[From: Paul W.Howerton and David C.Weeks (ed.) *Vistas in information handling* (Washington, D.C.: Spartan Books, 1963)]

CHAPTER 5

MACHINE TRANSLATION AND AUTOMATIC LANGUAGE DATA PROCESSING

LÉON E. DOSTERT
Georgetown University

This chapter discusses machine or automatic translation of natural languages. It reviews the status of the art at present; explains its basic operations, methods and procedures; indicates its objectives and uses, and situates machine translation or MT in the general field of automatic language data processing. Finally, it suggests its possible role in language communication as a whole.

Machine translation is a relatively new area of automatic language data processing. It came about in part as a result of the conjuncture of three trends: (1) the development of structuralist procedures in linguistics; (2) the increasing sophistication of programming techniques, and (3) the growing capabilities and versatility of computation devices. It also became a subject of interest in the scientific and managerial communities as a result of the increasing volume and diversity of scientific and technical writings in the several languages of scientifically creative cultures, and the lengthening lag between the publication of information in a given language and its accessibility in one or several other languages.

A decade ago machine translation was of interest to a relatively small group of people coming from such apparently unrelated fields as philosophy, physics, mathematics, sociology, logic, computational engineering, chemistry, and of course linguistics and languages. This diversity of background among the early comers was to bring about a widely diversified and divergent set of notions as to what automatic translation is or should be, what it ought to try to do, how and why it should do it.

Notwithstanding these divergences, MT research today is pursued in a number of centers and laboratories in some twenty countries, including besides the United States, where oriented research may be said to have originated, Great Britain, the U.S.S.R., Japan, Italy, France, Belgium, Germany, and others. The first public demonstration of feasibility was carried out jointly by Georgetown University and IBM in January, 1954, on the basis of an experiment for the transfer of a small corpus of Russian

to English on an IBM 701. The methods and results of this experiment, necessarily limited in scope, were made public.

Today there are, to my knowledge, three regular MT professional journals, one in this country, one in France and a third in the U.S.S.R.[1] At least two general national symposia have been held, one in the U.S., the other in the Soviet Union and the first International Conference on MT, was held at the National Physical Laboratory, Teddington, England, in September 1961. A first international seminar was held under the auspices of NATO in Venice last summer. A large number of reports and studies have been published by individuals and groups in several countries. Experimental and trial runs of increasing scope have been reported and some of their text output published for examination and review. Finally, courses in MT (or courses related to basic research in the field) have been introduced in the curricula of a number of universities in several countries and programs of studies leading to the doctorate in the field are under development in several places.

This brief inventory of the results of efforts over a decade, limited to the salient aspects of the work done, suffices to show the growing importance of machine translation. After ten significant years it is worthwhile therefore to attempt to situate MT research in the general field of which it is a part—that of automatic language data processing, to see what experience appears to teach us, and where we may expect to go in the immediate future.

Natural languages, unlike conventional sign systems such as those of chemistry or mathematics, are culturally based. Natural languages are structured and can be and are being increasingly formally described. The systems of notation used to represent the nature and functions of language signs may be such as to make them appear to be mathematical formulas. But linguistics is obviously not a part of mathematics, though the algorithmic representation of some generalized usages or "rules" may give that impression. Also, in our language manuals, when we speak of "rules," we are in fact describing the habits confirmed by "proper" usage among the largest groups of users of a given language.[2]

MT is now directing its efforts to the translation of written languages. What this involves will be described in some detail below. As for the oral form, while some preliminary steps have been taken, we are still far from oriented research.

There are several ways in which we can think of machine translation. We can of course dream of a day when "perfect" translation from and into

[1] A new publication, "Information Processing Journal," contains in its first monthly issue seven abstracts of four articles and three books under the heading "Natural Languages, Linguistics and Mechanical Translation" (p. 39).

[2] See David G. Hays, *Research Procedures in Machine Translation,* Memorandum RM-2916-MR, December 1961, The Rand Corporation, introductory pages.

several languages will be obtained with speed or at a relatively low cost in almost any realm of language. This is a rather remote objective, though a valid one, provided the quest for high quality, wide diversity and great speed does not preclude the attainment of more modest, practical, useful results attainable in reasonable time and at acceptable cost. Of course, any early systems should be so conceived and constructed as to remain open-ended and likely to yield results applicable to the refinement and broadening of the initial procedures.

We should accept the fact that "perfect" translation is neither *humanly nor mechanically* achievable. We should likewise recognize that although a so-called word-for-word translation can be carried out on a computer with relative ease, it is of no practical value. What then should we aim for? The experience of the last decade seems to argue for focussed and coordinated programs extending over two or three years for one-directional translation involving two languages, and in restricted areas such as the physical or natural sciences, general technology, economics, law, engineering, etc. These limited programs should strive for maximum results in terms not only of usability, but of revealing unanticipated problems, confirming the validity of initial procedures, and bringing nearer the general solution of only partially resolved problems.

In terms of practical use, what is a reasonable norm of acceptability for machine output attainable in the near future? Can we accept the following standard?

> The output text should convey the same information as the input text; if it describes a chemical experiment, a chemist should be able to read the translation and reproduce the experiment with no more difficulty than if he had read the original report. Moreover, he should be able to read the translation as easily as if it had been written by a person fluent in the output language—Russian documents should be translated into versions that might have been written by Americans, for example.[3]

Isn't this in fact requiring of the machine more than is generally performed by human translators? There is nothing wrong with this objective in terms of long-range efforts and those who are familiar with the vagaries of human translators look forward to some day when the machine may do better.

A more immediately achievable standard of practical performance is given by Ida Rhodes of the National Bureau of Standards.

> If, for example, a translated article enables a scientist to reproduce an experiment described in a source paper and to obtain the same results—such a translation may be regarded as a practical one. Perhaps the translation is not couched in elegant terms; here and there several alternative meanings are given for a target word; a word or two may appear as a mere transliteration of original source words.

[3] David G. Hays, *ibid.*

Nevertheless, this translation has served its main purpose: a scholar in one land can follow the work of his colleague in another.[4]

This does not require of machine output the same ease of readability elegance of style or completeness of language transfer, but it does insist on reliability of information transfer.

Have we reached a point, or are we about to reach it, where the language of a scientifically creative culture can be processed on a computer so as to produce an intelligible, reliable, and therefore acceptable, output so that "a scholar in one land can follow the work of his colleague in another"?

The following pages present a machine translated section of a Russian book in the field of cybernetics aimed at the general reader.[6] From the 120-page text we have reproduced only that section which deals with machine translation. The book was translated into English on an IBM 7090 computer on the basis of the current Georgetown MT program, and this is a second cyclical run, i.e., it includes lexical and structural improvements based on the review of a first output.

The reason we are reproducing a section from this book is that it is by and large nontechnical. Another reason is that this chapter deals with the subject of MT.

It will be seen that the text is presented in two columns. The column on the left contains every detail of the output as it actually was produced by the computer. The column on the right presents the same text with a minimum of editing to facilitate reading. A study of both columns will, we believe, show that the translation adheres to the standard of reliability of information-transfer mentioned above. The lexical and structural inadequacies which are readily noticeable in the left-hand column, and which have in part been eliminated in the human-revised version on the right-hand column, are at present the subject of careful study. As a result of this review, the program will again be modified to incorporate generalized solutions to permit a higher level of output quality. This is the principle of cyclical-improvement procedure which will be described in more detail later on. Thus, by comparing the quality level of the output of a given text at a certain phase of production with the same text printed out on the basis of a partially improved, and repeatedly improved program, it is possible to demonstrate two things to which again we will refer in greater detail later: (1) that the system is constituted so as to remain open-ended, and (2) that relatively short time focused on improvement research permits the up-grading of the output product in a significant manner.

[4] Ida Rhodes, "A New Approach to the Mechanical Syntactic Analysis of Russian," *Mechanical Translation,* November 1961.

[5] The original Russian text of this passage is reproduced as appendix I to this paper. It it taken from Z. Rovensky, A. Uemov and E. Uemova "Mashina i Mysl' " (Machine and Thought), State Publishing House, Moscow 1960.

A machine translation

By by one from the first practical applications of logcal capabilities of machines was their utilization for the translation of texts from an one tongue on other. Linguistic differences represent the serious hindrance on a way for the development of cultural, social - political and scientific connections between nations. Automation of the process of a translation, the application of machines, with a help which possible to effect a translation without a knowledge of the corresponding foreign tongue, would be by an important step forward in the decision of this problem.

A machine, as already repeatedly stressed, does not penetrate into a sense being carried out by it the operations, mashinaperevodchik, in particular, cannot be attracted to a to a content being converted by it a text; it operates only pure by formal relations. The effecting of an automatic translation from an one tongue on other supposes a composition such the programs, in which the agreement between by both tongues represented in the form of *a* system strict formal relations, which were installed on the basis of structural analysis of this and other tongue, (page 88) Assumptions for such an analysis, principal possibilities the establishments of an abstract system formal of agreements in tongues, as was showed above, have.

On the example of a translation of the phrase from a greek tongue on latin we saw, by which way possible in the principle to convert from an one tongue on other, not knowing any one from them. But there we had fact only with by an one proposition and with the dictionary, which consists in all from some 10 elements. For the translation of texts of more solid dimensions, naturally, necessary the dictionary of agreements, significantly more voluminous and complex. The composition of such dictionaries for whole tongues would be a problem practically unrealizable, if there were no

Machine Translation

One of the first practical applications of logical capabilities of machines was their utilization for the translation of texts from one tongue on other. Linguistic differences present a serious hindrance on the way for the development of cultural, social - political and scientific relations between nations. Automation of the process of translation, the application of machines, with the help of which it is possible to effect a translation without knowledge of the corresponding foreign tongue, would be an important step forward in the resolution of this problem.

A machine, as already repeatedly stressed, does not penetrate into the sense carried out |by it of the operations. Machine translator, in particular, cannot handle the content converted by |it |of a text |; it operates purely by formal relations. The effecting of an automatic translation from one tongue on other supposes the composition of such program, in which the agreement between both tongues represented in the form of a system of strict formal relations, which were established on the basis of structural analysis of this and the other tongue, (page 88) Assumptions for such an analysis, the principal possibilities of the establishment of an abstract system of formal agreements in tongues, as was showed above exist.

On the example of a translation of the phrase from Greek tongue on Latin we saw, in which way it is possible in principle to convert from one tongue on other, not knowing either one of them. But there we dealt only with one sentence and with the dictionary, which consists in all of some 10 elements. For the translation of texts of larger dimensions, naturally, is necessary a dictionary of agreements, significantly more voluminous and complex. The composition of such dictionaries for whole tongues would be a problem practically unrealizable, if there were no such possibilities of

such possibilities of a formalization of a tongue, concerning which spoke above.

On the basis of detachment from material, the lexical values of the words, there are installed the formal - grammatical agreements between by tongues. This makes it possible to constitute the separate dictionary of the grammatical indexes — the kind, the numbers, the case, a time, parts of a speech etc., instead of this to introduce in the program of a machine all of the words with by the corresponding attachments, by the suffixes and by completions.

From an other standpoint, the separation of the prefixes, the suffixes and completions permits isolating in a word this its part, which is kept upon all its modifications and contains the lexical value of the given word.

All this significantly reduces the volume of the memory apparatus, necessary for an automatic dictionary. . . .

The selection of the words for a dictionary produces on the basis of statistical calculation of their use - in a tongue in general language, if constitutes a general dictionary, and in given to the branch of a science or a technology upon the preparation of special dictionaries.

Itself is understood, the words of a converted text cannot introduce into a machine in the form of combinations known to us the letters of a written tongue. For them necessary the special system of designations, a special code, "clear" to a machine. The symbols of this code must correspond to the elements of a tongue, not changing the sense of the latter. As far as for the logical operations, and also for a translation, are used the usual electronic - numerical machines, which realize different actions over 1 and also 0, combinations of these two symbols express upon the composition of the program for the machine - trans-lator also and a word of a tongue. Each basis, the each prefix, suffix, completion and also any other introduced into a

formalization of a tongue, concerning which we spoke above.

On the basis of detachment from basic lexical values of the words, there are established the formal-grammatical agreements between_tongues. This makes it possible to constitute the separate dictionary of the grammatical indexes— the gender, the number, the case, tense, parts of_speech etc., rather than to introduce in the program of the machine all of the words with_corresponding attachments,_the suffixes and_ endings.

From another standpoint, the separation of the prefixes, the suffixes and endings permits isolating in a word the part which is kept in all its modifications and contains the lexical value of the given word.

All this significantly reduces the volume of the memory apparatus, necessary for an automatic dictionary. . . .

The selection of the words for a dictionary is produced on the basis of statistical calculation of their use_in general language, if_a general dictionary is made, and in a given_branch of a science or_technology for the preparation of special dictionaries.

It is understood, the words of a translated text cannot be introduced into a machine in the form of combinations known to us as the letters of the written tongue. For them is necessary the special system of designations, a special code, "clear" to a machine. The symbols of this code must correspond to the elements of a tongue, not changing the sense of the latter. For the logical operations, and also for a translation, are used the usual electronic-numerical machines, which realize different actions over 1 and_0, combinations of these two symbols are expressed upon the composition of the program for the machine-translator, and also the words_of the tongue. Each base,_each prefix, suffix, ending and

"memory" a machine an element is compared with the definite combination of units and zeroes. These combinations are transferred on the punched tape in the form of alternations of holes and gaps, which as a result of a series of closings and also openings are converted in the corresponding combination of conducting and also nonconducting states of electronic lamps. Upon the help of a whole series of a relay symbols on a punched tape are compared with the dictionary, which are found in "a memory" a machine. Comparison occurs in the form of a subtraction of dictionary combinations from combinations on a punched tape.

If compare words do not coincide, as a result this operation receives which anyone a number, but not 0. In this case occurs switchin on the following word of a dictionary, and so up to these being time, meanwhile upon a subtraction does not receive 0. 0 signifies, that a machine finding in a dictionary a combination, equal with given. Now necessary to know, which corresponds to it in by friend a tongue. Side-by-side with each word of a converted tongue is indicated the number of the cell, containing the corresponding combination of this tongue, on which convert. A when subtraction gives as a result 0, switching occurs already not on the following word of a dictionary, but on this cell of the second tongue, a number which is side-by-side with by the given word, (page 90) A combination of states of the apparatus "of a memory", concluded in this cell, gives upon an exit definite the alternation of holes and gaps on a blade, which is converted then on the usual tongue of letters. The when memory apparatus contains not the wholly words, but their bases and the grammatical indexes, then the machine seeks at first in the dictionary of the bases the maximum combination, which agrees with the first part of the given words, but then in the dictionary of the suffixes and also completions finds other its also any other element introduced into the "memory" of the machine is compared with the definite combination of units and zeroes. These combinations are trans- (105)
ferred on the punched tape in the form of alternations of holes and gaps, which as a result of a series of closings and openings are converted in the corresponding combination of conducting and noncon- (110)
ducting states of electronic lamps. With the help of a whole series of a relay symbols on a punched tape are compared with the dictionary, which are found in the "memory" of the machine. Comparison occurs in (115)
the form of a subtraction of dictionary combinations from combinations on a punched tape.

If compared words do not coincide, as a result of this operation a number (120)
is obtained, but not 0. In this case occurs switching to the following word of dictionary, and so on until a subtraction does not receive 0. 0 signifies that a machine finds in dictionary a combination (125)
equal with given. Now it is necessary to know, what corresponds to it in another tongue. Side-by-side with each word of a converted tongue is indicated the number of the cell containing the (130)
corresponding combination of the tongue, into which to translate. When a subtraction gives as a result 0, switching occurs no longer to the following word of dictionary, but on this cell of the (135)
second tongue the number of which is side-by-side with the given word. (page 90) A combination of states of the apparatus of the "memory" included in this cell, gives as an exit a definite alter- (140)
nation of holes and gaps on a tape, which is converted then in the usual language of letters. When the memory apparatus contains not the complete words, but their bases and the grammatical indexes, (145)
then the machine seeks at first in the dictionary of the bases the maximum combination, which agrees with the first part of the given words, and then in the (150)
dictionary of the suffixes and endings it finds its other part. For instance, on the punched tape of a machine converting

part. We permit, on the punched tape of a machine, converting from an english tongue on russian, perforated a word "letterless". In the dictionary of the bases turn out to be words "summer, Lett", "letter". A machine stops only on the latter, as far as it coincides with by a maximum part of the given in word and gives its translation: a letter, a learning, literacy. Then searches for a value which left to a part of a word - - less, indicating a negation, as a result of which on an exit is obtained a russian word "an ignorance", "illiteracy".

Thus, all of the process of a translation dismembers on the totality of the simplest problems, as this is done/ made upon the completion of the logical and arithmetical operations. The number of these problems is very great, but upon this rate, a which a machine resolves each from them, a translation is effected very rapid.

The main hindrance for the wide application of translating machines are the difficulties of not no technical, but linguistic order, connected not so many with the completion of the program, how many with by its composition. These difficulties are not exhausted by abundance of bases and grammatical resources of an expression in each tongue. Fact is complicated by this, that, on the one hand, the elements of a tongue often have not one, but somewhat different values, from an other standpoint, one and the same lexical sense can express by the different bases, one and the same grammatical value - by different formal resources. . . .

But the main difficulty is created by this circumstance, that between the elements of different tongues there is no univocal agreement. To a basis or to the suffix, univocal in an one tongue, can correspond in by friend elements, which exist besides given still somewhat values. The russian to word "a month" In english, german, french and other tongues correspond two words, one from which there indicates "a moon", second - • "12 • ya a part of a year". Russian combinations with the pretext "to" and

from English tongue to Russian is perforated the word "letterless." In the dictionary (155)
of the bases turn out to be words "let", "Lett" and "letter." A machine stops only on the latter, as far as it coincides with a maximum part of the given word and gives its translation: _letter,_learning, literacy. (160)
Then searches for a value of the part of word left - less, indicating a negation, as a result of which in the output is obtained Russian word "ignorance", "illiteracy". (165)

Thus, all_the process of translation is dismembered into the totality of the simplest problems, as this is done_upon the (170)
completion of the logical and arithmetical operations. The number of these problems is very great, but at the rate at which a machine resolves each from them a translation is effected very rapidly. (175)

The main hindrance for the wide application of translating machines are the difficulties |of |not technical but linguistic order, connected not so much with the completion of the program as (180)
with_its composition. These difficulties are not exhausted by abundance of bases and grammatical resources of_expression in each tongue. Fact is complicated by this_that on the one hand, the elements (185)
of a tongue often have not one, but several different values; from an other standpoint, one and the same lexical sense can be expressed by_different bases, one and the same grammatical value — by (190)
different formal resources. . . .

But the main difficulty is created by this circumstance, that between the elements of different tongues there is no univocal agreement. To a basic or to the (195)
suffix, univocal in one tongue, can correspond in_ another elements which have besides the given one, several other values. To the Russian word_"month" in English, German, French and other tongues (200)
correspond two words, one of which indicates "moon", second — "12th part of a year". Russian combinations with the preposition "k" and the dative case

the dative case without the pretext equal transfer in english with by the help of the pretext "this". In a russian tongue exist the categories of the kind and the case, into english them not, and has an article, which is absent in a russian tongue.

All this requires the developments of a special system of indicators for dictionaries, which introduce into the memory apparatus of an automatic translator.

without the preposition equal transfer in (205)
English with_the help of the preposition
"to". In Russian tongue exist the
categories of the gender and the case, in
English_not, but it has an article,
which is absent in Russian tongue. (210)

All this requires the development _
of special system of indicators for
dictionaries, which are introduced into
the memory apparatus of an automatic (215)
translator.

|Here the authors wander into the realm of words, clusters called idioms, to no lucid results as far as the automatic translation of the passage is concerned.]

To effect automation of a translation with the calculation all of these peculiarities of a tongue by the help of the usual dictionary of the bases and completions, of course, impossible. Such a dictionary can satisfy only in the case of a translation special the selected text which consists exceptional or preferred from univocal words. In order a machine could convert any text, necessary the formalization of all elements of a tongue, and also these, a sense which depends on their surrounding, from the context. In this connection rises a question concerning the creation of such dictionaries, in which there would enter not only separate words, but phrases and whole propositions, that, of course, is a problem significantly more complex, than the composition of dictionaries of the separate words or their bases and completions, (page 92)

A tendency to simplify work, which was connected with automation of a translation, led to an idea concerning the artificial tongue - middleman, free from a multiple validity, tyue idiomatiki and to this of similar phenomena, which hamper and even, possibly, excluding a translation without the understanding of sense converted. To establish the formal relations between two tongues, from which at least one strict logical and does not require the

To effect automation of a translation
with the calculation |all| of these
peculiarities of a tongue with the help of
the usual dictionary of the bases and (220)
endings, of course, is impossible.
Such a dictionary can satisfy only in the
case of a translation of a specially selected
text, which consists exclusively or
mainly of monovalent words. In order (225)
a machine could convert any text, necessary
the formalization of all elements of a
tongue, and also these, the sense of which
depends on their surrounding, or context.
In this connection a rises question con- (230)
cerning the creation of such dictionaries,
in which there would enter not only words,
but phrases and whole sentences. That,
of course, is a problem significantly
more complex, than the composition of (235)
dictionaries of the separate words or
their bases and endings. (page 92)

A tendency to simplify work, which
was connected with automation of_trans- (240)
lation, led to an idea concerning the
artificial tongue—intermediary free
from_multiple value, idiomatics and of
similar phenomena, which hamper and even,
possibly, exclude translation without the (245)
understanding of sense translated. To
establish the formal relations between two
tongues, from which at least one is
strictly logical and does not require the
special study and calculation of the (250)

special study and the calculation of the different kind of exceptions, divergence from formal rules, of course, significantly more easily, than between by the two usual tongues. Besides this, the presence of such a tongue in a many time would decrease the number of the necessary dictionaries, if rises a question concerning a translation from any tongue on any. We take at least 10 tongues. In such case would require 90 dictionaries as far as each tongue must be converted on 9 other, while by the presence of the tongue - middleman would be sufficient 20: 10 - for a translation on the tongue - middleman and 10 - for a back translation on any from these tongues. In this connection arguments the creation of the tongue - middleman is completely feasible.

different kind of exceptions, divergence from formal rules, of course, is significantly more easy, than between two natural tongues. Besides this, the presence of such a tongue many times (255) would decrease the number of dictionaries necessary, if a question arises concerning translation from any one tongue into any other. Let us take 10 tongues. In such case we would require 90 (260) dictionaries, since each tongue must be converted on 9 others, while by the presence of the intermediate tongue- would be sufficient 20: 10 - for a 2 x 10— translation to the |tongue| intermediate (265) and 10 — for a back translation to any of these tongues. In this connection arguments for the creation of the inter- mediate tongue - are completely feasible. (270)

A review of the right-hand column of the revised text, in which underscoring indicates the corrections, reveals that the largest number of errors (about one fourth of the total in the passage given) involves insertion of the definite or indefinite article in English. In spite of different approaches during the last four years, we have not yet been able to develop a reliable formal program for article insertion in Russian-English MT, nor do we know of any program on the article which has been tested. (Most of the material published on MT is in the form of theoretical studies rather than reports on actual computer experiments on specific or general procedures.) The second largest source of errors in the passage given involves the use of prepositions in general and in particular of the selection of "of" and "from" in English, as well as the use of the preposition "by" as an oversimplified procedure in the transfer of the Russian instrumental case. The third most frequent error occurs in lexical equivalents and a few lexical gaps. These inadequacies, and other deficiencies, will be eliminated to a considerable degree in a next cyclical run.

A separate random[6] text in the field of cybernetics was run recently against the program. Being more technical and therefore less complicated lexically and stylistically, the quality of the output is reported to be equal or superior to that of the book itself, of which the sample reproduced above is quite representative.

The processing of the signs of natural language by electronic computers for translation of a given source language into a chosen target language presents problems basically different from those presented by the automatic processing of conventional language signs, such as those of mathematics or chemistry. When in the field of numbers, for instance, we deal with different systems of representation such as the decimal or duodecimal, the binary or even the Morse system (essentially binary) we have obviously a fixed and generalized transfer of relationships between the different systems of signs. This is not true, or at least we have not yet established such correlation between any two natural language systems.

The signs (in natural language can be characterized as "unspecific" or "unstable" in the sense that a given letter or groups of letters between spaces (a word) will carry different information in terms of the language in which it operates, and also in terms of its possible equivalents in another language. Contextual factors, either phrase groups, sentences, or larger discourse units, or again in the sense of different fields or disciplines, are the basis for the reduction or elimination of ambiguity. Further, communication through natural languages involves in effect the

[6] By "random text" we mean a passage or article in a given discipline which is run without previous lexical or structural study or abstraction for incorporation in the program. A copy of this "random text" is available on request.

obtaining of information from signs, oral or written, which are "known" to the hearer or reader. In a monolingual situation, different levels or fields of discourse are accessible only to a person knowing the fields and level of discourse. In natural languages (and for that matter in conventional sign systems) the two basic means for the discernment of information are the forms of the signs and their distribution.

In the translating of one language into another, we are confronted with two distinct sign structures. Our problem is one of transfer. This transfer means that the information in a source text will be represented in another language form or text. What is involved in this transfer when we speak of machine translation? Some years back I attempted an explanation to the effect that MT involved the transfer of meaning by computers from the signs of the source language to those of the target. It is more accurate to describe the operation as involving the systematic substitution by computing devices of the signs of the target language for those of the source language, with the obvious aim of maximum information transfer.

Since we are dealing, for the present at least, with two structures, and since we are seeking to effect the systematic substitution of signs, it follows that we must establish as complete a correlation or correspondence between the structures involved as is required for effective information transfer. Thus we may say that MT research is, or should be, oriented to a specialized area of linguistic investigation, which is called transfer linguistics. While this may seem obvious to some, considerable research effort has been aimed primarily at intrinsic or monostructural investigation rather than at bistructural transfer programs. The thesis has been upheld by many that a "complete" analysis of a given structure should precede any approach to the problem of transfer, and that exhaustive monostructural automation is a prerequisite to systematic bi- or multilingual transfer. More pragmatic and empirical procedures are being gradually developed by other groups with reasonably promising results and prospects in terms of relatively modest and immediately practical aims. It is with this approach that we shall be concerned in the remaining part of this study.

Машинный перевод,

Одним из первых практических применений логических способностей машин явилось использование их для перевода текстов с одного языка на другой. Языковые различия представляет серьезное препятствие на пути к развитию культурных, общественно-политических и научных связей между народами. Автоматизация процесса перевода, применение машин, с помощью которых можно осуществлять перевод без знания соответствующего иностранного языка, было бы важным шагом вперед в решении этой проблемы.

Машина, как уже неоднократно подчеркивалось, не вникает в смысл производимых ею операций. Машина-переводчик, в частности, не может обращаться к содержанию переводимого ею текста; она оперирует только чисто формальными отношениями. Осуществление автоматического перевода с одного языка на другой предполагает составление такой программы, в которой соответствие между обоими языками представлено в виде системы строго формальных соотношений, установленных на основе структурного анализа того и другого языка. Предпосылки для такого анализа, принципиальные возможности установления абстрактной системы формальных соответствий в языках, как было показано выше, имеются.

На примере перевода фразы с греческого языка на латинский мы видели, каким образом можно в принципе переводить с одного языка на другой, не зная ни одного из них. Но там мы имели дело только с одним предложением и со словарем, состоящим всего-навсего из каких-нибудь 10

элементов. Для перевода текстов более солидных размеров, естественно, необходим словарь соответствий, значительно более объемистый и сложный. Состазление таких словарей для целых языков было бы задачей практически неосуществимой, если бы не было таких возможностей формализации языка, о которых говорилось выше.

На основе отвлечения от вещественных, лексических значений слов устанавливаются формально-грамматические соответствия между языками. Это дает возможность составить отдельный словарь грамматических показателей - рода, числа, падежа, времени, частей речи и т.д., вместо того чтобы вводить в программу все слова с соответствующими приставками, суффиксами и окончаниями.

С другой стороны, отделение префиксов, суффиксов и окончаний позволяет выделить в слове ту его часть, которая сохраняется при всех его видоизменениях и заключает в себе лексическое значение данного слова.

Все это значительно сокращает объем запоминающего устройства, необходимого для автоматического словаря. Например, мы имеем такой ряд слов: строить, строение, строящий, строивший, настроить, настроив, настроивший, настроенный, построить, построив, построивший, построенный, устроить, устроив, устроивший, устроенный, застроить, застроив, застроивший, застроенный. Здесь далеко не все возможные производные от слова "строить", но мы ограничимся хотя бы этими. Аналогичные проиаводные можно образовать и от других подобных глаголов, например "говорить", "солить" и т.д. Если бы глаголы "строить", "говорить", "со-

лить" и их производные ввести в словарь целиком, это составило бы более 70 слов. Между тем, в формализованном виде они будут выражены 16 элементами: 3 основами (стро-, говор-, сол-), б суффиксами (-ить, -енн-, -ящ-, -ив-, -ибш-, -енн-), 2 окончаниями (-е, -ий), 5 префиксами (на-, под-, у-, за-). Можно себе представить, во сколько раз упростится словарь, если речь пойдет не о трех, а о тысяче или более слов.

Отбор слов для словаря производится на основе статистического подсчета их употребительности - в языке вообще, если составляется общий словарь, и в данной отрасли науки или техники при подготовке специальных словарей.

Само собой разумеется, слова переводимого текста не могут вводиться в машину в виде сочетаний известных нам букв письменного языка. Для них нужна особая система обозначений, специальный код, "понятный" машине. Знаки этого кода должны соответствовать элементам языка, не меняя смысла последних. Поскольку для логических операций, в том числе и для перевода, используются обычные электронно-цифоровые машины, осуществляющие различные действия над 1 и 0, комбинациями этих двух знаков выражаются при составлении программы для машины-переводчика также и слова языка. Каждая основа, каждый префикс, суффикс, окончание и любой другой вводимый в "память" машины элемент сопоставляется с определенным сочетанием единиц и нулей. Эти сочетания переносятся на перфоленту в виде чередований отверстий и пропусков, которые в результате ряда замыканий и размыканий преобразуются в соответствующие комби-

нации проводящих и непроводящих состояний электронных ламп. При помощи целого ряда реле знаки на перфоленте сравниваются со словарем, находящимся в "памяти" машины. Сравнение происходит в виде вычитания словарных комбинаций из комбинаций на перфоленте.

Если сравниваемые слова не совпадают, в результате этой операции получится какое угодно число, но не 0. В этом случае происходит переключение на следующее слово словаря, и так до тех пор, пока при вычитании не получится О. О означает, что машина нашла в словаре комбинацию, одинаковую с данной. Теперь нужно узнать, что соответствует ей в другом языке. Рядом с каждым словом переводимого языка указывается номер ячейки, содержащий соответствующую комбинацию того языка, на который переводят. Когда вычитание дает в результате 0, переключение происходит ухе не на следующее слово словаря, а на ту ячейку второго языка, номер которой стоит рядом с данным словом. Комбинация состояний устройства "памяти", заключенная в этой ячейке, дает при выходе определенное чередование отверстий и пропусков на ленте, которое переводится затем на обычный язык букв. Когда запоминающее устройство содержит не целиком слова, а их основы и грамматические показатели, тогда машина ищет сначала в словаре основ наибольшую комбинацию, совпадающую с первой частью данного слова, а затем в словаре суффиксов и окончаний находит остальную его часть. Допустим, на перфоленте машины, переводящей с английского языка на русский, пробито слово "letterless ." В словаре основ оказываются слова "let," "Lett", "letter." Машина останавливается только на последнем, поскольку оно совпадает с наибольшей частью данного слова и дает его перевод:

буква, ученость, грамотность. Затем отыскивается значение оставшейся части слова — -less , обозначающей отрицание, в результате чего на выходе получается русское слово "неученность", "неграмотность."

Таким образом, весь процесс перевода расчленяется на совокупность простейших задач, подобно тому, как это делается при выполнении логических и арифметических операций. Число этих задач очень велико, но при той скорости с которой машина решает каждую из них, перевод осуществляется очень быстро.

Главным препятствием для широкого применения переводческих машин являются затруднения не технического, а лингвистического порядка, связанные не столько с выполнением программы, сколько с ее состаявлением. Эти трудности не исчерпываются изобилием основ и грамматических средств выражения в каждом языке. Дело осложняется тем, что, с одной стороны, элементы языка часто имеют не одно, а несколько разных значений, с другой стороны, один и тот же лексический смысл может выражаться разными основами, одно и то же грамматическое значение--разными формальными средствами.

Но главное затруднение создается тем обстоятельством, что между элементами разных языков нет однозначного соответствия. Основе или суффиксу, однозначном в одном языке, могут соответствовать в другом элементы, имеющие кроме данного еще несколько значений. Русскому слову "месяц" в английском, немецком, французском и других языках соответствуют два слова, одно из которых обозначает "луна", второе-"12-я часть года". Русские сочетания с предлогом "к" и дательный падеж без предлога одинаково передаются в английском с помощью предлога "to". В русском языке существуют категории рода и падежа, в англий-

ском их нет, а имеется артикль, отсутствующий в русской языке.

Все это требует разработки особой системы помет для словарей, которые вводятся в запоминающее устройство автоматического переводчика.

Во всех языках имеется много так называемых идиом, не допускающих буквального перевода. Например, французское выражение, соответствующее русскому "Он вылитый отец", при буквальном переводе дает: "Он тец, совершенно выплюнутый"; мы говорим: "С глазу на глаз", а французы в том же смысле скажут: "Голова к голове", немцы: "Между четырьмя глазами", англичане: "Лицо к лицу."

Осуществить автоматизацию перевода с учетом всех этих особенностей языка ири помощи обычного словаря основ и окончаний, конечно, невозможно. Такой словарь может удовлетворить лишь в случае перевода специально подобранного текста, состоящего исключительно или преимущественно из однозначных слов. Чтобы машина могла переводить любой текст, необходима формализация всех элементов языка, в том числе и тех, смысл которых зависит от их окружения, от контекста. В связи с этим встает вопрос о создании таких словарей, в которые входили бы не только отдельные слова, но и словосочетания и целые предложения, что, конечно, представляет собой задачу значительно более сложную, чем составление словарей отдельных слов или их основ и окончаний.

Стремление упростить работу, связанную с автоматизацией перевода, привело к идее об искусственном языке-посреднике, свободном от многозначности, идиоматики и тому подобных явлений, затрудняющих и

даже, возможно, исключающих перевод без понимания смысла переводимого. Установить формальные соотношения между двумя языками, из которых хотя бы один строго логичен и не требует специального изучения и учета разного рода исключений, отклонений от формальных правил, конечно, значительно легче, чем между двумя обычными языками. кроме того, наличие такого языка во много раз уменьшило бы число необходимых словарей, если встанет вопрос о переводе с любого языка на любой. Возьмем хотя бы 10 языков. В таком случае потребовалось бы 90 словарей, поскольку каждый язык должен переводиться на 9 остальных, тогда как при наличии языка-посредника было бы достаточно 20: 10— для перевода на любой из этих языков. В связи с этими соображениями создание языка-посредника представляется вполне целесообразным.