[*Proceedings of the National Symposium on Machine Translation*, UCLA February 1960]

Session 2: CURRENT RESEARCH

MT AT THE MASSACHUSETTS INSTITUTE OF TECHNOLOGY[1]

Victor H. Yngve
Massachusetts Institute of Technology

Mechanical translation has had a long history at M.I.T. Shortly after the Warren Weaver memorandum of 1949, Yehoshua Bar-Hillel became the first full-time worker in the field. He contributed many of the early ideas and will be well remembered for this. He organized the first conference on mechanical translation, held at M.I.T. in June of 1952. It was an international conference, and although there were only 18 persons registered, nearly everyone interested in MT in the world at that time was there. Of those 18 people, 4 are on the program of this conference, Leon Dostert, Victor Oswald, Erwin Reifler, and myself. The number of people here today gives a measure of how the field has grown in the intervening 7- 1/2 years. And this is a national, not an international conference. The second conference, also held at M.I.T. and also an international conference, took place in October of 1956. At that conference there were about 30 in attendance.

The reports or proceedings of both these conferences were published in the journal Mechanical Translation. This journal was founded at M.I.T. in 1954 when it became obvious that there was a need for better communication between those interested in MT and to prevent needless duplication of effort. The journal has continued to grow. The first volume contained 57 pages. The current volume, volume five, will contain well over twice that number. Starting with the next volume we will abandon the electric typewriter and photo-offset format, and go to letter press. This will give us a more attractive journal, will allow it to expand naturally, and will speed up the process of publication. We feel at M.I.T. that we are holding the journal in trust until the field comes of age. When the field has grown to the point where it becomes desirable to found a professional society, the journal can become its official organ.

Let us now turn to the research on mechanical translation at M.I.T. The group at M.I.T. has always stressed a basic, long-range

1 This work was supported in part by the National Science Foundation, and in part by the U.S. Army (Signal Corps), the U.S. Air Force (Office of Scientific Research, Air Research and Development Command), and the U.S. Navy (Office of Naval Research).

approach to the problem. We are placing an emphasis on completeness where completeness is possible and on the attempt to find out how to do a complete job where completeness is not now possible. We are not looking for short-cut methods that might yield partially adequate translations at an early date, an important goal pursued by other groups. Instead we are looking for methods that will be capable of yielding fully adequate results wherever they apply. We are thus seeking definitive solutions that will constitute permanent advances in the field rather than ad hoc or temporary solutions that may eventually have to be discarded because they are not compatible with improved systems.

The framework within which we are working was described about a year and a half ago in Mechanical Translation.[2] There were two main points in that paper. The first one was concerned with the aspect of completeness and with the point that it is essential for us to understand and use as much as possible of the syntax of the languages being translated. For many years the M.I.T. group has been working in the field of syntax. The other point in the paper was that it is possible, and perhaps necessary, to divide the problem of mechanical translation into six parts, each one fairly independent of the others. We are pleased that other groups are also adopting this same split, because we think it has a lot of merit.

A split of the problem into six more or less separate problems is a great advantage, because not only can more people work in parallel on the over-all problem by a division of effort, but also each part is easier to solve than the whole problem. The six-way split consists in reality of a two-way split and a three-way split. The two-way split is between the program or manipulative aspect of the problem and the static or stored knowledge aspect of the problem. Such a split would, for example, separate a recognition routine or a sentence-production routine from the grammar or rules of the language. With a split of this nature in a program it becomes much easier to make additions to the grammar rules without having to reprogram the routines. Another advantage is that the programs are easier to Understand and thus easier to improve. The three-way split is

2 "A Framework for Syntactic Translation", Mechanical Translation, vol. IV, no. 3.

between the problems concerning the input language only, the problems concerning the output language only, and the problems concerning the two languages simultaneously. We thus conceive of a three-step translation routine. The first step is recognition of the structure of the input text, the second step is the selection of the structure for the output text that will give the best translational equivalence, and the third step is the production of the actual output text from the specification of its structure. Again, the advantages of this split of the program are great. Here, particularly, there is a great simplification in keeping the monolingual phenomena separate from the bilingual translation phenomena. The result is an increased clarity of the issues. They are easier to cope with separately.

Since we start with input-language text and go through a three-step program resulting in output-language text, there are two intermediate encoded forms of the message, the coded form of the message that passes from the first or recognition step to the second or structure-transfer step, and the coded form of the message that passes from this structure transfer step to the third or text-production step. These two forms of the message we call specifiers. These specifiers are in no way to be considered as intermediate languages or universal languages. The specifier that passes from the recognition step to the structure-transfer step is an explicit representation of the structure of the input text in terms of the categories appropriate to the input language. It is merely a recoding of the input text with everything of importance made explicit. Similarly, the specifier that passes from the structure-transfer step to the text-production step is an explicit coded form of the structure of the output text. If we were to consider translating through an intermediate language so as to save on programming, a course of action that we do not recommend, we would have to use a six-step program instead of a three-step program. Our three-step program already involves some of the advantages usually attributed to the use of an intermediate language. Our first step, or recognition routine, can be common to all programs translating out of that language, and our last, or text-production step, can be common to all programs translating into that language. Only the middle, or transfer step, needs to be different for every pair of languages and for each direction of translation between the languages of the pair.

In order to write adequate translating routines, we need, among

other things, an adequate and detailed knowledge of the languages in question -- a knowledge of their formal properties as codes and a knowledge of how they are used to communicate. Linguistic research on the structure of individual languages thus constitutes an important part of our effort. German and English are being given primary attention. French is being studied also. We have no work going on in Russian. Each language is being studied as an isolated system. The relationship between languages is a separate question and is being given separate consideration.

Work on English grammar is being carried out by Edward S. Klima, David Lieberman, and V. H. Yngve. The work of Edward Klima, following the theoretical work of Noam Chomsky, has been most detailed, and extensive. He has done work on the imperative, on the use of "ing", on the relative clause, on pronouns, and on negation. Some of this work has already been submitted for publication and should appear shortly.

Work on German grammar is being carried out by Joseph Applegate, John Bross, Rosemarie Straussnigg, and John Viertel. Some of the work of Joseph Applegate on the German noun phrase will be presented in a later report at this conference. Some work has also been done on German grammar by visitors in our regular summer program for visiting scholars. This includes the work on the German adverb by James Gough of Georgia Institute of Technology, and work by Leonard Brandwood of England, Bjarne Ulvestad of Norway, and Stanley Werbow of the University of Texas.

Our work on French is being carried out by David Dinneen. He is writing a French sentence production routine in COMIT. With such a routine he will be able to study certain questions of French syntax with the help of the computer.

General research on the logical structure of language is being carried out by Elinor Charney. She has started from the work that she did with Hans Reichenbach on the analysis of conversational language, and particularly on the tense forms. The results promise to be an opening wedge into many interesting problems in semantics. She is being assisted to some extent by several other members of the group.

In addition to the basic research effort into language problems, considerable effort is being made to provide adequate tools for

research. At present these tools include two major sets of programs for the IBM 704 computer. The first of these is the COMIT system, a powerful programming aid which enables the linguist to do his own programming without the difficulties inherent in working through the intermediary of a professional programmer. The system will be described in a later talk. The other tool that is being provided is a method of handling large quantities of text that can be obtained from the publishing industry in the form of punched paper tape. This system of programs, which allows the computer to search through text for particular words or groups of words, is an invaluable aid to the linguist in his study of the structure of languages since it gives him ready access to his data.

The programming of the COMIT system is completed and the final check-out is in progress. We expect that it will be available for use soon. The programming has been done in a cooperative arrangement with the M. I. T. Computation Center.

When the COMIT system is finished, it will be made generally available. It is hoped that the availability of the system, will materially increase the productivity not only of our own group but of many others as well. We have already been using the COMIT notation extensively in mechanical translation research at M.I.T. even though programs cannot yet be run. We have used it to write down in an unambiguous fashion our ideas on translation. This has aided greatly in clarifying our own thoughts and in communicating them to each other. We have come to realize that without an adequate notational system, research becomes very difficult.

The other set of programs, for handling large quantities of text, has now been completed and is already in use. Texts currently available include 100, 000 words of American newspaper text and 100, 000 words of German newspaper text, both derived from punched paper tape obtained from the publisher. A third text, consisting of U. S. Patents, is being punched by the U. S. Patent Office in a cooperative arrangement whereby they are providing text which we can use and we are providing programs tailored to their text. The design of an appropriate transliteration scheme was carried out by Kenneth Knowlton of M. I. T. and Simon Newman and Rowena Swanson of the Patent Office. A description of the system is available in a Patent Office Report. K. C. Knowlton has written the required transliteration

programs. Again, text and programs are available to all legitimate users.

The field of automatic programming from problems stated in English has much in common with mechanical translation. Recognizing this, we are sponsoring some work in this field jointly with the M.I.T. Solid State and Molecular Theory group. This work is being done by Michael Barnett and John Carter.

Work is being done on methods of programming translating routines. William Cooper, now at Berkeley, worked out a method of argument compression for dictionaries. Anthony Phillips has worked out a method of splitting German noun compounds for dictionary lookup.

Recognition routines have engaged our attention for a number of years. You will hear later about the most recent one from Hugh Matthews. An earlier one by Matthews and Syrell Rogovin, one of our summer visitors, is being published in the current issue of Mechanical Translation. We are trying to explore as many of the possible ways of recognition routines as we can. The availability of COMIT makes it relatively easy to explore new types of programs.

A sentence-production routine has been devised that is quite simple and appears to have a lot of promise. It operates with a grammar expressed in terms of an unordered list of constituent-structure rules. Thus it shares with the recognition routine of Matthews the important property of having the program separated from the grammar of the language. The routine will work for any language for which we have an appropriate grammar. The sentence-production routine works from left to right, and from top to bottom in the constituent-structure tree of the sentence. The grammar that the program works with gives the constituents for every construction in the language. The program is capable of handling discontinuous constituents as well as continuous ones. In producing a sentence, the program must remember somehow what it is committed to do next by the rules of the language. Having expanded a symbol S for sentence into a subject and a predicate, it must remember that when it has finished expanding the subject, it is committed to expand a predicate of the appropriate type. And while it is expanding the article of the subject, it must also remember that it must later expand the noun.

One of the interesting things about this program is that it led

to the discovery of some unsuspected aspects of English structure. It appears that the English sentences that occur never require more than about seven items to be remembered for future expansion. This is startling because it had previously been thought that one could have clauses within clauses without limit in English. It turns out that one can have clauses within clauses without limit only if most of them are the right—hand constituents of the construction they are a part of. In this way the speaker is relieved of having to remember to complete an indefinite number of constructions. It appears that most of the complications of English syntax can be attributed to phenomena associated with this restriction imposed by a person's memory. Some of the phenomena of English syntax that appear to be thus explained include: the hierarchy of sentence, clause, noun phrase, adjective and adverb; the different behavior of subject and object clauses; the phrase structure of the active and the passive with the "by" phrase; the reversal of order of direct and indirect object; the shifting of the position of the separable verb particle; the function of the anticipatory "it"; the first position of the interrogative pronoun; the discontinuous nature of adjectival and adverbial phrases; the position of certain adverbs before the article; the fact that when the genitive marker follows its noun phrase, it is an affix " 's", and when it precedes it is a separate word "of"; and that derivational affixes are suffixes, and prepositions, articles, and conjunctions are separate words.

This work will be published soon in the Proceedings of the American Philosophical Society.

So you see that the mechanical translation research at M.I. T. is proceeding simultaneously on a number of fronts, and that some progress is being made toward a solution of the very difficult problems facing us in the development of mechanical translation to the point where mankind can count on it as a reliable means of bridging the language barriers.