[Fifth International Symposium on Computers in Literary and Linguistic Research, 1978]

SHIU CHANG LOH
HING-SUM HUNG
LUAN KONG

A dual language translator

One of the major problems in designing a 'multi-language machine translation system' for translating *N* natural languages into *N*-1 languages is how to minimize the number of the programs required. This is important mainly for economic reasons, but any economic achievements will very often require a better understanding of the differences of the linguistic characteristics of languages.

In approaching the problem, some researchers suggested the use of an intermediary or interchange language (Delavenay 1960). In this case, the number of programs required is reduced to 2*N* from *N* (*N*-1). If the intermediary language is not an artificially created language, instead it is one of the N languages, then we should need 2*N*-2 programs instead of 2*N*. But the problem is, how can we design such an artificial language, or, which of the *N* languages is the best to be an intermediary language? There has been no universally accepted solution so far.

On the other hand, some researchers, who are in favour of transformational grammar, believe that the deep structure of all languages may be similar (Aitchison); the only differences may be in the choice of transformational rules. If this is true, then what we need is a set of transformational rules and a general translation program. To translate texts from source language *L*1 into target language *L*2, the translation program would only have to know how to select the appropriate transformational rules to transform the source sentence in *L*1 into the deep structure, and from the deep structure into the target sentence in *L*2. These suggestions greatly simplify our program. However, no fully satisfactory proposals for universal deep structure have been made.

Being aware of the fact that fully automatic high-quality translation (FAHQT) is not feasible, not even for scientific texts (Loh, Kong, and Hung 1978), it is not our aim to develop a perfect machine translation system at the present stage. Instead, we are always trying to design a new system which is better than the existing one (Loh 1976b). In approaching the problem of designing a multi-language machine translation system, we started our research study by first analysing the possibilities of designing a dual system for Chinese-English and English-Chinese translation based on the design philosophy and the characteristics of CULT (Chinese University Language Translator) — a language translator developed by the machine translation group at the Chinese University of Hong Kong (Loh 1976a, Loh and Kong 1977, Loh, Kong, and Hung 1978). This then led to the design of our present system — the Dual Language Translator (DLT).

DLT is an experimental natural language translator which is capable of translating from Chinese into readable English as well as from English into readable Chinese. Generally speaking, DLT is a direct product of a research study into the further exploitation of the potential capabilities of the CULT language translator.

In this paper, our aim is to introduce DLT briefly, and to show the possibilities of extending the system into a multi-language machine translation system.

DUAL LANGUAGE TRANSLATOR (DLT)

THE CULT-LANGUAGE MODEL

The design and implementation of DLT is mainly based on the CULT-language model which was developed in conjunction with the design of the CULT system in the past years. The CULT-language model is one suitable for translation purposes, and its design is based on the results of a series of simulation studies on the structure of Chinese as well as English sentences.

The language model regards a sentence as a string of correlative syntactic semantic items $S_i (i = 1, 2, \ldots. n)$. Each S_i has a unique function in the sentence, and according to their functions, they are grouped together to form groups. Each group is then assigned to a grammatical category such is noun group (NG), verb group (VG), or modifier (MD). These grammatical categories are called the basic components of the language model, and they are defined as follows:

NG (noun group) = noun, pronoun, or items which have a function similar to a noun.
VG (verb group) = the verb (or verbs) of a sentence.
MD (modifier) = the group of items which is used to modify NG or VG.
AU (auxiliary) = items not belonging to the above categories (e.g. conjunctions).

Any sentence accepted by the language model must at least have one noun group and one verb group; 'modifier' and 'auxiliary' are optional components. In general, a sentence may have more than one noun group and modifier, but may only have one verb group except for compound sentences. A representation of the language model showing the components and how they are related to each other is shown in Figure 1. The relationships are represented by the symbols G_i, R_i, T, U_i, and V_i. Each of these symbols indicates a different type of relationship:

1. The *classifying relations* G_i $(i = 1, 2, 3, 4)$ specify how the syntactic/semantic items are classified into the different components.

2. The *modifying relations* R_i $(i = 1,2,...)$ specify how the modifiers are used to modify the noun groups and the verb group. No one single modifier may be used to modify more than one component, but the relationship need not be one to one, i.e. more than one modifier may be used to modify

one component.

3. The *noun group-verb group relations T* specify how the noun groups are related to the verb group (e.g. subject and verb relation).

4. The *inclusive relations* I_i (i = 1, 2, . . .) specify that a modifier may include noun groups, another modifier, or another set of components (NG, VG, MD, AU).

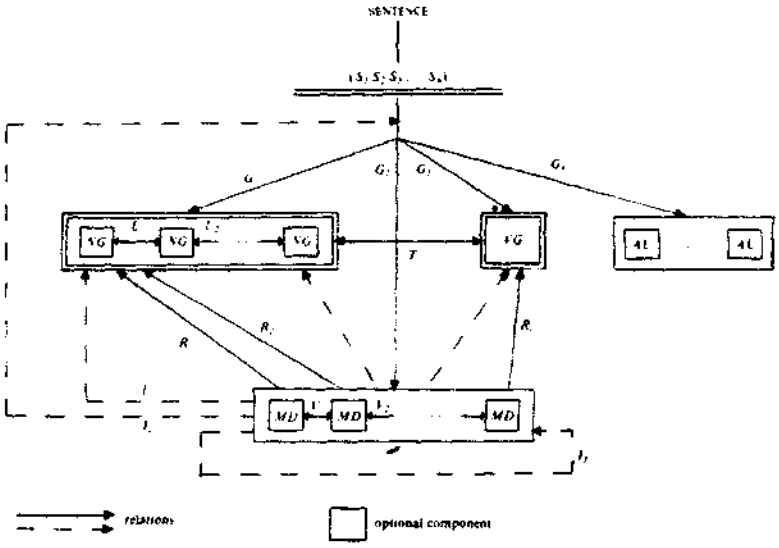

***Figure* 1. Components of the language model and their relations.**

5. The *internal relations* U_i (i= 1,2,...)and V_i (i = 1, 2, . . .) specify the interrelations of noun groups and the modifiers, respectively.

These relations are called the parameters of the language model. They are determined by the linguistic characteristics of the languages concerned (i.e. the source language and the target language). Because of the differences of the linguistic characteristics of languages, the parameters defined for a particular translation may not necessarily be suitable for other translations. Therefore, any translators designed based on the language model must have a facility to allow for the modification of the definitions of the parameters. DLT is an example of such a translator. The duality of DLT is achieved by redefining the parameters. A model of DLT is given below.

BASIC STRUCTURE OF DLT

Basically, DLT may be regarded as consisting of the following modules: CONTROL MODULE, TRANSLATION PROGRAMS MODULE, DICTIONARY MODULE, and PARAMETER DEFINITION MODULE (see Figure 2).

1. The CONTROL MODULE controls the overall information flow. It

instructs the TRANSLATION PROGRAMS MODULE what type of translation is to be carried out (i.e. Chinese-English or English-Chinese translation), and assigns the associated source and target dictionaries and the parameter definitions to the translation.

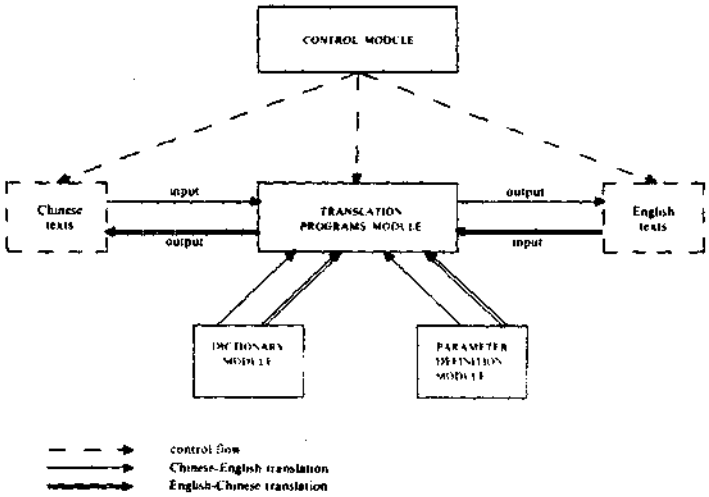

Figure 2. A model of the Dual Language Translator (DLT)

2. The TRANSLATION PROGRAMS MODULE is a collection of subprograms, each belonging to one of the following sub-modules: INPUT SUB-MODULE, ANALYSIS SUB-MODULE, OUTPUT SUB-MODULE, and SUPERVISOR SUB-MODULE (Figure 3).

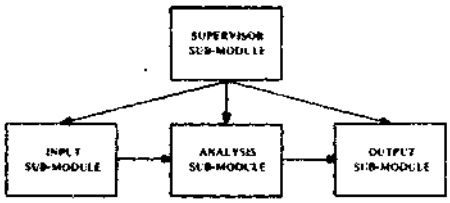

Figure 3. The TRANSLATION PROGRAMS MODULE.

The INPUT SUB-MODULE reads in the source sentence, converts the source codes if required, and performs lexical analysis by making use of the source dictionary. The resulting string of items is then input to the ANALYSIS SUB-MODULE. By consulting the dictionaries and the parameter definitions, the ANALYSIS SUB-MODULE analyses the items so as to determine the target codes required. Finally, the OUTPUT SUB-MODULE consults the target

dictionary, finds the required actual target codes, arranges them in a predetermined form, and outputs to the output device. The whole process of translation is under the supervision of the SUPERVISOR SUB-MODULE.

3. The DICTIONARY MODULE consists of three components (or sub-modules): DICTIONARY ADMINISTRATOR, CHINESE DICTIONARY, and ENGLISH DICTIONARY (Figure 4).

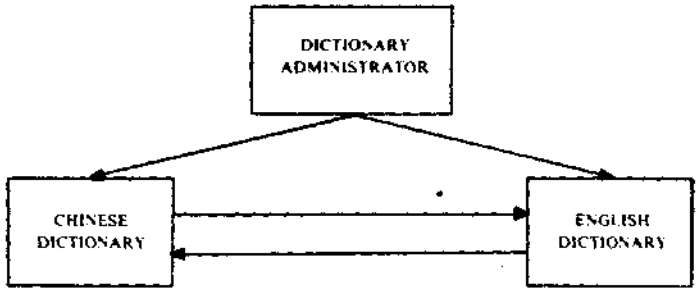

Figure 4. The DICTIONARY MODULE.

The DICTIONARY ADMINISTRATOR is responsible for the housekeeping of the two dictionaries (e.g. updating, insertion, deletion, and listing of the entries (or records)), and the determination of the functions of the two dictionaries for a translation (e.g. which dictionary is the source dictionary). The CHINESE DICTIONARY contains entries of Chinese words (items). Associated with each item is a set of grammatical assignments (syntactic and semantic information) of the item, and a set of pointers which point to the entries of the ENGLISH DICTIONARY where the translations of the item can be found. The ENGLISH DICTIONARY is constructed in a similar manner, but it contains the information for English words (items) instead of Chinese words (items).

4. The PARAMETER DEFINITION MODULE mainly consists of a set of definitions of the relations described in the above section. Like the DICTIONARY MODULE, this module also includes a sub-module called ADMINISTRATOR to take care of the housekeeping of the definitions and to assign the required definitions for a particular translation.

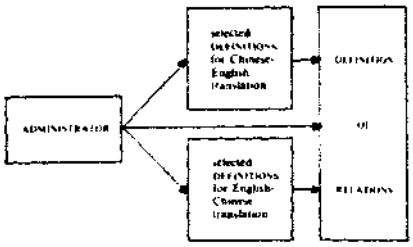

Figure 5. The PARAMETER DEFINITION MODULE.

The programs for the system are written in ANSI STANDARD FORTRAN and run on the ICL 1904A computer system. The system is used to translate English into Chinese for the purpose of evaluation.

TRANSLATION PROCEDURE

The steps involved in the translation using DLT are outlined in Figure 6.
These steps are illustrated for the case of English-Chinese translation.

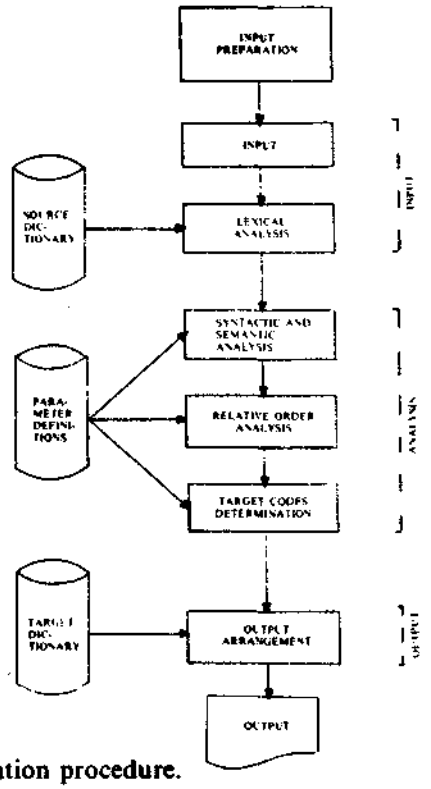

***Figure* 6. Translation procedure.**

The input preparation for English sentences is considerably simpler than Chinese sentences. For a computer system without a Chinese character input device, the Chinese characters would have to be coded (normally using telecodes) before submission to the computer for translation. In general, English words can be input to the computer directly. However, for simplicity in compiling the dictionaries at the present experimental stage, all English nouns and verbs are treated as having a unique form. That is, plural forms of nouns and the various forms of verbs are specially coded, e.g.

NOUNS ----------> NOUN+S
RETURNED ---------> RETURN+ED

Non-English characters and mathematical formulas are treated with special formats. After all the sentences are coded, the coded texts are then input to the computer for translation.

Once translation is initiated by the CONTROL. MODULE, the SUPERVISOR sub-module of the TRANSLATION PROGRAMS MODULE takes over the control of the translation process. SUPERVISOR instructs the INPUT, ANALYSIS, and OUTPUT sub-modules in turn to translate the coded texts one sentence at a time until all the sentences are translated. During the translation of a sentence, if an error occurs an error message will be printed together with some analytical information; the sentence is deleted and the translation of a new sentence is started. Otherwise, the translation will be carried out as follows:

1. After the sentence is read in by the INPUT sub-module, routine LEXANA is initiated to perform lexical analysis. LEXANA searches the ENGLISH DICTIONARY for the items of the input sentence by using the largest match principle (Loh, Kong, and Hung 1978). As a result, the grammatical assignments and the target code pointers of the lexical items formed are transferred to the core memory. By attaching to each lexical item its first grammatical assignment, a primitive string of syntactic, semantic items is then formed. This string is passed over to the ANALYSIS sub-module for syntactic and semantic analysis.

2. By making use of the CLASSIFYING RELATIONS G_i (i = 1, 2, 3, 4) and the INCLUSIVE RELATIONS I_i (i = 1, 2, 3), ANALYSIS makes the first step to analyse the relations of the items so as to determine their relative functions in the sentence, and classify them into different grammatical categories (i.e. the basic components of the language model). The MODIFYING RELATIONS R_i (i = 1, 2, ...) are then used to determine the relative order of the components (based on the structure of the Chinese sentence). The items of the sentence are then arranged in such a way that it is a pattern of a Chinese sentence. Finally. ANALYSIS determines the required Chinese equivalences of each item by using the relations R_i (i = 1, 2,...), U_i (i = 1,2, ...), V_i (i = 1,.2, ...), and T.

3. Once the target codes of the source sentence are determined, the OUTPUT sub-module is initiated to move the target codes to the output array and arrange them in a pre-determined form.

The output of Chinese characters is done by CalComp plotter, but may readily be accomplished by any matrix printer.

REMARKS

Further studies in other languages are needed in extending the capabilites of DLT into a multi-language machine translation system. However, the construction of DLT allows other translations to be incorporated into the system without difficulty.

REFERENCES

Aitchison, J. (1971) *General Linguistics.* The English University Press Ltd.

Bar-Hillel, Y. (1960) The present status in automatic translation of languages. *Advances in Computers 1.* New York: Academic Press.

Delavenay, E. (1960) *An Introduction to Machine Translation.* London: Thames & Hudson.

Loh, S.C. (1972) 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.*

Loh, S.C., Lam, M.N., & Chan, W.C. (1974) Machine-aided translation from Chinese to English. *United College Journal.*

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

Loh, S.C. (1976a) CULT (Chinese University Language Translator). FBIS Seminar on Machine Translation. 1976. *American Journal of Computational Linguistics,* microfiche 46. 46-51.

Loh. S.C. (1976b) Machine translation: past, present, and future. Presented at the Expert Group Meeting. UN Economic and Social Commission for Asia and The Pacific. Bangkok, December 1975. *ALLC Bulletin, 4,* 105-14.

Loh, S.C. & Kong. L. (1977) Computer translation of Chinese scientific journals. *Proceedings of the Third European Congress on Information Systems and Networks: Overcoming the Language Barrier.* Luxembourg.

Loh, S.C. Kong. L... & Hung H.S. (1978) *Machine Translation of Chinese Mathematics Articles.* Machine Translation Project, CUHK. Presented at a meeting jointly sponsored by the Specialist Group for Machine Translation of the British Computer Society and the Association for Literary and Linguistic Computing on 13 September 1977. *ALLC Bulletin.* 6, 111-18.

Example. English text (edited).

FUTURE

IN THE NEXT FIVE TO TEN YEAR + S, DEVICE + S FOR INPUT+ING AND OUTPUT+ING CHINESE AND OTHER NON-ALPHABETIC CHARACTER+S MIGHT BE AVAILABLE AT A REASONABLE COST. THIS IS, AT PRESENT, THE PRINCIPAL PROBLEM FACE + ING THE PROCESS+ING OF LANGUAGE + S OTHER THAN THOSE USE+1NG ALPHABET+S.

MORE LINGUISTIC RESEARCH APPLICABLE TO MACHINE TRANSLATION MAY YIELD USEFUL RESULT+S SO THAT MORE COMPREHENSIVE GRAMMATICAL RULE+S AND SENTENCE STRUCTURE+S BETWEEN LANGUAGE + S CAN BE FORMULATE + ED. *

AT THE CHINESE UNIVERSITY, WE ARE AT PRESENT WORK + ING ON A LANGUAGE TRANSLATOR WHICH WILL ENABLE US TO TRANSLATE CHINESE INTO ENGLISH AND AT THE SAME TIME TO TRANSLATE ENGLISH INTO CHINESE. A SIMULATION STUDY HAS BEEN DONE WHICH INDICATE + S THE POSSIBILITY OF SUCCESS. WE ARE HOPE + ING THAT. IN A FEW MONTH + S, A MODEL MAY BE DESIGN + ED TO DEMONSTRATE THE SYSTEM. IF SUCCESSFUL, WE ARE OF THE OPINION THAT A UNIVERSAL TRANSLATOR MAY BE DESIGN + ED SO THAT ONLY ONE TRANSLATOR WILL BE REQUIRE + ED TO DO THE TRANSLATION BETWEEN LANGUAGE + S.

ASK, NOT WHAT THE COMPUTER CAN DO FOR YOU, ASK YOURSELF WHAT YOU, WITH THE AID OF A COMPUTER, CAN DO FOR THE BETTERMENT OF MANKIND.

For Chinese text please see overleaf.