[*Science* 8 May 1964, volume 144, number 3619]

Mechanical Translation and Related Language Research

Research has provided a framework for solving many linguistic and information-processing problems.

Richard See

To many scientists the mechanization of the translation process is an intriguing problem and one worthy of serious study. To others it has seemed a will-o'-the-wisp. Jerome Wiesner, for example, said *(1)*, "Language, which most people use with little or no effort, at least when speaking, is an extremely complicated process which is only partially understood. This complexity and lack of understanding is shown very clearly when we attempt to tell a computing machine how to translate text from one language to another. This is something which a machine should be able to do, though so far it has not been done successfully: the difficulty is man's, not the machine's. We do not have an adequate understanding of the organization and operation of language to describe the translation process to the machine. There is a similar difficulty when we attempt to deal with problems of information retrieval and data storage." On the other hand Yehoshua Bar-Hillel, a well-known critic of this field, wrote (2), "Fully automatic, high quality translation is not a reasonable goal, not even for scientific texts. A human translator, in order to arrive at his high quality output, is often obliged to make intelligent use of extra-linguistic knowledge which sometimes has to be of considerable breadth and depth. Without this knowledge he would often be in no position to resolve semantical ambiguities. At present, no way of constructing machines with such a knowledge is known, nor of writing programs which will ensure intelligent use of this knowledge." Whatever the ultimate limit of mechanizability of translation may be—and it very probably falls short of such creative activity as the artistic translation of literature, poetry, and the like—that limit has not yet been probed. The origins of mechanical translation and related research, its present status, and its possible future directions and applications are discussed here.

The beginnings of such research may be dated from 1949, when Warren Weaver sent a memorandum to some 200 of his acquaintances in various fields suggesting that language translation by computer techniques might be possible *(3)*. The idea attracted considerable interest, but research was slow in starting. In 1954 the feasibility of carrying out prescribed linguistic operations in electronic digital computers was demonstrated by workers at Georgetown University and the International Business Machines Corporation. In the same year the first federal grant for a mechanical translation research project was made; this was a grant to the Massachusetts Institute of Technology by the National Science Foundation. The number of projects in the United States has gradually grown to ten or so, most of them at universities. Projects have also been established in a number of foreign countries—in Great Britain, the Soviet Union, France, Japan, Italy, Czechoslovakia, Mexico, Belgium, Poland, and others.

The author is program director for information systems in the National Science Foundation's Office of Science Information Service, Washington,D.C.

A Mechanical Translation System

There has gradually developed a conceptual model of a mechanical translation system for simulating the human act of translating from one language to another. The system is envisaged as consisting of three basic parts —an input, a processor, and an output. Although the possibility of translating from and into spoken languages, or between spoken and written language, has been considered, attention has thus far centered almost exclusively on translating to and from written languages. One reason for this emphasis is that the need to translate written documents is a more pressing problem, particularly where scientific information is concerned.

The input component would automatically recognize, and transcribe in some mechanically usable form, the letters, numbers, and other symbols of which a text is made up. Devices for this purpose already exist for certain limited styles and sizes of type under certain restricted conditions of page size, spacing, and so on. Furthermore, the state of the art in this field (called "optical character recognition") is such that fairly general character-reading devices for mechanical translation could be constructed in a few years if the need justified the cost. Such devices are in fact being constructed for other purposes, but since the central, or processing, component of a mechanical translation system is even further from realization, the lack of such an input device is not the main obstacle to the achievement of mechanical translation

at this time. For purposes of research, the necessary linguistic material is prepared by punching a record in a paper tape or card in such a way that the result simulates the product of such an input device. Until very large quantities of text are required for research, this method of double simulation—that is, simulating the proposed device, which itself is to simulate the human act of reading—will be more economical than construction of a generally useful print reader.

Before the processor—the second, and central, component in a mechanical translation system—can be designed, it is necessary to have an explicit understanding of the way in which the translation process may be carried out. Once this has been achieved, it will, of course, be necessary to design a physical device, the processor, which can accept the information produced by the input device and process it in such a way that a text in the desired output language is produced in a form that can be accepted by an output device, the third component required in a mechanical translation system.

Generally useful output devices are already available and in use in data-processing systems. Printing output devices are available today which can print upper- and lower-case letters and a variety of symbols. Still more versatile devices are being designed and constructed, but these must be considered unnecessary refinements until mechanical translation has been achieved and has been shown to be economically feasible.

The second component, the processor, is the heart of the problem, since the problems of input and output have been solved, at least in principle. For research purposes it has been demonstrated that a general-purpose digital computer is adequate to implement the procedures that have been considered up to the present, and virtually all research on mechanical translation is based on the assumption that such computers will be used for this work for some time to come. The reason for this is that the intellectual problems involved in the simulation of human translation are great and only partially understood, and the problems are conceptual ones concerning the automatic analysis and synthesis of text rather than practical ones concerning the design of economical systems based on well-understood processes. Thus, at present, computers cannot be programmed to produce output comparable to good, or even fair, human translation. Present computer programs, however, serve to record partial procedures already developed and are useful for experimentation.

In mechanical translation research, language processing of gradually increasing complexity has been considered. This increase in complexity can be described in terms of five levels of structure and of processing: item-for-item, morphological, syntactic, transformational, and semantic. These are briefly discussed later. To the extent that these levels of processing can be made to correspond to the structure of actual languages, it should become possible to simulate the translation process more closely. For most researchers now agree, as some have believed from the beginning, that ultimate success in this field must come from a more fundamental understanding of the problems involved rather than from a "black box" approach in which ad hoc procedures are linked together in various ways in an attempt to achieve good output by noting the effects of successive small changes in the system and incorporating those changes which seem to be beneficial.

Item-for-Item Substitution

The first, and most primitive, level of processing is that of item-for-item substitution, in which linguistic items, whether they be parts of words, whole words, or groups of words, are processed independently and converted one after another into corresponding items in another language. This sort of substitution is often called "word-for-word translation," although it may involve items larger or smaller than words and is certainly not true translation. Successful processing at this most primitive level was achieved at an early date. Computer programs were prepared which could utilize a list, or "dictionary," of full word forms in one language and equivalent forms in another and cause a computer to carry out the operations necessary for performing item-for-item substitution on textual information. Although the process is straightforward in principle and easily implemented for a few thousand items, when dictionaries of ten or 100,000 items are made there are serious problems arising from clerical inaccuracies which occur in the course of revising or expanding the dictionary and which lead to listing of more than one substitute for the same item, to loss of items already in the dictionary, and to the introduction of incorrect items.

Although it falls far short of actual translation, item-for-item substitution has been found useful in lieu of translation by some scientists, even though it has not been broadly accepted as a final product. One reason for the lack of general acceptance is that any given word form is likely to have not just one but a number of possible equivalents, and one must either program the computer to print out all equivalents (or at least a representative collection) or run the risk of suppressing the appropriate equivalent in some contexts. These multiple equivalents may be useful to some scientists but may present a bewildering array of choices to others. On the other hand, recent studies in the Federal Republic of Germany indicate that item-for-item substitution can serve as an aid to professional translators and increase the quality while decreasing the cost of translation, and plans are being made to implement automatic language processing on this level in actual translation work.

Morphological Processing

Second in order of complexity is what we may call morphological processing, which takes into account the morphology of a language—that is, the rules concerning the relationships between derived forms and the basic dictionary forms of words. This refinement of item-for-item processing permits the use of a smaller dictionary but requires a larger and more complex program. Here again, as in the case of item-for-item substitution, the presence of multiple equivalents for many words, or their absence when they should have been given, is an obstacle to direct use of the output.

Such processing, although more complex internally, still only performs item-for-item substitution of words or phrases. Thus, the increase in complexity in this case, theoretically at least, does not lead to any improvement in the quality of the output but merely represents a reorganization of the program. In practice, however, the labor of listing all of the forms of some words, such as Russian verbs, has led researchers to work with in-

complete lists of forms in item-for-item substitution, and the coverage provided by the dictionary would be less in such cases than the coverage in morphological processing based on the same words.

One example of morphological processing, in which there is substitution both of items smaller than words and of items consisting of phrases of several words, is the experimental system providing Russian-to-English substitution, which is now being evaluated at Wright-Patterson Air Force Base. The output is being subjected to considerable revision before being distributed. This system was developed with the support of the U.S. Air Force, largely by International Business Machines, Inc. Much of the work of compiling the special dictionary of over 100,000 items for this system was done at the Library of Congress.

Syntactic Translation

Syntactic processing, the third level of language processing in mechanical translation, is a process based on the syntax of languages—on the rules which govern the arrangements of words to form sentences. Clearly, the syntax of both the language from which the translation is being made (called the source language) and the language into which the source language is being translated (called the target language) must be taken into account in attempting to simulate translation. The natural tendency in mechanical translation research has been to concentrate first on the syntax of the source language in order to be able to analyze, or parse, automatically the source-language text that is being processed. Automatic parsing turned out to be a difficult problem, and only after many years of effort have reasonably complete computer programs for automatic parsing been completed. To put it very simply, such a program performs two basic functions: it (i) adds to each word of a source text the set of possible grammatical categories into which the word might fall, and (ii) derives from this sequence of sets of categories one or more possible syntactic analyses, or parsings, of the text. One example of such a program is the Harvard program for syntactic analysis of English, which has recently been included in the SHARE system for distribution of computer programs *(4)*.

Experimental simulated translations from Russian to English have thus far been based largely on syntactic analysis of the Russian, with a view to resolving some of the ambiguities in the source text. Only modest attempts have been made to follow up the analysis with generation of the English simulated translation in accordance with English syntax. As a result, simulated translations have been largely ungrammatical. In one important way, however, improvement can be, and has been, made on this level through the use of parsing to eliminate those multiple equivalents that correspond to grammatical categories inconsistent with the syntactic analysis or analyses of the text. Even were the syntactic analysis complete, however, a considerable number of the multiple equivalents would not be eliminable by this method, since alternative translational equivalents often are grammatically identical. It follows that in any simulated translations where no alternatives are presented, many equivalents will be incorrect, since there is at present no general method known for making a correct selection.

Had it not been for the eagerness of those requiring translations, the automatic simulation of translation might have been postponed until more was known about the problem. In fact, most researchers have not yet attempted the simulation of translations. Thus far, as I have pointed out, only one-sided syntactical processing has been attempted. Probably it will be possible in a few years to experiment for the first time with two-sided syntactical processing, in which syntactic analysis of the source language is linked to syntactic generation of the target language. This processing should produce a closer simulation of translation than has been achieved in the past, though probably still not sufficiently close a simulation to be widely usable as a substitute for translation.

Several possible applications of syntactical analysis programs have been studied and experimented with. A program devised by the University of Pennsylvania was applied, by the Radio Corporation of America, to the compression or condensation of English texts through the suppression of certain subordinate constructions (5). Another example is the research now being done at Harvard (6) on information retrieval through the use of syntactic analysis.

Transformational Translation

The fourth level of language processing which has been considered is transformational processing. Transformations are rules which indicate how various syntactic constructions in a language are related to, or may be derived from, other constructions which are regarded as simpler or more basic. The collection of transformational rules for a language is called a transformational grammar. Transformational processing, or transformational simulated translation, is language processing based on the use of transformational grammars both for the analysis of the source language and for the generation of the target language.

One attractive feature of attempting to simulate translation on this fourth level is that, for the first time, there appears to be available a reasonable mechanism for performing the heart of the translation process, the step between the analysis of the source language and the generation of the target language.

The description of this step that I gave earlier in discussing simulated translation on the syntactic level was necessarily quite vague. This is because it is not entirely clear just how the "linking" will take place. On the transformational level, however, the situation is different in that the representation of the input structure obtained after analysis can be made very concise, and it appears likely that correspondences can be set up between the source and target transformational representations which can be used as a basis for simulated translation.

There is considerable interest in transformational grammar outside the field of mechanical translation research, and work by linguists in this field should prove useful in mechanical translation research. Fairly complete transformational grammars of English have been developed, but present knowledge of the transformations of other languages is very incomplete.

Some preliminary work on transformational translation from German to English has been done at the University of Pennsylvania. Also in progress there is work aimed at using transformational representations to normalize English text for the purpose of searching for information. It should also be possible to translate from English to English—that is, to paraphrase—once this stage has been reached.

Semantic Translation

The fifth level of language processing which has been considered is the semantic level. Although transformational translation should represent a considerable advance over processing on the lower levels, it seems likely that some account must be taken of the meaning or content of language in order to produce satisfactory simulated translations, for in translation it is, after all, the meaning which must be conveyed.

Research on the problems of meaning in translation is under way, and some insight into the nature of possible solutions has been gained, but much work lies ahead. In summary, then, for some pairs of languages—for example, in simulated translation from Russian to English—language processing can now be done at a level somewhat below syntactic simulated translation, the third level described. Processing on the first two levels may lead to increased translation capacity if such techniques can be suited to the use of translators. Evaluation of the usefulness of future processing on the third and higher levels will have to await future developments.

Some Research Projects

The oldest mechanical translation project in the United States is that at the Massachusetts Institute of Technology. From its inception, the work has been concentrated on fundamental principles, and one discovery, that of the so-called depth hypothesis, has been recognized widely as being of great potential significance to mechanical translation. In addition, a special programming language for mechanical translation, called COMIT, has been devised by workers on this project and is being used by many other groups, both in the United States and abroad. Grammars for analysis and synthesis of a number of languages have been studied, and work has begun on various approaches to semantics, or content. This project is supported by the National Science Foundation.

The only mechanical translation research project within a government laboratory is that supported by the U.S. Army at the National Bureau of Standards. The goal of the project is the development of a practical procedure for translating from Russian to English. The project is known for having originated the method of parsing called predictive analysis (7).

A project at Harvard is supported by the National Science Foundation and was in the past also supported by the U.S. Air Force. It has studied the problems of constructing dictionaries for automatic mechanical translation from Russian to English and, more recently, it has concentrated on methods of parsing Russian and English that are based on the Bureau of Standards' method of predictive analysis.

Workers at the University of California, Berkeley, have been conducting research on mechanical translation from Russian to English and from Chinese to English, with emphasis on texts in the field of chemistry. This project, also supported by the National Science Foundation, has stressed the application of principles of linguistics to the analysis and simulation of the translation process. The Chinese characters that are important in scientific literature have been collected and thoroughly cross-indexed for the first time *(8)*.

A project at the University of Texas is supported both by the National Science Foundation and by the U.S. Army; it is concerned with research on English, German, Russian, and Chinese from the standpoint of mechanical translation and related linguistic processing. An integrated system of computer programs has been completed which permits the efficient gathering and organization of linguistic data for research.

At Wayne State University the U.S. Navy supports a project concerned with translation of mathematical texts from Russian to English. This project has concentrated its study on a small sample of Russian mathematics texts and has intensively analyzed certain important grammatical constructions.

A comparative study of Chinese and English transformational grammars is being supported at Ohio State University by the National Science Foundation.

In addition to the projects which have concentrated on basic research, there are three projects in the United States which, besides carrying on research, have brought to a point of operation programs that produce crude output which resembles human translation in some respects. All three have been concerned with translating from Russian to English. One of these projects, that at the Bunker Ramo Corporation (9), has concentrated on Russian texts in the field of nuclear physics, but it is broadening its work to include several other fields as well, The output produced by this project is based on a parsing of the Russian text, a limited effort being made to produce correct English through mechanical rearrangement of word order, insertion of articles, and the like. Thus, the output, in English words, in some cases is understandable, but often the word order is that of the Russian and the output is unreadable as English. In many cases, more than one possible meaning is provided for a given Russian word and one must try to select the appropriate meaning from context. Research is being carried on to improve the quality of the program, particularly through comparison of the experimental output with an actual translation. This project was supported earlier by the Air Force and is currently supported by the National Science Foundation.

The second such project, at Georgetown University, was supported during the research phase by the National Science Foundation, in part with funds transferred from the Central Intelligence Agency. Between 1959 and 1962 the project was supported directly by the Central Intelligence Agency, and an effort was made to produce a large number of translations economically by post-editing the computer output. The method used is similar to that of the Bunker Ramo Corporation in that it is based on parsing and procedures for some rearrangement of word order. After careful evaluation it was determined that the combined cost of preparing text for input, of computer processing, of output printing, and of post-editing made the process uneconomical at present. More recently the project has been supported by Euratom and by the Atomic Energy Commission.

Third, the I.B.M. Corporation has, with the support of the Air Force; completed a unique, but nonetheless general-purpose, computer for mechanical translation. In conjunction with this project the Air Force contracted for development of an experimental input device for mechanical translation from Russian to English—that is, a Russian print reader—as well as for adaptation of a photocomposition device for the output. The Russian print reader has never become operational, and manually operated tape-typewriters provide the input. As mentioned earlier, the system is being evaluated

at Wright-Patterson Air Force Base. In addition to constructing and experimenting with equipment for mechanical translation, I.B.M. has conducted research on the translation process as well.

Related Language Research

As mechanical translation research progressed, it became clear that many of the problems with which it was concerned were also of importance to other research areas, such as automatic abstracting, automatic indexing, and automatic information storage and retrieval. In particular, the techniques for thorough analysis of the source language that were sought as a step toward the mechanical translation of foreign languages should, if applied to English, provide a starting point in developing methods of information retrieval. Interestingly, many researchers on mechanical translation, after an initial period of experimentation on specific languages, came to seek generalized methods of syntactic analysis applicable to a number of languages without change in the underlying principles. Because of the close relationship between programs for analysis and programs for synthesis, several translation projects began applying their methods to the analysis of English. Thus it appears that programs for syntactic analysis should be divisible into two parts, a language-independent procedural part and a language-dependent data part which provides the specific facts about the language to be analyzed. One advantage of this approach is that a procedure which can be shown to be applicable to a second language can be said to have been experimentally verified. Attempts to apply essentially the same procedures to more than one language are being made at the University of Texas (English, German, Russian, and Chinese), at the University of California (Russian and Chinese), and at Harvard (English and Russian).

Among research groups working on information retrieval and also studying English syntactic analysis are groups at the University of Pennsylvania, Indiana State University, and the System Development Corporation. A group supported by the National Science Foundation at the University of Pennsylvania, mentioned earlier, has for a number of years been developing procedures for the syntactic and transformational analysis of English. These workers have investigated various procedures for syntactic analysis, one of which is closely related to the method of predictive analysis mentioned earlier. In addition, they have found it desirable to study certain foreign languages as well, to gain further insight into the problems found in studying syntactic analysis of English.

The group at the University of Indiana is working on mechanical analysis of scientific English in order to develop ways of storing it and retrieving information from it. Work is now beginning there on the automatic semantic analysis of English. This project is supported both by the Air Force and by the National Science Foundation.

At the System Development Corporation the Synthex research project has been concerned with programs to automatically index English text and to select answers in response to questions in English about the text, which includes 16 volumes of the *Golden Book Encyclopedia.* This group is also working on the automatic generation of grammatically correct nonsense in order to better understand the role of syntax in language.

These projects, together with the mechanical-translation projects, represent a broad area of research concerned with the formal and computational aspects of language processing. Researchers in this area share a belief in the importance of discrete algorithmic procedures, as opposed to procedures based on probabilistic and statistical calculations, feeling that formal computation on language information should be carried as far as possible before other methods are resorted to. As a result of this community of interest, a group of interested researchers recently formed the Association for Mechanical Translation and Computational Linguistics. As its name implies, its members include scientists in the field of mechanical translation and workers in other kinds of linguistic research based on similar computational techniques and a similar philosophy.

Support and Coordination of Research

Research on mechanical translation has been supported over the past 9½ years principally by five government agencies: the National Science Foundation, the Central Intelligence Agency, the Army, the Navy, and the Air Force. The variation in the definitions of the scope of this research has made it difficult to gather consistent data about support. This has led some to make extravagant estimates as to funding. For example, Yehoshua Bar-Hillel, in a paper presented at the International Federation for Information Processing Congress in 1962 *(10),* stated that "tens of millions of dollars have already been spent on machine translation research."

To arrive at a reasonable figure it is necessary to distinguish between the support of projects engaged in the design and construction of hardware, such as computers and optical character readers, and the support of research and experiments on the linguistic aspects of the problem. Authoritative figures for the United States and for certain foreign research supported by the United States for the years through fiscal year 1960 are to be found in *Research on Mechanical Translation,* a transcript of hearings before the Special Investigating Subcommittee on Science and Astronautics, U.S. House of Representatives, 86th Congress, 11, 12, 13, and 16 May 1960. The figures total approximately $3 million for the five agencies for research on mechanical translation proper (that is. linguistic research) and slightly more for development of hardware. By the end of fiscal year 1963 the figure had grown to about $8 million for mechanical translation research and perhaps to a like amount for development of hardware. (Expenditures in other countries would add only a fraction to this total.)

There are good reasons for not lumping the hardware figure with the linguistic figure. First, it is difficult to decide which kinds of hardware development should be included. Second, hardware developed specifically for mechanical translation has been used in other fields, and it is difficult to assign any particular portion of the funds to this field. Finally, because linguistics and hardware are so different, an "apples and pears" addition of the two is artificial and apt to be misleading.

On 19 July 1962, representatives of the National Science Foundation, the Department of Defense, and the Central Intelligence Agency, in order to strengthen coordination in the field of mechanical translation and related language research, agreed upon a "Joint Research and Development Program for Automatic Language Processing."

Automatic language processing under the plan includes mechanical translation, computational linguistics, and related work in areas such as automatic abstracting and development of hardware. The adoption of this plan constitutes recognition by the agencies involved that "fully automatic high-quality language processing, including mechanical translation, is a long-range goal," and that cooperation in planning and research are necessary to progress in this field.

A recent important step under the Joint Automatic Language Processing Program has been the appointment by the National Academy of Sciences of John R. Pierce as chairman of an advisory committee for automatic language processing. When fully established, the committee will provide the agencies participating in the joint program with advice which will aid them in planning future research, development, and evaluation in this field.

Warren Weaver concludes his note "Translation," referred to earlier (3), by indicating four possible types of attack on the mechanical translation problem. As for the fourth, he says,

> Indeed, what seems ... to be the most promising approach of all is one based on ... an approach that goes so deeply into the structure of languages as to come down to the level where they exhibit common traits. . . .
>
> Such a program involves a presumably tremendous amount of work in the logical structure of languages before one would be ready for any mechanization. . . . But it is along such general lines that it seems likely that the problem of translation can be attacked successfully. Such a program has the advantage that, whether or not it lead to a useful mechanization of the translation problem, it could not fail to shed much useful light on the general problem of communication.

The history of mechanical translation research has shown Warren Weaver's insight of 15 years ago to have been remarkably prophetic.

References and Notes

1. J. B. Wiesner, "Communication sciences in a university environment," paper presented at the Conference on Scientific Information, San Jose, California, 1958, and published in *IBM J. Res. Develop. 2*, 271 (1958).
2. Y. Bar-Hillel, "The present status of automatic translation of languages," in *Advances in Computers*, F. L. Alt, Ed. (Academic Press, New York, 1960), vol. 1, p. 135.
3. The entire note is reproduced in *Machine Translation of Languages*, W. N. Locke and A. D. Booth, Eds. (Massachusetts Institute of Technology Press, Cambridge, 1955), pp. 15-23.
4. *Sci. Inform. Notes 5*, No. 6 and 7 (1963-1964).
5. W. D. Climenson, N. H. Hardwick, S. N. Jacobson, *Am. Doc.* 12, No. 3, 178 (1961).
6. G. Salton, *ibid.* 14, No. 3, 213 (1963).
7. I am grateful to the National Bureau of Standards and to the project leader, Mrs. Ida Rhodes, for the opportunity to participate for a time in research there on mechanical translation.
8. C. Dougherty, S. M. Lamb, S. E. Martin, "Chinese Character Indexes" (Univ. of California Press, Berkeley, 1963).
9. The Bunker Ramo Corporation was formerly the Ramo-Wooldridge Division of Thompson Ramo Wooldridge, Inc.
10. Y. Bar-Hillel, "Machine translation: The end of an illusion," in *Information Processing 1962* (proceedings of the International Federation for Information Processing Congress, held in Munich, 1962), C. M. Popplewell, Ed. (North-Holland, Amsterdam, 1963), p. 331.