[*Proceedings of the Eighth International Congress of Linguists*, Oslo, August 5-9, 1957
(Oslo: Oslo University Press, 1958)]

WEDNESDAY AUGUST 7

Section A

MACHINE TRANSLATION

Chairman: YEHOSHUA BAR-HILLEL
Secretary: KENNETH G. CHAPMAN
Assistant: KNUT KRISTIANSEN
Rapporteurs: PAUL L. GARVIN
WILLIAM N. LOCKE and VICTOR H. YNGVE
ERWIN RE1FLER

a. Machine Translation

Report by

PAUL L. GARVIN
Institute of Languages and Linguistics,
Georgetown University

Machine translation is an extremely new field of research. It is only 8 years ago that the idea was formulated by Warren Weaver,[1] and hardly more than five years ago that work on the problem was developed in an organized manner.[2] It is therefore rather difficult to look at MT with complete objectivity and in perspective; I shall, as far as possible, present the problems in the field as I see them and attempt to avoid evaluative criticism.

The term 'machine translation' is self-explanatory. I prefer it the earlier term 'mechanical translation' because of the ambiguous connotations of that adjective; the Russians use the term 'auto-

[1] Translation, a memorandum written by Warren Weaver on 15 July 1949, reprinted in: William N. Locke and A. Donald Booth, eds., *Machine Translation of Languages, pp.* 15-23 (The Technology Press of the Massachusetts Institute of Technology, John Wiley & Sons, New York, Chapman & Hall, London, 1955).

[2] The MIT Conference on Machine Translation, June, 1952, and the Discussion Meeting on MT at the 7th International Congress of Linguists, London, September, 1952.

matic translation'. The machine which is expected to perform translation is a logical machine; either a high-speed, high-capacity general-purpose computer, or a machine especially designed for the purpose. The translation which is to be performed by the machine is, certainly for the present, intended to be of technical and scientific texts only, in view of the obvious additional complexities by, for instance, colloquial or literary texts.

Machine translation problems can be discussed in terms of the two components of the term: machine problems, and translation problems. Let me follow this breakdown in my discussion.

A logical machine, in order to translate, has to perform the following sets of operations: it has to read the input text in the source language, it has to manipulate the input translationally, and it has to furnish a usable output in the target language.

Reading the input and furnishing the output as machine operations are not fundamentally different from the input and output operations that a logical machine has to perform in handling any problem: since most digital computers operate with binary digits, the input operation has to include a transposition of the properly formulated source data into binary code, and the output operation includes the transposition of the binary result into more common symbols (decimal numerals and/or letters). The output can easily be equipped with a printer, resulting in legible printed text — this is common in modern computers, and hence no new equipment is required even for the ultimately contemplated translation program.

The input of modern computers consists of previously prepared punched cards, punched tape, magnetic tape, or the like. This requires preparatory equipment, such as a card punch or tape punch, which has to be operated manually. For a translation program, this means that a human operator has to read the source text and, say, punch it on cards or tape, before it can be fed into the input proper of the machine. In order to eliminate this preparatory human operation, the input of the machine would have to be equipped with an electronic scanning device, optical or acoustic, capable of reading printed text or perceiving speech sounds, and transposing them directly into binary code. At the present time, only the beginnings exist of either a visual or auditory scanner, and the technological difficulties are considerable, in view of the need for separating the relevant from the redundant features of the printed or acoustic stimulus by a machine analog of graphemic or phonemic analysis. Thus, machine translating input will in the immediate future have to be prepared by a human operator, in the same way in which input for any computer operation is now prepared. This input preparation does, however, not in any way constitute a preliminary editing of the text, but is a simple secretarial job, equivalent to retyping.

The machine manipulation of the text fundamentally involves two types of computer operations: table lookup and algorithmic (that is, properly computative) operations.[3] The table lookup operation consists in matching the sensed (that is, machine-read) input against a set of data stored in the memory unit(s) of the machine, and delivering these stored data to the arithmetic unit(s) of the machine for algorithmic processing. The result of this processing is the translated output, which is then fed into the printer and delivered to the user.

The ratio of table lookup to algorithmic operations will depend on the type of translation program prepared; it seems that modern computers are, or can be made to be, capable of performing either, and thus there are no foreseeable machine limitations on the choice of translation program. In an impressionistic way, it can be asserted that table lookup and algorithmic operations will always be in a roughly inverse proportion. Whether this impression can be technically validated, that is, whether an emphasis on table lookup will reduce algorithmic operations or *vice versa,* only detailed programming research will tell.

From the standpoint of machine design, the table lookup operation requires extensive memory storage capacity with very rapid access. It is quite obvious that, in any type of translation program, some kind of bilingual dictionary will have to be stored in the machine memory; in order to allow more than trivial translatability, a dictionary of considerable size will have to be contemplated. It is equally obvious that in any translation program input units will have to be matched one after the other in rapid succession against the units contained in the stored glossary. The rapid-access storage requirements of machine translation are far in excess of those required in current mathematical and logical computations; in the

[3] Gilbert King, The Problem of Lexical Storage, in: *Report of **the** Eighth Annual Round, Table on Linguistics and Language Teaching,* in press with Georgetown University Monographs on Linguistics and Language Teaching, No. 10, 1957.

latter, extensive storage may be required, but without rapid universal access. Bulk data, when required, are usually of the sort that can be fed into the machine consecutively from, say, a storage tape.

Technologically then, the problem is not the extent of storage but the requirement that any unit stored in the extensive memory be available for immediate lookup. At the present time, extensive storage is possible economically on devices with slow access; rapid-access memory devices are as yet of somewhat limited capacity. Research is, however, progressing extremely rapidly in this field, and it is quite thinkable that by the time an extensive translation program has been devised, a memory device will exist which can meet the requirements of the program adequately.

Regarding the algorithmic part of a translation program, the major difficulty lies in the fact that the algorithmic requirements of a translation operation are rather different from those of the mathematical and logical operations which modern computers are built to perform. The instructional details for which computer circuits are designed at present require long chains of addition, subtraction, shift and similar operations in order to accomplish what, for translation purposes, could be a single operation. Present-day translation programming, though already demonstrated to be feasible, is exceedingly cumbersome. A translation machine proper thus might contain algorithmic units rather different in design from the arithmetic units now in use in digital computers. Such an alteration in design presents no radical engineering problem, but can certainly not be undertaken 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. Engineering opinion here on the whole agrees with linguistic opinion that the translation program has to be formulated first, before any problems of machine design can be attacked realistically.[4]

A translation program, to be successful, has to accomplish more than merely the one-by-one transfer of units from the source

[4] I gather as much from discussions with Dan A. Belmore, Programming Consultant to the Georgetown MT Project, and from comments by computer engineers visiting the MT seminar at Georgetown.

language into the target language. It has to include some solution to the problems of choice implicit in the fact that (a) a unit in the source language may have more than one possible equivalent in the target language, and (b) that the order of source-language units in the input may not be suitable for the output in the target language. I have discussed these problems in some detail elsewhere under the headings of selection and arrangement;[5] the gist of the discussion is that the required selection and arrangement decisions can be programmed only if the contextual conditions can be determined under which any given decision from among several possible ones is to be implemented. The linguist's major contribution to MT research consists in the discovery of these conditions, and in the formulation of a routine for basing a decision on it.

There appears to be a certain correlation, on the one hand between lexical conditions and selection decisions, and on the other hand between syntactic conditions and arrangement decisions, but it is by no means to be assumed that selection decisions are based on lexical conditions only, nor that arrangement decisions are based on syntactic — or, more generally, grammatical — conditions only. One of the first results of my own research in MT has been that translation decisions cut across the various levels of linguistic analysis. On the one hand, the same decision may be based on a mixed set of conditions in the source and target languages: lexical, morphological, and syntactic. On the other hand, a given set of conditions — though assignable to one linguistic level only — may require both a selection and an arrangement decision.

A generally accepted example of the above is the translation of a Russian case suffix by one of several possible English prepositions. The translation of the case suffix separately from the base to which it is attached can be assigned to a set of morphological conditions in the source language; the choice from among several prepositions can be termed a lexical choice; the necessary rearrangement of the order of the translations of the base by an English noun, and of the suffix by a preposition, can be said to result from syntactic conditions in the target language. Thus, for this particular translation situation, there exists a mixed set of conditions — lexical, morphological, and syntactic, and both a selection and an arrangement decision are required in a single routine.

[5] Some Linguistic Problems in Machine Translation, in: *For Roman Jakobson,* Essays on the Occasion of his sixtieth birthday, Mouton & Co., the Hague, 1956, pp. 180-6.

Of particular interest to linguists, as well as a source of a good deal of discussion in the MT field, is whether all linguistic conditions have to be accounted for before a given translation decision can be made. My own opinion — in which I am supported by Martin Joos —[6] is that only some of these conditions are translationally relevant, and that one of the by-products of MT research will be a relevance scale of linguistic factors in terms of their effect on translation decisions.

Once the conditions for a given translation decision (or a set of conjoint decisions) have been ascertained, a routine must be formulated to recognize the appropriate conditions and to implement the required decision; the formulation must be logically flawless in order to allow for programming for a given existing computer, or for a yet-to-be-built translation machine.

As I visualize it in terms of my own experience, the first part of such a translation routine is a recognition routine: the place in the text requiring a decision (that is, the decision point) must be recognized as such, and subsequent to it the conditions for the choice of the appropriate decision (the decision cue or cues) must be found. The recognition routine is then followed by the implementation routine: selection and/or arrangement at the decision point are effected in terms of the decision cue(s).

Of the two routines above, I consider the recognition routine more difficult to formulate (and more fundamental), since a logical machine can not be expected to operate in terms of linguistic instructions (such as ‘find the noun in the nominative’, or ‘if no verb is present’). A code must therefore be devised which, based on linguistic information, allows the recognition of decision points and cues by the type of instructions proper to a logical machine (such as ‘if A is present, implement decision X, if not, implement decision Y’). The potential decision points and cues in the source language must thus be provided with appropriate code diacritics, and the code must be stored together with the source units to be matched and with the target units to be channeled into the output.

The formulation of a translation routine must thus include data of the required logical precision for the programming of the following: the source units to be matched against the input, the target unit(s) corresponding to each source unit, the necessary decision point and cue recognition code, and the implementation

[6] Final discussion, Eighth Annual Round Table, *op. cit.* in fn. 3.

instructions required to effect the needed decisions on the basis of the recognition routines.[7] All the above data must be stored in the appropriate memory compartments of the machine in order to bring about the necessary manipulation between input and output; the specific apportionment of storage space for this information depends on the technology of the particular machine used, and is therefore no longer part of the formulation but of the machine program itself.

The major centers in machine translation research at the present time are, to my knowledge, the Language Research Unit at Cambridge University in England, the Institute of Precision Mechanics and Computer Technology of the Academy of Sciences of the USSR, and in the United States, the machine translation projects of the International Telemeter Corporation of Los Angeles in cooperation with the University of Washington, of the Massachusetts Institute of Technology, and of the Institute of Languages and Linguistics of Georgetown University. In addition, individuals and small groups are active at various universities. The Linguistic Institute of 1957, held at the University of Michigan, will conduct a seminar in machine translation.

Of the above-mentioned centers, the project of the International Telemeter Corporation and of the University of Washington is primarily machine-oriented, the other projects are primarily language-oriented.

The International Telemeter Corporation is now engaged in the planning of a logical machine which is to include translation as one of its objectives, with primary emphasis on storage capacity, and without envisioning too much complex manipulation between input and output. The University of Washington group is preparing a translation program suited to the limitations of the intended machine. The purpose of the research is to investigate whether such a deliberately circumscribed operation may not be adequate for certain practical ends.

Of the language-oriented groups, that at the Massachusetts Institute of Technology differs from the others in its point of view, in that its research design contemplates a fairly self-contained linguistic analysis of the source and target languages as a necessary initial research phase, before approaching the translation

[7] For a detailed discussion of the formulation of such a translation routine, see my Linguistic Analysis and Translation Analysis, *op. cit.* in fn. 3.

problem proper. The source language is German; some rather interesting syntactic analysis of German has resulted.

The Cambridge group, the Soviet group, and the Georgetown group, on the other hand, have been trying to approach the translation problem directly, without an intervening phase of primarily linguistic analysis, but subordinating the analysis of the source and target languages to the requirements of the analysis of the translation process.

The Cambridge group differs from the other two by its emphasis on mathematical logic. Its research design contemplates the transfer of grammatical patterns from source to target language by means of a Boolean 'Phi' operation, This, if I understand them correctly, is to ensure the identification of translation units, and their proper manipulation as wholes for translation purposes, by the assignment of appropriate binary code digits to each table lookup unit (chunk) such that the 'Phi' operation performed on the code digits of any series of possible components of a translation unit of a given type will always result in a product equal to the code digits assignable to the head of that unit (thus, the product of the digits for all possible components of a noun phrase will equal the digits for its head, the noun). The transfer of semantic content is to be accomplished by a thesaurus routine, that is, a routine in which the semantic ranges of adjacent translation units are matched against each other by using coincident definitions in a thesaurus. The major problem of the Cambridge approach is, as I see it, given by the divergence of the relations of formal logic from the structural relations found within and between natural languages.

The work of the Soviet group, to the extent to which it can be assessed on the basis of information available, consists in effecting an analysis of the English source text, followed by a synthesis of the Russian output. Their approach seems to utilize English inflectional suffixes and word order as cues to bring about the appropriate inflected forms in Russian; in my opinion, the separation of the translation process into the analysis of English and synthesis of Russian is appropriate to this particular translation problem, but might not be suitable at all with a different source and target language.

The Soviet group has apparently been successful in testing its translation program on computing equipment.

The Georgetown group is basing its present work to some extent on the machine translation experiment performed in cooperation with International Business Machines Corporation in January, 1954. Although some diversity of opinion exists, and is encouraged, within the project, most participants agree on the postulate of effecting translation choices by ascertaining contextual cues, and on the basic design of subordinating linguistic research to the objective of translation. The major emphasis at the present time is on developing a properly coded machine glossary for a second more extensive test, in the sense in which I have defined the problem further above. The practical potentialities of machine translation have recently been summed up by L. E. Dostert;[8] I agree with him that, although for the present the major emphasis must of necessity remain on basic research, one may well be optimistic about future achievements and the possibility of their practical exploitation.

b. Research in Translation by Machine at M.I.T.*

Report by

W. N. LOCKE and V. H. YNGVE
Massachusetts Institute of Technology

In this paper we shall recount briefly the genesis and present state of work on the translation of language by machine at the Massachusetts Institute of Technology.

Speculation on the possibility of translation by machine from one of the languages of man into another is undoubtedly very old. Still, it is probably safe to assert that until the last decade no one actually envisaged the replacement of human translators by machines as we do today.

The first written suggestion that we have been able to find, to the effect that languages might be translated by computer, is in a 1947 letter from Warren Weaver, then Secretary of the Rockefeller Foundation, to Norbert Wiener of M.I.T. ' ... I have won-

[8] Practical Objectives in MT Research, *op. cit.* in fn. 3.

* This work is being carried out at the R L E, M. I. T., and is supported in part by the U. S. Army (Signal Corps), the U. S. Air Force (Office of Scientific Research, Air Research and Development Command), and the U. S. Navy (Office of Naval Research); and in part by the National Science Foundation.

dered if it were unthinkable to design a computer which could translate ...’. Weaver had discussed the idea with A. Donald Booth of Birkbeck College, London, a year or so earlier. Booth and R. H. Richens of Cambridge subsequently gave some careful thought to schemes for translating word stems and identifying flexional endings. They prepared a paper which, when presented at the first Conference on Mechanical Translation at M.I.T. in 1952, gave an illustration of the combined linguistic and engineering thinking which is one of the most significant characteristics of the field. Even earlier than that, the work of Booth and Richens had spread to the U.S.A. through Weaver’s 1949 memorandum ‘Translation’ [by machine] that communicated to others his enthusiasm and his faith in the ability of modern computers to recode from one natural language into another. That memorandum was the stimulus that started active work at M.I.T. In January 1950 Dr. Weaver met at M.I.T. with a dozen men from nearly as many different fields, including the heads of our Research Laboratory of Electronics, of our Digital Computer Laboratory, and of the Department of Modern Languages, and professors who were interested from one point of view or another in communication across language boundaries. The conclusion was cautious: the possibility of translation by machine was worth examining.

Yehoshua Bar-Hillel was given a full-time appointment at the Research Laboratory of Electronics in 1951, to study the question, first by a survey of the current thought and activity, then by planning what course our work should take. Bar-Hillel organized the 1952 Conference in the conviction that the isolated thinkers in England and the U.S. would profit by exchanging ideas. The Conference led to the conclusion that limited translation by existing computers was possible and that the research required to prepare the way for more complete translation was primarily linguistic. It is an extraordinary fact that, in our rapidly developing field, the same conclusion still holds true; progress in computer design and storage capacity has been far more rapid than progress in linguistics applied to translation by machine.

In 1953 Bar-Hillel returned to his position at Jerusalem University. Victor H. Yngve came to M.I.T. from the University of Chicago and recruited a group of linguists specifically to develop a new applied linguistics. The project has been financed by a series of grants from the National Science Foundation.

In 1955 the book, *Machine Translation of Languages,* appeared, and at about the same time the journal, *Mechanical Translation,* was started as a medium of communication among all interested in the field.

Yngve and his group set about the task of trying to find out how translations that are better than word-for-word translations could be achieved by machine. The well-known difficulties of multiple meaning and word order occupied their attention. At this point two possible paths presented themselves. The first was to make word-for-word translations, with modifications where the imperfections were most glaring or where it seemed possible to find some rather simple *ad hoc* or empirical rules. The more challenging approach was to try to find correct rules for translation that are based on an understanding of the structure of the language and their interrelations. These two philosophies of research will bear closer scrutiny.

The *ad hoc* philosophy holds that one should start on a word-for-word basis, and amend it, wherever there is a problem of multiple meaning or word order (a 'decision point') by searching the context in the neighborhood of the word in question for a 'clue' that will allow the mechanism to choose between alternative translations. By coincidence the proponents of this approach seem to be dealing with Russian and English. Both in the U.S.A. and Russia they have used computers to demonstrate a few translations. They have shown that a computer can substitute вода for 'water' or for 'the water' and *vice versa,* and that these substitutions will make sense most of the time. They have translated a number of sentences on the same basis. An example of an *ad hoc* rule for German-to-English translation that is 95% effective is: '*der* is to be translated "of the" when it follows a capitalized form without an intervening comma'. A rule of this sort will improve word-for-word translation considerably because 95% of the time it will give the right meaning for *der* but it will be wrong the other 5% of the time when *der* is nominative or dative. To be effective all the time a rule will have to provide a method (a 'recognition routine') that will enable the machine to recognize the case of noun phrases and take it into account in the translation.

The proliferation of *ad hoc* rules as one tries to deal with more and more of the troublesome items, plus the complications involved in each rule when one tries to increase the percentage of times that it gives good results, leads us to reject the empirical approach.

According to the structural approach that we are following, translation rules will be based on the grammatical and syntactic structure of each incoming sentence. The machine will make a structural analysis of each sentence in turn. This requires that we provide for the machine a comprehensive recognition routine that will enable it to recognize all structural features: the case of the noun phrases, the limits of the phrases and clauses, what modifies what, whether the sentence is active or passive, and so on. As in every type of recognition, an accurate description of what is to be recognized is required. That description must specify the essential features that distinguish the object or pattern to be recognized from all other objects or patterns. For language, this implies a detailed, accurate description of the grammatical and syntactic categories of the language, together with a precise statement of how they combine to form sentences. Descriptions adequate for this purpose do not exist. They have never been produced because there has never before been a need for such detail and such precision of statement. It is to the production of grammars of this type that we are giving our attention.

Our analysis of a language must be completely explicit and must be limited to the shape or form of the structures. For example, a human being can be told that " 's " is the English translation of the German genitive and he will be able to use this really incorrect rule, making due allowance for exceptions, often without the exception's being explicitly stated. If there were no ambiguity in the genitive endings of articles, adjectives, and nouns, if the genitive were unequivocally indicated every time, there would be no problem; but, as everyone knows, this is not so. Our machine will have to base its recognition of cases, and of the other morphological and syntactical constructions, on a description that lists the structures existing in German and the words or units associated in each structure. At M.I.T. we are now devoting ourselves to this description of German and an allied description of English. When we have these parallel structural statements, then we can look into the relationships between them, comparing the structure of each sentence with the structure of its translation. This is a completely new area: comparative syntax.

Upon a comparative syntax for pairs of languages can be based rules for acceptable translation — rules that can be followed by men or by machines — but it is to be emphasized that our aim extends beyond mere routines for mechanical translation into a more intimate understanding of the structure of human languages. We see language in a new perspective from the vantage point of the memory of a computer. It can know nothing, understand nothing except that which we store in it in minute detail, both as to separate items and as to relations between items. We believe that valuable new insights into language will come out of our basic work on structure.

c. The Machine Translation Project at the University of Washington, Seattle, Washington, U.S.A.

Report by

ERWIN REIFLER
University of Washington

1. *Introduction*

One of the largest U. S. Government contracts for machine translation development has been awarded to the University of Washington. Financed by the U. S. Air Force, the total value of this contract on the day of its fulfillment sometime in October this year will amount to $ 115,000. Our work falls into the following two phases:

1. *The Initial Project* which, financed by a grant of $ 30,000, was initiated in May 1956 and by March 15th this year supplied approximately 14,000 Russian-English operational entries for a translation machine memory.

2. *The Expanded Project* which is financed by an additional $ 85,000 and is to be completed by October 30th this year. It will increase the contents of this memory to approximately 200,000 Russian-English operational entries.

We are fortunate that our team has been chosen by the Air Force to work for the most advanced translation machine system under construction — that being built by the International Telemeter Corporation of Los Angeles. Some time in the fall this year the capabilities of this translation machine system will be demonstrated to representatives of the Air Force.

2. *The Limitations of the Translation Machine System Under Construction*

In order to understand the linguistic problems with which we are faced at present, it is necessary to say a few words about the limitations of the translation machine system under construction. This first machine will have a memory device with a practically unlimited storage capacity and an exceedingly low access time. But it will not yet have any logical equipment whatsoever for linguistic purposes. Consequently, not all of the linguistic problems involved in machine translation are at present accessible to a mechanical solution. We shall have no difficulty whatsoever with source-target semantic idioms. We shall however with this first machine not yet be able to reduce all grammatical and non-grammatical ambiguities to the grammatical or 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 of the output text will have to make his choice in consideration of the context. In many cases we are, however, able to reduce the number of these alternatives to such a degree that the output reader does not find it too difficult and time-consuming to arrive at the correct choice. This is done by making full use of the tremendous storage capacity of the photoscopic memory device of the International Telemeter Corporation and by certain editorial symbols which appear in the output text and help the reader in his choice. With this first machine we shall not yet be able to re-shuffle automatically the Russian word order into the word order required by conventional English. In many cases this does 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. Again in some of these cases we can alleviate the difficulty by changes in the form of the operational entry or entries concerned. But in other cases the source-target linguistic problems can only be resolved by the addition to the machine of logical equipment.

Another limitation of this first machine is that of its electronic reading device which automatically reads the tape-recorded Russian input text portions before they are compared with the entries in the memory of the machine. This reading device will not be able to read portions of the Russian input text which contain more than 16 symbols. This difficulty is being overcome by treating such portions as if they were compound words and dissecting them at points chosen from the point of practicality. That is, we are here applying a procedure which I had developed in 1952 for genuine compounds. But since the problem of these pseudo-compounds belongs rather to the field of pseudo-linguistics than to genuine linguistics, and since future translation machines will not have this limitation, there is no need here to go into this problem any further. Moreover, in about two years we shall have a device which will at one glance read a whole page and feed what it has read into a tape recorder and thus remove all human cooperation on the input side of the translation machines.

3. *The Remaining Linguistic Problems*

What linguistic problems do then still remain for the machine translation linguist? His main problems fall into the following two groups:

1. The problems of source-target syntax and morphology.
2. The problems of source-target semantics.

The peculiarities of our field require that these two groups 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'.

One of the basic problems in our linguistic research is that of form classes. Here we soon found it necessary to formulate the concept of 'operational form classes' as different from the traditional form classes. We are not only interested in what they look like, but also and especially in what they are doing. As a result we found it necessary to distinguish different *groups* of form classes and to change the membership in some of them. The details are found in my paper on the MT form classes filtering system. On the whole we are forced by the peculiarity of our field to keep in view the totality of a set of two languages, the total ascertainable vocabulary and the ascertainable totality of possible constructions and can never be satisfied with a so-called 'representative sample', although we also start out with representative samples. We shall be able to get very far in approaching this ideal of totality of possible constructions *because from the time the first machine becomes available, we shall be able to make use of machines to supply us with the material at a terrific rate.*

But we cannot even limit ourselves to the total ascertainable vocabulary. We even have to consider a certain type of future vocabulary, namely the so-called unpredictable compounds. As the result of my research in the summer of 1952 which was financed by a Rockefeller grant 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 there are only 30 types of theoretically possible compounds of which only 10 types are linguistically possible. I found moreover that only three matching procedures and 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.

We have no difficulty whatsoever with idioms. As a matter of fact, even with this first machine which will not have any logical equipment for linguistic purposes, idioms will get an idiomatic translation *which no human translator could do better.* But, as I have indicated earlier, we still have the problems of the grammatical ambiguities of non-distinctive paradigmatic forms and of the non-grammatical ambiguities of source language words with multiple target equivalents. I have to emphasize that we can here speak of ambiguities only if we consider the words concerned in isolation. If, however, we consider them in their environment, then they are in most cases not ambiguous at all. We also have the problems of disagreements in the word order of the two languages concerned in the translation process. Also here do we have to consider the environment in both languages if we want to elaborate the linguistic prerequisites for an automatic reshuffling of word order. Researches are already being carried on aiming at a solution of these problems. Mr. Robert E. Wall, Jr., an instructor in our Electrical Engineering

Department, together with a graduate student of his department, is at present working on the elaboration of a so-called ‘tag system’ which he is testing in experiments with the IBM 650 computer our University acquired recently. This research is based on ideas I developed and outlined earlier in a published paper and on language material becoming available in our research project. Another research aiming at a mathematical solution of these problems is being pursued by Mr. Aristotelis D. Stathacopoulos, another graduate student of the Electrical Engineering Department. He is carrying on this research in close cooperation with the linguistic members of our research team. Since Greek is his native language, I advised him to use the Greek language for his material, since it shares many characteristics, important for machine translation, with the Russian language.

In conclusion I wish to state my belief that it will not be very long before the remaining linguistic problems in machine translation will be solved for a number of important languages.

Discussion

PAUL L. GARVIN: (The speaker outlined the basic problem area as he sees it, based on his previous publications ‘Some Linguistic Problems in Machine Translation’ *(For Roman Jakobson,* The Hague 1956, 179—86) and ‘Linguistic Analysis and Translation Analysis’ *(Georgetown University Monograph on Linguistics and Language Teaching,* No. 10, 1957). He dwelled particularly on the necessity for working out a suitable code to handle problems of selection and arrangement, and exemplified the treatment of some problems of translation choice and syntactic identification in the code now being worked out by the Georgetown group.)

MICHAEL ZARECHNAK*: One of the research units of the current Georgetown University project in Machine Translation focuses its attention on problems of structural transfer from the source to the target language. This structural-syntactic orientation reflects the dissimilarity of linguistic systems, in that words can not be

* Report on the work in MT at Georgetown University, Washington, IXC. This material was mimeographed and distributed at the Congress. Ed.

translated in linear order. Because any linguistic system has its own structure which can be described by analytic procedures, the structural transfer operation can likewise be expressed in a code adaptable to a machine recognition technique.[1]

Linguistic structure can not be deduced from the linear arrangement pattern of its partials. In terms of symbolic logic, this would indicate that translation analysis should proceed along the lines of logical analogy, in that form is not identical with content.[2] In translation, operations of selection and arrangement are necessary at all linguistic levels, lexical, syntagmatic, and syntactic. Because these levels are not mutually exclusive, just as they are not linearly distinguishable, any machine translation program must be provided with means to analyze and synthesize functional units in successive inclusion. By 'functional units' we mean the various structural units of language, as opposed to the linear commutativeness of individual words or morphemes.

A procedure for effecting inter-structural transfer has been reported in one of the series of Georgetown work papers on MT.[3] The research was based on an exhaustive analysis of a sample of Russian chemical discourse and its English translation. The corpus utilized by the entire Georgetown project is a section of the Journal of General Chemistry of the USSR. Because translation implies equivalence, it is possible to compare sentences and paired items within the sentence, and thus discover regularities in structural transfer. In general, sentence boundaries define the domain of search for delimiting functional units and effecting transfer, in that the sentence constitutes the domain of grammatical relations. However, certain translationally ambiguous partials, such as a tense category or a pronominal item, may require search in preceding

[1] The programmer has at his disposal various means of handling a linguistic formulation. For a discussion of where a linguist's job ends and that of a programmer begins, see the Georgetown MT work paper MT-47, 'Formulation', by Dan A. Belmore. All MT work papers are available on request.

[2] A logical form not resembling its content can only through analogy represent the structure of a linguistic system, for example. We entirely agree with Suzanne K. Langer that 'Perhaps the most elaborate structure ever invented for purely representative purposes is the syntactical structure of language'; *An Introduction to Symbolic Logic,* Dover Publications, Inc., New York. Second Edition, 1953, p. 30.

[3] M. Zarechnak and Jane A. Pyne, 'Syntactic Transfer Procedures', MT-48.

sentences. The lexical and syntagmatic translation units may consist of one unit sensed (a word) or part of one or more than one. The search area rarely constitutes the complete sentence. Conversely, syntactic translation units are determined by examining the entire sentence including intermediate punctuation and any inserted structures. It should be noted that a sentence recognition technique, whereby sentence boundaries are established, must precede the delimitation of functional units.

An important consideration at this stage of MT research is the provision for rapid expansion of the glossary without basic changes in the intelligence assembly, the translation operation. This can be accomplished by separating constant from shifting structural features.[4] Such an approach results in a mixed logical system.[5] By the term 'mixed system' we understand a procedure whereby only *constant* diacritics are added to the individual glossary items, since these constant features are implicit in the word stem. Diacritics for shifting features must be added during analysis. For example, the gender of a Russian noun can be indicated in the glossary, but an adjective must be assigned a gender category.

At the present time the research unit under my direction is preparing to test the applicability of the various steps in the translation process in their logically formulated stage as a prerequisite to machine programming. We assume that the translation operation involves three major steps, morphological analysis,[6] syntag-

[4] See Roman Jakobson, 'Shifters, Verbal Categories, and the Russian Verb', Russian Language Project, Department of Slavic Languages and Literatures, Harvard University, 1957.

[5] 'A system wherein some truth-values may be deduced, but others neither imply anything nor are implied, is a *mixed system*'. Langer, *op. cit.* fn. 2, p. 80.

[6] Morphological analysis of Russian applies particularly to the recognition and division of nominal, pronominal and verbal base forms which are inflected by various infixes and suffixes. It is proposed that the main glossary should contain only base forms carrying non-shifting information. Infixes and suffixes are listed separately and are matched mechanically with the appropriate desinence type. Note that certain infixes and suffixes can not be treated according to traditional morphemic classification in that graphemes are not identical with morphemes; for example, we have found it necessary to employ the infix 'ENI' and 'NI' as identification signals for verbal nouns.

A set of about 150 logical formulas have been devised for the identification of inflected items. For further information, see MT-20, 'Review of Noun, Adjective and Verb Suffixes', and MT-32, 'Identification of Russian Items by Machine Procedures', both by M. Pacák.

matic and syntactic analysis,[7] and English synthesis.[8] Some elaboration of these steps and a list of pertinent available work papers are given in the references below.

WILLIAM N. LOCKE: In presenting to you the paper which Prof. Victor H. Yngve and I prepared for this Congress I would like to mention some general considerations. The first and perhaps the most important is the relationship of machine translation to the larger field of machine processing of information. May I emphasize that translation is a special case of information processing; that, therefore, progress in these two will go hand in hand. Unfor-

[7] Structural transfer may require one or more of the following operations :(1) choice between positional variants, (2) insertion, (3) rearrangement. Syntactic and syntagmatic analysis makes it possible to predict when and how a transformational operation is needed in translating from Russian to English. On the basis of morphemic analysis, forms are converted into functions; Russian units are translated into English according to their particular function.

Details of syntactic theory and research in structural transfer are reported in the following papers:

M. Zarechnak, 'Basic Syntactic Concepts of Russian', MT-29.

Jane A. Pyne, 'English Syntactic Concepts for MT', MT-38.

M. Zarechnak and Jane A. Pyne, 'Syntactic Transfer Procedures', MT-48.

M. Zarechnak, 'Types of Russian Sentences', and Jane A. Pyne, 'Some Ideas on Inter-Structural Transfer', in press, *Monograph Series on Languages and Linguistics,* No. 10, Georgetown University.

[8] The English synthesis program consists of resolving lexical ambiguity and transferring grammatical affixes, on the basis of the syntagmatic and syntactic analysis of Russian plus any specific stylistic requirements.

M. Zarechnak and Jane A. Pyne, 'The Range of Machine Search for Translation of Russian Pronouns', MT-22.

M. Sushko, 'Russian Phraseological Expressions', MT-27.

Nancy Fargo and Joan Rubin, 'Three Russian Prepositions, 'OT', 'DLYA', 'DO'', MT-30.

Nancy Fargo and Joan Rubin, 'The Russian Preposition 'K'', MT-31.

Nancy Fargo and Joan Rubin, 'Pronominal 'SHTO'', MT-37.

Nancy Fargo and Joan Rubin, 'Tentative Statement for Choice in the Translation of Noun Suffixes', MT-40.

M. Pacák and E. Pantzer, 'The Transfer of Russian Reflexive Verbs', MT-50.

M. Pacák, 'Impersonal and Infinitive Structures in Russian and their Transfer into English', MT-54.

Nancy Fargo and Joan Rubin, 'Prepositions 'S' and 'IZ'', MT-57.

Statements have been prepared on the transfer of Russian adverbs, conjunctions, and verbs; these will be available as soon as possible.

tunately, those interested in the two tend to come from widely different backgrounds with a possible bridge through mathematics and philosophy, but with an insufficient number of mathematicians and philosophers to provide for realty satisfactory communication between the linguists on the one hand and the librarians on the other. May I urge you as linguists to take an interest and participate in the general field of information processing; for we have much to gain from that study. Moreover, progress in that field depends, I am convinced, on the cooperation of linguists, since most of the information to be processed is expressed in natural languages. We shall also gain by the application of our techniques to concrete problems in a new domain.

Then, may I recall a few restrictions that, as far as I know, all the workers in the world have placed on their studies. In the first place they are limiting their work to scientific and technical material; that is, material where it is content rather than form which is of primary importance.

Another major restriction on our efforts is that no one, as far as I know, is actively working on machine translation of spoken language. This is not by choice. It was this aspect of the question which first attracted my own interest. But until we can identify speech sounds by machine we have no way of getting an input from speech into a translating machine. Indeed, the identification of speech sounds by machine is a translation problem in its own right; for we have to translate a non-linear symbolic system, speech, into a. linear machine code. So for the moment we have to be satisfied with working from the written language, through an operator who copies the text to be translated on a keyboard to provide a machine input.

In our work at M.I.T. we feel that one of the most important considerations for the present is not the solution of individual problems but the development of a new methodology for the analysis of language with a view to machine processing. To this end we are giving consideration to the theory of grammar. What are the characteristics of an ideal operational grammar? May I recommend to you in this connection a recent book, *Syntactic Structures* by Noam Chomsky (Mouton, ’S-Gravenhage, 1957). On the application side we are now working with the phrase as a syntactical unit, studying how individual components enter into phrases, noun phrases and verb phrases. Of course, the concept of noun and verb have to be refined as do all traditional grammatical concepts before they become satisfactory elements of a programmable machine syntax.

It is a pleasure to be able to announce to you that M.I.T. this summer accepted the first student in a new field, Communication Sciences and Linguistics, thus giving recognition to the mutual interdependence and fruitful association of these different disciplines in the field of machine translation.

In his introductory remarks Prof. Garvin mentioned that some aspects of machine translation seem like science fiction. May I remind him and all of you that the science fiction of today is the science of tomorrow.

ERWIN REIFLER: At the University of Washington we are working towards a solution of the linguistic problems of machine translation in 2 stages. The aim of the first stage is to determine how many of the bilingual linguistic problems can be solved by lexicography alone — that is, without any logical machine procedures for linguistic purposes whatsoever.

Already during this first stage we are creating the wherewithal for the work of the second stage, namely the analysis of the translation process itself. We believe that it would be very uneconomical, indeed, and of doubtful consequence to analyse the structures of the source and the target languages separately and then to try to correlate somehow the divergent problems. We have extracted all the *general language* material current in modern Russian scientific publications and are elaborating a large number of simulated machine translations which tell our linguists and engineers at one glance where the source and the target language are in perfect agreement and where they disagree. We have almost completed the mechanization of the elaboration of these predictions so that they will in a few weeks become available in large quantities. These predictions are serving as the basis for the analysis of the translation process itself. They are being studied by our linguists and engineers, and as a result of this study logical machine procedures are being developed by our engineers for the automatic resolution of those linguistic problems which can not be solved by lexicography alone.

In the second stage of our project which will begin after October 31st this year we shall concentrate exclusively on the development of these logical procedures already begun during the first stage.

I should like to use this opportunity to supplement and correct some of the statements Dr. Garvin has made in his published report. Some are of a more general nature, others concern the *University of Washington Project:*

1. On p. 504 Dr. Garvin says:

'In an impressionistic way, it can be asserted that table lookup and algorithmic operations will always be in a roughly inverse proportion. Whether this impression can be technically validated, that is, whether an emphasis on table lookup will reduce algorithmic operations or *vice versa,* only detailed programming research will tell.'

To this I have to say that it has been well demonstrated in the research at the University of Washington at Seattle that emphasis upon table lookup (increased storage) will reduce the required algorithmic (logical) operations. (Examples: paradigmatic forms.)

2. On p. 505 he says:

'At the present time, extensive storage is possible economically on devices with slow access; rapid-access memory devices are as yet of somewhat limited capacity .... it is quite thinkable that by the time an extensive translation program has been devised, a memory device will exist which can meet the requirements of the program adequately.'

Against this I have to point out that the International Telemeter storage device (large storage, low random access time) will be operative long before the completion of any programs,

3. On the same page Dr. Garvin says:

'A translation machine proper thus might contain algorithmic units rather different in design from the arithmetic units now in use in digital computers. Such an alteration in design ….. can certainly not be undertaken 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.'

It is the considered opinion of the engineering members of the

University of Washington project, arrived at in consultation with our linguistic members, that general algorithmic (translation logic or arithmetic) operations use selected existent units in normal computers; i.e. these units need 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 program is completed before preliminary work can begin.

There is, however, also another aspect to this problem. In the United States, and probably also in other countries, there are a number of public and private organizations which for the time being are still satisfied with a much humbler, much less sophisticated machine translation output as long as this output is already 'accurately intelligible'. These organizations have funds at their disposal which could, I believe, become available for further machine translation development if the MT pioneers are ready to combine their academic interests with the satisfaction of more immediate urgent requirements. Our project at the University of Washington is an example. The money for our project comes ultimately from the U.S. Air Force. However, the U. S. Air Force is primarily not interested in machine translation, but in an efficient information retrieval system permitting quick access to the enormous amount of information stored in its files. This automatic system is being developed by the International Telemeter Corporation of Los Angeles which, in turn, is very much interested in machine translation to which this automatic system is applicable. Consequently, I believe, if we are less dogmatic about what should be done first, we have a better chance to get money for what we should like to do in the first place.

4. On page 503 Dr. Garvin says:

'Of the above mentioned centers, the project of the International Telemeter Corporation and of the University of Washington is primarily machine-oriented, the other projects are primarily language-oriented.'

This is quite erroneous. The International Telemeter Corporation which is building the machine naturally is strongly machine oriented, but the University of Washington project is primarily language-oriented. It is sometimes difficult to harmonize these two interests, but until now we have succeeded. Because of the large staff working in our project we have a large number of staff members (6 or 7) doing completely ‘language. Oriented’ research, but they cooperate, of course, with the engineering members of our staff.

5. On the same page Dr. Garvin says:

‘The International Telemeter Corporation is now engaged in the planning of a logical machine....’

This should not be termed a logical machine, but primarily a storage device with certain logical capabilities.

6. On page 508 of Dr. Garvin's report we also read:

‘The University of Washington group is preparing a translation program suited to the limitations of the intended machine. The purpose of the research is to investigate whether such a deliberately circumscribed operation may not be adequate for certain practical ends.’

This is not the purpose of our research. Our research is not limited to the study of applications to the particular International Telemeter machine, but also extends into more pure forms of research.

7. On p. 509 Dr. Garvin says:

‘The Cambridge group, the Soviet group, and the Georgetown group, on the other hand, have been trying to approach the translation problem directly, without an intervening phase of primarily linguistic analysis, but subordinating the analysis of the source and target languages to the requirements of the analysis of the translation process.’

I do not have sufficiently detailed information about what the Soviet groups are doing. But I do know that, as far as all the other groups are concerned, the University of Washington group has assembled the largest bilingual material and simulated translation sample on which to base the analysis of the translation process. And I assure you, we are making full use of it. We have an almost complete store of semantic units belonging to the general language vocabulary in modern Russian scientific publications, and we are about to complete the automation of the production of simulated machine translations.

At last I should like to stress that the University of Washington accepted the development of an operational lexicography for the Telemeter device because we felt that this would serve as the basis for our further research in the analysis of the translation process.

M. A. K. HALLIDAY*: 0.0 The Cambridge Language Research Group was formed with the purpose of analysing language (1) by the collaboration of linguists, mathematicians and logicians, and (2) by the application of mechanical and particularly computer techniques. Work is at present being concentrated on machine translation, both because of the practical desirability of success in this field and because of the theoretical importance of machine translation for language research, which is becoming increasingly apparent as the work proceeds. The present report attempts to outline the Group's approach to machine translation from the standpoint of linguistics. The account falls into three sections: (1) the general programme for machine translation, (2) grammar and lattice theory, and (3) lexis and the thesaurus method.

0.1 Translation is regarded as a form of comparative descriptive linguistics; but whereas translation between a given pair of languages requires only particular (one language) and comparative (in this case transfer, i. e. two languages) description, we envisage it as a requirement of machine translation that the programme should be applicable to translation among all languages, and therefore we must face the necessity of universal (all languages) description. At the same time we must cope with the different levels of linguistic analysis, including the 'substance' (phonic or graphic, in this case graphic), the grammar, and the lexis, each of which in comparative analysis (including translation), as in particular description, has different techniques appropriate to it.

1.0 Clearly if work was concentrated on a one-one translation field, where only a straight transfer description is required, results might be expected much more quickly. But the whole programme might have to be remade for each pair of languages, and it seems preferable to aim at a universal-linguistic translation programme applicable to translation between any pair of languages. Linguists are rightly sceptical about the possibility of universal description, and if this wider aim is to be achieved it can only be by a rigorous separation of the particular from the comparative-universal ranges of validity (in MT terminology, of monolingual from interlingual features), and by their separate handling in the programme.[1]

* Report on the work in MT by the Cambridge Language Research Group, Cambridge, England.

[1] M. A. K. Halliday, 'Monolingual and Interlingual Chart for Italian Operators'. Work Paper, C. L. R. G., July 10th, 1957.

Furthermore the total programme consists of a number of operations in which are mechanized the various processes involved in translation; these ‘processes’ are neither theoretically independent nor chronologically sequential in practice, but in the devising of translation schedules it is useful to keep them apart so that research can proceed and modifications be made in one without prejudice to the others.[2] But the reference here to a distinct process does not imply that it forms an independent stage in the programme.

1.1 The input consists of the graphic substance of the source text, one input unit being one sentence. The treatment of the substance needs little comment, as it is in its linguistic aspects shared by all machine translation programmes; it may be worth noting that the data include some features which may be handled interlingually, such as capitalization and italics (where applicable), and some punctuation features such as quotation marks. The, text-substance is then identified as a sequence of ‘MT-units’ (which we call ‘chunks’) by matching against a dictionary; in languages with institutionalized words, with spaces between, the dictionary chunk will be less than, or coextensive with, the word, so that the matching involves the mechanical recognition of units within the word. Matching proceeds in reverse alphabetical order beginning with the longest segment that could be a chunk (the word) and proceeding by curtailment until a match is made. If no match is made, the source word will appear unchanged in the output.

1.2 The dictionary forms the bridge from substance to grammar and lexis, the chunk as ‘heading’ being followed by a reading giving monolingual and interlingual grammatical and lexical information about it. In the grammatical processes, which we envisage will involve both dictionary matching and subsequent mathematical operations, first the chunks and then the larger grammatical units are identified in the source language and then transformed into Interlingua (‘Nude’, because we do not clothe it in a substance) grammar.[3] Nude grammar, both of chunks and of larger units, has undergone considerable modifications, and has by no means

[2] R. H. Richens, ‘The Thirteen Steps; Basic Interlingual Syntax Programme’. Work Paper, C. L. R. G., July 8th, 1957.

[3] R. H. Richens, ‘A general programme for mechanical translation between any two languages via a notional interlingua’. Paper presented at the Second International Conference on Machine Translation, October 1956 (to be published).

reached its final form. But it seems clear that considerable use can be made, both in the dictionary entry and in the operations, of the descriptive distinction between those chunks which can be fully identified in the grammatical analysis (the grammatical chunks or 'operators') and those only partially identified in the grammar and requiring further, lexical, information (the lexical chunks or 'arguments'). This is of course an arbitrary distinction made for machine translation purposes; it reflects the different fields of application of the grammar and the dictionary in descriptive linguistics, whose boundary is similarly vague and is varied for different descriptive purposes. (See below, 2.)

1.3 In the lexical processes those chunks which cannot be fully described in the grammar alone, the 'arguments', require a lexical translation. The problem here lies in the systematization of the lexis in such a way as to yield some form of interlingual lexical unit. The usual type of bilingual dictionary entry presents a list of translation equivalents, the selection of one among which in machine translation involves a system of choices depending on a variety of contextual factors. The choice among real homonyms in the source language is not the chief problem here; this can be assisted by the sort of broad context indication envisaged in most machine translation programmes. The main difficulty lies in the choice among near-synonyms in the target language, and one proposed solution to this is that the lexis should be described in thesaurus series, the thesaurus 'head' or 'key-word' being then the form taken by the Nude lexis. (See below, 3.)

1.4 The processes of transformation from Nude into the target language are largely a repetition of the source language — Nude processes, though not necessarily with the same or a directly reversed order. Redundancy has a different effect in the two halves of the programme; some particular grammatical features, such as various types of concord (i. e. multiple exponence of one context-grammar category), which can be ignored in the source language, must be catered for, while others are used only for identification purposes within the source language. Broadly speaking, Nude represents the stage of translation between the shedding of the redundancy of the source language and the building in of the redundancy of the target language; and the difference between the two halves of the programme represents the different requirements of these two processes.

2.0 The basic problem in the grammar is the setting up of relations among the particular grammatical structures of different languages. Comparative linguistics has the theoretical equipment for doing this, by reference to categories of context-grammar; and the systems of context-grammar categories set up for machine translation make up a grammatical interlingua such that any single language is capable of comparison with them. This grammatical interlingua, or Nude grammar, is not an artificial language, which would merely turn the number of languages we have to deal with from n to n + 1, but a set of systems of grammatical relations identified in context-grammar, of the type that one sets up for the comparative identification of grammatical categories in descriptive linguistics, in order to look for the exponence or non-exponence of these systems, and of their terms, in the group of languages under study.

2.1 The monolingual information entered in the dictionary with each chunk, operator or argument, identifies it in the *particular* grammatical system of its own languages, including word class indication, flexion and sandhi class, and the like. Much of this can be ignored with a source language, by the use of the internal redundancy of the language; but it is required for each language as a target language. Similarly larger monolingual units are built up out of chunk sequences according to the class information by a number of reductive cycles, until the syntax of the whole sentence is analysed monolingually, permitted monolingual sequences are entered in a separate dictionary against which the reduction is checked.

2.2 The identification of chunks and sequences in Nude grammar has evolved through many stages. At first, chunks were simply given an interlingual word class indication, but the criteria for the classificatory type of identification are complex and the result generally unsatisfactory. The method which seems at present likely to be most fruitful, and is being tried out on a limited number of languages (Italian, Chinese, English, Russian, and Malay in the first instance), is to establish a rigid operator/argument distinction, and to identify the operators by their placing in a number (provisionally about 60) of two-term grammatical systems, each term being a yes-or-no function, so that each system carries two bits of information. The arguments are then classified by reference to groupings of these systems.

2.3 Grammatical structures were at first identified in Nude in a bracketing system of 0-ad, 1-ad and 2-ad relations, either with or, later, without distinctions of category within the elements. But the simple expression of grammatical relations by order of combination of elements is likely to give place to an approach of a more structural nature, the linking together of the Nude chunk-equivalents by various kinds of 'bond' independent of word-order; the topological configurations of these 'bonds' give the Nude sentence-grammar.

2.4 In this form interlingual grammar can in principle be represented by a structure having the mathematical properties of a lattice. This has been the, basis of the attempts so far made to programme the syntactic analysis of text material. The lattice sub-structures are built up in the computer on the basis of the interlingual group information contained in the dictionary readings, which for this purpose serve as 'Lattice Position Indicators'; and at least a substantial part of these sub-structures will constitute, without further transformation, an interlingual system.[4] These interlingual structures can then serve as guides for programming the process of building up the grammatical forms of the target language. Moreover, the lattice sub-structures may supplement the interlingual information provided by the dictionary about the chunks themselves; if the lattice position can be assigned by elimination to a given chunk, some of this information may be made to follow from this.

3.0 What primarily distinguishes lexical from grammatical analysis is the absence from the former of structurally defined closed systems, so that lexical items are usually described in arbitrarily ordered lists as in a dictionary. An alternative possibility[5] is to describe the lexis in 'series', groups of words having

[4] Margaret Masterman, 'Comparative Analysis of a Chinese Sentence'. Presented at the Second International Conference on Machine Translation, October 1956; available from the Editors of *Mechanical Translation.*

Report of Colloquium of Cambridge Language Research Group, August 1955; *Mechanical Translation.* III No. 1, Cambridge Language Research Group Issue.

[5] Margaret Masterman, 'The Potentialities of a Mechanical Thesaurus'. Presented at the Second International Conference on Machine Translation, October 1956 (to be published).

M. A. K. Halliday, 'The Linguistic Basis of a Mechanical Thesaurus'. Presented at the Second International Conference on Machine Translation, October 1956. To be published in *Mechanical Translation,* III No. 3.

A. F. Parker-Rhodes, 'An Algebraic Thesaurus'. Presented at the Second International Conference on Machine Translation, October 1956. (An enlarged version of this paper is to be published,) contextual commutability — the potentiality of operating in one and the same context with different lexical meaning. The exact limits of the range of any one series are of course arbitrary, and depend in particular on the range of the context taken into account, any expansion of which increases the number of words of which it is possible to say that they lie within or outside it. A lexical description using the series method is a 'thesaurus', a standard example being Roget's *Thesaurus of English Words and Phrases,* whose divisions (sub-paragraph, paragraph, section, etc.) correspond to different ranges of context. The thesaurus is an attempt to systematize the lexis, and its use in machine translation will, it is hoped, lead to some parallelism between the various grammatical and lexical processes enabling many of these to be carried out simultaneously.

3.1 The mechanical thesaurus would consist of a selected set of 'heads' each of which represents a series of chunks associated by contextual commutability or some other lexical relationship. In the first place such lexical relationships are established monolingually, and we envisage that initially the thesaurus will be set up in the target language. But the possibilities are being explored of using the thesaurus also for comparative lexical statement, by establishing an interlingual set of thesaurus heads. This must be done if there is to be any considerable degree of conflation of the grammatical and lexical processes.

3.2 The dictionary readings required by such a thesaurus would ideally consist in the yes-no placing of each argument chunk with reference to the total inventory of interlingual heads, the sequence of placings forming the lexical definition of the chunk. In practice, of course, the number of heads would be such that the dictionary readings would be far too long, but it should be possible with code compression to retain the principle while yet reducing the number of bits to a manageable size. This form of entry has the advantage of parallelism with the grammatical section of the dictionary reading. It would still be possible to group the thesaurus heads (as the grammatical systems are grouped); the group of heads would correspond to a wider range of context than the single head.

The syntagmatic analogue of the 'ranges' of context represented by the heads (sub-heads, groups of heads, etc.) is the hierarchy of structures up to and beyond the sentence into which the chunk enters textually. Each of these structures brings the given chunk into association with a different set of neighbour chunks, from the lexical (thesaurus) part of whose dictionary readings context indications can be constructed. This hierarchy of structures may be termed the ‘layers’ of the ‘lesser’ context, and contrast with the ‘greater’ context as defined by such things as the title of the text and by general indications unrestricted by syntagmatic relation to any given chunk.

3.3 With a monolingual thesaurus the heads are lexical key-words of the target language; the source chunk dictionary reading is a lexical item of the target language from which one is led by the key-word associated with it into a series; the final output word is determined by the context (including collocation) indications. This process corresponds to the human translator’s use of a book thesaurus such as Roget’s. With an interlingual thesaurus, however, the heads would be categories of context-lexis (just as the interlingual grammatical systems are categories of context-grammar), and the source language chunks would be assigned directly to thesaurus heads. The reading, together with context indication, could be made to yield an output taking into account the thesaurus placings of chunks in the lesser context, and this output would be matched with the target language dictionary. The nearest approximation to an exact match would be the translation-equivalent of the chunk. If the process yields no output, the syntagm is one which cannot be analysed comparatively; and this would be one means of recognizing idioms, these being defined as syntagms in the translation of which no chunk has a one-one correlate.

V. IVANOV*: There are three groups working on machine translation in Moscow.[1] The first group is working in the Institute of Exact Mechanics and Computing Technique. This group has worked on a program for machine translation from English into Russian based on word-by-word analysis of an English text. The results of the work are published in Panov’s book *Automatic Translation* (Moscow 1956) and in a report read at the meeting of the Academy of Sciences (‘Sessija Akademii Nauk SSSR po naučnym problemam avtomatizatsii proizvodstva’ Moscow 1957). This group has also

* Report on the MT work in the USSR.

[1] The work on machine translation has started in Leningrad. There a seminar on machine translation is organized by N. D. Andreev.

started a work on translation from German, Chinese, and Japanese into Russian.

The second group works in Steklov's Mathematical Institute. The main principles of its work were expressed in the paper read at the same meeting ('Sessija...'). The work of the second group is based on formal analysis of the linguistic structure. This group has elaborated a program for machine translation from French into Russian (see an article by O. Kulagina and I. Melchuk in *Voprosy jazykoznanija* 1956, No. 5; in the same issue of the magazine other articles on machine translation may be found). The program for English-Russian translation was worked out by T. Moloshnaya. In this program formal classification of English words was given and the rules of analysis were constructed according to syntagmatic theory of the structure of a sentence (see Moloshnaya's papers in *Voprosy jazykoznanija* 1957, No. 4, and in *B'uleten' ob'edinenija po mashinnomu perevodu).* Recently I. Melchuk has elaborated a program for machine translation from Hungarian into Russian. In his recent works I. Melchuk suggests the necessity of the analysis of whole groups of words and of constructing an intermediate language (interlingua) by giving indexes of similar syntactic structures found in different languages.

The third group is working on the problems of machine information retrieval in the Electro-modelling Laboratory of the Academy of Sciences (see articles on the subject in *Vestnik Akademii Nauk SSSR).* This group considers construction of an abstract machine language as necessary both for information retrieval and for machine translation. This abstract language should be an effective system of encoding scientific information. The problems of constructing such a language were discussed at a special conference in May 1957 where V. Uspenskij's report on the mathematical aspects of the problem and my report on linguistic questions of constructing machine language were read. We suggest that this language can be constructed as a metalanguage for the ordinary languages.

Finally I should like to stress the importance of machine translation for the analysis of the structure of a language and of its functioning in the process of constructing messages. Recently experiments on machine translation from French into Russian were carried out in Steklov's Institute. Some forms in the Russian translations can be found that are similar to forms built by analogy in real language. Thus it seems possible to construct a 'grammaire des fautes' of a machine that will be very useful for general linguistics.

JOHN P. CLEAVE*: 1. Research on the basic structure of MT programmes has been carried out. A programme of a type capable of indefinite extension has been constructed for A. P. E. X. C., which is a general purpose medium speed computer with a magnetic drum store of 1024 locations, each location capable of handling 32 bits. The program compares incoming words with the main dictionary by means of the bracketing procedure and stores the 'first equivalents' thus obtained in order in a special track T. The first equivalents each occupy one storage location and consist of three sets of ten digits, the remaining two of the 32 being spare. The first set of digits, ten in number, specify an address A which is the address of the next routine to be obeyed after the equivalent has been stored. Thus if a complete word has been identified, address A is merely the address of the input routine which begins the whole cycle of operations again. If, however, a word is input which it has been found convenient to split into stem and ending, then the comparison of input word with the dictionary yields a first equivalent (for the stem) whose A-address is the address of a routine which compares the word minus stem to an ending dictionary. From this is extracted the first equivalent for the ending whose A-address is that of the first operation of the input routine.

The symbols signifying the end of a sentence are entered in the main dictionary as other words. The A-address of these first equivalents specify a routine for processing track T containing the first equivalents. Track T thus contains a series of items uniform in size and structure, which is convenient for any subsequent operations.

The occurrence of an end-of-sentence symbol initiates a routine for systematically operating the routines — called 'condition routines' — specified by the second address, the B-address, of each of the first equivalents in track T. The routines are used to produce a code number specifying the grammatical function of each of the first equivalents. These are stored in order in another track V. The condition routines are also used to resolve ambiguities. The B-address of the final first equivalent, which is an end of sentence

* Report on the work in MT at Birkbeck College, University of London. Read by Erwin Reifler. Ed.

symbol, is used to initiate a routine for changing the order of the first equivalents. Groups of code-numbers from the track V are compared to a dictionary of sequences of code-numbers — the structure dictionary — the equivalents of which are used to effect change of order in the track T containing the first equivalents.

Following the change of order, the printing routine is brought into operation. This takes the first of the first equivalents in track T and prints out the contents of the location whose address on the drum is given by the third group of ten digits comprising the first equivalent. The conclusion of the printing operation initiates the input routine again.

Distinctive features of the programs are

(1) the uniform structure of the first equivalent. This facilitates handling problems. It is always easier to deal with the address of a word than with the word itself which almost always contains a larger number of digits. Thus the processing of the first equivalents can be systematized and the actual TL (target language) words handled only at the last moment by the output routine.

(2) the systematic use of a dictionary procedure for change of word order.

(3) the open structure of the program. This means that further more complex routines may be added to the program by increasing the number of entries in the dictionaries or by changing the addresses in the first equivalents. Both these processes do not necessitate any alteration of the program itself at all. Thus by increasing the number of entries in the structure dictionary more changes of order can be dealt with. By changing the B-address of a first equivalent a new condition routine can be added to establish a finer resolution of an ambiguity.

Limitations to the program are

(1) the unsystematic nature of the condition routines themselves. Though they are systematically brought into operation, at the moment, each condition routine of the type ‘Perform such and such an operation if the sequence p, q, r... of first equivalents occurs’ is programmed separately, whereas they might profitably be united in a dictionary procedure of some form.

(2) The type of rearrangement of word order possible by the above means is confined to permutation of at the most eight neighbouring elements at present. This is well suited to pronoun-verb inversion in French, but of course not for German.

(3) The main limitation, however, is the small storage capacity available. This limits both the number of dictionary entries and also the number of letters in each.

2. Another line of research is on the possibility of the mechanical translation of German into English. Analysis of German texts (at the moment confined due to lack of staff to the field of Electron-microscopy only) is being prepared by punching the German texts and English translations onto teletype tape. Statistical analyses will then be performed under control of a computer program.

P. MEILE: Il est un peu étonnant que l'on envisage généralement un automatisme complet de la traduction: n'a-t-on pas songé à des étapes intermédiaires consistant en semi-automatismes ou automatismes partiels?

Puisque la machine à traduire doit résoudre une pluralité de problèmes, très difficiles les uns et les autres, ne sera-t-elle pas, une fois achevée, la réunion de plusieurs machines ou dispositifs? Certains de ces dispositifs partiels sont-ils prêts dès maintenant? Ne peut-on nous en communiquer les résultats? Sinon, ne pourrait-on consacrer quelques efforts à les mettre au point séparément? Par exemple un «lecteur», à lui seul, rendrait des services. Et nous aurions du même coup une machine à translittérer, qui semble relativement aisée à réaliser et qui serait utile par elle-même. De même, il existe des analogies entre une machine à traduire et une machine à résumer. Ne convient-il pas de considérer en lui-même le problème du «résumé»?

Chacun de ces automatismes partiels pourrait peut-être apporter une aide immédiate aux traducteurs: ne pourrait-on considérer avant tout l'homme, dont il y a lieu de soulager l'effort, en augmentant son rendement? Les linguistes traducteurs aimeraient que les technologues s'intéressent aussi à des projets plus modestes qui, tout en leur facilitant leur tâche quotidienne, s'avéreront peut-être avoir été des étapes nécessaires vers l'automatisme complet.

Il conviendrait d'observer le traducteur humain opérant: par exemple, un traducteur efficient travaille sans dictionnaire. Le vocabulaire dont il dispose ne paraît pas être emmagasiné sous la forme d'un dictionnaire courant; or, dans les essais actuels de traduction automatique, il ne semble pas qu'on soit arrivé, jusqu'ici, à une formule réaliste de «magasinage» lexical. Ou alors nous admettrons, comme on l'a déjà fait, que le dictionnaire emmagasiné par la machine sera d'une formule nouvelle, recourant largement à l'analogie, par exemple.

Il est indispensable de dire nettement si l'on vise une machine à traduire universelle, ou une machine réduite à une certaine catégorie de textes. Donc la nature du texte importe. Et il faut tenir compte aussi du destinataire, puisque toute communication comprend une pré-information en même temps qu'une information.

Enfin, il est à souligner que les exigences nouvelles posées par ces recherches intéressent toute la linguistique, dans la mesure où les règles opérationnelles que l'on s'efforce de dégager pourront s'identifier ou correspondre à des lois réelles du langage, dont on tirerait parti de diverses façons, à la fois dans la description des langues et dans leur enseignement pratique.

WILLIAM N. LOCKE: Numerous of the comments made during this discussion period reflect the progression of ideas which has led the workers on machine translation to the present stage. It is fascinating to see how a new group of linguists follows step by step almost the identical path. May I recommend to those of you who have not seen it the book *Machine Translation of Languages,* published by John Wiley and the Technology Press in the United States, Chapman Hall in London, in 1955. In this book A. Donald Booth and I collected practically all the early papers on the subject. For those interested, the book will give a bibliography complete up to the date of its publication. An annotated bibliography is continued in the journal *Machine Translation,* published at M.I.T. We also publish articles and should be happy to have any of you submit papers. We are particularly eager to receive papers from our Russian colleagues, whom we now know to be so very active in the field. Their journal, *Problems of Linguistics,* should not be overlooked; for it is publishing a series of important articles on machine translation.

In conclusion may I say that this field is one in which there are too few active workers. It is my sincere conviction that linguistics as a science stands to reap enormous benefits from the type of objective analysis of language which these studies are forcing us to do. We need the help of all of you and of many more linguists to carry this work to a successful conclusion.

PAUL L. GARVIN: The field of machine translation is not yet accorded the status of a serious discipline by all linguists. In order to earn the respect of the linguistic profession, machine translators will have to learn to apply relevance criteria to their work more rigorously than is now the case.