RICHARD I. KITTREDGE
Université de Montréal

The Development of Automated Translation Systems in Canada

1. A historical sketch of Canadian research and development in automated translation

Canada is a relative new-comer to the field of computer-assisted translation, having set up its first research projects in 1965, some fifteen years after the first efforts began in the United States, Britain, and the Soviet Union. If, in 1980, it is now at the forefront of research and development in this field, this is due to the special priority given to bilingualism in Canada, plus two other important factors:

(1) During the first years of research, the helpful collaboration of linguists and computer scientists from abroad (particularly France and the United States) made it possible to profit from earlier mistakes and begin with the best existing theories and practical knowledge.

(2) The universities and governmental sponsors gave the university research team the long-term support they needed to reach the stage where practical results were possible. This support was particularly critical during the "lean and lonely" years of 1971-75 when only Canada and France, among Western countries, had substantial research projects in machine translation.

By 1973 the TAUM* research group at the Université de Montréal had demonstrated an experimental English-to-French system which could handle most of the complexities of English input sentences and make the proper translation of individual words and structures to produce acceptable French sentences. One major difficulty, however, was that the number and complexity of the syntactic rules required for doing this was so great that no machine could load the entire program into working memory. And these grammars were still not complete!

In the face of such complexity it was decided to make the first practical application of the research results within a specialized type of language, where the size of the vocabulary, possible word meanings, and grammatical rules are much more limited than in the language as a whole. In 1976, after two years of feasibility studies and design, the TAUM group delivered to the Canadian government an operational system for English-to-French translation of weather bulletins. This METEO system, described in detail in section 2, became the world's first system capable of taking large quantities of unedited text and producing translation which normally did not require revision. (In the small percentage of cases where a sentence cannot be analyzed by the program, the system automatically submits the sentence to a human translator.)

The operational success of the METEO system over the past three years has shown that useful machine-assisted translation is already achievable within small restricted *sublanguages.* Since 1976 the TAUM researchers have been working on the much larger and more difficult sublanguage of aircraft maintenance manuals. Their AVIATION system, outlined in section 3, is in an advanced stage of development, having already demonstrated the ability to provide revisable technical translation for a substantial portion (somewhat over half) of a typical manual.

Although many theoretical and practical problems remain to be solved before any kind of technical translation can be completely automated, it is clear that computer-assisted translation is rapidly becoming economically feasible in sublanguages far more complex than weather bulletins. Since the quality of machine output is fast approaching that of firstdraft human translation *for the sentences which the machine can handle,* the critical problem lies in reducing the percentage of sentences which fail altogether to be analyzed because of their length or structural and semantic complexity.

2. The METEO system

2.1. Overview

Before 1976 the translation of English weather bulletins into French at Montreal's weather translation service proceeded as follows: A specialized meteorological translator would monitor a teletype on which a large variety of weather data was being received. He would separate an incoming English weather bulletin (originating in Toronto, for example) from other data and retire to his work table where he proceeded to type out his translation. On finishing, he would hand the French text to a teletype operator, who dispatched it to Toronto.

The translator was under pressure to send each translation back within 30 minutes after receiving the original. Since most sentences repeated stereotyped patterns, he quickly became bored with such repetitious material, leading to a high error rate and general job dissatisfaction.

The current METEO system has greatly improved the efficiency and comfort (as well as the interest) of the translator's work, carrying out most of the previous activity automatically and requiring his intervention only as a consultant for the most difficult problems. A computer program now extracts weather bulletins from other incoming data by recognizing a coded prefix which indicates the type of text and city of origin. The bulletin is forwarded to the translation program for processing. Within the translation program itself (see 2.2. below) each text word is looked up in a dictionary and replaced by its dictionary representation(s). Grammatical rules then apply to the dictionary representations, calculating the probable role of each

* TAUM is an acronym for .Traduction Automatique - Université de Montréal.

word in the sentence from among the possible ones. When this succeeds (over 80% of the time), a so-called "sentence tree" is constructed. The sentence tree contains enough information in the proper order so that translation rules can be applied, giving the correct lexical choice and word order for French. In cases where the translation program cannot build a sentence tree, this means that the sentence structure, or at least one word in the sentence, is not recognized by the available rules of dictionary or grammar. Most often, this is due to misspelling, improper punctuation or improper formatting in the original bulletin. In such cases the supervisor program detects the failure to build a sentence tree and sends the original English sentence to a human translator via his cathode ray terminal. When the human translation has been keyed in for each unanalyzed English sentence, the completed French text is assembled in the proper order. The city code is then used to expedite the translation automatically to the city which provided the English original.

2.2. Linguistic principles of the translation program

A brief look at the sample weather bulletin in (1) below reveals that "sentences" in this sublanguage, while relatively simple, are not formed according to the same rules that govern "normal" English sentences.

(1) WWY273
FPCN11 CYXY 111230
FORECASTS FOR YUKON AND NORTHWESTERN BC ISSUED BY ENVIRONMENT CANADA AT 5.30 AM PDT FRIDAY JULY 11 1980 FOR TODAY AND SATURDAY.

KLONDIKE
BEAVER CREEK
STEWART RIVER.
RAIN OCCASIONALLY MIXED WITH SLEET TODAY CHANGING TO SNOW THIS EVENING. HIGHS 2 TO 4. WINDS INCREASING TO STRONG NORTHWESTERLY THIS AFTERNOON. CLOUDY WITH A FEW SHOWERS SATURDAY. HIGHS NEAR 6.

In particular, these telegraphic sentences lack a tensed verb. Since the verb is normally the pivotal point for both syntactic and semantic analysis, it is clear that the computerized grammar for analyzing weather bulletins must be based on rather special rules, which are in fact unique to this sublanguage. Virtually all sentences fall into one of five basic types depending on subject matter: atmospheric conditions, wind speeds, temperature minima and maxima, etc. For each sentence type there is an identifiable set of possible head words, around which the major constituent is built. For example, in the most important and potentially complex type, concerning atmospheric conditions, one of the words *clear, sunny, rain, snow, cloudiness,* etc., marked *AC ("atmospheric condition") in its dictionary entry, must be present in order for the structure-building rules to apply:

(2) INCREASING *CLOUDINESS* THIS AFTERNOON.
RAIN TODAY ENDING LATE THIS AFTERNOON.
PARTLY *CLOUDY* WITH SOME CHANCE OF THUNDERSHOWERS TODAY.
FOG ALONG THE COAST THIS EVENING.

The structure of the sentences in (2) follows a general pattern which can be stated in terms of the following four constituents (in typical order):

(3) (atmospheric condition) (modification of condition) (place) (time)

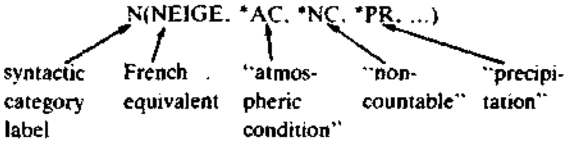
Only the first of these constituents is obligatory. Within each constituent, the possible combinations of words are described using rules which refer to traditional word classes (noun, verb, adjective, etc.) and to semantic subclasses (precipitation, time interval, geographic feature, etc.).

The word class and subclass information is provided when each word of an input bulletin is subjected to dictionary lookup. For example,

(4) SNOW

is replaced by its dictionary representation:

(5)

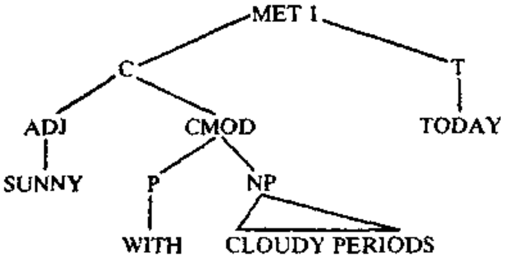


The syntactic rules which "search" for allowable combinations of words refer to their category label and to one or more of their semantic subclass labels. Upon finding a legal combination, these rules build a structure tree, first for each constituent, and then, using the individual constituent trees, for the whole sentence. Under this procedure, a sentence such as:

(6) SUNNY WITH CLOUDY PERIODS TODAY.

becomes:

(7)



except that in actual practice the French words have already been substituted at this point, along with their grammatical features.

A tree structure such as the one in (7) contains all the information necessary for producing a French "sentence" (also without verb) of meteorology. A set of generation rules applies to the tree structure to give the proper linear sequence of words in French (which may differ from the left-to-right order of words in the tree structure). These French words must still undergo inflection rules and agreement rules to produce the appropriate French sentence in readable form.

2.3. Everyday operation of the METEO system

On a typical day, the METEO translation system processes a load of some 30.000 words of weather bulletins. The great majority of the input sentences, as seen above, are translated and given only "spot checking" from time to time by a human translator.

One of the great advantages of the system is the reliability of its translated output. This is the result of a very conservative approach to the writing of linguistic rules in the analysis program. For a sentence to pass the analysis stage (and be assigned a tree structure), all of its words must be found in the dictionary and every grammatical structure must be recognizable. No metaphorical extensions are allowed and the program does not try to "guess" about what it doesn't "understand". Under these conditions the rules of translation can be made to function quite well.

When a sentence fails to be analyzed, this can be for one of several reasons:

(1) a new or infrequent word is not found in the dictionary,
(2) the sentence is improperly punctuated or the spacing or formatting of the whole text does not obey norms.
(3) a grammatical construction is used which is not allowed (recognizable) by the syntactic rules of analysis.

The majority of untranslated sentences fail analysis for the second reason. Unfound words also account for a substantial number of parse failures. The rarest, but most interesting type of failure is due to unusual syntactic constructions.

It occasionally happens that the author of the weather bulletin is faced with unusual or dangerous conditions and must abandon the normal telegraphic style so characteristic of these bulletins. The following sentence recently appeared in a report from Winnipeg:

(6) "Persons in or near this area should be on the lookout for these severe weather conditions and watch for updated warnings"

Note that the presence of a tensed verb sets this sentence apart from all those recognized by the METEO grammar. It would be possible in principle to enlarge the dictionary and grammar to cover many such sentences. But since these "emergency" sentences are much less predictable in word choice and syntactic structure, this could only be done at the cost of a very great increase in program size. And the translation quality could not be guaranteed for emergency sentences. Since less than 1% of a typical text falls outside the stereotyped style, it is more efficient to let a human translator handle the unusual sentences.

3. Automatic translation of technical manuals

3.1. Technical translation: a prime candidate for automation

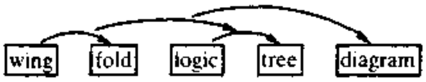
Technical documents constitute one of the most troublesome and expensive areas for human translation. A Canadian technical translator specializing in aircraft maintenance manuals must be familiar with a vast technical vocabulary in both English and French. His knowledge of the aircraft systems has to be virtually equal to that of the technician who will read the translation. It goes without saying that competent technical translators are a rare breed. This scarcity, in addition to making their services expensive, also limits the rate of translation in a given field. In a recent case where the documentation (in English) for a single airplane amounted to 90 million words, it was estimated that the four qualified translators would require well over ten years to complete the French translation, about the time it would take the airplane to become obsolete!

The high cost, the enormous volume and the urgency of technical translation have spurred the search for computerized aids. Although many technical documents are nearly impenetrable for the uninitiated reader, there are reasons why this kind of material may be relatively amenable to automatic translation. Despite the size of technical vocabularies it is easier to predict both the possible uses and correct translations of technical terms within a sublanguage than is the case for typical terms of unrestricted language. Thus the complexity of each word's dictionary entry in the machine is much less and the rules governing word translation much easier to discover and program. Furthermore, there is usually a much greater conformity of style between two languages in technical domains than in. say. newspaper articles.

3.2. Aircraft manuals - linguistic difficulties

The sublanguage of aircraft maintenance manuals is one of the most complex in which automatic processing has yet been attempted. This complexity has several sources, just one of which is the size of the technical vocabulary. On the level of sentence syntax two kinds of construction require special attention. First, several nouns may occur before a head noun as modifiers of it (and of each other), forming a so-called “empilage”. An expression such as *wing fold logic tree diagram* may be interpreted in many different ways, depending on the order in which the nouns and groups of nouns can modify each other. Just one of the possible interpretations is:

(7)



That is. a *wing fold logic tree diagram* is a diagram of a *wing fold logic tree.* A *wing fold logic tree* is a *logic tree* for a *wing fold.* A *wing fold* is a *folding* of the *wing*, etc.

Finding general rules which assign the correct hierarchical structure to each particular empilage requires an intricate semantic study of the possible relations between objects in that restricted domain. One of the ingredients of a successful solution is an analysis dictionary in which each word is classified according to a large number of possible semantic distinctions (for example, a feature designating “easily replaceable parts” may be ascribed to *filter* but not to *wing).*

A second serious problem arises in determining the scope of conjunction when two complex noun phrases are joined by *and* or *or.* The effect of conjunction in a complex expression such as *swivel joint and door hinge centre-line* could be analyzed by the machine in several ways, including the following three:

swivel [(joint) and (door)] hinge centre-line
(8) [(swivel joint) and (door hinge)] centre-line
[(swivel joint) and (door hinge centre-line)]

This conjunction scope problem is closely related to the problem of empilage structure. The solution of both problems requires specifying for each word in the dictionary (separately for each sense of the word) the semantic subcategories to which the word (sense) belongs. The analysis dictionary plays a crucial role during the assignment of the proper structure to each input sentence. The information ascribed to each dictionary entry (both syntactic features and semantic subcategory labels) can be determined only after a long linguistic study of the way language behaves in the specialized texts. In a complex domain such as aircraft manuals it may take a linguist-lexicographer several months to make a thousand dictionary entries. Since the technical domain may use tens of thousands of terms, it is easy to understand why dictionary-building is the most time-consuming aspect in the construction of technical translation systems.

*3.3. Processing sequence in the AVIAT1ON system**

Each English text treated by the AVIATION system is divided into segments called *processing units.* A processing unit may be a full sentence, the title of a section, or the label on a diagram. Each processing unit is handled separately by the program, passing through the following stages automatically:

Pre-edition and morphological analysis

After initial formatting, the inflected form of words (e.g.. *locks]* is replaced by their possible dictionary representations (e.g.. plural of noun *lock* or third person singular of verb *lock}.* The dictionary representations contain all information on the syntactic and semantic potential of the words.

Syntactic analysis

A syntactic processor scans the string of words, together with their dictionary information, from left to right to find a single sentence structure (or title structure, etc.) which will be compatible with each word's possible uses in the sentence. A “tree structure” is built for the sentence containing all information needed to translate it.

Lexical transfer

Each English word appearing in the tree structure is replaced by the appropriate French word, depending on complex rules which search the word's immediate environment in the tree structure. For example, when choosing the proper translation for the English verb *replace,* the lexical transfer rules check whether the direct object of the verb (easily found in the tree) is a noun which is marked with the semantic feature “easily replaceable part”. If so, the translation is French *remplacer* if not, the translation is *remettre en place.*

Syntactic transfer

In some cases, the syntactic structure of the English sentence must undergo major changes before an acceptable French sentence can be produced. For example, an English passive sentence *(This experiment was performed several times)* may correspond to a French active sentence *(On a fait cette expérience plusieurs fois).* After any such structural changes are carried out by the transfer rules, the new structure tree, complete with French words, corresponds to an acceptable French sentence.

Syntactic synthesis

In this stage, certain transformations convert the French structure into a linear string of French word forms. The word forms are not yet inflected, but carry coded morphological information to determine uniquely the proper inflected forms.

Morphological synthesis

The string of word forms is converted into a readable French sentence by rules which derive the proper inflected form from the coded information and base form. The translation is now complete.

3.4. Performance of the AVIAT1ON system

It is important to emphasize that the AVIATION system, unlike METEO, is designed only as a prototype or pilot translation system. This means that, in its present implementation, operating efficiency is not a primary goal. Rather, the system is designed to be as flexible

* For a fuller discussion of the stages of the AVIATION system, see the references cited at the end.

as possible, allowing changes and additions to be made in all stages as quickly as possible. While this flexibility is necessary during the building and testing of a new system, it comes at a heavy cost in operating efficiency. Once the grammars and dictionaries are relatively complete, it is possible to greatly improve efficiency and user comfort in a number of ways. One should therefore view with extreme caution any attempt to predict a prototype system's ultimate performance using statistics of its "unstreamlined" performance.

The AVIATION system has recently undergone an independent test based on a substantial corpus of new aircraft manuals pertaining to hydraulic systems. The evaluator's report reveals that somewhat more than half the translation units passed through the system, giving generally acceptable French, although the style was often heavy. The quality of this translation was rated at 80% of that for a human first-draft translation. Using a less subjective measure, the evaluators found that it took at least twice as long for a human revisor to correct the machine output as it did for the revision of the human translator's first draft. Although some of this additional time is due to difficulties in working with raw machine print-out, it nevertheless remains a striking difference and explains why the direct cost of machine translation (plus human revision) was calculated at over 18 cents per word compared to human translation (plus revision) of the same material at 14.5 cents per word. That is. even though the machine translation eliminates the need for the human first draft, the product takes much longer for the revisor (who is paid much more than any of the three human translators he normally revises) to correct.

The evaluators' report also makes clear that it is dangerous to conclude from these cost comparisons that the point of economic feasibility is close at hand. The user of a potentially commercial system must calculate the development cost of constructing large dictionaries and somewhat different grammatical rules for each new sublanguage. Even if the direct cost of machine translation plus revision can be reduced to 10 cents per word, it still requires a volume of several million words to recover the cost of developing a system as complex as the AVIATION system.

It would be unfair to conclude that the AVIATION system cannot *become* economically viable. During the recent evaluations most of the problems which resulted in mistranslation or in failure to translate a sentence were of the type whose solution is known. Coding errors in the dictionary accounted for a substantial proportion of these. Any multi-stage system is only as strong as its weakest link (or cumulation of weak links). What is difficult to estimate is the rate at which improvements can be made in a complex system where there is substantial interdependence between the parts undergoing change. When no firm basis can be given for predicting the time until economic viability can be achieved, economic planners cease to see a system as the solution to their pressing short-term problems of translation.

The AVIATION system should therefore not be viewed as a production system and probably not as the immediate prototype of a production system. Instead, it should be considered a "full-scale" experimental system of the most sophisticated type now in existence. The very enterprise of building a system which anticipates each and every linguistic problem which can arise in the most difficult kind of text imaginable has given a much better idea of what is involved in automated technical translation.

4. Future perspectives

We can draw several conclusions from the experience of the METEO and AVIATION systems. On one end of the scale of complexity, the METEO system clearly shows that computer-assisted translation is already economically advantageous in certain kinds of restricted texts. Weather bulletins, with an annual volume of over 10,000,000 words from Canadian sources alone, can be analyzed with a relatively small investment in dictionary construction and linguistic rule-building. On the other end of the complexity scale, the AVIATION system seems to show that, despite a large number of important linguistic discoveries, the most complex technical texts are not yet ripe for commercially viable machine translation. Translation services should therefore plan to use simpler computer aids, such as word banks, text editors and the like, in these linguistically complex areas.

This leaves open the important question as to whether some *intermediate level* applications exist for automated translation. A number of sublanguages exist which are less complex than aircraft manuals but more difficult than weather bulletins, and in which there may be sufficient volume to justify investing in a computerized system. Linguistic studies now in progress at Montreal's Contrastive Syntax Project indicate that stock market reports and other texts of the same complexity may be in the right range. The next few years will almost certainly see some intermediate-level systems come into useful production in Canada, as elsewhere in the industrialized world.

REFERENCES & ACKNOWLEDGEMENTS

For the purposes of this brief summary article, it has been necessary to simplify many complex issues. The reader interested in further details on the METEO system may consult Chevalier, M.. Dansereau. J. & Poulin, G. *TAUM-METEO* Groupe de recherches pour la traduction automatique, Université de Montréal, 1978. Linguistic aspects of the AVIATION system are given thorough treatment in Lehrberger, J. *"Automatic Translation and the Concept of Sublanguage"* to appear in Kittredge & Lehrberger (eds) *Sublanguage: Studies of Language in Restricted Semantic Domains* Walter de Gruyter, 1981 Many aspects of the cost and operation efficiency of the AVIATION system are covered in the *Final Report - Evaluation of the TAUM-AVIATION Machine Translation Pilot System* Secretary of State, Government of Canada, June 30, 1981. In addition to personal participation in the feasibility studies and first demonstrations of the TAUM and AVIATION systems during 1974, I have profited from many conversations with present and former members of the TAUM research group during the more recent evolution of these systems. Thanks are due in particular to John Chandioux, Pierre Isabelle, John Lehrberrger and Marcel Paré for information and documentation provided by them. Any errors in interpreting the current data or in predicting the future of such systems are entirely my own.