MACHINE TRANSLATION OF CHINESE MATHEMATICAL ARTICLES

Shiu-Chang Loh, Luan Kong, and Hing-Sum Hung
(Department of Computer Science,
The Chinese University of Hong Kong)

Abstract

A practical machine translation system called CULT (Chinese University Language Translator), capable of translating Chinese mathematical texts into readable English, has been developed during the period 1969-77 at the Chinese University of Hong Kong. The design of CULT is based on the algorithm discussed in Section 3; programs for the system are written in Standard FORTRAN and run on the ICL1904A computer system. This system has been modified, improved, and rigorously tested, and its potential and capabilities have been amply demonstrated. Since January 1975 CULT has been used on a regular basis to translate ***Acta Mathematica Sinica,*** a scientific journal which is published by the Chinese Academy of Science in Peking.

* *

Introduction

Work in Chinese-English machine translation (MT) has been undertaken by a number of research groups in the United States {[1]}. Recently, MT systems have been developed at Berkeley and LATSEC, Inc. for translating Chinese into English {2,3}. The Chinese University of Hong Kong began its formal studies in machine translation around 1969 with primary emphasis being placed on Chinese-English MT of scientific texts in the field of mathematics {[4,5,6,7,8]}. By confining ourselves to the translation of materials in such a specific field the task of dictionary compilation, and writing analysis and synthesis routines is made considerably simpler. The dictionary coverage can be limited to the required vocabulary and it is an observed fact that the syntactic style of scientific authors is usually simple. It is generally acknowledged, as first pointed out by Bar-Hillel, that fully automatic high-quality translation (FAHQT) is not feasible, not even for scientific texts {[9]}. The best possible solution for MT, both theoretically and practically, is symbiosis between man and machine {[1]}. Most research groups working on Chinese-English machine translation take the approach of post-editing machine-produced texts. The basic approach of the CUHK group centres around the machine-pre-editor partnership. By pre-editing the source text, the semantic and lexical analysis by machine will be much simpler and more effective. Post-editing the target text is unnecessary. Research in Chinese-English machine translation at the Chinese University has led to the development of a practical MT system called CULT.

Compilation of the MT Dictionary

The CUHK group has compiled a Chinese-English computer dictionary to be incorporated in the process of machine translation. This dictionary is primarily for mathematics, and currently contains approximately 30,000 items. An item consists of one or more Chinese characters which can be a word or a phrase. Only a single meaning can be assigned to each item and special efforts have been made

to select the most appropriate definition. The dictionary resides on disc, items in the dictionary being entered arbitrarily. Entries for items beginning with the same character are grouped together in such a way that allows for the Largest Match Principle (see below), the addition of new entries to the dictionary, and the deletion of entries from the dictionary in a simple manner.

Each entry of the computer dictionary includes the source codes, the function codes, and the target codes. They are represented in different formats in the dictionary as appropriate for their specific use in MT automatic procedures. Source codes are used to represent the source item, each character of the source item being represented by its equivalent telegraph code (see below). Function codes provide information for grammatical analysis of the source item and the selection of the target codes. Target codes give only a single meaning for the source item but all inflected forms of the English equivalent are listed.

Translation Procedure

Pre-editing is the work done before inputting the Chinese source text into the computer for translation. The amount of work done is very little; it involves only the insertion of syntactic and semantic indicators. With the pre-editing of the the source text, post-editing of the target text can be eliminated altogether since we can expect the output to be correct. .

After the Chinese source text has been scanned and pre-edited it must be coded, for which standard 4-digit Chinese telegraph codes (telecodes) are used. This method of coding provides a unique, one-to-one representation for each character. Non-Chinese characters and mathematical formulas are treated with special formats.

After the pre-edited source text has been coded, it is input to the machine for translation. As soon as the text is loaded and the control program is activated CULT will perform each step of the translation procedure from dictionary look-up, through source language analysis, to target language synthesis, until a complete syntactical analysis has been achieved and the entire input text has been translated and printed out. The average speed of translation is approximately 60 words per CPU second. Finally, appropriate target language items (those treated with special formats) missing in the source text are inserted in the computer printout.

CULT consists of a number of routines which will be called by the control program, MASTER. Under the supervision of MASTER the source text is read in by the input routine, INPUT, routine MATCH is then activated to search the MT dictionary for the items of the input sentence by using the Largest Match Principle. A one-character item is first tested against the input sentence. When a match is found, the next character is tested in combination with the first, and so on. Items of more than six characters are very rare, whereas items of two or three characters are the most common. The time required to look up an item in the dictionary is therefore quite short by using this method. If an item is found during the matching process, then the function codes and the target codes associated with this item are transferred from the dictionary into the code memory; otherwise, the error indicator in routine ERROR is set. This process continues until the end of the sentence. If no error occurs the routine PARSER is initiated to perform syntactic and semantic analysis based on the information transferred; otherwise, routine ERROR is called to print out error messages and the translation of a new sentence is started.

The analysis of the sentence mainly consists of three parses performed by various routines under the instruction of the routine PARSER.

Routine PARSER makes the first parse to analyze the relationship of the items so as to determine their relative parts of speech. Then they are classified into different grammatical categories such as noun group, verb group, phrase, or clause. In the course of analysis, if the source sentence is found to be ungrammatical, routine ERROR is called to print out messages together with some analytical information; further analysis of the sentence will not be carried out and PARSER passes control over to MASTER. In any case, the results of the analysis are saved in arrays for further reference.

In the second parse, the relative item orders are analyzed, based on the structure of an English sentence. In doing so, routine PARSER first determines the relative item orders of different grammatical categories by analyzing the actual functions of the phrases or clauses and their relationships with the noun groups and verb groups. The relative item orders of verbs and adverbs are then determined depending on the type of sentence and the kind of adverbs. Items of each noun group are also analyzed and their relative item orders are determined accordingly. At the end of this parse, the items of the sentence are rearranged in such a way that it is a pattern of an English sentence.

Finally, PARSER makes the third parse to determine the English equivalent (target codes) of the source sentence. The determination of the target codes of a certain item in the source sentence largely depends on its part of speech and to which grammatical category it belongs. For each category, PARSER will identify the principal item of the category. The rest of the items of the same category are capable of giving constructive information for translating the principal item. The target codes of these items can be determined individually. In dealing with the principal item, PARSER determines not only its target codes but also some of its important properties which will then be used to translate items in other categories of the sentence. Once the target codes of the source sentence are determined, the routine PRINT is called to move the target codes to the output array, arrange them in a pre-determined form, and then output on the line printer. Control is then transferred back to MASTER and the process is repeated until the source text is completely translated. For the entire translation procedure discussed in this section see Figure 1.

Further Remarks

Many of the difficulties which we experienced in translating the Chinese mathematics articles were not with the language translator itself but with the proper or correct translation of the mathematical terms. Consequently, a bilingual glossary of mathematical and computing science terms was compiled, a total of 33,000 terms being collected and translated.

Obviously, many unsolved language problems, such as the insertion of the definite and indefinite articles, and multiple meanings, to name but two, are still present and pre-editing is used to resolve these problems.

At present CULT is used to translate articles not only in mathematics but also in other subjects. A study has been made of the translation of English text into Chinese.

Acknowledgements

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

References

{1} *H.H. Josselson, 'Automatic Translation of Languages since 1960: A Linguist's View',* Advances in Computers, *11 (1971).*

{2} *W. S.-Y. Wang, S.W. Chan, et al.,* Chinese-English Machine Translation System, *Report RADC-TR-75-109 (University of California, Berkeley, California, 1975).*

{3} *P. Toma,* The SYSTRAN System, *FBIS Seminar on MT (1976).*

{4} *S.-C. Loh, 'Machine Translation at the Chinese University of Hong Kong',* Proceedings of the CETA (Chinese-English Translation Assistance) Workshop on Chinese Language and Chinese Research Materials, *CETA-72-01 (1972).*

{5} *S.-C. Loh,* Final Report on Machine Translation, Machine Translation Project *(CUHK, 1975).*

{6} *S.-C. Loh, 'CULT (Chinese University Language Translator), FBIS Seminar on MT, 1976',* American Journal of Computational Linguistics *(1976).*

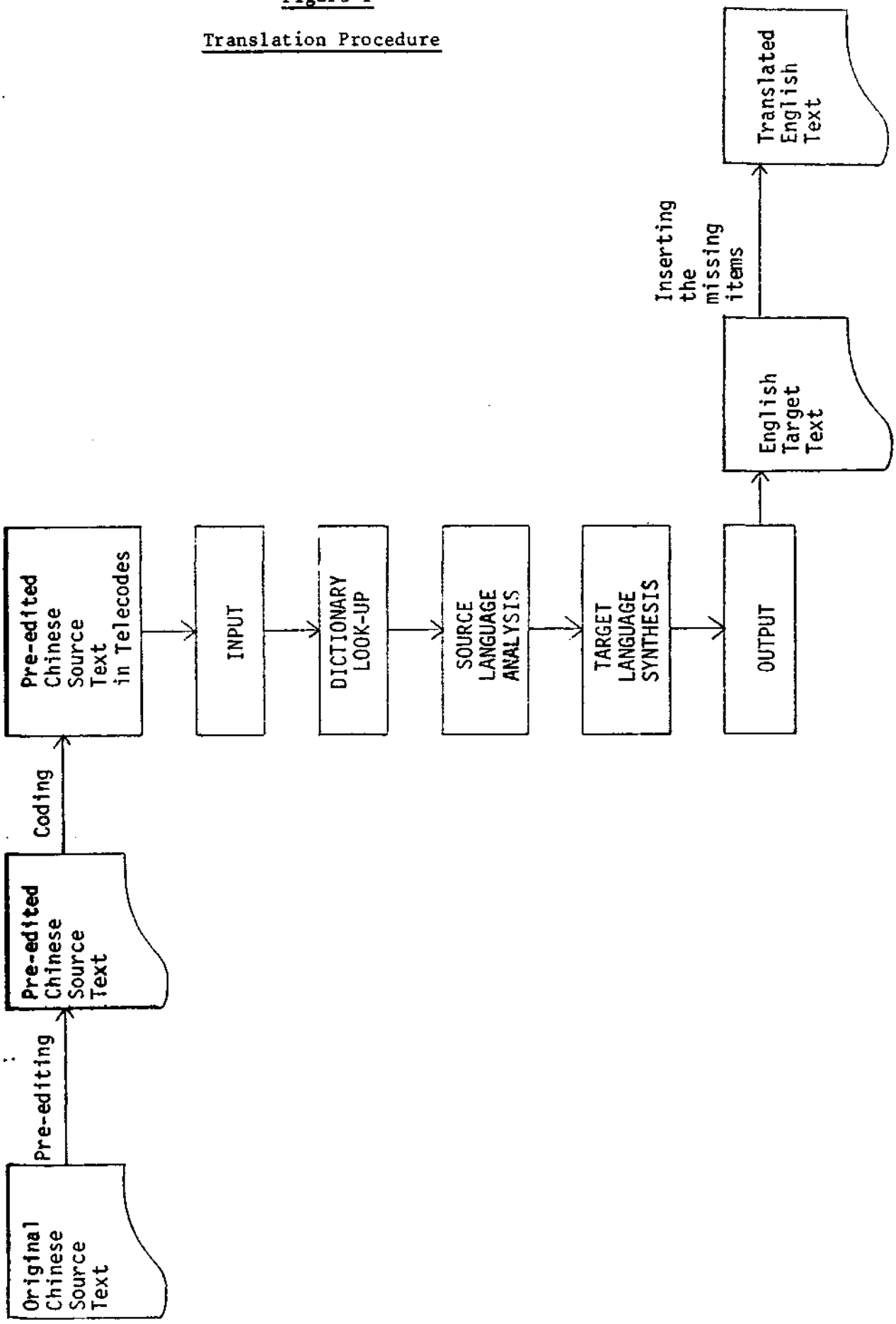
{7} *S.-C. Loh, 'Machine Translation: Past, Present, and Future',* ALLC Bulletin *(March, 1976).*

{8} *S.-C. Loh, and L. Kong, 'Computer Translation of Chinese Scientific Journals',* Proceedings of the Third European Congress on Information Systems and Networks Overcoming the Language Barrier' *(Luxembourg, 1977).*

{9} *Y. Bar-Hillel, 'The Present Status in Automatic Translation of Languages',* Advances in Computers, *1 (1960).*

Figure 1

Translation Procedure



Example

The following example illustrates the translation procedure.

Source sentence :

每 一 阿 貝尔 群 G 可 被 視 作 爲 一 Z - 模 如 果 對 于 $\alpha \varepsilon G$ 和 $n \varepsilon Z$ 我 們 定 義 $n\alpha = \alpha^n$.

Step 1. Pre-editing

每 一 阿 貝尔 群 G 可 被 視 作 爲 一 Z - 模 如 果 對 于 $\alpha \varepsilon G$ 和 $n \varepsilon Z$] 我 們 定 義 $n\alpha = \alpha^n$.

In the above pre-edited sentence, the syntactic indicator, "]" is used to indicate the end of the prepositional phrase, 對 于 $\alpha \varepsilon G$ 和 $n \varepsilon Z$.

Step 2. Coding

Converts all the Chinese characters into telegraph codes, and uses B00i to reserve i blanks for later insertion of G, Z-, $\alpha \varepsilon G$, $n \varepsilon Z$, $n\alpha = \alpha^n$, respectively, in Step 8.

每	一	阿	貝	尔	群	G	可	被	視
3020	0001	7093	6296	1422	5028	B001	0668	5926	6018
作	爲	一	Z -	模	如	果	對	于	$\alpha \varepsilon G$
0155	3634	0001	B002	2875	1172	2654	1417	0060	B003
和	$n \varepsilon Z$	]	我	們	定	義	$n\alpha=\alpha^n$	.	
0735	B003	]	2053	0226	1353	5030	B005	9978	

Step 3. Input

Punches the codes of the source sentence on cards and inputs them to the translator.

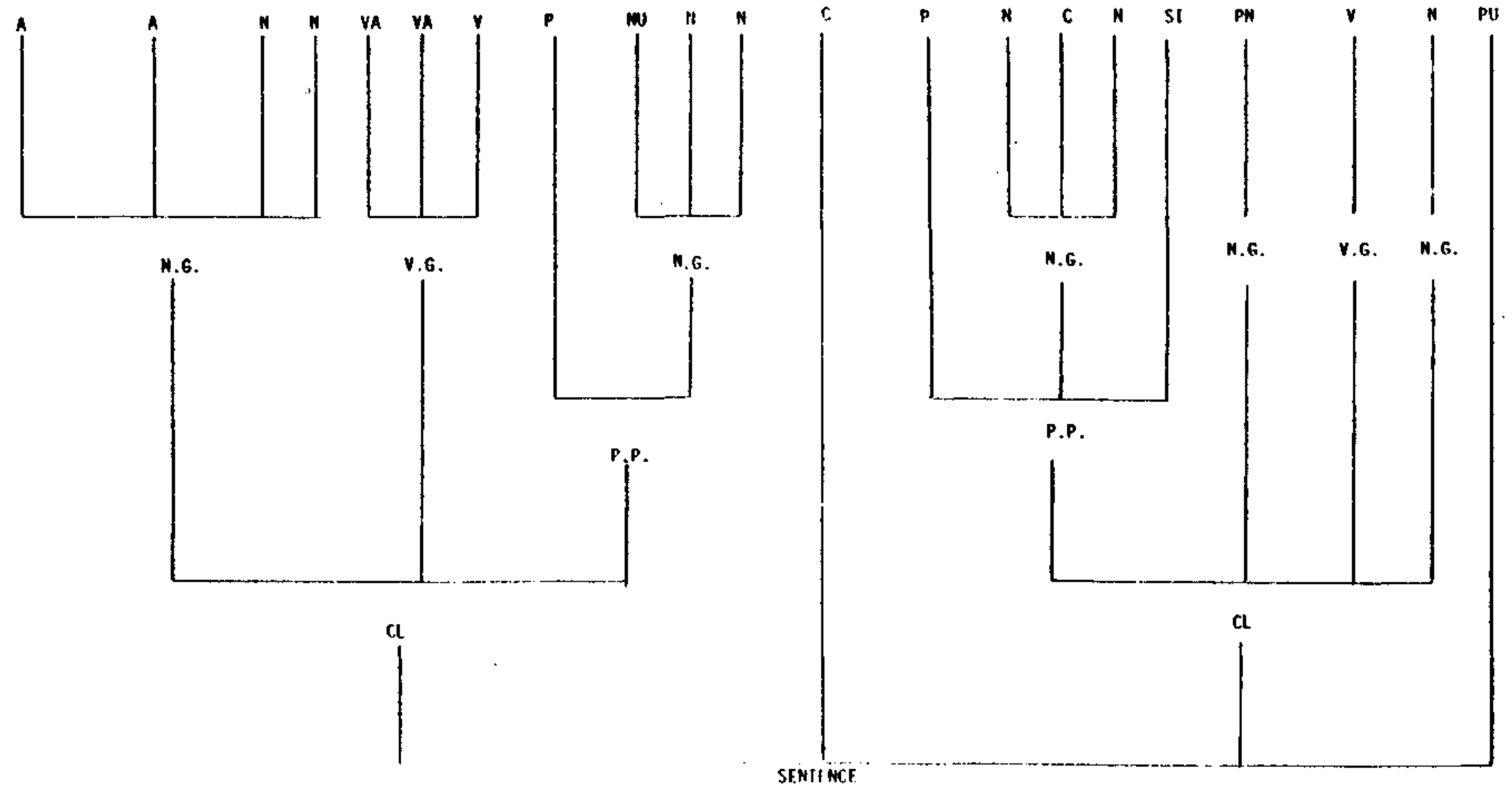
Step 4. Dictionary look-up

The dictionary look-up procedure is carried out based on the Largest Match Principle.

每	一	阿	貝	尔、	群	G	可	被	視
3020	0001	7093	6296	1422	5028	B001	0668	5926	6018
作	爲	一	Z -	域	如	果	對	于	$\alpha\varepsilon G$
0155	3634	0001	B002	2875	1172	2654	1417	0060	B003
和	$n\varepsilon Z$	]	我	們	定	義	$n\alpha=\alpha^n$	.	
0735	B003	]	2053	0226	1353	5030	B005	9978	

STEP 5. Source Language Analysis (First parse)

Pre-edited Source Sentence	每 一 阿 貝 尔 群 G 可 被 視 作 為 一 Z- 模 如 果 對 于 n∈G 和 n∈Z] 我 們 定 義 na=a^n .
Source Codes	3020 0001 7093 6296 1422 5028 B001 0668 5926 6018 0155 3634 0001 B302 2875 1172 2654 1417 0060 B003 0735 B003] 2053 0226 1353 5030 B005 9978



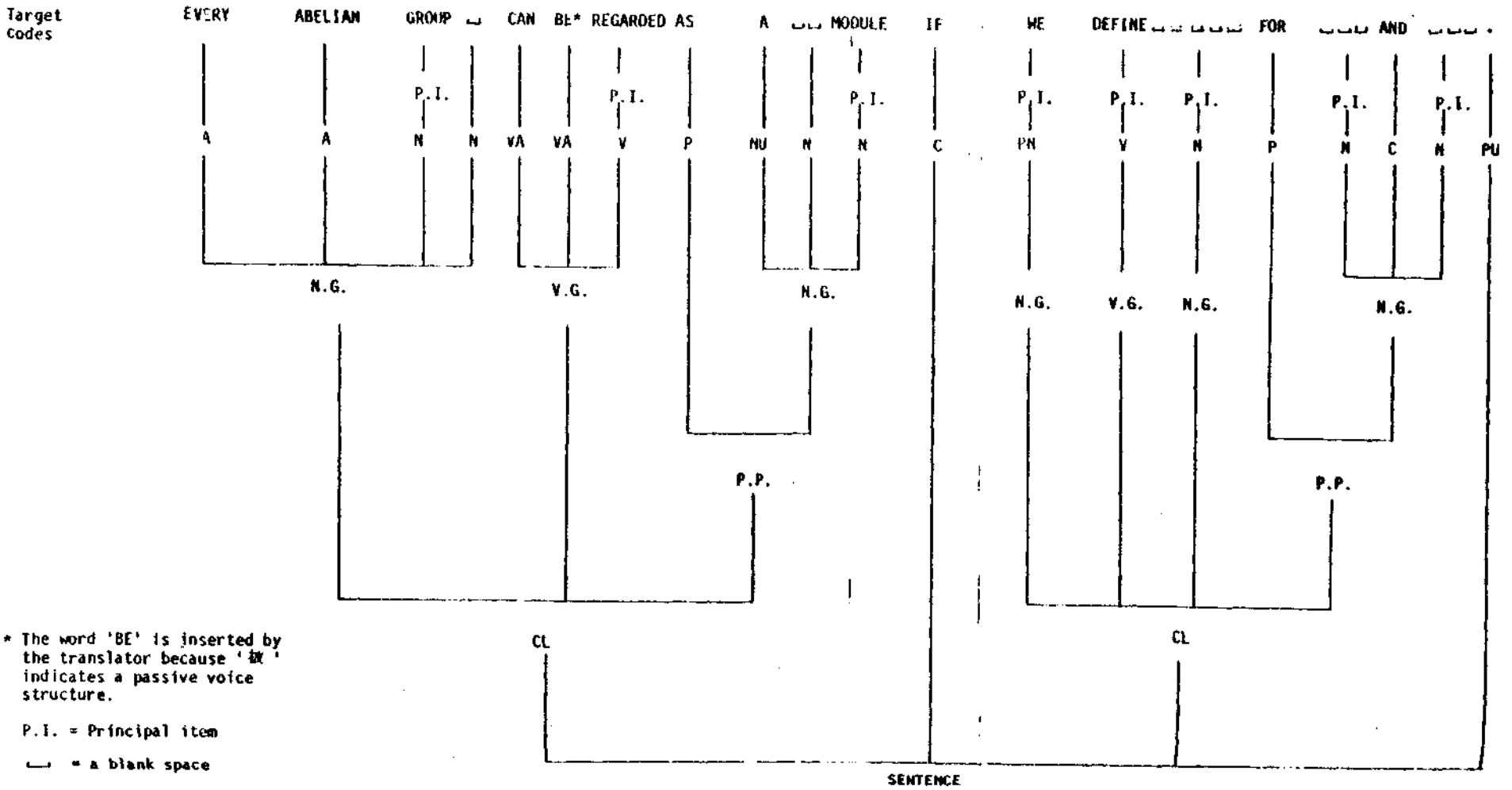
A = Adjective VA = Auxiliary verb P = Preposition C = Conjunction PN = Pronoun N.G. = Noun group P.P. = Prepositional phrase

N = Noun V = Verb NU = Numeral SI = Syntactic indicator PU = Punctuation V.G. = Verb group CL = Clause

Step 6. Target language synthesis (Second parse and third parse)

每 一 阿 貝 尔 群 G 可 较 雄 作 成 一 Z- 设 如 果 我 們 定 義 $na=a^n$ 對 于 aεG 和 nεZ .

3020 0001 7093 6296 1422 5028 B001 0668 5926 6018 0155 3634 0001 B002 2875 1172 2654 2053 0226 1353 5030 B005 1417 0060 B003 0735 B003 9978



Step 7. Output

EVERY ABELIAN GROUP _ CAN BE REGARDED AS A _ _ MODULE

IF WE DEFINE _ _ _ _ _ FOR _ _ _ AND _ _ _.

Step 8. Insertion of appropriate target language items missing in the source text.

EVERY ABELIAN GROUP G CAN BE REGARDED AS A Z- MODULE IF WE DEFINE $n\alpha = \alpha^n$ FOR $\alpha \varepsilon G$ AND $n \varepsilon Z$.