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

CREATION OF A SECOND-GENERATION SYSTEM FOR MACHINE TRANSLATION OF TECHNICAL MANUALS

John Chandioux

Consultant to the University of Montreal

Abstract

The extension of the Canadian government's policy of biculturalism is creating a demand for translation which exceeds the capacity of the State Department's Translation Bureau. It is for technical translation, particularly in aeronautics, that there is the greatest demand, and the natural, indeed necessary, solution is to have recourse to machine translation. On the one hand, the first-generation systems are not sufficiently reliable for the translation of texts on which human lives depend. On the other, there is no operational second-generation system in this field. The Translation Bureau has chosen to entrust the development of a second-generation system to the University of Montreal's TAUM project. This paper presents the various parts of the system and explains the choices that have been made.

1. Introduction

The Canadian government's policy of biculturalism is intended to allow federal employees to work with equal ease in either English or French. This results in a demand for translation which far surpasses the capacity of the market, especially as far as technical translation is concerned. The State Department's Translation Bureau, which is anxious to set matters right and feels satisfied with the results obtained by the University of Montreal's TAUM project in the field of hydraulics, has undertaken to develop a second-generation system for machine translation of technical manuals. The system is to be put into operation in November 1978, and the first series of technical documents translated will probably be that for the Aurora missile which has just been acquired by the Canadian armed forces.

The grammars and software which make up the translation system itself are being developed under contract by the TAUM project, while the dictionary is being drawn up at the Translation Bureau by a team of technical translators specializing in aeronautics, under the direction of André Petit. Installing the system will be the responsibility of the computer department of the State Department, which has already seconded Benoît Thouin, previously with GETA in Grenoble, to the TAUM project. It should be noted that this venture is in keeping with the policy of providing Canada within the next five years with a computer centre devoted to mechanized documentation, terminology and machine translation.

2. Description of the system

2.1 General

It is now generally accepted that the first-generation systems are inadequate for the translation of texts intended for publication. However, the ALPAC report led to the disappearance of a large number of the research centres, so that the second generation is still basically at the development stage, with two exceptions:

- the LOGOS III system (USA), which is often wrongly classed as a Georgetown derivative;
- the METEO system (Canada) which is used for the translation of weather forecasts intended for the general public

Moreover, there is no question, in the limited time available, of carrying out further research, rather than constructing a system based on the existing techniques developed by university research centres, such as those in Austin, Berkeley, Grenoble or Montreal.

While realizing the limitations of the second generation from the semantic point of view, the TAUM project thus stands by its achievements to date:

- complete syntactical analysis of English sentences;
- English-French transfer;
- independent generation of the source language;
- writing of grammars in high-level language.

Finally, the system is designed for technical translation, as regards both terminology and syntax and semantics. The development of a more generally applicable system would demand a much greater investment and there is as yet no proof that it is feasible.

Before agreeing on the final shape of the system, the researchers on the TAUM project studied each stage of processing in the light of the problems posed by technical texts. The Aviation system comprises six parts:

- morphological analysis of the English text;
- dictionary searching
- syntactic analysis;
- English-French transfer;

- syntactic generation;
- morphological generation

2.2 Morphological analysis of English

In the absence of a language comparable to ATEF (Grenoble) which would allow morphological analysis in interaction with dictionary searches, the first problem which presents itself is that of the place of morphology. This is why TAUM came to carry out the morphological analysis after a preliminary dictionary search so as to avoid breaking down words unnecessarily which is particularly costly with the use of Q systems. This solution is unacceptable in view of the fact that the dictionary search is the second bottleneck in the sequence of operations of a machine translation system after the collection of the data. Two solutions have been considered:

2.2.1 Morphological analysis before search

The drawback is the generation of superfluous breakdowns which call for a corresponding number of additional dictionary searches. In fact, if the number of searches is one per word in the case of a dictionary of full words, it is only 25% higher in the case of basic forms with a morphology restricted to removing the endings S, ED and ING. If, on the other hand, the ending ER were also to be removed, the ratio would be considerably less favourable, since comparatives represent only a minute percentage of cases.

2.2.2 Morphological generation during updating

This eliminates morphology from the translation sequence, while at the same time the translator is not obliged to enter into the dictionary all the forms of each word. On the other hand, on account of the extremely high number of verbs in technical language, the dictionary contains three times more entries than in the preceding case and many of the forms generated are rare, not to say out of use. Finally, the updating of such a dictionary requires a complex management system for the automatic retrieval of the word forms for which the entry is to be modified or erased.

The first solution proves to be the more economical and the morphology of English is to be programmed directly in PASCAL II, as has already been done for French.

2.3 Dictionary searching

First of all, it should be noted that no dictionary searching is carried out during transfer; this does not mean a rethinking of the basic principles of the second generation, but an economy which is quite logical for an exclusively bilingual system. The data necessary for analysis and for transfer remain distinct. During dictionary searching, the data needed for analysis are superposed onto the input text, while the transfer data are placed in a file to be consulted later. Three types of search are possible:

2.3.1 General memory search

This is rapid, but once the dictionary reaches a certain size it is necessary to use segmented loading or a division into microglossaries. The concept of microglossaries is much disputed: it is possible to give preference to a particular field, but it can never be said that a given text refers to one single field.

2.3.2 Tape search

Since sequence access is used, the words in the portion of text being processed must be presented in alphabetical order if good results are to be obtained, which comes down to carry-out one sort before the search and a reverse sort afterwards. This method, which was adopted for the first-generation systems at a time when discs did not exist, is very outdated.

2.3.3 Disc search

The speed of a disc search is closely linked to the number of disc access operations needed to locate an entry and the mechanical inertia of this type of memory. A dichotomizing search is virtually impossible in the case of a large dictionary since on average one extra disc access operation is needed per word each time the number of entries is doubled. Scattered or affinitive addressing must thus be used. At best, we can not expect to find a word in less than 50 ms, in other words a maximum of 70 000 words per hour.

The solution chosen is a hybrid one. The dictionary is to be placed on discs, but the functioned words and the words most frequently used in the preceding few minutes of processing will be stored in the central memory. This results in a reduction of 50 to 75% in the number of disc access operations needed. Moreover, the technical terms will be broken down using a code based on the ATA 100 breakdown used in aeronautics.

2.4 Syntactical analysis

In technical language the noun-verb ambiguity is more or less systematic and clusters of from 2 to 6 nouns or more are very common. This gives rise to two distinct problems:

2.4.1 Restricting the number of possible combinations

The longer the phrases and the more frequent the cases of syntactical homography, the more important it is to restrict the number of possible combinations. To give an idea of the order of magnitude, the number of possible analyses for a given phrase is equal to the product of the ambiguities. Although technical phrases are generally clear and free from stylistic effects, they are nonetheless not short, witness the phrases of more than 50 words in TAUM's working corpus. It is thus necessary to find a way of restricting the number of possible combinations in order to avoid saturating the memory, as can happen fairly easily when Q-systems are used. The analyser will be written in REZO, an adaptation of Woods' transition networks written by Gilles Stewart. This programme, written in PASCAL II and much easier to master, enables deterministic grammars to be written, while non-determinism can be introduced at any level at the linguist's discretion. The writing of partially deterministic grammars should be seen not as a way of restricting the number of analyses, but as a way of eliminating abortive analyses at a very early stage and preventing the repetition of identical analyses of the same syntagm. This latter problem was never satisfactorily solved by Woods and his collaborators.

2.4.2 Analysis of clusters

The problem of clusters is basically a semantic one and can only be solved by means of highly detailed subcategorization of nouns and adjectives, together with rules of compatibility and non-compatibility in order to identify what qualifies what. If suitable solutions are to be found, the linguist must firstly work in close collaboration with the translator, and secondly these rules must be specific to technical language. The only other alternative is to regard these clusters as idioms and enter them en masse in the dictionary, which inevitably interferes with the smooth operation of the analyser.

2.5 Transfer

The transfer operation is in the form of an interpret programme for the translation rules coded for each dictionary entry which applies these rules to the tree structure of the phrase being processed. For example, the verb 'to replace' will be translated by 'remplacer' if the object of the verb is a noun group of which the top term belongs to the class of consumables (filter, joint, etc) or is a pronoun of which the antecedent has been identified as belonging to this class. Otherwise, it will be translated by 'remettre en place', unless the object noun group contains an adjective implying replacement (cracked, damaged etc.), and so on. The translation of purely technical words is based to a large extent on the breakdown code.

None of the existing tools in the TAUM project seemed really adequate for this task; on the other hand, the transfer experiments carried out in connection with the experimental TAUM *76* system and presented at that year's COLING give a good idea of the type of tool needed. As for the other parts of the system, this will be programmed in PASCAL 11.

2.6 Syntactic generation

In view of the nature of the transfer tool, it seems advisable to put off until the generation phase certain tree structure manipulations for which there are excellent languages, such as Q-systems or CETA (Grenoble). Since syntactic generation is a thoroughly deterministic process, the use of a combinational language poses no problems provided it is effective. The FORTRAN version of Q-systems has proved its worth in the METEO system and seems to be the best choice.

2.7 Morphological generation

The TAUM experimental system has included an exhaustive morphology of French since the 75 version, and this will be incorporated as it stands into the Aviation system. The only difference is that the generation of rectional forms (e.g. formation of nouns) is to be abandoned and replaced by use of the dictionary, since theoretical research carried out in this field is not sufficiently promising to be taken up in the framework of a development contract.

3. Operation

Since the team responsible for putting the system into operation has not yet been created, this can only be discussed hypothetically.

The documents to be translated will be divided into processing batches corresponding both to natural divisions in the text and to the optimum load for the machine. In the operating sequence several units will be processed simultaneously, but at different phases. The probable phases are as follows:

- Data acquisition. In general, there is no problem here since the manufacturers are in a position to provide the technical documents on magnetic tape;
- Correction of errors and updating the dictionary for each new processing batch, using an interactive programme;
- Actual translation;
- Revision of the machine output by revisers trained for machine translation;
- Correction of the translated text via a terminal;
- Input of tables, parts lists or portions of text translated manually;
- Production of the final tape with microcodes for photocomposition and, in certain cases merging with the original tape for the production of bilingual texts;
- Supply of negatives for the illustrations, translated if necessary.

Initially, the system will be installed on a computer similar to that of the University of Montreal (CDC 173), since the simultaneous transfer and updating of such a complex system would present problems. However, the transfer of the system to another computer would be greatly facilitated by choosing the same programming language for writing most of the software.