ENGLISH-FRENCH TRANSLATION ON A COMPUTER

K. H. V. BOOTH
C. A. BROWN
C. STOCK
University of Saskatchewan
Saskatoon, Saskatchewan
Canada

For some years a group has been working on the production of a computer program to translate from English to French. This is of particular interest in Canada because of the large volume of official literature which must, by law, be published in both languages.

In order to achieve results in a finite time, two limitations were imposed; firstly, only the translation of "government" type documents is considered as opposed to more literary compositions or transcriptions of the spoken word. Secondly, it is not claimed that translations will be perfect, but rather a good approximation. The implication of this is that a post-editor will be required to tidy up the machine output, select the most apposite translation of a word when more than one is available and, hopefully, correct cases where the program has gone astray and misinterpreted a sentence.

In line with this philosophy a "statistical" approach has been adopted in many problems such as, for example, determining the grammatical role of words in a context, or deciding the subject of verbs.

With these restrictions in mind, the program can be outlined under the following main sections:

1. Input, dictionary look-up and idiom identification.
2. Identification of the most probable grammatical code of each word using a statistical method.
3. Translation program and output.

The input program is composed of several routines which read text from computer-readable medium, analyze the string of characters, identify the separate constituents of text and store these together with determined information. The Reader performs the actual character input and provides facilities for examining either prior or post context of the current-input character. The Analyzer examines

the string, determines the bounds of words and sentences and resolves the ambiguous use of certain characters (e.g., the period or dot may be used as a full-stop, or to denote an abbreviation or as a decimal point of a number, etc.). In many cases, it is necessary for the Analyzer to consult the dictionary in order to determine the extent of a word, while in other cases, the Analyzer requires that the text rigorously obey such orthographic conventions as placing two or more spaces after a full-stop. The Store routine merely arranges the constituents of text and related information into the high-speed memory of the computer.

Because of the immediate nature of the information required for the machine translation process, the dictionary must be provided on-line. Direct access devices such as disks or drums provide efficient storage and search capabilities together with high-speed data transmission. In order to facilitate an efficient search method, we have physically separated the relatively fixed-length English word (search key) from the highly variable length translations: the English words are blocked, in alphabetic order, on full tracks of the direct access device with the lengths and track addresses of the corresponding translations, which are packed on separate tracks. In addition, a table containing the last word of each " English " track with pointers to the corresponding track is retained in memory. A binary search of this table isolates the appropriate track for a given word; a subsequent binary search of that track yields the location and length of the desired translation. This dictionary arrangement appears to be most suitable for an experimental translation process because of its simplicity and relative efficiency.

Certain phrases, idioms and proper names, present difficulty for machine translation because they do not obey regular grammatical rules. In order to identify proper names, an isolation routine and a verification routine employs a Bayesian statistic on isolated groups of words.[1] This routine has been approximately 95 percent successful on a large sample of text. Following this routine is the idiomatic phrase resolution; flags on selected keywords initiate a search on a prepared binary list table.[1] This method has advantages over others in that it allows an unlimited number of idioms to be associated with any keyword, is capable of resolving discontinuous idioms as well as the continuous form and is extremely efficient in terms of speed. In addition, the idiom resolution routine provides the facility for parsing the resolved idioms.

The second phase, consisting of the identification of grammatical codes has been described in a previous paper.[2] Suffice it to say

[1] C. Stock, M. Sc. thesis, University of Saskatchewan, Saskatoon, Saskatchewan, 1970.

[2] K. H. V. Booth, "Statistical Parsing by Computer", *Pensiero e linguaggio in operazioni*, I. 1. 1970.

that at the end of this phase the translation program is presented with a text in which each word has been allotted one of 16 grammatical codes (noun, preposition, etc.) and the appropriate translation (or translations) into French supplied. The error rate in this process is about 4 percent.

The translation program of the third stage then takes this information and by correcting word order, inflecting verbs and adjectives, etc. and so on produces the final French translation. This part of the program is not yet complete, and indeed probably never will be if the criterion for completion is that no improvement is possible. However, the output is now such that a preliminary evaluation of the program is possible, and this will be described later in the paper.

A detailed description of this phase would be too timeconsuming, but an idea of the mechanism can be obtained from the following break-down into subsections: At each stage the sentence is taken as the semantic unit.

1) The sentence is scanned for verb groups, which are tagged, and rearrangement of verb-adverb order is made where necessary.

2) A scan is made for noun groups which are tagged according to gender and number. Plural nouns are inflected during this scan.

3) Noun groups are " processed " by rearranging noun-adjective order where necessary choosing the correct form of determiners (le/la/les, etc.) and inserting them where they are omitted in English and replacing preposition-determiner pairs by the correct form (de le=du, etc.). Adjectives are also inflected at this stage, using the information on noun groups obtained in the preceeding scan.

4) Verbs are " processed ". This involves a preliminary scan to determine the subject which is then tagged. Compound tenses are associated (Exports *have grown...*) and negatives rearranged (it is not=il n'est pas) and verbs inflected according to their tense and subject. At this stage, too, relative pronouns standing as subjects are given their correct form.

5) Finally the translation is output together with any alternative translations of individual words.

Several comments can be made on this phase of the program. Firstly, the arrangement of the sections is to some extent arbitrary and was dictated by the fact that the program " just grew " that

way. Thus it might be more logical to place the first section immediately before section 4 and deal with verbs all the same time. This would pose no difficulties.

Secondly, no semantic analysis of the text has been attempted and this leads to difficulties in places. In the demarcation of noun groups and rearrangement of noun-adjective order for example, it is occasionally impossible to decide on the correct structure without recourse to meaning. Consider the sentences:

> Strata of different kinds and belts of intrusive rocks form northeasterly trending banks.
>
> The surface consists of hills, ridges and valleys containing innumerable lakes and streams.

In the first sentence " different " applies only to " kinds " but not to " belts " while in the second " innumerable " applies to both " lakes and streams "; without some form of semantic analysis it is impossible to distinguish these cases. Inspection of a large section of text, taken from the Canada Year Book, 1962, led to the conclusion that fewer mistakes would be made by always assuming that an adjective is attached only to the immediately following noun, and this is what the program does.

No attempt is made to decide between different translations for the same word. Thus " time " used as a noun can have at least 6 different translations depending on the context and the program merely prints all of them out and leaves it to the post-editor to choose the most appropriate. To attempt a choice would involve a program and dictionary at least an order of magnitude more complicated than those at present.

Throughout this work a sample from the Canada Year Book has been taken as a basis for the dictionary, idiom list and for the type of text we are attempting to translate. This has led to certain welcome simplifications in the problems involved. Thus the text is written entirely in the third person (in common with most official and scientific type documents) and this makes the task of verb inflection easier. Also, as such texts tend to make little use of pronouns, they have not been dealt with extensively in the program.

The question of the passive tense always causes much agitation in translating from English into French, the point being that in some cases a passive construction in English must be turned into an active one in French:

> He was given a book by his father.
> Son père lui a donné un livre.

Here again the type of text with which we are dealing largely eliminates this problem as very few instances of this construction appear.

A more serious problem is that of distinguishing the incomplete form of the passive. Thus in the sentence:

> Methods developed in Canada have been applied to Precambrian Shields of other continents.

" developed " is actually a shortened form of the passive " which have been developed ", but without a more extensive analysis of the syntax of the sentence it cannot be distinguished from the simple past tense.

Work is continuing on this problem, and it is hoped that a more refined classification for verbs will clear up many doubtful cases. A discussion of some of the problems involved is given in reference.[3]

Finally, some practical remarks on the program may be of interest. The translation section is written in COBOL, largely because this was the only high level language which was at all suitable and available when we began the project. Actually, apart from the annoying restrictions on subscripts, it has been surprisingly convenient in use especially considering its design for very different applications.

The program in its present form contains about 1500 statements, and the total translation time for 1800 words, the largest segment of text yet processed, is about 5 minutes. To give an idea of the quality of the translation, a few sentences are shown in Figure 1.

The object of this work is to provide assistance to the overworked Government translation services in Ottawa, and we are therefore interested in assessing how far an output of the quality shown in Figure 1 could be understood by someone having no access to the original document, and indeed not necessarily knowing English.

Early this year (1970), we felt we had reached a point in our work at which it would be worth while having some of our French output corrected by well-educated native speakers with little or no knowledge of English. Since our corpus, 25,000 words from the *Canada Year Book 1962*, is being used at a test sample, we chose at random two short passages from the section on Geology for our experiment. There was a total of forty sentences, a " sentence " ending with either a period or a semi-colon.

[3] C. Brown, "Some of the problems connected with passive constructions in English-French computer translation", *Pensiero e linguaggio in operazioni*, II, 5, 1971.

*NORTH *AMERICA COMPRISES SIX MAIN NATURAL REGIONS WHICH ARE BOTH PHYSIOGRAPHIC AND GEOLOGICAL BECAUSE THE AGES, KINDS AND STRUCTURES OF THE UNDERLYING ROCKS DETERMINE THE NATURES OF THE LAND SURFACES. *KNOWLEDGE OF THESE REGIONS IS IMPORTANT BECAUSE THEIR GEOLOGICAL CHARACTERISTICS HAVE MUCH INFLUENCE ON THE SUITABILITY OF DIFFERENT AREAS FOR SUCH ACTIVITIES AS AGRICULTURE, MINING, PETROLEUM PRODUCTION AND SPORTS, AND CONTRIBUTE AS WELL TO THE VARIED SCENERY OF THE COUNTRY. *THE SIX REGIONS ARE: THE *CANADIAN *SHIELD, A VAST AREA OF ANCIENT ROCKS THAT IS MAINLY IN *CANADA; THE *INTERIOR *PLAINS AND *LOWLANDS, THE LARGEST AREA OF WHICH EXTENDS THROUGHOUT THE MID-*CONTINENT FROM THE *GULF OF *MEXICO TO THE *ARCTIC *OCEAN; THE *APPALACHIAN *REGION, MAINLY IN THE *UNITED *STATES BUT ALSO FORMING AN IMPORTANT PART OF *EASTERN *CANADA; THE *CORDILLERAN *REGION, EXTENDING ALONG THE ENTIRE WEST COAST OF THE *CONTINENT; THE *ATLANTIC *COASTAL *PLAIN ALONG THE EASTERN SEABOARD OF THE *UNITED *STATES; AND THE *INNUITIAN *REGION, A MOUNTAINOUS BELT IN THE *ARCTIC *ARCHIPELAGO. *CANADA INCLUDES PARTS OF FOUR OF THESE REGIONS AND ALL OF THE *INNUITIAN *REGION, BUT NONE OF THE ATLANTIC *COASTAL *PLAIN.

L'*AMERIQUE DU NORD COMPREND SIX R1EGIONS PRINCIPALES NATURELLES/PROPRE/ QUI SONT 2A LA FOIS/ AUSSI BIEN QUE/ PHYSIOGRAPHIQUES ET GEOLOGIQUES PARCE QUE LES 3AGES, LES SORTES ET LES STRUCTURES DES ROCHES SOUS JACENTES/FONDAMENTAL/ D1ELIMITENT/ D1ECIDER/ LES TERRAINS/NATURE/ DES TERRES DE SURFACE. DE LA/LA/ CONNAISSANCE DE CES R1EGIONS EST IMPORTANTE PARCE QUE LEURS CARACT1ERISTIQUES G1EOLOGIQUES INFLUENT BEAUCOUP SUR L'ADAPTATION DE ZONES/SUPERFICIE/ DIVERSES POUR DES ACTIVITIES TELLE QUE DE L'/L'/ EXPLOITATION MINI2ERE, DE LA/LA/ PRODUCTION P1ETROLI2ERE/DU P1ETROLE/ ET DES/LES SPORTS, ET FOURNISSENT/CONTRIBUER/ EN PLUS AU PAYSAGE VARI1E DU PAYS/R1EGION/. LES SIX R1EGIONS SONT: LE *BOUCLIER *CANADIEN, UNE ZONE/SUPERFICIE/ VASTE DE ROCHES ANCIENNES QUI EST PRINCIPALEMENT/ DANS L'ENSEMBLE/AU *CANADA: LES PLAINES INT1ERIEURES ET LES BASSES TERRES, LA PLUS GRANDE ZONE/SUPERFICIE/ DE QUI S'1ETEND/PROLONGER/ 2A TRAVERS LE CENTRE DU CONTINENT DU/DEPUIS/GOLFE DU *MEXIQUE 2A L'OC1EAN ARCTIQUE: LA R1EGION *APPALACHIENNE, PRINCIPALEMENT/DANS L'ENSEMBLE/AUX *1ETATS *UNIS MAIS FORMANT/ ASSURER/ AUSSI UNE PARTIE/LET1ENDUE/ IMPORTANTE D'*EST CANADIEN: LA R1EGION DE LA *CORDILL2ERE, S'1ETENDANT/PROLONGER/ LE LONG DE TOUTE LA C3OTE OUEST DU CONTINENT: LA *PLAINE C3OTI2ERE DE L'*ATLANTIQUE LE LONG DU LITTORAL EST/ORIENTAL/ DES *1ETATS *UNIS: ET LA R1EGION *INNUITIENNE, UNE ZONE/CHA3INE/ MONTAGNEUSE DANS/AU COURS DE/EN/ L'ARCHIPEL ARCTIQUE. LE *CANADA COMPREND DES/LES/ PARTIES/1ETENDUES/ DE QUATRE DE CES R1EGIONS ET TOUTE LA R1EGION *INNUITIENNE, MAIS NE-AUCUN DE LA *PLAINE C3OTI2ERE DE L'ATLANTIQUE.

Note: Numerals represent accents as follows:
1 acute
2 grave
3 circumflex
Slashes / / indicate alternative translations
* indicates capital letters.

Fig. 1

We assumed that nearly all well-educated Quebecers would have some knowledge of English so decided to use people from France as our first " post-editors ". Twenty were given a copy of the two passages and a sheet of instructions; nineteen have returned the corrected version. Nine or ten of these had never had any instruction in English and none of the others had, to the best of our knowledge, studied it beyond first year university level. All of our post-editors had completed high school while a number were university graduates.

The thirty-nine sentences which we can consider (the fortieth was incomplete) varied greatly in length and in difficulty from the point of view of the post-editors, and a quantitative assessment of their performance is therefore rather difficult, and possibly misleading. To give an idea of the general standard reached, over half of the editors interpreted at least 31 out of the 39 sentences correctly and all interpreted at least 24 correctly. Four particular sentences were responsible for over half of all the errors made.

In a paper yet to be completed we intend to look at the work of the post-editors from a different standpoint. A good translation will render faithfully the meaning of the original. What was there then in our output which caused some post-editors to make choices between alternatives and changes which caused the corrected output to do other than convey the meaning of the English text? In some cases phrases or sentences had little or no meaning in French and the post-editors were then expected to write in something which they thought would fit. Very often in these cases the French did not translate the idea expressed originally in English.

In the paper mentioned above errors will be categorized and their percentage calculated when that is considered appropriate. In places where *a priori* we considered errors of interpretation to be " highly probable " we will give the percentage of post-editors who actually went wrong. We will do this also in cases where we considered errors to be " possible ". The " unanticipated " errors will simply be left as figures.

In this present paper we can, however, discuss and give examples of a few of the difficulties faced by the post-editors of our output. It is interesting to see what they made of sentences where French syntax was not respected or where a " reference " was not clear. So far, we have made no attempt to deal with semantics, with the result that a post-editor may have to choose one of several nouns, verbs, adjectives, prepositions, etc. to suit the context. Our system of statistical parsing sometimes goes wrong and this nearly always causes misinterpretation of the sentence.

In the field of syntax, one problem which remains to be tackled is that of the position of relatives in subordinate clauses.

In English it is more flexible for some relatives: " the Interior Plains and Lowlands, the largest area *of which* extends throughout the mid-Continent ", in French " dont " must immediately follow its antecedent. When our programme translated "of which" as "de (ce)qui " and retained English word order only two post-editors out of the nineteen made the necessary changes. When an incorrect translation of a relative is not combined with the disruption of syntax results are much better: the translation of " igneous processes *of which* good evidences remain " was correctly interpreted by seventeen of the post-editors.

There was general misinterpretation of a long sentence beginning with " il/elle " and with a number of past participles and adjectives given in their four forms because the references had not been determined.

The translation of the preposition " from " as the set of alternatives *de/depuis/d'après* caused five post-editors out of nineteen to change the meaning of the following English sentence in their correction of the French output: " The Innuitian Region is known mainly *from* reconnaissance surveys. " They left " depuis " and crossed out " de " and " d'après ", the latter, of course, rendering the meaning intended.

More than half of the post-editors did not understand our translation of the following when certain words were incorrectly parsed as marked: " The Shield continues into the United States $\frac{\text{west}}{\text{adv.}}$ and $\frac{\text{south}}{\text{noun}}$ of Lake Superior and $\frac{\text{east}}{?}$ of the upper St. Lawrence River, etc. "

It is unlikely that errors of this type can be eliminated entirely and in these cases we shall have to fall back on the post-editor to provide corrections. In our tests the post-editors had no access at all to the original English text, and in any case, many of them were not familiar with the language. The textual matter itself was highly technical and none of the editors were specialists in the field. In practice it seems likely that an official translator would be fluent in English and therefore able to refer to the text to clear up puzzling cases.

The authors wish to thank the National Research Council of Canada for generous support in this project.