[*Mechanical Translation*, Vol.6, November 1961]

The Machine Translation Project at the University off Washington, Seattle, Washington: Outline of the Project

by Erwin Reifler, University of Washington

This report traces the history of MT research at the University of Washington which, as elsewhere in this country, was sparked by Dr. Warren Weaver's memorandum of July 15th, 1949. After preliminary grants from the Rockefeller Foundation in 1952 and 1953 for research in German-English MT, Dr. Reifler's research was sponsored by the Rome Air Development Center, Rome, N. Y., of the USAF and supported by several grants for Russian-English MT which at the time of the termination of the project in March of 1960 totaled $235,500.00.

The lexicographical, linguistic and engineering research of the project was carried out in consideration of the capabilities of the photoscopic translation system designed by Dr. Gilbert W. King and first developed by International Telemeter Corporation of Los Angeles. It is at present being improved by IBM under Dr. King's supervision.

The aim of the first phase of the project was an optimum of lexicography intended to solve as many bilingual linguistic problems as possible by lexicographical means alone. The second phase of the project was concerned with the elaboration of linguistic and engineering procedures aimed at the solution of the remaining bilingual problems. Details concerning the research and its results will be found in LINGUISTIC AND ENGINEERING STUDIES IN AUTOMATIC LANGUAGE TRANSLATION OF SCIENTIFIC RUSSIAN INTO ENGLISH, PHASE I (June, 1958) and PHASE II (March, 1960), University of Washington Press, Seattle.

1.0 History of the Project

Machine translation (MT) research at the University of Washington in Seattle began in November 1949, when I received Dr. Warren Weaver's now historic memorandum of July 15th of that year, in which he suggested the possibility of automatic translation by computer techniques. My research in 1952 and 1953 was supported by grants from the Rockefeller Foundation.

Since early in 1952 Professor W. Ryland Hill, Dr. Thos. M. Stout and Mr. Robert E. Wall, Jr., of our Electrical Engineering Department, have contributed their engineering knowledge to this research, which has since then been a joint enterprise of their Department and our Department of Far Eastern and Slavic Languages and Literature. The result of this teamwork was the designing and construction in 1954 of the First Pilot Model for German-English MT[1] and the gradual elaboration of a new MT terminology.[2] Our team was subsequently further strengthened by the addition of two other staff members, Dr. Lew R. Micklesen, Russian language expert of the Far Eastern Department, and Dr. David L. Johnson, computer expert of the Electrical Engineering Department.

In June 1956 the University of Washington was awarded a U. S. Air Force subcontract for the purpose of preparing the lexical contents of a photoscopic memory device and translation system designed by Dr. Gilbert W. King and then being developed for the Air Force by the International Telemeter Corporation of Los Angeles. This *Initial Project* was financed by a grant of $30,000 and was concerned with the linguistic investigation and analysis of 111 Russian texts from 40 fields of science. It supplied more than 14,000 Russian-English operational entries consisting of Russian “semantic units” belonging to the *technical* and *general-language* vocabulary occurring in these texts, of additional lexical units selected from high-frequency lists, and of their target equivalents.

The initial grant was subsequently increased to $108,500 to finance the *Expanded Project* which began in March 1957. Its object was the supplementation of all paradigmatic forms already in the initial store of 14,000 “semantic units”. It was then estimated that this would increase the store to approximately 200,000.*

In March 1958 the University of Washington was awarded a prime contract from the U.S. Air Force for the continuation of the project. Under this contract the University received an additional grant totaling $127,000 to cover the expenses of the project.

* The final count at the conclusion of the lexicographical phase of the project in June, 1959, showed that our estimate had been too generous. By that time, our sponsors had received from us a Russian-English MT lexicon of 170,563 entries on 556,141 IBM punch cards.

2.0 Theoretical Considerations

2.1 THE SCOPE

The modern structural linguist defines language as "... a system of arbitrary vocal symbols by which members of a social group cooperate and interact."[3] If Professor Edgar H. Sturtevant in this definition had omitted the word "vocal", this definition could be used for our purposes without any further change; for although machines will be built in the future also for the automatic-translation of *speech,* we are at present concerned only with the automatic translation of *printed* texts. The reasons for this present limitation are of both an engineering and a linguistic nature:

(a) Certain engineering problems have still to be solved before an acoustic device can be designed whose speech recognition is not limited to the speech of an individual speaker.

(b) *Published written texts* are more likely to conform closely to the rules of their language than spoken ones.

(c) The number of *homophones* is in many languages much larger than the number of their *homographs*. English examples are *to, too, two,* which have exactly the same pronunciation but whose written forms are highly distinctive. There are of course also many cases where the situation is reversed. Examples are English *the bow* and *to bow, the wind* and *to wind, the sow* and *to sow, the tear* and *to tear,* etc. etc. But by and large the written form presents the smaller mechanization problem. Moreover, as a result of research I conducted during the summer of 1953 under a grant from the Rockefeller Foundation, I can state that there is a relatively simple solution for these ambiguous cases. This solution is based on highly distinctive differences in the grammatical function of these words which affect their position in their clauses and sentences. These homographic words are thus clearly recognizable by their characteristic immediate environment.

Thus we are at present concerned only with printed texts. There is, however, a further limitation. In order to speed up progress on all levels of human civilization it is necessary that wasteful duplication of research be avoided as far as possible. It is, therefore, of the greatest importance that the scientists of all countries be continuously kept up-to-date on scientific developments elsewhere.

Now in countries like the U.S.A., Great Britain, the Soviet Union, Sweden, Italy, Germany, France and Japan, to name the most important ones, progress in all fields of human knowledge is proceeding at such a rate that there simply are not enough qualified *human* translators to translate into other important languages the ever-increasing avalanche of scientific publications.

The immediate purpose of creating translation machines is therefore to make up for this deficiency which already constitutes a serious problem. The primary purpose of these machines will be to supply the scientists of the world with reliable translations of *scientific publications* at a speed and price which cannot be obtained otherwise.

2.2 STORAGE PROBLEMS

Early MT pioneering was dominated by the fact that the mechanical and electronic equipment then available offered a storage capacity too small or too expensive to allow for the storing of all essential words or word groups of any set of two languages concerned in the translation process. Saving of storage space played therefore an important role in the thinking of the MT pioneers. Consequently, they thought of two measures which could result in a substantial saving of storage space.

One of these measures consisted in the separate storing of stems and endings. Thus the English pioneers Drs. A. D. Booth and R. H. Richens as early as 1948 experimented with the automatic dissection of words into their stems and endings to which separate positions could be assigned in the machine memory.[4] This procedure is still being followed in some MT schemes.

Another measure took into consideration the fact that many meaningful units of a language permit a multiplicity of translations with unintended meanings, if narrow or wider context is ignored. In the absence of logical procedures and equipment which would enable a MT system to pinpoint intended meaning and select the appropriate target equivalent in consideration of context, the system would in such cases supply multiple output alternatives. In any case, all possible target alternatives would have to be coded into the machine memory and thus much more storage space would be required. In order to cope with this problem—that is, in order to reduce the amount of the information to be stored—the suggestion was made *that an individual MT system be developed for each sub-branch of science,* to be based on "idio-glossaries" containing nothing but the bilingual vocabulary current in the particular field. Such a procedure, it was thought, would substantially reduce the number of possible target alternatives and result in an appreciable saving of storage space.[5] Also this procedure still has adherents among MT pioneers.

Subsequent phenomenal developments in MT engineering (i.e., the photoscopic memory system designed by Dr. Gilbert W. King) have in the meantime removed the limitations in storage capacity, at least for all practical purposes. It would be very *uneconomical* to aim at specialized translation machines for different branches or sub-branches of science because to do so would mean extensive duplication in lexicography and engineering and would thus be very wasteful in time and money.

It is, furthermore, *not necessary* to aim at such specialized translation machines because:

(a) the overwhelming majority of the *not specialized vocabulary* is shared by all branches and sub-branches of science;

(b) a large amount of the *technical vocabulary* is, both graphically and in meaning, either shared, or for all practical purposes shared, by each set of two languages belonging to the same civilization, and need, therefore, not to be coded into the machine memory. The problem of the *technical vocabulary* which is not shared both graphically and semantically can be solved by the following arrangement:

(c) a general purpose translation machine can easily be so rigged that it selects only those target alternatives which are characteristic of the branch or sub-branch of science concerned. Any inadequacies of this procedure are the same as those possible in the specialized translation machine and idio-glossary approach.

Consequently, we should from the very beginning concentrate rather on the creation of *general-purpose translation machines* for scientific publications.

I was also at first influenced by the limitations in the capabilities of the equipment initially available. During a research in the summer of 1952, which was supported by the first of the two Rockefeller grants mentioned above, I extended the procedure of Booth and Richens for the dissection of complex forms into their stems and endings *to a dissection of compound words.* Here some very interesting and highly intricate problems presented themselves which I shall outline further below. There is, however, an important difference between the dissection of complex forms and that of compounds. The first results in a *destruction* of information about the meaning of the form concerned which the automatic system has then to recover again somehow. This can be easily demonstrated with the very language material used for the Georgetown University—IBM MT experiment in New York, January 7, 1954.

On the other hand, my scheme for the dissection of compound forms has no such disadvantages. Although, as pointed out above, subsequent phenomenal developments in MT engineering have, at least for all practical purposes, removed all limitations in storage capacity, my solution for the automatic identification of the linguistic constituents of compounds will continue to play a role in future MT, namely, in the case of extemporized compounds (see 3.4).

2.3 ENGINEERING AND LINGUISTIC TIME TABLE

One more problem deserves to be mentioned here. Some hold that MT engineering design should not begin "until translation programming research has been advanced to the point where a detailed routine has been stabilized to such an extent that no radical changes due to further research can be anticipated, and engineers can begin to design circuits without having to fear a revocation of specifications once given."[6] Now digital computers probably contain units not at all or not very suitable for MT, and future translation machines may be based on designs not found in these computers. It is nevertheless the considered opinion of the MT engineers on the staff of the University of Washington project that it would be a mistake to thus delay MT engineering development. General algorithmic (translation logic or arithmetic) operations use selected *existent* units in normal computers. These units need therefore not be completely redesigned. As there is no reason to believe that any profoundly new designs will be required, there is no reason to wait until the total MT linguistic program is completed before preliminary engineering work can begin.

2.4 INTERMEDIATE MT SYSTEMS AS RESEARCH TOOLS

We are here, in the last analysis, faced with a fundamental decision. Should we first seek a purely academic, theoretical solution to all MT-linguistic problems and then turn over our results to the engineers for their translation into hardware? Or should we pursue a combined theoretical and empirical line of attack and push simultaneously our linguistic and engineering researches in close correlation? We believe that the second procedure will ultimately be more fruitful, speed up the perfection of MT, and prove to be more economical. We are here not concerned with the linguistic analysis of individual languages, but with the linguistic analysis of the translational agreements and divergences of at least two languages. These translational agreements and divergences can best be studied in a corpus of translated material in consideration of the source text and the conventional requirements of the target language. Therefore, we believe that we should from time to time build an intermediate translation machine on the basis of preliminary research results, feed large quantities of source text material into this machine and study the output. This will tell us at once to what extent we have succeeded and what still remains to be done. These intermediate machines could already be of practical use for purposes not requiring a highly finished translation product. They would certainly be very valuable research tools for the improvement of MT.

This has been the view of the University of Washington MT group from the very beginning. We feel that MT is best developed in terms of better and better pilot models. Such a procedure, which means the utilization of machines for the linguistic research still necessary, will shorten the time of this research tremendously. That is why we first designed the *University of Washington Pilot Model for German-English MT.* That is why our group was glad to have been asked to do the linguistic work for the photoscopic translation system designed by Dr. King.

3.0 The Linguistic Problems

3.1 THE MATERIAL

As I have already pointed out earlier, we are in MT not concerned with the linguistic analysis of individual languages, but with the linguistic analysis of the translational agreements and divergences of at least two languages. The peculiarity of our field forces us, moreover, to consider and keep in view *the totality* of a set of at least two languages as far as it concerns scientific publications, the total ascertainable vocabulary and the ascertainable totality of possible forms and constructions. We can actually never be satisfied with a so-called "representative sample", although we also start with representative samples. We shall be able to go very far in approaching this ideal of a consideration of the totality of relevant phenomena, because from the time the first machine becomes available, we shall be able to make use of machines to supply us with a quantity of additional research material at a terrific rate.

But we cannot even limit ourselves to the total *ascertainable* vocabulary. We even have to and can consider a certain type of *future* vocabulary, namely, the so-called *unpredictable* compounds which play an important role in certain languages, especially in scientific publications.

Furthermore, many structural linguists feel that their concern is the analysis of only the *formal* structure of language. The linguists working in MT have no quarrel with this opinion. But they feel that *for them* the analysis of the *formal* structure is only the prerequisite for the analysis of the *semantic* structure of language. It is true that we are only just beginning to approach this new horizon. But we are confident that we shall be able to make contributions in this field which will ultimately make possible the automatic determination of intended meaning and result in the creation of translation machines supplying not only accurately intelligible translations, but translations which are also in the conventional form required by the target language. We are confident because we have already solved a number of these semantic problems.

Since we are at present concerned only with the written language, the symbolization of language by vocal noises does not now play any role in our researches. Consequently, we do not now have to deal with *phonemes*. One should, however, expect that something like *graphemes* would in our thinking take the place of phonemes. But it appears that we, the MT linguists, will not even have to spend much time on the problems of differences in distinctiveness of single letters or groups of letters in different environments. This seems to turn out to be almost entirely a job for electronic engineers, who are in fact already working on electronic reading devices which will almost completely eliminate human cooperation on the input side of the translation system.

3.2 MORPHOLOGY AND SYNTAX

Grammatical ambiguity of non-distinctive paradigmatic forms of the source language constitute an important problem in MT research. It has, however, to be emphasized, that we can speak here of ambiguity only if we consider the words concerned in isolation. But if we consider them in their environment, then they are in most cases not ambiguous at all. Research aiming at an automatic solution of this problem has been carried on at the University of Washington for some time. Mr. Robert E. Wall, Jr., of our Electrical Engineering Department, whom I have mentioned earlier, together with a graduate student of his Department, worked on the elaboration of the so-called *tag system* for logical processing, which he is testing in experiments with an IBM computer at our University.[7] This research is based on ideas I outlined earlier in a published paper[8] and on linguistic research material (output predictions) becoming available in our project research. Another research aiming at a mathematical solution of these problems has been pursued by Mr. Aristotelis D. Stathacopoulos, another graduate student of our Electrical Engineering Department.[9]

Another problem of great importance in MT is that of disagreements in the word order of the two languages concerned in the translation process. Also here we have to consider environmental factors in both languages if we want to elaborate the linguistic prerequisites for an automatic reshuffling of the word order of the source language text into that required by the conventions of the target language. Much thought and research is being dedicated to this problem at the University of Washington as well as at all other MT research centers in this country and abroad. I think I can already say today with confidence that the mechanization of the resolution of the word order problem is only a question of time.

3.3 MEANING

As pointed out earlier, the MT linguist has to concern himself not only with the *formal structure* of language, but also with its *meaning aspects*. The aim of all translations is, after all, to determine the meaning intended by the original author coded in one system of symbols, and to trans-code it into another system of symbols. Meaning has to be considered not in terms of the semantic behaviour of *one* language, but in terms of what I call "source-target semantics". Moreover, the peculiarities of our field require that the problems of source-target morphology and syntax, and those of source-target semantics be not dealt with in isolation from one another. It is, in fact, very useful not to think here in terms of the contrasts of form and meaning, but rather in terms of something like a unified field theory: *We are always dealing with meaning* of which we distinguish two kinds, namely:

(a) grammatical meaning.
(b) non-grammatical meaning.

This enables us to do without the very bad term of "lexical meaning".

The problem of non-grammatical meaning we subdivide for practical MT purposes into several categories. In many sets of two languages we find a large number of meaningful units characterized by one-to-one correspondence of their single or multiple meanings. Many of these are shared by both languages, not only with regard to their meaning or meanings but even in their *graphic form*. These I call "diglots" as opposed to the "monoglots"—that is, meaningful units of the source language not shared by the target language.[10] Since these diglots present no problem at all in MT, we can ignore them. In the University of Washington MT project they are at present actually ignored to the extent that they are even omitted from our store of entries and will thus not be coded into the machine memory. The photoscopic translation system designed by Dr. Gilbert W. King of the IBM Research Center, New York, is arranged in such a way that all meaningful units of the Russian text which are not identified in the machine memory—that is, which are not found there—are transferred untranslated but in an easily readable phonetization in the Latin alphabet, and in red print, to the output. Thus all diglots of the source text will appear in the output text in a form the output reader can read and understand, although they do not occur in the machine memory.

Then there are a large number of meaningful units of the source text which have two or more alternative equivalents in the target language. Among these we have to distinguish between those which belong to the general language, to the specialized language and to both. I have already pointed out earlier that we shall have no difficulty with the purely technical vocabulary which has different meanings in different fields of science. I have said there (2.2c) that a general purpose translation machine can easily be so rigged that it selects only those target alternatives which are characteristic of the branch or sub-branch of science concerned. For the other two groups in this category we have to develop logical equipment and procedures for the automatic selection of the appropriate target alternative in consideration of the narrower and wider context. The photoscopic translation system for which our project elaborated the bilingual lexicography did not yet have such logical accessories. This system will therefore in these cases still supply multiple target alternatives. But in a cooperative research undertaken in the summer of 1956 by a number of linguistic experts and graduate students of our Department of Far Eastern and Slavic Languages and Literature the number of all possible English alternative equivalents of Russian semantic units was substantially reduced by a selection of only *those which are absolutely necessary for accurate intelligibility in all possible contexts.* Thus although the output of this translation system will still be cluttered up by multiple English alternatives and will show a word order which in many instances does not agree with the conventions of English, it will nevertheless already be *accurately intelligible.*

Another highly interesting category is one in which the intended non-grammatical meaning is simultaneously pinpointed by the automatic resolution of the grammatical ambiguity.[11] Since we know that the resolution of grammatical ambiguities can be mechanized (cf. 3.2), this category no longer constitutes a problem for MT.

The last category is represented by word sequences of the source language which are idiomatic in terms of the target language. Apart from these *genuine* idiomatic sequences, we have been able to treat many other word sequences which, linguistically speaking, are not source-target idioms, lexicographically *as if they were source-target idioms.* We call these "bilingual pseudo-idiomatic sequences". Such a procedure is of great advantage for MT because we have no difficulty whatsoever with bilingual idiomatic sequences. The photoscopic translation system will without logical equipment for linguistic purposes give these an *idiomatic translation no human translator can do better.*

The reason for this is the following. We distinguish two kinds of bilingual idiomatic sequences, namely the *non-paradigmatic* and *paradigmatic* ones. An example of the first is English *at any rate* which does not permit a word-for-word translation in any of the languages I know, except in those cases where it is followed by *of.* In this idiom only *rate* has other paradigmatic forms *(rate's, rates),* but only *rate* is permissible in the idiom. If it were changed to *rates,* the sequence would lose its idiomatic value and would permit a word-for-word translation in the case of many target languages.

An example of a paradigmatic idiom is *to hold one's tongue.* Here many paradigmatic changes are possible (*I hold my tongue, you held your tongue, she was holding her tongue, they will hold their tongues,* etc., etc.) without any change in the idiomatic status of the sequence.

Now since the photoscopic translation system has a practically unlimited storage capacity, we can code all idiomatic sequences *in toto* into its memory device, and in the case of the paradigmatic ones we can code into it all their meaningful paradigmatic forms, and beside them we code their *idiomatic* translation into the target language. This simple lexicographical arrangement permits us to obtain idiomatic machine translations for all source-target idiomatic sequences.

3.4 FORM CLASSES

I have said above that the peculiarities of our field require that the problems of source-target morphology and syntax, and those of source-target semantics be not dealt with in isolation from one another. The automatic determination of intended grammatical meaning is in fact a prerequisite for the automatic determination of intended non-grammatical meaning. Research which I also carried on in the summer of 1952 revealed that it is actually very easy to devise a scheme by which an automatic system could extract the relevant grammatical information from a source text without the necessity of human intervention. In order to achieve this it would be necessary only to arrange for a kind of *filtering procedure* in which *identification coincides with grammatical determination.* A basic problem here is that of the determination of the form class to which a meaningful input unit belongs. Here it was found necessary to formulate the concept of "operational form classes" as different from the traditional form classes. In MT we are not only interested in what meaningful units look like but also whether they pinpoint the grammatical and non-grammatical meaning of others or are pinpointed by them, and also what has to be done with them machine-operation-wise. This required *a change in the membership, the creation of new form classes,* and *the distinction of different groups of form classes.*[12] One of these new form classes is, for example, the distributional class of source-target idiomatic sequences. It consists of the two sub-classes of paradigmatic and non-paradigmatic sequences (cf. above).

Compound forms play an important role in this context. They are again divided into the two classes of pure and hybrid compounds. A complicating factor is here the phenomenon of what I call *the X-factor problem in compounds.* This factor would permit a multiplicity of dissections of a compound, theoretically possible in terms of predictable constituents occurring in the machine memory, unless linguistic considerations and special procedures based on them limit the dissections to the linguistically correct one. As a result of my research in the summer of 1952, I was able to demonstrate how a translation machine can be given the wherewithal to deal with *all unpredictable* future compounds composed of predictable constituents. I found that in any language only 30 types of compounds are possible *theoretically,* of which, however, only 10 types are *linguistically* possible. I found moreover that only three matching procedures and a maximum of four matching steps are necessary to deal effectively—that is, to machine-translate correctly————any of these ten types of compounds of any language in which they occur although none of these compounds occur in the machine memory.[13]

4.0 Application of Research Results in the Present MT Project

4.1 VAST STORAGE CAPACITY VERSUS LOGICAL LIMITATION

All these linguistic research results and procedures have been considered first in the design of the *University of Washington German-English MT Pilot Model* and subsequently in the USAF-financed Russian-English MT project. Some modifications were, however, necessary because of the special capabilities and limitations of the photoscopic translation system for which our project was carrying out the linguistic, lexicographical, and engineering research. As already indicated above, this translation system has a memory device with practically unlimited storage capacity (permanent storage of 30 x 10^6 bits). It also has an exceedingly low access time (random access time of the order of 0.05 seconds). But it does not yet have any logical equipment for linguistic purposes. Consequently, not all of the linguistic problems involved in MT are at present accessible to a mechanical solution. We have, however, already been working for some time on the elaboration of logical procedures which will allow the development of accessory equipment for such linguistic purposes.[14]

We decided, however, to make full use of the vast storage capacity and to achieve an automatic solution of as many of our linguistic problems as possible *through an optimum of lexicography.* The vast storage capacity permits us to treat even a whole string of words and each of its paradigmatic variations together with its target equivalents as individual entries. The result is an idiomatic translation by automatic means of all bilingual idioms. Furthermore, it permits us to treat punctuation marks and even the graphically very distinctive space between words as letters of an extended alphabet and as part of a "semantic unit". This extension of the concepts of alphabet and word provides additional graphic and semantic distinctiveness which greatly improves the translation product.

On this basis a program for machine translation has been devised which:

(a) provides for the translation of words and word sequences,
(b) permits the automatic dissection of complex forms and the identification and translation of prefixes,
(c) permits the automatic dissection of compounds and a translation even of unpredictable compounds which is "accurately intelligible".

Each unit of input is compared serially with the entries of the store to find the longest possible memory equivalent that matches an initial portion. This is accomplished by a logical ordering of the store which places any memory equivalent that is an initial portion of a longer one behind the longer one. Each entry consists of the memory equivalent of a "semantic unit" of the source language, its target equivalent or equivalents, the control symbols for operating the machine, and the editing symbols intended to help the reader of the output text. Once logical accessories have been developed and added to the translation system, the editing symbols will be replaced by logical tags which will be processed in a computing device to edit the information extracted from the memory. The result will be a better translation product.

4.2 THE MT PRODUCT

As already pointed out above, with the present translation system we shall not yet be able to reduce all grammatical and non-grammatical ambiguities to the grammatical and non-grammatical meaning intended by the Russian author. The English output text will be cluttered up with "strings" of grammatical or non-grammatical alternatives from which the English reader will have to make his choice in consideration of context. In many cases we have, however, been able to reduce the number of these alternatives to such a degree that the output reader will not find it too difficult and time-consuming to arrive at the correct choice. This is done by making full use of the large storage capacity of the photoscopic memory and by the editorial symbols mentioned above. The word order of the English output text will still mostly show the Russian word order. In many cases this will not matter at all because of agreements in the word order of both languages or because the difference does not at all impede the *accurate intelligibility* of the output text. There are, however, cases where this difference does play a role and constitutes a serious obstacle to an accurate and quick understanding. In some of these cases we are able to alleviate the difficulty by changes in the form of the operational entry or entries concerned. The output form of all source-target idiomatic sequences will, of course, have the correct English word order. But in other cases the source-target linguistic problems can only be resolved by the addition to the machine of logical equipment. A detailed discussion and exemplification of these problems can be found in PROCEDURAL REPORT by Dr. Micklesen and in papers of the engineering members of the project staff.

The output of such a translation system already deserves the name of a real translation, because by and large it will already be *accurately intelligible,* although the output reader will not always get this accurate intelligibility at first glance but in a number of cases will get the meaning intended by the original author by what we may call "an understanding on second thought". The loss of time involved in this "delayed action" understanding will, I believe, be more than compensated by the speed with which large quantities of foreign scientific material will become accessible to scientists not familiar with the foreign language concerned. The output of such a translation system is comparable to that of a foreigner who has completely memorized the contents of a comprehensive bilingual dictionary, translates with superhuman speed, only knows the word order of his own language and completely lacks the intelligence necessary to decide himself which of the many possible target alternatives he remembers is appropriate in any particular context. Therefore he supplies every one of them. This is because the translation system in question lacks, as pointed out frequently, certain additional logical equipment that would make those decisions possible.

4.3 LEXICOGRAPHY

This description outlines the framework within which we had to try to achieve an optimum of lexicographical work. Our first problem was the collection of a representative vocabulary. The time and money available for our project excluded the possibility of the lexicographical preparation of a complete bilingual word list or, as has been suggested, the consideration of the contents of a small bilingual dictionary. We had neither the time nor the encyclopedic knowledge necessary to select from a bilingual dictionary those entries which, on account of the high frequency of occurrence in *scientific* texts of the Russian words concerned, would be of greatest practical value. How could we know which Russian words had this high frequency in scientific texts? The most sensible procedure was to select a representative sample of Russian scientific literature, extract from it the running general language vocabulary and the comparatively limited number of technical terms which happened to occur there, and to add to them general language words from published high frequency lists which did not happen to occur in the selected texts. Only in this way could we collect a really representative "operational" vocabulary. This is exactly what we did. By March 15, 1957, we supplied approximately 14,000 entries collected in this manner.

Another very difficult problem was that presented by the multiple English alternatives of the majority of Russian semantic units occurring in the selected texts and taken from the high frequency lists. I knew, however, from my experience with a number of languages, that, if mere "accurate intelligibility" was wanted, one to three alternatives could very often represent all of them in all possible contexts. Thus we were faced with the task of making a wise selection of such representative alternatives. The bulk of this task was accomplished during the summer of 1956 during which we availed ourselves of the combined knowledge of a number of native speakers of Russian, both faculty members and graduate students of our Department of Far Eastern and Slavic Languages and Literature. The result of their work was then checked and corrected in the subsequent months and was being continuously revised in the light of the lessons taught by the simulated machine translations we prepared.

4.4 SIMULATED MACHINE TRANSLATIONS

This brings us to the subject of simulated machine translations—that is, predictions of what kind of translations we expected from the automatic system for which we were preparing the lexicography. The uninitiated may think that the elaboration of such predictions is a waste of time since the machine, when completed, can supply such samples of its output at a terrific rate. And yet, these predictions were of the greatest value for us because they told us to what extent our lexicographical work was successful and where it failed, and on the basis of this information we were able to improve our lexicography before it was incorporated in the memory device of the machine.

These predictions are important in yet another respect. They clearly show up the cases in which changes in the *lexicography alone* are not enough to remove difficulties from the output text. They furnish numerous examples of those problems for the solution of which additional logical equipment is imperative. Without these predictions it would be a time-consuming task, indeed, to find all those problems and representative samples for them. With these predictions the task is a relatively simple one. The engineering members of our research group, in cooperation with the linguistic members, concentrated on those cases in which lexicographical changes do not solve the problem and, as mentioned earlier, began developing logical machine programs, which will lead to the designing of additional logical equipment for these purposes.

4.5 SUPPLEMENTATION OF PARADIGMATIC FORMS

If we had limited ourselves to the word list extracted from the selected Russian texts and the high frequency lists, the machine for which we were working would have been able to supply "accurately intelligible" outputs only for these selected texts. In this word list only some of all possible paradigmatic forms of every semantic unit occurred. In order to extend the scope to *all* Russian scientific texts it was imperative to supplement all the relevant paradigmatic forms which did not happen to occur in the selected texts and in the high frequency lists. This was the aim of our expanded program. Under this program the initial approximately 14,000 entries were increased to more than 170,000 entries.

5.0 Further Output Improvements by Mechanical Means

All output improvements achieved so far were obtained by lexicographical means. But, as pointed out earlier, our linguistic and engineering staff was already thinking and working in terms of a better machine—that is, a machine which will have the additional logical equipment necessary for achieving an output that will agree more closely with the conventional requirements of the target language. We distinguish three phases:

> *Phase 1.* The development of logical programs for the automatic pinpointing of intended *grammatical* meaning, resulting in the removal from the output of all superfluous grammatical clutter. *Phase 2.* The development of logical programs for the automatic reordering of the word order of the Russian language into that required by the conventions of the English language. *Phase 3.* The development of logical programs for the automatic pinpointing of intended *non-grammatical* meaning, resulting in the removal from the output of as many superfluous non-grammatical alternatives as possible.

Of these three the first one should have precedence because, firstly, the so-called superfluous grammatical clutter is due to bilingual phenomena of the highest frequency; secondly, because it is the easiest of the three to deal with; and thirdly, because the successful conclusion of the first phase is a prerequisite for work on the two other phases.

Work on these three phases of machine translation development requires an exhaustive investigation of Russian context in consideration of the coinciding and diverging requirements of the English language, the extraction and formulation of operational rules of the morphology and syntax of both languages, and the transformation of these rules into logical operations.

6.0 Conclusion

Even without these improvements brought about by logical procedures and equipment, the output as predicted by our simulated machine translations may already have valuable applications. It is, however, important to stress that the staff members of the University of Washington MT Project were themselves by no means satisfied with it. But this is the optimum attainable under the conditions dictated by the prevailing specifications of Dr. King's automatic system and the well-known differences between the Russian and the English language, and in view of the limitations set by available funds and time.

On the other hand, the preliminary results of our research in logical procedures were very promising. They make us feel confident that we shall ultimately be able to achieve an MT output satisfactory not only from the point of view of costs, efficiency, and "accurate intelligibility", but also from that of *readability.* We consider therefore our present results only as an intermediate step on the way to the ultimate goal, in fact, as a very important prerequisite for the attainment of that goal.

References

1. *Demonstration Machine,* Mechanical Translation, Vol. II, No. 2, November 19,55, p. 27; Geoff Douthwaite, *The UW Automatic Language Translator,* Washington Engineer, February 1956.
2. Erwin Reifler, *Some New MT Terms;* MT, Vol. 4, No. 3, Dec. 1957.
3. E. H. Sturtevant, *An Introduction to Linguistic Science,* 1947, p. 2.
4. R. H. Richens and A. D. Booth, *Some Methods of Mechanized Translation,* Machine Translation of Languages, 1955.
5. Victor A. Oswald, Jr., and Richard H. Lawson, *An Idioglossary for Mechanical Translation,* Modern Language Forum, Vol. 38, No. 3-4, September-December 1953.
6. Paul L. Garvin, *Machine Translation,* Reports for the Eighth International Congress of Linguists, Oslo, 1957, Vol. I, p. 106.
7. Robert E. Wall, Jr., and Udo K. Niehaus, *Russian to English Machine Translation With Simple Logical Processing,* AIEE, Paper No. 57-1062, August 1957.
8. Erwin Reifler, *The Mechanical Determination of Meaning,* pp. 152-154, in Machine Translation of Languages, 1955.
9. Aristotelis D. Stathacopoulos, *Noun Phrase Analysis Using The IBM 650 Computer,* The TREND in Engineering at the University of Washington, April 1957.
10. Erwin Reifler, *op. cit.,* pp. 147 and 148.
11. Erwin Reifler, *Studies in Mechanical Translation,* April 1951 (mimeographed), #80, (pp. 40-42).
12. Erwin Reifler, *The Mechanical Determination of Meaning,* pp. 154-159, in Machine Translation of Languages, 1955.
    Lew R. Mickelsen, *Form Classes: Structural Linguistics and Mechanical Translation,* in For Roman Jakobson, 1956.
13. Erwin Reifler, *op. cit.,* pp. 144-148, and *Mechanical Determination of the Constituents of German Substantive Compounds,* Mechanical Translation, Vol. II, No. 1, July 1955.
14. Robert E. Wall, Jr., and Udo K. Niehaus, *loc. cit.*