[*Eighth Annual Round Table Meeting on Linguistics and Language Studies,*
Georgetown University, 1957]

Brief History Of Machine Translation Research

Leon Dostert
Georgetown University

Only five years ago the idea of using electronic computers to effect the translation of language seemed to many to be highly premature, if not actually fanciful. In the few short years that have elapsed since the first formal meeting on the subject in June of 1952 at the Massachusetts Institute of Technology, research in the field of machine translation has become widespread and has achieved academic and scientific respectability.

The story of the genesis of machine translation — the transference of meaning from one patterned set of signs occurring in a given culture into another set of patterned signs occurring in another related culture by means of an electronic computer — has been traced with care in the first compendium of essays on the subject entitled *Machine Translation of Languages,* edited by William Locke and A. Donald Booth.

From the outset it is understood that the type of machine to be used for translation is properly a logical machine. It must read the input text in the source language, manipulate the input translationally, and furnish a usable output in the target language. Since most digital computers operate with binary digits, the input operation has to include a transposition of the properly formulated source data into binary code, and the output operation includes the transposition of the binary result into more common symbols — decimal numerals or letters. The output can easily be equipped with a printer, resulting in legible printed text. The input of modern computers consists of previously prepared punched cards, punched tape, magnetic tape, or the like. For a translation program, this means that a human operator has to read the source text and, say, punch it on cards or tape, before it can be fed into the machine. In order to eliminate this preparatory human operation, the input of the machine would have to be equipped with an electronic scanning device capable of reading printed text and transposing it directly into binary code. Such a device has been the object of continued research for several years, and partial results have been attained.

Translation has to accomplish more than merely the one-by-one transfer of units from the source language into the target language. It must include some solution to the problems of choice implicit in the fact that: A) a unit in the source language may have more than one possible equivalent in the target language, and B) that the order of source language units in the input may not be suitable for the output in the target language. The machine manipulation of the text fundamentally involves two types of computer operations: table look-up and algorithmic (that is, properly computative). The table look-up operation consists in matching the machine-read input against a set of data stored in the memory of the machine, and in delivering these stored data to the arithmetic unit of the machine for algorithmic processing. The result of this processing is the translated output, which is then fed into the printer and delivered to the user.

The Georgetown-IBM experiment of January 1954, which succeeded for the first time in effecting machine translation from Russian into English on a limited basis gave, notwithstanding its critics, considerable impetus to research in the field. Admittedly, a number of the operations in the first trial were formulated on an *ad hoc* basis and the area of search and choice, as well as the extent of manipulation were strictly limited. However, as a result of this experiment, certain essential concepts were formulated which have largely remained valid. The fact that the experiment had attracted wide interest to the general problem was evidenced when in 1956 the Institute of Precision Mechanics and Computer Technology of the U.S.S.R. Academy of Sciences announced the successful performance of translation of English into Russian on their BESM computer, and acknowledged the relationship between their undertaking and the Georgetown-IBM experiment. This announcement was not unrelated to a renewal of interest and support for work in MT in the United States. In June of 1956 Georgetown University received a substantial grant from the National Science Foundation to undertake intensive research for the translation of Russian scientific materials into English. This grant has been renewed for a second year of continued research.

I shall give you a brief description of the work presently being conducted at the most important centers of organized machine translation research in the United States and abroad.

I have already mentioned the Academy of Sciences of the U.S.S.R., where a group of mathematicians and linguists have been working on various aspects of the problem, including the translation of limited segments of English scientific texts into Russian, and further, the investigation of machine translation of Chinese, French, and German into Russian. Two additional groups have been working on diversified techniques both in Moscow and Leningrad. The information available is not complete, but it appears that the fundamental technique of these groups involves what they call the analysis of the source language and a synthesis (necessitated obviously by the inflectional character of Russian structure) process in respect to the target language. Their approach seems to utilize English inflectional suffixes and word order as cues to bring about the appropriate inflected forms in Russian. A second point which seems to emerge from published material is that their plan is to make Russian a sort of “core” language, so that machine translation from, for example, Chinese into French, would ultimately be effected through the medium of Russian. The Soviet group has apparently been successful in testing its translation program on computing equipment.

The approach of the Cambridge Language Research Unit in England differs from that of other groups by its emphasis on mathematical logic. Its plan of research involves the transfer of grammatical patterns from source to target language by means of Boolean algebra. This ensures the identification of translation units and their proper manipulation as wholes for translation purposes. Another feature of their approach is the use of a thesaurus routine to achieve semantic transfer from one language to another. That is to say, the semantic ranges of adjacent translation units are matched against each other by using coincident definitions in a thesaurus.

In this country, the group at the Massachusetts Institute of Technology, so far as is known, is concentrating on an exhaustive linguistic analysis of German on the one hand as source language, and English on the other as target language. The assumption of the MIT group appears to be that an analysis as complete as possible of the structure of the languages involved must first be made before the specific problems of meaning transference are approached.

On the West Coast, the International Telemeter Corporation of Los Angeles is now engaged in the planning of a translation machine of limited scope, with primary emphasis on storage capacity, and without envisioning too much complex manipulation between input and output. In conjunction with the development of this project, a group of linguists at the University of Washington is preparing a translation program which appears to be specifically geared to the characteristics or limitations of the machine under construction. The objective of their research is to investigate whether or not such a deliberately circumscribed operation may not be adequate for certain practical purposes.

At U.C.L.A., Dr. Kenneth Harper has completed a preliminary project which goes considerably beyond Russian-English dictionary searching. Research is also presently being conducted at the following universities: Illinois, Michigan and Texas.

In October 1956 a meeting was held at the Massachusetts Institute of Technology, where American centers engaged in this work were represented, including groups from the University of Washington, UCLA, the International Telemeter Corporation, Harvard University, and Georgetown University, as well as MIT. Also present at this meeting were representatives of the Cambridge Language Research Unit in England. The U.S.S.R., though invited, was unable to send representatives to the meeting.

Next summer a seminar on machine translation will be held as part of the Linguistic Institute at the University of Michigan. It is supported by the National Science Foundation under the auspices of the American Council of Learned Societies.

Next August two reports on machine translation by members of the Georgetown project will be presented at the Eighth International Congress of Linguists at Oslo.

At Georgetown, after the 1954 experiment our research continued on a very limited scale until we received the grant from the National Science Foundation which enabled us in the fall of 1956 to engage in a full-scale project employing more than twenty senior and junior researchers. It was decided to focus on the translation of Russian texts in the field of organic chemistry.

The project is under the direction of a group of seven members of the regular faculty who represent varying types of linguistic specialization. Three are general linguists, two are Slavicists, one is an Arabist, and one is specialized in Germanic linguistics. Within the group competence in Romance and non-Indo-European linguistics is also found. While the focus is on the problem of translating Russian into English, the diversity of background of the group of linguists makes it possible for us to go ahead with the preliminary formulation of approaches to the problem of translation from and into other languages on the basis of the techniques developed for the Russian-English transfer.

The linguists work in close cooperation with nine research associates or assistants, all of them trained in linguistics, several of them competent in Slavics, others in other linguistic fields, and a few of them bilingual in Russian and English. Programming consultants have been working with the group so that the emerging linguistic solutions may remain within the limitations of programming requirements for machine operations. Consultations with members of the University's department of mathematics have been held.

Assisting the linguists and the associates is a group of three bilingual translation analysts. Their task is to develop, with the guidance of the linguists, a lexicon from an existing English translation of portions of the *Soviet Journal of General Chemistry.* The particular text that has been chosen is concerned with experiments in organic chemistry which have application to physical chemistry. The English-language version was found to be inadequate for machine translation purposes and two persons were assigned the task of preparing a standardized translation which would be free of stylistic idiosyncrasies and as consistent as possible. This is not to be confused with pre-editing, or simplifying the text to any kind of basic language. In this work, the analysis group has made use of the services of a graduate fellow in organic chemistry to assure the correct translation of technical terminology. It was found that the problem involved in producing a translation especially suited to the purposes of machine translation is largely one of consistency.

The staff is divided into working groups and assigned individual topics. From the outset our policy has been not only to permit, but to encourage diversity of approach.

Several members of the project are following and amplifying into increasingly broader formulae the partly empirical and partly analytical technique of the 1954 Georgetown-IBM experiment. As a first step, this group analyzed and reclassified the data of the first experiment. As a result of this work, it was decided to focus on one set of problems, namely, those involved in the translation of the Russian noun phrase. These problems were approached by the Experimental group in terms of the translation of the nominal and adjectival case suffixes. At the present time, the Experimental group is engaged in working out the code needed to present a considerably expanded and more advanced sample suitable for a second experiment. The code now in preparation is intended to cover precisely and in a generalized way the translation of the noun phrase, within exactly stated tractability limits. They intend to apply this code to a normal technical text in Russian, and translate the entire text by giving, in addition to the precise solutions described above, intuitive, or *ad hoc*, solutions for the problems not so far included in the research, with the two types of solutions clearly differentiated in the sample.

Individual researchers have been given specific analytical problems, such as the formulation of a program for the insertion of the article in the English translation from Russian. Rules have been formulated for the prediction of the occurrence of *the* or "zero" article in English and have been demonstrated to have about 80% accuracy on the basis of the corpus tested. Russian prepositions have been analyzed in order to find the determiners of their variant translations into English and to study their correlation with the English prepositional pattern. An analysis of Russian noun and verb endings has been completed with a program consisting of some 500 statements of the type required for programming formulation. Specifically, a list was made of noun-adjective and verb suffixes, arranged in such a way as to allow for easy and economical storage in a glossary. After the preparation of the list, work was started with the consultant programmer, on a method by which the results obtained in the preliminary stage might be utilized in machine translation. The grammatical identification of Russian items can be achieved by matching stem with suffix, thus making it possible for the machine itself to recognize the grammatical function of a given item.

This identification program is directly connected with research on the correlation of syntactic operations in Russian and English. The underlying assumption is that syntactic analysis of successively included constructions within the sentence is essential for any method of translation that is to be reducible to mechanical procedures. The syntax function approach amounts to instructing the machine to proceed on the basis of syntactic analysis. This is achieved by approaching the translation analysis in terms of syntactic hierarchies and their sensing equivalents for a logical machine. The sentence must be handled in terms of its constituents in successive inclusion, so that the composition and order of smaller constituents can be adequately translated.

One member of the project has undertaken the study of a French chemical text and developed a strictly empirical technique based on the concept that the exhaustive translational analysis of a series of sentences within a continuous text would yield a number of formulations adequate to constitute the basis for a general program.

Finally, a research associate, with the assistance of one of the linguists, has undertaken the investigation of a technique in which a series of code affixes based on traditional grammatical categorization is added to Russian words so that ambiguity is resolved within a segment of the source language through a matching technique. The appropriate target item is then elicited by a second matching of the code affixes corresponding to identical affixes occurring as part of the target lexical items.

In order to ensure coordination among the several groups and individuals pursuing specific assignments, the group of linguists meets as required and special coordination meetings are held as circumstances warrant.

In addition, the entire staff participates in a weekly two-hour general seminar which is open to the public. As a result of the examination of specific problems through discussion, review, and comments by the other members of the project, certain conclusions are formulated which serve as orientation for further research. Individual members present papers summing up their own work or progress, or that of a group with which they are working. These papers are reproduced and distributed to those interested or working in the field.

One of the difficulties which has handicapped progress in the field of machine translation has been the lack of communication on the one hand, and the assumption of somewhat rigid positions on the other. To help to remedy the first difficulty, we hope that the publication and distribution of our seminar work papers will be one step toward greater communication among the different research groups in this country and in England. The publication of the proceedings of this meeting next fall will be another step in seeking broader communication of results and techniques. In respect to the second problem, it is somewhat more difficult to suggest remedies. It seems to me that since we are still barely past the threshold of our investigation, it would be both premature and unscientific to cling narrowly to a given hypothesis or theory as to the most efficient manner in which the problem can be resolved. It can already be seen that the various techniques and approaches which have characterized the work of different individuals and groups are beginning to reach a point where results can be integrated. It is planned that as each technical approach reaches a certain level of formulation we shall conduct tests on existing computers which will give us a valid basis for determining which of the several approaches is the most efficient and on the basis of the results the possibility of a one-line approach will be considered.

On behalf of the faculty of The Institute and of my associates in MT research, I am happy to welcome you, and I hope that our exchange of ideas will prove of value to our common objective.