MECHANICAL TRANSLATION

DEVOTED TO THE TRANSLATION OF LANGUAGES WITH THE AID OF MACHINES

VOLUME ONE, NUMBER TWO August 1954

NEWS

MECHANICAL TRANSLATION COMES OF AGE

What is believed to be the first thesis ever written in the field of mechanical translation was presented in May to the Harvard University Division of Applied Science by Anthony G. Oettinger. The thesis, which earned Dr. Oettinger the degree of Ph.D. in Applied Mathematics, was entitled "A Study for the Design of an Automatic Dictionary." An abstract of the thesis is to be found on page 35 of this issue.

BOOK

W. N. Locke of M.I.T. and A. D. Booth of Birkbeck College, London, are co-editing a book of up to date essays on mechanical translation. There will be some dozen chapters written by nearly all of the active workers in the field. It is hoped that the book, to be published jointly by the Technology Press of M.I.T. and John Wiley & Sons, will be ready in the fall.

WORD COUNT BIBLIOGRAPHY

William Bull, Department of Spanish, University of California at Los Angeles, writes that over the years he has collected some 200 titles dealing with word counts. This may be the most complete list in existence. He wonders if there would be sufficient interest for him to go to the trouble of completing it and publishing it in perhaps a year or two. The completed bibliography would probably run to some 50 pages. We wonder if our readers would be sufficiently interested to see this in a future issue of MT.

ERWIN REIFLER, University of Washington, Seattle, Washington, read a paper on "MT, its psychological aspects and its significance for the human society" before the SECOND CONFERENCE ON GENERAL SEMANTICS in St. Louis June 12.

PETER SHERIDAN, Scientific Computing Service, International Business Machines Corporation, 590 Madison Avenue, New York 22, N.Y. presented a detailed account of the computer programming for the IBM - Georgetown University demonstration at the June 24 session of the Association for Computing Machines conference at Ann Arbor, Michigan.

LEON E. DOSTERT, Director, Institute of Languages and Linguistics, Georgetown University, Washington 6, D. C., spoke July 16 at the Summer Speech Institute, University of Michigan, on the subject of "Technical Aids in Interlingual Communication." The major part of his paper was devoted to mechanical translation.

THE MACHINE AND THE MAN*

Victor H. Yngve
Research Laboratory of Electronics
Massachusetts Institute of Technology
Cambridge, Massachusetts

WHEN extensive mechanical translation becomes a reality, many new jobs will be created. Some of these jobs will be closely related to existing occupations. In this category are those occupations connected with the construction of the machines — electronic design and construction, machine shop work, and the like. Then there will be others involved in the daily running of the machine — typists, operators, office workers, and administrative personnel. In addition to these rather obvious occupations, there are some that may be less obvious. In the following article we shall discuss several of the less obvious roles that humans may play in relation to a translating machine.

Man in the role of creator of the machine, the designer of the system by which it translates, was one of the earliest concepts to be found in MT literature. This idea is implicit in practically all of the work that has been done on mechanical translation. The machines that have been considered are slave machines, built by man and tirelessly carrying out to the letter the instructions originally given them. The burden that this throws upon man is the task of designing the machine and instructing it in detail in the routine it is to use to translate everything fed into it.

Perhaps Y. Bar-Hillel has given the most detailed statement of the tacit assumption that underlies the thought of many others when he writes of the necessity for man to provide "an operational syntax" for the machine. By this he means a program that the machine can carry out in sequence, at each point being given the exact criteria for determining what to do next. This program is to be capable of translating all possible sentences from the input language to the output language. Furthermore Dr. Bar-Hillel has outlined the things that he considers necessary for man to do before the machine can get to work. He envisions the compilation of a complete word index giving the stem-ending analysis; a complete dictionary giving for each word the various meanings and all the other information that will be needed for the grammatical analysis; and an operational syntax "giving a complete sequential program for the analysis of every sentence." The construction of this program constitutes a great challenge to the linguist, since it requires him to consider language as it actually is and to specify exactly and completely all the operations necessary for translation. As has been pointed out, the machine will be in the position of a person trying to translate from language A to language B, using a set of rules expressed in a third language and never knowing the meaning of what is being translated. The challenge to the linguist and to man as the creator and designer of the machine is to provide this set of rules.

Another widely held assumption is that a machine may never be able to produce a perfect translation. For this reason, a good deal of thought has gone into the possibility of man-machine combinations. One of the great difficulties that man as the creator of the machine will have to face is the fact that the input language does not have sufficient semantic explicitness in many cases to provide a machine with enough information to solve the many problems in grammar, syntax, and multiple meanings. Prof. Erwin Reifler pursued this problem and suggested a number of ways in which a human pre-editor could make the input text more explicit. The job description of the pre-editor is to be found in Reifler's first paper, abstracted in the last issue. "Whatever the native reader has to do by way of interpretation in the case of non-distinctive features of the FL (foreign or input language) text, can at least at the present stage of computer development, not be mechanized. Therefore, all that an FL text leaves to the FL reader to determine concerning lexical meaning, connotations, grammatical meaning, and word order, has to be added to the FL text before it is fed into the computer. And it has to be added in a form that the computer can 'digest'."

Perhaps his most far-reaching suggestion, as far as its possible impact on man, was his universal MT orthography. He proposed that the pre-editor capitalize the first letter of nouns, as in German, the second letter of verbs, the third letter of attributive adjectives, and so on. Reifler further proposed that this orthography could become universal and be applied to all languages that are written in scripts that allow capitalization. Thus the machine would have at the input a specification of the grammatical categories of the words to assist it in making a proper translation. This orthography would be taught in the schools. Here we have MT changing our conventional script, and thus affecting nearly everyone by

* This work was supported in part by the Signal Corps; the Office of Scientific Research, Air Research and Development Command; and the Office of Naval Research

requiring a change in the conventional method of writing. This concept of changing the input language to fit the needs of the machine is carried to the extreme by Stuart Dodd, who proposed that English and other languages be regularized along the lines of his proposed "Model English." Writers of material to be translated would be required to write according to the rules of Model English. The output of the machine could also be in a "modelized" language.

It seems to be a fair statement, however, that the idea of the pre-editor, and all other tampering with the input text or language, is nearly dead. Most workers now seem to consider that probably all of the tasks formerly assigned to the pre-editor can be mechanized. Perhaps the greatest stimulus to this thinking came from the work of Oswald and Fletcher, who proposed routines by which a machine could recognize blocks of words of a German text, and by which "the fluid German word order is resolved into a rigid English sequence." This suggestion, together with the suggestion of Booth and of Oswald and Lawson of strictly limiting the dictionary of the machine to those words and meanings required to translate in a particular field, brain surgery, for example, was supposed to eliminate the pre-editor for all but a very few routine problems, such as the splitting of long German compounds into their component parts. Even this problem seems amenable to solution by methods suggested by Reifler.

With or without the pre-editor, the output of the translating machine may still be no literary masterpiece. But it may be satisfactory for some purposes. For example, it might be adequate for the use of the scientist in keeping up with the foreign literature in his field. Much of the problem of keeping up with the literature is concerned with looking over articles in a rather cursory manner and deciding which ones merit more careful attention. For every important article, there are usually many that are unimportant for that particular person. If the scientist or engineer can scan and discard 100 documents by seeing only a rough translation made by a machine, and can select the one in which he is particularly interested, this one can be translated for him carefully by an expert human translator. If imperfect mechanical translations are given a fairly wide circulation to people who are interested in following the literature in a given field, the demand for translations of good quality, made by standard methods, will increase greatly. Thus the wide use of imperfect but useful mechanical translations may actually increase the demand for human translators.

The output of the machine itself, of course, could be made the basis for the more careful job of translation. This leads us to the concept of the post-editor, which has also been discussed in detail in the MT literature, particularly by Reifler.

A post-editor is a person skilled in the output language but who may be entirely ignorant of the input language. His task is to take the imperfect output from the machine and edit it into a polished or at least easily comprehensible document. This puts man in the role of partner with the machine. Or, as some would have it, the machine helps him produce the output text by doing much of the routine work that he would otherwise have to do to produce an acceptable translation. Although man has been reduced to a link in the chain, he does not have to solve the large number of routine problems, but can concentrate on the real difficulties. It has been shown that the post-editor is better able to do his job if he also knows the input language; thus we have the bilingual post-editor. It has also been shown that the post-editor is better able to do his job if he is an expert in the particular field of knowledge. If a mathematics text is being translated, the post-editor should be an expert in mathematics. Various authors have specified different ideal qualifications for the post-editor. It seems obvious that the amount of work done by a post-editor depends upon the ultimate purpose for which the translation is being made. If the purpose is to provide a translation in a literary style that could be published in a journal, possibly with large circulation, the post-editor might have a big job. If the purpose is to provide a rough copy that can be used by experts to determine whether or not the material is of interest to them, the post-editor would have a smaller job, or might not be needed at all. His utility depends upon how perfect a translation the machine makes and how perfect a translation is desired.

If the output of translating machines is imperfect, but adequate for screening purposes, the ultimate user or reader of the translation can be regarded as his own post-editor. He may be strongly motivated to acquire the skills necessary to do his own post-editing as he is now motivated to learn several languages so that he can keep up with the literature in his field. Thus there may be a considerable change in language teaching in the schools, with more emphasis on the skills of post-editing and less emphasis on reading ability of foreign scientific material.

Let us at this point dispose of the post-editor by saying that a machine can probably be constructed which will give a translation that is sufficiently accurate for any purpose that we happen to have in mind, if we don't have in mind a translation which reflects accurately the literary quality of the original. We now inquire what is the relation of man to the machine under these circumstances.

We still have man as designer and creator of the machine; but let us not be so demanding as to say that he must create the machine and the translation system in its final form before the switch is thrown and the machine starts carrying out its built-in destiny. Let us

suppose that man as the creator does not do as good a job as this, but first designs and builds a machine that can translate some things, but not all things. To be specific: the machine may have only a limited vocabulary; it may be able to handle only a limited number of grammatical or syntactic problems. Man in this new role, which we might call monitor and program adjuster, watches the machine translate, checks the output, notes its shortcomings, and alters the design or the program or the contents of the memory of the machine in such a way that the machine gradually builds a larger vocabulary, gradually becomes more proficient. Such a man may actually post-edit, but if the output is already satisfactory he will not have to do this. His duty is to instruct the machine, taking his cues from the machine's shortcomings as revealed by its output. We might say that the man is providing feedback of the type required for learning and that he is altering the machine in such a way that it behaves as if it were learning by its mistakes.

There are some who believe that this learning loop can be closed inside the machine, that the machine can be programmed to learn by its own mistakes with no human intervention other than the original design and construction of the machine . Perhaps experiments with the more deterministic type of machine will help show how to realize a learning type of translating machine at some time in the more distant future.

We have briefly discussed some of the ways in which man and machine may be related in the future mechanical translation industry. Beside these more or less obvious connections, the easy availability of mechanical translations of the most important foreign scientific and cultural writings is bound to have a great effect upon international communication and understanding; on our own culture, science and technology; and thus on nearly all of the occupations of man.

THE FIRST CONFERENCE ON MECHANICAL TRANSLATION

Erwin Reifler
Department of Far Eastern and Slavic Languages and Literature
University of Washington, Seattle, Wash.

THE FOLLOWING is a report on the proceedings of the first MT Conference, held at the Massachusetts Institute of Technology, Cambridge, Mass., June 17-20, 1952, and my own reactions.[1]

At the Conference individuals working on MT in this country and in England met for the first time and presented their different approaches. A detailed list of participants appears on the next page. The important point is that at this Conference linguists and electronic engineers joined for the first time to survey the linguistic and engineering problems presented by MT. At the end of the Conference it was the general impression of the participants that, for certain types of source material, a mechanization of the translation process is now a distinct possibility. Thus Dr. Warren Weaver's ideas about the possibility of MT in our time ceased to be a dream and moved into the realm of reality.

As a matter of fact, the engineers envisaged the creation of pilot machines within the next few years; that is, machines with limited storage for the translation of a limited quantity of scientific material from a foreign language into intelligible English, built for the purpose of convincing the general public and, especially, foundations and other organizations able to support new ventures, of the feasibility of MT, in order to obtain the funds necessary for further research and improvements.

The Conference was ably organized by Dr. Y. Bar-Hillel of the Research Laboratory of Electronics at M.I.T. Half a year earlier Dr. Bar-Hillel had visited the different groups working on MT in this country and published an excellent REPORT ON THE PRESENT STATE OF RESEARCH ON MECHANICAL TRANSLATION.[2] There can be no doubt that much of the success of the Conference was due to Dr. Bar-Hillel's efforts, and it is, I believe, no overstatement to say that MT, if and when it materializes, will be very much indebted to him.

The Conference decided that the papers of the participants should be published together with the discussions.[3]

Automatic Dictionary

Of greatest interest to the Conference was Dr. Booth's report on the translation experiments he and Dr. R, H. Richens had programmed on a computer in London. Dr. Warren Weaver had previously, in his first memorandum on MT (July 15, 1949), referred to their work. According to him "their interest was, at least at that time, confined to the problem of the mechanization of a dictionary which in a reasonably efficient way would handle all forms of all words." In a longer paper, SOME METHODS OF MECHANIZED TRANSLATION, which Dr. Booth submitted to the Conference he and Dr. Richens explain their approach. The translation they envisage is a word-for-word translation maintaining the word order of the input text and, in the case of multiple meanings, supplying alternative English equivalents. The machine determines by itself the stems and endings of the words of the input text and compares them with the entries in its separate stem and ending memories. These furnish not only the (often multiple) English equivalents for the input words, but also the (sometimes multiple) grammatical meanings involved. The latter are indicated in the output of the machine by abbreviations of the terms for the grammatical meaning concerned. At present only scientific material is considered for MT. Idioglossaries are used for the various fields, which means a considerable decrease in the number of possible meanings of each technical

1 This report was written in July, 1952. Opinions and facts are of that date.

2 AMERICAN DOCUMENTATION, 2:229 - 237, 1951.

3 Lack of sufficient funds has prevented the carrying out of this plan. However, a publisher has now been found for a volume of up-to-date essays reflecting present thinking on MT. This volume is scheduled to be published in the fall of 1954 jointly by the Technology Press of M.I.T, and John Wiley & Sons. It is being edited by A. D. Booth and W. N. Locke.

Participants in the Conference on Mechanical Translation

Dr. A. D. Booth, Director, Electronic Computer Section, Birkbeck College, London

Prof. William E. Bull, Department of Spanish, University of California, Los Angeles

Prof. Stuart C. Dodd, Director, Washington Public Opinion Laboratory, University of Washington, Seattle

Prof. Leon Dostert, Director, Institute of Languages and Linguistics, Georgetown University, Washington, D. C.

Dr. Olaf Helmer, Director of Research, Math, Division, Rand Corporation, Santa Monica, Calif.

Dr. Harry D. Huskey, Assistant Director, National Bureau of Standards, Institute for Numerical Analysis, University of California, Los Angeles

Mr. Duncan Harkin, Department of Defense, Washington, D. C.

Prof. Victor A, Oswald, Department of Germanic Languages, University of California, Los Angeles

Prof. Erwin Reifler, Far Eastern and Russian Institute, University of Washington, Seattle

Mr. Victor H. Yngve, University of Chicago, Chicago

Dr. Yehoshua Bar-Hillel, Research Associate, Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge

Mr. Jay W. Forrester, Director of Digital Computer Laboratory, Massachusetts Institute of Technology, Cambridge

Prof. William N. Locke, Department of Modern Languages, Massachusetts Institute of Technology Cambridge

Mr, James W. Perry, Research Associate, Center of International Studies, Massachusetts Institute of Technology, Cambridge

Dr. Vernon Tate, Director of Libraries, Massachusetts Institute of Technology, Cambridge

Dr. Jerome B. Wiesner, Director, Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge

Mr. A. Craig Reynolds, Jr., Endicott Laboratories, I.B.M., Endicott, N. Y.

Mr. Dudley A. Buck, Research Assistant, Electrical Engineering Department, Massachusetts Institute of Technology, Cambridge

term and an appreciable reduction both in the amount of storage required and in the access time. A number of sample products of this machine show the degree of intelligibility of the mechanical translation product and demonstrate how much this solution of MT leaves to the interpretation of a post-editor. There can be no doubt as to the value of Richens' and Booth's approach. It is, however, as they themselves are, I believe, very ready to admit, still far from the ideal of MT which I would define as follows: A complete mechanization of the translation process - that is, a mechanical system which, without the intervention of either a pre- or post-editor, outputs translations satisfactory with regard to both semantic accuracy and intelligibility. [4]

Some of the participating linguists indicated in private conversations that the samples of automatic dictionary output were unintelligible to them. My own impression is that the time required for the interpretation of the meaning of the output of this machine will be a serious factor in the evaluation of its practicality. This time has to be added to the time required by the machine itself for its operations. People who know classical Chinese will, for obvious reasons, have less difficulty than others with the interpretation of the products of this machine.

"Word-by-Word" or "Block-by-Block" Translations

Other very valuable contributions were made by Professor Victor A. Oswald, Jr., of the UCLA who, together with Stuart L. Fletcher, Jr., had previously published PROPOSALS FOR THE MECHANICAL RESOLUTION OF GERMAN SYNTAX PATTERNS.[5] In his conference paper WORD-BY-WORD TRANSLATIONS Dr. Oswald exemplified the inadequacies of such translation, even going so far as to assert that such a "translation is literally impossible." He suggested instead "block-by-block transverbalizatlon, in which process, problems of syntactic ambiguity are solved by the connection of syntactic segments with each other, and the fluid German word order is resolved into a rigid English sequence." This he had previously demonstrated in the PROPOSALS, "...and," he added, "we now know that a recognition of syntactic connection can be built into the 'memory' of machines of the high speed computer type."

Idio-Glossaries

Another important suggestion made in his paper and elaborated in a second paper entitled MICROSEMANTICS is his "micro-glossaries - glossaries which will reduce the range of choice of meaning from a bewildering multiplicity to a matter of - at the most - two or three." It has to be emphasized here that on every page of almost every scientific text scientific terms are rare islands in an ocean of general language. Consequently his scheme envisages "micro-glossaries" for the non-technical vocabulary of a whole domain of a particular science. This may reduce the number of non-grammatical meaning alternatives of the general language portions of scientific material in a number of cases. In the majority of cases, however, the non-grammatical incident meaning i.e., the particular meaning of the word in a given context, of these portions of the vocabulary is by no means determined or generally definable by the branch of science to which the material belongs, but has to be inferred from the meaning of co-occurrences of the narrow context. Therefore, although "micro-glossaries" (for which I suggested the obviously better term "idio-glossaries" - it is also preferable to speak of "idiosemantics" rather than of "micro-semantics") will certainly play a significant role in the ultimate solution of MT, in the case of scientific source material we are still faced with all the problems of multiple non-grammatical meaning presented by general language. Micro-glossaries "could," as Professor Oswald says, "serve to replace a team of specialists (on the post-editor side) in our proposed process of MT." But they will, I am afraid, not enable us to dispense with a human editor or editors for general language problems, whether on the input or on the output side, or on both sides of the MT assembly line. Moreover, Professor Oswald is well aware that "It is possible that it might be prohibitively expensive* to produce such glossaries.

Vocabulary Frequencies and Distribution

Of the greatest importance for the development of MT will be a conference paper by Professor William E. Bull of the UCLA, entitled PROBLEMS OF VOCABULARY FREQUENCY AND DISTRIBUTION. He exposes a number of "fallacies which are current in most discussions of word frequencies" From this highly techni-

4 See my chapter in the volume mentioned in footnote 3.

5 MODERN LANGUAGE FORUM, 36:1 - 24, 1951

cal paper I quote only the following passages of great relevance for the problem of "macro-" and "micro-glossaries":

> "There exists no scientific method of establishing a limited vocabulary which will translate any predictable percentage of the content (not the volume) of heterogeneous material. An all-purpose mechanical memory will have to contain something approaching the total available vocabulary of both the foreign (original) language and the target (final) language. In order to cover most semantic variations several millions of items would be needed. At the present time we have no machine which can manage such a number at a profitable speed."
>
> "A micro-vocabulary appears feasible only if one is dealing with a micro-subject, a field in which the number of objective entities and the number of possible actions are extremely limited. The number of such fields is, probably, insignificant."
>
> "The limitations of machine translation which we must face are, vocabularywise, the inadequacy of a closed and rigid system operating as the medium of translation within an ever-expanding, open continuum."

Operational Syntax and Teaching Foreign Languages

Extremely valuable not only for MT, but also for all those interested in improving the teaching of languages is Professor Bull's second paper entitled TEACHING FOREIGN LANGUAGES. I can here only quote some of the important suggestions made in his paper:

> "In teaching languages we should either replace rules by operational instructions or spell out in simple terms the operations necessary to make a rule work. I should like to stress in this connection, that the signs which may be used in teaching (and in the instruction of a machine) do not necessarily have to have any logical connection with the meaning. I shall give just two examples from Spanish. First, there are two verbs in Spanish commonly used to translate an English locative "to be": estar and haber. They are synonymous and even the educated native does not know what determines his choice. The signal is fundamentally non-semantic and the result of useless specialization in form usage. The problem, however, can be solved both for the machine and the student by isolating the fact that "the" in English takes estar and "a" takes haber.
>
> The man is here. El hombre esta aqui.
>
> A man is here. Hay un hombre aqui." [6]

It is Interesting here to note that Professor Bull's rule is perfectly applicable to the use of modern Chinese 有 (haber). In the first case one cannot use 有, in the second case one has to use it. Incidentally, Dr. Bar-Hillel also strongly advocates the development of what he calls "operational syntax" for language teaching as well as for MT.

Other important statements in Bull's paper are the following:

> "The total volume of the high frequency words is established by counting their uses with the words included in the selection and all their uses with the rare words excluded from the selection. The student, consequently, who learns this vocabulary is over-supplied with cement and under-supplied with things to be cemented together. He is like a builder who is given ten tons of cement and 500 bricks and told to build a home. If he keeps his proportions proper he has to be contented with an elegant privy. I submit that this is one of the major sources of irritation and frustration in our elementary courses in foreign languages. The reason our students cannot say anything much after a year of language is not because they haven't studied; they haven't got a vocabulary whose proportions permit them to say anything but the obvious banalities." (The underscoring is mine.)
>
> "The principle of excessive repetition cannot be sustained by the evidence of how a native is forced to learn his own language. This suggests strongly that we should increase the number of items given to the student and decrease, if possible, the number of repetitions of high frequency vocabulary."

---

6 TEACHING FOREIGN LANGUAGES, p.3. For the second example, see the original.

In his conclusion Professor Bull suggests the following points for consideration in the improvement of language teaching:

" (l) the abandonment of outmoded elementalism, and research directed at language as a structural whole
(2) a clear analysis of what is actually mechanical in language
(3) the description of what the native's language-feel actually is
(4) the substitution of operational instructions, whenever necessary for abstract rules
(5) research to discover the mechanical signposts which are guides to usage
(6) a new approach to the selection and teaching of vocabulary based on demonstrable facts"

Pivot Languages

Of the many valuable suggestions made by Professor Leon Dostert of Georgetown University I would especially like to mention one which will certainly become an important feature of future MT. Describing his experiences in multiple translations, he stressed the advantage of a "pivot language" or "pivot languages." General MT (mechanical translation from one into many languages), he said, should be so developed that one translates first from the input language into one "pivot" language (which in our case will, most likely, be English) and from that pivot language into any one of the output languages desired. This will, I believe, be very beneficial for MT, as will become clear from the following.

Model Target Languages

Professor Stuart C. Dodd of the University of Washington in Seattle addressed the Conference on MODEL TARGET LANGUAGES, (i.e., a regularized form of the languages into which one translates). His paper caused a very lively discussion as a result of which I can say that "model TL-s," especially his "model target English" will constitute an important item in the mechanization of the translation process. As I pointed out in the first of my two papers (MT WITH A PRE-EDITOR AND WRITING FOB MT), if we aim at a practical solution of MT, then we can interfere neither with the language nor the conventional spelling (speaking here entirely with respect to alphabetized languages) of the original language. But on the output side we can, within certain definable li mits, plan the form of the output language. We can put a selected vocabulary and a regularized morphology and syntax into the machine and, moreover, within the limitations of intelligibility, adjust the final language to certain peculiarities of each of the original languages.

Irregular Original Language - Model Pivot Language - Model Output Language

Now in General MT, if we do not work with a "pivot language," we shall (except in the case of original languages like Chinese and Japanese which by nature are very regular) in every case be faced with a mechanical correlation between one irregular and one regularized language. But if we do use a pivot language, then only at the first step will this be the case; that is, in the MT from a natural language into the pivot language. From here on, however, - that is, in the MT from the pivot language into any of the model output languages - we would in every case have a mechanical correlation between two regularized languages. Thus the use of a pivot language in General MT as suggested by Professor Dostert will mean a further simplification of the engineering problems involved.

Mechanical Abstraction of Grammatical-Information

In my paper quoted above I also demonstrated how the graphic indication by a human agent of certain types of grammatical meaning in the input text might enable the machine to determine incident non-grammatical meaning. Drs. Bull and Oswald, however, in their papers foresaw the possibility that a machine might be designed to determine grammatical meaning by itself, on the basis of nothing more than the conventional graphic form of input texts. If this is possible, then that kind of pre-editorial work which my idea necessitates can be dispensed with. It will mean much for MT if it can be demonstrated that operational instructions can be abstracted from a language on which we can base the programming of a machine for the mechanical determination of certain types of grammatical meaning. But even so it is important to point out the following:

a) even if this is possible for some types of grammatical information, it may not be possible for other types. In his MICROSEMANTICS Dr. Oswald mentions one kind of grammatical information for which he can - at least for the present - see only a human supplier. He says:

"The German system of noun compounding

> is such that a glossary based on the graphic forms would be both unwieldy and grossly inefficient because of unnecessary repetition. Almost any sequence of nouns in German not syntactically connected is automatically made into a compound, and your German noun strays gaily about appearing now as the "head" and now as the "tail" of a compound .... In a word, you must break up German compounds if you want to make any sort of efficient German-English glossary.... We know no mechanized process by which this could be accomplished, but an intelligent....pre-editor could indicate the dissection for any sort of context."[7]

b) even though it is possible for some languages, it may not be possible for some others.

c) the machinery required may be so complex and expensive that we may ultimately prefer to have a human agent indicate the relevant grammatical information of the input text by some system of symbolization (pre-editor).

d) if, as in the case of German compounds (see under a), no mechanized process can supply the information relative to one grammatical situation, so that this information has to be supplied anyway by a pre-editor, then the latter might as well add "seam-signals" to indicate the position of the "seam" (Oswald's "fracture-surfaces") in different types of compounds. The same signal would thus serve to indicate more than one type of grammatical meaning. This might result in a simplification of the mechanism designed for the determination of grammatical meaning because then the machine has more instructions on the basis of which to supply less information.

Mechanical Determination of Incident Non-Grammatical Meaning and the Limited Storage Capacity of the Mechanical Memory

A most serious objection to my suggestion of a mechanical determination of incident non-grammatical meaning was voiced by Dr. Bar-Hillel. He said that such a plan would require a storage of billions or trillions of entries - obviously quite impossible to achieve. However, appearances are misleading here. Before I can show this, I have first to introduce a few new concepts:

In the following I shall call "clue-sets" a set of co-occurrent words of which one or one group "pinpoints" the meaning of the remainder. I shall name "pinpointers" the pinpointing words and "pinpointees" those whose meaning is pinpointed by such "pinpointers." Furthermore, I wish to remind the reader of the phenomenon of "Shared Transferred Meanings" discussed in # H/6 of my first paper on mechanical translation and of the vast possibilities of "Pseudo-One-To-One Correlations" exemplified in my second Conference paper. Lastly I shall speak about "Pinpointees with a Manageable or Unmanageable Number of Pinpointers" and about "Pinpointee Meanings Stable or Unstable in the Light of Source-Target Semantics" (I beg the indulgence of the reader for the freak terms "pinpointer" and "pinpointee." I could not think of any other terms more "to the point.")

Now Dr. Bar -Hillel's objection remains valid only if we are thinking of putting into the mechanized memory all possible clue-sets. This is, however, neither intended nor necessary. We have to consider here the following facts:

1. Each set of two languages shares a considerable number of semantic parallels (shared transferred meanings). For example English will which, like Chinese 要, is used in the sense of "to want, to wish" and also as an auxiliary verb, expressing future; French ça va, German es geht and Chinese 行, meaning "to go" and also used in the sense of "that does" or "that will do"; Latin noli, "don't," a contraction of non voli, meaning "not want," and Chinese 不要, meaning "not want" and "don't"; etc., etc.

2. In an extremely large number of cases a literal translation, though resulting in an unaccustomed output form, is still perfectly intelligible either in the narrower or in the wider context. For example, in playing Chinese chess, a player may say 我吃你的象 which even in its literal translation, "I eat your elephant" (I take your elephant; the elephant is something like the bishop in Western chess), is perfectly intelligible to the English reader. We are in very many cases able to create artificial one-to-one correlations by selecting from the available output alternatives one which, though it may be customary or "good" only for certain context, is still intelligible in others. For example, Chinese 作, "to create, make, do,

---

7 Shortly after distributing my report on the conference I completely solved this problem of the mechanical dissection and identification of all predictable and unpredictable compounds. A detailed description of this solution, first reported in my SIMT Nos. 6 & 7 (mimeographed) will be included in the forthcoming volume mentioned in footnote 3.

act, etc.", is also used in contexts where the English translator usually prefers to render it by forms of the verb "to be." If we translate "make" also in these contexts, the result will often be horrible for the English hearer or reader, but it will still be intelligible. Thus "he is a teacher, student, father, son, etc.,etc." would appear in the English translation as "he make teacher, student, father, son, etc.", which in its context, for example in answer to questions meaning something like "what is his profession, position, what is he doing? etc." or when discussing somebody's duties in relation to his position, will be perfectly intelligible. A speaker of standard English does not need to learn pidgin English in order to understand what "this master makee teacher" (this gentleman is a teacher) means.

3. In every language there is a large number of words which may co-occur with a large number of other words "pinpointing" their incident meanings, but among these we have to distinguish several groups:

a) "Pinpointees" whose meanings in the light of source-target semantics (semantic relationships between the pair of languages) are the same with all "pinpointers," either in fact (semantic parallel, cf. point 1} or in terms of artificial one-to-one correlations (cf. point 2). Here no clue-set entries are necessary. The number of possible "pinpointers" is here, of course, of no consequence whatsoever for MT. For example German kaufen "to buy", verkaufen "to sell", schreiben "to write", essen "to eat", in terms of German-English and German-Chinese semantics.

b) "Pinpointees" the number of whose "pinpointers" is comparatively small and whose meanings in the light of source-target semantics are, in terms of points 1 and 2 above, different with all "pinpointers." Here all clue-sets should and can be entered into the mechanized memory.

c) 'Pinpointees" the number of whose "pinpointers" is large and whose meanings in the light of source-target semantics are, in terms of points 1 and 2, the same in the case of a very large number of "pinpointers," but different in the case of a small number of "pinpointers." Here no clue-set entry is necessary in the first case, whereas in the second all clue-sets should and can be entered.

d) "Pinpointees" the number of whose "pinpointers" is large and whose meanings in the light of source-target semantics are, in terms of points 1 and 2, the same in the case of a comparatively small number of "pinpointers," but different with regard to a large number of "pinpointers." Here no clue-set entry is necessary in the first case, whereas for the second the decision has to be deferred until we know more about the size of the total residual problem.

e) "Pinpointees" the number of whose "pinpointers" is large and whose meanings in the light of source-target semantics are, in terms of points 1 and 2, different with regard to different groups of "pinpointers." Here we can certainly enter all clue-sets relative to one of the groups, preferably the group with the largest still manageable number of "pinpointers," whereas for the remainder the decision has to be deferred until we know more about the size of the total residual problem.

f) "Pinpointees" the number of whose "pinpointers" is large and whose meanings in the light of source-target semantics are, in terms of points 1 and 2, different with regard to every "pinpointer" (this situation will be either rare or not occur at all). Here the decision has to be deferred until we know more about the size of the total residual problem.

Thus wherever transferred meanings are shared or wherever we can artificially create one-to-one correlations, no consideration of "pinpointers" is necessary and, consequently, we need not worry about the entry of clue-sets. Wherever transferred meanings are not shared, or wherever we can not artificially create one-to-one correlations, and where the number of "pinpointers" is comparatively small, we certainly can enter all clue-sets. Thus we are ultimately concerned only with the residual problem of those cases where "pinpointers" have to be considered and are very numerous. No research has ever been done for any set of two languages to determine the size of the residual problem. It is, therefore, not possible to decide on its treatment at present. If it still required more than, say, 10 million entries, one would naturally hesitate to consider recording in the mechanized memory. What is important, however, is that, assuming the residual problem required too many entries to permit mechanization, the machine would leave only this residual group of multiple meanings to a pre- or post-editor. The editor would have much less editing to do and in the case of a post-editor the difficulty of semantic determination might well be diminished to a degree he would certainly appreciate: the larger the number of semantic decisions the machine makes for him, the clearer the output context he has to consider for the solution of the remaining riddles! Certainly, in MT wherever mechanization is practical, it should be carried out!

Pre-editor Versus Post-editor

In this context I should like to add some remarks to the problem "pre-editor versus post-editor." In my first two papers on MT 1 burdened the pre-editor not only with the signalization of the grammatical, but also with that of the incident non-grammatical meaning; that is, wherever source-target semantics presented a problem of multiple meaning. In #81 of the first paper I had actually previously considered the alternative possibility of using a post-editor to whom, in the case of multiple meanings, the machine would supply the various alternatives from which he would have to make the correct selection. I had said there that from the point of view of complete mechanization this may seem to be preferable because then no human factor would interrupt the purely mechanical side of MT. However, from the point of view of MT as a whole, using a pre-editor is still much quicker for the following reasons: whereas the reader of the original text (i.e., pre-editor) has to select the meaning that "makes sense" in an original context which is completely intelligible to him, the output text reader (i.e., post-editor) has to do this in an output context which will necessarily contain a large number of non-distinctive words with transferred meanings different from those of the corresponding original language words, that is in a context that will often not be clear."

Dr. Bar-Hillel, on the other hand, advocates the determination of such incident meanings by a post-editor and has found much support for his idea. As a matter of fact, at this early stage of MT research I, too, cannot completely rule out the possibility that a MT post-editor (not to be confounded with a general post-editor concerned with stylistic improvements of the output text) may be necessary for the solution of at least some of the semantic problems involved.

Professor Oswald in his WORD-BY-WORD TRANSLATION voiced his scepticism concerning both the pre- and the post-editorial approach. "I do not believe," he says, "that his (i.e., Reifler's) combination of pre-editor with a mechanical dictionary constitutes the ultimate solution of our problem. In fact, I am of the opinion that we must grapple with the problem precisely at the point where Mr. Reifler abandons it. His proposals are most enlightening for the solution of problems of general language, but he has excluded problems of specific language (the jargon of medicine, mathematics, linguistics, geology, etc.) from the domain of mechanical solution. We shall be much closer to the realization of mechanical translation if we can mechanize the components of his "mechanized" dictionary....A pre-editor can do much to simplify syntactic connection for mechanical 'digestion,' but I do not see how, as an operator in the FL (i.e., foreign or original language), he can effectively guide either the machine, or the machine plus a post-editor, through the mazes of multiple meaning on the TL (target or final language). Nor do I think we can hope for much accurate help from one monolingual post-editor or even from one bilingual consultant. What has been overlooked is the fact that the competence required in the post-editor, even if he be bilingual, is only partially linguistic. The real prerequisite for him is an intimate knowledge of the field to which the translated text pertains" (pp. 3-5).

Apart from the fact that I have in no way "excluded problems of specific language...from the domain of mechanical solution" (I am fully aware of the urgency of the translation of scientific material, but would point out that even in such material we have to solve problems of general language), I fully agree with Professor Oswald. But he had, when he wrote his paper, not yet seen my third paper (the first submitted to the Conference) in which I indicated my radical departure from my previous position, demonstrated the possibility of mechanizing the determination of incident non-grammatical meaning on the basis of information relative to certain types of grammatical meaning, and limited the work of the pre-editor to the signalization of these types of grammatical meaning. Both Drs. Oswald and Bull have, on the other hand, mentioned the possibility that the determination of incident grammatical meaning may be mechanized. If this can be done, then there would remain only the question whether the solution of all multiple meaning problems (in case no portion of this problem can be mechanized) or of the semantic problems left over by the machine is - from the point of view of all-round practicality - better done by a pre- or a post-editor. I still feel that this task is easier for the pre-editor. The post-editor is faced with a non-conventional form of output context in which he has to make a selection from each of a number of conglomerations of output alternatives in consideration of one or more other conglomerations of output alternatives. He does, in fact, not fully understand the narrow output context before he has made at least some correct selections. The pre-editor, on the other hand, is confronted with a familiar linguistic medium without any conglomerations of alternative words and under-

stands the contexts before he is informed about the existence of a multiple meaning problem in terms of source-target semantics and before he has chosen the appropriate supplementary signal from the dictionary entry supplied by the mechanized dictionary. If we assume that a large portion of the multiple meaning problems can be solved mechanically along the lines I have suggested and that the pre-editor would thus be faced only with the residual semantic problems, then the combined man-machine procedure would be something like the following. The pre-editor sends the original text into the dictionary mechanism. In all cases of multiple meanings in which the dictionary mechanism can itself determine the incident meaning and supply the appropriate output equivalent on the basis of the supplementary grammatical signals which the pre-editor has added to the conventional graphic form of the original text (or on the basis of the grammatical information Bull's and Oswald's "grammar mechanism" has abstracted and supplied to the dictionary mechanism), the pre-editor would never have to know that multiple meanings in terms of source-target semantics are involved. The machine would do the work without giving any hint that there are such multiple meaning problems. In the case of a residual problem, however, the machine would in every case notify the pre-editor in some way and supply him with a dictionary entry (in his own language!) indicating the meaning alternatives in the light of source-target semantics. From these the pre-editor would have to choose and then add the appropriate supplementary signal to the portion of the input text involved. As pointed out above, he can make such a choice much quicker than a post-editor because he is dealing with a familiar linguistic medium and understands the output context before he makes his choice.

I should like to add that I am keeping an open mind with regard to this problem of pre-editor versus post-editor. It is, in fact, quite possible that, in terms of the time and money spent on linguistic and engineering research (linguistic research is probably less expensive than engineering research), mechanical complexity and construction time, speed and accuracy of translation, etc., etc., the optimum may be reached in an arrangement in which a pre-editor signalizes certain types of grammatical information, the machine abstracts some other types of grammatical information and on the basis of this information from two sources determines certain types of incident non-grammatical meaning and reshuffles the word order. A post-editor then solves the residual semantic problems on the basis of an output context which, because it does not contain too many clusters of alternatives, is much clearer.

Pilot Machines

Professor Dostert suggested the early creation of a pilot machine or of pilot machines proving to the world not only the possibility, but also the practicality of MT. Since the time necessary for the creation of such machines is an important factor, it will be best to develop a plan based on the simplest possible conditions. When this problem was raised at the Conference, the general opinion seemed to be that the simplest conditions are found in the mechanical correlation of certain European languages (Germani) with the English language. I pointed out, however, that contrary to appearances, a German-into-English scheme can not in the least compete with a Chinese (or Japanese) into English scheme. In the case of these two languages nature has already provided us with highly regular languages. Moreover, both in morphology and syntax Chinese and English happen to have more in common than German (or any other European language) and English. If we put into the translation mechanism a regularized Englich which is, furthermore, within the limitations of intelligibility, adjusted to certain peculiarities of Chinese, we have an ideal situation: a correlation between two regular and in many respects very similar languages. It is true that - as was stressed at the Conference - certain government agencies may be readier to supply the funds necessary for further research and improvements if the first pilot machine is designed for mechanical translation from Russian into English. But such a machine will be more complex and more expensive and the work necessary for its creation more time-consuming than in the case of a Chinese-English MT unit.

Thus the first pilot machine should, I feel, be programmed for a MT from Chinese into Englich. Moreover, if we want to go further and show the possibility and practicality of General MT (mechanical translation from one into many languages) on the basis of the concept of "pivot languages" as suggested by Dr. Dostert, our simplest proposition would be one in which we add to the Chinese-English unit a second unit for the translation of the English output of the first unit into Japanese. Then we would have a mechanical correlation merely between a regularized language (English) and another language (Japanese) which by nature is highly regular.

The Conference ended on an optimistic note

with the suggestion by Professor Booth that the next conference be held in London.

Chinese Characters Versus Alphabetization

I should like to add here a valuable suggestion which has come to me from Dr. Fang-kuei Li. With regard to languages with a non-alphabetic script I had hitherto thought of making use of an alphabetized form. I had pointed to the fact that, wherever different alphabetization systems have been suggested or are actually used, the graphio-semantically most distinctive one would be most beneficial for MT. For Chinese this would be the I.R. (Interdialect Romanization). But even in this romanization some additional differentiation is necessary in order to further reduce the still large number of homographs. Dr. Li suggested that, since even the I.R. requires further adjustments for purposes of graphio-semantic distinctiveness, it may be worthwhile to consider the development of sino-foreign MT on the basis of the Chinese characters themselves, which are graphio-semantically more distinctive than the I.R. He added that he had heard that a machine supplying the corresponding characters for the Chinese telegraph code numbers has already been developed in this country. There should be no reason why a machine which reverses this process could not be built. A pre-editor could add the supplementary grammatical signals just as well to a Chinese character text as to an alphabetized form of this text. The supplementary signals would be typed into the character-(code) number machine together with the characters to which they refer. Such an approach would eliminate the transcription into an alphabetization and thus save time.[8]

---

8 For dates and references to Dr. Reifler's papers on MT, see Vol. I, No. 1 of MT, March 1954.

HAROLD L. BARNEY, Bell Telephone Laboratories, Murray Hill, N. J.
(Active) "Work is being carried on in our area on machine recognition of phonetic elements, and of limited vocabularies of words such as the digits. Using techniques of pattern matching which have been described by K. H. Davis, a limited vocabulary automatic word recognizer might be adapted to a 'word for word' translation between two languages."

A. D. BOOTH, Birkbeck College Computer Laboratory, London., W.C.1, England
R. H. RICHENS, School of Agriculture, Cambridge, England
(Active) "The improvement of input-output mechanisms suitable for MT; the preparation of detailed micro-glossaries for scientific topics."

COLIN CHERRY, City and Guilds College, South Kensington, London, S.W.7, England
(Active) "I am working upon an Introductory volume to the Series on Communication to be published through M.I.T. by John Wiley. This is essentially at introductory level and deals, to some extent, with problems of describing language, mechanical translation, etc."

FRANKLIN S. COOPER, Haskins Laboratories, 305 East 43 Street, New York 17, N. Y.
(Active) "Analysis and synthesis of speech sounds and their perception."

D. G. ELLISON, Department of Psychology, Indiana University, Bloomington, Indiana
(Active) "Working on a mathematical model of certain behavior. A computer performing the operations specified by this model can learn problems of the type represented in MT. Its solutions are not limited to those which have been exactly duplicated in its 'experience.' I have worked in some detail on the application of a model to a computer which can read printed letters and correct some misprints when first encountered."

KENNETH E. HARPER, Department of Slavic Languages, University of California, Los Angeles 24, Calif.
(Active) "I hope to do some additional work on MT during my stay in Washington, D.C. this summer. It will be a study of Russian vocabulary and syntax and concerns the right-to-left structure in Russian words."

MORRIS HALLE and WILLIAM N. LOCKE, Department of Modern Languages, M.I.T., Cambridge 39, Mass.
(Active) "Automatic identification of speech sounds and use of such identification as input to translator."

RICHARD H. LAWSON, Department of Foreign Languages, State College of Washington, Pullman, Washington
(Interested) "The frequency in scientific German of idiomatic expressions not translatable by micro-glossary."

W. H. MEYER and J. WILKINSON, University of Chicago, Chicago 37, Illinois
(Active) "Have underway an extensive collection of papers that deal with aspects of the general theme 'Language and meaning.' One part of the anthology is intended to treat various 'structural' theories of meaning, another to attend to revisions of theoretical grammar in light of 'structural' meaning and the concepts of modern logic."

ERWIN REIFLER, Far Eastern Institute, University of Washington, Seattle, Washington
(Active) "1. The treatment in MT of words of dual nationality.
2. The mechanical determination of proper names.
3. The mechanical determination of incident non-grammatical meaning."
(Planning) "A further refinement of the MT form class filtering system."

PETER SHERIDAN, Scientific Computing Service, International Business Machines Corporation, 590 Madison Avenue, New York 22, N. Y.
(Active) "As you may already know from widely publicized accounts, a joint machine translation research project was initiated – under the aegis of the IBM program of endowed research in computation – and entered into by the Institute of Languages and Linguistics of the Georgetown University School of Foreign Service, under the direction of Prof. Leon Dostert, and the Applied Science Division of IBM. Dr. Paul Garvin of the Georgetown staff was assigned the task of formulating, in linguistic terms, the lexical and operational syntactic requirements of a pilot Russian - English conversion scheme, and this writer the task of reformulating that scheme in terms of logical syntactic patterns digestible by the IBM 701 calculator."

R. J. SOLOMONOFF, 1317 Franklin Avenue, Bronx 56, N. Y.
(Active) "I am interested in learning or 'thinking' machines. Present efforts show how useful learning might occur by extensions of statistical inference techniques. Sequences of question-answer pairs are used to test the learning ability of the method."

LUITGARD WUNDHEILER (Mrs.), 6031 S. Dorchester, Chicago 37, 111.
(Active) " 1. A study of ambiguity including a discussion of 'context-bound' words and morphemes.
2. Invariant syntax as a prerequisite of every translation."
(Planning) "On the concept of part of speech."
(Interested) "Mechanical resolution of words into morphemes: the best strategy."

VICTOR H. YNGVE, Research Laboratory of Electronics, 20B-101B, M.I.T., Cambridge, Mass.
(Active) "Application of statistical techniques to provide a description of syntax. Syntax and multiple meaning problems in general and in German-English translation. Application of information theory to problems of MT."
(Planning) "Lexical research: German to English for partial translations."
(Interested) "How to secure wide publication of translations; financial and copyright questions; what should be translated."

Victor A. Oswald, Jr. and 42
Richard H. Lawson
An Idioglossary for Mechanical Translation
Modern Language Forum, Vol. 38, No. 3-4, pages 1-11 (Sept. - Dec. 1953)

See abstract number 33, last issue.

Kenneth E. Harper 43
The Mechanical Translation of Russian: A Preliminary Report
Modern Language Forum, Vol. 38, No. 3-4, pages 12-29 (Sept. - Dec. 1953)

See abstract 32, last issue.

Anthony G. Oettinger 44
A Study for the Design of an Automatic Dictionary
Doctoral Thesis, Harvard University, April 1954

An automatic dictionary is the fundamental component of an automatic translator, and may be used independently to produce rough translations of technical texts. Experiments show that these translations can effectively be used directly by specialists in the subject matter. The use of an automatic dictionary also insures that only texts which actually merit translation and for which the output of the automatic dictionary is not fully satisfactory will be brought to the attention of professional translators. The

Oettinger- Automatic Dictionary (Cont.)

methods of logical design evolved in the study of mathematical machines are applied in an original fashion to the Russian language to obtain design specifications, and the application of a digital computer to the study of the structure of Russian is described.

Author

T. M. Stout 45
Computing Machines for Language Translation
The Trend in Engineering at the University of Washington, Vol. 6, No. 3, page 11 ff. (July 1954)

This is a general article which is intended to suggest some of the linguistic problems to the engineer and to explain some of the engineering ideas for the linguist. The multiple meaning of words is discussed. Alphabetic coding, input and output devices, magnetic drum storage and methods of dictionary search are described.

V. H. Y.