[From: *Statistical Methods in Linguistics* 1976]

B. Vauquois

AUTOMATIC TRANSLATION — A SURVEY OF DIFFERENT APPROACHES

Origin and motivations of automatic translation.

As this international conference COLING 76, like the others since 1965, is devoted to "Computational Linguistics", it may be a good opportunity to recall the role of automatic translation (A.T.) in the development of this field.

At the beginning, when Y. BAR-HILLEL, at M.I.T. was the first full time researcher, the motivation for A.T. was curiosity. The use of computers was almost restricted to computation in numerical analysis ; few scholars were thinking at other activities. Translating from a natural language into another with a computer appeared as a feasible and a very attractive task. At that time (1951-53), A.T. was the most important subject (perhaps the only subject) in the field of what has been called later "computational linguistics". During the following years, many laboratories in different countries, were created to survey and to make concrete experiments in the new area.

Moreover, beside the original curiosity, the increasing demand for translation brought a practical goal to A.T. which was believed to be a powerful and economical substitute to human translation.

In fact, for almost 15 years, this need for translation was considered exclusively for use in information gathering (for reading scientific and technical literature as well as newspapers published in foreign countries in their own language).

That is the main reason why in United States and later on in Great-Britain the target language was invariably English, in USSR it was Russian, in France and in Japan it was French and Japanese respectively.

In the early sixties the situation was the following :

On the one hand, experimental designs of automatic translation had been checked on some pairs of languages with limited corpus as one of these, more

developed, was almost ready to provide the users of translation with a large program, using enormous dictionaries, for translation from Russian into English. That was the system made by the Georgetown University which can be considered as the leader of what we call "the first generation" of A.T. Systems. In fact, this system operational since 1963, is still in use at the Atomic Energy Commission, Oak Ridge (Tennessee), and at the Euratom Common Research Center, Ispra (Italy). Similar other programs have been derived from this initial work at Georgetown.

On the other hand for many scholars, A.T. was rather considered as a source of inspiration for more academical studies. Realizing an A.T. system became a long range project ; and systematic research both in linguistics (analysis, generation, and comparison of languages) and in computer sciences (formal models of languages, algorithms for parsing, adequate programming languages, ...) took precedence of all considerations of utilization. That attitude reflected the research priorities followed by the so-called "second generation".

The characteristic features of the first generation

We have to keep in mind that the purpose of all programs characteristic of this generation is practical automatic translation, available as soon as possible. Furthermore, the period of designing such programs spreads from 1955 to 1960, at a time where the statements of linguistic models and their formalization could not be very helpful and the software available on computers offered poor facilities. Also, at the beginning of automatic processing of natural languages (devoted also to lexicography and quantitative linguistics), only the medium of punched cards was used. So, the activity of encoding extended to the level of processing by computers where programing became a subtle manipulation with positional codes for linguistic features which were chosen "a priori". Consequently the programs reflected the complete heuristics of the designer effectively including the grammar of the language by means of hierarchical questions represented by flow-charts.

The basic component of such a first generation system is the dictionary which furnishes all lexical and syntactic information, and the translation (or multiple translations) in the target language for each entry in the source language. Given an input text, the first step of such a system

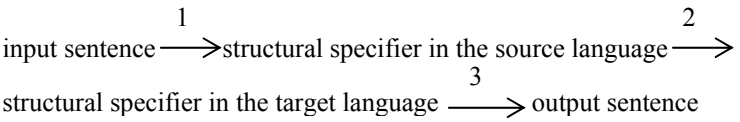
is a dictionary look up. In most cases this operation is performed by matching the form in the dictionary with the unanalysed occurrence in the text. However, in case of idiomatic expressions or particular strings of words, the longest match is usually preferred. For those source languages having many inflected forms for a single word, an elementary morphological analysis is sometimes performed by cutting the occurrence into a stem and an ending. After this dictionary look up, each occurrence (or sequence of occurrences) is replaced by the information found in the corresponding entry of the dictionary. Then, the next step consists in solving lexical ambiguities. Among the many kinds of ambiguities, the highest priority is given to the multiple syntactic class. For instance, for a word "MATCHES", it has to be decided whether it is a noun or a verb. A table of all specific syntactic ambiguities has been previously constructed and for each case, (verb-noun, verb-noun-adjective, verb-conjunction...), an appropriate sequence of questions about the preceding and following words, represented in the computer as a subroutine, has been established to find the correct solution. Some difficulties appear when many such ambiguities occur in the same sentence ; the order of application of the different subroutines is sometimes important to avoid either a lost of alternate solutions or a blocking of the system.
Nevertheless, the translating process goes on and consists in applying a sequence of translation routines dealing with words or groups of words. These routines make a reordering of the words based on a restricted context ; many of them are called by specific lexical entries. Finally, if it can be solved within the selected context, a morphological routine computes the grammatical agreement (for instance : verbal conjugation) and morphological alterations.

In conclusion, the strategy of the first generation system is based on a catalogue of linguistic facts which are locally relevant for a given pair of languages considered from the point of view of translation in one direction. The major guide for the composition of this catalogue and also for its use is the designer's knowledge of grammar and the experience of human translation. More sophisticated cases are solved, when possible, either by ad hoc subroutines (one for each case) or by direct translation in the dictionary, considering each such case as an extension of an idiomatic expression.

The characteristic features of the second generation

As early as 1957, V. YNGVE proposed "a framework for syntactic translation". The basic concepts of such a framework can be stated as follows :

First, the fragment of text considered as a whole is the sentence. Then, it is assumed that for each language a sentence may be adequately described by a structural specifier ; in fact, only limited indications about the kind of required information for such a specifier were provided as it was too early to define an adequate formalization. Nevertheless, the idea arose of a system proceeding in these three general steps. This is the first characteristic of second generation systems.

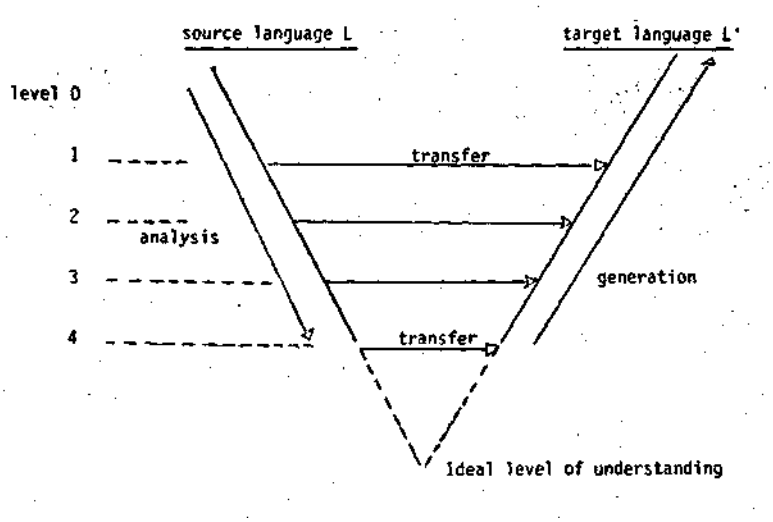


Step 1 deals only with the source language, it is the analysis (parsing) procedure ; step 3 deals only with the target language (generation from the specifier) ; step *2* involves both languages at some abstract level (for the moment, transfer restricted to syntactic structures).

The immediate consequences of such an approach were very important ; in spite of a large amount of investigation and fruitful development for many years, the research still continues in this framework, and further results are expected.

Indeed, this strategy matches the theory of stratification in natural languages and is well suited to the realization of computable models.

By a simple extrapolation we can imagine as many levels as we wish from the zero level (level of the text considered as a string of characters), asymptotically towards a level of understanding. At each level a formalization of the input sentence can be defined. Then, it may be assumed that the deeper the level chosen, the easier the transfer is. At the limit, if the ideal level of understanding could be reached for a given sentence in one language, the same structural specifier would represent all the paraphrases of this sentence in all languages.



During the 60's many laboratories worked within this framework, the selected level for transfer being more and more ambitious (from surface syntactic structure, to deep syntactic structure, sememic level, approximations of pivot languages, ...)
The second characteristic feature of this second generation concerns the way of representing linguistic data and the algorithmic approach.
The stratification description of natural language implies a kind of representation at each level by means of some artificial language ; Then, access to level n+1 from level n in analysis (or level n from level n+l in generation) needs a transduction from one artificial language to another.
For such a purpose, the concept of "model", based on formalized and computable languages, appears to be fundamental. Also, the notion of artificial language, considered as a set of strings over some vocabulary, has to be extended to set of rooted-trees as soon as deep levels are investigated.

By increasing degree of complexity, we can say that from level 0 to morphemic level a finite state string-to-string transducer is powerful enough to ensure a perfect matching of the model to the linguistic reality ; then a context-free parser has been used almost everywhere to simulate a string to tree transducer. The results of these parsings matches for the first time the idea of structural specifiers assigned to sentences of the input text ; even if a context-free model is not adequate for some sentences, the approximation is interesting.

Some laboratories performed automatic translations, at an experimental stage, using this level for transfer. That was the case of experiments conducted at the University of Kyoto and the University of Kyushu ; The autonomic division of the National Physical Laboratory (Teddington) had a similar project. At a higher degree of sophistication, many attempts have been made to extend a little more the power of context-free models in order to reach a higher level of adequacy for the structural specifiers ; in particular, interesting results have been obtained in many places (Rand, Harvard, IBM, Leningrad, Moscow, Grenoble,...). Finally, by the end of the 69's and in the early 70's, automatic translation has been obtained by a transfer at a deeper level. That is the case of Grenoble system introducing "pivot languages" and Montreal system which identifies relations between words and syntagms at a sememic level. The latter has recently achieved a system translating weather forecasts from English into French for the Canadian meteorological network. That is the only example of second generation A.T. system in use beyond an experimental demonstration.

A third characteristic feature of this generation concerns the way of programming. The different "Context-Free Parsers" and other transducers are generalized routines for which grammars and dictionaries are considered as data along with the text to be translated. Programs written for the first generation applied linguistic information directly to the input text ; on the contrary, the 2nd generation parsers and synthetisers compile for each grammar a program which operates in turn on the text.

The new look of Automatic Translation

At the present time, a few A.T. systems of the first generation are available; only one of the second generation is running (MONTREAL) and few others are expected in the near future (Leibniz group).

Research on a third generation began a few years ago. In contrast with the past, when we could seldom predict, either what would be available, or when, we now have a better idea of what can be expected in the near future (within 2-3 years) and in the further future (5-7 years).

For a long time. A.T. has been torn between two opposite goals : a concrete and efficient system for commercial use on one hand, and on the other, scientific research on computational linguistics. It seems, now, that

some concrete applications are being obtained from the scientific side. Certainly, a "fully automatic high quality translation" is not reachable in any foreseeable future, but we can expect feasible systems in which the complete process of translation is shared between human translator and computer. Before surveying the current trends which are developing in different places, we have to mention the new motivations for A.T. research and production.

About 20 years ago, the primary demand was devoted exclusively to the translation for information gathering ; this motivation still persists with an increasing intensity ; but in addition for the last few years, translations are needed more and more for the dissemination of information ; in particular, that is the case for multilingual organizations (for example : the European Community, bilingual agencies in Canada, ...). Certainly, the possibilities of human translation cannot face the totality of this demand any longer ; the only solution lies in computerized systems, in which the knowledge amassed in the past 15 years from the scientific approach should bring positive results.

If we consider what has been done within the framework of the second generation and what could bring immediate results in artificial intelligence, we can see how to satisfy simultaneously, the requirements of scientific research and the expected demands for translation.

The artificial intelligence approach to natural language processing is mainly (if not exclusively) oriented toward the semantic interpretation of texts. This does not mean that all of the second generation systems were restricted to syntactic description without any semantic considerations. On the contrary, the most sophisticated systems included some "semantic features" assigned to the lexical entries of their dictionaries and these features were used by their grammars for "semantic agreement", the same way as syntactic codes were used for grammatical agreement. In other words, the grammar rules were stated in terms of conditions on the combinatorial properties between classes which could be semantic classes as well as grammatical classes. More ambitious are the goals of semantic computation in artificial intelligence. Inference rules should be applicable to deduce new statements. The semantic consistency of a sentence or of a sequence of sentences should be open to evaluation. Furthermore, if a data base consisting of an appropriate description of knowledge about the "world" is stored in the computer, then any part of the input text (or deductions computed from it) should be open to evaluation for consistency with respect to that data base.

This approach seems to be the only means to solve the remaining ambiguities occurring at the end of syntactic analysis (even strengthened by semantic features) ; in particular, the reference of a pronoun outside the sentence cannot be determined by a sentence to sentence translation,

The introduction of such a semantic component in a A.T. system is the characteristic feature of the so-called "third generation".
However, all experiments conducted within this artificial intelligence approach are restricted to "microworlds" ; it is too early to consider seriously a generalized use of this method, given the amount of information about the world needed for a real translation of large amounts of text. But research in this third generation certainly needs to be extended.

Considering now the feasibility of A.T. systems which merge human translators and the computer in a hybrid process, we can imagine several different strategies.

Let us assume a large translation service (about 1 500 translators and revisors) where most of the texts have to be translated into several languages. In such a case, it seems, at least for the moment, that the work of the human translators must be separated from the work of the computer.

The complete process would be first a preedition of the text (by inserting specific disambiguating markers) ; then, the automatic translation system operates on this edited text ; and finally the output produced by the computer is revised to get the desired quality of translation.
Of course, the balance between preedition and revision must be optimized according to the following considerations : ambiguities in the source text which are not solvable by the A.T. system, quality which is desired for the final result, number of target languages..... It is necessary to find a preedition code flexible enough to ensure such a balance and compatible with the automatic device.

Another strategy would be a machine translation aided by human translator in a conversational way. It is certainly the ideal way for the future ; it would be interesting to develop experiments on a small scale (with a few users in time sharing) to improve the best ways of doing so.

In both eases, as far as the automatic system is concerned, several remarks about the way of progressing can be enumerated :

a- The third generation approach can be considered as an extension of the second generation systems with a greater ambition concerning the transfer level (on the contrary, first generation systems cannot be assimilated). So, there is no obstacle to build a system of second generation with the possibility of increasing its power every time a suitable progress in artificial intelligence is available.

b- The stratified models of the second generation are extremely rigid ; the system applies this models in sequence without any interference between them. A new approach consists in an arrangement of the grammars in such a way that the computation of the structural specifier at each level is not strictly sequential, but interfering each other. A formalized representation of specifiers sharing the same graph structure but identified by different labels on the nodes has been proposed by the A.T. section of the LEIBNIZ Group two years ago.

c- As a consequence of the preceding remark, the usual parsers of the 60's (responsible for this rigidity) are not the suitable software any longer. Among the new software systems designed for a flexible computation at different levels, let us mention M. KAY's system, WOOD'S extended transition networks, Q-systems, REZO and TARZAN developed at Montreal, ATEF, CETA, SYGMOR at Grenoble.

In conclusion, the activity in A.T. got a renewed interest just a few years ago. All the progresses realized in the processing of languages and in specialized software contribute to successful realizations in the near future.