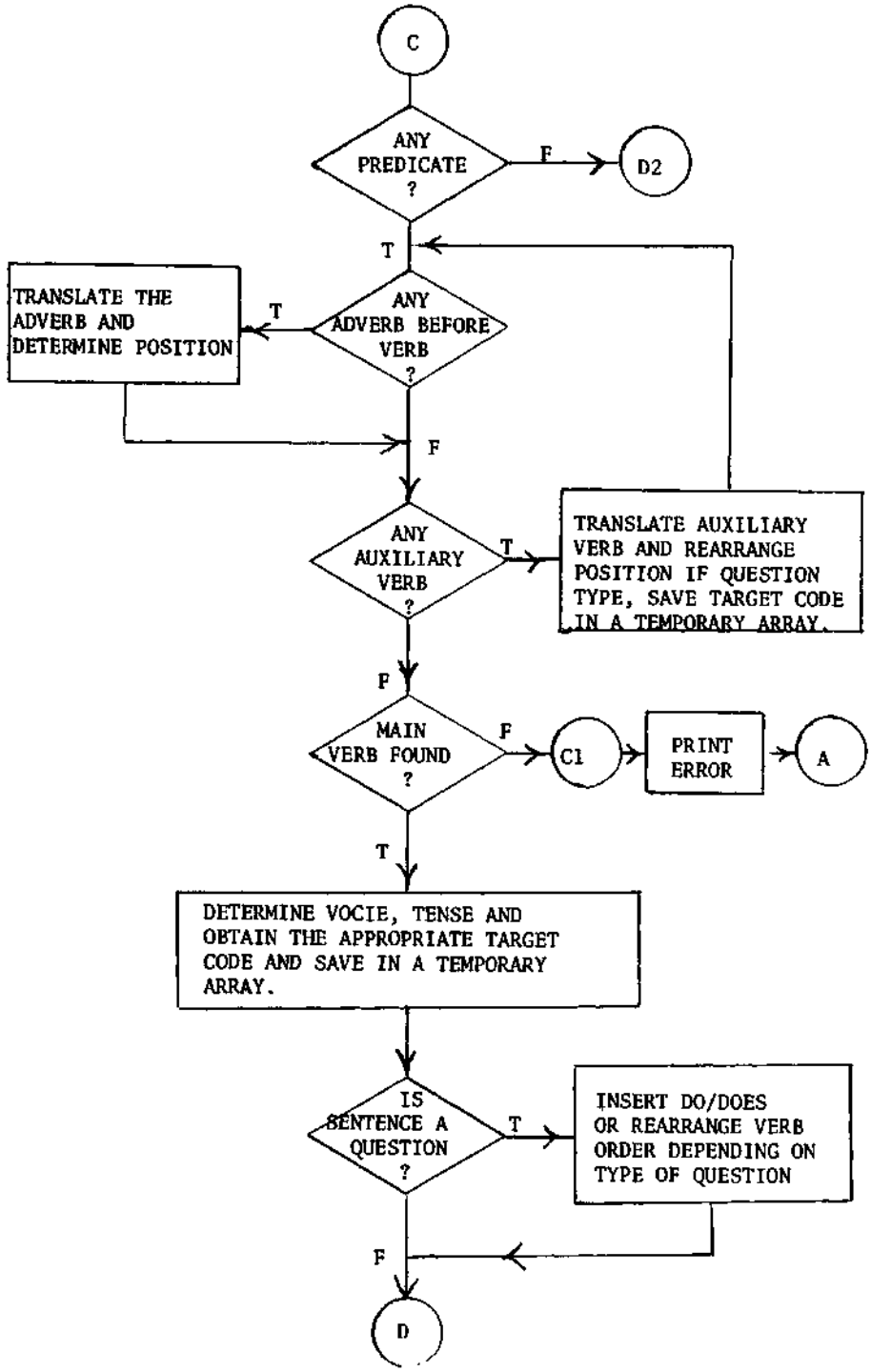
Finally, the whole sentence is analysed and translated. As the various stages of the translation process stores the translated target words in different storage arrays, these are now checked so that all output is placed in a single, final storage. The punctuation mark is obtained from the dictionary also. An essay-forming subroutine arranges the translated sentence, together with any already translated sentences, into paragraphs. Any equations that need to be inserted in the paragraph are also taken from the storage file and the paragraph is stored in an output file. The MT program then checks if all the input sentences have been translated. The final step is to print out the target sentences, properly arranged, onto paper from the output file. The program then stops.

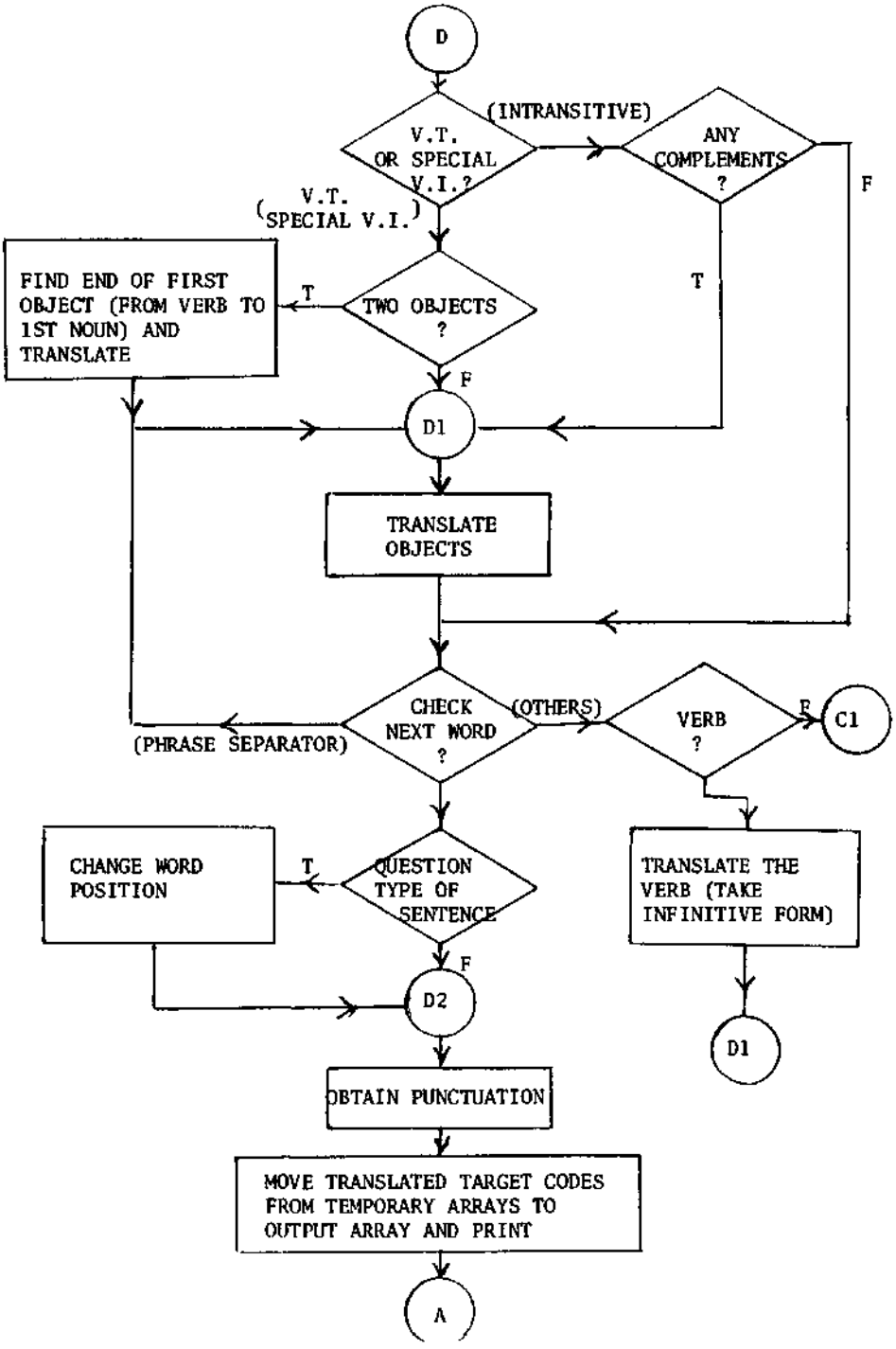
For simplicity, the error detection has not been shown in this flowchart, nevertheless, if an error occurs at any stage of the translation, an error routine will be called and the types of errors will be printed together with the grammatical information of the words. The present sentence then deleted and a new sentence is read.

TRANSLATION PROCEDURE




B
QUESTION TYPE OF SENTENCE
T
DETERMINE TYPE OF QUESTION
F
CHECK FOR PREPOSITIONAL PHRASES (P.P.)
PRESENT
EACH WORD IN P.P. IS TRANSLATED AND STORED IN A TEMPORARY ARRAY
NO PREPOSITIONS
TRANSLATE THESE CONJUNCTIONS AND STORE IN OUTPUT ARRAY
T
STARTING WITH CONJ. ?
F
ANY SUBJECTS ?
F
T
TRANSLATE THE SUBJECT. (EACH WORD IS ANALYSED BY THE APPROPRIATE SUBROUTINE DEPENDING ON THE FUNCTION CODES. THE MOST SUITABLE TARGET WORD IS CHOSEN. MARKS FOR ORDER, ARTICLE & AUXILIARY WORDS ARE DETERMINED
C


C
ANY PREDICATE ?
F
D2
T
TRANSLATE THE ADVERB AND DETERMINE POSITION
T
ANY ADVERB BEFORE VERB ?
F
ANY AUXILIARY VERB ?
T
TRANSLATE AUXILIARY VERB AND REARRANGE POSITION IF QUESTION TYPE, SAVE TARGET CODE IN A TEMPORARY ARRAY.
F
MAIN VERB FOUND ?
F
C1
PRINT ERROR
A
T
DETERMINE VOCIE, TENSE AND OBTAIN THE APPROPRIATE TARGET CODE AND SAVE IN A TEMPORARY ARRAY.
IS SENTENCE A QUESTION ?
T
INSERT DO/DOES OR REARRANGE VERB ORDER DEPENDING ON TYPE OF QUESTION
F
D


D
V.T. OR SPECIAL V.I.?
(INTRANSITIVE)
ANY COMPLEMENTS ?
F
T
V.T. (SPECIAL V.I.)
FIND END OF FIRST OBJECT (FROM VERB TO 1ST NOUN) AND TRANSLATE
T
TWO OBJECTS ?
F
D1
TRANSLATE OBJECTS
CHECK NEXT WORD ?
(OTHERS)
VERB ?
F
C1
(PHRASE SEPARATOR)
CHANGE WORD POSITION
T
QUESTION TYPE OF SENTENCE
F
TRANSLATE THE VERB (TAKE INFINITIVE FORM)
D2
D1
OBTAIN PUNCTUATION
MOVE TRANSLATED TARGET CODES FROM TEMPORARY ARRAYS TO OUTPUT ARRAY AND PRINT
A

3. AN EXAMPLE

The following example illustrates the procedures of translation.

SOURCE SENTENCE :

定理3的結果證明在所有的情況下每一個充份地大的偶數

都能夠表示爲一個素數及另一個實數之總和。

(A) FIRST-STEP : INPUT

Convert all the Chinese characters and quotations into telegraphic codes, punch the codes on 80 columns cards and input to the translator.

定	理	3	的	結	果	證	明
1353	3810	9963	4104	4814	2654	6086	2494
在	所	有	的	情	況	下	每
0961	2076	2589	4104	1906	0400	0007	3008
一	個	充	份	地	大	的	偶
0001	0222	0339	0118	0966	1129	4104	0260
數	都	能	夠	表	示	爲	一
2422	6757	5174	1124	5903	4355	3634	0001
個	素	數	及	另	一	個	實
0222	4790	2422	0644	0659	0001	0222	1395
數	之	總	和	。			
2422	0037	4920	0735	10700			

Each code occupies 5 columns, and a maximum of 15 codes can be punched on one card. The last five columns of the card are reserved for "card identification."

If an illegal character is found in a card, the translator will output the error message :

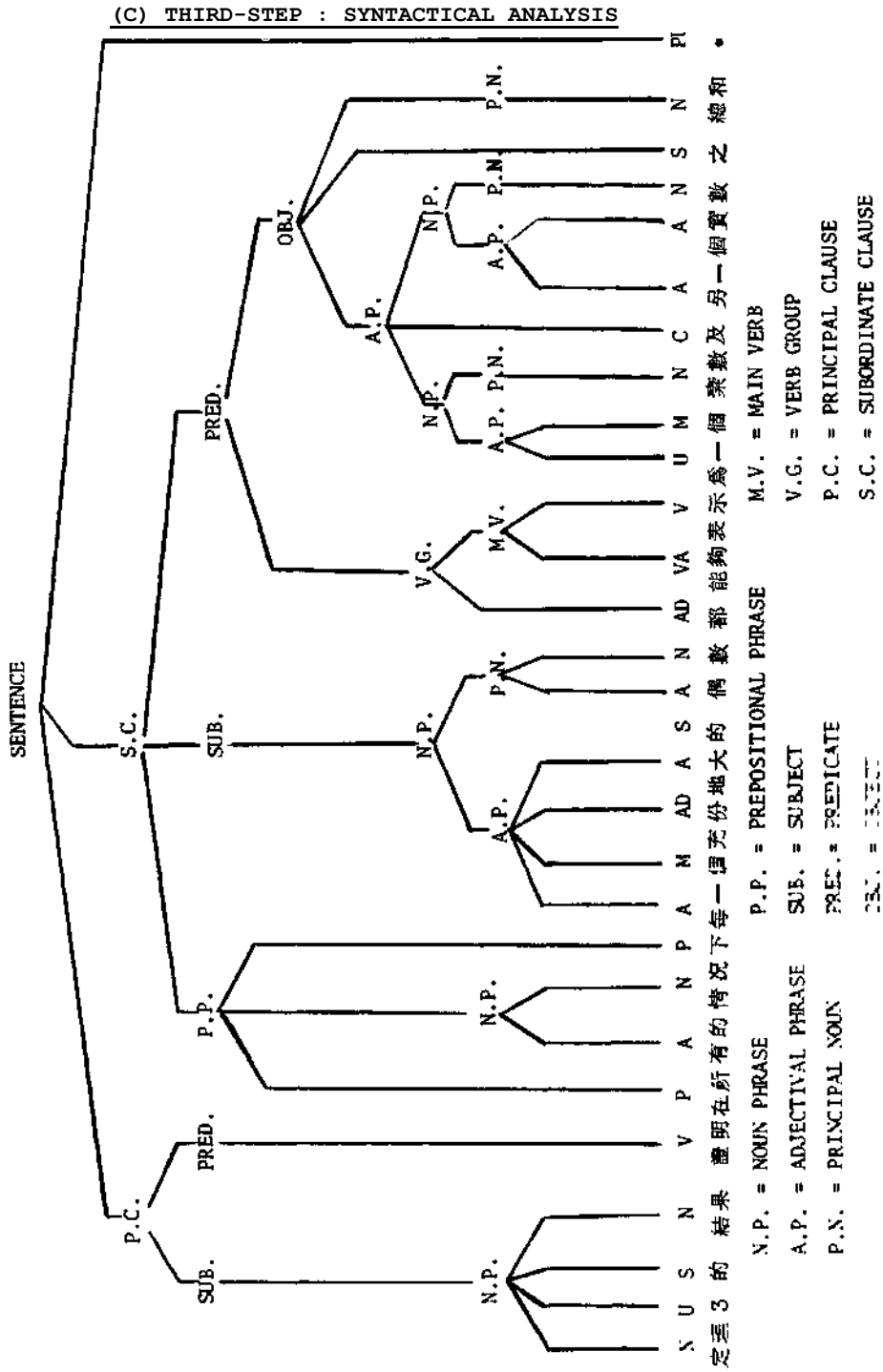
"ILLEGAL CHARACTER FOUND :" followed by the card identification of the card and the card is rejected.

(B) SECOND-STEP : LARGEST MATCH

Every characters of the input sentence are grouped together to form words by using the "largest-match" principle. After a largest word is found, the grammatical information and the English equivalent of this word is transferred from dictionary into core memory.

WORD NO.	WORDS	PART OF SPEECH	TARGET CODES
1	定理	NOUN	THEORM
2	3	NUMERAL	3
3	的	SPECIAL	
4	結果	NOUN	RESULT
5	證明	VERB	PROVE
6	在	PREPOSITION	IN
7	所有的	ADJECTIVE	ALL
8	情況	NOUN	CASE
9	下	PREPOSITION	IN
10	每一	ADJECTIVE	EVERY
11	個	MEASURE	
12	充份地	ADVERB	SUFFICIENTLY
13	大	ADJECTIVE	LARGE
14	的	SPECIAL	
15	偶	ADJECTIVE	EVEN
16	數	NOUN	NUMBER
17	都	ADVERB	ALL
18	能夠	AUX. VERB	CAN
19	表示爲	VERB	REPRESENT
20	一	NUMERAL	ONE, A, AN
21	個	MEASURE	
22	素數	NOUN	PRIME NUMBER
23	及	CONJUNCTION	AND
24	另一個	ADJECTIVE	ANOTHER
25	實	ADJECTIVE	REAL
26	數	NOUN	NUMBER
27	之	SPECIAL	
28	總和	NOUN	SUM
29	。	PUNCTUATION	.

(C) THIRD-STEP : SYNTACTICAL ANALYSIS



(D) FOURTH-STEP : RE-ARRANGEMENT

Re-arrange Prepositional Phrases

定理3的結果證明每一個充份地大的偶數都能夠表示
爲一個素數及另一個實數之總和在所有的情況下。

(E) FIFTH-STEP ; OUTPUT

The corresponding target codes of each word are moved to the output array and print out.

THE RESULT OF THE THEORM 3 PROVES THAT EVERY SUFFICIENTLY LARGE EVEN NUMBER CAN ALL BE REPRESENTED AS THE SUM OF A PRIME NUMBER AND ANOTHER REAL NUMBER IN ALL CASES.

The average speed of translation is approximately 2-3 sentences per CPU second.

4. SOME REMARKS

A great deal of difficulties which have been experienced have not been with the machine translation itself, but with the proper or correct translation of scientific and technical terms used in the publications. This is true particularly in the field of Physics, as no up-to-date appropriate dictionary is available.

The application of CULT to other subject fields is being studied as well as to other natural languages.

5. ACKNOWLEDGEMENT

The authors are grateful to The Asia Foundation and Rockefeller Brothers Fund for their financial grant given to the Machine Translation Project.